Predictive Uncertainty Estimation via Prior Networks

2018 NeurIPS

Xin Gao

2023.12.29

Introduction

- This work proposes a new framework for modeling predictive uncertainty called Prior Networks (PNs) which explicitly models distributional uncertainty.
- PNs do this by parameterizing a prior distribution over predictive distributions.

model uncertainty, data uncertainty and distributional uncertainty

- Model uncertainty, Epistemic uncertainty (reducible)
- Data uncertainty, Aleatoric uncertainty (irreducible)
- Distributional uncertainty arises due to mismatch between the training and test distributions (also called dataset shift)

Bayesian uncertainty

Distribution $p(\boldsymbol{x}, y)$ over input features $\boldsymbol{x}$ and labels y

A classification model $P(\omega_c|\boldsymbol{x}^*, \mathcal{D})$ trained on a finite dataset $\mathcal{D} = \{\boldsymbol{x_j}, y_j\}_{j=1}^{N} \sim p(\boldsymbol{x}, y)$

$$\mathrm{P}(\omega_c \mid \boldsymbol{x}^*, \mathcal{D}) = \int \underbrace{\mathrm{P}(\omega_c \mid \boldsymbol{x}^*, \boldsymbol{\theta})}_{\text{Data}} \underbrace{\mathrm{p}(\boldsymbol{\theta} \mid \mathcal{D})}_{\text{Model}} d\boldsymbol{\theta}$$

Approximation $q(\theta)$

$$\mathrm{p}(\boldsymbol{\theta} \mid \mathcal{D}) \approx \mathrm{q}(\boldsymbol{\theta})$$

Sampling

$$\mathrm{P}(\omega_c \mid \boldsymbol{x}^*, \mathcal{D}) \approx \frac{1}{M} \sum_{i=1}^{M} \mathrm{P}\left(\omega_c \mid \boldsymbol{x}^*, \boldsymbol{\theta}^{(i)}\right), \boldsymbol{\theta}^{(i)} \sim \mathrm{q}(\boldsymbol{\theta})$$

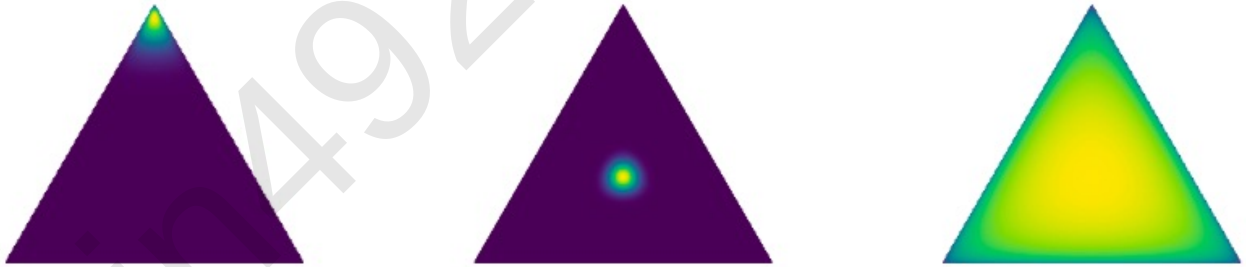
Distribution of distribution

A categorical distribution μ over class labels y

$$\left\{\mathrm{P}\left(\omega_c \mid \boldsymbol{x}^*, \boldsymbol{\theta}^{(i)}\right)\right\}_{i=1}^{M}$$

(a) Ensemble (b) Distribution

Figure 1: Distributions on a Simplex



(a) Confident Prediction (b) High data uncertainty (c) Out-of-distribution

Figure 2: Desired behaviors of a distribution over distributions

Explicitly parameterize a distribution over distributions on a simplex

- **Confident in-distribution data**: a sharp distribution centered on one of the corners of the simplex
- **Noise or class overlap (data uncertainty)**: a sharp distribution focused on the center of the simplex
- **Confident out-of-distribution**: a flat distribution, large uncertainty

Prior Network

In Prior Networks **data uncertainty** is described by the point-estimate categorical distribution μ and **distributional uncertainty** is described by the distribution over predictive categoricals $p(\mu|\boldsymbol{x}^*, \theta)$

$$\mathrm{P}(\omega_c \mid \boldsymbol{x}^*, \mathcal{D}) = \iint \underbrace{\mathrm{p}(\omega_c \mid \boldsymbol{\mu})}_{\text{Data}} \underbrace{\mathrm{p}(\boldsymbol{\mu} \mid \boldsymbol{x}^*, \boldsymbol{\theta})}_{\text{Distributional}} \underbrace{\mathrm{p}(\boldsymbol{\theta} \mid \mathcal{D})}_{\text{Model}} d\boldsymbol{\mu} d\boldsymbol{\theta}$$

Distributions over a simplex: a Dirichlet, Mixture of Dirichlet distributions or the Logistic-Normal distribution

$$\mathrm{Dir}(\boldsymbol{\mu} \mid \boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\prod_{c=1}^{K} \Gamma(\alpha_c)} \prod_{c=1}^{K} \mu_c^{\alpha_c - 1}, \quad \alpha_c > 0, \alpha_0 = \sum_{c=1}^{K} \alpha_c$$

