Are Uncertainty Quantification Capabilities of Evidential Deep Learning a Mirage?

Maohao Shen, J. Jon Ryu, Soumya Ghosh, Yuheng Bu, Prasanna Sattigeri, Subhro Das, Gregory W. Wornell

MIT, IBM Research and University of Florida

November 15th, 2024

Presented by Boyao Li

Recall: Evidential Deep Learning

- Place a Dirichlet distribution on the output

$$D(\mathbf{p} \mid \boldsymbol{\alpha}) = \begin{cases} \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{K} p_i^{\alpha_i - 1} & \text{for } \mathbf{p} \in \mathcal{S}_K, \\ 0 & \text{otherwise,} \end{cases}$$

- Minimize the MSE loss function with regularizing terms.

$$\mathcal{L}(\Theta) = \sum_{i=1}^{N} \mathcal{L}_i(\Theta) + \lambda_t \sum_{i=1}^{N} \mathrm{KL}\left[D(\mathbf{p}_i \mid \tilde{\boldsymbol{\alpha}}_i) \parallel D\left(\mathbf{p}_i \mid \langle 1, \ldots, 1 \rangle\right)\right]$$

$$\mathcal{L}_i(\Theta) = \int \|\mathbf{y}_i - \mathbf{p}_i\|_2^2 \frac{1}{B(\boldsymbol{\alpha}_i)} \prod_{j=1}^{K} p_{ij}^{\alpha_{ij}-1} \, \mathrm{d}\mathbf{p}_i$$

~~Introduction~~ Conclusion

- Uncertainty learned by most of the EDL methods has **no** statistical meaning.
- Problems with EDL arise from ignoring the **model uncertainty** for computational effiency.
- The authors suggest that EDL methods can be better interpreted as energy-based out-of-distribution (OOD) detection algorithms.

Recent work has reported EDL limitations

- Bengs et al. (NeurIPS 2022): non-vanishing distributional uncertainty
- Bengs et al. (ICML 2023): a possibility of non-existence of proper scoring rules for meta distributions
- Jürgens et al. (ICML 2024): a gap between learned uncertainty and an ideal meta distribution
- What is proposed by the authors in this paper (NeurIPS 2024)?

Introduction

This paper:

- Unifies various objective functions (loss function to minimize) of a wide class of EDL methods.
- Provides empirical evidence to point out the fundamental limitations of the learned uncertainties by EDL.
- Presents several findings showing that existing EDL methods are essentially OOD detectors.

Problem Setting

- Aim to learn $p(y \mid x)$, given data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ drawn from underlying $p(x, y)$ over $\mathcal{X} \times \mathcal{Y}$.
  - $\mathcal{Y} = [C] := \{1, 2, ..., C\}$ for classification, $\mathcal{Y} \in \mathbb{R}$ for regression.
  - Here we mainly discuss classification settings.
- With a trained parametric classifier $p(y|x, \psi)$, the predictive posterior distribution is given by:

$$p(y \mid x, \mathcal{D}) := \int p(y \mid x, \psi) p(\psi \mid \mathcal{D}) \, \mathrm{d}\psi$$

$p(y \mid x, \psi)$ captures *aleatoric uncertainty*, while $p(\psi \mid \mathcal{D})$ describes *epistemic uncertainty*.

Problem Setting

For EDL approach:

- Decompose $p(y \mid x, \psi) = \int p(\boldsymbol{\pi} \mid x, \psi) p(y \mid \boldsymbol{\pi}) \, \mathrm{d}\boldsymbol{\pi}$, where $\pi \in \Delta^{C-1}$ is a probability vector.
  - $p(\boldsymbol{\pi} \mid x, \psi)$ is a *meta distribution* over predictions at x. $p(y \mid \boldsymbol{\pi}) = \pi_y$ is a fixed likelihood model.
  - Usually $p(\boldsymbol{\pi} \mid x, \psi)$ is a conjugate prior (in the same probability distributions family) of $p(y \mid \boldsymbol{\pi})$, but sometimes not.
- The full decomposition:

$$p(y \mid x, \mathcal{D}) = \iint \underbrace{p(y \mid \boldsymbol{\pi})}_{\text{aleatoric}} \underbrace{p(\boldsymbol{\pi} \mid x, \psi)}_{\textit{distributional}} \underbrace{p(\psi \mid \mathcal{D})}_{\text{model}} \, \mathrm{d}\psi \, \mathrm{d}\boldsymbol{\pi}$$

Problem Setting

Consider marginalizing out ψ:

$$p(y \mid x, \mathcal{D}) = \iint p(y \mid \boldsymbol{\pi}) p(\boldsymbol{\pi} \mid x, \mathcal{D}) \mathrm{d}\boldsymbol{\pi}$$

EDL assumes the best single model ψ^* learned with data $\mathcal{D}$, without any randomness in $p(\psi \mid \mathcal{D})$:

$$p(\psi \mid \mathcal{D}) = \delta(\psi - \psi^*) \Longrightarrow p(\boldsymbol{\pi} \mid x, \mathcal{D}) \approx p(\boldsymbol{\pi} \mid x, \psi^*)$$

This simplification allows its computational efficiency but leads to fake distributional uncertainty.

