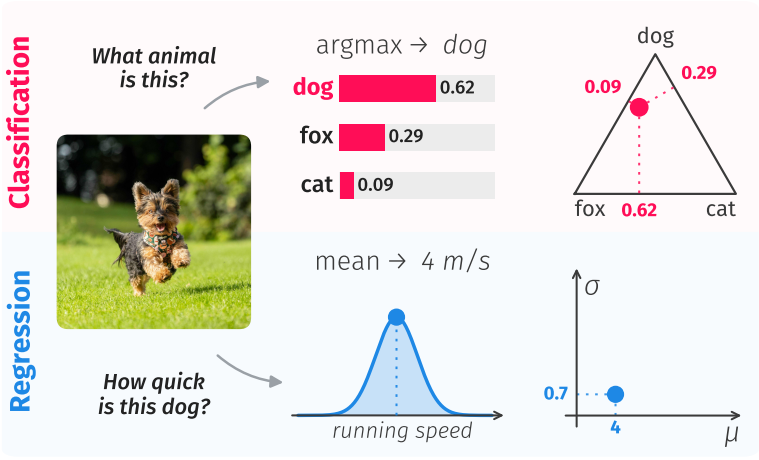
Uncertainty in Machine Learning: From Aleatoric to Epistemic
Day 3: Second-order Uncertainty Representation
Day 4: Uncertainty Quantification and Evaluation

Eyke Hüllermeier[1,2,3]

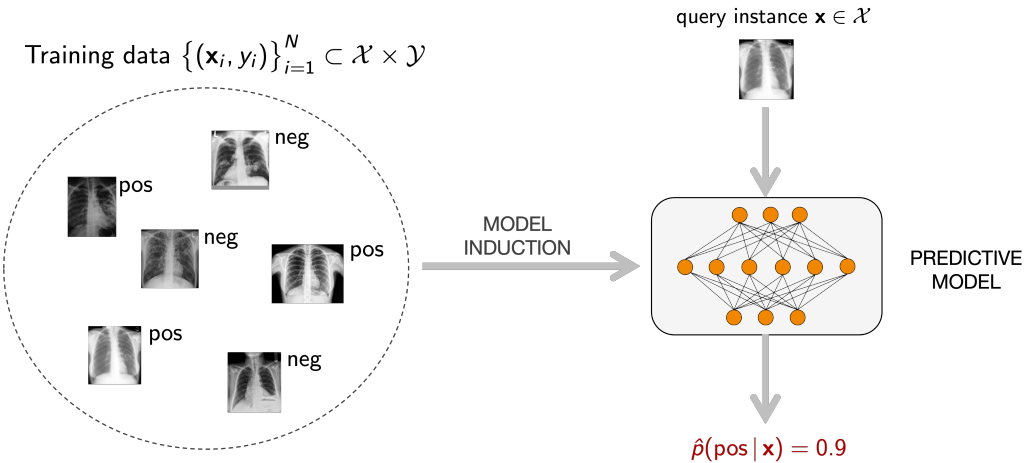
[1]Institute of Informatics, LMU Munich
[2]Munich Center for Machine Learning (MCML)
[3]German Research Center for Artificial Intelligence (DFKI)

European Summer School on Artificial Intelligence (ESSAI), Vienna, July 2026

Probabilistic predictions in machine learning



Need for uncertainty-awareness of ML systems



Lack of uncertainty-awareness of ML systems

- Predictions by state-of-the-art neural network: “umbrella” with confidence 97 % (left), “skirt” with confidence 96 % (right)

Lack of uncertainty-awareness of ML systems

- Hallucination in large language models (Shorinwa *et al.*, 2024):

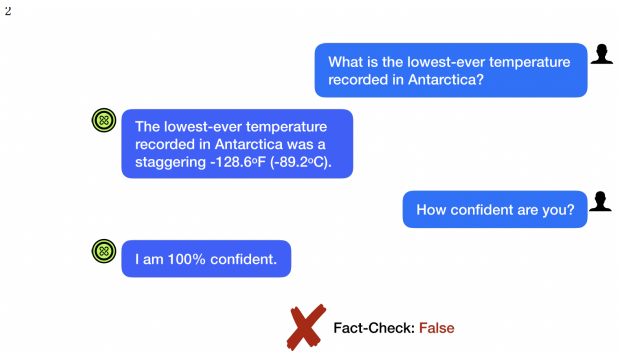
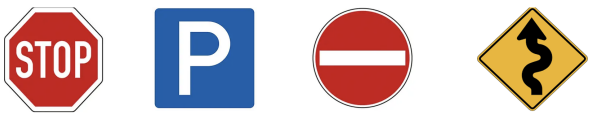


Fig. 1. A user asks an LLM the question: *What is the lowest-ever temperature recorded in Antarctica?*; in response, the LLM answers definitively. Afterwards, the user asks the LLM how confident the LLM is. Although the LLM states that it is "100% confident," the LLM's response fails to pass a fact-check test. Confidence scores provided by LLMs are generally miscalibrated. UQ methods seek to provide calibrated estimates of the confidence of LLMs in their interaction with users.

Out-of-distribution data: “don’t know”

Training data:

Query data:



Uncertainty-aware ML systems

- Ideally, when making a **prediction**, the learner knows what it knows and, perhaps more importantly, **what is does not know**.
- This requires an adequate **representation** of predictions Q (e.g., in terms of distributions or sets) and **quantification** (e.g., in terms of entropy) of their uncertainty, ...
- ... as well as suitable **learning algorithms** to produce Q-valued predictors.
- In this regard, a distinction between different **sources and types of uncertainty** turns out to be meaningful (H. and Waegeman, 2021), notably
  - **aleatoric** (inherent randomness, irreducible),
  - **epistemic** (lack of knowledge, reducible).

Aleatoric versus epistemic uncertainty

H

$1/2$

T

$1/2$

H

$1/2$

T

$1/2$

“Not knowing the chance of mutually exclusive events and knowing the chance to be equal are two quite different states of knowledge"

Ronald Fisher (1890-1962)

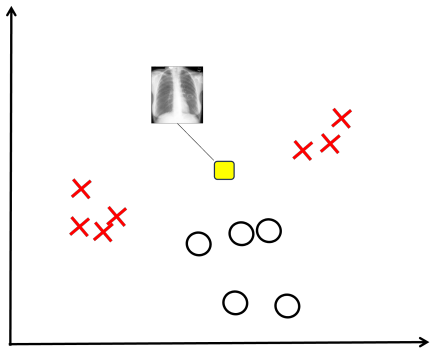
Aleatoric versus epistemic uncertainty

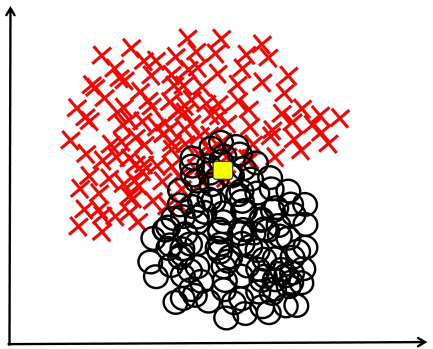
- **Aleatoric** (statistical) uncertainty
  - refers to the notion of **randomness**, that is, the variability in the outcome which is due to inherently random effects,
  - is a property of the **data-generating process**,
  - and as such **irreducible**.