Higher values of α_0 lead to sharper distributions

Uncertainty $\frac{K}{\alpha_0}$

Dirichlet Prior Network

A Prior Network which parametrizes a Dirichlet will be referred to as a Dirichlet Prior Network (DPN). A DPN will generate the concentration parameters α of the Dirichlet distribution.

$$\mathrm{p}\left(\boldsymbol{\mu} \mid \boldsymbol{x}^{*}; \hat{\boldsymbol{\theta}}\right) = \mathrm{Dir}(\boldsymbol{\mu} \mid \boldsymbol{\alpha}), \quad \boldsymbol{\alpha} = \boldsymbol{f}\left(\boldsymbol{x}^{*}; \hat{\boldsymbol{\theta}}\right)$$

The posterior over class labels will be given by the mean of the Dirichlet:

$$\mathrm{P}\left(\omega_c \mid \boldsymbol{x}^{*}; \hat{\boldsymbol{\theta}}\right) = \int \mathrm{p}(\omega_c \mid \boldsymbol{\mu}) \mathrm{p}\left(\boldsymbol{\mu} \mid \boldsymbol{x}^{*}; \hat{\boldsymbol{\theta}}\right) d\boldsymbol{\mu} = \frac{\alpha_c}{\alpha_0}$$

If an exponential output function is used for the DPN, where $\alpha_c = e^{z_c}$, then the expected posterior probability of a label ω_c is given by the output of the softmax

$$\mathrm{P}\left(\omega_c \mid \boldsymbol{x}^{*}; \hat{\boldsymbol{\theta}}\right) = \frac{e^{z_c(\boldsymbol{x}^{*})}}{\sum_{k=1}^{K} e^{z_k(\boldsymbol{x}^{*})}}$$

Training loss

Minimize the KL divergence between

- the model and a sharp Dirichlet distribution focused on the appropriate class for in-distribution data
- the model and a flat Dirichlet distribution for out-of-distribution data

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{\mathrm{p}_{\mathrm{in}}(\boldsymbol{x})}[KL[\mathrm{Dir}(\boldsymbol{\mu} \mid \hat{\boldsymbol{\alpha}}) \| \mathrm{p}(\boldsymbol{\mu} \mid \boldsymbol{x}; \boldsymbol{\theta})]] + \mathbb{E}_{\mathrm{pout}(\boldsymbol{x})}[KL[\mathrm{Dir}(\boldsymbol{\mu} \mid \tilde{\boldsymbol{\alpha}}) \| \mathrm{p}(\boldsymbol{\mu} \mid \boldsymbol{x}; \boldsymbol{\theta})]]$$

It is simple to specify a **flat Dirichlet distribution** by setting all $\tilde{\alpha}_c = 1$

The **in-distribution** target $\hat{\alpha}_c$, $\hat{\mu}_c = \frac{\hat{\alpha}_c}{\hat{\alpha}_0}$

$$\hat{\mu}_c = \begin{cases} 1 - (K-1)\epsilon & \text{if } \delta(y = \omega_c) = 1 \\ \epsilon & \text{if } \delta(y = \omega_c) = 0 \end{cases}$$

Measures

Expected predictive categorical $\mathrm{P}(\omega_c \mid \boldsymbol{x}^*; \mathcal{D})$

Max probability: measure of confidence in the prediction

$$\mathcal{P} = \max_c \mathrm{P}(\omega_c \mid \boldsymbol{x}^*; \mathcal{D})$$

Entropy: entropy of the predictive distribution, behaves similar to max probability, represents the uncertainty encapsulated in the entire distribution

$$\mathcal{H}[\mathrm{P}(y \mid \boldsymbol{x}^*; \mathcal{D})] = -\sum_{c=1}^{K} \mathrm{P}(\omega_c \mid \boldsymbol{x}^*; \mathcal{D}) \ln(\mathrm{P}(\omega_c \mid \boldsymbol{x}^*; \mathcal{D}))$$

Measures: MI

Mutual Information (MI) between the categorical label y and the parameters of the model $\boldsymbol{\theta}$ is a measure of the spread of an ensemble $\left\{\mathrm{P}\left(\omega_c \mid \boldsymbol{x}^*, \boldsymbol{\theta}^{(i)}\right)\right\}_{i=1}^{M}$ which assess uncertainty in predictions due to model uncertainty.

Here, MI implicitly captures elements of distributional uncertainty.

$$\underbrace{\mathcal{I}[y,\boldsymbol{\theta}|\boldsymbol{x}^*,\mathcal{D}]}_{Model\ Uncertainty} = \underbrace{\mathcal{H}[\mathbb{E}_{\mathrm{p}(\boldsymbol{\theta}|\mathcal{D})}[\mathrm{P}(y|\boldsymbol{x}^*,\boldsymbol{\theta})]]}_{Total\ Uncertainty} - \underbrace{\mathbb{E}_{\mathrm{p}(\boldsymbol{\theta}|\mathcal{D})}[\mathcal{H}[\mathrm{P}(y|\boldsymbol{x}^*,\boldsymbol{\theta})]]}_{Expected\ Data\ Uncertainty}$$