For the following formulas, we use $p_\psi(\boldsymbol{\pi} \mid x)$ instead of $p(\boldsymbol{\pi} \mid x, \psi)$ since a single model ψ is assumed in this context.

EDL Taxonomy

Criterion 1. Parametric Form of Meta Distribution

The parametric form of $\alpha_\psi(x)$ in Dirichlet distribution $p_\psi(\boldsymbol{\pi} \mid x) = \mathrm{Dir}(\boldsymbol{\pi}; \boldsymbol{\alpha}_\psi(x))$:

- Direct parametrization: parameterize $\boldsymbol{\alpha}_\psi(x))$ by a direct output of a neural network. But it can arbitrary values on the OOD data points.
- Density parametrization: $\boldsymbol{\alpha}_\psi(x) \leftarrow \boldsymbol{\alpha}_0 + \boldsymbol{N}_\psi(x)$, where for $y \in [C]$, $(\boldsymbol{N}_\psi(x))_y := N_y p_{\psi_2}(f_{\psi_1}(x) \mid y)$. $f_{\psi_1}(x)$ is a feature extractor, and $p_{\psi_2}(z \mid y)$ is a tractable density model such as normalizing flows.

EDL Taxonomy

Criterion 2. Objective Function Function to minimize for representative EDL methods:

- Prior Networks (FPriorNet/RPriorNet): Given $\nu \gg 1$, $\boldsymbol{\alpha}_0 = \mathbf{1}_C$ and $\boldsymbol{e}_y$ as the one-hot true label, minimize

$$\mathbb{E}_{p(x,y)}\left[D\left(\mathrm{Dir}(\boldsymbol{\pi};\boldsymbol{\alpha}_0+\nu\boldsymbol{e}_y), p_\psi(\boldsymbol{\pi}\mid x)\right)\right]+\gamma_{\text{ood}}\mathbb{E}_{p_{\text{ood}}(x)}\left[D\left(\mathrm{Dir}(\boldsymbol{\pi};\boldsymbol{\alpha}_0), p_\psi(\boldsymbol{\pi}\mid x)\right)\right]$$

  Here $D(p,q) = D(p \parallel q)$ in FPriorNet, and $D(p,q) = D(q \parallel p)$ in RPriorNet.

- EDL: minimize the MSE loss with a reverse KL regularizer

$$\ell_{\text{MSE}}(\psi;x,y) := \mathbb{E}_{p_\psi(\boldsymbol{\pi}|x)}\left[\|\boldsymbol{\pi}-\boldsymbol{e}_y\|^2\right]+\lambda D\left(p_\psi(\boldsymbol{\pi}\mid x) \parallel \mathrm{Dir}(\boldsymbol{\pi};\boldsymbol{\alpha}_0)\right).$$

EDL Taxonomy

- Belief Matching: minimize VI loss justified by variational inference framework

$$\ell_{\mathrm{VI}}(\psi; x, y) := \mathbb{E}_{p_\psi(\boldsymbol{\pi}|x)}\left[\log \frac{1}{\pi_y}\right] + \lambda D\left(p_\psi(\boldsymbol{\pi} \mid x) \,\|\, \mathrm{Dir}(\boldsymbol{\pi}; \boldsymbol{\alpha}_0)\right).$$

- Posterior Networks (PostNet and NatPN): minimize the uncertainty-aware cross entropy (UCE) loss

$$\ell_{\mathrm{UCE}}(\psi; x, y) := \mathbb{E}_{p_\psi(\boldsymbol{\pi}|x)}\left[\log \frac{1}{\pi_y}\right] - \lambda h\left(p_\psi(\boldsymbol{\pi} \mid x)\right).$$

Unifying EDL Objectives for Classification

Define the *tempered likelihood*: for $\nu > 0$, define

$$p^{(\nu)}(\boldsymbol{\pi} \mid y) := \frac{p^{(\nu)}(\boldsymbol{\pi}, y)}{\int p^{(\nu)}(\boldsymbol{\pi}, y)\, \mathrm{d}\boldsymbol{\pi}}, \quad \text{where} \quad p^{(\nu)}(\boldsymbol{\pi}, y) := \frac{p(\boldsymbol{\pi}) p^{\nu}(y \mid \boldsymbol{\pi})}{\int p(\boldsymbol{\pi}) \sum_y p^{\nu}(y \mid \boldsymbol{\pi})\, \mathrm{d}\boldsymbol{\pi}}.$$

