Credal Deep Ensembles for Uncertainty Quantification

Machine Learning in Practice Reading Group

Duke B&B

December 5, 2025

Presented by Yuqi Li

Wang et al., NeurIPS 2024

Section 1: Introduction

Why Uncertainty Quantification Matters in Classification

Problem with Standard Neural Networks (SNNs)

- SNNs output a **single probability distribution** over classes, hiding how reliable the prediction is
- Cannot distinguish between different sources of uncertainty

Two Types of Uncertainty

- **Aleatory Uncertainty (AU)**: Inherent randomness in data — *irreducible*
- **Epistemic Uncertainty (EU)**: Lack of knowledge — *reducible*

Why does this matter?

- Model should signal when their predictions may be unreliable
- Out-of-Distribution (OOD) detection

Section 2: Background

Core Intuition: From Points to Intervals

Key Idea

- Standard networks give point estimates: "$P(\text{cat}) = 0.8$"
- Credal models give intervals: "$P(\text{cat}) \in [0.7, 0.9]$"
- **Wider intervals** suggest higher epistemic uncertainty

Deep Ensembles (DEs) [Lakshminarayanan et al., 2017]

- Train M neural networks independently with different random seeds
- Final prediction: Average the probability distributions

$$\bar{\boldsymbol{q}} = \frac{1}{M} \sum_{m=1}^{M} \boldsymbol{q}_m$$

From Ensembles of Points to Ensembles of Intervals

- Deep Ensembles: average point predictions from multiple SNNs
- Credal Deep Ensembles: aggregate probability intervals from multiple CreNets

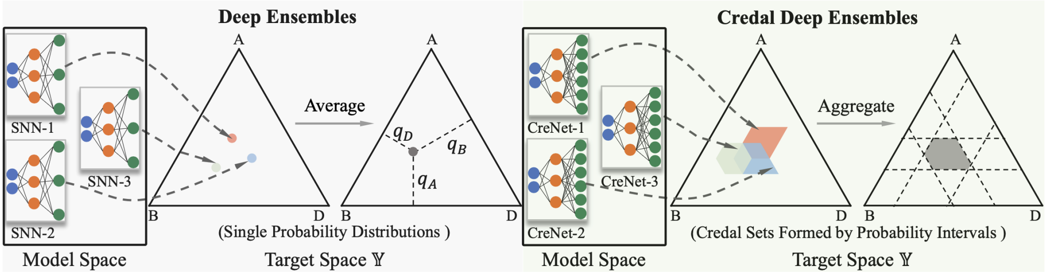

Deep Ensembles
SNN-1
SNN-2
SNN-3
Model Space
A
B
D
Average
q_D
q_B
q_A
(Single Probability Distributions)
Target Space $\mathbb{Y}$
Credal Deep Ensembles
CreNet-1
CreNet-2
CreNet-3
Model Space
Aggregate
(Credal Sets Formed by Probability Intervals)
Target Space $\mathbb{Y}$

Figure 1: Deep Ensembles vs Credal Deep Ensembles

Section 2: Background

Deep Ensembles vs Credal Deep Ensembles

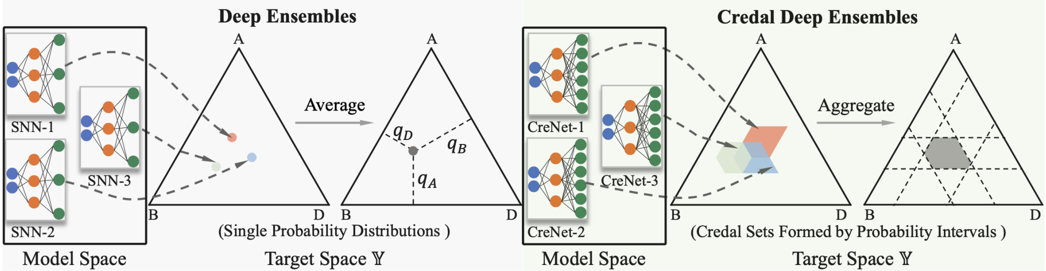


Figure 1 Explanation:

- **Left (Deep Ensembles)**: Each SNN outputs a point on the probability simplex; final prediction is the average point
- **Right (CreDEs)**: Each CreNet outputs probability intervals defining a credal set (shaded region); intervals are aggregated

Definition

- A **credal set** $\mathbb{Q}$ is a *convex set of probability distributions*
- Represents uncertainty as a *set* of plausible distributions (not just one)

Credal Set from Probability Intervals

Given lower bounds $\boldsymbol{q}_L$ and upper bounds $\boldsymbol{q}_U$:

$$\mathbb{Q} = \left\{ \boldsymbol{q} \,\middle|\, q_i \in [q_{L_i}, q_{U_i}], \sum_{i=1}^{C} q_i = 1 \right\}$$

Validity Conditions:

- $q_{L_i} \leq q_{U_i}$ for all classes i
- $\sum_{i=1}^{C} q_{L_i} \leq 1 \leq \sum_{i=1}^{C} q_{U_i}$ (ensures non-empty credal set)

Section 2: Background

Distributionally Robust Optimization (DRO)

Standard Training

- Implicitly assumes training and test distributions are identical

DRO Approach

- Minimizes **worst-case expected risk** over uncertain distributions:

$$\min_{\theta} \sup_{U \in \mathcal{U}} \mathbb{E}_{(\boldsymbol{x},\boldsymbol{t}) \sim U} \mathcal{L}((\boldsymbol{x}, \boldsymbol{t}), \theta)$$

- Prepares model for scenarios where test data differs from training
- Chooses θ that performs well under the *worst* distribution in a neighborhood of the training distribution

Key Insight for CreNets:

- Use classic training for “optimistic” upper bounds
- Use DRO for “pessimistic” lower bounds
- Interval width reflects uncertainty about distribution shift

Uncertainty Quantification in DEs

- Total Uncertainty (TU): $H(\bar{\boldsymbol{q}})$ — entropy of averaged prediction
- Aleatoric Uncertainty (AU): $\tilde{H}(\boldsymbol{q}) = \frac{1}{M}\sum_{m=1}^{M} H(\boldsymbol{q}_m)$
- Epistemic Uncertainty (EU): $H(\bar{\boldsymbol{q}}) - \tilde{H}(\boldsymbol{q})$

Note: $H(\boldsymbol{q}) = -\sum_k q_k \log q_k$ is the entropy of the class probability vector $\boldsymbol{q}$, higher values mean more spread (more uncertainty).

Limitation: Empirical evidence suggests DEs yield **low-quality EU estimates**

Section 3: Methods

Overview

This section covers:

1. **CreNet Architecture**: Modified final layer outputting probability intervals
2. **Training Procedure**: Composite loss with classic CE + DRO components
3. **Class Prediction**: Maximin and maximax criteria
4. **Uncertainty Quantification**: Upper/lower entropy for EU estimation

Section 3: Methods

CreNet Architecture

Key Modification: C nodes → $2C$ nodes

- First C nodes: Interval **midpoints** $\boldsymbol{m}$
- Last C nodes: Interval **half-lengths** $\boldsymbol{h}$

Computation (let $\boldsymbol{z}$ = input to final layer):

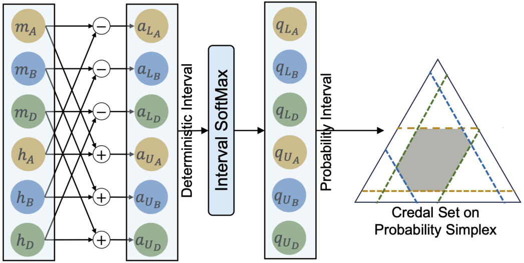
$$\boldsymbol{m} = g(\boldsymbol{W}_{1:C} \cdot \boldsymbol{z} + \boldsymbol{b}_{1:C})$$

$$\boldsymbol{h} = \zeta(\boldsymbol{W}_{C+1:2C} \cdot \boldsymbol{z} + \boldsymbol{b}_{C+1:2C})$$

where $\zeta(\cdot)$ is Softplus (ensures $\boldsymbol{h} \geq 0$)