MI between y and μ, the spread is now explicitly due to distributional uncertainty

$$\underbrace{\mathcal{I}[y,\boldsymbol{\mu} \mid \boldsymbol{x}^*;\mathcal{D}]}_{\text{Distributional Uncertainty}} = \underbrace{\mathcal{H}\left[\mathbb{E}_{\mathrm{p}(\boldsymbol{\mu}|\boldsymbol{x}^*;\mathcal{D})}[\mathrm{P}(y \mid \boldsymbol{\mu})]\right]}_{\text{Total Uncertainty}} - \underbrace{\mathbb{E}_{\mathrm{p}(\boldsymbol{\mu}|\boldsymbol{x}^*;\mathcal{D})}[\mathcal{H}[\mathrm{P}(y \mid \boldsymbol{\mu})]]}_{\text{Expected Data Uncertainty}}$$

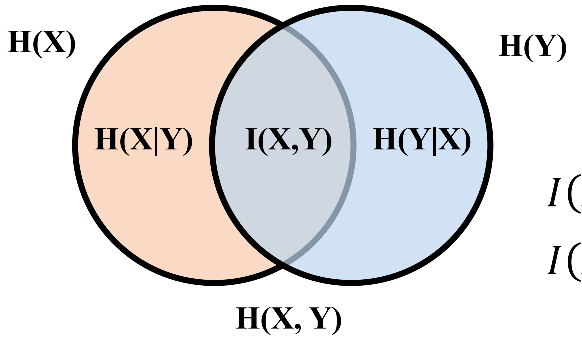
Entropy, MI

$$
\begin{aligned}
\mathcal{I}[y,\boldsymbol{\theta} \mid \boldsymbol{x}^*,\mathcal{D}] &= \mathcal{H}[y \mid \boldsymbol{x}^*,\mathcal{D}] - \mathcal{H}[y \mid \boldsymbol{\theta},\boldsymbol{x}^*] \\
&= -\int P(y \mid \boldsymbol{x}^*,\mathcal{D}) \log P(y \mid \boldsymbol{x}^*,\mathcal{D}) dy + \int\int P(y,\boldsymbol{\theta} \mid \boldsymbol{x}^*,\mathcal{D}) \log P(y \mid \boldsymbol{x}^*,\boldsymbol{\theta}) dy d\boldsymbol{\theta} \\
&= -\int \left(\int P(y \mid \boldsymbol{x}^*,\boldsymbol{\theta}) P(\boldsymbol{\theta} \mid \mathcal{D}) d\boldsymbol{\theta}\right) \log\left(\int P(y \mid \boldsymbol{x}^*,\boldsymbol{\theta}) P(\boldsymbol{\theta} \mid \mathcal{D}) d\boldsymbol{\theta}\right) dy \\
&\quad + \int\int P(y \mid \boldsymbol{x}^*,\boldsymbol{\theta}) P(\boldsymbol{\theta} \mid \mathcal{D}) \log P(y \mid \boldsymbol{x}^*,\boldsymbol{\theta}) dy d\boldsymbol{\theta} \\
&= \mathcal{H}\left[\mathbb{E}_{P(\boldsymbol{\theta}|\mathcal{D})}[y \mid \boldsymbol{x}^*,\boldsymbol{\theta}]\right] + \int P(\boldsymbol{\theta} \mid \mathcal{D}) \left(\int P(y \mid \boldsymbol{x}^*,\boldsymbol{\theta}) \log P(y \mid \boldsymbol{x}^*,\boldsymbol{\theta}) dy\right) d\boldsymbol{\theta} \\
&= \mathcal{H}\left[\mathbb{E}_{P(\boldsymbol{\theta}|\mathcal{D})}[y \mid \boldsymbol{x}^*,\boldsymbol{\theta}]\right] - \mathbb{E}_{P(\boldsymbol{\theta}|\mathcal{D})}[\mathcal{H}[y \mid \boldsymbol{x}^*,\boldsymbol{\theta}]]
\end{aligned}
$$

Model uncertainty

Total uncertainty

Expected Data Uncertainty



$$I(X,Y) = H(X) - H(X|Y)$$
$$I(X,Y) = H(Y) - H(Y|X)$$

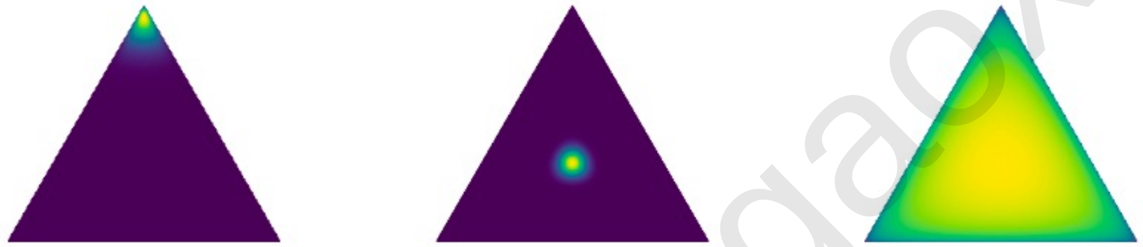
$$
\begin{aligned}
\int\int P(x,y) \log \frac{P(x,y)}{P(x)P(y)} dxdy &= -\int\int P(x,y) \log P(x) dxdy - \int\int P(x,y) \log P(x \mid y) dxdy \\
&= -\int P(x) \log P(x) dx - -\int\int P(x,y) \log P(x \mid y) dxdy \\
&= H(X) - H(X|Y)
\end{aligned}
$$

