Uncertainty Estimation by Fisher Information-based Evidential Deep Learning

Danruo Deng [1 2] **Guangyong Chen** [3] **Yang Yu** [1 2] **Furui Liu** [3] **Pheng-Ann Heng** [1 2]

Abstract

Uncertainty estimation is a key factor that makes deep learning reliable in practical applications. Recently proposed evidential neural networks explicitly account for different uncertainties by treating the network's outputs as evidence to parameterize the Dirichlet distribution, and achieve impressive performance in uncertainty estimation. However, for high data uncertainty samples but annotated with the one-hot label, the evidence-learning process for those mislabeled classes is over-penalized and remains hindered. To address this problem, we propose a novel method, *Fisher Information-based Evidential Deep Learning* ($\mathcal{I}$-EDL). In particular, we introduce Fisher Information Matrix (FIM) to measure the informativeness of evidence carried by each sample, according to which we can dynamically reweight the objective loss terms to make the network more focus on the representation learning of uncertain classes. The generalization ability of our network is further improved by optimizing the PAC-Bayesian bound. As demonstrated empirically, our proposed method consistently outperforms traditional EDL-related algorithms in multiple uncertainty estimation tasks, especially in the more challenging few-shot classification settings.

1. Introduction

Uncertainty estimation is crucial not only for safety decision-making in high-risk domains such as medical image analysis (Seeböck et al., 2019; Nair et al., 2020) and autonomous vehicle control (Feng et al., 2018; Choi et al., 2019), but also in the general fields where data is highly heterogeneous or scarcely annotated (Gal et al., 2017; Ablain et al., 2019). Predictive uncertainty is quite diverse and can be divided into *data uncertainty, model uncertainty*, and *distributional uncertainty* (Gal, 2016; Malinin & Gales, 2018). Data uncertainty, or aleatoric uncertainty, is caused by the natural complexity of the data, such as class overlap and label noise. Model uncertainty, or epistemic uncertainty, measures the uncertainty in estimating model parameters given training data. Model and data uncertainty are sometimes referred to as reducible and irreducible uncertainty, respectively, since model uncertainty can be reduced with more training data, while data uncertainty cannot. Distributional uncertainty arises from a distribution mismatch between the training and test distributions, i.e., the test data is out-of-distribution (OOD) (Quinonero-Candela et al., 2008). Quantifying the different uncertainty is obviously a key factor that makes deep learning reliable.

Softmax is the most widely used normalization function that maps the continuous activations of the output layer to a probability distribution. However, *Softmax* is notorious for inflating the probabilities of predicted classes (Szegedy et al., 2016; Guo et al., 2017; Wilson & Izmailov, 2020). Some methods calibrate network predictions to improve the reliability of uncertainty estimation (Guo et al., 2017; Liang et al., 2018). However, these methods can still not distinguish between different types of uncertainty, which seriously limits the practical usage of deep learning in challenging domains without enough training samples. Recently notable progress has been made in estimating uncertainty in DNNs. One class of methods stems from Bayesian neural networks, which quantify uncertainty by learning a posterior over weights (Gal & Ghahramani, 2016; Ritter et al., 2018; Kristiadi et al., 2022). Other methods often combine predictions from several independently trained networks to estimate statistics for class probability distributions (Lakshminarayanan et al., 2017; Zaidi et al., 2021). However, these approaches still cannot distinguish distributional uncertainty from other uncertainties (Malinin & Gales, 2018).

To address this limitation, Dirichlet-based uncertainty models quantify different types of uncertainty by modeling the output as the concentration parameters of a Dirichlet distribution (Malinin & Gales, 2018; 2019; Nandy et al., 2020; Charpentier et al., 2020). Evidential deep learning (EDL) (Sensoy et al., 2018) adopts Dirichlet distribution and treats the output as evidence to quantify belief mass and uncer-

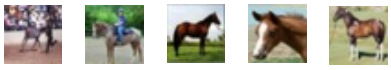
[1]Department of Computer Science and Engineering, The Chinese University of Hong Kong [2]Institute of Medical Intelligence and XR, The Chinese University of Hong Kong [3]Zhejiang Lab. Correspondence to: Guangyong Chen <gychen@zhejianglab.com>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

(a)

EDL	0.63	0.43	0.72	0.63	0.14
$\mathcal{I}$-EDL	1.54	1.49	0.98	0.82	0.47

(b)



EDL	0.23	0.12	0.15	0.35	0.09
$\mathcal{I}$-EDL	1.48	1.44	1.36	1.35	1.28

Figure 1. Data uncertainty for (a) digit "4" in MNIST, (b) "horse" in CIFAR10. $\mathcal{I}$-EDL has the ability to distinguish between hard samples (orange) and easy samples (green), but EDL cannot.

tainty by jointly considering the Dempster–Shafer Theory of Evidence (DST) (Dempster, 1968; Shafer, 1976) and subjective logic (SL) (Jøsang, 2016). The evidential network proposed by EDL can be represented as a probabilistic graphical model, where the observed labels $\boldsymbol{y}$ are generated from the Dirichlet distribution with its parameter $\boldsymbol{\alpha}$ calculated by passing the input sample $\boldsymbol{x}$ through the network. EDL, which learns optimal parameters by maximizing the expected likelihood of the observed labels, shows an impressive performance in uncertainty quantification and is widely used in various applications, such as graph neural networks (Zhao et al., 2020), open set recognition (OSR) (Bao et al., 2021; 2022), molecular property prediction (Soleimany et al., 2021), meta-learning (Pandey & Yu, 2022), and active learning (Hemmer et al., 2022). However, for samples with high data uncertainty but annotated with one-hot vectors, the learning process of evidence for those mislabeled classes is over-penalized and remains hindered. Data uncertainty is ubiquitous in real-world applications, such as the indistinguishable "4" and "9", "1" and "7" in the MNIST dataset, images containing multiple objects in ImageNet, and the unavoidable noise labels in almost all datasets. Figure 1 illustrates that EDL cannot correctly distinguish MNIST and CIFAR10 image samples with different data uncertainties. Although EDL can model different types of uncertainty, its training process will underestimate the irreducible data uncertainty, thereby reducing the availability of uncertainty estimation.

In this paper, we propose a simple and novel method, *Fisher Information-based Evidential Deep Learning* ($\mathcal{I}$-EDL), to weigh the importance of different classes for each training sample. In particular, we introduce Fisher Information Matrix (FIM) to measure the information amount of evidence carried by the categorical probabilities $\boldsymbol{p}$ for each sample $\boldsymbol{x}$. According to the derivation of Dirichlet distribution's FIM, we found that the higher the evidence for a certain

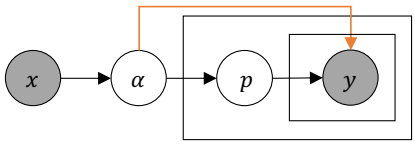


Figure 2. Graphical model representation of $\mathcal{I}$-EDL. The generative process of observed labels $\boldsymbol{y}$ is not only related to the predicted categorical probability $\mathbf{p} \sim Dir(\boldsymbol{\alpha})$, but also to the parameters $\boldsymbol{\alpha}$ of the Dirichlet distribution.

class label, the less corresponding information. Thus, we can set up a Gaussian distribution to help generate observed labels $\boldsymbol{y}$ by setting a larger variance for these less informative classes. As shown in Figure 2, the generative process of observed labels $\boldsymbol{y}$ is not only related to the predicted categorical probability $\mathbf{p}$, but also to the parameters $\boldsymbol{\alpha}$ of the Dirichlet distribution. Take the example of an informative image containing both a dog and a cat. The evidence for the dog and cat classes should be high if the neural network learns correctly. Our proposed generative model allows the observed label to be either dog or cat while retaining evidence for the other label, when the observed labels are set as one-hot vectors. From the perspective of optimization, compared with classical EDL, our proposed model encourages the network to focus more on the accuracy of classifying uncertain classes during the training process. From a generative model perspective, compared to classical EDL which only considers a single observation $\boldsymbol{y}$, the FIM we introduce can be seen as a type of ground truth distribution. These improvements help enhance the classification performance and uncertainty estimation of the evidential neural networks. Finally, we further improve the generalization ability of our network by optimizing the PAC-Bayesian bound.

To our knowledge, we are the first to explicitly leverage evidence to improve uncertainty reliability by dynamically weighing objective loss terms. Our main contributions can be summarized as follows: (1) We propose a novel method to quantify uncertainty by combining Fisher information and evidential neural networks. (2) We introduce PAC-Bayesian bound to further improve the generalization ability. (3) Our proposed method achieves excellent empirical performance in confidence evaluation and OOD detection, especially in the more challenging few-shot classification setting.

2. Preliminary

Evidential deep learning (EDL) (Sensoy et al., 2018) models the neural network by using the theory of subjective logic (SL) (Jøsang, 1997; 2016), a type of probabilistic logic that explicitly takes into account epistemic uncertainty and source trust. The opinion representation used to represent beliefs in SL provides greater expressive power than Boolean truths and probabilities, as it can express *"I don't*

know" as an opinion on the truth of possible states. The concept of beliefs is derived from the Dempster–Shafer Theory of Evidence (DST), which is a generalization of the Bayesian theory to subjective probabilities (Dempster, 1968; Shafer, 1976). The main idea behind DST is to abandon the additivity principle of probability theory, which means that the sum of belief masses can be less than 1, and the remainder is supplemented by uncertainty mass, i.e., lack of evidence about the truth of state values.

More specifically, considering a state space consisting of K mutually exclusive singletons (e.g., class labels), SL provides a belief mass b_k for each singleton $k \in [K]$ and an overall uncertainty mass u. The $K+1$ mass values satisfy $u + \sum_{k=1}^{K} b_k = 1$, where $u \geq 0, b_k \geq 0, \forall k \in [K]$. Belief mass depends on the evidence for the corresponding singleton, which measures the amount of support collected from data. In the absence of evidence, the belief for each singleton is 0, and uncertainty is 1. Conversely, an infinite amount of evidence leaves no room for uncertainty, yielding belief masses that sum to 1. SL formalizes the belief assignment of DST as a Dirichlet distribution with concentration parameters $\alpha_k = e_k + 1$, where e_k denotes the derived evidence for the k-th singleton. That is, the belief b_k and the uncertainty u can easily be derived from the parameters of the corresponding Dirichlet distribution by using

$$b_k = \frac{\alpha_k - 1}{\alpha_0}, \quad u = \frac{K}{\alpha_0},$$

where $\alpha_0 = \sum_{k=1}^{K} \alpha_k$ is referred to as the precision of the Dirichlet distribution. Higher values of α_0 lead to sharper, more confident distributions.

The Dirichlet distribution is the conjugate prior of the categorical distribution. It is parameterized by the concentration parameters $\boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_K], \forall \alpha_k > 0$, defined as:

$$Dir(\boldsymbol{p}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} p_k^{\alpha_k - 1}, \alpha_0 = \sum_{k=1}^{K} \alpha_k$$

where $\boldsymbol{p} \in \Delta^{K-1}$, and $\Gamma(\cdot)$ is the *gamma* function. Given an opinion, the expected probability of the k-th singleton is equivalent to the mean of the corresponding Dirichlet distribution, calculated as

$$\hat{p}_k = \frac{\alpha_k}{\alpha_0} = \frac{e_k + 1}{\sum_{c=1}^{K} e_c + K}.$$

When the evidence for one of the K attributes is observed, the corresponding Dirichlet parameter is incremented to update the Dirichlet distribution with the new observation.

EDL encourages neural networks to formalize the multiple opinions for the classification of a given sample as a Dirichlet distribution (Sensoy et al., 2018). Compared to standard neural networks, EDL replaces the last softmax layer with an activation layer, e.g., ReLU or Softplus, to obtain a non-negative output, which is the evidence vector used to parameterize the Dirichlet distribution. Standard DNNs for classification with a softmax output function can be viewed as predicting the expected classification distribution of EDL with an exponential output function. This means it is not sensitive to arbitrary scaling of α_k. Compared with the standard neural network classifiers that directly output the classification probability distribution of each sample, evidential neural networks obtain the density of classification probability assignments by parameterizing the Dirichlet distribution. Therefore, EDL models second-order probability and uncertainty (Jøsang, 2016), which can use the properties of Dirichlet distribution to distinguish different types of uncertainties, but classical EDL hinders the learning of evidence, especially for samples with high data uncertainty.

3. Method

3.1. Generative Model of Evidential Network

Evidential neural networks are typically trained using a combination of the expected mean squared error (MSE) and a Kullback-Leibler (KL) divergence term as a loss function, where the KL term penalizes evidence for classes that do not fit the training data. Note that, MSE performs best compared with the cross-entropy loss and the negative log marginal likelihood as empirically demonstrated in Sensoy et al. (2018). Let $f_{\boldsymbol{\theta}} : \mathbb{R}^d \to \mathbb{R}_+^K$ denotes the evidential neural network. Given a sample $(\boldsymbol{x}, \boldsymbol{y})$, where $\boldsymbol{y}$ is the one-hot encoded ground-truth class of observation $\boldsymbol{x}$, the loss function of EDL is expressed as:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) = &\mathbb{E}_{\boldsymbol{p} \sim Dir(\boldsymbol{\alpha})} \left[(\boldsymbol{y} - \boldsymbol{p})^T (\boldsymbol{y} - \boldsymbol{p}) \right] \\ &+ \lambda D_{\mathrm{KL}}(Dir(\boldsymbol{p}|\hat{\boldsymbol{\alpha}}) \| Dir(\boldsymbol{p}|\mathbf{1})),\end{aligned}$$

where $\boldsymbol{\alpha} = f_{\boldsymbol{\theta}}(\boldsymbol{x}) + 1, \hat{\boldsymbol{\alpha}} = \boldsymbol{\alpha} \odot (1 - \boldsymbol{y}) + \boldsymbol{y}, \lambda \geq 0$, $\mathbf{1} = [1; \cdots; 1] \in \mathbb{R}^K$, and $Dir(\boldsymbol{p}|\mathbf{1})$ denotes the uniform Dirichlet distribution.