Deterministic Intervals:

$$[\boldsymbol{a}_L, \boldsymbol{a}_U] = [\boldsymbol{m} - \boldsymbol{h}, \boldsymbol{m} + \boldsymbol{h}]$$



CreNet final layer for 3 classes

Problem: Standard SoftMax on $\boldsymbol{a}_L$ and $\boldsymbol{a}_U$ separately can produce invalid intervals ($q_{L_i} > q_{U_i}$)

Interval SoftMax [Wang et al., 2024]:

$$q_{L_i} = \frac{\exp(a_{L_i})}{\exp(a_{L_i}) + \sum_{k \neq i} \exp\left(\frac{a_{U_k} + a_{L_k}}{2}\right)}, \quad q_{U_i} = \frac{\exp(a_{U_i})}{\exp(a_{U_i}) + \sum_{k \neq i} \exp\left(\frac{a_{U_k} + a_{L_k}}{2}\right)}$$

Guaranteed Properties:

- $q_{L_i} \leq q_{U_i}$ for all classes i
- $\sum_{i=1}^{C} q_{L_i} \leq 1 \leq \sum_{i=1}^{C} q_{U_i}$

$\Rightarrow$ Always produces **valid credal sets**

Section 3: Methods

Loss Function Design

Goal: Interval width should reflect epistemic uncertainty about train-test divergence

Two-Component Loss Strategy

Component	**Applied to**	**Intuition**
Classic CE	Upper probability $\boldsymbol{q}_U$	Optimistic: assumes test $\approx$ train
DRO-inspired	Lower probability $\boldsymbol{q}_L$	Pessimistic: accounts for distribution shift

Result:

- Upper bound q_{U_i}: "Best case" if test matches training
- Lower bound q_{L_i}: "Worst case" if distribution shifts
- **Interval width** reflects epistemic uncertainty

Section 3: Methods

CreNet Loss Function

Complete Loss Function:

$$\mathcal{L}_{\text{CreNet}} = \underbrace{\frac{1}{N}\sum_{n=1}^{N} \text{CE}(\boldsymbol{q}_{U_n}, \boldsymbol{t}_n)}_{\text{Classic Component}} + \underbrace{\max_{\boldsymbol{w}\in\mathbb{S}} \frac{1}{N}\sum_{n=1}^{N} w_n \cdot \text{CE}(\boldsymbol{q}_{L_n}, \boldsymbol{t}_n)}_{\text{DRO Component}}$$

Component Breakdown:

- **Classic Component**: encourages sharp upper bounds when the model is confident
- **DRO Component**: weights w_n emphasize "hard-to-learn" samples

Note: Upper and lower bounds are *correlated* through Interval SoftMax

Section 3: Methods

Training Procedure

Practical Implementation of DRO Component

For each batch, select the $\delta \in [0.5, 1)$ fraction of samples with **highest** $\mathrm{CE}(\boldsymbol{q}_L, \boldsymbol{t})$

Algorithm 1 CreNet Training Procedure

```
Require: Training data $\mathcal{D}$, portion $\delta \in [0.5, 1)$, batch size $\eta$
while training do
    Compute $\mathrm{CE}(\boldsymbol{q}_{U_n}, \boldsymbol{t}_n)$ and $\mathrm{CE}(\boldsymbol{q}_{L_n}, \boldsymbol{t}_n)$ for each sample
    Sort samples by $\mathrm{CE}(\boldsymbol{q}_{L_n}, \boldsymbol{t}_n)$ in descending order
    Define $\eta_\delta = \lfloor \delta \cdot \eta \rfloor$
    Minimize: $\mathcal{L} = \frac{1}{\eta} \sum_{n=1}^{\eta} \mathrm{CE}(\boldsymbol{q}_{U_n}, \boldsymbol{t}_n) + \frac{1}{\eta_\delta} \sum_{j=1}^{\eta_\delta} \mathrm{CE}(\boldsymbol{q}_{L_{m_j}}, \boldsymbol{t}_{m_j})$
end while
```

Hyperparameter δ: Controls pessimism level (default: $\delta = 0.5$)