Measures: the differential entropy

The differential entropy: **maximized when the Dirichlet Distribution is flat**

$$\mathcal{H}[\mathrm{p}(\boldsymbol{\mu}|\boldsymbol{x}^*;\mathcal{D})] = -\int_{\mathcal{S}^{K-1}} \mathrm{p}(\boldsymbol{\mu}|\boldsymbol{x}^*;\mathcal{D}) \ln(\mathrm{p}(\boldsymbol{\mu}|\boldsymbol{x}^*;\mathcal{D}))d\boldsymbol{\mu}$$

Distribution Uncertainty

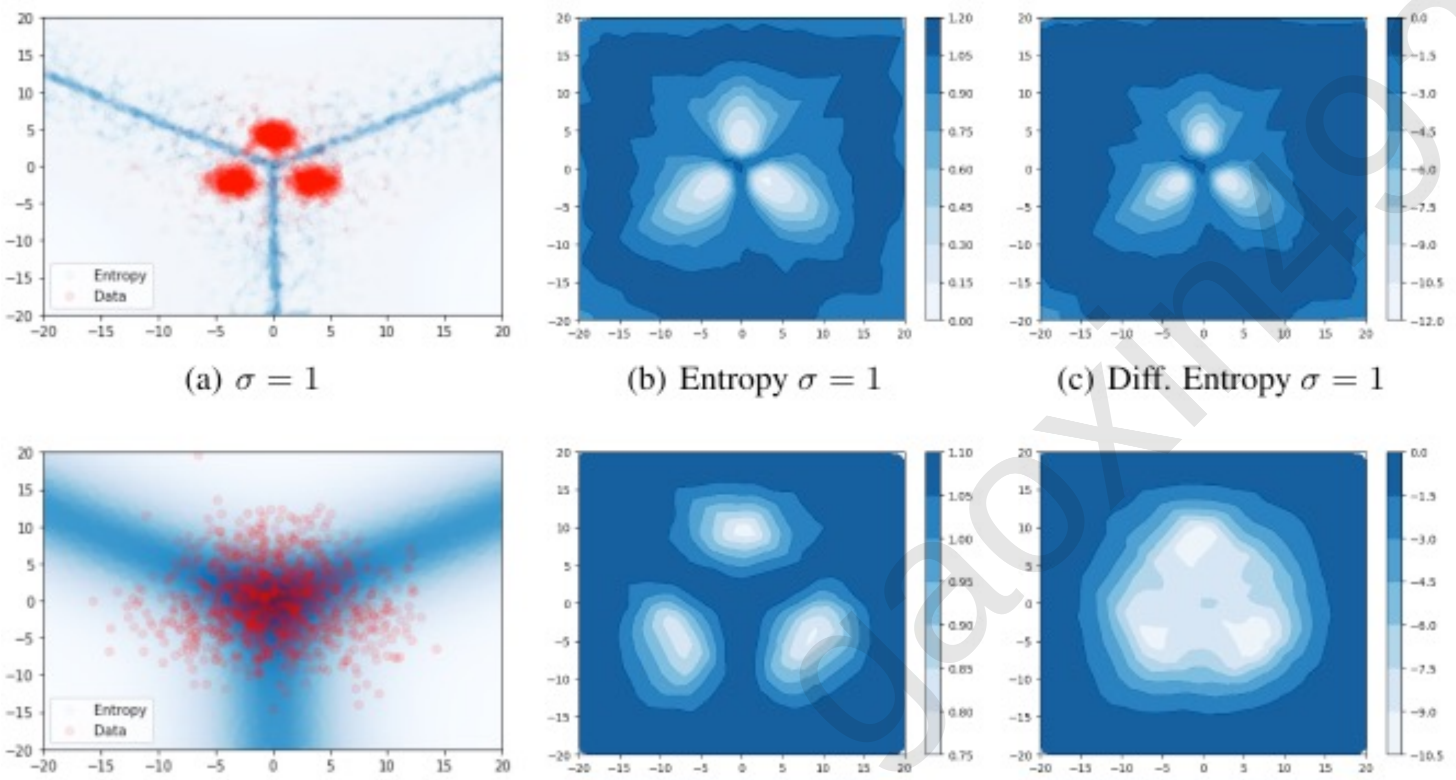


(a) Confident Prediction (b) High data uncertainty (c) Out-of-distribution

Figure 2: Desired behaviors of a distribution over distributions

(a) Sharp distribution, concentrated categorical prediction
(b) Sharp distribution, equiprobable categorical prediction
(c) Flat distribution, equiprobable categorical prediction

Experiments and results



(a) $\sigma = 1$ (b) Entropy $\sigma = 1$ (c) Diff. Entropy $\sigma = 1$

(d) $\sigma = 4$ (e) Entropy $\sigma = 4$ (f) Diff. Entropy $\sigma = 4$

- **class overlap**

Entropy is high both in region of class overlap and far from training data

- difficult to distinguish out-of-distribution samples and in-distribution samples at a decision boundary

Differential entropy is low over the whole region of training data and high outside

- allowing the in-distribution region to be clearly distinguished from the out-of-distribution region

Experiments and results

MNIST and CIFAR-10 are low data uncertainty datasets - all classes are distinct

Differential entropy of the Dirichlet prior will be able to distinguish in-domain and out-of-distribution data better than entropy when the classes are less distinct.

OOD: positive class   ID: negative class

Table 3: MNIST vs OMNIGLOT. Out-of-distribution detection AUROC on noisy data.