Actually, when trained with MSE loss, the evidential network proposed by EDL can be understood as a new probabilistic graphical model. Specifically, let $\mathbf{x}, \mathbf{y}$ denote random variables whose unknown probability distribution generates inputs $\boldsymbol{x} \in \mathbb{R}^d$ and labels $\boldsymbol{y} \in \mathbb{R}^K$, respectively. EDL (Sensoy et al., 2018) supposes the observed labels $\boldsymbol{y}$ were drawn *i.i.d.* from an isotropic Gaussian distribution, i.e.

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{p}, \sigma^2 \boldsymbol{I}),$$

where $\boldsymbol{p} \sim Dir(f_{\boldsymbol{\theta}}(\boldsymbol{x}) + 1)$. Then, training evidential neural networks by minimizing the expected MSE can be viewed as learning model parameters that maximize the expected likelihood of the observed labels. Since the observed labels $\boldsymbol{y}$ are one-hot encoded, and the Gaussian distribution of

the generated labels is isotropic, i.e. each class is treated independently with the same variance, for high data uncertainty samples, the learning process of evidence for those mislabeled classes is over-penalized and remains hindered. This results in the total amount of evidence being underestimated and make the network overfitting. Furthermore, recently proposed work (Bengs et al., 2022) also argues that classical EDL does not incentivize learners to faithfully predict their epistemic uncertainty due to its sensitivity to the regularization parameter λ.

3.2. Fisher Information-based Evidential Network

In this work, based on the fact that the information of each class carried in categorical probabilities $\boldsymbol{p}$ is different, we argue that the generation of each class for a specific sample should not be isotropic. Intuitively, a certain class label with higher evidence is allowed to have a larger variance, so that the evidence for missing labels can be preserved while maximizing the likelihood of the observed labels. The Fisher information matrix (FIM) is chosen to measure the amount of information that the categorical probabilities $\boldsymbol{p}$ carry about the concentration parameters $\boldsymbol{\alpha}$ of a Dirichlet distribution that models $\boldsymbol{p}$. Formally, the FIM is defined as:

$$\mathcal{I}(\boldsymbol{\alpha}) = \mathbb{E}_{Dir(\boldsymbol{p}|\boldsymbol{\alpha})}\left[\frac{\partial \ell}{\partial \boldsymbol{\alpha}}\frac{\partial \ell}{\partial \boldsymbol{\alpha}^T}\right], \tag{1}$$

where $\ell = \log Dir(\boldsymbol{p}|\boldsymbol{\alpha})$ is the log-likelihood function. Under weak conditions (see Lemma 5.3 in Lehmann & Casella (2006)), the FIM can be expressed as $\mathcal{I}(\boldsymbol{\alpha}) = \mathbb{E}_{Dir(\boldsymbol{p}|\boldsymbol{\alpha})}\left[-\partial^2\ell/\partial\boldsymbol{\alpha}\boldsymbol{\alpha}^T\right]$. After applying a series of derivation steps (see Appendix A.1 for details), $\mathcal{I}(\boldsymbol{\alpha})$ can be simplified to:

$$\mathcal{I}(\boldsymbol{\alpha}) = \text{diag}([\psi^{(1)}(\alpha_1), \cdots, \psi^{(1)}(\alpha_K)]) - \psi^{(1)}(\alpha_0)\mathbf{1}\mathbf{1}^T,$$

where $\psi^{(1)}(\cdot)$ denotes the *trigamma* function, defined as $\psi^{(1)}(x) = d\psi(x)/dx = d^2 \ln\Gamma(x)/dx^2$. Since $\psi^{(1)}(x)$ is a monotonically decreasing function when $x > 0$, the class label with higher evidence corresponds to less Fisher information. Hence, we use the inverse of the FIM ($\mathcal{I}(\boldsymbol{\alpha})^{-1}$) as the variance of the generative distribution of $\boldsymbol{y}$.

Thus, we assume that the target variable $\boldsymbol{y}$ follows a multivariate Gaussian distribution with the following closed form:

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{p}, \sigma^2\mathcal{I}(\boldsymbol{\alpha})^{-1}),$$

where $\boldsymbol{\alpha} = f_{\boldsymbol{\theta}}(\boldsymbol{x}) + 1, \boldsymbol{p} \sim Dir(\boldsymbol{\alpha})$, σ^2 is the scalar used to adjust covariance value, $\mathcal{I}(\boldsymbol{\alpha})$ is the FIM of $Dir(\boldsymbol{\alpha})$, defined as Eq.(1). In MLE, we aim to learn model parameters $\boldsymbol{\theta}$ that maximize the marginal likelihood obtained by integrating the class probabilities, i.e.

$$\max_{\boldsymbol{\theta}} \ \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{P}}\left[\log \mathbb{E}_{\boldsymbol{p}\sim Dir(\boldsymbol{\alpha})}[\mathcal{N}(\boldsymbol{y}|\boldsymbol{p}, \sigma^2\mathcal{I}(\boldsymbol{\alpha})^{-1})]\right]. \tag{2}$$

Due to the concavity of the log function, by Jensen's inequality, our objective of Eq.(2) can be achieved by minimizing the expected negative log-likelihood loss function:

$$\begin{aligned} \min_{\boldsymbol{\theta}} \quad & \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{P}}\mathbb{E}_{\boldsymbol{p}\sim Dir(\boldsymbol{\alpha})}\left[-\log p(\boldsymbol{y}|\boldsymbol{p}, \boldsymbol{\alpha}, \sigma^2)\right]. \\ \text{s.t.} \quad & \boldsymbol{\alpha} = f_{\boldsymbol{\theta}}(\boldsymbol{x}) + 1 \\ & \mathcal{I}(\boldsymbol{\alpha}) = \mathbb{E}_{Dir(\boldsymbol{p}|\boldsymbol{\alpha})}\left[-\frac{\partial^2 \log Dir(\boldsymbol{p}|\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}\boldsymbol{\alpha}^T}\right] \\ & p(\boldsymbol{y}|\boldsymbol{p}, \boldsymbol{\alpha}, \sigma^2) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{p}, \sigma^2\mathcal{I}(\boldsymbol{\alpha})^{-1}) \end{aligned} \tag{3}$$

3.3. Learning with PAC-Bayesian Bound

Since the PAC-Bayesian theory (McAllester, 1999) provides data-driven generalization bounds computed on the training set and are simultaneously valid for all posteriors on network parameters, it is often used as a criterion for model selection or as an inspiration for learning algorithm conception. In the PAC-Bayes setting, it assumes that the predictor $f_{\boldsymbol{\theta}}$ has prior knowledge of the hypothesis space $\boldsymbol{\Theta}$ in the form of a prior distribution π. After the training dataset $\mathcal{D}$ is fed to the predictor, the prior is updated to a posterior distribution ρ. The full bound theorem is restated below, derived from the theorems in Germain et al. (2009), Alquier et al. (2016), Masegosa (2020), and we give the proof in appendix A.3 for completeness.

Theorem 3.1 ((Germain et al., 2009; Alquier et al., 2016; Masegosa, 2020))**.** *Given a data distribution $\mathcal{P}$ over $\mathcal{X} \times \mathcal{Y}$, a hypothesis set $\boldsymbol{\theta}$, a prior distribution π over $\boldsymbol{\Theta}$, for any $\delta \in (0, 1]$, and $\lambda > 0$, with probability at least $1 - \delta$ over samples $\mathcal{D} \sim \mathcal{P}^n$, we have for all posterior ρ,*

$$\begin{aligned} \mathbb{E}_{\rho(\boldsymbol{\theta})}[\mathcal{L}(\boldsymbol{\theta})] \leq \mathbb{E}_{\rho(\boldsymbol{\theta})}[\hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{\theta})] \\ + \frac{1}{\lambda}\left[D_{\text{KL}}(\rho\|\pi) + \log\frac{1}{\delta} + \Psi_{\mathcal{P},\pi}(\lambda, n)\right], \end{aligned}$$

where $\Psi_{\mathcal{P},\pi}(\lambda, n) = \log \mathbb{E}_{\pi(\boldsymbol{\theta})}\mathbb{E}_{\mathcal{D}\sim\mathcal{P}^n}\left[e^{\lambda(\mathcal{L}(\boldsymbol{\theta}) - \hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{\theta}))}\right]$.

In this paper, we treat $Dir(\boldsymbol{p}|\boldsymbol{\alpha})$ as the posterior distribution, and the prior as $Dir(\boldsymbol{p}|\boldsymbol{\mu})$, where $\boldsymbol{\mu}$ is set to $\beta \gg 1$ for the corresponding class and 1 for all other class. Given training set $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^N$, π, λ and δ, by the Theorem 3.1, the upper bound of Eq.(3) can be expressed as

$$\frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_i(\boldsymbol{\theta}) + \frac{1}{\lambda}D_{\text{KL}}(Dir(\boldsymbol{p}_i|\boldsymbol{\alpha}_i)\|Dir(\boldsymbol{p}_i|\boldsymbol{\mu}_i)), \tag{4}$$

where $\mathcal{L}_i(\boldsymbol{\theta}) = \mathbb{E}_{Dir(\boldsymbol{p}_i|\boldsymbol{\alpha}_i)}\left[-\log\mathcal{N}(\boldsymbol{y}_i|\boldsymbol{p}_i, \sigma^2\mathcal{I}(\boldsymbol{\alpha}_i)^{-1})\right]$, $\boldsymbol{\alpha}_i = f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) + 1$ and $\boldsymbol{p}_i \sim Dir(\boldsymbol{\alpha}_i)$.

The first term in Eq.(4) can be reformulated as a trade-off between the expected FIM-weighted MSE ($\mathcal{I}$-MSE) and a penalty of negative log determinant of the FIM ($|\mathcal{I}|$), since

$$\mathcal{L}_i(\boldsymbol{\theta}) \propto \underbrace{\mathbb{E}\left[(\boldsymbol{y}_i - \boldsymbol{p}_i)^T\mathcal{I}(\boldsymbol{\alpha}_i)(\boldsymbol{y}_i - \boldsymbol{p}_i)\right]}_{\mathcal{L}_i^{\mathcal{I}\text{-MSE}}} - \sigma^2\underbrace{\log|\mathcal{I}(\boldsymbol{\alpha}_i)|}_{\mathcal{L}_i^{|\mathcal{I}|}}.$$

Among them, since the class label with less evidence corresponds to the larger Fisher information, minimizing $\mathcal{L}_i^{\mathcal{I}\text{-MSE}}$ can be viewed as: for a certain class with low evidence, we expect its corresponding prediction probability to be more accurate, regardless of whether this class is the ground-truth label or not. The penalty for adding $-\mathcal{L}_i^{|\mathcal{I}|}$ can be seen as avoiding overconfidence caused by excessive evidence. It is worth noting that these loss terms are not only related to classes but also related to samples. Thus, the FIM-weighted term can be considered as an adaptive weight that can self-adjust the corresponding MSE loss based on the information of each class contained in the sample.

For the second term in Eq.(4), $\boldsymbol{\mu}_i$ varies with sample labels. To simplify this Kullback-Leibler (KL) divergence, we set $\hat{\boldsymbol{\alpha}}_i = \boldsymbol{\alpha}_i \odot (1 - \boldsymbol{y}_i) + \boldsymbol{y}_i$ to remove the predicted concentration parameter of the true label corresponding to the sample $\boldsymbol{x}_i$ following Sensoy et al. (2018), thereby the KL term is converted to

$$\mathcal{L}_i^{\text{KL}} = D_{\text{KL}}(Dir(\boldsymbol{p}_i|\hat{\boldsymbol{\alpha}}_i)\|Dir(\boldsymbol{p}_i|\mathbf{1})).$$

Finally, the objective function Eq.(4) can be reformulated as

$$\min_{\boldsymbol{\theta}} \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_i^{\mathcal{I}\text{-MSE}} - \lambda_1 \mathcal{L}_i^{|\mathcal{I}|} + \lambda_2 \mathcal{L}_i^{\text{KL}}, \tag{5}$$

where

$$\mathcal{L}_i^{\mathcal{I}\text{-MSE}} = \sum_{j=1}^{K}\left((y_{ij} - \frac{\alpha_{ij}}{\alpha_{i0}})^2 + \frac{\alpha_{ij}(\alpha_{i0}-\alpha_{ij})}{\alpha_{i0}^2(\alpha_{i0}+1)}\right)\psi^{(1)}(\alpha_{ij}),$$

$$\mathcal{L}_i^{|\mathcal{I}|} = \sum_{j=1}^{K}\log\psi^{(1)}(\alpha_{ij}) + \log\left(1 - \sum_{j=1}^{K}\frac{\psi^{(1)}(\alpha_{i0})}{\psi^{(1)}(\alpha_{ij})}\right),$$

$$\begin{aligned}\mathcal{L}_i^{\text{KL}} = &\log\Gamma(\sum_{j=1}^{K}\hat{\alpha}_{ij}) - \log\Gamma(K) - \sum_{j=1}^{K}\log\Gamma(\hat{\alpha}_{ij}) \\ &+ \sum_{j=1}^{K}(\hat{\alpha}_{ij}-1)\left[\psi(\hat{\alpha}_{ij}) - \psi(\sum_{k=1}^{K}\hat{\alpha}_{ik})\right],\end{aligned}$$

and $\lambda_1, \lambda_2 \geq 0$. The detailed derivation steps are left in Appendix A.2. The complete pseudo-code of our method is outlined in Algorithm 1. Actually, classical EDL can be viewed as a degenerate version of $\mathcal{I}$-EDL without $\mathcal{L}_i^{|\mathcal{I}|}$ and with $\mathcal{L}_i^{\text{MSE}}$ instead of $\mathcal{L}_i^{\mathcal{I}\text{-MSE}}$. Furthermore, we introduce the KL term from the PAC-Bayesian bound to make it more reasonable and interpretable. A detailed comparison of $\mathcal{I}$-EDL and classical EDL is given in Table 1.

