Learn to Accumulate Evidence from All Training Samples: Theory and Practice

Deep Pandey [1] **Qi Yu** [1]

Abstract

Evidential deep learning, built upon belief theory and subjective logic, offers a principled and computationally efficient way to turn a deterministic neural network uncertainty-aware. The resultant evidential models can quantify fine-grained uncertainty using the learned evidence. To ensure theoretically sound evidential models, the evidence needs to be non-negative, which requires special activation functions for model training and inference. This constraint often leads to inferior predictive performance compared to standard softmax models, making it challenging to extend them to many large-scale datasets. To unveil the real cause of this undesired behavior, we theoretically investigate evidential models and identify a fundamental limitation that explains the inferior performance: existing evidential activation functions create *zero evidence regions*, which prevent the model to learn from training samples falling into such regions. A deeper analysis of evidential activation functions based on our theoretical underpinning inspires the design of a novel regularizer that effectively alleviates this fundamental limitation. Extensive experiments over many challenging real-world datasets and settings confirm our theoretical findings and demonstrate the effectiveness of our proposed approach.

1. Introduction

Deep Learning (DL) models have found great success in many real-world applications such as speech recognition (Kamath et al., 2019), machine translation (Singh et al., 2017), and computer vision (Voulodimos et al., 2018). However, these highly expressive models may easily fit the noise in the training data, which leads to overconfident predictions (Nguyen et al., 2015). The challenge is further compounded when learning from limited labeled data, which is common for applications from specialized domain (*e.g.,* medicine, public safety, and military operations) where data collection and annotation is highly costly. Accurate uncertainty quantification is essential for successful application of DL models in these domains. To this end, DL models have been augmented to become uncertainty-aware (Gal & Ghahramani, 2016; Blundell et al., 2015; Pearce et al., 2020). However, commonly used extensions require expensive sampling operations (Gal & Ghahramani, 2016; Blundell et al., 2015), which significantly increase the computational costs (Lakshminarayanan et al., 2017).

The recently developed evidential models bring together evidential theory (Shafer, 1976; Jøsang, 2016) and deep neural architectures that turn a deterministic neural network uncertainty-aware. By leveraging the learned evidence, evidential models are capable of quantifying fine-grained uncertainty that helps to identify the sources of 'unknowns'. Furthermore, since only lightweight modifications are introduced to existing DL architectures, additional computational costs remain minimum. Such evidential models have been successfully extended to classification (Sensoy et al., 2018), regression (Amini et al., 2020), meta-learning (Pandey & Yu, 2022a), and open-set recognition (Bao et al., 2021) settings.

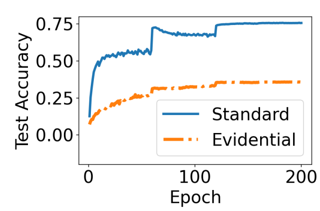


Figure 1. Cifar100 Result

Despite the attractive uncertainty quantification capacity, evidential models are only able to achieve a predictive performance on par with standard deep architectures in relatively simple learning problems. They suffer from a significant performance drop when facing large datasets with more complex features even in the common classification setting. As shown in Figure 1, an evidential model using ReLU activation and an evidential MSE loss (Sensoy et al., 2018) only achieves 36% test accuracy on Cifar100, which is almost 40% lower than a standard model trained using softmax. Additionally, most evidential models can easily break down with minor architecture changes and/or have a much stronger dependency on hyperparameter tuning to achieve reasonable predictive performance. The experiment section provides more details on these failure cases.

[1]Rochester Institute of Technology. Correspondence to: Qi Yu <qi.yu@rit.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

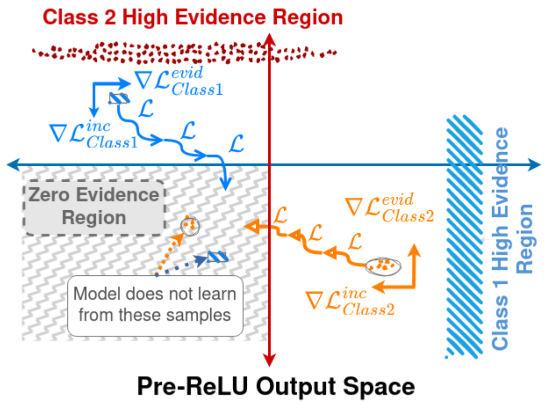


Figure 2. Visualization of zero-evidence region for evidential models with ReLU activation in a binary classification setting. Existing models fail to learn from samples that are mapped to such zero-evidence region (shared area at the bottom left quadrant).

To train uncertainty-aware evidential models that can also predict well, we perform a novel theoretical analysis with a focus on the standard classification setting to unveil the underlying cause of the performance gap. Our theoretical results show that existing evidential models learn sub-optimally compared to corresponding softmax counterparts. Such sub-optimal training is mainly attributed to the inherent *learning deficiency* of evidential models that prevents them from learning across all training samples. More specifically, they are incapable to acquire new knowledge from training samples mapped to "zero-evidence regions" in the evidence space, where the predicted evidence reduces to zero. The sub-optimal learning phenomenon is illustrated in Figure 2 (detailed discussion is presented in Section 4.2). We analyze different variants of evidential models present in the existing literature and observe this limitation across all the models and settings. Our theoretical results inspire the design of a novel **R**egularized **E**vidential mo**d**el (**RED**) that includes positive evidence regularization in its training objective to battle the learning deficiency. Our major contributions can be summarized as follows:

- We identify a fundamental limitation of evidential models, *i.e.,* lack the capability to learn from any data samples that lie in the "zero-evidence" region in the evidence space.
- We theoretically show the superiority of evidential models with $\exp$ activation over other activation functions.
- We conduct novel evidence regularization that enables evidential models to avoid the "zero-evidence" region so that they can effectively learn from all training samples.
- We carry out experiments over multiple challenging real-world datasets to empirically validate the presented theory, and show the effectiveness of our proposed ideas.

2. Related Works

Uncertainty Quantification in Deep Learning. Accurate quantification of predictive uncertainty is essential for development of trustworthy Deep Learning (DL) models. Deep ensemble techniques (Pearce et al., 2020; Lakshminarayanan et al., 2017) have been developed for uncertainty quantification. An ensemble of neural networks is constructed and the agreement/disagreement across the ensemble components is used to quantify different uncertainties. Ensemble-based methods significantly increase the number of model parameters, which are computationally expensive at both training and test times. Alternatively, Bayesian neural networks (Gal & Ghahramani, 2016)(Blundell et al., 2015)(Mobiny et al., 2021) have been developed that consider a Bayesian formalism to quantify different uncertainties. For instance, (Blundell et al., 2015) use Bayes-by-backdrop to learn a distribution over neural network parameters, whereas (Gal & Ghahramani, 2016) enable dropout during inference phase to obtain predictive uncertainty. Bayesian methods resort to some form of approximation to address the intractability issue in marginalization of latent variables. Moreover, these methods are also computationally expensive as they require sampling for uncertainty quantification.

Evidential Deep Learning. Evidential models introduce a conjugate higher-order evidential prior for the likelihood distribution that enables the model to capture the fine-grained uncertainties. For instance, Dirichlet prior is introduced over the multinomial likelihood for evidential classification (Bao et al., 2021; Zhao et al., 2020), and NIG prior is introduced over the Gaussian likelihood (Amini et al., 2020; Pandey & Yu, 2022b) for the evidential regression models. Adversarial robustness (Kopetzki et al., 2021) and calibration (Tomani & Buettner, 2021) of evidential models have also been well studied. Usually, these models are trained with evidential losses in conjunction with heuristic evidence regularization to guide the uncertainty behavior (Pandey & Yu, 2022a; Shi et al., 2020) in addition to reasonable generalization performance. Some evidential models assume access to out-of-distribution data during training (Malinin & Gales, 2019; 2018) and use the OOD data to guide the uncertainty behavior. A recent survey (Ulmer, 2021) provides a thorough review of the evidential deep learning field.

In this work, we focus on evidential classification models and consider settings where no OOD data is used during model training to make the proposed approach more broadly applicable to practical real-world situations.

3. Learning Deficiency of Evidential Models

3.1. Preliminaries and problem setup

Standard classification models use a softmax transformation on the output from the neural network $\mathcal{F}_\Theta$ for input $\mathbf{x}$ to obtain the class probabilities in K-class classification problem. Such models are trained with the cross-entropy based loss.

For a given training sample $(\mathbf{x}, \mathbf{y})$, the loss is given by

$$\mathcal{L}_{cross} = -\sum_{k=1}^{K} \mathbf{y}_k \log(\mathtt{sm}_k) \tag{1}$$

where $\mathtt{sm}_k$ is the softmax output. These models have achieved state-of-the-art performance on many benchmark problems. A detailed gradient analysis shows that they can effectively learn from all training data samples (see Appendix A). Nevertheless, these models lack a systematic mechanism to quantify different sources of uncertainty, a highly desired property in many real-world problems.

Figure 3. Graphical model for Evidential Deep Learning

Evidential classification models formulate training as an evidence acquisition process and consider a higher-order Dirichlet prior $\mathtt{Dir}(\mathbf{p}|\boldsymbol{\alpha})$ over the predictive Multinomial distribution $\mathtt{Mult}(\mathbf{y}|\mathbf{p})$. Different from a standard Bayesian formulation which optimizes *Type II Maximum Likelihood* to learn the Dirichlet hyperparameter (Bishop & Nasrabadi, 2006), evidential models directly predict $\boldsymbol{\alpha}$ using data features $\mathbf{x}$ and then generate the prediction $\mathbf{y}$ by marginalizing the Multinomial parameter $\mathbf{p}$. Figure 3 describes this generative process. Such higher-order prior enables the model to systematically quantify different sources of uncertainty. In evidential models, the softmax layer of the standard neural networks is replaced by a non-negative activation function $\mathcal{A}$, where $\mathcal{A}(\mathbf{x}) \geq 0 \quad \forall x \in [-\infty, \infty]$, such that for input $\mathbf{x}$, the neural network model $\mathcal{F}_\Theta$ with parameters Θ can output evidence $\mathbf{e}$ for different classes. Dirichlet prior $\boldsymbol{\alpha}$ is evaluated as $\boldsymbol{\alpha} = \mathbf{e} + \mathbf{1}$ to ensure $\boldsymbol{\alpha} \geq 1$. The trained evidential model outputs Dirichlet parameters $\boldsymbol{\alpha}$ for input $\mathbf{x}$ that can quantify fine-grained uncertainties in addition to the prediction $\mathbf{y}$. Mathematically, for $K-$class classification problem,

$$\mathtt{Evidence}(\mathbf{e}) = \mathcal{A}(\mathcal{F}_\Theta(\mathbf{x})) = \mathcal{A}(\mathbf{o}) \tag{2}$$

$$\mathtt{Dirichlet\ Parameter}(\boldsymbol{\alpha}) = \mathbf{e} + \mathbf{1} \tag{3}$$

$$\mathtt{Dirichlet\ Strength}(S) = K + \sum_{k=1}^{K} \mathbf{e}_k \tag{4}$$

The activation function $\mathcal{A}(\cdot)$ assumes three common forms to transform the neural network output into evidence: (1) $\mathtt{ReLU}(\cdot) = \max(0, \cdot)$, (2) $\mathtt{SoftPlus}(\cdot) = \log(1 + \exp(\cdot))$, and (3) $\exp(\cdot)$.

Evidential models assign input sample to that class for which the output evidence is greatest. Moreover, they quantify the confidence in the prediction for K class classification problem through vacuity ν (*i.e.,* measure of lack of confidence in the prediction) computed as

$$\mathtt{Vacuity}(\nu) = \frac{K}{S} \tag{5}$$

For any training sample $(\mathbf{x}, \mathbf{y})$, the evidential models aim to maximize the evidence for the correct class, minimize the evidence for the incorrect classes, and output accurate confidence. To this end, three variants of evidential loss functions have been proposed (Sensoy et al., 2018): 1) Bayes risk with sum of squares loss, 2) Bayes risk with cross-entropy loss, and 3) Type II Maximum Likelihood loss. Please refer to equations (21), (22), and (23) in the Appendix for the specific forms of these losses. Additionally, incorrect evidence regularization terms are introduced to guide the model to output low evidence for classes other than the ground truth class (See Appendix C for discussion on the regularization). With evidential training, accurate evidential deep learning models are expected to output high evidence for the correct class, low evidence for all other classes, and output very high vacuity for unseen/out-of-distribution samples.

3.2. Theoretical Analysis of Learning Deficiency in Evidential Learning

To identify the underlying reason that causes the performance gap of evidential models as described earlier, we consider a K class classification problem and a representative evidential model trained using Bayes risk with sum of squares loss given in (21). We first provide an important definition that is critical for our theoretical analysis.

Definition 1 **(Zero-Evidence Region).** A *Zero-evidence sample* is a data sample for which the model outputs zero evidence for all classes. A region in the evidence space that contains *zero-evidence samples* is a *zero-evidence region*.

For a reasonable evidential model, novel data samples not yet seen during training, difficult data samples, and out-of-distribution samples should become zero-evidence samples.

Theorem 1. *Given a training sample* $(\mathbf{x}, \mathbf{y})$*, if an evidential neural network outputs zero evidence* $\mathbf{e}$*, then the gradients of the evidential loss evaluated on this training sample over the network parameters reduce to zero.*

Proof. Consider an input $\mathbf{x}$ with one-hot ground truth label $\mathbf{y}$. Let the ground truth class index be gt, *i.e.,* $y_{gt} = 1$, with corresponding Dirichlet parameter α_{gt}, and $y_{\neq gt} = 0$. Moreover, let $\mathbf{o}, \mathbf{e}$, and $\boldsymbol{\alpha}$ represent the neural network output vector before applying the activation $\mathcal{A}$, the evidence vector, and the Dirichlet parameters respectively.

In this evidential model, the loss is given by

$$\mathcal{L}^{\text{MSE}}(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{K} (y_j - \frac{\alpha_j}{S})^2 + \frac{\alpha_j(S - \alpha_j)}{S^2(S+1)} \tag{6}$$

Now, the gradient of the loss with respect to the neural network output can be computed using the chain rule:

$$\begin{aligned}\frac{\partial \mathcal{L}^{\text{MSE}}(\mathbf{x},\mathbf{y})}{\partial o_k} &= \frac{\partial \mathcal{L}^{\text{MSE}}(\mathbf{x},\mathbf{y})}{\partial \alpha_k}\frac{\partial e_k}{\partial o_k} \\ &= \Big[\frac{2\alpha_{gt}}{S^2} - 2\frac{y_k}{S} - \frac{2(S-\alpha_k)}{S(S+1)} + \\ &\quad + \frac{2(2S+1)\sum_i\sum_j \alpha_i\alpha_j}{(S^2+S)^2}\Big] \times \frac{\partial e_k}{\partial o_k}\end{aligned} \tag{7}$$

Based on the actual form of $\mathcal{A}$, we have three cases:

Case I: $\texttt{ReLU}(\cdot)$ to transform logits to evidence

$$e_k = \text{ReLU}(o_k) \implies \frac{\partial e_k}{\partial o_k} = \begin{cases} 1 & \text{if} \quad o_k > 0 \\ 0 & \text{otherwise}\end{cases} \tag{8}$$

For a zero-evidence sample, the logits o_k satisfy the relationship $o_k \leq 0 \;\forall\; k \implies \frac{\partial e_k}{\partial o_k} = 0 \implies \frac{\partial \mathcal{L}^{\text{MSE}}(\mathbf{x},\mathbf{y})}{\partial o_k} = 0$

Case II: $\texttt{SoftPlus}(\cdot)$ to transform logits to evidence

$$e_k = \log(\exp(o_k)+1) \implies \frac{\partial e_k}{\partial o_k} = \text{Sigmoid}(o_k) \tag{9}$$

For a zero-evidence sample, the logits $o_k \to -\infty \implies \text{Sigmoid}(o_k) \to 0$ & $\frac{\partial e_k}{\partial o_k} \to 0$.

