RIT | Rochester Institute of Technology

Uncertainty in Machine Learning

Sheeraja Rajakrishnan

February 20, 2025

Confidence or Uncertainty?

Confidence vs. Uncertainty

Confidence	Uncertainty
Model's certainty about prediction	How much model's prediction is variable
High confidence variance	Low uncertainty
Low confidence variance	High uncertainty

Examples:

Image 1: (Cat: 0.8, Dog: 0.1, Rabbit: 0.1) - high variance, high confidence, low uncertainty

Image 2: (Cat: 0.4, Dog: 0.35, Rabbit: 0.25) - low variance, low confidence, high uncertainty

Uncertainty in Machine Learning

1. Model needs to know the unknown
2. High confidence for OOD data
3. Quantifies model's trust and usefulness

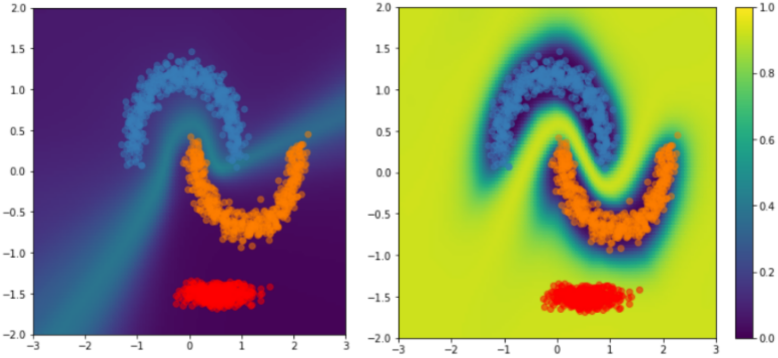
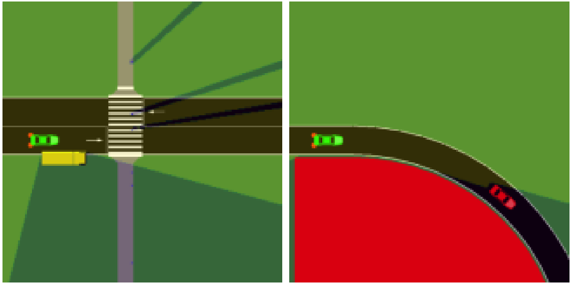
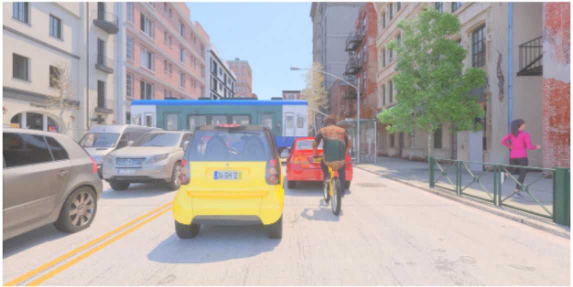


Image: Simple and Principled Uncertainty Estimation with Deterministic Deep Learning via Distance Awareness, Liu et. al (2020)

Uncertainty in Safety-Critical Domains



(a) Occlusion at a pedestrian crossing and occlusion due to a curvy road can cause uncertainty[1]



(b) Image with an OOD bicyclist introduces uncertainty

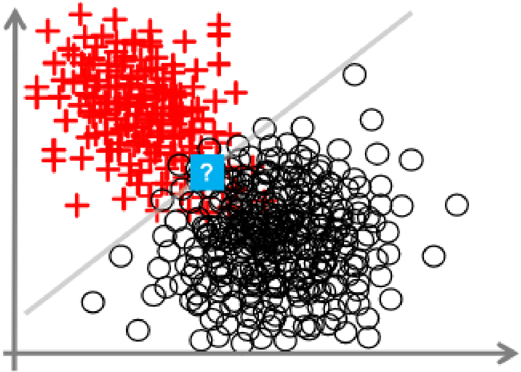
(a) Motion Planning for Autonomous Vehicles in the Presence of Uncertainty Using Reinforcement Learning, Rezaee et. al (2021), (b) MUAD: Multiple Uncertainties for Autonomous Driving, a benchmark for multiple uncertainty types and tasks, Franchi et. al (2022)
[1] https://www.iihs.org/news/detail/self-driving-vehicles-could-struggle-to-eliminate-most-crashes