Algorithm 1 $\mathcal{I}$-Evidential Deep Learning

Input: λ, Training set $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{N}$, batch size b, learning rate β, total epochs T
Initialize $\boldsymbol{\theta}$
for $t = 0, 1, \cdots, T$ **do**
  $\lambda_t = \min(1.0, t/T)$
  for $\mathcal{D}_b \sim \mathcal{D}$ **do**
    for $(\boldsymbol{x}_i, \boldsymbol{y}_i) \sim \mathcal{D}_b$ **do**
      $\boldsymbol{\alpha}_i = f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) + 1$
      $\hat{\boldsymbol{\alpha}}_i = \boldsymbol{\alpha}_i \odot (1 - \boldsymbol{y}_i) + \boldsymbol{y}_i$
      $\mathcal{L}_i = \mathcal{L}_i^{\mathcal{I}\text{-MSE}} - \lambda\mathcal{L}_i^{|\mathcal{I}|} + \lambda_t\mathcal{L}_i^{\text{KL}}$ //see Eq.(5)
    end for
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \beta\nabla_{\boldsymbol{\theta}}\mathcal{L}$ with $\mathcal{L} = \frac{1}{b}\sum_{i=1}^{b}\mathcal{L}_i$
  end for
end for

Table 1. Given a sample $(\boldsymbol{x}_i, \boldsymbol{y}_i)$, the difference in loss function between $\mathcal{I}$-EDL and EDL are marked in blue.

	EDL	**$\mathcal{I}$-EDL**
MSE	$\sum_{j=1}^{K}(y_{ij} - \frac{\alpha_{ij}}{\alpha_{i0}})^2 + \sum_{j=1}^{K}\frac{\alpha_{ij}(\alpha_{i0}-\alpha_{ij})}{\alpha_{i0}^2(\alpha_{i0}+1)}$	$\sum_{j=1}^{K}(y_{ij} - \frac{\alpha_{ij}}{\alpha_{i0}})^2\psi^{(1)}(\alpha_{ij}) + \sum_{j=1}^{K}\frac{\alpha_{ij}(\alpha_{i0}-\alpha_{ij})}{\alpha_{i0}^2(\alpha_{i0}+1)}\psi^{(1)}(\alpha_{ij})$
KL	$D_{\text{KL}}(Dir(\hat{\boldsymbol{\alpha}}_i)\|Dir(\mathbf{1}))$	$D_{\text{KL}}(Dir(\hat{\boldsymbol{\alpha}}_i)\|Dir(\mathbf{1}))$
$\mathcal{I}$	-	$-\log\|\mathcal{I}(\boldsymbol{\alpha}_i)\|$

4. Related Work

Dirichlet-based uncertainty models (DBU) predict the parameters of the Dirichlet distribution, which allows the computation of closed-form classical uncertainty metrics such as differential entropy, mutual information, etc. These metrics can be used to distinguish among data, model, and distributional uncertainty. Different DBU models differ in the parameterization and training strategy of the Dirichlet distribution. For example, **KL-PN** (Malinin & Gales, 2018) proposes the Prior Networks (PN) trained with two KL divergence terms. The first term is used to learn sharp Dirichlet parameters for ID data, while the other learns flat Dirichlet parameters for OOD data. Since the forward KL divergence is zero-avoiding, **RKL-PN** (Malinin & Gales, 2019) introduces the reverse KL divergence to avoid undesired multimodal target distributions. Posterior Network (**PostN**) (Charpentier et al., 2020) uses Normalizing Flows to predict the posterior distribution of any input sample without training with OOD data. Evidential Deep Learning (**EDL**) (Sensoy et al., 2018) treats the network's outputs as belief masses based on the Dempster-Shafer Theory of Evidence (DST) (Sentz & Ferson, 2002) and derives the loss function using subjective logic (Jøsang, 2016). Moreover, deep evidential regression (Amini et al., 2020; Soleimany et al., 2021) introduces evidential priors over the original Gaussian likelihood function to model the uncertainty of

regression networks. Meinert et al. (2022) further analyze why DER can produce reasonable results in practice despite overparameterized representations of uncertainty. Zhao et al. (2020) propose a multi-source uncertainty framework combined with DST for semi-supervised node classification with GNNs. Bao et al. (2022) propose a general framework for Open Set Temporal Action Localization (OSTAL) based on EDL. Although EDL shows an impressive performance in uncertainty quantification and is widely used in various applications, recently proposed work (Bengs et al., 2022) argues that classical EDL does not motivate learners to faithfully predict their epistemic uncertainty because it is sensitive to the regularization parameter. Compared with previous efforts, our method is the first to exploit evidence for training to improve the performance of uncertainty quantification. The Fisher information matrix (FIM) we introduce can be seen as some type of first-order distribution information, which can help learners make more accurate predictions and better estimate uncertainty.

Bayesian Neural Networks (BNNs) explicitly model network parameters as random variables, quantifying uncertainty by learning a posterior over parameters. Since the posterior inference of BNNs is intractable, many posterior approximation schemes have been developed to improve scalabilities, such as variational inference (VI) (Graves, 2011; Blundell et al., 2015), stochastic gradient Markov Chain Monte Carlo (Welling & Teh, 2011; Ma et al., 2015), and Laplace approximation (Ritter et al., 2018; Kristiadi et al., 2021). Furthermore, the integral of marginalizing the likelihood with the posterior distribution is also intractable and is typically approximated via sampling. A well-known method is Monte Carlo Dropout (**MC Dropout**) (Gal & Ghahramani, 2016), which treats the dropout layer as a Bernoulli distributed random variable, and training a network with dropout layers can be interpreted as an approximate VI. However, these methods require significant modifications to the training process and are computationally expensive, and more importantly, cannot distinguish between distributional uncertainty and other uncertainties.

Calibration methods aim to reduce over-confidence by calibrating models. For example, Guo et al. (2017) introduce temperature scaling as a post-hoc calibration to mitigate overconfidence. ODIN (Liang et al., 2018) uses a mix of temperature scaling at the softmax layer and input perturbations. Pereyra et al. (2017) penalize the low-entropy output distribution in the loss function. Karandikar et al. (2021) propose differentiable losses to improve calibration based on a soft version of the binning operation underlying popular calibration-error estimators. Roelofs et al. (2022) focus on assessing statistical bias in calibration. Since these methods cannot distinguish between different types of uncertainty, they are often combined with the first two types of methods, such as Lakshminarayanan et al. (2017) combined calibration results with deep ensembles. Additionally, there are also some methods of modeling with label noise. For example, Collier et al. (2021) propose input-dependent noise losses for label noise in classification. Cui et al. (2022) provides a unified framework for reliable learning under the joint (image, label)-noise. Compared to these works, we focus on the evidence underestimation problem in evidential networks, and more importantly, we are the first to address this problem by introducing the Fisher information matrix.

5. Experiments

In this section, we conduct extensive experiments to compare the performance of our proposed method with previous methods on multiple uncertainty estimation-related tasks. See Appendix C for additional results and more details. [1]

5.1. Experimental Setup

Datasets We evaluate our algorithm on the following image classification datasets: **MNIST** (LeCun, 1998), **CIFAR10** (Krizhevsky et al., 2009), and **mini-ImageNet** (Vinyals et al., 2016). For OOD detection experiments, we use **KMNIST** (Clanuwat et al., 2018) and **FashionMNIST** (Xiao et al., 2017) for MNIST, the Street View House Numbers (**SVHN**) (Netzer et al., 2018) and **CIFAR100** (Krizhevsky et al., 2009) for CIFAR10, and the Caltech-UCSD Birds (**CUB**) dataset (Wah et al., 2011) for mini-ImageNet. More details are given in Appendix C.1.

Implementation details Following Charpentier et al. (2020), we use 3 convolutional layers and 3 dense layers (ConvNet) on MNIST and VGG16 (Simonyan & Zisserman, 2014) on CIFAR10. For all experiments on both datasets, we split the data into train, validation, and test sets. We use a validation loss-based early termination strategy to train up to 200 epochs with a batch size of 64. For the mini-ImageNet dataset, we conduct experiments on the more challenging few-shot classification setting. We use WideResNet-28-10 pre-trained backbone from Yang et al. (2021) as the feature extractor to train a 1-layer classifier. Refer to Appendix C.2 for more details.

Baselines We focus on comparing our algorithm with other Dirichlet-based uncertainty methods, since only DBU methods can distinguish different types of uncertainty compared to BNNs and calibration methods, as mentioned previously. In particular, we compare to following baselines: Prior Networks (PN) trained with KL divergence (**KL-PN**) (Malinin & Gales, 2018) and Reverse KL divergence (**RKL-PN**) (Malinin & Gales, 2019), Posterior Network (**PostN**) (Charpentier et al., 2020), and Evidential Deep Learning

[1] The code is available at: https://github.com/danruod/IEDL

Table 2. AUPR scores of OOD detection (mean $\pm$ standard deviation of 5 runs). † indicates that the first four lines are the results reported by Charpentier et al. (2020). Bold and underlined numbers indicate the best and runner-up scores, respectively.

	MNIST $\rightarrow$ KMNIST†		MNIST $\rightarrow$ FMNIST†		CIFAR10 $\rightarrow$ SVHN†		CIFAR10 $\rightarrow$ CIFAR100	
Method	**Max.P**	α_0	**Max.P**	α_0	**Max.P**	α_0	**Max.P**	α_0
MC Dropout	94.00 ± 0.1	-	96.56 ± 0.2	-	51.39 ± 0.1	-	45.57 ± 1.0	-
KL-PN	92.97 ± 1.2	93.39 ± 1.0	$\underline{98.44 \pm 0.1}$	$\underline{98.16 \pm 0.0}$	43.96 ± 1.9	43.23 ± 2.3	61.41 ± 2.8	61.53 ± 3.4
RKL-PN	60.76 ± 2.9	53.76 ± 3.4	78.45 ± 3.1	72.18 ± 3.6	53.61 ± 1.1	49.37 ± 0.8	55.42 ± 2.6	54.74 ± 2.8
PostN	95.75 ± 0.2	94.59 ± 0.3	97.78 ± 0.2	97.24 ± 0.3	$\underline{80.21 \pm 0.2}$	77.71 ± 0.3	81.96 ± 0.8	82.06 ± 0.8
EDL	97.02 ± 0.8	$\underline{96.31 \pm 2.0}$	98.10 ± 0.4	98.08 ± 0.4	78.87 ± 3.5	$\underline{79.12 \pm 3.7}$	$\underline{84.30 \pm 0.7}$	$\underline{84.18 \pm 0.7}$
$\mathcal{I}$-EDL	$\mathbf{98.34 \pm 0.2}$	$\mathbf{98.33 \pm 0.2}$	$\mathbf{98.89 \pm 0.3}$	$\mathbf{98.86 \pm 0.3}$	$\mathbf{83.26 \pm 2.4}$	$\mathbf{82.96 \pm 2.2}$	$\mathbf{85.35 \pm 0.7}$	$\mathbf{84.84 \pm 0.6}$

Table 3. AUPR scores and accuracy of CIFAR10 with VGG16 in misclassified image detection and image classification, respectively. Each experiment is run with 5 seeds. † denotes results reported by (Charpentier et al., 2020). Bold and underlined numbers indicate the best and runner-up scores, respectively.

Method	**Max.P**	**Max.α**	**Acc.**
MC Dropout†	97.15 ± 0.0	-	82.84 ± 0.1
KL-PN†	50.61 ± 4.0	52.49 ± 4.2	27.46 ± 1.7
RKL-PN†	86.11 ± 0.4	85.59 ± 0.3	64.76 ± 0.3
PostN†	97.76 ± 0.0	97.25 ± 0.0	$\underline{84.85 \pm 0.0}$
EDL	$\underline{97.86 \pm 0.2}$	$\underline{97.86 \pm 0.2}$	83.55 ± 0.6
$\mathcal{I}$-EDL	$\mathbf{98.72 \pm 0.1}$	$\mathbf{98.63 \pm 0.1}$	$\mathbf{89.20 \pm 0.3}$

(**EDL**) (Sensoy et al., 2018). Furthermore, we compare the dropout model (**MC Dropout**) (Gal & Ghahramani, 2016), which is often state-of-the-art in many uncertainty estimation tasks (Ovadia et al., 2019). Since other methods except KL-PN and RKL-PN do not require OOD data for training. For a fair comparison, we use the uniform noise instead of actual OOD test data as OOD training data for the former two methods following Charpentier et al. (2020).

5.2. Confidence Evaluation

We first measure the availability of uncertainty estimates in the confidence evaluation tasks that aim to answer an interesting question "*Are more confident (i.e., less uncertain) predictions more likely to be correct?*". We use the area under the precision-recall curve (**AUPR**) metric. For DBU methods, we represent $\max_c p_c$ (Max.P) and $\max_c \alpha_c$ (Max.α) respectively as the scores with labels 1 for correct and 0 for incorrect predictions. Since the dropout model does not have concentration parameters, we only provide results with Max.P as scores. Table 3 shows our proposed method achieves state-of-the-art performance in all measurements. In particular, our method improves image classification by about **5.2%** and confidence estimation by about **0.9%** compared to the runner-up methods.

5.3. OOD detection

We then measure the usability of uncertainty quantification in the OOD detection task. The performance of OOD detection is also measured by **AUPR** with labels 1 for ID data and 0 for OOD data. The scores of the DBU methods are given by $\max_c p_c$ (Max.P) and α_0 ($\sum_c \alpha_c$) respectively, while Dropout uses Max.P as scores. We compare our method with other methods on four OOD detection tasks, including MNIST against KMNIST and FMNIST, and CIFAR10 against SVHN and CIFAR100. Table 2 shows our proposed method achieves superior performance in all tasks without training with additional OOD data. More specifically, $\mathcal{I}$-EDL outperforms the second-placed method by about **1.3%**, **0.5%**, **3.8%** and **1.2%** on four OOD detection tasks, respectively. Note that EDL does not achieve suboptimal performance on all OOD detection tasks. We also evaluate performance using differential entropy (**D.Ent.**) and mutual information (**M.I.**) as scores and area under a ROC curve (**AUROC**). All these results can be seen in Appendix C.3. Given lots of efforts contributed to OOD detection (Liang et al., 2018; Sastry & Oore, 2020), here we mainly focus on the comparisons with DBU models, which solve OOD detection by distinguishing different types of uncertainty.

5.4. Few-shot Learning

We next conduct more challenging few-shot experiments on mini-ImageNet. We use the WideResNet trained following Yang et al. (2021) to obtain pre-trained features, and then train the 1-layer classifier under N-way K-shot setting. We evaluate $\{5, 10\}$-way $\{1, 5, 20\}$-shot classification, confidence estimation and OOD detection. The performance of classification and uncertainty estimation are reported in the average accuracy(%, top-1) and AUPR(%), respectively, with 95% confidence interval over $10,000$ few-shot episodes. Each episode contains randomly sampled N classes and K samples per class for adaptation, $\min(15, K)$ query samples per class for image classification and confidence evaluation, and the same number of query samples

Table 4. Classification accuracy (Acc.), AUPR scores for both confidence evaluation (Conf.) and OOD detection (OOD) under $\{5, 10\}$-way $\{1, 5, 20\}$-shot settings of mini-ImageNet. CUB is used for OOD detection. Each experiment is run for over $10,000$ few-shot episodes.