Case III: $\exp(\cdot)$ to transform logits to evidence

$$e_k = \exp(o_k) \implies \frac{\partial e_k}{\partial o_k} = \exp(o_k) = \alpha_k - 1 \tag{10}$$

For a zero-evidence sample, $\alpha_k \to 1 \implies \frac{\partial e_k}{\partial o_k} \to 0$. Moreover, there is no term in the first part of the loss gradient in (7) to counterbalance these zero-approaching gradients. So, for *zero-evidence training samples*, for any node k,

$$\frac{\partial \mathcal{L}^{\text{MSE}}(\mathbf{x},\mathbf{y})}{\partial o_k} = 0 \tag{11}$$

Since the gradient of the loss with respect to all the nodes is zero, there is no update to the model from such samples. This implies that the evidential models fail to learn from a zero-evidence data sample. □

For completeness, we present the analysis of standard classification models in Appendix A, detailed proof of the evidential models trained using Bayes risk with sum of squares error along with other evidential lossses in Appendix B, and impact of incorrect evidence regularization in Appendix C.

Remark: Evidential models can not learn from a training sample that the model has never seen and for which the model accurately outputs "I don't know", *i.e.,* $e_k = 0 \quad \forall k \in [1, K]$. Such samples are expected and likely to be present during model training. However, the supervised information in such training data points is completely missed by evidential models so they fail to acquire any new knowledge from all such training data samples (*i.e.,* data samples in zero-evidence region of the evidence space).

Corollary 1. *Incorrect evidence regularization can not help evidential models learn from zero-evidence samples.*

Intuitively, the incorrect evidence regularization encourages the model to output zero evidence for all classes other than the ground truth class and the regularization does not have any impact on the evidence for the ground truth class. So, the regularization updates the model parameters such that the model is likely to map input samples closer to zero-evidence region in the evidence space. Thus, the regularization does not address the failure of evidential models to learn from zero evidence samples.

Theorem 2. *For a data sample* $\mathbf{x}$*, if an evidential model outputs logits* $\mathbf{o}_k \leq 0 \;\forall k \in [0, K]$*, the exponential activation function leads to a larger gradident update on the model parameters than* `softplus` *and* `ReLu`*.*

Limited by space, we present the proof of Theorem 2 along with additional analysis in the Appendix D. The proof follows the gradient analysis of the exponential, `Softplus`, and `ReLU` based models. It implies that the the training of evidential models is most effective with the exponential activation function. Intuitively, the `ReLU` based activation completely destroys all the information in the negative logits, and has largest region in evidence space in which training data have zero evidence. `Softplus` activation improves over the `ReLU`, and compared to `ReLU`, has smaller region in evidence space where training data have zero evidence. However, `Softplus` based evidential models fail to correct the acquired knowledge when the model has strong wrong evidence. Moreover, these models are likely to suffer from vanishing gradients problem when the number of classes increases (*i.e.,* classification problem becomes more challenging). Finally, exponential activation has the smallest zero-evidence region in the evidence space without suffering from the issues of `SoftPlus` based evidential models.

4. Avoiding Zero-Evidence Regions Through Correct Evidence Regularization

We now consider an evidential model with exponential function to transform the logits into evidence. We propose a novel vacuity-guided correct evidence regularization term

$$\mathcal{L}_{\texttt{cor}}(\boldsymbol{x},\boldsymbol{y}) = -\lambda_{\texttt{cor}}\log(\alpha_{gt}-1) \tag{12}$$

where $\lambda_{\texttt{cor}} = \nu = \frac{K}{S}$ represents the regularization term whose value is given by the magnitude of the vacuity output by the evidential model and $\alpha_{gt} - 1$ represents the predicted evidence for the ground truth class. The regularization term $\lambda_{\texttt{cor}}$ determines the relative importance of the correct evidence regularization term compared to the evidential loss and incorrect evidence regularization and is treated as constant during model parameter update.

Theorem 3. *Correct evidence regularization $\mathcal{L}_{\texttt{cor}}(\boldsymbol{x}, \boldsymbol{y})$ can address the issue of learning from zero-evidence training samples.*

Proof. The proposed regularization term $\mathcal{L}_{\texttt{cor}}(\boldsymbol{x}, \boldsymbol{y})$ does not contain any evidence terms other than the evidence for the ground truth node. So, the gradient of the regularization for nodes other than the ground truth node will be 0 i.e. $\frac{\partial \mathcal{L}_{\texttt{cor}}(\boldsymbol{x},\boldsymbol{y})}{\partial o_k}\Big|_{k \neq gt} = 0$ and there will be no update on these nodes. For the ground truth node $gt, y_{gt} = 1$, the gradient is given by

$$\frac{\partial \mathcal{L}_{\texttt{cor}}(\boldsymbol{x}, \boldsymbol{y})}{\partial o_{gt}} = \frac{\partial\big(-\lambda_{\texttt{cor}} \log(\alpha_{gt} - 1)\big)}{\partial o_{gt}} \tag{13}$$

$$= -\lambda_{\texttt{cor}} \frac{\partial \log(\alpha_{gt} - 1)}{\partial \alpha_{gt}} \times \frac{\partial \alpha_{gt}}{\partial o_{gt}} \tag{14}$$

$$= -\frac{\lambda_{\texttt{cor}}}{(\alpha_{gt} - 1)}(\alpha_{gt} - 1) = -\lambda_{\texttt{cor}} \tag{15}$$

The gradient value equals the magnitude of the vacuity. The vacuity is bounded in the range $[0, 1]$, and *zero-evidence sample*, the vacuity is maximum, leading to the greatest gradient value of $\frac{\partial \mathcal{L}_{\texttt{cor}}(\boldsymbol{x},\boldsymbol{y})}{\partial o_{gt}} = -1$. In other words, the regularization encourages the model to update the parameters such that the correct evidence $\alpha_{gt} - 1$ increases. As the model evidence increases, the vacuity decreases, and the contribution of the regularization $\mathcal{L}_{\texttt{cor}}(\boldsymbol{x}, \boldsymbol{y})$ is minimized. Thus, the proposed regularization enables the evidential model to learn from *zero-evidence samples*. □

4.1. Evidential Model Training

We formulate an overall objective used to train the proposed **R**egularized **e**vidential mo**d**el (**RED**). Essentially, the evidential model is trained to maximize the correct evidence, minimize the incorrect evidence, and avoid the *zero-evidence* region during training. The overall loss is

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) = \mathcal{L}^{\texttt{evid}}(\boldsymbol{x}, \boldsymbol{y}) + \eta_1 \mathcal{L}^{\texttt{inc}}(\boldsymbol{x}, \boldsymbol{y}) + \mathcal{L}^{\texttt{cor}}(\boldsymbol{x}, \boldsymbol{y}) \tag{16}$$

where $\mathcal{L}^{\texttt{evid}}(\boldsymbol{x}, \boldsymbol{y})$ is the loss based on the evidential framework given by (21), (23), or (22) (See Appendix B), $\mathcal{L}^{\texttt{inc}}(\boldsymbol{x}, \boldsymbol{y})$ represents the incorrect evidence regularization (See Appendix Section C), $\mathcal{L}^{\texttt{cor}}(\boldsymbol{x}, \boldsymbol{y})$ represents the proposed novel correct evidence regularization term in (12), and $\eta_1 = \lambda_1 \times \min(1.0, \text{epoch index}/10)$ controls the impact of incorrect evidence regularization to the overall model training. In this work, we consider the forward-KL based incorrect evidence regularization given in (42) based on (Sensoy et al., 2018).

4.2. Evidence Space Visualization

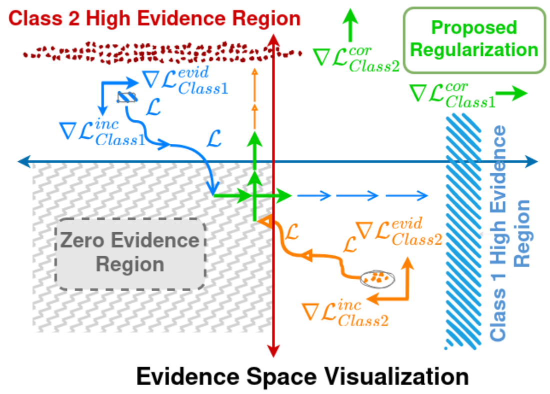


Figure 4. Evidence space visualization to demonstrate the effectiveness of the proposed method.

Figure 2 visualizes the evidence space in `ReLU`-based evidential models by considering the pre-ReLU output in a binary classification setting. Ideally, all samples that belong to Class 1 should be mapped to the blue region (region of high evidence for Class 1, low evidence for all other classes), all samples that belong to Class 2 should be mapped to the red region, and all out-of distribution samples should be mapped to the zero-evidence region (no evidence for all classes). To realize this goal, the models are trained using the evidential loss $\mathcal{L}^{evid}$ with incorrect evidence regularization $\mathcal{L}^{inc}$. However, there is no update to the evidential model from such samples of *zero-evidence region*. Model's prior belief of "I don't know" for such samples does not get updated even after being exposed to the true label. For the samples with high incorrect evidence and low correct evidence, evidential model aims to correct itself. However, many such samples are likely to get mapped to the zero-evidence region (as shown by blue and orange arrows in Figure 2) after which there is no update to the model. Such fundamental limitation holds true for all evidential models.

The evidence space visualization for RED is shown in Figure 4 to illustrate how it addresses the above limitation. Correct evidence regularization (indicated by green arrows) is weighted by the magnitude of the vacuity and is maximum in the zero-evidence region. In this problematic region, the proposed regularization fully dominates the model update as there is no update to the model from the two loss components ($\mathcal{L}^{evid}$ and $\mathcal{L}^{inc}$) in (16). As the sample gets far away from the zero evidence region, the vacuity decreases proportionally, the impact of the proposed regularization to model update becomes insignificant, and the evidential losses ($\mathcal{L}^{evid}$ & $\mathcal{L}^{inc}$) guide the model training. In this way, RED can effectively learn from all training samples irrespective of the model's existing evidence.

5. Experiments

Datasets and setup. We consider the standard supervised classification problem with MNIST (LeCun, 1998), Cifar10, and Cifar100 datasets (Krizhevsky et al., 2009), and few-shot classification with *mini*-ImageNet dataset (Vinyals et al., 2016). We employ the LeNet model for MNIST, ResNet18 model (He et al., 2016) for Cifar10/Cifar100, and ResNet12 model (He et al., 2016) for *mini*-ImageNet. We first conduct experiments to demonstrate the learning deficiency of existing evidential models to confirm our theoretical findings. We then evaluate the proposed correct evidence regularization to show its effectiveness. We finally conduct ablation studies to investigate the impact of evidential losses on model generalization and the uncertainty quantification of the proposed evidential model. Limited by space, additional clarifications, experiment results including few-shot classification experiments, experiments over challenging tiny-Imagenet datasett with Swin Transformer, hyperparameter details, and discussions are presented in the Appendix.

5.1. Learning Deficiency of Evidential Models

Sensitivity to the change of the architecture. We first consider a toy illustrative experiment with two frameworks: 1) standard softmax, 2) evidential learning, and experiment with the LeNet (LeCun et al., 1999) model considered in EDL (Sensoy et al., 2018) with a minor modification to the architecture: no dropout in the model. To construct the toy dataset, we randomly select 4 labeled data points from the MNIST training dataset as shown in the Figure 5. For the evidential model, we use ReLU to transform the network outputs to evidence, and train the model with MSE-based evidential loss (Sensoy et al., 2018) given in (21) without incorrect evidence regularization. We train both models using only these 4 training data points.

Figure 6 compares the training accuracy and training loss trends of the evidential model with the standard softmax model (trained with the cross-entropy loss). Before any training, both models have 0% accuracy and the loss is high as expected. For the evidential model, in the first few iterations, the model learns from the training dataset, and the model's accuracy increases to 50%. Afterward, the evidential model fails to learn as the evidential model maps two of the training data samples to the *zero-evidence region*. Even in such a trivial setting, the evidential model fails to fit the 4 training data points showing their learning deficiency that empirically verifies the conclusion in Theorem 1. It is also worth noting that the range of the evidential model's loss is significantly smaller than the standard model. This is mainly due to the bounded nature of the evidential MSE loss(*i.e.,* it is bounded in the range $[0, 2]$) (a detailed theoretical analysis of the evidential losses is provided in the Appendix). In contrast, the standard model trained with cross-entropy loss

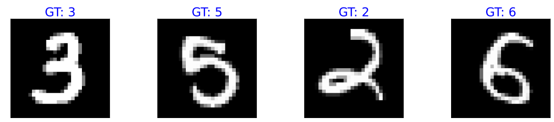
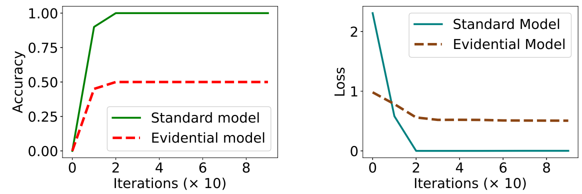


Figure 5. Toy dataset with 4 data points.



(a) Training accuracy trend (b) Training loss trend

Figure 6. Training of standard and evidential models

easily fits the trivial dataset, obtains near 0 loss, and perfect accuracy of 100% after a few iterations of training.

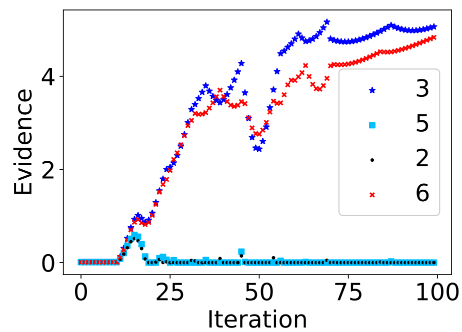


Figure 7. Zero-evidence trend during model training

Additionally, we visualize the zero-evidence data samples for the toy dataset setting. We plot the total evidence for each training sample as training progresses for the first 100 iterations. The total evidence trend as training progresses for the first 100 iterations is shown in Figure 7. The evidential model's predictions are correct for data samples with ground truth labels of 3 and 6, and incorrect for the remaining two data samples. After few iterations of training, the remaining two samples have zero total evidence (i.e. samples are mapped to zero evidence region), the model never learns from them, and the model only achieves overall 50% training accuracy even after 100 iterations. Clearly, the evidential model continues to output zero evidence for two of the training examples and fails to learn from them. Such learning deficiency of evidential models limits their extension to challenging settings. In contrast, the standard model easily overfits the 4 training examples and achieves 100% accuracy.

Sensitivity to hyperparameter tuning. In this experiment, evidential models are trained using evidential losses given in (21), (22), or (23) with incorrect evidence regularization to guide the model for accurate uncertainty quan-

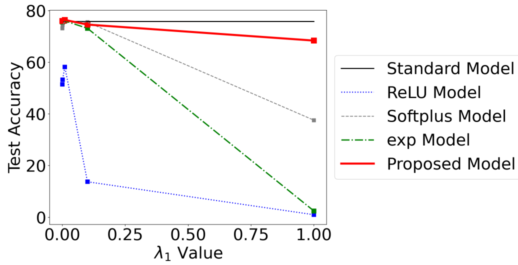


Figure 8. Impact of different incorrect evidence regularization strengths to the test set accuracy on Cifar100 dataset

tification. We study the impact of the incorrect evidence regularization λ_1 to the evidential model's performance using Cifar100. The result shows that the generalization performance of evidential models is highly sensitive to λ_1 values. To illustrate, we consider the Type II Maximum Likelihood loss in (23) with different λ_1 to control KL regularization (results on other loss functions are presented in the Appendix). As shown in Figure 8, when some regularization is introduced, evidential model's test performance improves slightly. However, when strong regularization is used, the model focuses strongly on minimizing the incorrect evidence. Such regularization causes the model to push many training samples into or close to the zero-evidence regions, which hurts the model's learning capabilities. In contrast, the proposed model can continue to learn from samples in zero-evidence regions, which shows its robustness to incorrect evidence regularization. Moreover, our model has stable performance across all hyperparameter settings as it can effectively learn from all training samples.