Section 3: Methods

Class Prediction from Credal Sets

Reachable Probabilities

Not all marginal bounds (q_{L_i}, q_{U_i}) are jointly attainable on the simplex. q_{L_i}, q_{U_i} = predicted bounds; $q^*_{L_i}, q^*_{U_i}$ = reachable bounds within the credal set.

$$q^*_{U_i} = \min\left(q_{U_i}, 1 - \sum_{j \neq i} q_{L_j}\right), \qquad q^*_{L_i} = \max\left(q_{L_i}, 1 - \sum_{j \neq i} q_{U_j}\right)$$

Prediction Criteria

Criterion	Formula	Interpretation
Maximin	$\hat{i}_{\text{min}} = \arg\max_i q^*_{L_i}$	Conservative: highest reachable lower bound
Maximax	$\hat{i}_{\text{max}} = \arg\max_i q^*_{U_i}$	Optimistic: highest reachable upper bound

Generalized Entropy for Credal Sets

Upper Entropy (Total Uncertainty): the most disordered distribution inside the credal set

$$\overline{H}(\mathbb{Q}) = \max_{\boldsymbol{q}} \sum_{i=1}^{C} -q_i \cdot \log_2 q_i \quad \text{s.t. } q_{L_i}^* \leq q_i \leq q_{U_i}^*, \ \sum_{i=1}^{C} q_i = 1$$

Lower Entropy (Aleatoric Uncertainty): the most concentrated distribution inside the credal set

$$\underline{H}(\mathbb{Q}) = \min_{\boldsymbol{q}} \sum_{i=1}^{C} -q_i \cdot \log_2 q_i \quad \text{(same constraints)}$$

Epistemic Uncertainty:

$$\text{EU} = \overline{H}(\mathbb{Q}) - \underline{H}(\mathbb{Q})$$

Intuition: Wide interval → large gap between max/min entropy → high EU

Section 3: Methods

Credal Deep Ensembles (CreDEs)

Ensemble Construction
Train M CreNets with different random seeds, then aggregate by averaging:

$$\tilde{q}_L^* = \frac{1}{M}\sum_{m=1}^{M} q_{L_m}^*, \quad \tilde{q}_U^* = \frac{1}{M}\sum_{m=1}^{M} q_{U_m}^*$$

Key Property: Averaged intervals still satisfy credal set validity conditions

Why Averaging?

- Reduces uncertainty from random initialization
- Remaining interval width reflects **train-test distribution divergence**

Standard practice: $M = 5$ ensemble members

Section 4: Experimental Results

Overview

Goal: Evaluate CreDEs vs Deep Ensembles on uncertainty quantification quality

Datasets

- In-Distribution (ID): CIFAR10, CIFAR100, ImageNet
- Out-of-Distribution (OOD): SVHN, Tiny-ImageNet, CIFAR10-C, ImageNet-O

Backbones: ResNet50, VGG16, ViT Base

Setup: 15 models trained per method, 5-member ensembles

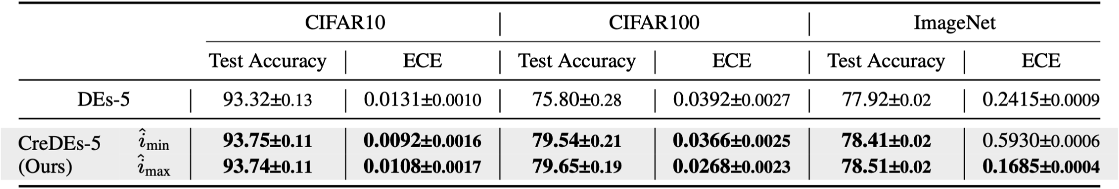
Comparison:

- DEs: EU $= H(\bar{\boldsymbol{q}}) - \tilde{H}(\boldsymbol{q})$
- CreDEs: EU $= \overline{H}(\mathbb{Q}) - \underline{H}(\mathbb{Q})$