- **Epistemic** (systematic) uncertainty
  - refers to uncertainty caused by a **lack of knowledge**, i.e.,
  - to the epistemic state of the **agent** (e.g., learning algorithm),
  - can in principle be **reduced** on the basis of additional information.

Aleatoric versus epistemic uncertainty in ML

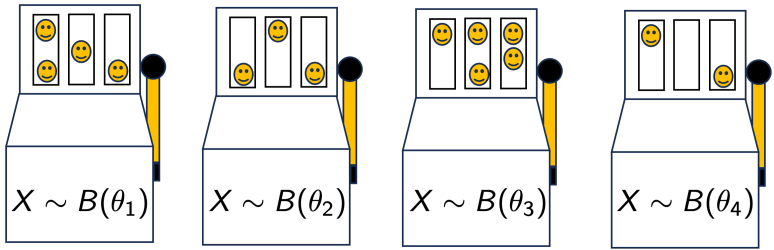
- This distinction is also relevant for machine learning, where the learner's state of knowledge depends (*inter alia*) on the amount of data seen so far:





Machine learning and the explore-exploit dilemma

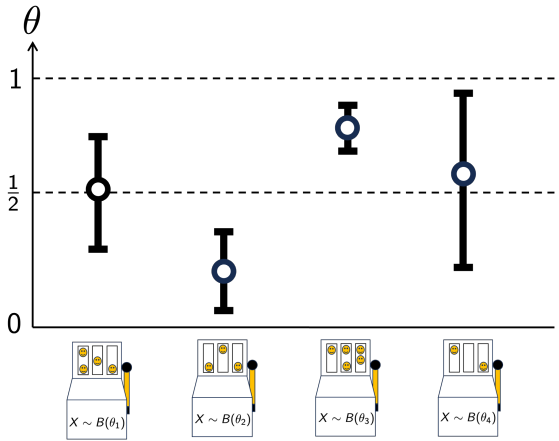
- It is also needed to deal with the **explore-exploit dilemma** — like in multi-armed bandits, where the goal is maximise accumulated reward over time.



- **Exploration** reduces epistemic uncertainty, and **epistemic uncertainty** steers exploration ...

Machine learning and the explore-exploit dilemma

- To play the game optimally, representation of **epistemic uncertainty** is essential and (provably) needed (e.g., upper confidence bounds in UCB).



Agenda

1. Introduction and first-order uncertainty representation (Willem)
2. Conformal prediction (Willem)
3. **Second-order uncertainty representation**
   - The Bayesian approach
   - Second-order loss minimisation
   - Credal prediction
4. Uncertainty quantification and evaluation
5. Summary and outlook

Notation

- Input/feature space $\mathcal{X}$, outcome/label space $\mathcal{Y}$
  - Classification: $\mathcal{Y} = [K] := \{1, \ldots, K\}$
  - Regression: $\mathcal{Y} = \mathbb{R}$
- Generic probability P, i.e., $P(E) =$ probability of event E
- $\mathbb{P}(\mathcal{Y})$ denotes the set of probability distributions on $\mathcal{Y}$
- Vector notation for classification:

$$\boldsymbol{p} = (p_1, \ldots, p_K) \in \Delta_K := \left\{ \boldsymbol{q} \in \mathbb{R}_+^K \,|\, q_1 + \ldots + q_K = 1 \right\}$$

- Training data $\mathcal{D}_N = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N \subset \mathcal{X} \times \mathcal{Y}$ (sampled i.i.d. according to joint P)
- Predictor $h : \mathcal{X} \longrightarrow \mathbb{P}(\mathcal{Y})$ induced from training data
- $\hat{\boldsymbol{p}} = p_h(y \,|\, \boldsymbol{x}) = p(y \,|\, \boldsymbol{x}, h) = h(\boldsymbol{x})$ denotes prediction of ground-truth conditional $p(y \,|\, \boldsymbol{x})$

Supervised machine learning

- Given **training data** $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N \subset \mathcal{X} \times \mathcal{Y}$, typically assumed to be i.i.d. according to some (unknown) measure P on $\mathcal{X} \times \mathcal{Y}$, a **hypothesis space** $\mathcal{H}$ of predictors $\mathcal{X} \longrightarrow \mathbb{P}(\mathcal{Y})$, and a **loss function**

$$L : \mathcal{Y} \times \mathbb{P}(\mathcal{Y}) \longrightarrow \mathbb{R}\,,$$

the learner seeks to find

$$h^* \in \operatorname*{arg\,min}_{h \in \mathcal{H}} R(h)\,,$$

i.e., a predictor $h : \mathcal{X} \longrightarrow \mathbb{P}(\mathcal{Y})$ with **minimal risk** (expected loss)

$$R(h) := \mathbb{E}_{(\boldsymbol{x},y)\sim P} L(y, h(\boldsymbol{x})) = \int_{\mathcal{X}\times\mathcal{Y}} L(y, h(\boldsymbol{x}))\, d\, P(\boldsymbol{x}, y)\,.$$

Aleatoric versus epistemic uncertainty in ML

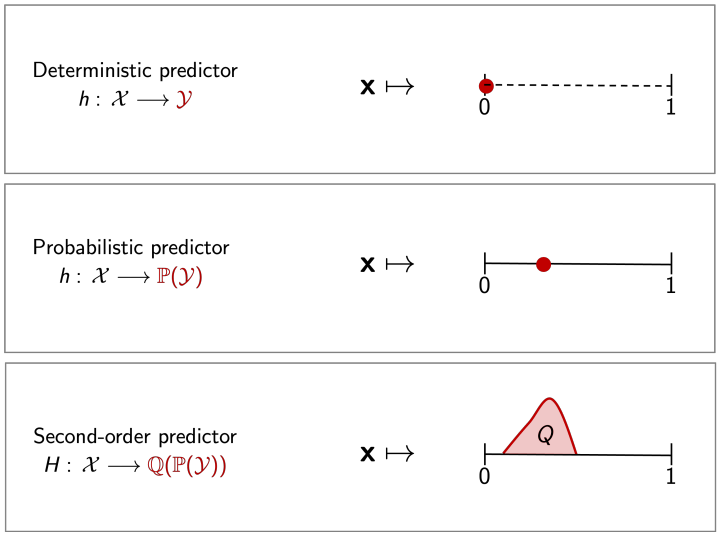
- Assuming that data is generated (i.i.d.) according to a **joint probability distribution** $p \in \mathbb{P}(\mathcal{X} \times \mathcal{Y})$, the true probability of outcome $y \in \mathcal{Y}$ given $\boldsymbol{x} \in \mathcal{X}$ is given by

$$p(y \,|\, \boldsymbol{x}) = \frac{p(\boldsymbol{x}, y)}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{x}, y)}{\sum_{y' \in \mathcal{Y}} p(\boldsymbol{x}, y')} \,. \qquad (\star)$$