	Ent.		M.I.		D.Ent.	
σ	0.0	3.0	0.0	3.0	0.0	3.0
DNN	98.8	58.4	-	-	-	-
MCDP	98.8	58.4	99.3	79.1	-	-
DPN	100.0	51.8	99.5	22.3	100.0	99.8

total   model   distribution

zero mean isotropic Gaussian noise with a standard deviation σ=3 noise

Uncertainty Estimation by Fisher Information-based Evidential Deep Learning

2023 ICML

Xin Gao

2024.1.5

Introduction

- It is not sensitive to arbitrary scaling of α_k classical EDL hinders the learning of evidence, especially for samples with high data uncertainty annotated with the one-hot label.
- We propose a simple and novel method, Fisher Information-based Evidential Deep Learning (I-EDL), to weigh the importance of different classes for each training sample.
- We introduce PAC-Bayesian bound to further improve the generalization ability.
- Our proposed method consistently outperforms traditional EDL-related algorithms in multiple uncertainty estimation tasks, in the confidence evaluation, OOD detection, and few-shot classification.

DUM and EDL

- Dirichlet-based uncertainty models quantify different types of uncertainty by modeling the output as the concentration parameters of a Dirichlet distribution.
- Evidential deep learning (EDL) adopts Dirichlet distribution and treats the output as evidence to quantify belief mass and uncertainty by jointly considering the Dempster–Shafer Theory of Evidence (DST) and subjective logic (SL).

State space: K mutually exclusive singletons (e.g., class labels)

=> belief mass, uncertainty mass $u + \sum_{k=1}^{K} b_k = 1$

=> Dirichlet prior, evidence $\mathrm{Dir}(\boldsymbol{p} \mid \boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} p_k^{\alpha_k - 1}, \alpha_0 = \sum_{k=1}^{K} \alpha_k$ $\quad \alpha_k = e_k + 1$

=> assign belief and uncertainty $b_k = \frac{\alpha_k - 1}{\alpha_0}, \quad u = \frac{K}{\alpha_0}$

=> point-estimated categorical prediction $\hat{p}_k = \frac{\alpha_k}{\alpha_0} = \frac{e_k + 1}{\sum_{c=1}^{K} e_c + K}$

Graphic Representation

- EDL supposes the observed labels y were drawn i.i.d. from an isotropic Gaussian distribution, i.e.

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

where $p \sim Dir(f_\theta(x) + 1)$.

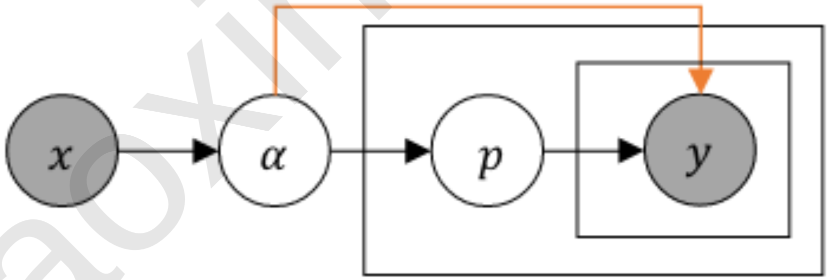
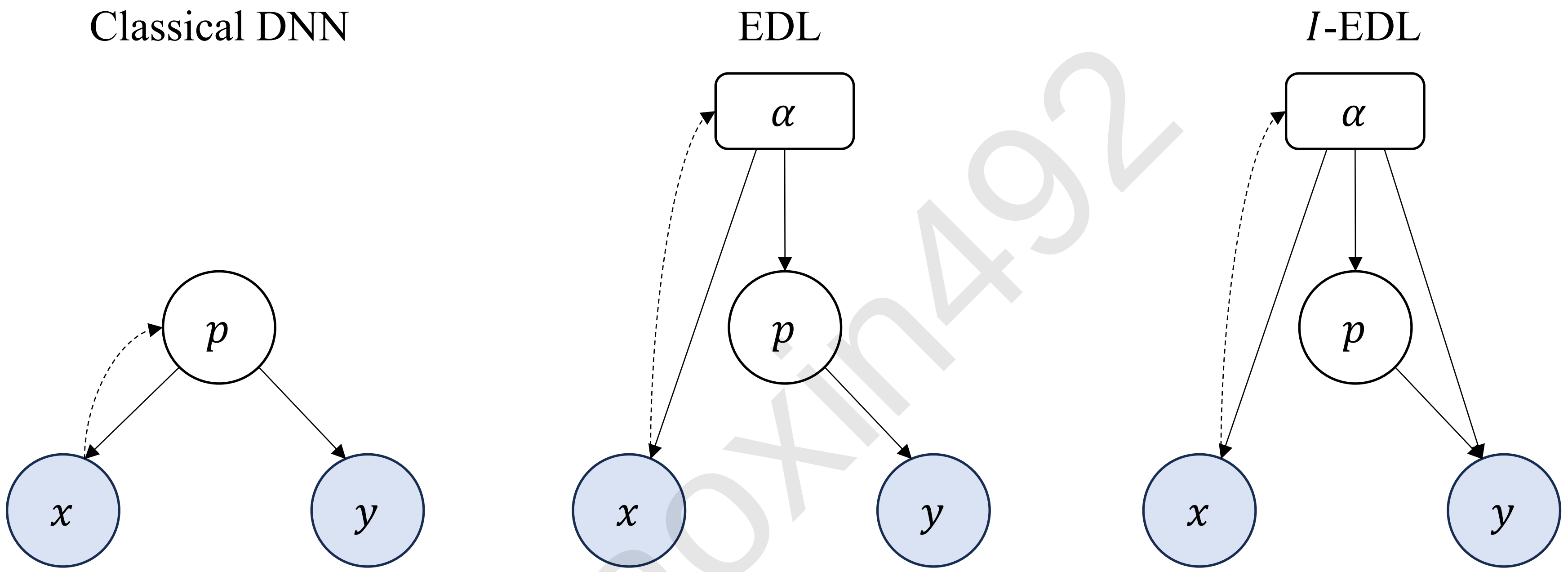