	5-Way 1-Shot			5-Way 5-Shot			5-Way 20-Shot		
Method	**Acc.**	**Conf. (Max.α)**	**OOD (α_0)**	**Acc.**	**Conf. (Max.α)**	**OOD (α_0)**	**Acc.**	**Conf. (Max.α)**	**OOD (α_0)**
EDL	61.00 ± 0.22	80.59 ± 0.23	65.40 ± 0.26	80.38 ± 0.15	93.92 ± 0.09	76.53 ± 0.27	85.54 ± 0.12	97.51 ± 0.04	79.78 ± 0.23
$\mathcal{I}$-EDL	63.82 ± 0.20	82.00 ± 0.21	74.76 ± 0.25	82.00 ± 0.14	94.09 ± 0.09	82.48 ± 0.20	88.12 ± 0.09	97.54 ± 0.04	85.40 ± 0.19
Δ	**2.82**	**1.41**	**9.36**	**1.62**	**0.17**	**5.95**	**2.58**	**0.04**	**5.62**
	10-Way 1-Shot			**10-Way 5-Shot**			**10-Way 20-Shot**		
Method	**Acc.**	**Conf. (Max.α)**	**OOD (α_0)**	**Acc.**	**Conf. (Max.α)**	**OOD (α_0)**	**Acc.**	**Conf. (Max.α)**	**OOD (α_0)**
EDL	44.55 ± 0.15	65.97 ± 0.20	67.83 ± 0.24	62.52 ± 0.16	86.81 ± 0.10	76.34 ± 0.20	69.29 ± 0.17	94.21 ± 0.06	76.88 ± 0.17
$\mathcal{I}$-EDL	49.37 ± 0.13	68.29 ± 0.19	71.95 ± 0.20	67.89 ± 0.11	87.45 ± 0.09	82.29 ± 0.19	78.60 ± 0.08	94.40 ± 0.04	82.52 ± 0.14
Δ	**4.82**	**2.32**	**4.12**	**5.37**	**0.64**	**5.95**	**9.31**	**0.19**	**5.64**

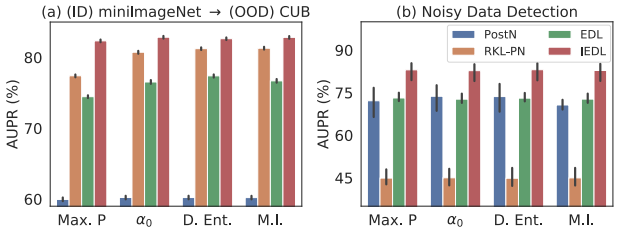


Figure 3. (a) Comparison of OOD detection (AUPR) on mini-ImageNet against CUB under 5-way 5-shot. (b) Noisy data detection on CIFAR10.

from CUB for OOD detection.

As shown in Table 4, the average test accuracy, Max.α-based confidence evaluation, and α_0-based OOD detection show impressive improvements on $\mathcal{I}$-EDL over EDL on all the N-way K-shot tasks. More specifically, all average test accuracy improvements of our method exceed **1.62%**, up to **9.31%** under 10-way 20-shot. In confidence evaluation, $\mathcal{I}$-EDL also shows better performance than EDL, especially the improvement over **2.32%** under 10-way 1-shot. Moreover, $\mathcal{I}$-EDL shows excellent performance in OOD detection, where all improvements are between **4.12%** and **9.36%**. Furthermore, we compare the OOD detection performance of our method and other DBU methods under 5-way 5-shot in Fig. 3(a). All of these results demonstrate that our method not only improves classification accuracy but also greatly improves the availability of uncertainty estimation in the more challenging few-shot scenarios. More results including AUPR and AUROC scores based on M.I. and D.Ent are provided in Appendix C.4.

5.5. Noisy data detection

We finally evaluate our method on noisy examples. Noisy examples are generated by adding zero-mean isotropic Gaussian noise with standard deviation $\sigma = 0.1$ to the test data of the ID dataset. As shown in Fig. 3(b), our method demonstrates excellent detection of noisy data across all metrics, outperforming the runner-up method by more than 11%.

Table 5. Ablation studies under mini-ImageNet 5-way 5-shot for image classification, confidence evaluation, and OOD detection against CUB.

$\mathcal{I}$-MSE	$\lvert\mathcal{I}\rvert$	**Acc.**	**Conf. (Max.α)**	**OOD (α_0)**
		80.38 ± 0.15	93.92 ± 0.09	76.53 ± 0.27
	✓	81.82 ± 0.14	93.97 ± 0.09	79.68 ± 0.25
✓		81.27 ± 0.14	$\mathbf{94.42} \pm 0.08$	81.75 ± 0.22
✓	✓	$\mathbf{82.00} \pm 0.14$	94.09 ± 0.09	$\mathbf{82.48} \pm 0.20$

5.6. Ablation Study

We further investigate our method performance with an ablation study and summarize it in Table 5. We respectively ablate the effects of expected FIM-weighted MSE ($\mathcal{I}$-MSE) and FIM's negative log determinant ($|\mathcal{I}|$). Note that classical EDL is equivalent to $\mathcal{I}$-EDL without using $\mathcal{I}$-MSE and $|\mathcal{I}|$. From the result, we observe that both optimizations are beneficial for image classification, confidence evaluation, and OOD detection. In particular, with the only usage of $|\mathcal{I}|$ ($\mathcal{I}$-MSE), the improvements over EDL for classification and α_0-based OOD detection are $\sim$ **1.9%** (**1.3%**) and $\sim$ **4.1%** (**6.8%**). If both optimizations are used together, the improvements increase to about **2.0%** and **7.8%**. Thus, the combination of two optimizations achieves a win-win effect. More ablation studies are provided in Appendix C.5.

5.7. Analysis of Uncertainty estimation

Figure 4 represents density plots of the predicted differential entropy and mutual information[2]. Lower entropy or mutual information represents the model yields a sharper distribution, indicating that the sample has low uncertainty. We also

[2]Detailed formulas refer to Eq.(18) and Eq.(16) in Appendix B

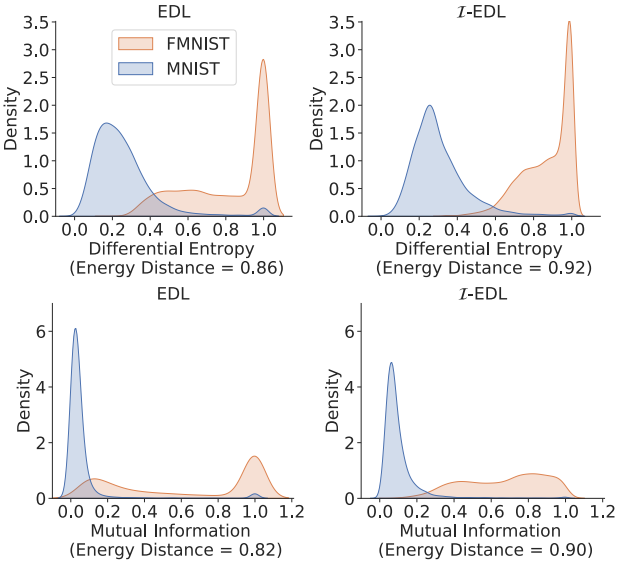


Figure 4. Uncertainty representation for ID (MNIST) and OOD (FMNIST). More results are shown in Appendix C.6.

report the energy distance (Székely & Rizzo, 2013) of two distributions (Formula is given in Appendix B.4.), which shows that our method provides more separable uncertainty estimates. More specifically, $\mathcal{I}$-EDL produces sharper prediction peaks than EDL, both in the low uncertainty region of ID samples and the high uncertainty region of OOD samples. Furthermore, our method also reduces the occurrence of ID samples in high-uncertainty regions.

6. Conclusion

In this paper, we found that the classical EDL trained with mean square error would hinder the learning of evidence, especially for high data uncertainty samples. To address this issue, we propose a novel and simple method, *Fisher Information-based Evidential Deep Learning* ($\mathcal{I}$-EDL) to alleviate the over-penalization of the mislabeled classes by considering importance weights with different classes for each sample. More specifically, we introduce the perspective of generative models to model evidential networks, where the observed label is jointly generated by the predicted categorical probability and the informativeness of each class contained in the sample. The categorical probabilities are generated from the Dirichlet distribution with its concentration parameter calculated by passing the input sample through the evidential network, while the information is obtained from FIM. Extensive experiments on various image classification, confidence evaluation and OOD detection tasks, as well as comparisons with some state-of-the-art algorithms, demonstrate the effectiveness of our approach in achieving high classification and uncertainty quantification. Since our method is designed based on Dirichlet distribution, it cannot be directly applied to regression tasks. A naive approach is to discretize the regression task, but this is not a good solution because it loses information about continuous labels, such as order. However, we believe that avoiding overfitting caused by data uncertainty in regression tasks and how to apply our ideas to regression tasks is a very attractive problem worthy of exploration in future work.

Acknowledgments

We would like to thank Prof. Chang-Yu Hsieh for his helpful discussions on Evidential Deep Learning, and anonymous reviewers for their feedback and constructive comments to improve this work. This work is supported by the National Key R&D Program of China (2022YFE0200700), the Research Grants Council of the Hong Kong Special Administrative Region, China (Project Reference Number: T45-401/22-N and 14201321), the National Natural Science Foundation of China (Project No. 62006219), the Natural Science Foundation of Guangdong Province (2022A1515011579), and the Hong Kong Innovation and Technology Fund (Project No. ITS/170/20 and ITS/241/21).

References

Ablain, M., Meyssignac, B., Zawadzki, L., Jugier, R., Ribes, A., Spada, G., Benveniste, J., Cazenave, A., and Picot, N. Uncertainty in satellite estimates of global mean sea-level changes, trend and acceleration. *Earth System Science Data*, 11(3):1189–1202, 2019.

Alquier, P., Ridgway, J., and Chopin, N. On the properties of variational approximations of gibbs posteriors. *The Journal of Machine Learning Research*, 17(1):8374–8414, 2016.

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.

Bao, W., Yu, Q., and Kong, Y. Opental: Towards open set temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2979–2989, 2022.

Bengs, V., Hüllermeier, E., and Waegeman, W. Pitfalls of epistemic uncertainty quantification through loss minimisation. In *Advances in Neural Information Processing Systems*, 2022.

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.

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

Clanuwat, T., Bober-Irizar, M., Kitamoto, A., Lamb, A., Yamamoto, K., and Ha, D. Deep learning for classical japanese literature. *arXiv preprint arXiv:1812.01718*, 2018.

Collier, M., Mustafa, B., Kokiopoulou, E., Jenatton, R., and Berent, J. Correlated input-dependent label noise in large-scale image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1551–1560, 2021.

Cui, P., Yue, Y., Deng, Z., and Zhu, J. Confidence-based reliable learning under dual noises. *Advances in Neural Information Processing Systems*, 35:35116–35129, 2022.

Dempster, A. P. A generalization of bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(2):205–232, 1968.

Feng, D., Rosenbaum, L., and Dietmayer, K. Towards safe autonomous driving: Capture uncertainty in the deep neural network for lidar 3d vehicle detection. In *2018 21st international conference on intelligent transportation systems (ITSC)*, pp. 3266–3273. IEEE, 2018.

Gal, Y. *Uncertainty in Deep Learning*. PhD thesis, University of Cambridge, 2016.

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.

Gal, Y., Islam, R., and Ghahramani, Z. Deep bayesian active learning with image data. In *International Conference on Machine Learning*, pp. 1183–1192. PMLR, 2017.

Germain, P., Lacasse, A., Laviolette, F., and Marchand, M. Pac-bayesian learning of linear classifiers. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 353–360, 2009.

Ghaffari, S., Saleh, E., Forsyth, D., and Wang, Y.-X. On the importance of firth bias reduction in few-shot classification. *arXiv preprint arXiv:2110.02529*, 2021.

Graves, A. Practical variational inference for neural networks. *Advances in neural information processing systems*, 24, 2011.

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

Hemmer, P., Kühl, N., and Schöffer, J. Deal: deep evidential active learning for image classification. In *Deep Learning Applications, Volume 3*, pp. 171–192. Springer, 2022.

Jøsang, A. Artificial reasoning with subjective logic. In *Proceedings of the second Australian workshop on commonsense reasoning*, volume 48, pp. 34. Citeseer, 1997.

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

Karandikar, A., Cain, N., Tran, D., Lakshminarayanan, B., Shlens, J., Mozer, M. C., and Roelofs, B. Soft calibration objectives for neural networks. *Advances in Neural Information Processing Systems*, 34:29768–29779, 2021.

Kristiadi, A., Hein, M., and Hennig, P. Learnable uncertainty under laplace approximations. In *Uncertainty in Artificial Intelligence*, pp. 344–353. PMLR, 2021.

Kristiadi, A., Hein, M., and Hennig, P. Being a bit frequentist improves bayesian neural networks. In *International Conference on Artificial Intelligence and Statistics*, pp. 529–545. PMLR, 2022.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. *cs.toronto.edu*, 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.

Lehmann, E. L. and Casella, G. *Theory of point estimation*. Springer Science & Business Media, 2006.

Liang, S., Li, Y., and Srikant, R. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018.

Ma, Y.-A., Chen, T., and Fox, E. A complete recipe for stochastic gradient mcmc. *Advances in neural information processing systems*, 28, 2015.

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.

Masegosa, A. Learning under model misspecification: Applications to variational and ensemble methods. *Advances in Neural Information Processing Systems*, 33: 5479–5491, 2020.

McAllester, D. A. Some pac-bayesian theorems. *Machine Learning*, 37(3):355–363, 1999.

Meinert, N., Gawlikowski, J., and Lavin, A. The unreasonable effectiveness of deep evidential regression. *arXiv preprint arXiv:2205.10060*, 2022.

Nair, T., Precup, D., Arnold, D. L., and Arbel, T. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. *Medical image analysis*, 59:101557, 2020.