- Thus, even if p is known, **aleatoric uncerainty** about the outcome remains, unless $p(y \,|\, \boldsymbol{x})$ degenerates to a Dirac distribution (deterministic case).
- Additional **epistemic uncertainty** is caused by the learner's lack of knowledge of p.
- Instead of $(\star)$, the learner delivers an **approximation** that comes from an empirically induced predictor $h : \mathcal{X} \longrightarrow \mathbb{P}(\mathcal{Y})$:

$$\hat{p}(y \,|\, \boldsymbol{x}) = h(\boldsymbol{x})$$

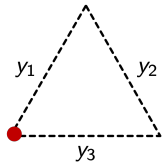
Uncertainty representation and levels of uncertainty-awareness



Uncertainty representation and levels of uncertainty-awareness

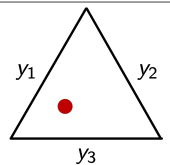
Deterministic predictor
$h : \mathcal{X} \longrightarrow \mathcal{Y}$

$\mathbf{x} \mapsto$



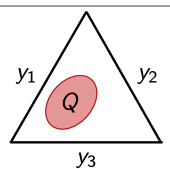
Probabilistic predictor
$h : \mathcal{X} \longrightarrow \mathbb{P}(\mathcal{Y})$

$\mathbf{x} \mapsto$



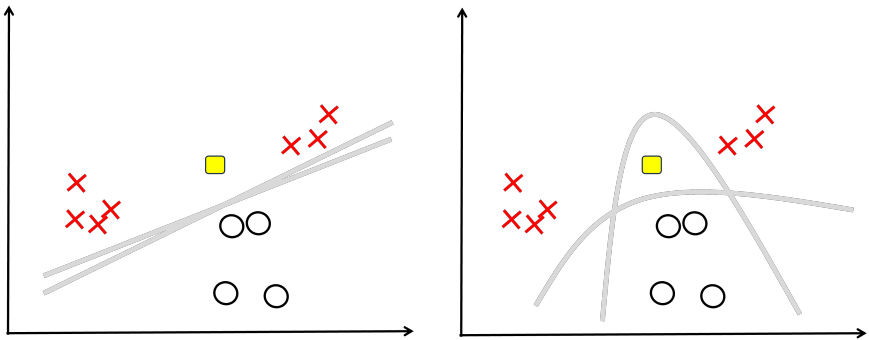
Second-order predictor
$H : \mathcal{X} \longrightarrow \mathbb{Q}(\mathbb{P}(\mathcal{Y}))$

$\mathbf{x} \mapsto$



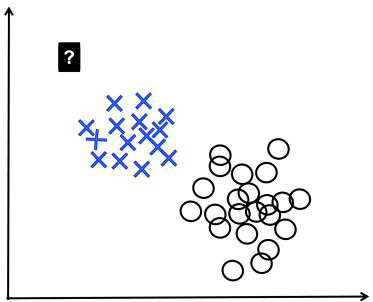
Assumptions and prior knowledge

- Assuming a linear model (left), the query can be safely classified as positive, while relaxing the model assumptions leads to increased uncertainty (right).



Assumptions must be made explicit

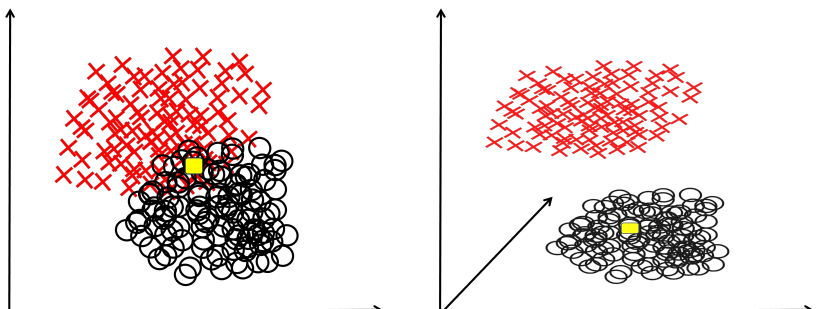
- More generally, UQ requires **assumptions**, which should be made explicit; uncertainty measures and AU/EU separation are only valid **conditional** to these assumptions



- Low uncertainty under the closed world assumption
- High uncertainty under the open world assumptions

Assumptions must be made explicit

- In particular, this includes any means to reduce uncertainty, such as adding data or adding features — what is reducible and what is not?



Predictive uncertainty

- In the standard setting of **supervised learning**, we are mainly interested in (per-instance) **predictive uncertainty**: Instead of a deterministic prediction $\hat{y}$ of the outcome for a query instance $\boldsymbol{x}$, we seek a prediction

$$Q = H(\boldsymbol{x})$$

  adequately representing the learner's uncertainty about the prediction.

- Various **approaches** have been proposed in the literature:
  - The Bayesian approach
  - Post-hoc assessment
  - Second-order loss minimisation