Challenging datasets and settings. We next consider standard classification models for the Cifar100 dataset and 1-shot classification with the *mini*-ImageNet dataset. We develop evidential extensions of the classification models using Type II Maximum Likelihood loss given in (23) without any incorrect evidence regularization and use `ReLU` to transform logits to evidence. As shown in Figure 10, compared to the standard classification model, the evidential model's predictive performance is sub-optimal (almost 20% lower for both classification problems). This is mainly due to the fact that evidential model maps many of the training data points to *zero-evidence region*, which is equivalent to the model saying "I don't know to which class this sample belongs" and stopping to learn from them. Consequently, the model fails to acquire new knowledge (*i.e.,* update itself), even after being exposed to correct supervision (the label information). In these cases, instead of learning, the evidential model chooses to ignore the training data on which it does not have any evidence and remains to be ignorant.

Visualization of zero-evidence samples. We next show the 2-dimensional visualization of the latent representation for the randomly selected 500 training examples based on

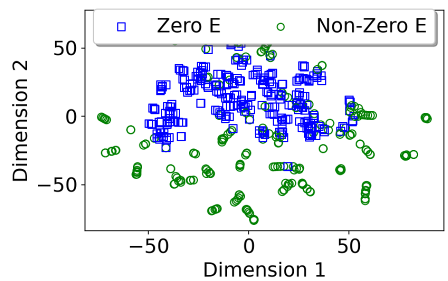
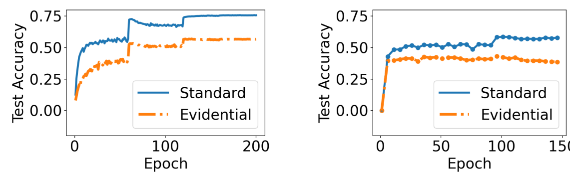


Figure 9. Zero-Evidence Sample Visualization



(a) Cifar100 Results (b) 1-Shot Results

Figure 10. Learning trends in complex classification problems

the tSNE plot for ReLU based evidential model trained on the Cifar100 dataset with $\lambda_1 = 0.1$. Figure 9 plot visualizes the latent embedding of zero evidence (Zero E) training samples with non-zero evidence (Non-Zero E) training samples. As can be seen, both zero and non-zero evidence samples appear to be dispersed, overlap at different regions, and cover a large area in the embedding space. This further confirms the challenge of effectively learning from these samples

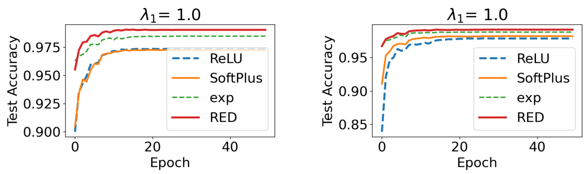
5.2. Effectiveness of the RED

Evidential activation function. We first experiment with different activation functions for the evidential models to show the superior predictive performance and generalization capability of exp activation validating our Theorem 2. We consider evidential models trained with evidential log loss given by (23) in Table 1 (Additional results along with hyperparameter details are presented in Appendix Section F). As can be seen, exp activation to transform network outputs into evidence leads to superior performance compared to `ReLU` and `Softplus` based transformations. Furthermore, our proposed model with correct evidence regularization further improves over the exp-based evidential models as it enables the evidential model to continue learning from *zero-evidence* samples.

Table 1. Classification performance comparison

Model	MNIST	Cifar10	Cifar100
`ReLU`	$98.19_{\pm 0.08}$	$41.43_{\pm 19.60}$	$61.27_{\pm 3.79}$
`SoftPlus`	$98.21_{\pm 0.05}$	$95.18_{\pm 0.11}$	$74.48_{\pm 0.17}$
exp	$98.79_{\pm 0.02}$	$95.11_{\pm 0.10}$	$76.12_{\pm 0.04}$
RED(Ours)	$\mathbf{99.10}_{\pm \mathbf{0.02}}$	$\mathbf{95.24}_{\pm \mathbf{0.06}}$	$\mathbf{76.43}_{\pm \mathbf{0.21}}$

We next present the test set performance change as training progresses with MNIST dataset and two different evidential losses in Figure 11 where we observe similar results. The exp activation shows superior performance, as it has smallest *zero-evidence region*, and does not suffer from many learning issues present in other activation functions.

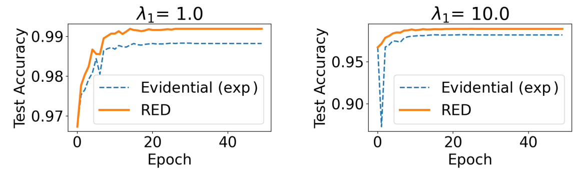


(a) Evidential MSE loss (b) Evidential Log loss

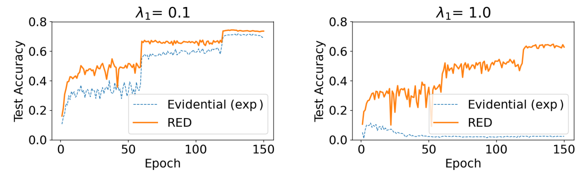
Figure 11. Impact of evidential activation functions to the Test Accuracy

Correct evidence regularization. We now study the impact of the proposed correct evidence regularization using the MNIST and Cifar100 classification problems. We consider the evidential baseline model that uses exp activation to acquire evidence, and is trained with Type II Maximum Likelihood based loss with different incorrect evidence regularization strengths. We introduce the proposed novel correct evidence regularization to the model. As can be seen in Figure 12, the model with correct-evidence regularization has superior generalization performance compared to the baseline evidential model. This is mainly due to the fact that with proposed correct evidence regularization, the evidential model can also learn from the zero-evidence training samples to acquire new knowledge instead of ignoring them. Our proposed model considers knowledge from all the training data and aims to acquire new knowledge to improve its generalization instead of ignoring the samples on which it has no knowledge. Finally, even though strong incorrect evidence regularization hurts the model's generalization, the proposed model is robust and generalizes better, empirically validating our Theorem 3. Limited by space, we present additional results in Appendix F.3.2.

Zero-evidence Sample Anaysis. Similar to the toy MNIST zero-evidence analysis, we consider the Cifar100 dataset, and carry out the analysis for this complex dataset/setting. Instead of focusing on a few training examples, we present the average statistics of the evidence ($\mathcal{E}$) for the 50,000 training samples in the 100 class classification problem for a model trained for 200 epochs using a log-based evidential loss in (23) with $\lambda_1 = 1.0$. For reference, the samples with less than 0.01 average evidence (*i.e.,* $\mathcal{E} \leq 0.01$) are samples on which the model is not confident (*i.e.,* having a high vacuity of $\nu \geq 0.99$), and are close to the ideal zero-evidence region. Our proposed RED model effectively avoids such zero evidence regions, and has the lowest number of samples (i.e. only 0.06% of total training dataset compared to 58.96% of SoftPlus based,



(a) Trend for $\lambda_1 = 1.0$ (b) Trend for $\lambda_1 = 10.0$



(c) Trend for $\lambda_1 = 0.1$ (d) Trend for $\lambda_1 = 1.0$

Figure 12. Impact of correct evidence regularization to test accuracy: (a), (b) - MNIST Results; (c), (d) - Cifar100 Results

and 100% of ReLU based evidential models) in very low evidence regions.

Table 2. Zero-Evidence Analysis for Complex Dataset-Setting

Model	$\mathcal{E} \leq .01$	$\mathcal{E} \leq 0.1$	$\mathcal{E} \leq 1.0$	$\mathcal{E} > 1.0$
ReLU	50000	50000	50000	0
SoftPlus	29483	32006	49938	62
Exp	48318	49881	49949	51
RED	30	16322	25154	24846

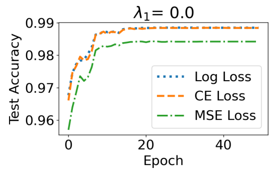
5.3. Ablation Study

Impact of loss function. We next study the impact of the evidential loss function on the model's performance using MNIST and CIFAR100 classification problems. We consider all three activations: ReLU, SoftPlus, and exp to transform neural network outputs to evidence and carry out experiments over CIFAR100 with identical model and settings. As seen in Table 3, the generalization performance of evidential model is consistently sub-optimal when trained with evidential MSE loss given by (21) compared to the two other evidential losses (22) & (23). This is consistent across all three evidence activation functions. This is mainly due to the bounded nature of the evidential MSE loss (21): for all training samples, evidential MSE loss is bounded in the range of $[0, 2]$. Type II Maximum Likelihood loss given in (23) and cross-entropy based evidential loss given in (22) show comparable empirical results.

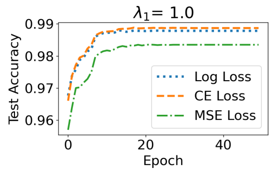
Next, we consider exp activation and conduct experiments over the MNIST dataset for incorrect evidence regularization strengths of $\lambda_1 = 0\&1$. We again observe similar results where the training with the Evidential MSE loss in (21) leads to sub-optimal test performance. Additional results, along with theoretical analysis are presented in the Appendix. In the subsequent experiments, we consider the Type II Maximum Likelihood loss (23) for evidential model training due to its simplicity and some theoretical advantages (see Appendix E). We leave a thorough investigation of these two evidential losses ((22) & (23)) as future work.

Table 3. Impact of evidential losses on classification performance

Loss	ReLU	SoftPlus	exp	**RED(Ours)**
MSE(21)	$31.49_{\pm 0.3}$	$15.74_{\pm 0.5}$	$42.95_{\pm 0.7}$	$\mathbf{75.73_{\pm 0.3}}$
CE (22)	$68.62_{\pm 2.4}$	$74.44_{\pm 0.1}$	$76.23_{\pm 0.1}$	$\mathbf{76.35_{\pm 0.1}}$
Log(23)	$61.27_{\pm 3.8}$	$74.48_{\pm 0.1}$	$76.12_{\pm 0.1}$	$\mathbf{76.43_{\pm 0.2}}$



(a) Trend for $\lambda_1 = 0.0$



(b) Trend for $\lambda_1 = 1.0$

Figure 13. Impact of evidential losses on test set accuracy

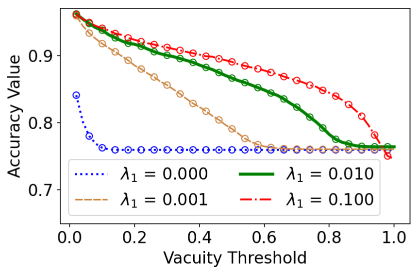


Figure 14. Accuracy-Vacuity curve

Study of uncertainty information. We now investigate the uncertainty behavior of the proposed evidential model with Cifar100 experiments. We present the Accuracy-Vacuity curve for different incorrect evidence regularization strengths (λ_1) in Figure 14. Vacuity reflects the lack of confidence in the predictions, and the accuracy of effective evidential model should increase with lower vacuity threshold. Without any incorrect evidence regularization (*i.e.,* $\lambda_1 = 0$), the evidential model is highly confident on its predictions and all test samples are concentrated on the low vacuity region. As the incorrect evidence regularization strength is increased, the model outputs more accurate confidence in the predictions. Strong incorrect evidence regularization hurts the generalization over the test set as indicated by low accuracy when all test samples are considered (*i.e.,* vacuity threshold of 1.0). In all cases, the evidential model shows reasonable uncertainty behavior: the model's test set accuracy increases as the vacuity threshold is decreased.

Next, we look at the accuracy of the evidential models on their top-K % most confident predictions over the test set. Table 4 shows the accuracy trend of Top-K (%) confident samples. Consider the most confident 20% samples (corresponding to 2000 test samples of Cifar100 dataset). The proposed model leads to highest accuracy (of 99.35%) compared to all the models. Similar trend is seen for different K values where the proposed model shows comparable to superior results demonstrating its accurate uncertainty quantification capability.

Table 4. Accuracy on Top-K% confident samples (%)

Model	10%	20%	30%	50%	80%	100%
`ReLU`	98.50	98.30	97.27	90.60	71.54	61.27
`SoftPlus`	99.10	98.75	98.30	95.86	85.56	74.48
`exp`	99.40	98.95	98.50	96.52	86.46	76.12
RED	99.60	99.35	98.83	96.24	86.38	76.43

We next consider out-of-distribution (OOD) detection experiments for the Cifar100-trained evidential model using SVHN dataset (as OOD) (Netzer et al., 2011). As seen in Table 5, the evidential models, on average, output very high vacuity for the OOD samples, showing the potential for OOD detection.

Table 5. Out-of-Distribution sample detection

Model	InD Vacuity	OOD Vacuity (SVHN)
exp	0.3227	0.7681
RED (Ours)	0.2729	0.7552

We present the AUROC score for Cifar100 trained models with SVHN dataset test set as the OOD samples in Table 6. In AUROC calculation, we use the maximum softmax score for the standard model, and predicted vacuity score for all the evidential models. As can be seen, the `exp`-based model outperforms all other activation functions, and the proposed model RED can learn from all the training samples that leads to the best performance.

Table 6. AUROC for Cifar100-SVHN experiment

Model	ReLU	SoftPlus	Standard	exp	**RED**
AUROC	0.7430	0.8058	0.8669	0.8804	**0.8833**

6. Conclusion

In this paper, we theoretically investigate the evidential models to identify their learning deficiency, which makes them fail to learn from zero-evidence regions. We then show the superiority of the evidential model with exp evidential activation over the `ReLU` and `SoftPlus` based models. We further analyze the evidential losses, and introduce a novel correct evidence regularization over the exp-based evidential model. The proposed model effectively pushes the training samples out of the zero-evidence regions, leading to superior learning capabilities. We conduct extensive experiments that empirically validate all theoretical claims while demonstrating the effectiveness of the proposed approach.

Acknowledgements

This research was supported in part by an NSF IIS award IIS-1814450 and an ONR award N00014-18-1-2875. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agency.

References

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

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

Bishop, C. M. and Nasrabadi, N. M. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

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

Charpentier, B., Zügner, D., and Günnemann, S. Posterior network: Uncertainty estimation without ood samples via density-based pseudo-counts. *Advances in Neural Information Processing Systems*, 33:1356–1367, 2020.

Chen, Y., Liu, Z., Xu, H., Darrell, T., and Wang, X. Meta-baseline: Exploring simple meta-learning for few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9062–9071, 2021.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1126–1135. JMLR. org, 2017.

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

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

Huynh, E. Vision transformers in 2022: An update on tiny imagenet. *arXiv preprint arXiv:2205.10660*, 2022.

Jøsang, A. *Subjective logic*, volume 3. Springer, 2016.

Kamath, U., Liu, J., and Whitaker, J. *Deep learning for NLP and speech recognition*, volume 84. Springer, 2019.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Knopp, K. Weierstrass's factor-theorem. In *Theory of Functions: Part II*, pp. 1–7. Dover, 1996.

Kopetzki, A.-K., Charpentier, B., Zügner, D., Giri, S., and Günnemann, S. Evaluating robustness of predictive uncertainty estimation: Are dirichlet-based models reliable? In *International Conference on Machine Learning*, pp. 5707–5718. PMLR, 2021.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. -, 2009.

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

LeCun, Y. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

LeCun, Y., Haffner, P., Bottou, L., and Bengio, Y. Object recognition with gradient-based learning. In *Shape, contour and grouping in computer vision*, pp. 319–345. Springer, 1999.