Nandy, J., Hsu, W., and Lee, M. L. Towards maximizing the representation gap between in-domain & out-of-distribution examples. *Advances in Neural Information Processing Systems*, 33:9239–9250, 2020.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. The street view house numbers (svhn) dataset. Technical report, Technical report, Accessed 2016-08-01.[Online], 2018.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., and Snoek, J. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

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*, pp. 14391–14400, 2022.

Pereyra, G., Tucker, G., Chorowski, J., Kaiser, Ł., and Hinton, G. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*, 2017.

Quinonero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. *Dataset shift in machine learning*. Mit Press, 2008.

Ren, M., Triantafillou, E., Ravi, S., Snell, J., Swersky, K., Tenenbaum, J. B., Larochelle, H., and Zemel, R. S. Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676*, 2018.

Ritter, H., Botev, A., and Barber, D. A scalable laplace approximation for neural networks. In *6th International Conference on Learning Representations, ICLR 2018-Conference Track Proceedings*, volume 6. International Conference on Representation Learning, 2018.

Roelofs, R., Cain, N., Shlens, J., and Mozer, M. C. Mitigating bias in calibration error estimation. In *International Conference on Artificial Intelligence and Statistics*, pp. 4036–4054. PMLR, 2022.

Sastry, C. S. and Oore, S. Detecting out-of-distribution examples with gram matrices. In *International Conference on Machine Learning*, pp. 8491–8501. PMLR, 2020.

Seeböck, P., Orlando, J. I., Schlegl, T., Waldstein, S. M., Bogunović, H., Klimscha, S., Langs, G., and Schmidt-Erfurth, U. Exploiting epistemic uncertainty of anatomy segmentation for anomaly detection in retinal oct. *IEEE transactions on medical imaging*, 39(1):87–98, 2019.

Sensoy, M., Kaplan, L., and Kandemir, M. Evidential deep learning to quantify classification uncertainty. *Advances in Neural Information Processing Systems*, 31, 2018.

Sentz, K. and Ferson, S. Combination of evidence in dempster-shafer theory. *US Department of Energy*, 2002.

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

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Soleimany, A. P., Amini, A., Goldman, S., Rus, D., Bhatia, S. N., and Coley, C. W. Evidential deep learning for guided molecular property prediction and discovery. *ACS central science*, 7(8):1356–1367, 2021.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Székely, G. J. and Rizzo, M. L. Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference*, 143(8):1249–1272, 2013.

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

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. The caltech-ucsd birds-200-2011 dataset. *caltech.edu*, 2011.

Welling, M. and Teh, Y. W. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 681–688. Citeseer, 2011.

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

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Yang, S., Liu, L., and Xu, M. Free lunch for few-shot learning: Distribution calibration. *arXiv preprint arXiv:2101.06395*, 2021.

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

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.

A. Derivation and Proof

This section provides the derivation of the Fisher Information Matrix (FIM) of Dirichlet distribution and the final objective function of Eq. 5. We also provide a brief overview of the proof of Theorem 3.1 from (Germain et al., 2009; Alquier et al., 2016; Masegosa, 2020).

A.1. FIM Derivation for Dirichlet Distribution

The Fisher Information Matrix (FIM) of Dirichlet distribution is defined as:

$$\mathcal{I}(\boldsymbol{\alpha}) = \mathbb{E}\left[\frac{\partial \ell}{\partial \boldsymbol{\alpha}} \frac{\partial \ell}{\partial \boldsymbol{\alpha}^T}\right] \in \mathbb{R}^{K\times K},$$

where $\ell = \log Dir(\boldsymbol{p}|\boldsymbol{\alpha})$ is the log-likelihood function. Under weak conditions (see Lemma 5.3 in (Lehmann & Casella, 2006)), the FIM can be expressed as $\mathcal{I}(\boldsymbol{\alpha}) = \mathbb{E}_{Dir(\boldsymbol{p}|\boldsymbol{\alpha})}\left[-\partial^2 \ell / \partial \boldsymbol{\alpha}\boldsymbol{\alpha}^T\right]$. Thus, we can calculate each element by

$$\begin{aligned}
[\mathcal{I}(\boldsymbol{\alpha})]_{ij} &= \mathbb{E}_{Dir(\boldsymbol{p}|\boldsymbol{\alpha})}\left[-\frac{\partial^2}{\partial\alpha_i\partial\alpha_j}\log Dir(\boldsymbol{p}|\boldsymbol{\alpha})|\boldsymbol{\alpha}\right] \\
&= \mathbb{E}_{Dir(\boldsymbol{p}|\boldsymbol{\alpha})}\left[-\frac{\partial^2}{\partial\alpha_i\partial\alpha_j}\left(\log\Gamma\left(\alpha_0\right) - \sum_{k=1}^{K}\log\Gamma\left(\alpha_k\right) + \sum_{k=1}^{K}(\alpha_k - 1)\log p_k\right)\right] \\
&= \mathbb{E}_{Dir(\boldsymbol{p}|\boldsymbol{\alpha})}\left[-\frac{\partial}{\partial\alpha_j}\left(\psi\left(\alpha_0\right) - \psi\left(\alpha_i\right) + \log p_i\right)\right] \\
&= \begin{cases} \psi^{(1)}\left(\alpha_i\right) - \psi^{(1)}\left(\alpha_0\right), & i = j, \\ -\psi^{(1)}\left(\alpha_0\right), & i \neq j, \end{cases}
\end{aligned}$$

where $\Gamma(\cdot)$ is the *gamma* function, $\psi(\cdot)$ is the *digamma* function, $\psi^{(1)}(\cdot)$ is the *trigamma* function, defined as $\psi^{(1)}(x) = d\psi(x)/dx = d^2 \ln\Gamma(x)/dx^2$. Then, we can get the matrix form of the FIM:

$$\begin{aligned}
\mathcal{I}(\boldsymbol{\alpha}) &= \begin{bmatrix} \psi^{(1)}\left(\alpha_1\right) - \psi^{(1)}\left(\alpha_0\right) & -\psi^{(1)}\left(\alpha_0\right) & \cdots & -\psi^{(1)}\left(\alpha_0\right) \\ -\psi^{(1)}\left(\alpha_0\right) & \psi^{(1)}\left(\alpha_2\right) - \psi^{(1)}\left(\alpha_0\right) & \cdots & -\psi^{(1)}\left(\alpha_0\right) \\ \vdots & \vdots & \ddots & \vdots \\ -\psi^{(1)}\left(\alpha_0\right) & -\psi^{(1)}\left(\alpha_0\right) & \cdots & \psi^{(1)}\left(\alpha_K\right) - \psi^{(1)}\left(\alpha_0\right) \end{bmatrix} \\
&= \text{diag}([\psi^{(1)}(\alpha_1), \cdots, \psi^{(1)}(\alpha_K)]) - \psi^{(1)}(\alpha_0)\mathbf{1}\mathbf{1}^T,
\end{aligned} \tag{6}$$

where $\mathbf{1} = [1; \cdots; 1] \in \mathbb{R}^K$. Let $\boldsymbol{b} = [\psi^{(1)}(\alpha_1), \cdots, \psi^{(1)}(\alpha_K)]^T$, by applying Matrix-Determinant Lemma, we have

$$|\mathcal{I}(\boldsymbol{\alpha})| = |\text{diag}(\boldsymbol{b})| \cdot (1 - \psi^{(1)}(\alpha_0)\mathbf{1}^T\text{diag}(\boldsymbol{b})^{-1}\mathbf{1}) = \prod_{i=1}^{K}\psi^{(1)}(\alpha_i)(1 - \sum_{j=1}^{K}\frac{\psi^{(1)}(\alpha_0)}{\psi^{(1)}(\alpha_j)}).$$

Therefore,

$$\log|\mathcal{I}(\boldsymbol{\alpha})| = \sum_{i=1}^{K}\log\psi^{(1)}(\alpha_i) + \log(1 - \sum_{i=1}^{K}\frac{\psi^{(1)}(\alpha_0)}{\psi^{(1)}(\alpha_i)}). \tag{7}$$

A.2. Derivation of the objective function Eq. 5

Given training set $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^N$, by applying Theorem 3.1, the upper bound of Eq.(3) can be expressed as

$$\begin{aligned}
\min_{\boldsymbol{\theta}} \quad & \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}}\left[\mathbb{E}_{\boldsymbol{p}\sim Dir(\boldsymbol{\alpha})}\left[-\log p(\boldsymbol{y}|\boldsymbol{p}, \boldsymbol{\alpha}, \sigma^2)\right] + \lambda D_{\text{KL}}(Dir(\boldsymbol{p}|\hat{\boldsymbol{\alpha}})\|Dir(\boldsymbol{p}|\mathbf{1}))\right] \\
\text{s.t.} \quad & \boldsymbol{\alpha} = f_{\boldsymbol{\theta}}(\boldsymbol{x}) + 1 \\
& \hat{\boldsymbol{\alpha}} = \boldsymbol{\alpha} \odot (1 - \boldsymbol{y}) + \boldsymbol{y} \\
& \mathcal{I}(\boldsymbol{\alpha}) = \text{diag}([\psi^{(1)}(\alpha_1), \cdots, \psi^{(1)}(\alpha_K)]) - \psi^{(1)}(\alpha_0)\mathbf{1}\mathbf{1}^T \\
& \boldsymbol{p} \sim Dir(\boldsymbol{\alpha}) \\
& p(\boldsymbol{y}|\boldsymbol{p}, \boldsymbol{\alpha}, \sigma^2) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{p}, \sigma^2\mathcal{I}(\boldsymbol{\alpha})^{-1}).
\end{aligned} \tag{8}$$

We first simplify the first term $\mathbb{E}_{Dir(\boldsymbol{p}|\boldsymbol{\alpha})}[-\log \mathcal{N}(\boldsymbol{y}|\boldsymbol{p}, \sigma^2 \mathcal{I}(\boldsymbol{\alpha})^{-1})]$ (abbreviated as $\mathbb{E}[-\log \mathcal{N}(\boldsymbol{y}|\boldsymbol{p}, \sigma^2 \mathcal{I}(\boldsymbol{\alpha})^{-1})]$),

$$
\begin{aligned}
&\mathbb{E}[-\log \mathcal{N}(\boldsymbol{y}|\boldsymbol{p}, \sigma^2 \mathcal{I}(\boldsymbol{\alpha})^{-1})] \\
=&\mathbb{E}\left[-\log\left((2\pi\sigma^2)^{-\frac{k}{2}}|\mathcal{I}(\boldsymbol{\alpha})|^{\frac{1}{2}}\exp(-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{p})^T\mathcal{I}(\boldsymbol{\alpha})(\boldsymbol{y}-\boldsymbol{p}))\right)\right] \\
=&\mathbb{E}\left[\frac{k}{2}\log(2\pi\sigma^2) - \frac{1}{2}\log|\mathcal{I}(\boldsymbol{\alpha})| + \frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{p})^T\mathcal{I}(\boldsymbol{\alpha})(\boldsymbol{y}-\boldsymbol{p})\right] \\
=&\frac{k}{2}\log(2\pi\sigma^2) - \frac{1}{2}\log|\mathcal{I}(\boldsymbol{\alpha})| + \frac{1}{2\sigma^2}\mathbb{E}\left[(\boldsymbol{y}-\boldsymbol{p})^T\mathcal{I}(\boldsymbol{\alpha})(\boldsymbol{y}-\boldsymbol{p})\right]
\end{aligned}
\tag{9}
$$

Then, $\mathbb{E}\left[(\boldsymbol{y}-\boldsymbol{p})^T\mathcal{I}(\boldsymbol{\alpha})(\boldsymbol{y}-\boldsymbol{p})\right]$ can be converted to:

$$
\begin{aligned}
&\mathbb{E}\left[(\boldsymbol{y}-\boldsymbol{p})^T\mathcal{I}(\boldsymbol{\alpha})(\boldsymbol{y}-\boldsymbol{p})\right] \\
=&\boldsymbol{y}^T\mathcal{I}(\boldsymbol{\alpha})\boldsymbol{y} - 2\boldsymbol{y}^T\mathcal{I}(\boldsymbol{\alpha})\mathbb{E}[\boldsymbol{p}] + \mathbb{E}\left[\boldsymbol{p}^T\mathcal{I}(\boldsymbol{\alpha})\boldsymbol{p}\right] \\
=&\boldsymbol{y}^T\mathcal{I}(\boldsymbol{\alpha})\boldsymbol{y} - 2\boldsymbol{y}^T\mathcal{I}(\boldsymbol{\alpha})\mathbb{E}[\boldsymbol{p}] + \mathbb{E}[\boldsymbol{p}]^T\mathcal{I}(\boldsymbol{\alpha})\mathbb{E}[\boldsymbol{p}] - \mathbb{E}[\boldsymbol{p}]^T\mathcal{I}(\boldsymbol{\alpha})\mathbb{E}[\boldsymbol{p}] + \mathbb{E}\left[\boldsymbol{p}^T\mathcal{I}(\boldsymbol{\alpha})\boldsymbol{p}\right] \\
=&(\boldsymbol{y}-\mathbb{E}[\boldsymbol{p}])^T\mathcal{I}(\boldsymbol{\alpha})(\boldsymbol{y}-\mathbb{E}[\boldsymbol{p}]) + \mathbb{E}\left[(\boldsymbol{p}-\mathbb{E}[\boldsymbol{p}])^T\mathcal{I}(\boldsymbol{\alpha})(\boldsymbol{p}-\mathbb{E}[\boldsymbol{p}])\right]
\end{aligned}
\tag{10}
$$