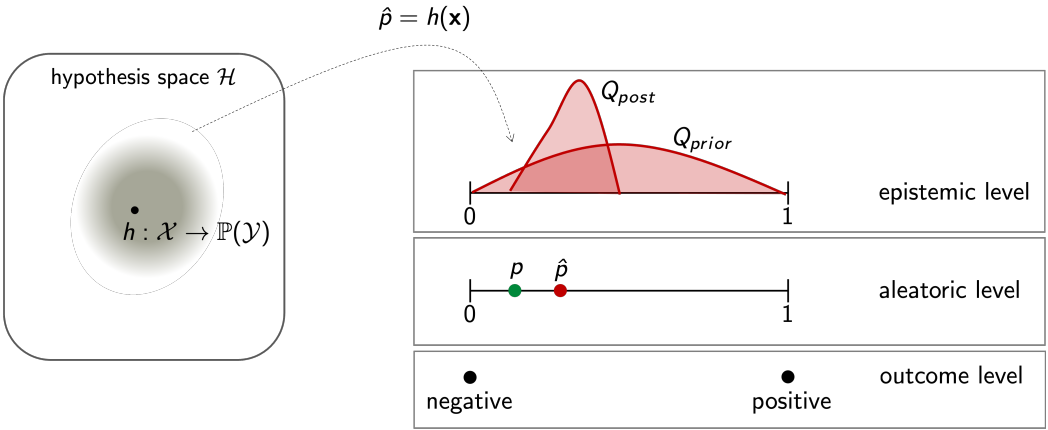
Notation

- Training data $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N} \subset \mathcal{X} \times \mathcal{Y}$
- First-order distributions $p \in \mathbb{P}(\mathcal{Y})$
- First-order predictor $h : \mathcal{X} \longrightarrow \mathbb{P}(\mathcal{Y})$
- Second-order distributions $Q \in \mathbb{Q}(\mathcal{Y})$ $(= \mathbb{P}(\mathbb{P}(\mathcal{Y})))$
- Second-order predictor $H : \mathcal{X} \longrightarrow \mathbb{Q}(\mathcal{Y})$

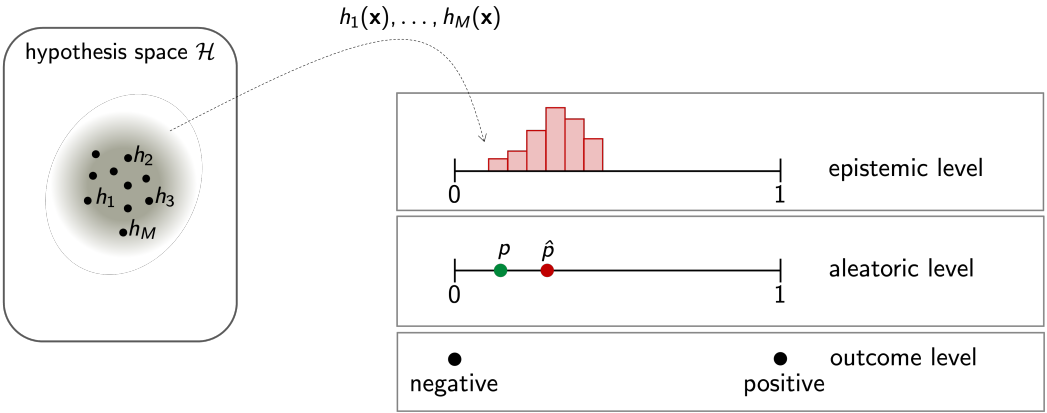
Agenda

1. Introduction
2. Basic statistical concepts
3. First-order uncertainty representation
4. **Second-order uncertainty representation**
   - **The Bayesian approach**
   - Second-order loss minimisation
   - Credal prediction
4. Uncertainty quantification and evaluation
5. Summary and outlook

The Bayesian approach: posterior predictive distribution

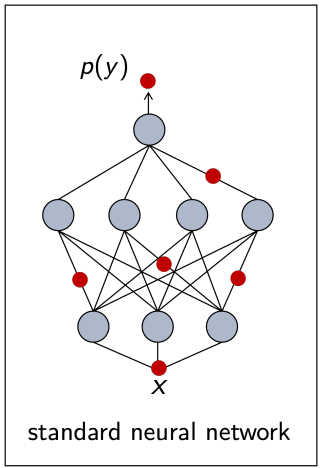


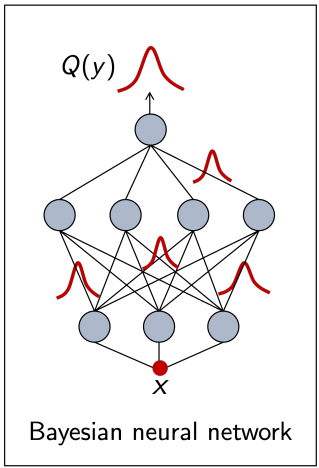
Ensemble methods

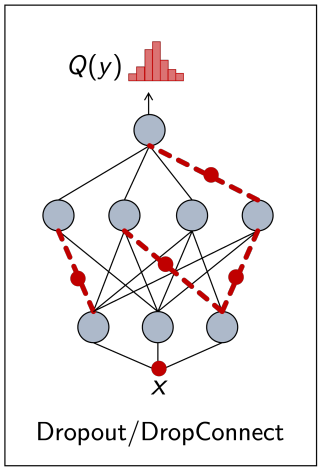


Bayesian neural networks

- Bayesian neural networks use probabilistic weights; approximate inference using sampling (MCMC), variational methods, Laplace approximation, ...

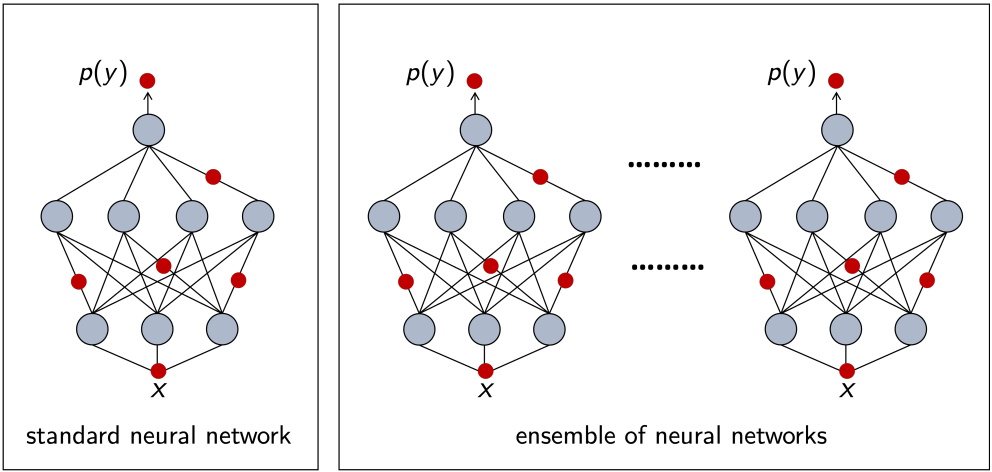






Ensemble methods

- Ensembles of (deep) neural networks (Lakshminarayanan *et al.*, 2017) is a simpler though still costly alternative:



The frequentist counterpart: Fisher information

- The distribution of the **maximum likelihood estimate** $\hat{\boldsymbol{\theta}}$ is known to be asymptotically normal: $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ converges to a normal with mean 0 and covariance matrix $\mathcal{I}_N^{-1}(\boldsymbol{\theta})$, where

$$\mathcal{I}_N(\boldsymbol{\theta}) = -\left[\mathbf{E}_{\boldsymbol{\theta}}\left(\frac{\partial^2 \ell_N}{\partial\theta_i\,\partial\theta_j}\right)\right]_{1\le i,j\le N}$$

  is the **Fisher information matrix** (negative Hessian of log-likelihood at $\boldsymbol{\theta}$).