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

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

Mobiny, A., Yuan, P., Moulik, S. K., Garg, N., Wu, C. C., and Van Nguyen, H. Dropconnect is effective in modeling uncertainty of bayesian deep networks. *Scientific reports*, 11(1):1–14, 2021.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. 2011.

Nguyen, A., Yosinski, J., and Clune, J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 427–436, 2015.

Pandey, D. S. and Yu, Q. Multidimensional belief quantification for label-efficient meta-learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14391–14400, June 2022a.

Pandey, D. S. and Yu, Q. Evidential conditional neural processes. *arXiv preprint arXiv:2212.00131*, 2022b.

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

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

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

Shi, W., Zhao, X., Chen, F., and Yu, Q. Multifaceted uncertainty estimation for label-efficient deep learning. *Advances in neural information processing systems*, 33, 2020.

Singh, S. P., Kumar, A., Darbari, H., Singh, L., Rastogi, A., and Jain, S. Machine translation using deep learning: An overview. In *2017 international conference on computer, communications and electronics (comptelix)*, pp. 162–167. IEEE, 2017.

Tomani, C. and Buettner, F. Towards trustworthy predictions from deep neural networks with fast adversarial calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 9886–9896, 2021.

Ulmer, D. A survey on evidential deep learning for single-pass uncertainty estimation. *arXiv preprint arXiv:2110.03051*, 2021.

Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.

Voulodimos, A., Doulamis, N., Doulamis, A., and Protopapadakis, E. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*, 2018, 2018.

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

Appendix

Organization of the Appendix

- In Section A, we present an analysis of standard classification models trained with cross-entropy loss to show their learning capabilities.
- In Section B, we present a complete proof of Theorem 1 for different evidential losses that demonstrates the inability of evidential models to learn from zero-evidence samples.
- In Section C, we describe different incorrect evidence regularizations used in the existing literature and carry out a gradient analysis to study their impact on evidential model learning.
- In Section D, we present the proof for Theorem 2 that shows the superiority of exp activation over the `SoftPlus` and `ReLU` functions to transform logits to evidence.
- In Section E, we analyze the evidential losses that reveals the theoretical limitation of evidential models trained using Bayes risk with sum of squares loss.
- In Section F, we present additional experiment results, clarifications, hyperparameter details, and discuss some limitations along with possible future works.

The source code for the experiments carried out in this work is attached in the supplementary materials and is available at the link: https://github.com/pandeydeep9/EvidentialResearch2023

A. Standard Classification Model

Consider a standard cross-entropy based model for $K-$class classification. Let the overall network be represented by $f_\Theta(.)$, and let $\mathbf{o} = f_\Theta(\mathbf{x})$ be the output from this network before the softmax layer for input $\mathbf{x}$ and one-hot ground truth label of $\mathbf{y}$. The output after the softmax layer is given by

$$\texttt{sm}_i = \frac{\exp(o_i)}{\sum_{k=1}^{K}\exp(o_k)} = \frac{\exp(o_i)}{S^{\text{ce}}} \tag{17}$$

Where $S^{\text{ce}} = \sum_{i=1}^{K}\exp(o_i)$. The model is trained with cross-entropy loss. For a given sample $(\mathbf{x}, \mathbf{y})$, the loss is given by

$$\mathcal{L}_{\texttt{cross-entropy}} = -\sum_{k=1}^{K} y_k \log(\texttt{sm}_k) = -\sum_{k=1}^{K}\Big[y_k o_k - y_k \log\Big(\sum_{i=1}^{K}\exp(o_i)\Big)\Big] \tag{18}$$

$$= \log S^{\text{ce}} - \sum_{k=1}^{K} y_k o_k \tag{19}$$

Now, looking at the gradient of this loss with respect to the pre-softmax values $\mathbf{o}$

$$\text{grad}_k = \frac{\partial \mathcal{L}_{\texttt{cross-entropy}}}{\partial o_k} = \Big(\frac{1}{S^{\text{ce}}}\frac{\partial S^{\text{ce}}}{\partial o_k} - y_k\Big) = \Big(\frac{\exp(o_k)}{S^{\text{ce}}} - y_k\Big) = \texttt{sm}_k - y_k \tag{20}$$

Analysis of the gradients For Standard Classification Model.

The gradient measures the error signal, and for standard classification models, it is bounded in the range [-1, 1] as $0 \leq \texttt{sm}_k \leq 1$ and $y_k \in \{0, 1\}$. The model is updated using gradient descent based optimization objectives. For input $\mathbf{x}$, the neural network outputs K values o_1 to o_K, and the corresponding ground truth is $\mathbf{y}$, $y_{gt} = 1, y_{\neq gt} = 0$.

When $y_i = 0$, the gradient signal is $\text{grad}_i = \texttt{sm}_i$ and the model optimizes the parameters to minimize this value. Only when $\texttt{sm}_i = 0$, the gradient is zero, and the model is not updated. In all other cases when $\texttt{sm}_i \neq 0$, there is a non-zero gradient dependent on $\texttt{sm}_i$, and the model is updated to minimize the $\texttt{sm}_i$ as expected.

When y_i = 1, the gradient signal is $\text{grad}_i = \texttt{sm}_i - 1$ and the model optimizes the parameters to minimize this value. As $\texttt{sm}_i \in [0, 1]$, only when the model outputs a large logit on i (corresponding to the ground truth class) and small logit for all other nodes, $\texttt{sm}_i = 1$, the gradient is zero, and the model is not updated. In all other cases when $\texttt{sm}_i < 1$, there is a non-zero gradient dependent on $\texttt{sm}_i$ and the model is updated to maximize the $\texttt{sm}_i$ and minimize all other $\texttt{sm}_{\neq i}$ as expected. The gradient signal in standard classification models trained with standard cross-entropy loss is reasonable and enables learning from all the training data samples.

B. Evidential Classification Models

Theorem 1: Given a training sample $(\mathbf{x}, \mathbf{y})$, if an evidential neural network outputs zero evidence $\mathbf{e}$, then the gradients of the evidential loss evaluated on this training sample over the network parameters reduce to zero.

Proof. In the main paper, we considered a $K-$class classification problem and a representative evidential model trained using Bayes risk with sum of squares loss (Eqn. 21) in the proof. Following 3 variants of evidential losses ((Sensoy et al., 2018)) have been commonly used in evidential classification works:

1. Bayes risk with sum of squares loss (*i.e.,* Evidential MSE loss) (Zhao et al., 2020)

$$\mathcal{L}^{\texttt{MSE}}(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{K} (y_j - \frac{\alpha_j}{S})^2 + \frac{\alpha_j(S - \alpha_j)}{S^2(S+1)} \tag{21}$$

2. Bayes risk with cross-entropy loss (*i.e.,* Evidential CE loss)(Charpentier et al., 2020)

$$\mathcal{L}^{\texttt{CE}}(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{K} y_k \Big(\Psi(S) - \Psi(\alpha_k) \Big) \tag{22}$$

3. Type II Maximum Likelihood loss (*i.e.,* Evidential log loss)(Pandey & Yu, 2022a)

$$\mathcal{L}^{\texttt{Log}}(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^{K} y_k \Big(\log(S) - \log(\alpha_k) \Big) \tag{23}$$

For completeness, we consider all three loss functions used in evidential classification models and carry out their analysis.

B.1. Gradient of Evidential Activation Functions $\mathcal{A}(.)$

Three non-linear functions are proposed and commonly used in the existing literature to transform the neural network output to evidence: 1) `ReLU` function, 2) `SoftPlus` function, and 3) Exponential function. In this section, we compute the gradients of the evidence output e_i from these non-linear activation functions with respect to the logit input o_i

1. $\mathcal{A}(.) = \texttt{ReLU}(.) = \max(0, .)$

$$e_k = \texttt{ReLU}(o_k) = \max(0, o_k) \implies \frac{\partial e_k}{\partial o_k} = \begin{cases} 0 & \text{if} \quad o_k \leq 0 \\ 1 & \text{otherwise} \end{cases} \tag{24}$$

2. $\mathcal{A}(.) = \texttt{SoftPlus}(.) = \log(1 + \exp(.))$

$$e_k = \log(\exp(o_k) + 1) \implies \frac{\partial e_k}{\partial o_k} = \frac{1}{1 + \exp(-o_k)} = \text{Sigmoid}(o_k) \tag{25}$$

3. $\mathcal{A}(.) = \exp(.)$

$$e_k = \exp(o_k) \implies \frac{\partial e_k}{\partial o_k} = \exp(o_k) = e_k = \alpha_k - 1 \tag{26}$$

B.2. Evidential Model Trained using Bayes risk with sum of squares loss (*i.e.,* Eqn. 21)

Proof. Consider an input $\mathbf{x}$ with one-hot ground truth label of $\mathbf{y}$. Let the ground truth class be g i.e. $y_{gt} = 1$, with corresponding Dirichlet parameter α_{gt}, and $y_{\neq gt} = 0$. Moreover, let $\mathbf{o}, \mathbf{e}$, and $\boldsymbol{\alpha}$ represent the neural network output vector before applying the activation $\mathcal{A}$, the evidence vector, and the Dirichlet parameters respectively.

In this evidential framework, the loss is given by

$$\mathcal{L}^{\text{MSE}}(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{K}(y_j - \frac{\alpha_j}{S})^2 + \frac{\alpha_j(S-\alpha_j)}{S^2(S+1)} = 1 - \frac{2\alpha_{gt}}{S} + \frac{\sum_k \alpha_k^2}{S^2} + \frac{2\sum_i\sum_j \alpha_i\alpha_j}{S^2(S+1)} \tag{27}$$

$$= 2 - \frac{2\alpha_{gt}}{S} - \frac{2\sum_i\sum_j \alpha_i\alpha_j}{S(S+1)} \tag{28}$$

Now, consider different components of the loss and compute the gradients of the components with respect to Dirichlet parameters α,

$$\frac{\partial \frac{\alpha_{gt}}{S}}{\partial \alpha_{gt}} = \frac{1}{S} - \frac{\alpha_{gt}}{S^2} \quad \& \quad \frac{\partial \frac{\alpha_{gt}}{S}}{\partial \alpha_{\neq gt}} = -\frac{\alpha_{gt}}{S^2} \implies \frac{\partial \frac{\alpha_{gt}}{S}}{\partial \alpha_k} = \frac{y_k}{S} - \frac{\alpha_{gt}}{S^2}$$

The gradient of the variance term is the same for all the K Dirichlet parameters and is given by

$$\frac{\partial \frac{\sum_i\sum_j \alpha_i\alpha_j}{S(S+1)}}{\partial \alpha_k} = \frac{(S-\alpha_k)}{S(S+1)} - \frac{(2S+1)\sum_i\sum_j \alpha_i\alpha_j}{(S^2+S)^2}$$

Now, the gradient of the loss with respect to the neural network output can be computed using the chain rule as

$$\frac{\partial \mathcal{L}^{\text{MSE}}(\mathbf{x}, \mathbf{y})}{\partial o_k} = \frac{\partial \mathcal{L}^{\text{MSE}}(\mathbf{x}, \mathbf{y})}{\partial \alpha_k}\frac{\partial e_k}{\partial o_k} = -\left[2\frac{\partial \frac{\alpha_k}{S}}{\partial \alpha_k} - 2\frac{\partial \frac{\sum_i\sum_j \alpha_i\alpha_j}{S(S+1)}}{\partial \alpha_k}\right] \times \frac{\partial e_k}{\partial o_k}$$

$$= \left[\frac{2\alpha_{gt}}{S^2} - 2\frac{y_k}{S} - \frac{2(S-\alpha_k)}{S(S+1)} + \frac{2(2S+1)\sum_i\sum_j \alpha_i\alpha_j}{(S^2+S)^2}\right] \times \frac{\partial e_k}{\partial o_k}$$

Case I: $\texttt{ReLU}(.)$ to transform logits to evidence

$$e_k = \text{ReLU}(o_k) = \max(0, o_k) \implies \frac{\partial e_k}{\partial o_k} = \begin{cases} 1 & \text{if} \quad o_k > 0 \\ o & \text{otherwise} \end{cases} \tag{29}$$

For zero-evidence sample with $\texttt{ReLU}(.)$ used to transform the logits to evidence, the logits o_k satisfy the relationship $o_k \leq 0 \;\forall\; k \implies \frac{\partial e_k}{\partial o_k} = 0 \implies \frac{\partial \mathcal{L}^{\text{MSE}}(\mathbf{x},\mathbf{y})}{\partial o_k} = 0$

Case II: $\texttt{SoftPlus}(.)$ to transform logits to evidence

$$e_k = \log(\exp(o_k) + 1) \implies \frac{\partial e_k}{\partial o_k} = \text{Sigmoid}(o_k) \tag{30}$$

Case II: $\exp(.)$ to transform logits to evidence

$$e_k = \exp(o_k) \implies \frac{\partial e_k}{\partial o_k} = \exp(o_k) = \alpha_k - 1 \tag{31}$$

For zero-evidence sample with $\texttt{SoftPlus}(.)$ used to transform the logits to evidence, the logits $o_k \to -\infty \implies \text{Sigmoid}(o_k) \to 0 \;\&\; \frac{\partial e_k}{\partial o_k} \to 0$. For zero-evidence sample with $\exp(.)$ used to transform the logits to evidence, $\alpha_k \to 1 \implies \frac{\partial e_k}{\partial o_k} \to 0$. Moreover, there is no term in the first part of the loss gradient (see Eqn. 29) to counterbalance these zero-approaching gradients. So, for zero-evidence samples,

$$\frac{\partial \mathcal{L}^{\text{MSE}}(\mathbf{x}, \mathbf{y})}{\partial o_k} = 0 \tag{32}$$

Since the gradient of the loss with respect to all the nodes is zero, there is no update to the model from such samples. Thus, the evidential models fail to learn from such zero-evidence samples. □

B.3. Evidential Model Trained using Type II Maximum Likelihood formulation of Evidential loss (*i.e.,* Eqn. 23)

Consider a $K-$class evidential classification model that trains the model using Type II Maximum Likelihood formulation of the evidential loss. Consider an input $\mathbf{x}$ with one-hot ground truth label of $\mathbf{y}$, $\sum_{k=1}^{K} y_k = 1$. For this evidential framework, the Type II Maximum Likelihood loss is given by

$$\mathcal{L}^{\text{Log}}(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^{K} y_k \Big(\log(S) - \log(\alpha_k) \Big) = \log S - \sum_{k=1}^{K} y_k \log \alpha_k \tag{33}$$

Taking the gradient of the loss with the logits $\mathbf{o}$, we get

$$\text{grad}_k = \frac{\partial \mathcal{L}^{\text{Log}}(\mathbf{x}, \mathbf{y})}{\partial o_k} = \frac{1}{S}\frac{\partial S}{\partial o_k} - y_k \frac{1}{\alpha_k}\frac{\partial \alpha_k}{\partial o_k} = \Big(\frac{1}{S} - \frac{y_k}{\alpha_k}\Big)\frac{\partial e_k}{\partial o_k} \tag{34}$$

Case I: $\texttt{ReLU}(.)$ to transform logits to evidence

For any zero-evidence sample with $\texttt{ReLU}(.)$ used to transform the logits to evidence, the logits o_k satisfy the relationship $o_k \leq 0 \; \forall \; k \implies \frac{\partial e_k}{\partial o_k} = 0 \implies \frac{\partial \mathcal{L}^{\text{Log}}(\mathbf{x}, \mathbf{y})}{\partial o_k} = 0 \; \forall k \in [1, K]$

Case II: $\texttt{SoftPlus}(.)$ to transform logits to evidence. Considering Eqn. 34 and Eqn 25, the gradient of the loss with respect to the logits becomes

$$\text{grad}_k = \frac{\partial \mathcal{L}^{\text{Log}}(\mathbf{x}, \mathbf{y})}{\partial o_k} = \Big(\frac{1}{S} - \frac{y_k}{\alpha_k}\Big)\text{Sigmoid}(o_k) \tag{35}$$

Case III: $\exp(.)$ to transform logits to evidence. Considering Eqn. 34 and Eqn 26, the gradient of the loss with respect to the logits becomes

$$\text{grad}_k = \frac{\partial \mathcal{L}^{\text{Log}}(\mathbf{x}, \mathbf{y})}{\partial o_k} = \Big(\frac{1}{S} - \frac{y_k}{\alpha_k}\Big)(e_k) = \Big(\frac{1}{S} - \frac{y_k}{\alpha_k}\Big)(\alpha_k - 1) \tag{36}$$

For zero-evidence sample with $\texttt{SoftPlus}(.)$ used to transform the logits to evidence, the logits $o_k \to -\infty \implies \text{Sigmoid}(o_k) \to 0$ & $\frac{\partial e_k}{\partial o_k} \to 0$. Similarly, for zero-evidence sample with $\exp(.)$ used to transform the logits to evidence, $\alpha_k \to 1 \implies \frac{\partial e_k}{\partial o_k} \to 0$. Moreover, there is no term in the first part of the loss gradient (see Eqn. 35 and Eqn. 36) to counterbalance these zero-approaching gradient terms.