Types of Uncertainty

- Aleatoric or Data Uncertainty
- Epistemic or Model Uncertainty

Aleatoric Uncertainty

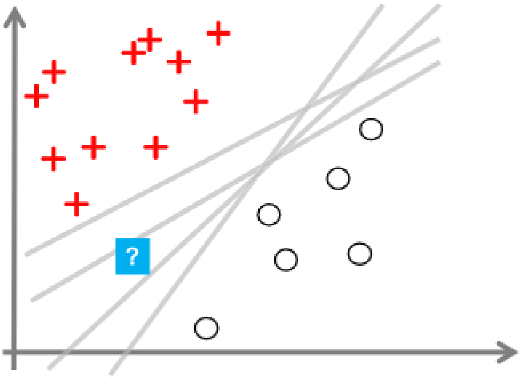
1. Data uncertainty
2. Cannot be reduced
3. More training data - No effect



Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods, Hullermeier and Waegeman (2020)

Epistemic Uncertainty

1. Model uncertainty
2. Can be reduced
3. More training data can reduce
4. Used to identify OOD data

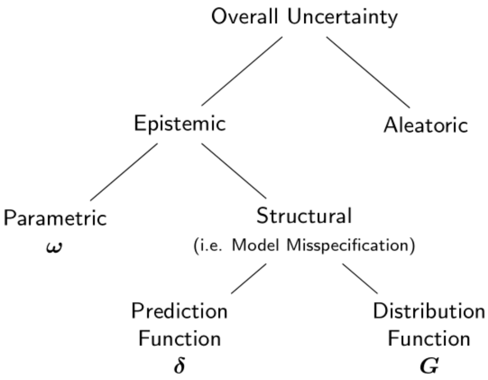


Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods, Hullermeier and Waegeman (2020)

Other Types of Uncertainty

Other less-explored uncertainties:

1. Parametric - model parameter estimations
2. Structural - model specs to describe data
3. Prediction function bias - systematic bias
4. Distribution function bias - distribution not capturing data stochasticity



Accurate Uncertainty Estimation and Decompositionin Ensemble Learning, Liu, Jeremiah, et al. (2019)

Bayesian Deep Learning

- BNNs learn probability distributions of the weights and activations - this overcomes the challenges of NN by providing point estimates
- Place priors over network weights
- Given a prior belief $p(\theta)$ and likelihood $p(D|\theta)$, Bayes' rule posterior is given by

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{\int p(D|\theta)p(\theta)d\theta}$$

- The denominator is intractable and the posterior predictive distribution can be used

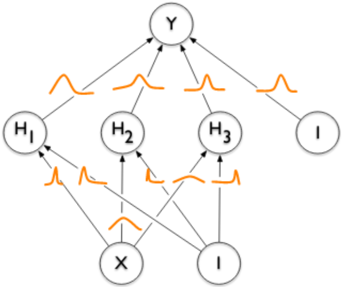
$$P(y|x, D) = \int P(y|x, \theta)P(\theta|D)d\theta$$

- The model performance depends on the approximation method
- Requires the training to be modified
- Expensive computations compared to non-Bayesian NNs

Prior and Posterior Networks: A Survey on Evidential Deep Learning Methods For Uncertainty Estimation, Ulmer et. al (2023)

Bayesian Deep Learning

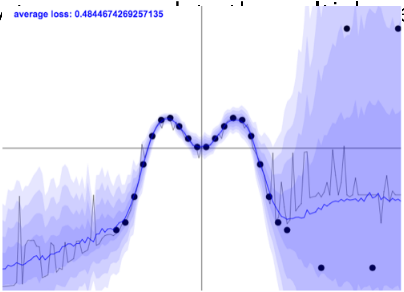
- Posterior approximations obtained using dropout, ensembles
- Requires expensive sample for variance predictions
- Model performance depends on approximation methods used



Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks (2015)

Ensemble Learning

- Use multiple versions of models or data
- Samples required for estimating uncertainty
- Better performance than BNNs
- Robust to OOD data
- Needs more memory [illegible]etwork parameters



https://github.com/yaringal/HeteroscedasticDropoutUncertainty

Prior Networks

- Does not require sampling
- Places prior distributions over hierarchical models
- Require OOD data for training
- Suitable for discrete learning scenarios