- The prior distribution $p(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi}; \boldsymbol{\alpha}_0)$ with $\alpha_0 = \mathbf{1}_C$.
- The likelihood model of Prior/Post Net and BM is categorical: $p(y \mid \boldsymbol{\pi}) = \pi_y$. It is easy to check $p^{(\nu)}(\boldsymbol{\pi} \mid y) = \mathrm{Dir}(\boldsymbol{\pi}; \boldsymbol{\alpha}_0 + \nu \boldsymbol{e}_y)$.
- The likelihood model of EDL is Gaussian: $p(y \mid \boldsymbol{\pi}) = \mathcal{N}(\boldsymbol{e}_y; \boldsymbol{\pi}, \sigma^2 \boldsymbol{I}_C)$, which does not admit a closed form expression for $p^{(\nu)}(\boldsymbol{\pi} \mid y)$.

Unifying EDL Objectives for Classification

Introduce a *unified objective function*:

$$\mathcal{L}(\psi) := \mathbb{E}_{p(x,y)}\left[D\left(p^{(\nu)}(\boldsymbol{\pi} \mid y), p_\psi(\boldsymbol{\pi} \mid x)\right)\right] + \gamma_{\text{ood}}\mathbb{E}_{p_{\text{ood}}(x)}\left[D\left(p(\boldsymbol{\pi}), p_\psi(\boldsymbol{\pi} \mid x)\right)\right]$$

for some divergence function $D(\cdot,\cdot)$, a tempering parameter $\nu > 0$, and an OOD regularization parameter $\gamma_{\text{ood}} \geq 0$ with a distribution p_{ood} for OOD samples.

The paper proves that under certain settings, the unified objective function is equivalent to different EDL objective functions.

"Optimal" Meta Distribution

Here we focus on the reverse-KL type EDL methods.

Theorem

For any prior $p(\boldsymbol{\pi})$ and likelihood $p(y \mid \boldsymbol{\pi})$, we have

$$\min_{\psi} \mathbb{E}_{p(x,y)} \left[D\left(p_\psi(\boldsymbol{\pi} \mid x) \parallel p_\nu(\boldsymbol{\pi} \mid y) \right) \right] \equiv \min_{\psi} \mathbb{E}_{p(x)} \left[D\left(p_\psi(\boldsymbol{\pi} \mid x) \parallel p^\star(\boldsymbol{\pi} \mid x) \right) \right]$$

where $p^\star(\boldsymbol{\pi} \mid x) := \dfrac{p(\boldsymbol{\pi}) \exp\left(\nu \mathbb{E}_{p(y|x)}[\log p(y|\boldsymbol{\pi})]\right)}{\int p(\boldsymbol{\pi}) \exp\left(\nu \mathbb{E}_{p(y|x)}[\log p(y|\boldsymbol{\pi})]\right) \mathrm{d}\boldsymbol{\pi}}$.

When the model meta distribution $p_\psi(\boldsymbol{\pi} \mid x)$ is trained with the reverse-KL objective, it is forced to fit a **fixed** target meta distribution $p^\star(\boldsymbol{\pi} \mid x)$.

Example: Categorical Likelihood

Consider $p(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi}; \boldsymbol{\alpha}_0)$ and $p(y \mid \boldsymbol{\pi}) = \pi_y$, we have $p^\star(\boldsymbol{\pi} \mid x) = \mathrm{Dir}\left(\boldsymbol{\pi}; \boldsymbol{\alpha}_0 + \nu\boldsymbol{\eta}(x)\right)$, where $\boldsymbol{\eta}(x) := \mathbb{E}_{p(y|x)}[\boldsymbol{e}_y] = [p(1 \mid x), \ldots, p(C \mid x)]$ denotes the true label distribution.

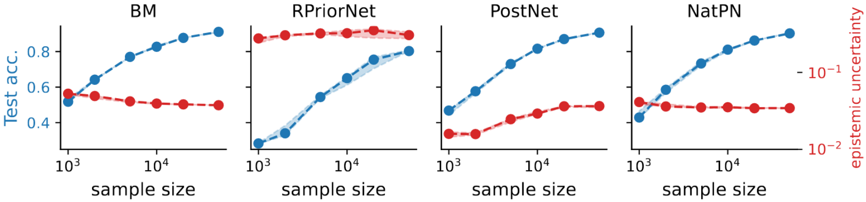
Theorem implies that

$$\min_{\psi} \mathbb{E}_{p(x,y)} \left[D\left(p_\psi(\boldsymbol{\pi} \mid x) \parallel p_\nu(\boldsymbol{\pi} \mid y) \right) \right] \equiv \min_{\psi} \mathbb{E}_{p(x)} \left[D\left(\mathrm{Dir}(\boldsymbol{\pi}; \boldsymbol{\alpha}_\psi(x)) \parallel \mathrm{Dir}(\boldsymbol{\pi}; \boldsymbol{\alpha}_0 + \nu\boldsymbol{\eta}(x)) \right) \right].$$

This shows that $\boldsymbol{\alpha}_\psi(x)$ is forced to match $\boldsymbol{\alpha}_0 + \nu\boldsymbol{\eta}(x)$ as *fixed* target.