Since the gradient of the loss with respect to all the nodes is zero, there is no update to the model from such samples. Thus, the evidential models trained with Type II Maximum Likelihood formulation of the evidential loss fail to learn from such zero-evidence samples.

B.4. Evidential Model Trained using Bayes risk with cross-entropy formulation of Evidential loss (*i.e.,* Eqn. 22)

Consider a $K-$class evidential classification model that trains model using Bayes risk with cross-entropy loss for evidential learning (Eqn. 22). Consider an input $\mathbf{x}$ with one-hot ground truth label of $\mathbf{y}$, $\sum_{k=1}^{K} y_k = 1$. For this evidential framework, the loss is given by

$$\mathcal{L}^{\text{CE}}(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{K} y_k \Big(\Psi(S) - \Psi(\alpha_k) \Big) = \Psi(S) - \Psi(\alpha_{gt}) \tag{37}$$

Where α_{gt} represents the output Dirichlet parameter for the ground truth class i.e. $y_{gt} = 1$, $y_{\neq gt} = 0$, and $\Psi(.)$ represents the Digamma function, and for $z \geq 1$, is given by

$$\Psi(z) = \frac{d}{dz}\log\Gamma(z) = \frac{d}{dz}\bigg(-\gamma z - \log z + \sum_{n=1}^{\infty}\Big(\frac{z}{n} - \log\big(1 + \frac{z}{n}\big)\Big)\bigg) = -\gamma - \frac{1}{z} + z\sum_{n=1}^{\infty}\frac{1}{n(n+z)}$$

Here, γ is the Euler–Mascheroni constant, and $\Gamma(.)$ is the gamma function, Using Weierstass's definition of gamma function (Knopp, 1996) for values outside negative integers that is given by

$$\Gamma(z) = \frac{e^{-\gamma z}}{z} \prod_{n=1}^{\infty} \Big(1 + \frac{z}{n}\Big)^{-1} e^{\frac{z}{n}}$$

Using the definition of the digamma functions, the loss updates as

$$\mathcal{L}^{\text{CE}}(\mathbf{x}, \mathbf{y}) = \Psi(S) - \Psi(\alpha_{gt}) = \frac{1}{\alpha_{gt}} - \frac{1}{S} + S \sum_{n=1}^{\infty} \frac{1}{n(n+S)} - \alpha_{gt} \sum_{n=1}^{\infty} \frac{1}{n(n+\alpha_{gt})} \tag{38}$$

The derivative of the digamma function is bounded and is given by

$$\frac{\partial \Psi(z)}{\partial z} = \frac{\partial}{\partial z}\Big(-\gamma - \frac{1}{z} + \sum_{n=1}^{\infty} \frac{1}{n} - \frac{1}{n+z}\Big) = \frac{1}{z^2} + \sum_{n=1}^{\infty} \frac{1}{(n+z)^2}$$
$$\frac{1}{z^2} < \frac{\partial \Psi(z)}{\partial z} < \frac{1}{z^2} + \frac{\pi^2}{6}, \quad z \geq 1$$

With this, we can compute the gradients of the loss with respect to the logits as

$$\text{grad}_k = \frac{\partial \mathcal{L}^{\text{CE}}(\mathbf{x}, \mathbf{y})}{\partial o_k} = \frac{\partial}{\partial \alpha_k}\big(\Psi(S) - \Psi(\alpha_{gt})\big)\frac{\partial \alpha_k}{\partial o_k} = \Big(\frac{1}{S^2} + \sum_{i=1}^{\infty} \frac{1}{(n+S)^2} - \frac{y_k}{\alpha_{gt}^2} - \sum_{i=1}^{\infty} \frac{y_k}{(n+\alpha_{gt})^2}\Big)\frac{\partial e_k}{\partial o_k} \tag{39}$$

Case I: $\texttt{ReLU}(.)$ to transform logits to evidence

For any zero-evidence sample with $\texttt{ReLU}(.)$ used to transform the logits to evidence, the logits o_k satisfy the relationship $o_k \leq 0 \;\forall\; k \implies \frac{\partial e_k}{\partial o_k} = 0 \implies \frac{\partial \mathcal{L}^{\text{CE}}(\mathbf{x}, \mathbf{y})}{\partial o_k} = 0 \;\forall k \in [1, K]$

Case II: $\texttt{SoftPlus}(.)$ to transform logits to evidence. Considering Eqn. 25 and Eqn 39, the gradient of the loss with respect to the logits becomes

$$\text{grad}_k = \frac{\partial \mathcal{L}^{\text{CE}}(\mathbf{x}, \mathbf{y})}{\partial o_k} = \Big(\frac{1}{S^2} + \sum_{i=1}^{\infty} \frac{1}{(n+S)^2} - \frac{y_k}{\alpha_{gt}^2} - \sum_{i=1}^{\infty} \frac{y_k}{(n+\alpha_{gt})^2}\Big)\text{Sigmoid}(o_k) \tag{40}$$

Case III: $\exp(.)$ to transform logits to evidence. Considering Eqn. 26 and Eqn 39, the gradient of the loss with respect to the logits becomes

$$\text{grad}_k = \frac{\partial \mathcal{L}^{\text{CE}}(\mathbf{x}, \mathbf{y})}{\partial o_k} = \Big(\frac{1}{S^2} + \sum_{i=1}^{\infty} \frac{1}{(n+S)^2} - \frac{y_k}{\alpha_{gt}^2} - \sum_{i=1}^{\infty} \frac{y_k}{(n+\alpha_{gt})^2}\Big)(\alpha_k - 1) \tag{41}$$

For zero-evidence sample with $\texttt{SoftPlus}(.)$ used to transform the logits to evidence, the logits $o_k \to -\infty \implies \text{Sigmoid}(o_k) \to 0$ & $\frac{\partial e_k}{\partial o_k} \to 0$. Similarly, for zero-evidence sample with $\exp(.)$ used to transform the logits to evidence, $\alpha_k \to 1 \implies \frac{\partial e_k}{\partial o_k} \to 0$. Moreover, there is no term in the first part of the loss gradient (see Eqn. 29) to counterbalance these zero-approaching gradient terms.

The gradient of the loss with respect to all the nodes is zero for all the considered cases. Since the gradient of the loss with respect to all the nodes is zero for all three cases, there is no update to the model from such samples. Thus, the evidential models fail to learn from such zero-evidence samples in all cases. □

C. Regularization in the Evidential Classification Models

Based on the evidence $\mathbf{e}$, beliefs $\mathbf{b}$, and the Dirichlet parameters $\boldsymbol{\alpha}$, various regularization terms have been introduced that aim to penalize the incorrect evidence/incorrect belief of the model, leading to the model with accurate uncertainty estimates. Here, we briefly summarize the key regurlaizations:

1. Introduce a forward KL regularization term as in EDL (Sensoy et al., 2018) that regularizes the model to output no incorrect evidence.

$$\mathcal{L}^{\texttt{EDL}}_{\texttt{reg}}(\boldsymbol{x},\boldsymbol{y}) = \mathrm{KL}\big(\texttt{Dir}(\boldsymbol{p}|\tilde{\boldsymbol{\alpha}})||\texttt{Dir}(\boldsymbol{p}|\mathbf{1})\big) = \log\Big(\frac{\Gamma\sum_{k=1}^{K}\tilde{\alpha}_k}{\Gamma(K)\prod_{k=1}^{K}\Gamma\tilde{\alpha}_k}\Big) + \sum_{k=1}^{K}(\tilde{\alpha}_k - 1)\Big[\psi(\tilde{\alpha}_k) - \psi\Big(\sum_{j=1}^{K}\tilde{\alpha}_j\Big)\Big] \tag{42}$$

Where $\tilde{\boldsymbol{\alpha}} = \boldsymbol{y} + (\mathbf{1} - \boldsymbol{y}) \odot \boldsymbol{\alpha} = (\tilde{\alpha}_1, \tilde{\alpha}_2, ...\tilde{\alpha}_N)$ parameterize a dirichlet distribution, $\tilde{\alpha}_{i=gt} = 1, \tilde{\alpha}_i = \alpha_i \forall i \neq gt$. Here, the KL regularization term encourages the Dirichlet distribution based on the incorrect evidence i.e., $\mathrm{Dir}(\boldsymbol{p}|\tilde{\boldsymbol{\alpha}})$ to be flat which is possible when there is no incorrect evidence. From Eqn. 42, we can see that the regularization term, introduces digamma functions for the loss and may require evaluation of higher-order polygamma functions for challenging problems (e.g. involving bi-level optimizations as in MAML (Finn et al., 2017)).

2. Introduce an incorrect evidence regularization term as in ADL (Shi et al., 2020) that is the sum of the incorrect evidence for a sample

$$\mathcal{L}^{\texttt{ADL}}_{\texttt{reg}}(\boldsymbol{x},\boldsymbol{y}) = \sum_{k=1}^{K}\big(\mathbf{e}\odot(\mathbf{1}-\mathbf{y})\big)_k = \sum_{k=1}^{K} e_k \times (1 - y_k) \tag{43}$$

Here, $\odot$ represents element-wise product. The evidence for a class e_k is only restricted to be non-negative and can take large positive values leading to large variation in the overall loss.

3. Introduce incorrect belief-based regularization as in Units-ML (Pandey & Yu, 2022a)

$$\mathcal{L}^{\texttt{Units}}_{\texttt{reg}}(\boldsymbol{x},\boldsymbol{y}) = \sum_{k=1}^{K}\big(\frac{\mathbf{e}}{S}\odot(\mathbf{1}-\mathbf{y})\big)_k = \sum_{k=1}^{K} \frac{e_k}{S} \times (1 - y_k) \tag{44}$$

The regularization value is bounded to be in a range of $[0, 1]$ for all the data samples, no matter how severe the mistake is.

All three regularizations aim to guide the model such that the incorrect evidence is minimized (ideally close to zero). These regularizations help the evidential model acquire desired uncertainty quantification capabilities in evidential models. Such guidance is expected to update the model such that it maps input samples near zero-evidence regions in the evidence space. Thus, the regularization does not help address the issue of learning from zero-evidence samples and is likely to hurt the model's learning capabilities.

C.1. Gradient Analysis of the Incorrect Evidence Regularizations

The regularization terms use ground truth information to consider only the incorrect evidence. Thus, the gradient of the regularization loss with respect to the ground truth node α_{gt} is 0. In this analysis, we consider the gradient with respect to non-ground truth nodes i.e. α_k, and $o_k, k \neq gt$.

1. Gradient for EDL regularization (Eqn. 42)

$$\begin{aligned}\mathcal{L}^{\texttt{EDL}}_{\texttt{reg}}(\boldsymbol{x},\boldsymbol{y}) &= \mathrm{KL}\big(\texttt{Dir}(\boldsymbol{p}|\tilde{\boldsymbol{\alpha}})||\texttt{Dir}(\boldsymbol{p}|\mathbf{1})\big) = \log\Big(\frac{\Gamma\sum_{k=1}^{K}\tilde{\alpha}_k}{\Gamma(K)\prod_{k=1}^{K}\Gamma\tilde{\alpha}_k}\Big) + \sum_{k=1}^{K}(\tilde{\alpha}_k - 1)\Big[\psi(\tilde{\alpha}_k) - \psi\Big(\sum_{j=1}^{K}\tilde{\alpha}_j\Big)\Big] \\ &= \log\Gamma(S-\alpha_{gt}) - \log\Gamma(K) - \sum_{k=1}^{K}\log\Gamma\tilde{\alpha}_k + \sum_{k=1}^{K}(\tilde{\alpha}_k - 1)\Big[\psi(\tilde{\alpha}_k) - \psi(S-\alpha_{gt})\Big]\end{aligned} \tag{45}$$

$$
\begin{aligned}
\frac{\partial \mathcal{L}^{\texttt{EDL}}_{\texttt{reg}}(\boldsymbol{x},\boldsymbol{y})}{\partial \alpha_k} &= \frac{\partial}{\partial \alpha_k}\bigg(\log\Gamma(S-\alpha_{gt}) - \log\Gamma(K) - \sum_{k=1}^{K}\log\Gamma\tilde{\alpha}_k + \sum_{k=1}^{K}(\tilde{\alpha}_k - 1)\Big[\psi(\tilde{\alpha}_k) - \psi(S-\alpha_{gt})\Big]\bigg) \\
&= \psi(S-\alpha_{gt}) - \psi(\alpha_k) + \frac{\partial}{\partial \alpha_k}\bigg(\sum_{k=1}^{K}(\tilde{\alpha}_k - 1)\Big[\psi(\tilde{\alpha}_k) - \psi(S-\alpha_{gt})\Big]\bigg) \\
&= \psi(S-\alpha_{gt}) - \psi(\alpha_k) + \psi(\alpha_k) - \psi(S-\alpha_{gt}) + (\alpha_k - 1)\frac{\partial}{\partial \alpha_k}\Big(\psi(\tilde{\alpha}_k) - \psi(S-\alpha_{gt})\Big) \\
&= (\alpha_k - 1)\frac{\partial}{\partial \alpha_k}\Big(\psi(\alpha_k) - \psi(S-\alpha_{gt})\Big) = (\alpha_k - 1)\big(\psi_1(\alpha_k) - \psi_1(S-\alpha_{gt})\big)
\end{aligned}
$$

Where ψ_1 is the trigamma function. Further, using the definition of trigamma function,

$$
\frac{\partial \mathcal{L}^{\texttt{EDL}}_{\texttt{reg}}(\boldsymbol{x},\boldsymbol{y})}{\partial \alpha_k} = (\alpha_k - 1)\big(\psi_1(\alpha_k) - \psi_1(S-\alpha_{gt})\big) = (\alpha_k - 1)\bigg(\sum_{n=0}^{\infty}\frac{1}{(n+\alpha_k)^2} - \frac{1}{(n+S-\alpha_{gt})^2}\bigg) \tag{46}
$$

Now, the gradients with respect to the logits o_k becomes

$$
\frac{\partial \mathcal{L}^{\texttt{EDL}}_{\texttt{reg}}(\boldsymbol{x},\boldsymbol{y})}{\partial o_k} = \frac{\partial \mathcal{L}^{\texttt{EDL}}_{\texttt{reg}}(\boldsymbol{x},\boldsymbol{y})}{\partial \alpha_k}\frac{\partial \alpha_k}{\partial o_k} = (\alpha_k - 1)\bigg(\sum_{n=0}^{\infty}\frac{1}{(n+\alpha_k)^2} - \frac{1}{(n+S-\alpha_{gt})^2}\bigg) \times \frac{\partial e_k}{\partial o_k} \tag{47}
$$

Case I: $\texttt{ReLU}(.)$ to transform logits to evidence. The gradients with respect to the logits o_k for zero evidence is zero. For all non-zero evidence, the gradient updates as $\frac{\partial e_k}{\partial o_k} = 1 \forall e_k > 0$ and

$$
\frac{\partial \mathcal{L}^{\texttt{EDL}}_{\texttt{reg}}(\boldsymbol{x},\boldsymbol{y})}{\partial o_k} = (\alpha_k - 1)\bigg(\sum_{n=0}^{\infty}\frac{1}{(n+\alpha_k)^2} - \frac{1}{(n+S-\alpha_{gt})^2}\bigg) \tag{48}
$$

Now, when $\alpha_k \to \infty$, the value of the gradient $\frac{\partial \mathcal{L}^{\texttt{EDL}}_{\texttt{reg}}(\boldsymbol{x},\boldsymbol{y})}{\partial o_k} \to 0$. There is close to zero model update from regularization for very large incorrect evidence.