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

Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior network (2018)

Evaluation and Calibration

1. Temperature Scaling - softens the NN output
   - higher T $\Rightarrow$ more confident, less calibrated predictions
   - lower T $\Rightarrow$ less confident, more calibrated predictions

$$P(\hat{y}) = \frac{e^{z/T}}{\sum_j e^{z_j/T}}$$

   where $\hat{y}$ is the prediction, z is the logit and T is the learned temperature.
2. Expected Uncertainty Calibration Error (UCE) is used to evaluate the model
   - NN output is split into M equal sized bins
   - Uncertainty values are compared with the values of the bin and placed in appropriate bins
   - B_m is the number of items in bin m, n is the total number of items, $\text{err}(B_m)$ is the mean error of bin m and $\text{uncert}(B_m)$ is the mean uncertainty of bin m
3. Model with lower UCE value is a well-calibrated model

$$UCE = \sum_{m=1}^{M} \frac{|B_m|}{n} |\text{err}(B_m) - \text{uncert}(B_m)|$$

Laves, Max-Heinrich, et al. "Well-calibrated model uncertainty with temperature scaling for dropout variational inference." arXiv preprint arXiv:1909.13550 (2019).
Guo, C., Pleiss, G., Sun, Y. and Weinberger, K.Q. On Calibration of Modern Neural Networks. In ICML, 2017

Expected Uncertainty Calibration Error (UCE)

Calibration - adjust model predictions to align with the ground truth

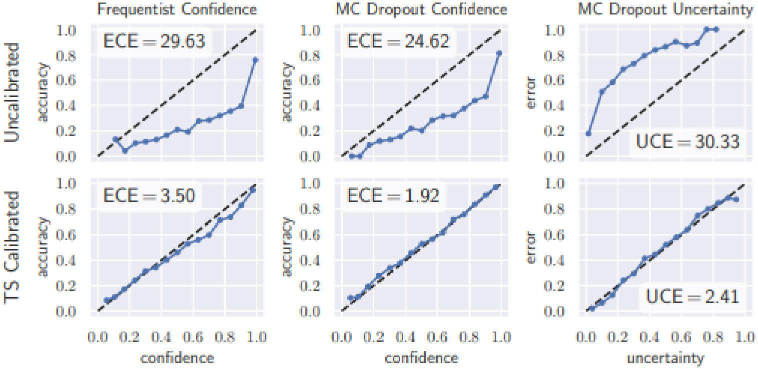


Figure 1: Reliability diagrams ($M = 15$ bins) for ResNet-101 on CIFAR-100. Top row: Uncalibrated frequentist confidence (left), and confidence and uncertainty obtained by dropout variational inference (right). Bottom row: Results from calibration with TS. Dashed lines denote perfect calibration.

Laves, Max-Heinrich, et al. "Well-calibrated model uncertainty with temperature scaling for dropout variational inference." arXiv preprint arXiv:1909.13550 (2019).

Evidential Deep Learning

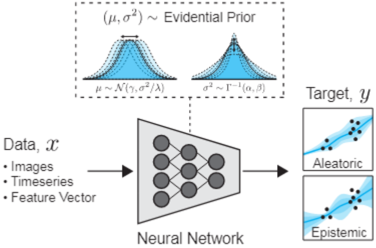
Evidential Deep Learning (EDL)

- Evidence-collecting process
- More evidence $\implies$ high confidence and low uncertainty of predictions
- Can learn evidence variables directly from the data
- Robust to different uncertainty sources
- Requires novel and complex loss function - uses approximation or regularization methods (e.g. softmax approximation)
- The regularization coefficient needs to be tuned to remove evidence, that is not misleading, from uncertainty calibration

Evidential Deep Learning to Quantify Classification Uncertainty, Sensoy et. al (2018)
Deep Evidential Regression, Amini et. al (2020)

Deep Evidential Regression (DER)

- Applies to a continuous regression problem
- Evidential prior distribution is placed over the likelihood function and the network is then trained to obtain the hyperparameters of this evidential distribution
- No sampling or training on OOD, single model training
- Predicts a uniform distribution for OOD data
- Misleading evidence is minimized for incorrect predictions to increase uncertainty



Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep Evidential Regression, (2020)

Basic Idea of DER

- Priors are placed on the unknown mean and variance of the target distribution
  - Mean: $\mu \sim \mathcal{N}(\gamma, \sigma^2\nu^{-1}) \rightarrow \textit{Gaussian}$
  - Variance: $\sigma^2 \sim \Gamma^{-1}(\alpha, \beta) \rightarrow \textit{Inverse} - \textit{Gamma}$
- The posterior distribution is obtained by factorizing the estimated distribution as the NIG distribution (Gaussian conjugate prior):

$$p(\mu, \sigma^2|\gamma, \nu, \alpha, \beta) = \frac{\beta^\alpha\sqrt{\nu}}{\Gamma(\alpha)\sqrt{2\pi\sigma^2}}\left(\frac{1}{\sigma^2}\right)^{\alpha+1} exp\left\{-\frac{2\beta + \nu(\gamma-\mu)^2}{2\sigma^2}\right\}$$

- The first order moments of the above distribution gives the uncertainties

$\mathbb{E}[\mu] = \gamma$	$\mathbb{E}[\sigma^2] = \frac{\beta}{\alpha-1}$	$Var[\mu] = \frac{\beta}{\nu(\alpha-1)}$
Prediction	Aleatoric Uncertainty	Epistemic Uncertainty

Model Learning

The evidential prior distributions are optimized in 2 ways:

1. Maximizing the model fit - analytical solution for intractable model evidence

$$\mathcal{L}_i^{NLL}(w) = \frac{1}{2} log\left(\frac{\pi}{\nu}\right) - \alpha log(\Omega) + \left(\alpha + \frac{1}{2}\right) log((y_i - \gamma)^2\nu + \Omega) + log\left(\frac{\Gamma(\alpha)}{\Gamma(\alpha + \frac{1}{2})}\right)$$

   This is the NLL of the model evidence, which is a Student-t distribution.

2. Minimizing the evidence on errors

$$\mathcal{L}_i^{R}(w) = |y_i - \gamma| \cdot (2\nu + \alpha)$$

   where $(2\nu + \alpha)$ is the total evidence. This regularization term imposes a penalty in the case of a wrong prediction.