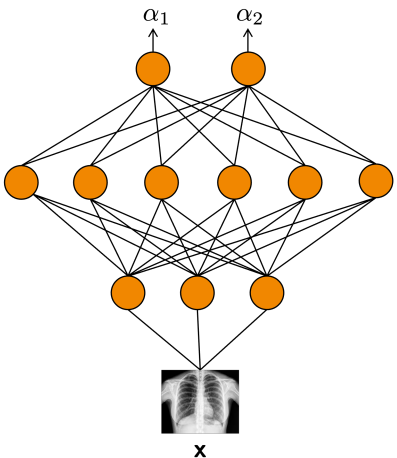
- The Fisher information matrix allows for constructing (approximate) **confidence regions** for $\boldsymbol{\theta}$ around the estimate $\hat{\boldsymbol{\theta}}$.

- Obviously, the larger this region, the higher the (approximation) uncertainty about the true model $\boldsymbol{\theta}$.

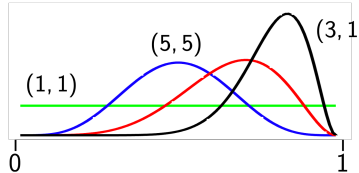
Agenda

1. Introduction and first-order uncertainty representation (Willem)
2. Conformal prediction (Willem)
3. **Second-order uncertainty representation**
   - The Bayesian approach
   - **Second-order loss minimisation**
   - Credal prediction
4. Uncertainty quantification and evaluation
5. Summary and outlook

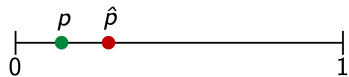
Second-order prediction through ERM (evidential deep learning)



$$D_{\boldsymbol{\alpha}}(p_1, \ldots, p_K) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{K} (p_i)^{\alpha_i - 1}$$



level 2 (epistemic)



level 1 (aleatoric)

negative positive

level 0 (outcome)

Second-order prediction through ERM

- Given training data $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N \subset \mathcal{X} \times \mathcal{Y}$, can we train a second-order predictor

$$H : \mathcal{X} \longrightarrow \mathbb{P}(\mathbb{P}(\mathcal{Y}))$$

via (variants of) **empirical risk minimisation** (ERM), i.e., by minimising

$$R_{emp}(H) = \sum_{i=1}^{N} L_E(H(\boldsymbol{x}_i), y_i) \,,$$

with a suitable **second-order (epistemic) loss function**

$$L_E : \mathbb{P}(\mathbb{P}(\mathcal{Y})) \times \mathcal{Y} \longrightarrow \mathbb{R} \,,$$

such that the predictor represents its epistemic uncertainty in a "faithful" way?

Representing epistemic uncertainty

- This approach is well motivated, as it works for learning (first-order) probability distributions, using **proper scoring rules**.
- Training a probabilistic predictor via **empirical risk minimisation**, i.e.,

$$h = \arg\min_{g \in \mathcal{H}} \sum_{i=1}^{N} L_A\left(g(\boldsymbol{x}_i), y_i\right) ,$$

  yields (theoretically) unbiased predictors if L_A is a (strictly) **proper scoring rule**, which incentivises the learner to predict the true $p = p(y \mid \boldsymbol{x})$.
- If the loss L_A is a (strictly) proper scoring rule, then the learner minimises expected loss (on samples $Y \sim p$) by predicting the true probability:

$$\underbrace{\mathbf{E}_{Y \sim p}[\, L_A(p, Y)\,]}_{S_1(p,p)} \leq \underbrace{\mathbf{E}_{Y \sim p}[\, L_A(\hat{p}, Y)\,]}_{S_1(\hat{p},p)} .$$

Representing epistemic uncertainty

- One way to define an epistemic loss L_E is in terms of an expectation over L_A:

$$L_E(Q, y) = \mathbf{E}_{p \sim Q} L_A(p, y) ,$$

- We obtained a series of **negative results**, showing that second-order loss minimisation is theoretically flawed, not only for losses of that kind but also in general (Bengs *et al.*, 2022, 2023; Jürgens *et al.*, 2024).

- Somewhat simplified, a loss L_E incentivising the learner to make meaningful second-order predictions Q **cannot exist**.

- In practice, epistemic uncertainty is imposed through **regularisation**:

$$L'_E(Q, y) = L_E(Q, y) + \lambda \times \underbrace{d_{KL}(Q, Q_0)}_{\text{deviation from uniform on } \mathbb{P}(\mathcal{Y})}$$

- While this appears undesirable, one has to recognise that epistemic uncertainty, by its very nature, cannot solely depend on the data, and arguably cannot be "objective" (compare with the need for **prior knowledge** in Bayesian inference).

Non-existence of second-order scoring rules

- A second-order loss L_E (such that $L_E(Q, \cdot)$ is $\mathbb{Q}(\mathcal{Y})$-quasi-integrable for all $Q \in \mathbb{Q}(\mathcal{Y})$) is a **proper second-order scoring rule** if, for all $\hat{Q}, Q \in \mathbb{Q}(\mathcal{Y})$,

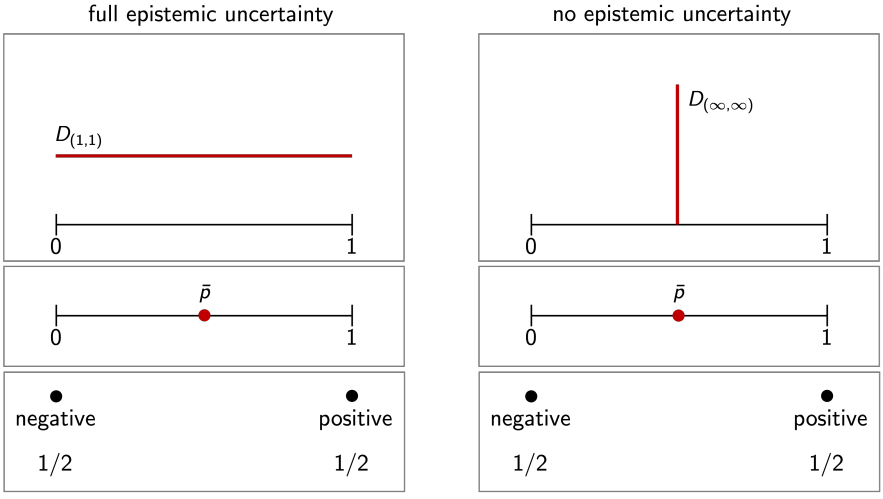
$$\underbrace{\mathbb{E}_{p\sim Q}\Big[\mathbb{E}_{Y\sim p}[\,L_E(Q, Y)\,]\Big]}_{S_2(Q,Q)} \leq \underbrace{\mathbb{E}_{p\sim Q}\Big[\mathbb{E}_{Y\sim p}[\,L_E(\hat{Q}, Y)\,]\Big]}_{S_2(\hat{Q},Q)}.$$

  If the learner holds "second-order belief" Q, and is penalised according to L_E, then it should report $\hat{Q} = Q$ as the (double-)expected loss-minimising prediction.