Since $\mathbb{E}[\boldsymbol{p}] = \boldsymbol{\alpha}/\alpha_0$, $\mathrm{Cov}(p_i, p_j) = (\delta_{ij}\alpha_i\alpha_0 - \alpha_i\alpha_j)/(\alpha_0^2(\alpha_0+1))$, where δ_{ij} is the Kronecker delta (i.e. $\delta_{ij} = 1$ if $i = j$, else $\delta_{ij} = 0$), combined with the value of the FIM (Eq. 6), we have

$$
\begin{aligned}
&(\boldsymbol{y}-\mathbb{E}[\boldsymbol{p}])^T\mathcal{I}(\boldsymbol{\alpha})(\boldsymbol{y}-\mathbb{E}[\boldsymbol{p}]) \\
=&(\boldsymbol{y}-\frac{\boldsymbol{\alpha}}{\alpha_0})^T\begin{bmatrix}\psi^{(1)}(\alpha_1)-\psi^{(1)}(\alpha_0) & -\psi^{(1)}(\alpha_0) & \cdots & -\psi^{(1)}(\alpha_0) \\ -\psi^{(1)}(\alpha_0) & \psi^{(1)}(\alpha_2)-\psi^{(1)}(\alpha_0) & \cdots & -\psi^{(1)}(\alpha_0) \\ \vdots & \vdots & \ddots & \vdots \\ -\psi^{(1)}(\alpha_0) & -\psi^{(1)}(\alpha_0) & \cdots & \psi^{(1)}(\alpha_K)-\psi^{(1)}(\alpha_0)\end{bmatrix}(\boldsymbol{y}-\frac{\boldsymbol{\alpha}}{\alpha_0}) \\
=&(\boldsymbol{y}-\frac{\boldsymbol{\alpha}}{\alpha_0})^T\left(\mathrm{diag}([\psi^{(1)}(\alpha_1),\cdots,\psi^{(1)}(\alpha_K)]) - \psi^{(1)}(\alpha_0)\mathbf{1}\mathbf{1}^T\right)(\boldsymbol{y}-\frac{\boldsymbol{\alpha}}{\alpha_0}) \\
=&\sum_{i=1}^{K}(y_i-\frac{\alpha_i}{\alpha_0})^2\psi^{(1)}(\alpha_i),
\end{aligned}
\tag{11}
$$

and

$$
\begin{aligned}
&\mathbb{E}\left[(\boldsymbol{p}-\mathbb{E}[\boldsymbol{p}])^T\mathcal{I}(\boldsymbol{\alpha})(\boldsymbol{p}-\mathbb{E}[\boldsymbol{p}])\right] \\
=&\sum_{i,j=1}^{K}\mathrm{Cov}(p_i,p_j)\mathcal{I}(i,j) = \sum_{i,j=1}^{K}\frac{\delta_{ij}\alpha_i\alpha_0-\alpha_i\alpha_j}{\alpha_0^2(\alpha_0+1)}(\delta_{ij}\psi^{(1)}(\alpha_i)-\psi^{(1)}(\alpha_0)) \\
=&\sum_{i=1}^{K}\frac{\alpha_i\alpha_0-\alpha_i^2}{\alpha_0^2(\alpha_0+1)}\psi^{(1)}(\alpha_i) - \sum_{i=1}^{K}\frac{\alpha_i\alpha_0}{\alpha_0^2(\alpha_0+1)}\psi^{(1)}(\alpha_0) + \sum_{i,j=1}^{K}\frac{\alpha_i\alpha_j}{\alpha_0^2(\alpha_0+1)}\psi^{(1)}(\alpha_0) \\
=&\sum_{i=1}^{K}\frac{\alpha_i(\alpha_0-\alpha_i)}{\alpha_0^2(\alpha_0+1)}\psi^{(1)}(\alpha_i).
\end{aligned}
\tag{12}
$$

Plugging (11) and (12) into (10), we have

$$
\begin{aligned}
&\mathbb{E}\left[(\boldsymbol{y}-\boldsymbol{p})^T\mathcal{I}(\boldsymbol{\alpha})(\boldsymbol{y}-\boldsymbol{p})\right] \\
=&\sum_{i=1}^{K}(y_i-\frac{\alpha_i}{\alpha_0})^2\psi^{(1)}(\alpha_i) + \frac{\alpha_i(\alpha_0-\alpha_i)}{\alpha_0^2(\alpha_0+1)}\psi^{(1)}(\alpha_i) = \sum_{i=1}^{K}\left((y_i-\frac{\alpha_i}{\alpha_0})^2 + \frac{\alpha_i(\alpha_0-\alpha_i)}{\alpha_0^2(\alpha_0+1)}\right)\psi^{(1)}(\alpha_i)
\end{aligned}
\tag{13}
$$

Furthermore, for the KL term, we have

$$
\begin{aligned}
&D_{\mathrm{KL}}(Dir(\boldsymbol{p}|\hat{\boldsymbol{\alpha}})\|Dir(\boldsymbol{p}|\mathbf{1})) \\
&= \log\Gamma(\sum_{k=1}^{K}\hat{\alpha}_k) - \log\Gamma(K) - \sum_{k=1}^{K}\log\Gamma(\hat{\alpha}_k) + \sum_{k=1}^{K}(\hat{\alpha}_k - 1)\left[\psi(\hat{\alpha}_k) - \psi(\sum_{j=1}^{K}\hat{\alpha}_j)\right]
\end{aligned} \tag{14}
$$

Plugging (7), (13) and (14) into (8), we can obtain the final objective of Eq. 5

$$
\begin{aligned}
&\mathbb{E}\left[-\log p(\boldsymbol{y}|\boldsymbol{p},\boldsymbol{\alpha},\sigma^2)\right] + \lambda D_{\mathrm{KL}}(Dir(\boldsymbol{p}|\hat{\boldsymbol{\alpha}})\|Dir(\boldsymbol{p}|\mathbf{1})) \\
&\propto \underbrace{\sum_{i=1}^{K}\left((y_i - \frac{\alpha_i}{\alpha_0})^2 + \frac{\alpha_i(\alpha_0-\alpha_i)}{\alpha_0^2(\alpha_0+1)}\right)\psi^{(1)}(\alpha_i)}_{\mathcal{L}_i^{\mathcal{I}\text{-MSE}}} - \lambda_1 \underbrace{\left(\sum_{i=1}^{K}\log\psi^{(1)}(\alpha_i) + \log(1 - \sum_{i=1}^{K}\frac{\psi^{(1)}(\alpha_0)}{\psi^{(1)}(\alpha_i)})\right)}_{\mathcal{L}_i^{|\mathcal{I}|}} \\
&+ \lambda_2 \underbrace{\left(\log\Gamma(\sum_{k=1}^{K}\hat{\alpha}_k) - \log\Gamma(K) - \sum_{k=1}^{K}\log\Gamma(\hat{\alpha}_k) + \sum_{k=1}^{K}(\hat{\alpha}_k - 1)\left[\psi(\hat{\alpha}_k) - \psi(\sum_{j=1}^{K}\hat{\alpha}_j)\right]\right)}_{\mathcal{L}_i^{\mathrm{KL}}}.
\end{aligned}
$$

A.3. Proof of Theorem 3.1

Theorem 3.1 ((Germain et al., 2009; Alquier et al., 2016; Masegosa, 2020)). Given a data distribution P over $\mathcal{X}\times\mathcal{Y}$, a hypothesis set $\boldsymbol{\theta}$, a prior distribution π over $\boldsymbol{\theta}$, for any $\delta\in(0,1]$, and $\lambda>0$, with probability at least $1-\delta$ over samples $D\sim P^n$, we have for all posterior ρ,

$$
\mathbb{E}_{\rho(\boldsymbol{\theta})}[L(\boldsymbol{\theta})] \le \mathbb{E}_{\rho(\boldsymbol{\theta})}[\hat{L}(\boldsymbol{\theta},D)] + \frac{1}{\lambda}\left[D_{\mathrm{KL}}(\rho\|\pi) + \log\frac{1}{\delta} + \Psi_{P,\pi}(\lambda,n)\right]
$$

where $\Psi_{P,\pi}(\lambda,n) = \log\mathbb{E}_{\pi(\boldsymbol{\theta})}\mathbb{E}_{D\sim P^n}\left[e^{\lambda(L(\boldsymbol{\theta})-\hat{L}(\boldsymbol{\theta},D))}\right]$.

Proof. The Donsker-Varadhan's change of measure states that for any measurable function $\phi:\boldsymbol{\theta}\to\mathbb{R}$, we have

$$
\mathbb{E}_{\rho(\boldsymbol{\theta})}[\phi(\boldsymbol{\theta})] \le D_{\mathrm{KL}}(\rho\|\pi) + \log\mathbb{E}_{\pi(\boldsymbol{\theta})}[e^{\phi(\boldsymbol{\theta})}]
$$

Thus, with $\phi(\boldsymbol{\theta}) := \lambda\left(L(\boldsymbol{\theta}) - \hat{L}(\boldsymbol{\theta},D)\right)$, we obtain $\forall\rho$ on $\boldsymbol{\theta}$:

$$
\begin{aligned}
\mathbb{E}_{\rho(\boldsymbol{\theta})}\left[\lambda\left(L(\boldsymbol{\theta}) - \hat{L}(\boldsymbol{\theta},D)\right)\right] &= \lambda\left(\mathbb{E}_{\rho(\boldsymbol{\theta})}\left[L(\boldsymbol{\theta})\right] - \mathbb{E}_{\rho(\boldsymbol{\theta})}\left[\hat{L}(\boldsymbol{\theta},D)\right]\right) \\
&\le D_{\mathrm{KL}}(\rho\|\pi) + \log\mathbb{E}_{\pi(\boldsymbol{\theta})}\left[e^{\lambda\left(L(\boldsymbol{\theta})-\hat{L}(\boldsymbol{\theta},D)\right)}\right]
\end{aligned}
$$

Next, we apply Markov's inequality on the random variable $\zeta_\pi(D) := \mathbb{E}_{\pi(\boldsymbol{\theta})}\left[e^{\lambda\left(L(\boldsymbol{\theta})-\hat{L}(\boldsymbol{\theta},D)\right)}\right]$:

$$
\Pr\left(\zeta_\pi(D) \le \frac{1}{\delta}\mathbb{E}_{D\sim P^n}\left[\zeta_\pi(D)\right]\right) \ge 1-\delta
$$

This implies that with probability at least $1-\delta$ over the choice of $D\sim P^n$, we have $\forall\rho$ on $\boldsymbol{\theta}$:

$$
\Pr\left(\mathbb{E}_{\rho(\boldsymbol{\theta})}[L(\boldsymbol{\theta})] \le \mathbb{E}_{\rho(\boldsymbol{\theta})}[\hat{L}(\boldsymbol{\theta},D)] + \frac{1}{\lambda}\left[D_{\mathrm{KL}}(\rho\|\pi) + \log\frac{1}{\delta} + \Psi_{P,\pi}(\lambda,n)\right]\right) \ge 1-\delta,
$$

where $\Psi_{P,\pi}(\lambda,n) = \log\mathbb{E}_{\pi(\boldsymbol{\theta})}\mathbb{E}_{D\sim P^n}\left[e^{\lambda(L(\boldsymbol{\theta})-\hat{L}(\boldsymbol{\theta},D))}\right]$.

B. Derivations for Uncertainty Measures and Energy Distance

The Dirichlet distribution is parameterized by its concentration parameters $\boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_K], \forall \alpha_c > 0$, defined as:

$$Dir(\boldsymbol{p}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\prod_{c=1}^{K} \Gamma(\alpha_c)} \prod_{c=1}^{K} p_k^{\alpha_c - 1}, \alpha_0 = \sum_{c=1}^{K} \alpha_c$$

where $\boldsymbol{p} \in \Delta^{K-1}$, and $\Gamma(\cdot)$ is the *gamma* function. Note that the parameters $\boldsymbol{\alpha}$ is calculated by passing the input sample through the evidential network ($\boldsymbol{\alpha} = f_{\boldsymbol{\theta}}(\boldsymbol{x}) + 1 \in \mathbb{R}_+^K$). The following derivation is adapted from the Appendix of (Malinin & Gales, 2018).

B.1. Expected Entropy of Dirichlet-based Uncertainty Models

The derivation of the expected entropy is as follows:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{p}\sim Dir(\boldsymbol{\alpha})}\left[\mathcal{H}[p(y \mid \boldsymbol{p})]\right] &= \int_{\mathcal{S}^{K-1}} Dir(\boldsymbol{p}|\boldsymbol{\alpha}) \left(-\sum_{c=1}^{K} p_c \ln p_c\right) d\boldsymbol{p} \\
&= -\sum_{c=1}^{K} \int_{\mathcal{S}^{K-1}} Dir(\boldsymbol{p}|\boldsymbol{\alpha}) \left(p_c \ln p_c\right) d\boldsymbol{p} \\
&= -\sum_{c=1}^{K} \int_{\mathcal{S}^{K-1}} \frac{\Gamma(\alpha_0)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} p_k^{\alpha_k - 1} \left(p_c \ln p_c\right) d\boldsymbol{p} \\
&= -\sum_{c=1}^{K} \int_{\mathcal{S}^{K-1}} \frac{\alpha_c}{\alpha_0} \frac{\Gamma(\alpha_0 + 1)}{\Gamma(\alpha_c + 1) \prod_{k=1,k\neq c}^{K} \Gamma(\alpha_k)} \prod_{k=1,k\neq c}^{K} p_k^{\alpha_k - 1} p_c^{\alpha_c} \ln p_c d\boldsymbol{p} \\
&= -\sum_{c=1}^{K} \frac{\alpha_c}{\alpha_0} \int_{\mathcal{S}^{K-1}} \mathbb{E}_{\boldsymbol{p}\sim Dir([\alpha_1, \cdots, \alpha_{c-1}, \alpha_c+1, \alpha_{c+1}, \cdots, \alpha_K])}[\ln p_c] d\boldsymbol{p} \\
&= -\sum_{c=1}^{K} \frac{\alpha_c}{\alpha_0} \left(\psi(\alpha_c + 1) - \psi(\alpha_0 + 1)\right)
\end{aligned} \tag{15}$$

The last third equation comes from the fact that $\Gamma(n) = (n-1)!$. Since the expected entropy captures the peaks of the output distribution $p(y \mid \boldsymbol{p})$, it is used to measure *data uncertainty*. More specifically, lower entropy means that the model concentrates all probability mass on one class, while high entropy indicates that all $\boldsymbol{p}$ generated by $Dir(\boldsymbol{\alpha})$ are more uniformly distributed.