3. Total loss is given by: $\mathcal{L}_i(w) = \mathcal{L}_i^{NLL}(w) + \mathcal{L}_i^{R}(w)$

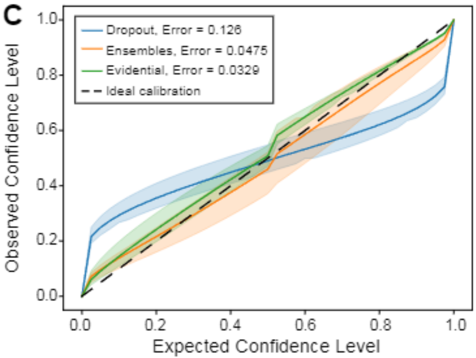
Benchmark Regression Tests

Dataset	RMSE			NLL			Inference Speed (ms)		
	Dropout	Ensembles	Evidential	Dropout	Ensembles	Evidential	Dropout	Ensemble	Evidential
Boston	**2.97 ± 0.19**	**3.28 ± 1.00**	**3.06 ± 0.16**	**2.46 ± 0.06**	**2.41 ± 0.25**	**2.35 ± 0.06**	3.24	3.35	**0.85**
Concrete	**5.23 ± 0.12**	6.03 ± 0.58	5.85 ± 0.15	3.04 ± 0.02	3.06 ± 0.18	**3.01 ± 0.02**	2.99	3.43	**0.94**
Energy	**1.66 ± 0.04**	2.09 ± 0.29	2.06 ± 0.10	1.99 ± 0.02	**1.38 ± 0.22**	**1.39 ± 0.06**	3.08	3.80	**0.87**
Kin8nm	0.10 ± 0.00	**0.09 ± 0.00**	**0.09 ± 0.00**	-0.95 ± 0.01	-1.20 ± 0.02	**-1.24 ± 0.01**	3.24	3.79	**0.97**
Naval	0.01 ± 0.00	**0.00 ± 0.00**	**0.00 ± 0.00**	-3.80 ± 0.01	**-5.63 ± 0.05**	**-5.73 ± 0.07**	3.31	3.37	**0.84**
Power	**4.02 ± 0.04**	**4.11 ± 0.17**	4.23 ± 0.09	**2.80 ± 0.01**	**2.79 ± 0.04**	**2.81 ± 0.07**	2.93	3.36	**0.85**
Protein	**4.36 ± 0.01**	4.71 ± 0.06	4.64 ± 0.03	2.89 ± 0.00	2.83 ± 0.02	**2.63 ± 0.00**	3.45	3.68	**1.18**
Wine	**0.62 ± 0.01**	**0.64 ± 0.04**	**0.61 ± 0.02**	**0.93 ± 0.01**	**0.94 ± 0.12**	**0.89 ± 0.05**	3.00	3.32	**0.86**
Yacht	**1.11 ± 0.09**	**1.58 ± 0.48**	**1.57 ± 0.56**	1.55 ± 0.03	**1.18 ± 0.21**	**1.03 ± 0.19**	2.99	3.36	**0.87**

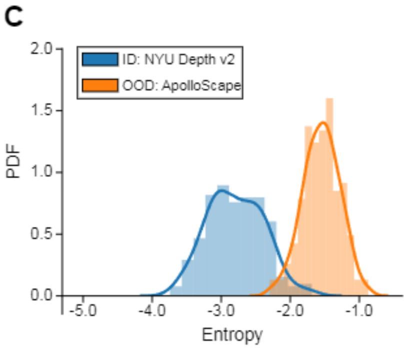
Benchmark models are outperformed by evidential models for NLL and the inference speed on all datasets

Monocular Depth Estimation

Predict the depth of pixels from a high-dimensional RGB image



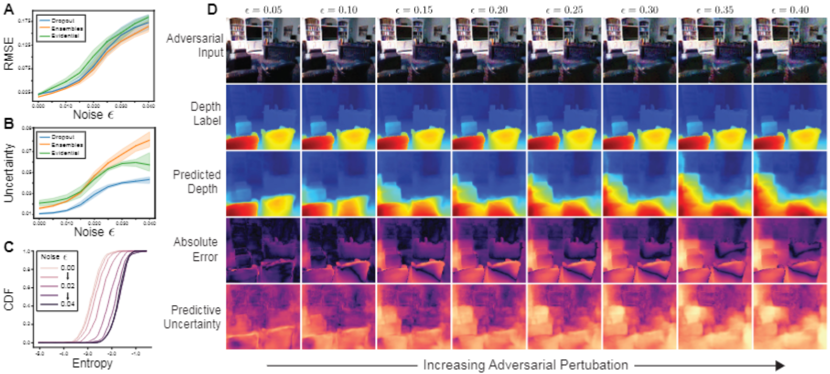
Model uncertainty calibration (ideal y=x)



Evidential entropy on ID and OOD data

Adversarial Noise

OOD detection from adversarially perturbed inputs



ϵ is the noise scale, (D) shows the effects of increasing perturbations on the predictions, error, and uncertainty, for evidential regression

Summary

1. DER is a scalable method for estimating aleatoric and epistemic uncertainty
2. An evidential regularizer enables OOD samples to be penalized
3. Evaluation of DER against state-of-the-art uncertainty estimation models
4. Evaluation of DER calibration on OOD data

Dempster-Shafer Theory

- Assigns set probability/belief masses
- Evidence for multiple events (any class is likely)
- 3 important functions:
  - Probability assignment function (m): a belief mass for each element of the power set
    $$m : P(X) \to [0,1]$$
    $$m(\phi) = 0; \sum_{A \in P(X)} m(A) = 1$$
  - Belief function (Bel): sum of all the masses of subsets of the set of interest
    $$Bel(A) = \sum_{B|B \subseteq A} m(B)$$
  - Plausibility function (Pl): sum of all the masses of the sets B that intersect the set of interest A
    $$Pl(A) = \sum_{B|B \cap A \neq \phi} m(B)$$

Sentz, K. & Ferson, S. Combination of Evidence in Dempster-Shafer Theory. (2002).
https://en.wikipedia.org/wiki/Dempster-Shafer_theory