Case II: $\texttt{SoftPlus}(.)$ to transform logits to evidence. The gradients with respect to the logits o_k is given by the sigmoid i.e. $\frac{\partial e_k}{\partial o_k} = \text{sigmoid}(o_k)$, $\lim_{o_k \to \infty}\frac{\partial e_k}{\partial o_k} = 1$, and

$$
\frac{\partial \mathcal{L}^{\texttt{EDL}}_{\texttt{reg}}(\boldsymbol{x},\boldsymbol{y})}{\partial o_k} = (\alpha_k - 1)\bigg(\sum_{n=0}^{\infty}\frac{1}{(n+\alpha_k)^2} - \frac{1}{(n+S-\alpha_{gt})^2}\bigg)\sigma(\alpha_k - 1) \tag{49}
$$

Now, similar to $\texttt{ReLU}$, when $\alpha_k \to \infty$, the value of the gradient $\frac{\partial \mathcal{L}^{\texttt{EDL}}_{\texttt{reg}}(\boldsymbol{x},\boldsymbol{y})}{\partial o_k} \to 0$. There is close to zero model update from regularization for very large incorrect evidence.

Case III: $\exp(.)$ to transform logits to evidence. When using exponential non-linearity to transform the neural network output to evidence, the α_k is given by $\alpha_k = \exp(o_k) + 1$, $\frac{\partial \alpha_k}{\partial o_k} = \alpha_k - 1$. Now the gradients with respect to the neural network output o_k becomes:

$$
\frac{\partial \mathcal{L}^{2}_{reg}(\boldsymbol{x},\boldsymbol{y})}{\partial o_k} = \frac{\partial \mathcal{L}^{2}_{reg}(\boldsymbol{x},\boldsymbol{y})}{\partial \alpha_k} \times \frac{\partial \alpha_k}{\partial o_k} = (\alpha_k - 1)^2\bigg(\sum_{n=0}^{\infty}\frac{1}{(n+\alpha_k)^2} - \frac{1}{(n+S-\alpha_{gt})^2}\bigg) \tag{50}
$$

Here, the gradient values increase as $\alpha_k \to \infty$, and the gradient values do not vanish. Simply, as the incorrect evidence becomes very large, the model updates also become large in the accurate direction.

Thus, considering Case I, II, and II, we see that the incorrect evidence-based regularization with forward KL divergence is not effective in regions of incorrect evidence when using `ReLu` and `SoftPlus` functions to transform logits to evidence. This issue of correcting very large incorrect evidence does not appear when using exp function to transform the logits into evidence.

2. Gradient for ADL regularization ((Shi et al., 2020))

$$\mathcal{L}_{\texttt{reg}}^{\texttt{ADL}}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{k=1}^{K} \big(\mathbf{e} \odot (\mathbf{1} - \mathbf{y})\big)_k = \sum_{k=1}^{K} e_k \times (1 - y_k) = S - K - \alpha_{gt} + 1 \tag{51}$$

Considering the gradient of the regularization with respect to the parameters $\alpha_k, k \neq gt$, and corresponding logits o_k, we get

$$\frac{\partial \mathcal{L}_{\texttt{reg}}^{\texttt{ADL}}(\boldsymbol{x}, \boldsymbol{y})}{\partial \alpha_k} = 1 \quad \implies \frac{\partial \mathcal{L}_{\texttt{reg}}^{\texttt{ADL}}(\boldsymbol{x}, \boldsymbol{y})}{\partial o_k} = \frac{\partial e_k}{o_k} \tag{52}$$

When considering the exp function to transform logits to evidence, $\frac{\partial e_k}{o_k} = e_k = \exp(o_k)$ and the gradient value becomes very large when the model's predicted incorrect evidence value is large. This may lead to exploding gradients and stability issues in the model training. For `ReLU` and `SoftPlus` functions, the gradients in positive evidence regions are $\frac{\partial e_k}{o_k} = 1$, and $\frac{\partial e_k}{o_k} = \sigma(o_k)$ respectably. Thus, the gradient and corresponding model updates for high incorrect evidence are as desired.

3. Gradient analysis of incorrect belief regularization term as in Units-ML(Pandey & Yu, 2022a)

$$\mathcal{L}_{\texttt{reg}}^{\texttt{Units}}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{k=1}^{K} \big(\frac{\mathbf{e}}{S} \odot (\mathbf{1} - \mathbf{y})\big)_k = \sum_{k=1}^{K} \frac{e_k}{S} \times (1 - y_k) = \frac{1}{S}\big(S - K - \alpha_{gt} + 1\big) \tag{53}$$

The regularization value is bounded to be in a range of $[0, 1]$ for all the data samples, no matter how severe the mistake which may limit its effectiveness. Next, the gradient of the regularization with respect to the parameters α_k, and logits o_k is given by

$$\frac{\partial \mathcal{L}_{\texttt{reg}}^{\texttt{Units}}(\boldsymbol{x}, \boldsymbol{y})}{\partial \alpha_k} = \frac{\partial\Big(\frac{1}{S}\big(S - K - \alpha_{gt} + 1\big)\Big)}{\partial \alpha_k} = \frac{\alpha_{gt} + K - 1}{S^2} = \frac{e_{gt} + K}{(K + \sum_{k=1}^{K} e_k)^2} \tag{54}$$

$$\frac{\partial \mathcal{L}_{reg}^{3}(\boldsymbol{x}, \boldsymbol{y})}{\partial o_k} = \frac{\partial \mathcal{L}_{\texttt{reg}}^{\texttt{Units}}(\boldsymbol{x}, \boldsymbol{y})}{\partial \alpha_k} \times \frac{\partial \alpha_k}{\partial o_k} = \frac{e_{gt} + K}{S^2} \times \frac{\partial e_k}{\partial o_k} \tag{55}$$

The gradient value decreases as the number of classes K in the classification problem increases. For all three transformations: `ReLU`, `SoftPlus`, and exp to transform logits to evidence, the gradients will go to zero as the incorrect evidence increases i.e. $e_k \to \infty$ and $S \to \infty \implies \frac{\partial \mathcal{L}_{reg}^{3}(\boldsymbol{x}, \boldsymbol{y})}{\partial o_k} \to 0$. So, the regularization may be ineffective when the incorrect evidence is very high.

D. Impact of Non-linear Transformation

Theorem 2: For a data sample $\mathbf{x}$, if an evidential model outputs logits $\mathbf{o}_k \leq 0 \; \forall k \in [0, K]$, the exponential activation function leads to a larger gradident update on the model parameters than `softplus` and `ReLu`.

Proof. Consider an evidential loss $\mathcal{L}$, which is formally defined in Eqns. (21), (22), and (23), is used to train the evidential model, let $\mathbf{o}, \mathbf{e} \in \mathbb{R}^K$ denote the neural network output vector before applying the activation $\mathcal{A}$, and the evidence vector, respectively, for a network with weight w. For a data sample $\mathbf{x}$, if the network outputs $o_k < 0, \forall k \in [K]$, we have:

1. `ReLu`:

$$\frac{\partial \mathcal{L}_1}{\partial w} = \sum_k \frac{\partial \mathcal{L}_1}{\partial e_k} \frac{\partial e_k}{\partial o_k} \frac{\partial o_k}{\partial w} = 0 \qquad \text{(see Eqn. 8)},$$

2. `SoftPlus`:

$$\frac{\partial \mathcal{L}_2}{\partial w} = \sum_k \frac{\partial \mathcal{L}_2}{\partial e_k} \frac{\partial e_k}{\partial o_k} \frac{\partial o_k}{\partial w} = \sum_k \frac{\partial \mathcal{L}_2}{\partial e_k} \frac{\partial o_k}{\partial w} \text{Sigmoid}(o_k) \qquad \text{(see Eqn. 9)},$$

3. Exponential:

$$\frac{\partial \mathcal{L}_3}{\partial w} = \sum_k \frac{\partial \mathcal{L}_3}{\partial e_k}\frac{\partial e_k}{\partial o_k}\frac{\partial o_k}{\partial w} = \sum_k \frac{\partial \mathcal{L}_3}{\partial e_k}\frac{\partial o_k}{\partial w}\exp(o_k) = \sum_k \frac{\partial \mathcal{L}_3}{\partial e_k}\frac{\partial o_k}{\partial w}\{[1+\exp(o_k)]\text{Sigmoid}(o_k)\} \qquad \text{(see Eqn. 10)}$$

Thus, we have $\frac{\partial \mathcal{L}_3}{\partial w} \geq \frac{\partial \mathcal{L}_2}{\partial w} \geq \frac{\partial \mathcal{L}_1}{\partial w}$, which implies that $\mathcal{A} = \exp$ leads to a larger update to the network than both Softplus and ReLu. This completes the proof. Now we carry out an analysis of the three activations. □

Analysis:

Consider a representative $K-$class evidential classification model that trains using Type II Maximum Likelihood evidential loss. Consider an input $\mathbf{x}$ with one-hot label of $\mathbf{y}$, $\sum_{k=1}^{K} y_k = 1$. For this evidential framework, the Type II Maximum Likelihood loss ($\mathcal{L}^{\text{Log}}(\mathbf{x},\mathbf{y})$) and its gradient with the logits $\mathbf{o}$ (Eqn. 34) are given by

$$\mathcal{L}^{\text{Log}}(\mathbf{x},\mathbf{y}) = \log S - \sum_{k=1}^{K} y_k \log \alpha_k \quad \& \quad \text{grad}_k = \frac{\partial \mathcal{L}^{\text{Log}}(\mathbf{x},\mathbf{y})}{\partial o_k} = \Big(\frac{1}{S} - \frac{y_k}{\alpha_k}\Big)\frac{\partial e_k}{\partial o_k} \tag{56}$$

Case I and II: `ReLU(.)` and `SoftPlus(.)` to transform logits to evidence.

- **Zero evidence region:** For `ReLU(.)` based evidential models, if the logits value for class k i.e. o_k is negative, then the corresponding evidence for class k i.e. $e_k = 0$, $\frac{\partial e_k}{\partial o_k} = 0$ & $\text{grad}_k = \frac{\partial \mathcal{L}^{\text{Log}}(\mathbf{x},\mathbf{y})}{\partial o_k} = 0$. So, there is no update to the model through the nodes that output negative logits value. In the case of `SoftPlus`(.) based evidential models, there is no update to the model when training samples lie in zero-evidence regions. This is possible in the condition of $o_k \to -\infty$. In other cases, there will be some small finite small update in the accurate direction from the gradient.
- **Range of gradients:** The range of gradients for both `ReLU(.)` and `SoftPlus`(.) based evidential models are identical. Considering the gradient for the ground truth node $i.e. y_k = 1$, the range of gradients is $[\frac{1}{K} - 1, 0]$. For all other nodes other than the ground truth node i.e. $y_k = 0$, the range of gradients is $[0, \frac{1}{K}]$. So, for classification problems with a large number of classes, the gradient updates to the nodes that do not correspond to the ground truth class will be bounded in a small range and is likely to be very small.
- **High incorrect evidence region:** If the evidence for class k is very large i.e. $e_k \to \infty$, then for `ReLU(.)`, $\frac{\partial e_k}{o_k} = 1$, and for `SoftPlus(.)`, $\frac{\partial e_k}{o_k} = \text{Sigmoid}(o_k) \to 1$, $\frac{1}{\alpha_k} = \frac{1}{e_k+1} \to 0$, $\frac{1}{S} \to 0$, & $\text{grad}_k = \frac{\partial \mathcal{L}^{\text{Log}}(\mathbf{x},\mathbf{y})}{\partial o_k} \to 0$. For large positive model evidence, there is no update to the corresponding node of the neural network. The evidence can be further broken down into correct evidence (corresponding to the evidence for the ground truth class), and incorrect evidence (corresponding to the evidence for any other class other than the ground truth class). When the correct class evidence is large, the corresponding gradient is close to zero and there is no update to the model parameters which is desired. When the incorrect evidence is large, the model should be updated to minimize such incorrect evidence. However, the evidential models with `ReLU` and `Softplus` fail to minimize incorrect evidence when the incorrect evidence value is large. These necessities the need for incorrect evidence regularization terms.

Case III: $\exp(.)$ to transform logits to evidence. Considering Eqn. Eqn. 34 and Eqn 26, the gradient of the loss with respect to the logits becomes

$$\text{grad}_k = \frac{\partial \mathcal{L}^{\text{Log}}(\mathbf{x},\mathbf{y})}{\partial o_k} = \Big(\frac{1}{S} - \frac{y_k}{\alpha_k}\Big)(e_k) = \Big(\frac{1}{S} - \frac{y_k}{\alpha_k}\Big)(\alpha_k - 1) \tag{57}$$

- **Zero evidence region:** In case of $\exp(.)$ based evidential models, except in the extreme cases of $\alpha_k \to \infty$, there will be some signal to guide the model. In cases outside the zero-evidence region (i.e. outside $\alpha_k \to \infty$), there will be some finite small update in the accurate direction from the gradient. Moreover, for same evidence values, the gradient of $\exp$ based model is larger than the `SoftPlus` based evidential model by a factor of $1 + \exp(o_k)$. Compared to `SoftPlus` models, the larger gradient is expected to help the model learn faster in low-evidence regions.
- **Range of gradients:** For the ground truth node $i.e. y_k = 1$, the range of gradients is $[-1, 0]$. For all nodes other than the ground truth node i.e. $y_k = 0$, the range of gradients is $[0, 1]$. Thus, the gradients are expected to be more expressive and accurate in guiding the evidential model compared to `ReLU` and `SoftPlus` based evidential models.

- **High evidence region:** If the evidence for class k is very large i.e. $e_k \to \infty$, then $\alpha_k - 1 \approx \alpha_k$ and $\text{grad}_k = \texttt{sm}_k - y_k$. In other words, the model's gradient updates become identical to the standard classification model (see Section A) without any learning issues.

Due to smaller zero-evidence region, more expressive gradients, and no issue of learning in high incorrect evidence region, the exponential-based evidential models are expected to be more effective compared to `ReLU` and `SoftPlus` based evidential models.

E. Analysis of Evidential Losses

Here, we analyze the three variants of evidential loss. As seen in Section D, exp function is expected to be superior to `ReLU` and `SoftPlus` functions to transform the logits to evidence. Thus, in this section, we consider exp function to transform the logits into evidence. However, the analysis holds true for all three functions.