Metrics: Test Accuracy, ECE, AUROC, AUPRC for OOD detection

Section 4: Experimental Results

Accuracy and Calibration



		CIFAR10		CIFAR100		ImageNet	
		Test Accuracy	ECE	Test Accuracy	ECE	Test Accuracy	ECE
DEs-5		93.32±0.13	0.0131±0.0010	75.80±0.28	0.0392±0.0027	77.92±0.02	0.2415±0.0009
CreDEs-5	$\hat{i}_{\text{min}}$	**93.75±0.11**	**0.0092±0.0016**	**79.54±0.21**	**0.0366±0.0025**	**78.41±0.02**	0.5930±0.0006
(Ours)	$\hat{i}_{\text{max}}$	**93.74±0.11**	**0.0108±0.0017**	**79.65±0.19**	**0.0268±0.0023**	**78.51±0.02**	**0.1685±0.0004**

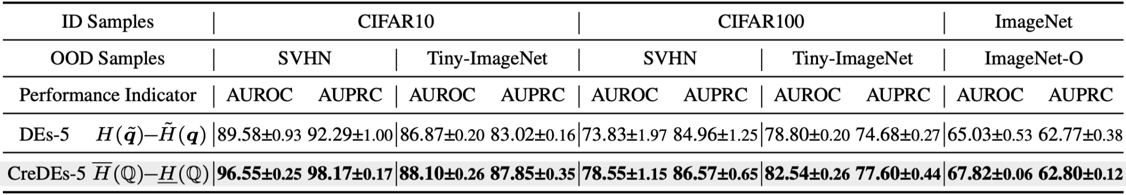
Table 1. Test accuracy and ECE of DEs-5 and CreDEs-5

Key Findings:

- CreDEs achieve **higher accuracy** than DEs
- CreDEs have **lower ECE** (better calibrated)

Section 4: Experimental Results

OOD Detection Performance



ID Samples	CIFAR10				CIFAR100				ImageNet	
OOD Samples	SVHN		Tiny-ImageNet		SVHN		Tiny-ImageNet		ImageNet-O	
Performance Indicator	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
DEs-5 $H(\tilde{\boldsymbol{q}})-\tilde{H}(\boldsymbol{q})$	89.58±0.93	92.29±1.00	86.87±0.20	83.02±0.16	73.83±1.97	84.96±1.25	78.80±0.20	74.68±0.27	65.03±0.53	62.77±0.38
CreDEs-5 $\overline{H}(\mathbb{Q})-\underline{H}(\mathbb{Q})$	**96.55±0.25**	**98.17±0.17**	**88.10±0.26**	**87.85±0.35**	**78.55±1.15**	**86.57±0.65**	**82.54±0.26**	**77.60±0.44**	**67.82±0.06**	**62.80±0.12**

Table 2. OOD detection AUROC and AUPRC

Key Findings:

- CreDEs **significantly outperform** DEs on OOD detection
- Better EU quantification → better OOD detection