- **Theorem:** There exists no second-order loss L_E which is a proper second-order scoring rule (Bengs *et al.*, 2023).

- **Ambiguity** due to missing first-order data: Different epistemic states (distributions over first-order models) are compatible with the observed (ground-level) data.

Ambiguity due to lack of first-order data



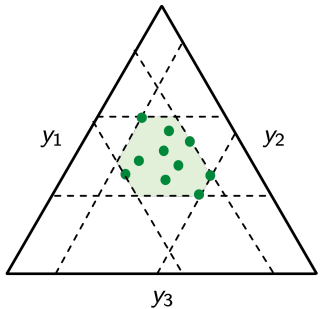
Agenda

1. Introduction and first-order uncertainty representation (Willem)
2. Conformal prediction (Willem)
3. **Second-order uncertainty representation**
   - The Bayesian approach
   - Second-order loss minimisation
   - **Credal prediction**
4. Uncertainty quantification and evaluation
5. Summary and outlook

Simple credal set prediction: the credal wrapper

- Wang *et al.* (2025) propose the **credal wrapper**:
  - train an ensemble $h_1, \ldots, h_M$ of probabilistic predictors, using any learner
  - given $\boldsymbol{x}$, produce predictions $h_1(\boldsymbol{x}), \ldots, h_M(\boldsymbol{x}) \in \Delta_K$
  - derive minimum and maximum probability $\underline{p}_k(\boldsymbol{x})$ and $\overline{p}_k(\boldsymbol{x})$ per class, $k = 1, \ldots, K$
  - predict the credal set

$$C(\boldsymbol{x}) = \left\{ p \in \Delta_K \mid \forall k : \underline{p}_k(\boldsymbol{x}) \leq p_k \leq \overline{p}_k(\boldsymbol{x}) \right\}.$$



Credal prediction based on relative likelihood (Löhr *et al.*, 2025)

- With likelihood function $L : \mathcal{H} \longrightarrow \mathbb{R}_+$, define **relative likelihood** as

$$\gamma(h) = \frac{L(h)}{\sup_{h \in \mathcal{H}} L(h)} \in [0, 1] \,.$$

- For $\alpha \in (0, 1]$, define the **plausible model set** (α-cut on relative likelihood)

$$\mathcal{C}_\alpha = \{\, h \in \mathcal{H} : \gamma(h) \geq \alpha \,\} \,.$$

- Given $\boldsymbol{x} \in \mathcal{X}$, the predictive image of $\mathcal{C}_\alpha$ is $\mathcal{Q}_{\boldsymbol{x},\alpha} = \{\, h(\boldsymbol{x}) \,|\, h \in \mathcal{C}_\alpha \,\} \subseteq \Delta_K$.
- A convenient class-wise summary of $\mathcal{Q}_{\boldsymbol{x},\alpha}$ is given by the **marginal extrema**

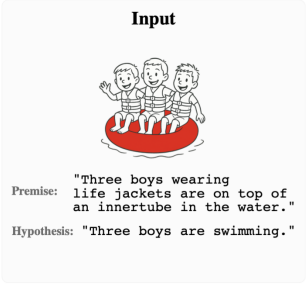
$$\underline{p}_k(\boldsymbol{x}) = \inf_{h \in \mathcal{C}_\alpha} h_k(\boldsymbol{x}), \qquad \overline{p}_k(\boldsymbol{x}) = \sup_{h \in \mathcal{C}_\alpha} h_k(\boldsymbol{x}), \quad k = 1, \ldots, K.$$

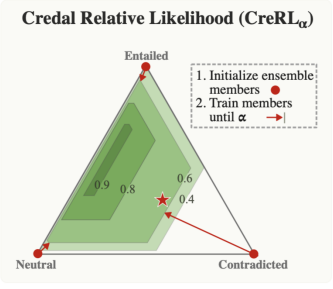
- Again, we then define the **box credal set** at $\boldsymbol{x}$ as

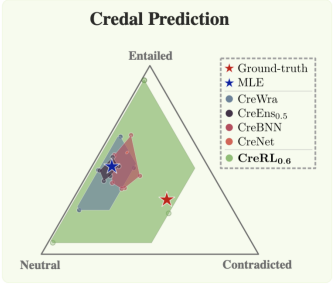
$$C(\boldsymbol{x}) = \Big\{\, p \in \Delta_K \,|\; \forall k : \underline{p}_k(\boldsymbol{x}) \leq p_k \leq \overline{p}_k(\boldsymbol{x}) \;\Big\}.$$

Credal prediction based on relative likelihood

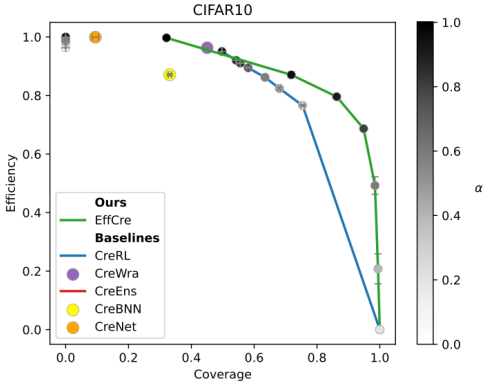
Example: ChaosNLI, three class classification problem: premise and hypothesis are: entailed, contradicted or neutral.