1. Bayes risk with the sum of squares loss (Eqn. 21)

$$\mathcal{L}^{\texttt{MSE}}(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{K} (y_j - \frac{\alpha_j}{S})^2 + \frac{\alpha_j(S - \alpha_j)}{S^2(S+1)} \tag{58}$$

The loss can be simplified as

$$\begin{aligned}
\mathcal{L}^{\texttt{MSE}}(\mathbf{x}, \mathbf{y}) &= \sum_{j=1}^{K} (y_j - \frac{\alpha_j}{S})^2 + \frac{\alpha_j(S - \alpha_j)}{S^2(S+1)} && (59)\\
&= 1 - \frac{2\alpha_{gt}}{S} + \frac{\sum_k \alpha_k^2}{S^2} + \frac{2\sum_i \sum_j \alpha_i \alpha_j}{S^2(S+1)} && (60)\\
&= 1 - \frac{2\alpha_{gt}}{S} + \frac{\sum_k \alpha_k^2 + 2\sum_i \sum_j \alpha_i \alpha_j}{S^2} + \frac{2\sum_i \sum_j \alpha_i \alpha_j}{S^2(S+1)} - \frac{2\sum_i \sum_j \alpha_i \alpha_j}{S^2} && (61)\\
&= 2 - \frac{2\alpha_{gt}}{S} + \frac{2\sum_i \sum_j \alpha_i \alpha_j}{S^2}\Big[\frac{1}{(S+1)} - 1\Big] && (62)\\
&= 2 - \frac{2\alpha_{gt}}{S} - \frac{2\sum_i \sum_j \alpha_i \alpha_j}{S(S+1)} && (63)
\end{aligned}$$

The range of the two components in the loss is $0 \leq \frac{2\alpha_{gt}}{S} + \frac{2\sum_i \sum_j \alpha_i \alpha_j}{S(S+1)} \leq 2$ and the loss is bounded in the range $[0, 2]$. In other words, the loss for any sample in the entire sample space is bounded in the range of $[0, 2]$ no matter how severe the mistake is. Such bounded loss is expected to restrict the model's learning capacity.

2. Bayes risk with cross-entropy loss (Eqn. 22)

$$\mathcal{L}^{\texttt{CE}}(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{K} y_k \Big(\Psi(S) - \Psi(\alpha_k)\Big) = \Psi(S) - \Psi(\alpha_{gt}) \tag{64}$$

Where $\Psi(.)$ is the Digamma function, and Γ is the gamma function. The functions and their gradients are defined as

$$\begin{aligned}
\Gamma(z) &= \frac{e^{-\gamma z}}{z} \prod_{n=1}^{\infty} \Big(1 + \frac{z}{n}\Big)^{-1} e^{\frac{z}{n}} && (65)\\
\Psi(z) &= \frac{d}{dz} \log \Gamma(z) = \frac{d}{dz}\bigg(- \gamma z - \log z + \sum_{n=1}^{\infty} \Big(\frac{z}{n} - \log\big(1 + \frac{z}{n}\big)\Big)\bigg) && (66)\\
&= -\gamma - \frac{1}{z} + \sum_{n=1}^{\infty} \frac{1}{n} - \frac{1}{n+z} && (67)\\
\frac{\partial \Psi(z)}{\partial z} &= \frac{\partial}{\partial z}\bigg(- \gamma - \frac{1}{z} + \sum_{n=1}^{\infty} \frac{1}{n} - \frac{1}{n+z}\bigg) = \frac{1}{z^2} + \sum_{n=1}^{\infty} \frac{1}{(n+z)^2} && (68)
\end{aligned}$$

Now, the Bayes risk with cross-entropy loss becomes

$$\mathcal{L}^{\text{CE}}(\mathbf{x}, \mathbf{y}) = \Psi(S) - \Psi(\alpha_{gt}) \tag{69}$$

$$= \frac{1}{\alpha_{gt}} - \frac{1}{S} + S \sum_{n=1}^{\infty} \frac{1}{n(n+S)} - \alpha_{gt} \sum_{n=1}^{\infty} \frac{1}{n(n+\alpha_{gt})} \tag{70}$$

Both the infinite sums ($\sum_{n=1}^{\infty} \frac{1}{n(n+S)}$ and $\sum_{n=1}^{\infty} \frac{1}{n(n+\alpha_{gt})}$) converge and lie in the range of 0 to $\frac{\pi^2}{6}$. The minimum possible value of this loss is 0 when $\alpha_{gt} \to \infty \& S \approx \alpha_{gt}$. The maximum possible value is ∞ when only $S \to \infty$. The loss lies in the range $[0, \infty]$ and is more expressive compared to MSE-based evidential loss.

Considering the gradient of the loss with respect to the ground truth node (i.e. $\alpha_{gt}, y_{gt} = 1$),

$$\frac{\partial \mathcal{L}^{\text{CE}}(\mathbf{x}, \mathbf{y})}{\partial \alpha_{gt}} = \frac{\partial}{\partial \alpha_{gt}} \Psi(S) - \Psi(\alpha_{gt}) = \frac{1}{S^2} + \sum_{n=1}^{\infty} \frac{1}{(n+S)^2} - \frac{1}{\alpha_{gt}^2} - \sum_{n=1}^{\infty} \frac{1}{(n+\alpha_{gt})^2} \tag{71}$$

As $\alpha_{gt} < S$, the gradient is always negative. Thus, the model aims to maximize the correct evidence α_{gt}. Considering the gradient of the loss with respect to nodes not corresponding to the ground truth (i.e. $\alpha_k, k \neq gt, y_k = 0$),

$$\frac{\partial \mathcal{L}^{\text{CE}} \mathbf{x}, \mathbf{y})}{\partial \alpha_k} = \frac{\partial}{\partial \alpha_k} \Psi(S) - \Psi(\alpha_{gt}) = \frac{\partial \Psi(S)}{\partial S} \frac{\partial S}{\partial \alpha_k} = \frac{1}{S^2} + \sum_{n=1}^{\infty} \frac{1}{(n+S)^2} \tag{72}$$

$$\frac{\partial \mathcal{L}^{\text{CE}} \mathbf{x}, \mathbf{y})}{\partial o_k} = \frac{\partial \mathcal{L}^{\text{CE}} \mathbf{x}, \mathbf{y})}{\partial \alpha_k} \times \frac{\alpha_k}{o_k} = \Big(\frac{1}{S^2} + \sum_{n=1}^{\infty} \frac{1}{(n+S)^2} \Big) (\alpha_k - 1) \tag{73}$$

The gradient at nodes that do not correspond to ground truth is always non-negative. However, this gradient is also minimum and 0 when $S \to \infty \,\&\, \alpha_k \to \infty$. This is an undesired behavior as the model may be encouraged to always increase the evidence for all the classes. Moreover, the gradient is zero and there is no update to the nodes when $S \to \infty, \& \,\alpha_k \to \infty$. So, the incorrect evidence regularization to penalize the incorrect evidence is essential for the evidential model trained with this loss.

3. Type II Maximum Likelihood loss (Eqn. 23)

$$\mathcal{L}^{\text{Log}}(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^{K} y_k \Big(\log(S) - \log(\alpha_k) \Big) = \log(S) - \log(\alpha_{gt}) \tag{74}$$

The loss is bounded in the range of $[0, \infty]$ as the loss is minimum and 0 when $\alpha_{gt} \to S \to \infty$, and maximum loss when $\alpha_{gt} << S \,\&\, S \to \infty$. Thus, the loss is more expressive compared to MSE based evidential loss. Now, the gradient of the loss is given by

$$\frac{\partial \mathcal{L}^{\text{Log}}(\mathbf{x}, \mathbf{y})}{\partial o_k} = \frac{1}{S} \frac{\partial S}{\partial o_k} - y_k \frac{1}{\alpha_k} \frac{\partial \alpha_k}{\partial o_k} = \Big(\frac{1}{S} - \frac{y_k}{\alpha_k} \Big) \frac{\partial e_k}{\partial o_k} = \Big(\frac{1}{S} - \frac{y_k}{\alpha_k} \Big) (\alpha_k - 1) \tag{75}$$

Here, when $S \to \infty \,\&\, \alpha_k \to \infty$, the gradient becomes $\frac{\partial \mathcal{L}^{\text{Log}}(\mathbf{x}, \mathbf{y})}{\partial o_k} \to (1 - y_k)$. This is highly desirable behavior for the model as it aims to minimize the evidence for the incorrect class and there will be no update to the node corresponding to the ground truth class if $\alpha_k = \alpha_{gt}, y_{gt} = 1$. Thus, the Type II based issue is expected to be superior to the other two losses as the range of loss is optimal (i.e. in the range $[0, \infty]$), and no learning issue arises for samples with high incorrect evidence.

F. Additional Experiments and Results

We first present the details of the models, hyperparameter settings, clarification regarding dead neuron issue, and experiments used in the work in Section F.1. We then present additional results and discussions, including Few-shot classification, and 200-class tiny-ImageNet Classification results, that show the effectiveness of the proposed model RED in Section F.3. Finally discuss some limitations and potential future works in Section F.4.

F.1. Hyperparameter details

For Table 1 results, $\lambda_1 = 1.0$ was used for MNIST experiments, $\lambda_1 = 0.1$ was used for Cifar10 experiments, and $\lambda_1 = 0.001$ was used for Cifar100 experiments. Table 8, 9, and 10 present complete results across the hyperparameter values and experiment settings. MNIST model was trained on the LeNet model (Sensoy et al., 2018) for 50 epochs, and Cifar10/Cifar100 models were trained on Resnet-18 based classifier (He et al., 2016) for 200 epochs. Few-shot classification experiments were carried out with $\lambda_1 = 0.1$ using Resnet-12 based classifier (Chen et al., 2021). All results presented in this work are from local reproduction. MNIST models were trained with learning rate of 0.0001 and Adam optimizer (Kingma & Ba, 2014), and all remaining models were trained with learning rate of 0.1 and Stochastic Gradient Descent optimizer with momentum. Tabular results represent the mean and standard deviation from 3 independent runs of the model. In the proposed model RED, correct evidence regularization is weighted by the parameter $\lambda_{\texttt{cor}}$ whose value is given by the predicted vacuity ν. $\lambda_{\texttt{cor}}$ is treated as hyperparameter, *i.e.,* constant weighting term in the loss during model update.

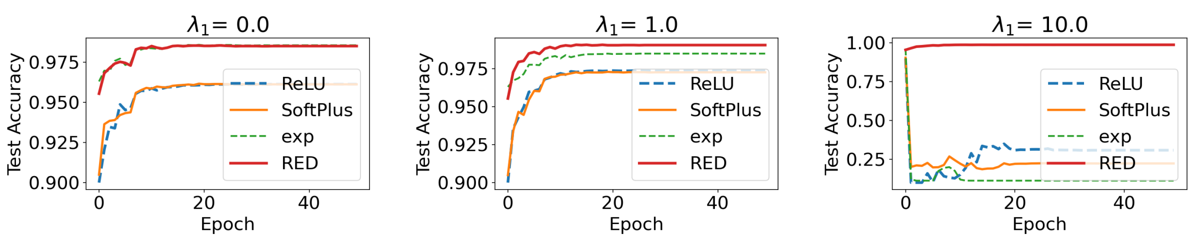
F.2. Dead Neuron Issue Clarification

Instead of using ReLU as an activation function in a standard deep neural network, evidential models introduce ReLU as non-negative transformation function in the output layer to ensure that the predicted evidence is non-negative to satisfy the requirement of evidential theory. This non-negative evidence vector parameterizes a Dirichlet prior for fine-grained uncertainty quantification that covers second-order uncertainty, including vacuity and dissonance. We theoretically and empirically show the learning deficiency of ReLU based evidential models and justify the advantage of using an exponential function to output (non-negative) evidence. We further introduce a correct evidence regularization term in the loss that addresses the learning deficiency from zero-evidence samples. The "dead neuron" issue in the activation functions has been studied, and ReLU variations such as Exponential Linear Unit, Parametric ReLU, and Leaky ReLU have been developed to address the issue. But, these activation functions will not be theoretically sound in the evidential framework as they are can lead to negative evidences. In this case, they can not serve as Dirichlet parameters that are interpreted as pseudo counts.

F.3. Effectiveness of Regularized Evidential Model (RED)

F.3.1. Evidential Activation Function.

In this section, we present additional results (for section 5.2) with the MNIST classification problem using the LeNet model to empirically validate Theorem 2. We carry out experiments for evidential models trained using all three evidential losses: Evidential MSE loss in (21), Evidential cross-entropy loss in (22), and Evidential Log loss in (23) with $\lambda_1 = \{0.0, 1.0, \&10.0\}$. As can be seen in Figure 15, 16, and 17, using $\exp$ activation for transforming logits to evidence leads to superior performance in all settings compared to ReLU and Softplus based evidential models that empirically validates Theorem 2.

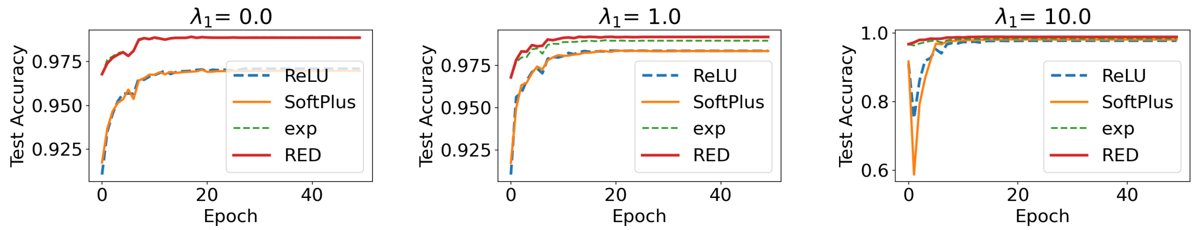


(a) Trend for $\lambda_1 = 0.0$ (b) Trend for $\lambda_1 = 1.0$ (c) Trend for $\lambda_1 = 10.0$

Figure 15. Impact of Evidential Activation to the test set accuracy of the model trained with MSE based evidential loss (Eqn. 21)

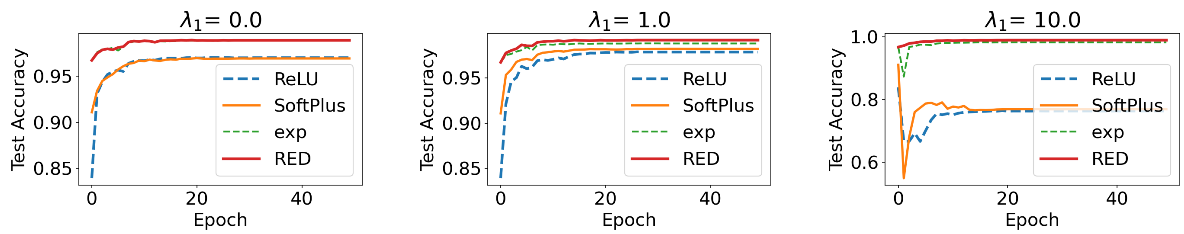
F.3.2. Correct Evidence Regularization

We introduce the novel correct evidence regularization term to train the evidential model (Section 4.1). In this section, we present additional results for the evidential model that uses $\exp$ activation. We trained the model using evidential losses with different incorrect evidence regularization strengths ($\lambda_1 = 0,\ 1.0\ \&\ 10.0$). As can be seen(Figure 18, and 19), the model with proposed correct-evidence regularization leads to improved generalization compared to the baseline model



(a) Trend for $\lambda_1 = 0.0$ (b) Trend for $\lambda_1 = 1.0$ (c) Trend for $\lambda_1 = 10.0$

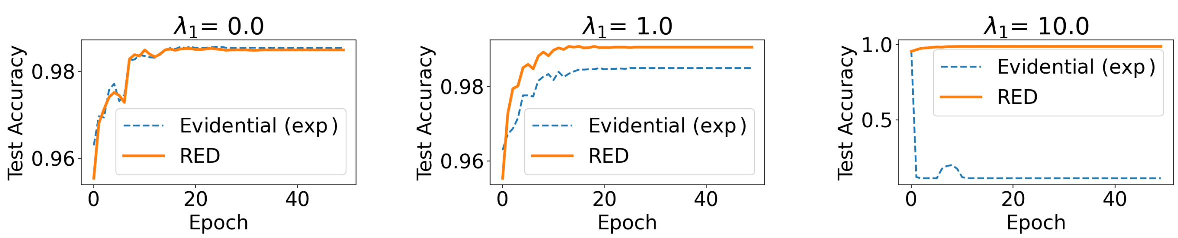
Figure 16. Impact of Evidential Activation to test set accuracy of the model trained with cross-entropy based evidential loss (Eqn. 22)



(a) Trend for $\lambda_1 = 0.0$ (b) Trend for $\lambda_1 = 1.0$ (c) Trend for $\lambda_1 = 10.0$

Figure 17. Impact of Evidential Activation to the test set accuracy of the model trained with Type II based evidential loss (Eqn. 23)

as the proposed correct-evidence regularization term enables the evidential model to learn from zero-evidence samples instead of ignoring them. Moreover, even though strong incorrect evidence regularization hurts both model's generalization, the proposed regularization leads to a more robust model that generalizes better. Finally, the MSE-based evidential model is hurt the most with strong incorrect evidence regularization as thee MSE based evidential loss is bounded in the range $[0, 2]$, and the incorrect evidence-regularization term may easily dominate the overall loss compared to other evidential losses. This can be seen in Figure 18(c) where the incorrect evidence regularization strength is large i.e. $\lambda_1 = 10.0$ and the evidential model fails to train. Due to strong incorrect evidence regularization, the model may have learned to map all training samples to zero-evidence region. However, with the proposed regularization, the model continues to learn and achieves good generalization performance.