B.2. Mutual Information of Dirichlet-based Uncertainty Models

In the DBU models, the mutual information between the labels y and the categorical $\boldsymbol{p}$ can be deduced by computing the difference between the entropy of the expected distribution and the expected entropy of the distribution, which can be viewed as the difference between the total amount of uncertainty and the data uncertainty.

$$\underbrace{I\left[y, \boldsymbol{p} \mid \boldsymbol{x}, \mathcal{D}\right]}_{\text{Distributional Uncertainty}} = \underbrace{\mathcal{H}\left[\mathbb{E}_{p(\boldsymbol{p}|\boldsymbol{x},\mathcal{D})}[p(y \mid \boldsymbol{p})]\right]}_{\text{Total Uncertainty}} - \underbrace{\mathbb{E}_{p(\boldsymbol{p}|\boldsymbol{x},\mathcal{D})}[\mathcal{H}[p(y \mid \boldsymbol{p})]]}_{\text{Expected Data Uncertainty}}. \tag{16}$$

Assuming that point estimate $p(\boldsymbol{\theta}|\mathcal{D}) = \delta(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \Rightarrow p(\boldsymbol{p}|\boldsymbol{x}, \mathcal{D}) \approx p(\boldsymbol{p}|\boldsymbol{x}, \hat{\boldsymbol{\theta}}) = Dir(\boldsymbol{p}|\boldsymbol{\alpha})$ is sufficient given appropriate regularization and training data size, the mutual information can be simplified by

$$\begin{aligned}
\underbrace{I\left[y, \boldsymbol{p} \mid \boldsymbol{x}, \mathcal{D}\right]}_{\text{Distributional Uncertainty}} &\approx \underbrace{\mathcal{H}\left[\mathbb{E}_{\boldsymbol{p}\sim Dir(\boldsymbol{\alpha})}[p(y \mid \boldsymbol{p})]\right]}_{\text{Total Uncertainty}} - \underbrace{\mathbb{E}_{\boldsymbol{p}\sim Dir(\boldsymbol{\alpha})}[\mathcal{H}[p(y \mid \boldsymbol{p})]]}_{\text{Expected Data Uncertainty}} \\
&= -\sum_{c=1}^{K} \frac{\alpha_c}{\alpha_0} \ln \frac{\alpha_c}{\alpha_0} + \sum_{c=1}^{K} \frac{\alpha_c}{\alpha_0} \left(\psi(\alpha_c + 1) - \psi(\alpha_0 + 1)\right) \\
&= -\sum_{c=1}^{K} \frac{\alpha_c}{\alpha_0} \left(\ln \frac{\alpha_c}{\alpha_0} - \psi(\alpha_c + 1) + \psi(\alpha_0 + 1)\right)
\end{aligned} \tag{17}$$

The second term in this derivation is from the results of B.1. The mutual information is often used to measure *distributional uncertainty*, as it captures the uniform output distribution that excludes data uncertainty. High mutual information indicates a uniform distribution of expected categorical probability with low data uncertainty.

B.3. Differential Entropy of Dirichlet-based Uncertainty Models

The derivation of the differential entropy is as follows:

$$\begin{aligned}
\mathcal{H}\left[Dir(\boldsymbol{p} \mid \boldsymbol{\alpha})\right] &= -\int_{\mathcal{S}^{K-1}} Dir(\boldsymbol{p}|\boldsymbol{\alpha}) \ln Dir(\boldsymbol{p}|\boldsymbol{\alpha}) d\boldsymbol{p} \\
&= -\int_{\mathcal{S}^{K-1}} Dir(\boldsymbol{p}|\boldsymbol{\alpha}) \left(\ln \Gamma\left(\alpha_0\right) - \sum_{c=1}^{K} \ln \Gamma\left(\alpha_c\right) + \sum_{c=1}^{K} (\alpha_c - 1) \ln p_k \right) d\boldsymbol{p} \\
&= \sum_{c=1}^{K} \ln \Gamma\left(\alpha_c\right) - \ln \Gamma\left(\alpha_0\right) - \sum_{c=1}^{K} \left(\alpha_c - 1\right) \mathbb{E}_{\boldsymbol{p} \sim Dir(\boldsymbol{\alpha})}[\ln p_c] \\
&= \sum_{c=1}^{K} \ln \Gamma\left(\alpha_c\right) - \ln \Gamma\left(\alpha_0\right) - \sum_{c=1}^{K} \left(\alpha_c - 1\right) \left(\psi\left(\alpha_c\right) - \psi\left(\alpha_0\right)\right)
\end{aligned} \tag{18}$$

The last equation comes from $\mathbb{E}_{\boldsymbol{p} \sim Dir(\boldsymbol{\alpha})}[\ln p_c] = \psi\left(\alpha_c\right) - \psi\left(\alpha_0\right)$. Lower entropy indicates the model yields a sharper distribution, while high entropy denotes a more uniform Dirichlet distribution. Thus, differential entropy is also a common measure of *distributional uncertainty*.

B.4. Energy Distance

Energy distance is a metric that measures the distance between the distributions of random vectors. Let X and Y be independent random vectors in $\mathbb{R}^d$, with cumulative distribution function (CDF) F and G, respectively. The energy distance can be defined in terms of expected distances between the random vectors,

$$D(F, G) = \left(2\mathbb{E}\|X - Y\| - \mathbb{E}\|X - X'\| - \mathbb{E}\|Y - Y'\|\right)^{1/2}, \tag{19}$$

where X and X' (resp. Y and Y') are independent random variables whose probability distribution is F (resp. G), $\|\cdot\|$ denotes the Euclidean norm. Energy distance is zero if and only if the distributions are identical.

C. Experimental Details and Additional Results

C.1. Datasets

MNIST (LeCun, 1998) is a database of handwritten digits 0 to 9, consisting of a training set of $60,000$ examples and a test set of $10,000$ examples. Each input is composed of a $1 \times 28 \times 28$ tensor. We use **(80%, 20%)** to split the training samples into training and validation sets. For OOD detection experiment, we use **KMNIST** (Clanuwat et al., 2018) and **FashionMNIST** (Xiao et al., 2017) containing images of Japanese characters and images of clothes, respectively.

CIFAR10 (Krizhevsky et al., 2009) consists of $60,000$ images in 10 classes, including airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck. Among them, there are $50,000$ training images and $10,000$ test images. Each input is composed of a $3 \times 32 \times 32$ tensor. We use **(95%, 5%)** to split the training samples into training and validation sets. Street View House Numbers (**SVHN**) dataset (Netzer et al., 2018), a dataset containing digital images, and **CIFAR100** are used for OOD detection.

mini-ImageNet (Vinyals et al., 2016) dataset was proposed for few-shot learning evaluation. It contains 100 classes with 600 samples of 84×84 color images per class. These 100 classes are divided into 64, 16, and 20 classes respectively for sampling tasks for meta-training, meta-validation, and meta-test. For the few-shot classification task, we evaluate N-way K-shot classification tasks for $N \in \{5, 10\}$ and $K \in \{1, 5, 20, 50\}$ and report the average accuracy(%, top-1) and 95% confidence interval over $10,000$ few-shot episodes on meta-test set. Each episode contains randomly sampled N classes and K samples per class for adaptation, and $\min(15, K)$ query samples per class for evaluation. For OOD detection experiments, we use Caltech-UCSD Birds (**CUB**) dataset (Wah et al., 2011), contains $11,788$ images of 200 subcategories belonging to birds. Note that we randomly sample the same number of OOD examples as the query sample for OOD detection.

tiered-ImageNet (Ren et al., 2018) dataset is a larger subset of ILSVRC-12. Compared with the 100 classes of mini-ImageNet, it contains 608 classes (779,165 images), which are grouped into 34 higher-level nodes in the ImageNet human-curated hierarchy. The nodes are divided into $20, 6$, and 8 disjoint sets of training, validation, and testing nodes, with the corresponding classes forming their respective meta-sets. We conduct the few-shot classification and OOD detection experiments for CUB on this dataset using the same settings as mini-ImageNet.

C.2. Implementation details

For the MNIST and CIFAR10 datasets, it is implemented by adapting the code provided by (Charpentier et al., 2020). Following (Charpentier et al., 2020), we use 3 convolutional layers with 3 dense layers and VGG16 (Simonyan & Zisserman, 2014), respectively. *Softplus* is used in the last layer to get the non-negative output. We use a validation loss-based early termination strategy to train up to 200 epochs with a batch size of 64. The learning rate is set to 0.001 for MNIST and FMNIST, 0.0005 for CIFAR10. The coefficient λ of -$|\mathcal{I}|$ is set by grid-search $(0.1, 0.05, 0.01, 0.005, 0.001)$. The last chosen hyperparameter is 0.005 for MNIST, 0.01 for FMNIST and 0.05 for CIFAR10.

For the mini-ImageNet and tiered-ImageNet few-shot classification experiments, it is implemented by adapting the code provided by (Ghaffari et al., 2021). More specifically, we use WideResNet-28-10 pre-trained backbone from (Yang et al., 2021) as feature extractor to train the 1-layer classifier. *Softplus* is used as the activation function to obtain the non-negative output. The coefficient λ is also set by grid-search on the meta-validation set. Table 6 reports the last chosen hyperparameter for few-shot settings. Figure 6 shows the impact of λ under 5-way mini-ImageNet setting.

Table 6. List of hyperparameters for our approach.

	5way1shot	**5way5shot**	**5way20shot**	**5way50shot**	**10way1shot**	**10way5shot**	**10way20shot**	**10way50shot**
λ	0.01	0.05	0.005	0.01	0.1	0.01	0.01	0.01

C.3. Additional Experimental Results on OOD detection

Table 7 and 8 displays the AUPR and AUROC scores of OOD detection on CIFAR10 against CIFAR100 and SVHN, MNIST against KMNIST and FMNIST. Table 9 compares our method and label smoothing on the few-shot setting. We use maximum probability (Max.P) and entropy (Ent.) to measure uncertainty for label smoothing, and Max.P and α_0 for $\mathcal{I}$-EDL. All of these results consistently demonstrate our proposed method's superior OOD detection performance.

C.4. Additional Experimental Results on Few-shot Learning

Table 10 shows the AUROC scores of OOD detection under $\{5, 10\}$-way $\{1, 5, 20, 50\}$-shot of mini-ImageNet. For all the N-way K-shot tasks, AUROC with Max.P, α_0, D.Ent, and M.I. show impressive improvements on $\mathcal{I}$-EDL over EDL. For example, the improvements of OOD detection after using $\mathcal{I}$-EDL are all above **3.37%**, especially the improvement of Max.P-based OOD detection is up to **12.98%** under 10-way 1-shot. The experimental results of tiered-ImageNet are shown in Figure 5. We can observe the same experimental results as mini-ImageNet, which indicates that our method improves the performance of OOD detection, especially in the more challenging few-shot setting.

C.5. Additional Ablation Study

Figure 6 shows the effect of coefficient λ of the negative log determinant of FIM (-$|\mathcal{I}|$) under mini-ImageNet classification and OOD detection experiments. We plot the results under 5-way $\{5, 20, 50\}$-shot. It can be observed that the best coefficients for OOD detection based on different uncertainty measures show consistency, but not with the best coefficients for accuracy. More specifically, the best coefficients for OOD detection are all around 0.01, but accuracy prefers 0.001. This is a non-trivial problem because it involves a multi-objective optimization problem. In this work, the coefficient is ultimately a compromise choice that combines the performance of image classification and OOD detection. Whether there is a better way to optimize multiple objectives at the same time remains to be explored in the future.

Table 7. OOD detection results on (ID) CIFAR10 against (OOD) CIFAR100. Each experiment is run with 5 seeds.

	AUPR				AUROC			
Method	**Max. P**	α_0	**D. Ent.**	**M.I.**	**Max. P**	α_0	**D. Ent.**	**M.I.**
KL-PN	61.41 ± 2.8	61.53 ± 3.4	60.21 ± 3.2	61.66 ± 3.4	57.89 ± 1.8	58.43 ± 2.6	55.94 ± 3.3	58.53 ± 2.6
RKL-PN	55.42 ± 2.6	54.74 ± 2.8	55.40 ± 2.9	54.86 ± 2.9	54.24 ± 2.4	53.25 ± 2.9	54.15 ± 3.0	53.41 ± 3.0
PostN	81.96 ± 0.8	82.06 ± 0.8	82.34 ± 0.8	78.64 ± 1.7	80.49 ± 0.9	81.17 ± 1.1	81.51 ± 1.0	80.22 ± 1.0
EDL	84.30 ± 0.7	84.18 ± 0.7	84.32 ± 0.7	84.19 ± 0.7	80.96 ± 0.8	80.63 ± 1.0	80.99 ± 0.8	80.65 ± 1.0
$\mathcal{I}$-EDL	**85.35 ± 0.7**	**84.84 ± 0.6**	**85.40 ± 0.6**	**84.95 ± 0.7**	**83.55 ± 0.7**	**82.15 ± 0.5**	**83.69 ± 0.7**	**82.44 ± 0.5**

Table 8. OOD detection results on (ID) MNIST against (OOD) KMNIST and FMNIST, (ID) CIFAR10 against (OOD) SVHN. Each experiment is run with 5 seeds.

	AUPR				AUROC			
Method	**Max. P**	α_0	**D. Ent.**	**M.I.**	**Max. P**	α_0	**D. Ent.**	**M.I.**
				MNIST $\rightarrow$ KMNIST				
EDL	97.02 ± 0.7	96.31 ± 0.2	96.92 ± 0.9	96.41 ± 1.8	96.59 ± 0.6	96.18 ± 1.3	96.49 ± 0.8	96.22 ± 1.3
$\mathcal{I}$-EDL	98.34 ± 0.2	98.33 ± 0.2	98.34 ± 0.2	98.33 ± 0.2	98.00 ± 0.3	97.97 ± 0.3	97.99 ± 0.3	97.97 ± 0.3
				MNIST $\rightarrow$ FMNIST				
EDL	98.10 ± 0.4	98.08 ± 0.4	98.10 ± 0.4	98.09 ± 0.4	97.39 ± 0.6	97.40 ± 0.5	97.48 ± 0.5	97.43 ± 0.5
$\mathcal{I}$-EDL	98.89 ± 0.3	98.86 ± 0.3	98.89 ± 0.3	98.87 ± 0.2	98.49 ± 0.3	98.41 ± 0.4	98.48 ± 0.4	98.42 ± 0.4
				CIFAR10 $\rightarrow$ SVHN				
EDL	78.87 ± 3.5	79.12 ± 3.7	78.91 ± 3.5	79.11 ± 3.7	80.64 ± 4.2	81.06 ± 4.5	80.71 ± 4.3	81.05 ± 4.5
$\mathcal{I}$-EDL	83.26 ± 2.4	82.96 ± 2.2	83.31 ± 2.5	83.06 ± 2.2	87.58 ± 2.0	86.79 ± 1.3	87.69 ± 2.1	87.01 ± 1.5

Table 9. Comparsion of label smoothing and $\mathcal{I}$-EDL on OOD detection AUPR of mini-ImageNet against CUB under 5-way and $\{1, 5, 20, 50\}$-shot settings. Each experiment is run for over $10,000$ few-shot episodes.