Evidential Deep Learning for Classification

- A frame of K mutually exclusive singletons
- Each singleton is assigned a belief mass $b_k \geq 0$
- Overall uncertainty mass is $u \geq 0$

$$u + \sum_{k=1}^{K} b_k = 1$$

- b_k is computed from the evidence e_k

$$b_k = \frac{e_k}{S} \text{ and } u = \frac{K}{S}$$

K is the number of classes and $S = \sum_{i=1}^{K}(e_i + 1)$

Sensoy, M., Kaplan, L. & Kandemir, M. Evidential Deep Learning to Quantify Classification Uncertainty. Arxiv (2018)

Basic Idea of DEC

- A Dirichlet distribution is fit over the probabilities of a neural network classification model
- The Dirichlet prior has a probability density function, parameterized by α for K categories and is given by,

$$\text{Dir}(x|\alpha) = \frac{1}{\beta(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1}$$

where $\beta(\alpha) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma(\alpha_0)}$, $\alpha_0 = \sum_{i=1}^{K} \alpha_i$ and $\{x_1, x_2, ..., x_K\}$ represent the support for the K categories and $x_i \epsilon [0, 1]$ where $\sum_{i=1}^{K} x_i = 1$.

- The Dirichlet distribution is a conjugate prior of the Multinomial distribution and the posterior is given by:

$$\text{P}(\theta|x) \propto \text{Dir}(X|x_{nk} + \alpha_k)$$

Model Learning

The evidential prior distributions are optimized in 2 ways:

1. Minimizing the NLL loss (sum of squares loss)

$$\mathcal{L}_i^{NLL}(\Theta) = \sum_{j=1}^{K} (y_{ij} - \hat{p}_{ij})^2 + \frac{\hat{p}_{ij}(1 - \hat{p}_{ij})}{(S_i + 1)}$$

   where $y_i j$ is the one-hot vector of the ground-truth observation class, $S_i = \sum_{i=1}^{K} \alpha_i$, $\alpha_i = e_i + 1$, e_i is the evidence from the neural network, and $\hat{p}_k = \frac{\alpha_k}{S}$.

2. Penalize states that do not contribute to data fit

$$\mathcal{L}_i^{R}(\Theta) = KL[D(p_i|\tilde{\alpha}_i)||D(p_i| < 1, ..., 1 >)]$$

   where $D(p_i| < 1, ..., 1 >)$ is the uniform Dirichlet distribution.

3. Total loss is given by: $\mathcal{L}_i(\Theta) = \mathcal{L}_i^{NLL}(\Theta) + \lambda_t \ \mathcal{L}_i^{R}(\Theta)$
   where $\lambda_t = min(1.0, t/10) \in [0, 1]$ is the annealing coefficient

EDL For Classification

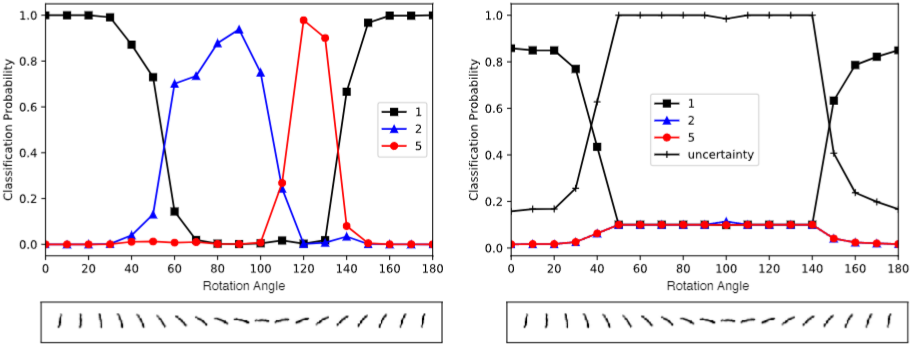
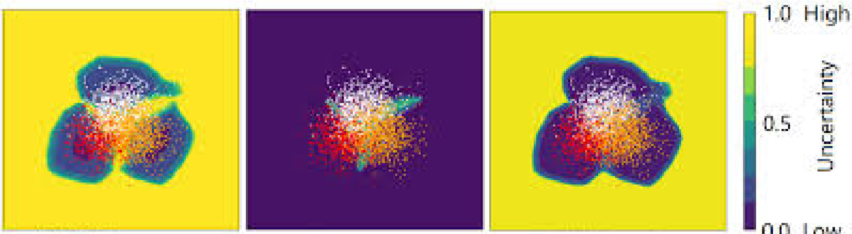