Section 4: Experimental Results

OOD Detection: CIFAR10 vs CIFAR10-C

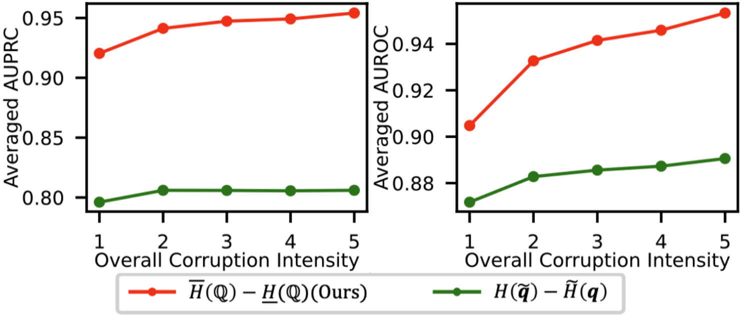


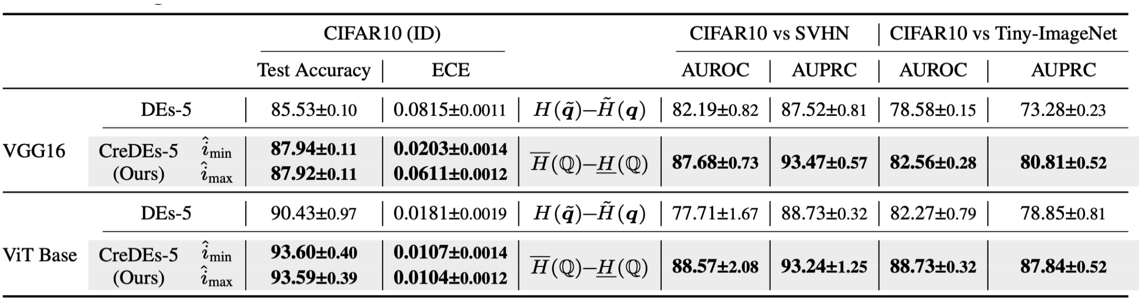
Figure 3: OOD detection vs corruption intensity]

Key Findings:

- CreDEs maintain advantage across all corruption intensities
- Performance gap increases with corruption severity

Section 4: Experimental Results

Ablation Study Highlights



			CIFAR10 (ID)			CIFAR10 vs SVHN		CIFAR10 vs Tiny-ImageNet	
			Test Accuracy	ECE		AUROC	AUPRC	AUROC	AUPRC
VGG16	DEs-5		85.53±0.10	0.0815±0.0011	$H(\tilde{\boldsymbol{q}})-\tilde{H}(\boldsymbol{q})$	82.19±0.82	87.52±0.81	78.58±0.15	73.28±0.23
	CreDEs-5 (Ours)	$\hat{i}_{\min}$	**87.94±0.11**	**0.0203±0.0014**	$\overline{H}(\mathbb{Q})-\underline{H}(\mathbb{Q})$	**87.68±0.73**	**93.47±0.57**	**82.56±0.28**	**80.81±0.52**
		$\hat{i}_{\max}$	**87.92±0.11**	**0.0611±0.0012**					
ViT Base	DEs-5		90.43±0.97	0.0181±0.0019	$H(\tilde{\boldsymbol{q}})-\tilde{H}(\boldsymbol{q})$	77.71±1.67	88.73±0.32	82.27±0.79	78.85±0.81
	CreDEs-5 (Ours)	$\hat{i}_{\min}$	**93.60±0.40**	**0.0107±0.0014**	$\overline{H}(\mathbb{Q})-\underline{H}(\mathbb{Q})$	**88.57±2.08**	**93.24±1.25**	**88.73±0.32**	**87.84±0.52**
		$\hat{i}_{\max}$	**93.59±0.39**	**0.0104±0.0012**					

Table 3. Accuracy, ECE, and OOD Detection Results over Different Backbones

Key Findings:

- **Architecture Robustness**: Improvements hold for VGG16 and ViT Base
- **Hyperparameter** δ: Model is robust to choice of δ
- Outperforms DEs with DRO, MCDropout, and BNN baselines

Strengths and Limitations

Strengths

- **Simple modification**: Only changes final layer ($C \to 2C$ nodes)
- **Theoretical foundation**: Valid credal sets guaranteed
- **Improvements**: Across architectures and datasets

Limitations

- **Training cost**: Slower per epoch (custom training loop)
- **Ensemble requirement**: Still needs M networks
- **Hyperparameter**: δ may need tuning

Section 5: Conclusion

Future Directions and Summary

Future Work:

- Statistical coverage guarantees via conformal prediction
- Extension to regression tasks
- Real-world validation in medical imaging

Key Takeaways:

1. **CreNets** predict probability intervals instead of point probabilities
2. **Training** uses composite loss: classic CE (upper) + DRO (lower)
3. **CreDEs** aggregate CreNets by averaging intervals
4. **EU** $= \overline{H}(\mathbb{Q}) - \underline{H}(\mathbb{Q})$

References

Key References:

- Lakshminarayanan, A., Pritzel, A., & Blundell, C. (2017). *Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles*. NeurIPS.
- Hüllermeier, E., & Waegeman, W. (2021). *Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction*. Machine Learning, 110(3), 457–506.
- Sagawa, S., Koh, P. W., Hashimoto, T., & Liang, P. (2019). *Distributionally Robust Neural Networks*. ICLR.

Questions?