	5-Way 1-shot		5-Way 5-shot		5-way 20-shot		5-Way 50-shot	
Method	**Max. P**	**Ent.-α_0**	**Max. P**	**Ent.-α_0**	**Max. P**	**Ent.-α_0**	**Max. P**	**Ent.-α_0**
Label smoothing	72.03 ± 0.2	73.00 ± 0.2	77.17 ± 0.2	77.11 ± 0.2	76.11 ± 0.2	75.35 ± 0.2	74.76 ± 0.2	73.86 ± 0.2
$\mathcal{I}$-EDL	71.95 ± 0.2	74.79 ± 0.2	82.04 ± 0.2	82.49 ± 0.2	84.31 ± 0.2	85.42 ± 0.2	84.68 ± 0.2	84.91 ± 0.2
Δ	**-0.08**	**1.79**	**4.87**	**5.38**	**8.20**	**10.07**	**9.92**	**11.05**

C.6. Additional Analysis of Uncertainty estimation

Figure 7 and 8 shows density plots of the normalized uncertainty measures for MNIST vs FMNIST, and MNIST vs KMNIST, respectively. The uncertainty measures include precision α_0, $\max_c p_c$, differential entropy and mutual information. We normalize each uncertainty value u_c by $u_c = (u_c - \min_i u_i)/(\max_i u_i - \min_i u_i)$. We also report the energy distance of two distributions, with higher values indicating more separability. It can be observed that $\mathcal{I}$-EDL produces sharper prediction peaks than EDL in the in-distribution (MNIST) region. Although not using OOD data, our method also makes the uncertainty of OOD data more aggregated.

Table 10. **AUROC** scores of OOD detection against CUB under $\{5, 10\}$-way and $\{1, 5, 20, 50\}$-shot settings. Each experiment is run for over $10,000$ few-shot episodes.

	5-way 1-shot				**10-way 1-shot**			
Method	**Max.P**	α_0	**D. Ent.**	**M.I.**	**Max.P**	α_0	**D. Ent.**	**M.I.**
EDL	61.88 ± 0.27	59.72 ± 0.31	63.60 ± 0.31	60.42 ± 0.30	55.83 ± 0.22	63.02 ± 0.29	63.06 ± 0.27	63.05 ± 0.29
$\mathcal{I}$-**EDL**	67.46 ± 0.28	70.08 ± 0.30	69.51 ± 0.30	70.03 ± 0.30	68.81 ± 0.22	69.29 ± 0.22	68.80 ± 0.22	69.29 ± 0.22
Δ	**5.58**	**10.36**	**5.91**	**9.61**	**12.98**	**6.27**	**5.74**	**6.24**
	5-way 5-shot				**10-way 5-shot**			
Method	**Max.P**	α_0	**D. Ent.**	**M.I.**	**Max.P**	α_0	**D. Ent.**	**M.I.**
EDL	69.71 ± 0.25	72.01 ± 0.33	72.97 ± 0.29	72.14 ± 0.32	68.59 ± 0.23	73.23 ± 0.23	72.70 ± 0.23	73.16 ± 0.23
$\mathcal{I}$-**EDL**	79.33 ± 0.22	79.81 ± 0.22	79.60 ± 0.22	79.79 ± 0.22	79.34 ± 0.23	80.29 ± 0.22	78.36 ± 0.20	79.91 ± 0.21
Δ	**9.62**	**7.80**	**6.63**	**7.65**	**10.75**	**7.06**	**5.66**	**6.75**
	5-way 20-shot				**10-way 20-shot**			
Method	**Max.P**	α_0	**D. Ent.**	**M.I.**	**Max.P**	α_0	**D. Ent.**	**M.I.**
EDL	76.59 ± 0.24	76.16 ± 0.29	76.75 ± 0.27	76.18 ± 0.29	71.69 ± 0.19	74.08 ± 0.20	73.88 ± 0.19	74.05 ± 0.20
$\mathcal{I}$-**EDL**	82.04 ± 0.21	83.38 ± 0.22	83.00 ± 0.21	83.29 ± 0.21	79.66 ± 0.16	80.74 ± 0.16	80.07 ± 0.15	80.61 ± 0.16
Δ	**5.45**	**7.22**	**6.25**	**7.11**	**7.96**	**6.66**	**6.19**	**6.56**
	5-way 50-shot				**10-way 50-shot**			
Method	**Max.P**	α_0	**D. Ent.**	**M.I.**	**Max.P**	α_0	**D. Ent.**	**M.I.**
EDL	79.30 ± 0.20	78.74 ± 0.24	79.21 ± 0.22	78.76 ± 0.24	73.67 ± 0.17	74.43 ± 0.18	74.32 ± 0.17	74.40 ± 0.18
$\mathcal{I}$-**EDL**	82.55 ± 0.17	83.17 ± 0.19	83.29 ± 0.18	83.20 ± 0.19	77.39 ± 0.15	77.80 ± 0.16	77.72 ± 0.15	77.78 ± 0.16
Δ	**3.25**	**4.43**	**4.08**	**4.44**	**3.72**	**3.37**	**3.40**	**3.38**

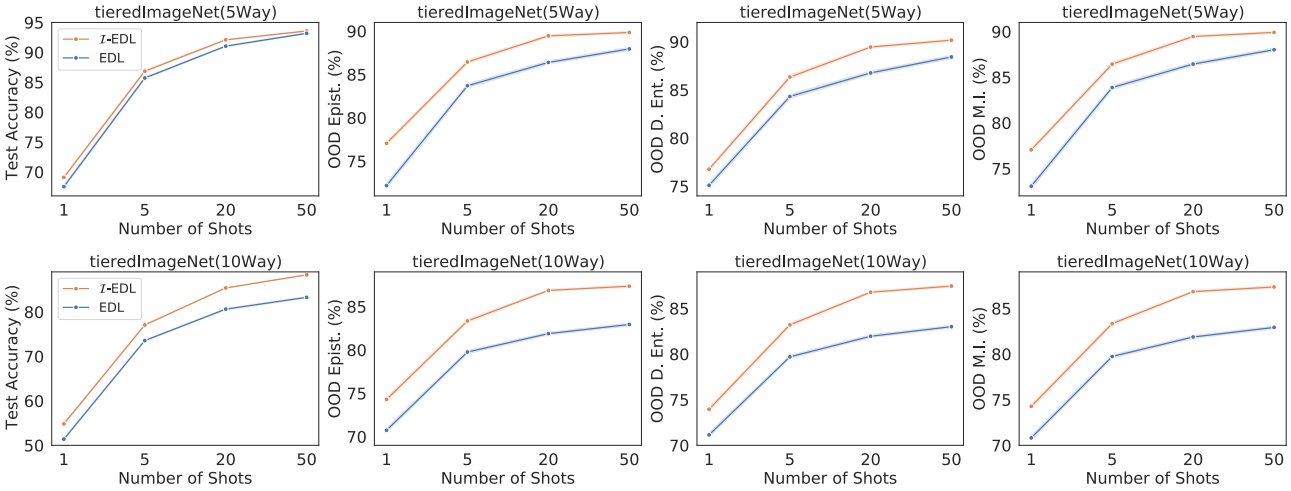


Figure 5. Results on tiered-ImageNet $\{5, 10\}$-way $\{1, 5, 20, 50\}$-shot classification (**Accuracy**) and OOD detection against CUB (**AUPR**). The error bars are almost non-existent (i.e., less than 0.03%), since over $10,000$ trials were performed for each point. $\mathcal{I}$-EDL produces statistically significant improvements over a wide range of the number of shots.

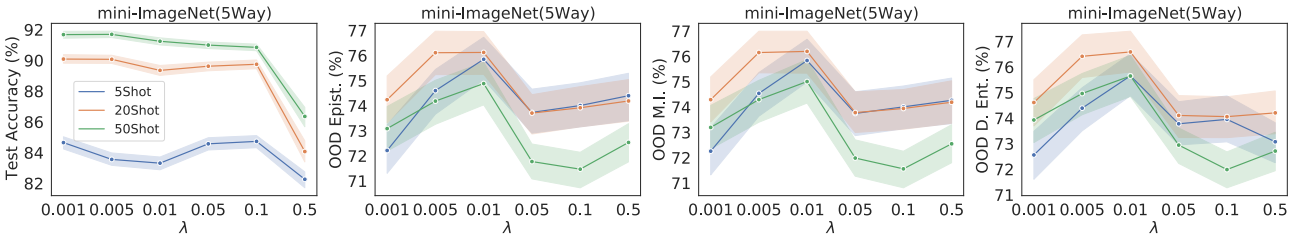


Figure 6. The impact of coefficient λ of the negative log determinant of the FIM (-$|\mathcal{I}|$) under 5-way mini-ImageNet classification and OOD detection experiments.

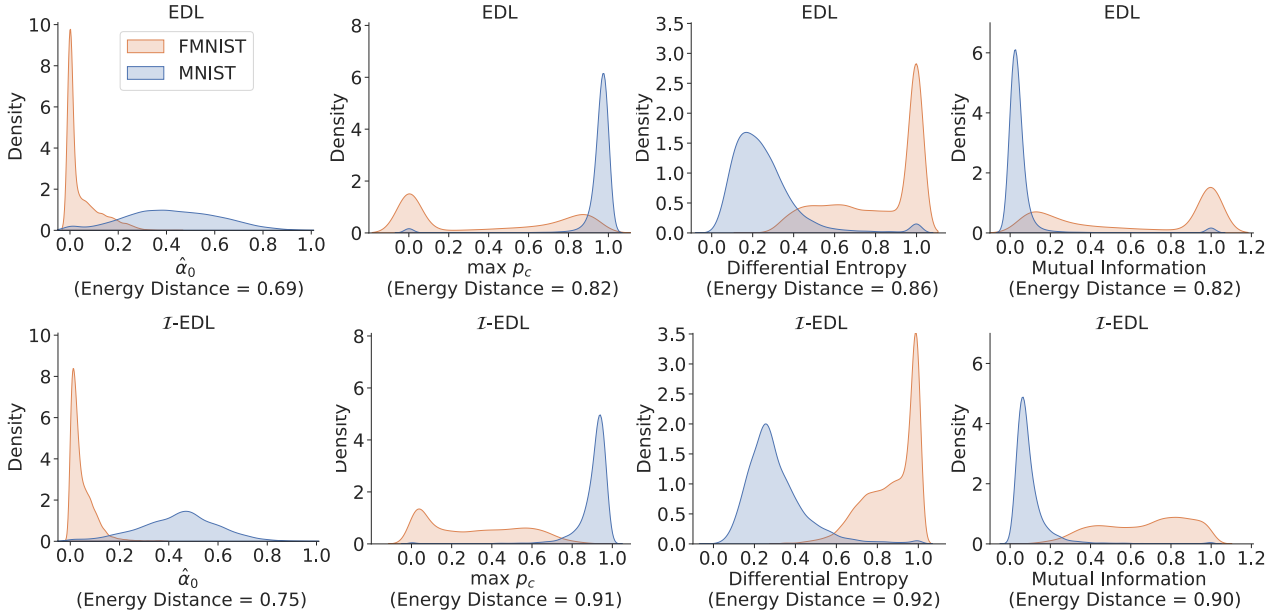


Figure 7. Uncertainty representation for ID (MNIST) and OOD (FMNIST).

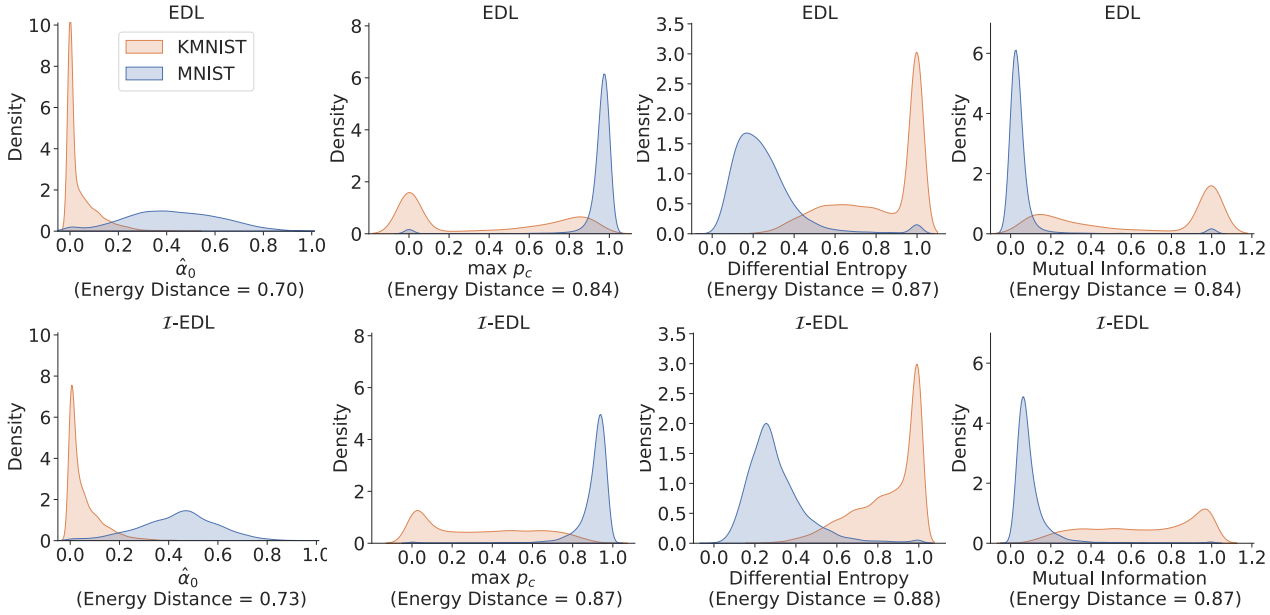


Figure 8. Uncertainty representation for ID (MNIST) and OOD (KMNIST).