EDL Methods Learn False Epistemic Uncertainty

Based on the theorem, even with infinite data, the learned "distributional uncertainty" would remain constant for ID data.

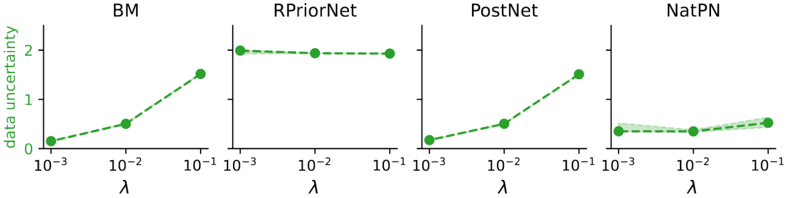


(a) Epistemic Uncertainty: CIFAR10

Figure: Epistemic uncertainty does not vanish with an increasing number of observed samples.

EDL Methods Learn False Aleatoric Uncertainty

EDL methods quantify aleatoric uncertainty as $\mathbb{E}_{p_\psi(\boldsymbol{\pi}|x)}\left[H\left(p(y \mid \boldsymbol{\pi})\right)\right]$. In classification example, the optimal meta distribution is $p_{\psi^\star} = \mathrm{Dir}(\boldsymbol{\pi}; \alpha_0 + \nu\boldsymbol{\eta}(x))$, suggesting that the aleatoric uncertainty would depend on the hyper-parameter $\lambda = \nu^{-1}$, which should be a fixed constant from the underlying label distribution $p(y \mid x)$.



(b) Aleatoric Uncertainty: CIFAR10

Figure: EDL methods learn model-dependent aleatoric uncertainty that depends on hyper-parameter λ.

EDL Methods Are EBM-Based OOD Detector

For EDL methods with Dirichlet prior and categorical model, the induced model predictive distribution is $p_\psi(y \mid x) = \mathbb{E}_{p_\psi(\boldsymbol{\pi}|x)}[p(y \mid \boldsymbol{\pi})] = \frac{\alpha_{\psi,y}(x)}{\mathbf{1}_C^\top \boldsymbol{\alpha}_\psi(x)}$.

In the OOD detection literature, there exists an energy-based model(EBM) based algorithm. Consider a standard classifier with exponentiated logits $\beta_\phi(x)$, whose prediction is given as $p_\phi(y \mid x) := \frac{\beta_{\phi,y}(x)}{\mathbf{1}_C^\top \boldsymbol{\beta}_\phi(x)}$.

The algorithm relates the denominator to a free energy $E_\phi(x) := -\log \mathbf{1}_C^\top \boldsymbol{\beta}_\phi(x)$. The model is trained to minimize

$$-\mathbb{E}_{p(x,y)}[\log p_\psi(y \mid x)] + \tau \left\{ \mathbb{E}_{p(x)}\left[\max(0, E_\phi(x) - m_{\text{id}})^2\right] + \mathbb{E}_{p_{\text{o.o.d}}(x)}\left[\max(0, m_{\text{ood}} - E_\phi(x))^2\right] \right\}$$

This reveals that this EBM-based OOD framework has an almost identical learning mechanism to the EDL methods.

EDL Methods Prefer Smaller λ

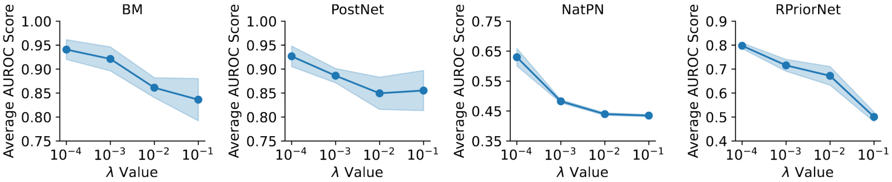


Figure: OOD Detection Performance v.s. λ on CIFAR10

The EDL model is encouraged to fit its output to a large target, so that the summation of the output is large for ID data, and small for OOD data.

Recommendations

It is worth reading, but as a **review** paper. Need to first read some papers about different EDL methods and limitations of them.