Figure 2. Graphical model representation of $\mathcal{I}$-EDL.

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.

Graphic Representation



x: Observed images

y: Observed labels

p: Probability map

α: Parameter of Dirichlet distribution

Solid arrows indicate generation while dashed ones refer to inference procedure from a neural network.

Higher evidence & Higher variance

- EDL supposes the observed labels y were drawn i.i.d. from an isotropic Gaussian distribution, i.e.

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

where $p \sim Dir(f_\theta(x) + 1)$.

$$-\log \mathcal{N}(\boldsymbol{y}_i \mid \boldsymbol{p}_i, \boldsymbol{\Sigma}) = \frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{p}_i)^T \boldsymbol{\Sigma}(\boldsymbol{y}_i - \boldsymbol{p}_i) + \frac{1}{2}\log|\boldsymbol{\Sigma}| + const$$

- The information of each class carried in categorical probabilities p is different, thus the generation of each class for a specific sample should not be isotropic.

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

Fisher information matrix

- The Fisher information is a way of measuring the amount of information that an observable random variable X carries about an unknown parameter θ of a distribution that models X.

$$x_1, x_2, \dots, x_n \rightarrow p(x; \theta)$$

- To assess the goodness of our estimate of θ we define a score function $\boldsymbol{s}(\theta) = \nabla_\theta \log p(x \mid \theta)$
- The expected value of score wrt. our model is zero

$$\underset{p(x|\theta)}{\mathbb{E}}[\boldsymbol{s}(\theta)] = \underset{p(x|\theta)}{\mathbb{E}}[\nabla \log p(x \mid \theta)] = 0$$

- The covariance of score function above is the definition of **Fisher Information Matrix**

$$\mathrm{F} = \underset{p(x|\theta)}{\mathbb{E}}\left[(\boldsymbol{s}(\theta) - 0)(\boldsymbol{s}(\theta) - 0)^T)\right] = \underset{p(x|\theta)}{\mathbb{E}}\left[\nabla \log p(x \mid \theta) \nabla \log p(x \mid \theta)^{\mathrm{T}}\right]$$

- The negative expected Hessian of log likelihood is equal to the Fisher Information Matrix F

$$\mathbf{F} = -\mathbb{E}_{p(x|\theta)}\left[\mathbf{H}_{\log p(x|\theta)}\right]$$

Insights about FIM

In our context, **the Fisher information matrix (FIM)** is chosen to measure the amount of information that the categorical probabilities p carry about the concentration parameters α of a Dirichlet distribution that models p.

$$\ell = \log \mathrm{Dir}(\boldsymbol{p} \mid \boldsymbol{\alpha})$$

$$\mathcal{I}(\boldsymbol{\alpha}) = \mathbb{E}_{\mathrm{Dir}(\boldsymbol{p}|\boldsymbol{\alpha})}\left[\frac{\partial \ell}{\partial \boldsymbol{\alpha}} \frac{\partial \ell}{\partial \boldsymbol{\alpha}^T}\right] = \mathbb{E}_{\mathrm{Dir}(\boldsymbol{p}|\boldsymbol{\alpha})}\left[-\frac{\partial^2 \ell}{\partial \boldsymbol{\alpha} \boldsymbol{\alpha}^T}\right]$$

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

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

$\alpha_k < \alpha_0$, trigamma function is a monotonically decreasing function when $x > 0$

MLE

- In MLE, we can learn model parameters θ by minimizing the expected negative log-likelihood loss function:

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

- General loss can improve generalization but is intractable $(x, y) \sim P$
- We can find an upper bound of this optimization problem, converting general loss into empirical loss.

PAC-Bayesian Bound

- This theory focuses on the upper bound of the probability of generalization error for a model output by a learning algorithm, given a certain data distribution.

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 Θ, 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,

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

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

$RV:\ \theta \Rightarrow p$

$p \sim Dir(p|\alpha)$

$\alpha = f_\theta(x) + 1$

1. **Prior Distribution,** π: The distribution over the hypothesis set before observing any data. It reflects our initial beliefs about the parameters.
2. **Posterior,** ρ: After observing data, our beliefs about the hypothesis set are updated, leading to the posterior distribution.