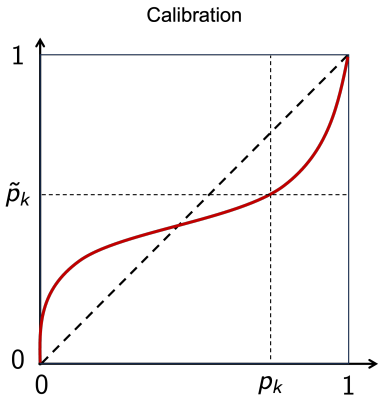
Credal prediction based on relative likelihood

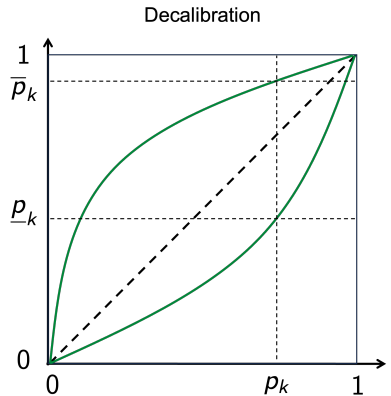


$$\text{Coverage}(\boldsymbol{x}) = \mathbb{I}(p(\cdot \mid \boldsymbol{x}) \in H(\boldsymbol{x}))\,, \qquad \text{Efficiency}(\boldsymbol{x}) = \frac{1}{K} \sum_{k=1}^{K} \overline{p}_k(\boldsymbol{x}) - \underline{p}_k(\boldsymbol{x})$$

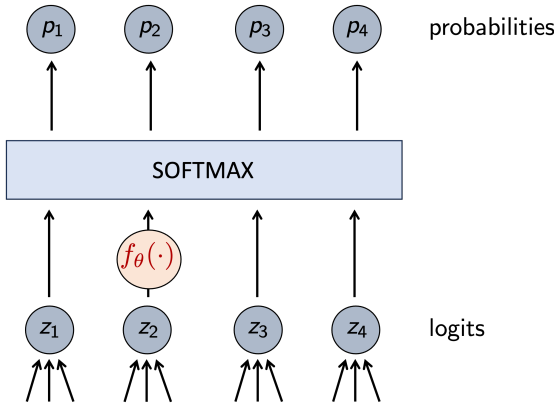
Efficient credal prediction through decalibration

- Hofman *et al.* (2025) propose to approximate the marginals through **decalibration**: how much can one deviate from maximum likelihood probabilities p_k, through mapping $h' = \text{dec}_k \circ h_{ML}$, without violating $\gamma(h') \geq \alpha$?

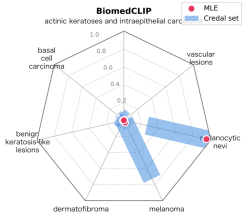


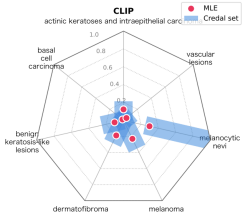


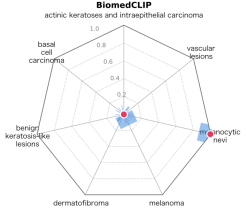
Efficient credal prediction through decalibration

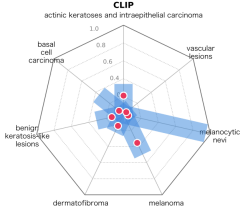


Credal prediction on DERMAMNIST









Comparison of credal sets for BIOMEDCLIP and CLIP on DERMAMNIST for a melanocytic nevi (above) and a melanoma (below).

Second-order methods

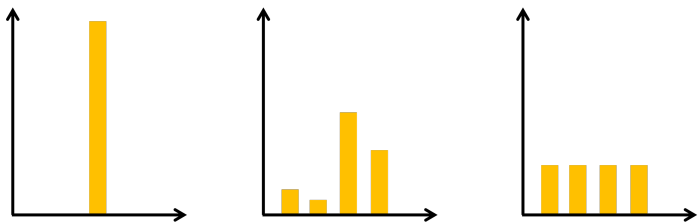
- (Approximate) Bayesian machine learning
  - BNNs
- Ensemble methods and pseudo-ensemble methods
  - Deep ensembles, Dropout, DropConnect, HyperDM
- Second-order loss minimisation
  - Evidential deep learning (ENN, PriorNN, PostNN, EDD)
- Evidential machine learning
  - DS theory, belief functions
- Direct Epistemic Uncertainty Prediction (DEUP)
  - Train additional model that predicts excess risk (gap to Bayes)
- Credal set prediction
  - Credal wrapper, credal BNN, naïve credal classifier, credal decision trees
- Interval-based methods
  - Interval parameters and/or interval predictions

Agenda

1. Introduction and first-order uncertainty representation (Willem)
2. Conformal prediction (Willem)
3. Second-order uncertainty representation
4. **Uncertainty quantification and evaluation**
   - Quantification
   - Evaluation
5. Summary and outlook

Uncertainty quantification

Shannon entropy: $H(p) = -\sum_i p_i \cdot \log(p_i)$



Shannon entropy: axioms

- Shannon entropy can be justified **axiomatically**.
- It is implied (up to a constant multiplicative factor) by the following properties:

  A1 **Continuity**: H is a continuous function

  A2 **Symmetry**: for any permutation τ,

  $$H(p_1, \ldots, p_K) = H(p_{\tau(1)}, \ldots, p_{\tau(K)})$$

  A3 **Additivity**: for all distributions p, p' and their product $p \times p'$,

  $$H(p \times p') = H(p) + H(p')$$

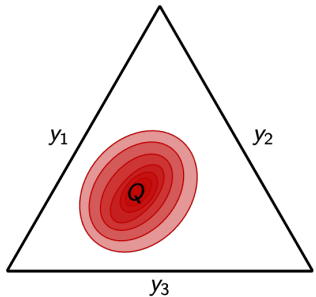
  A4 **Sub-additivity**: for any joint distribution p with marginals p' and p'',

  $$H(p) \leq H(p') + H(p'')$$

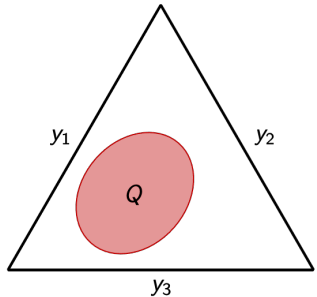
Uncertainty quantification

- **Uncertainty quantification** (UQ) seeks to measure the amount of total, aleatoric, and epistemic uncertainty of a prediction Q in terms of (axiomatically justified) numerical measures, ideally such that

$$\mathrm{TU}(Q) = \mathrm{AU}(Q) + \mathrm{EU}(Q)\,.$$



$\mathrm{TU} = 0.6$, $\mathrm{AU} = 0.2$, $\mathrm{EU} = 0.4$



$\mathrm{TU} = 0.5$, $\mathrm{AU} = 0.2$, $\mathrm{EU} = 0.3$

More axioms (Wimmer *et al.*, 2023; Sale *et al.*, 2024)