Figure 1: Classification of the rotated digit 1 (at bottom) at different angles between 0 and 180 degrees. **Left:** The classification probability is calculated using the *softmax* function. **Right:** The classification probability and uncertainty are calculated using the proposed method.

Sensoy, M., Kaplan, L. & Kandemir, M. Evidential Deep Learning to Quantify Classification Uncertainty. Arxiv (2018)

Vacuity and Dissonance



(a) Entropy (b) Dissonance (c) Vacuity

- Entropy high in ID and OOD regions
- Dissonance high on boundary (misclassification)
- Vacuity high in OOD region

Hu, Yibo, et al. "Multidimensional uncertainty-aware evidential neural networks." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 35. No. 9. 2021.

Vacuity and Dissonance

- Vacuity - Lack of support

$$u + \sum_{k=1}^{K} b_k = 1$$

u in the above equation is the uncertainty mass that represents the vacuity of evidence

- Dissonance - Conflicting evidence

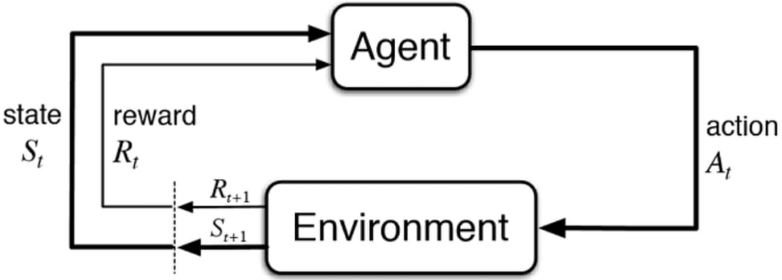
$$b_X^{Diss} = \sum_{x_i \in X} \left(\frac{b_X(x_i) \sum_{x_j \in \mathbb{X} \setminus x_i} b_X(x_j) Bal(x_j, x_i)}{\sum_{x_j \in \mathbb{X} \setminus x_i} b_X(x_j)} \right)$$

where $Bal(x_j, x_i)$ is the relative mass balance between a pair of belief masses, given by

$$Bal(x_j, x_i) = 1 - \frac{|b_X(x_j) - b_X(x_i)|}{b_X(x_j) + b_X(x_i)}$$

Shi, Weishi, et al. "Multifaceted uncertainty estimation for label-efficient deep learning." Advances in neural information processing systems 33 (2020): 17247-17257.
A. Josang, J. -H. Cho and F. Chen, "Uncertainty Characteristics of Subjective Opinions," 2018 21st International Conference on Information Fusion (FUSION), Cambridge, UK, 2018

Uncertainty in Reinforcement Learning



https://www.altexsoft.com/blog/reinforcement-learning-explained-overview-comparisons-and-applications-in-business/

Sources of Uncertainty

1. Aleatoric Uncertainty - random traps
2. Epistemic Uncertainty - actions that neglect exploration such as shortcuts
3. Hinders ability to gain knowledge of better rewards

Stutts, Alex Christopher, et al. "Echoes of Socratic Doubt: Embracing Uncertainty in Calibrated Evidential Reinforcement Learning." arXiv preprint arXiv:2402.07107 (2024).

Uncertainty-Aware Deep Q Network (UADQN)

- Estimates 50 quantiles and uses random MAP sampling to sample 2 anchor networks.
- Thompson sampling overcomes the exploration-exploitation dilemma and uses epistemic uncertainty to prioritize transitions to replay.

$$\tilde{\sigma}^2_{epistemic} = \frac{1}{2}\mathbb{E}_{i\sim U\{1,N\}}[y_i(\theta_A, s, a) - y_i(\theta_B, s, a)]^2$$

- The action mean is updated for risk-aversion

$$\mu = \mu - \lambda\tilde{\sigma}_{aleatoric}$$

where μ is the action mean, λ is a hyperparameter and $\tilde{\sigma}_{aleatoric}$ is the uncertainty calculated from the anchor networks

$$\tilde{\sigma}^2_{aleatoric} = cov_{i\sim U\{1,N\}}(y_i(\theta_A, s, a), y_i(\theta_B, s, a))$$

- The proposed UADQN model is tested in 5 game environments.