Upper Bound

- In this paper, we treat $Dir(p|\alpha)$ as the posterior distribution, and the prior as $Dir(p|\mu)$, where μ is set to $\beta \gg 1$ for the corresponding class and 1 for all other class.
- The upper bound of the optimization problem in MLE can be expressed as

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

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

- The first term is the expected FIM-weighted MSE subtract the negative log determinant of the FIM: $\mathcal{L}_i(\boldsymbol{\theta}) \propto \underbrace{\mathbb{E}\left[(\boldsymbol{y}_i - \boldsymbol{p}_i)^T\mathcal{I}(\boldsymbol{\alpha}_i)(\boldsymbol{y}_i - \boldsymbol{p}_i)\right]}_{\mathcal{L}_i^{\mathcal{I}\text{-MSE}}} - \sigma^2\underbrace{\log|\mathcal{I}(\boldsymbol{\alpha}_i)|}_{\mathcal{L}_i^{|\mathcal{I}|}}$
- The second term can be simplified by setting $\hat{\boldsymbol{\alpha}}_i = \boldsymbol{\alpha}_i \odot (1 - \boldsymbol{y}_i) + \boldsymbol{y}_i$ as Sensoy et al.

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

MLE & MSE & Cross-entropy

$$\max \log \mathcal{P}(y;\theta) = \max \sum_{i=1}^{n} \log \mathcal{P}(y_i;\theta)$$

- Gaussian

$$\sum_{i=1}^{n} \log \mathcal{N}(y_i \mid f_\theta(x_i), \Sigma) = -\frac{1}{2}\sum_{i=1}^{n}(y_i - f_\theta(x_i))^T \Sigma (y_i - f_\theta(x_i)) - \frac{n}{2}\log|\Sigma| + const$$

$$\sum_{i=1}^{n} \log \mathcal{N}(y_i \mid f_\theta(x_i), I) = -\frac{1}{2}\sum_{i=1}^{n}(y_i - f_\theta(x_i))^T (y_i - f_\theta(x_i)) + const$$ ⇨ MSE

- Bernoulli

$$y \sim B(y, f_\theta(x))$$

$$\sum_{i=1}^{n} \log [f_\theta(x_i)]^{y_i}[1 - f_\theta(x_i)]^{(1-y_i)} = \sum_{i=1}^{n} y_i \log f_\theta(x_i) + (1-y_i)\log(1 - f_\theta(x_i))$$

⇨ Cross entropy

Objective function

- Finally, the objective function Eq.(2) can be reformulated as

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

classical EDL can be viewed as a degenerate version of I-EDL

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

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

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

Table 1. Difference between $\mathcal{I}$-EDL and EDL are marked in blue.

	EDL	**$\mathcal{I}$-EDL**
MSE	$\sum_{i=1}^{K}(y_i - \frac{\alpha_i}{\alpha_0})^2 + \sum_{i=1}^{K}\frac{\alpha_i(\alpha_0-\alpha_i)}{\alpha_0^2(\alpha_0+1)}$	$\sum_{i=1}^{K}(y_i - \frac{\alpha_i}{\alpha_0})^2\psi^{(1)}(\alpha_i) + \sum_{i=1}^{K}\frac{\alpha_i(\alpha_0-\alpha_i)}{\alpha_0^2(\alpha_0+1)}\psi^{(1)}(\alpha_i)$
KL	$D_{\text{KL}}(Dir(\hat{\boldsymbol{\alpha}})\Vert Dir(\mathbf{1}))$	$D_{\text{KL}}(Dir(\hat{\boldsymbol{\alpha}})\Vert Dir(\mathbf{1}))$
$\mathcal{I}$	-	$-\log\lvert\mathcal{I}(\boldsymbol{\alpha})\rvert$

- For different labels in a sample

Though it has been correctly classified for a specific label, it still allows for more evidence for the overlapping labels.

Objective function

Uncertainty $\frac{K}{\alpha_0}$

```python
def compute_fisher_mse(self, labels_1hot_, evi_alp_):
    evi_alp0_ = torch.sum(evi_alp_, dim=-1, keepdim=True)

    gamma1_alp = torch.polygamma(1, evi_alp_)
    gamma1_alp0 = torch.polygamma(1, evi_alp0_)

    gap = labels_1hot_ - evi_alp_ / evi_alp0_

    loss_mse_ = (gap.pow(2) * gamma1_alp).sum(-1).mean()

    loss_var_ = (evi_alp_ * (evi_alp0_ - evi_alp_) * gamma1_alp / (evi_alp0_ * evi_alp0_ * (evi_alp0_ + 1))).sum(-1).mean()

    loss_det_fisher_ = - (torch.log(gamma1_alp).sum(-1) + torch.log(1.0 - (gamma1_alp0 / gamma1_alp).sum(-1))).mean()

    return loss_mse_, loss_var_, loss_det_fisher_
```

Table 1. Difference between $\mathcal{I}$-EDL and EDL are marked in blue.

	EDL	**$\mathcal{I}$-EDL**
MSE	$\sum_{i=1}^{K}(y_i - \frac{\alpha_i}{\alpha_0})^2 + \sum_{i=1}^{K} \frac{\alpha_i(\alpha_0-\alpha_i)}{\alpha_0^2(\alpha_0+1)}$	$\sum_{i=1}^{K}(y_i - \frac{\alpha_i}{\alpha_0})^2 \psi^{(1)}(\alpha_i) + \sum_{i=1}^{K} \frac{\alpha_i(\alpha_0-\alpha_i)}{\alpha_0^2(\alpha_0+1)} \psi^{(1)}(\alpha_i)$
KL	$D_{\mathrm{KL}}(Dir(\hat{\boldsymbol{\alpha}}) \Vert Dir(\mathbf{1}))$	$D_{\mathrm{KL}}(Dir(\hat{\boldsymbol{\alpha}}) \Vert Dir(\mathbf{1}))$
$\mathcal{I}$	-	$-\log \lvert\mathcal{I}(\boldsymbol{\alpha})\rvert$