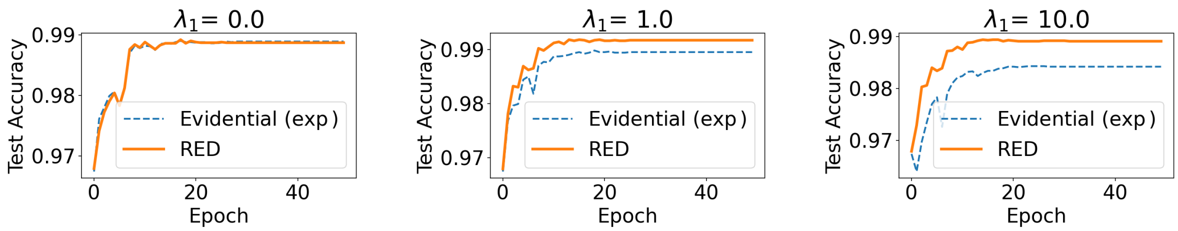


(a) Trend for $\lambda_1 = 0.0$ (b) Trend for $\lambda_1 = 1.0$ (c) Trend for $\lambda_1 = 10.0$

Figure 18. Impact of proposed Correct Evidence Regularization to the test set accuracy of the evidential model(Trained with Eqn. 21)

F.3.3. Few-Shot Classification Experiments

Ideas presented in this work address the fundamental limitation of evidential classification framework that enables the evidential model to acquire knowledge from all the training samples. Using these ideas, evidential framework can be extended to challenging classification problems to the reasonable predictive performance. To this end, we experiment with few-shot classification using 1-shot and 5-shot classification for the *mini*-ImageNet dataset (Vinyals et al., 2016). We



(a) Trend for $\lambda_1 = 0.0$ (b) Trend for $\lambda_1 = 1.0$ (c) Trend for $\lambda_1 = 10.0$

Figure 19. Impact of proposed Correct Evidence Regularization to the test set accuracy of the evidnetial model (Trained with Eqn. 22)

consider the ResNet-12 backbone, classifier-baseline model (Chen et al., 2021), and its evidential extension. Table 7 shows the results for 1-shot and 5-shot classification experiments. As can be seen, the `ReLU` and `Softplus` based evidential models have suboptimal performance as they avoid many training samples of the zero-evidence region. In contrast, the exp model has a better learning capacity that leads to superior performance. Finally, the proposed model RED can learn from all training samples, which leads to the best generalization performance among all the evidential models.

Table 7. Few-Shot Classification Accuracy comparison: *mini*-ImageNet dataset

Standard CE Model: 1 Shot: $57.9_{\pm 0.2}\%$; 5-Shot: $76.9_{\pm 0.2}\%$				
1-Shot Experiments				
Regularization	`ReLU`	`SoftPlus`	exp	**RED (Ours)**
$\lambda_1 = 0.000$	$38.78_{\pm 3.75}$	$51.60_{\pm 0.40}$	$57.11_{\pm 0.09}$	$56.27_{\pm 0.15}$
$\lambda_1 = 0.100$	$31.15_{\pm 1.69}$	$48.87_{\pm 0.21}$	$56.43_{\pm 0.03}$	$\mathbf{58.03_{\pm 0.39}}$
$\lambda_1 = 1.000$	$20.00_{\pm 0.00}$	$43.81_{\pm 0.56}$	$27.43_{\pm 0.88}$	$54.68_{\pm 0.45}$
5-Shot Experiments				
Regularization	`ReLU`	`SoftPlus`	exp	**Ours**
$\lambda_1 = 0.000$	$52.66_{\pm 5.32}$	$67.22_{\pm 0.17}$	$75.87_{\pm 0.09}$	$75.31_{\pm 0.13}$
$\lambda_1 = 0.100$	$43.95_{\pm 3.72}$	$66.14_{\pm 0.05}$	$74.08_{\pm 0.13}$	$\mathbf{76.05_{\pm 0.17}}$
$\lambda_1 = 1.000$	$20.00_{\pm 0.00}$	$61.96_{\pm 0.61}$	$34.01_{\pm 1.46}$	$72.32_{\pm 0.20}$

F.3.4. Complex Dataset/Model Experiments

We also carry out experiment for a challenging 200-class classification problem over Tiny-ImageNet based on (Huynh, 2022). We adapt the Swin Transformer to be evidential, and train all the models for 20 epochs with Evidential log loss (Eqn. 23). In this setting, ReLU based evidential model achieves 85.25% accuracy, softplus based model achieves 85.15 % accuracy, the exponential model improves over both to achieve 89.93 % accuracy, and our proposed model RED outperforms all the evidential models to achieve the greatest accuracy of 90.14%, empirically validating our theoretical analysis.

F.4. Limitations and Future works

We carried out a theoretical investigation of the Evidential Classification models to identify their fundamental limitation: their inability to learn from *zero evidence regions*. The empirical study in this work is based on classification problems. We next plan to extend the ideas to develop Evidential Segmentation and Evidential Object Detection models. Moreover, this work identifies limitations of Evidential MSE loss in (21), and we plan to carry out a thorough theoretical analysis to analyze other evidential losses given in (23) and (22)). The proposed evidential model, similar to existing evidential classification models, requires hyperparameter tuning for λ_1 i.e. the incorrect evidence regularization hyperparameter.

In addition, extending evidential models to noisy and incomplete data settings and investigating the benefits of leveraging uncertainty information could be interesting future work. Finally, It will be an interesting future work to extend the analysis and evidential models to tasks beyond classification, for instance to build effective evidential segmentation and object detection models.

Table 8. Classification performance comparison: MNIST dataset

Standard CE Model: $99.21_{\pm 0.03}\%$				
Log loss				
Regularization	`ReLU`	`SoftPlus`	exp	**RED (Ours)**
$\lambda_1 = 0.000$	$97.06_{\pm 0.19}$	$97.07_{\pm 0.24}$	$98.85_{\pm 0.03}$	$98.82_{\pm 0.04}$
$\lambda_1 = 1.000$	$98.19_{\pm 0.08}$	$98.21_{\pm 0.05}$	$98.79_{\pm 0.02}$	$\mathbf{99.10_{\pm 0.02}}$
$\lambda_1 = 10.000$	$83.17_{\pm 4.54}$	$80.37_{\pm 18.70}$	$98.14_{\pm 0.07}$	$98.84_{\pm 0.03}$
Evidential CE loss				
$\lambda_1 = 0.000$	$97.03_{\pm 0.21}$	$97.09_{\pm 0.21}$	$98.84_{\pm 0.02}$	$98.81_{\pm 0.01}$
$\lambda_1 = 1.000$	$98.27_{\pm 0.02}$	$98.36_{\pm 0.02}$	$98.87_{\pm 0.03}$	$\mathbf{99.12_{\pm 0.02}}$
$\lambda_1 = 10.000$	$97.46_{\pm 1.02}$	$97.14_{\pm 1.42}$	$98.31_{\pm 0.07}$	$98.84_{\pm 0.04}$
Evidential MSE loss				
$\lambda_1 = 0.000$	$96.18_{\pm 0.02}$	$96.20_{\pm 0.03}$	$98.42_{\pm 0.03}$	$98.41_{\pm 0.06}$
$\lambda_1 = 1.000$	$97.41_{\pm 0.22}$	$97.45_{\pm 0.16}$	$98.35_{\pm 0.05}$	$\mathbf{99.02_{\pm 0.00}}$
$\lambda_1 = 10.000$	$19.93_{\pm 6.98}$	$27.14_{\pm 6.37}$	$27.17_{\pm 3.72}$	$98.76_{\pm 0.03}$

Table 9. Classification performance comparison: Cifar10 Dataset

Standard CE Model: $95.43_{\pm 0.02}\%$				
Log loss				
Regularization	`ReLU`	`SoftPlus`	exp	**RED (Ours)**
$\lambda_1 = 0.000$	$43.83_{\pm 14.60}$	$95.19_{\pm 0.10}$	$95.35_{\pm 0.02}$	$95.03_{\pm 0.14}$
$\lambda_1 = 0.100$	$41.43_{\pm 19.60}$	$95.18_{\pm 0.11}$	$95.11_{\pm 0.10}$	$\mathbf{95.24_{\pm 0.06}}$
$\lambda_1 = 1.000$	$38.42_{\pm 15.64}$	$94.94_{\pm 0.22}$	$93.95_{\pm 0.06}$	$94.78_{\pm 0.17}$
$\lambda_1 = 10.000$	$10.00_{\pm 0.00}$	$32.42_{\pm 6.99}$	$23.29_{\pm 5.24}$	$90.96_{\pm 0.35}$
$\lambda_1 = 50.000$	$10.00_{\pm 0.00}$	$10.00_{\pm 0.00}$	$12.47_{\pm 3.49}$	$65.09_{\pm 0.74}$
Evidential CE loss				
$\lambda_1 = 0.000$	$79.19_{\pm 16.06}$	$95.32_{\pm 0.17}$	$95.38_{\pm 0.10}$	$\mathbf{95.40_{\pm 0.14}}$
$\lambda_1 = 0.100$	$75.97_{\pm 20.56}$	$95.12_{\pm 0.05}$	$95.33_{\pm 0.03}$	$95.08_{\pm 0.07}$
$\lambda_1 = 1.000$	$75.83_{\pm 20.74}$	$94.99_{\pm 0.08}$	$94.65_{\pm 0.04}$	$94.74_{\pm 0.11}$
$\lambda_1 = 10.000$	$10.00_{\pm 0.00}$	$89.63_{\pm 0.38}$	$56.54_{\pm 4.80}$	$91.71_{\pm 0.23}$
$\lambda_1 = 50.000$	$10.00_{\pm 0.00}$	$27.03_{\pm 2.62}$	$25.33_{\pm 6.66}$	$62.98_{\pm 0.84}$
Evidential MSE loss				
$\lambda_1 = 0.000$	$\mathbf{95.43_{\pm 0.05}}$	$95.35_{\pm 0.15}$	$95.10_{\pm 0.04}$	$94.92_{\pm 0.12}$
$\lambda_1 = 0.100$	$95.15_{\pm 0.10}$	$95.04_{\pm 0.05}$	$95.14_{\pm 0.03}$	$95.03_{\pm 0.13}$
$\lambda_1 = 1.000$	$49.68_{\pm 29.48}$	$93.51_{\pm 0.03}$	$18.98_{\pm 1.82}$	$94.90_{\pm 0.20}$
$\lambda_1 = 10.000$	$10.00_{\pm 0.00}$	$10.00_{\pm 0.00}$	$10.00_{\pm 0.00}$	$90.15_{\pm 0.71}$
$\lambda_1 = 50.000$	$10.00_{\pm 0.00}$	$10.00_{\pm 0.00}$	$10.00_{\pm 0.00}$	$27.11_{\pm 24.20}$

Table 10. Classification performance comparison: Cifar100 dataset

Standard CE Model: 75.67 ± 0.11				
Log loss				
Regularization	`ReLU`	`SoftPlus`	exp	**RED (Ours)**
$\lambda_1 = 0.000$	$56.69_{\pm 5.83}$	$73.85_{\pm 0.20}$	$76.25_{\pm 0.16}$	$76.26_{\pm 0.27}$
$\lambda_1 = 0.001$	$61.27_{\pm 3.79}$	$74.48_{\pm 0.17}$	$76.12_{\pm 0.04}$	$\mathbf{76.43_{\pm 0.21}}$
$\lambda_1 = 0.010$	$54.20_{\pm 5.93}$	$75.56_{\pm 0.43}$	$76.02_{\pm 0.16}$	$76.14_{\pm 0.09}$
$\lambda_1 = 0.100$	$20.29_{\pm 4.54}$	$75.67_{\pm 0.22}$	$72.72_{\pm 0.26}$	$74.62_{\pm 0.21}$
$\lambda_1 = 1.000$	$1.00_{\pm 0.00}$	$37.60_{\pm 0.82}$	$2.59_{\pm 0.52}$	$68.62_{\pm 0.03}$
$\lambda_1 = 2.000$	$1.00_{\pm 0.00}$	$1.57_{\pm 0.35}$	$0.97_{\pm 0.06}$	$62.33_{\pm 0.52}$
Evidential CE loss				
$\lambda_1 = 0.000$	$66.37_{\pm 3.47}$	$73.73_{\pm 0.38}$	$75.91_{\pm 0.20}$	$76.19_{\pm 0.22}$
$\lambda_1 = 0.001$	$68.62_{\pm 2.41}$	$74.44_{\pm 0.08}$	$76.23_{\pm 0.09}$	$\mathbf{76.35_{\pm 0.06}}$
$\lambda_1 = 0.010$	$71.94_{\pm 0.66}$	$75.45_{\pm 0.12}$	$75.95_{\pm 0.14}$	$76.13_{\pm 0.24}$
$\lambda_1 = 0.100$	$67.25_{\pm 1.84}$	$75.75_{\pm 0.21}$	$74.02_{\pm 0.09}$	$74.69_{\pm 0.13}$
$\lambda_1 = 1.000$	$1.00_{\pm 0.00}$	$73.10_{\pm 0.20}$	$37.36_{\pm 0.73}$	$69.40_{\pm 0.16}$
$\lambda_1 = 2.000$	$1.00_{\pm 0.00}$	$52.99_{\pm 0.56}$	$12.94_{\pm 1.11}$	$63.93_{\pm 0.34}$
Evidential MSE loss				
$\lambda_1 = 0.000$	$35.76_{\pm 2.81}$	$20.45_{\pm 1.41}$	$75.70_{\pm 0.47}$	$75.55_{\pm 0.24}$
$\lambda_1 = 0.001$	$31.49_{\pm 0.31}$	$15.74_{\pm 0.47}$	$42.95_{\pm 0.76}$	$\mathbf{75.73_{\pm 0.27}}$
$\lambda_1 = 0.010$	$13.60_{\pm 2.44}$	$1.00_{\pm 0.00}$	$1.00_{\pm 0.00}$	$75.35_{\pm 0.16}$
$\lambda_1 = 0.100$	$1.00_{\pm 0.00}$	$1.00_{\pm 0.00}$	$1.00_{\pm 0.00}$	$74.00_{\pm 0.13}$
$\lambda_1 = 1.000$	$1.00_{\pm 0.00}$	$1.00_{\pm 0.00}$	$1.00_{\pm 0.00}$	$66.61_{\pm 0.46}$
$\lambda_1 = 2.000$	$1.00_{\pm 0.00}$	$1.00_{\pm 0.00}$	$1.00_{\pm 0.00}$	$63.01_{\pm 0.83}$