Clements, William R., et al. "Estimating risk and uncertainty in deep reinforcement learning." arXiv preprint arXiv:1905.09638 (2019).

UA-DQN

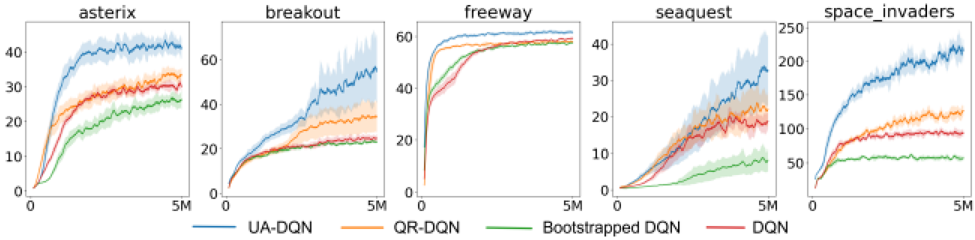


Figure 3. Learning curves over 5 million steps for different agents on the MinAtar testbed. Shaded areas correspond to the 95% confidence interval of the mean obtained from 10 training seeds.

Clements, William R., et al. "Estimating risk and uncertainty in deep reinforcement learning." arXiv preprint arXiv:1905.09638 (2019).

Calibrated Evidential Quantile Regression in Deep Q Network (CEQR-DQN)

- Closely comparable to UADQN, CEQR-DQN uses evidential deep learning to calculate uncertainties

$$\text{Aleatoric: } \mathbb{E}[\sigma^2] = \frac{\beta}{\alpha - 1}; \text{Epistemic: } \text{Var}[\mu] = \frac{\beta}{\nu(\alpha - 1)};$$

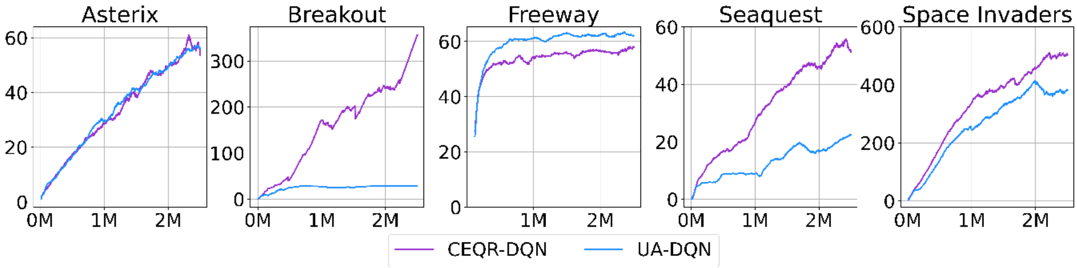
  where α, β and ν are parameters of the evidential distribution.
- Uncertainty is obtained from the 5^{th} and 95^{th} percentiles used to obtain an evidence-based confidence interval ($\mathcal{L}_{interval}$)
- The loss function to be minimized is

$$\mathcal{L}_{EL} = \mathcal{L}_{evi} + \mathcal{L}_{cal} + \mathcal{L}_{interval}$$

  where $\mathcal{L}_{evi}$ is the evidential loss, $\mathcal{L}_{cal}$ is the calibration loss and $\mathcal{L}_{interval}$ is the interval loss.
- The proposed CEQR-DQN model is tested in 5 game environments

Stutts, Alex Christopher, et al. "Echoes of Socratic Doubt: Embracing Uncertainty in Calibrated Evidential Reinforcement Learning." arXiv preprint arXiv:2402.07107 (2024).

CEQR-DQN



Stutts, Alex Christopher, et al. "Echoes of Socratic Doubt: Embracing Uncertainty in Calibrated Evidential Reinforcement Learning." arXiv preprint arXiv:2402.07107 (2024).

Thank You!