https://github.com/danruod/IEDL

Experiments

- OOD detection

Table 3. AUPR scores of OOD detection (mean ± 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 → KMNIST†		**MNIST → FMNIST**†		**CIFAR10 → SVHN**†		**CIFAR10 → CIFAR100**	
Method	**Max.P**	α_0	**Max.P**	α_0	**Max.P**	α_0	**Max.P**	α_0
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	98.44 ± 0.1	98.16 ± 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	80.21 ± 0.2	77.71 ± 0.3	81.96 ± 0.8	82.06 ± 0.8
EDL	97.02 ± 0.8	96.31 ± 2.0	98.10 ± 0.4	98.08 ± 0.4	78.87 ± 3.5	79.12 ± 3.7	84.30 ± 0.7	84.18 ± 0.7
$\mathcal{I}$**-EDL**	**98.34 ± 0.2**	**98.33 ± 0.2**	**98.89 ± 0.3**	**98.86 ± 0.3**	**83.26 ± 2.4**	**82.96 ± 2.2**	**85.35 ± 0.7**	**84.84 ± 0.6**

We mainly focus on the comparisons with DBU models, which solve OOD detection by distinguishing different types of uncertainty.

Experiments

- Few-shot Learning

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**

Our method not only improves classification accuracy but also greatly improves the availability of uncertainty estimation in the more challenging few-shot scenarios.

Experiments

- Density plots of the predicted differential entropy and mutual information (Last paper, distributional uncertainty)
- Lower entropy or mutual information represents the model yields a sharper distribution, indicating that the sample has low uncertainty.
- Our method provides more separable uncertainty estimates, I-EDL produces sharper prediction peaks than EDL

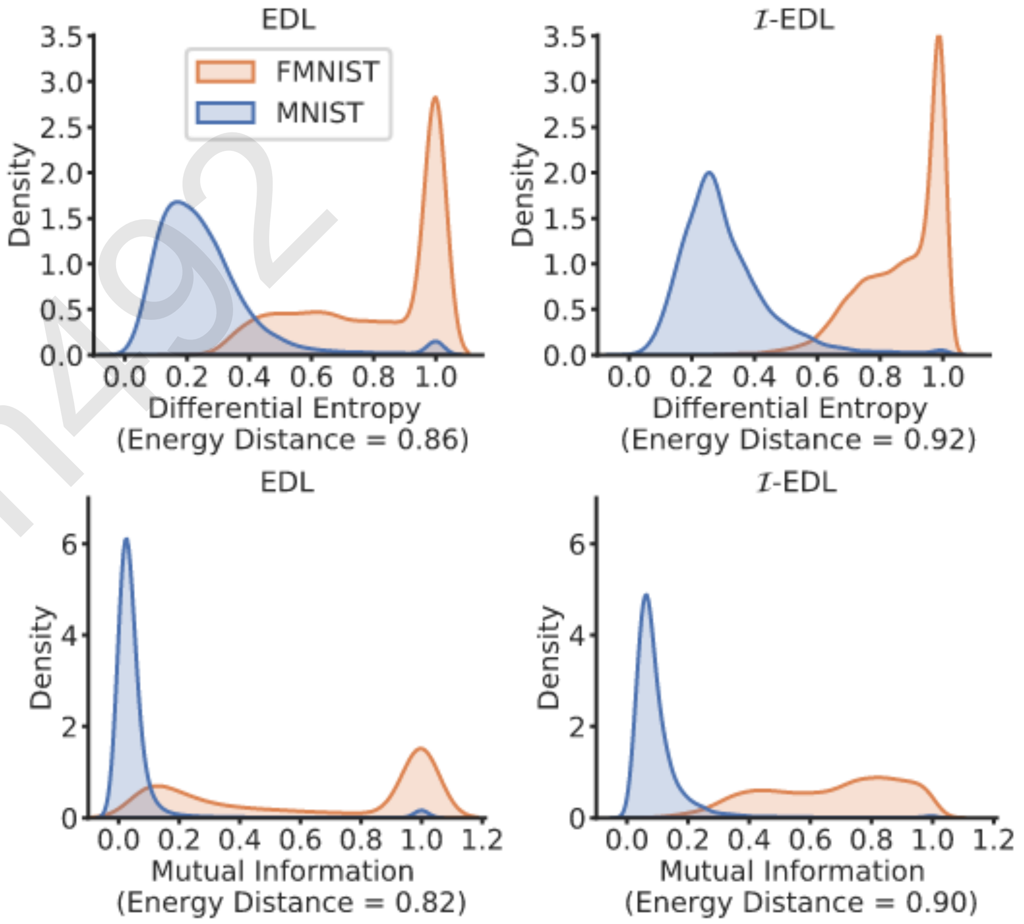


Figure 4. Uncertainty representation for ID (MNIST) and OOD (FMNIST). More results are shown in Appendix C.6.

Conclusion

- The observed label is jointly generated by the predicted categorical probability and the informativeness of each class contained in the sample.
- The informativeness is modeled by the uncertainty of the estimator of α (FIM), naturally including data uncertainty.