- A0 TU, AU, and EU are non-negative
- A1 $\mathrm{AU}(\delta_{\mathrm{Unif}(\mathcal{Y})}) \geq \mathrm{AU}(\delta_p) \geq \mathrm{AU}(\delta_{\delta_y}) = 0$ holds for any $y \in \mathcal{Y}$ and $p \in \mathbb{P}(\mathcal{Y})$
- A2 $\mathrm{EU}(Q) \geq \mathrm{EU}(\delta_p) = 0$ holds for any $Q \in \mathbb{P}(\mathbb{P}(\mathcal{Y}))$ and $p \in \mathbb{P}(\mathcal{Y})$; further, $\mathrm{AU}(Q) = 0$ implies $\mathrm{EU}(Q') \geq \mathrm{EU}(Q)$ for Q' such that $Q'(\delta_y) = \frac{1}{K}$ for all $y \in \mathcal{Y}$
- A3 $\mathrm{AU}(Q) \leq \mathrm{TU}(Q)$ and $\mathrm{EU}(Q) \leq \mathrm{TU}(Q)$ holds for any $Q \in \mathbb{P}(\mathbb{P}(\mathcal{Y}))$
- A4 $\mathrm{TU}(Q)$ is maximal for Q being the continuous second-order uniform distribution
- A5 If Q' is a mean-preserving spread of Q, then $\mathrm{EU}(Q') \geq \mathrm{EU}(Q)$ (weak version) or $\mathrm{EU}(Q') > \mathrm{EU}(Q)$ (strict version)
- A6 If Q' is a spread-preserving location shift of Q, then $\mathrm{EU}(Q') = \mathrm{EU}(Q)$
- A7 $\mathrm{TU}_{\mathcal{Y}}(Q) \leq \mathrm{TU}_{\mathcal{Y}_1}(Q_{|\mathcal{Y}_1}) + \mathrm{TU}_{\mathcal{Y}_2}(Q_{|\mathcal{Y}_2})$
- A8 $\mathrm{TU}_{\mathcal{Y}}(Q_{|\mathcal{Y}_1} \otimes Q_{|\mathcal{Y}_2}) = \mathrm{TU}_{\mathcal{Y}_1}(Q_{|\mathcal{Y}_1}) + \mathrm{TU}_{\mathcal{Y}_2}(Q_{|\mathcal{Y}_2})$, with $\otimes$ the product measure

Uncertainty quantification

- Common approach for second-order predictions $Q = H(\boldsymbol{x}) \in \mathbb{P}(\mathbb{P}(\mathcal{Y}))$, treating first-order predictions $p \in \mathbb{P}(\mathcal{Y})$ as random variables distributed according to Q:

$$Y \sim p \sim Q$$

  - TU = **Shannon entropy** of the probabilistic prediction $Y \sim \bar{p}$, where $\bar{p}$ is the predictive distribution (averaged over models):

$$\mathrm{TU} = \mathrm{ENT}(Y) = \mathrm{ENT}(\bar{p}) = \mathrm{ENT}\left(\int p \, dQ(p)\right)$$

  - AU = **conditional entropy** (of prediction given model):

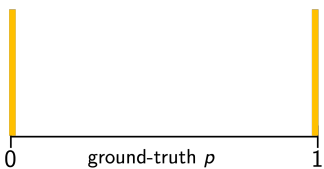
$$\mathrm{AU} = \mathrm{ENT}(Y \mid P) = \int \mathrm{ENT}(p) \, dQ(p)$$

  - EU = **mutual information** $I(Y, P) = \mathrm{ENT}(Y) - \mathrm{ENT}(Y \mid P)$.

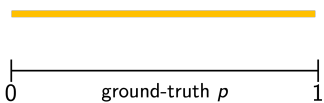
Violation of axiomatic properties

- This approach has recently been criticised by Wimmer *et al.* (2023).
- For example, when should EU be highest?

Wimmer *et al.* (2023)



Mutual information is maximal,
but EU should (arguably) not



Mutual information is lower, but
EU should (arguably) be maximal

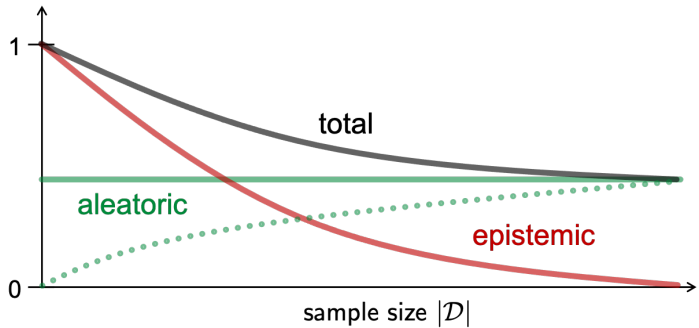
Disentangling aleatoric and epistemic uncertainty

- The **distinction between aleatoric and epistemic uncertainty** can be difficult.
- **Predict the next number**: 116, 304, 194, 341, 224, 654, 609, 625, 533, 91, 205, 35, 527, 611, 128, 235, 348, 912, 582, 52, 672, 20, 856, 904, 628, 273, 615, 105, 610, 862, 384, 705, 73, 794, 775, 156, **??**

$$x \leftarrow x \times 237 \bmod 971$$

- **Epistemic uncertainty implies uncertainty about** the data-generating process, and hence about the (true) **aleatoric uncertainty**.
- Hence precise quantification of AU is not possible unless EU $= 0$ ...

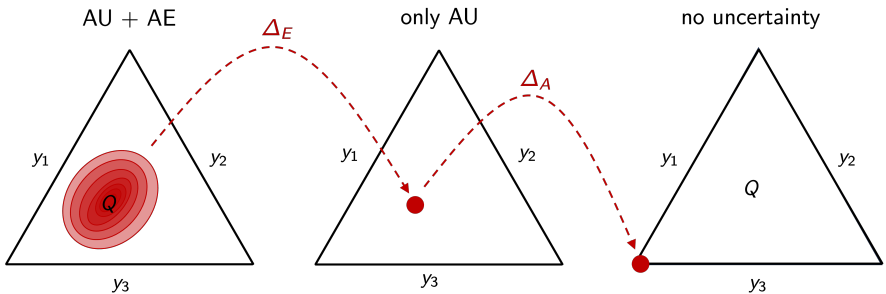
Additive decomposition



Qualitative behaviour of total, aleatoric, and epistemic uncertainty depending on the sample size. Dotted line = total − epistemic uncertainty (= aleatoric uncertainty when stipulating an additive decomposition).

Second-order uncertainty quantification

- Sale *et al.* (2024) proposed an alternative approach based on the notion of (Wasserstein) **distance**: How much transport is needed to turn a second-order distribution into a distribution representing
  - (i) no epistemic and
  - (ii) no aleatoric uncertainty?



Second-order uncertainty quantification

- Another proposal by Hofman *et al.* (2024) is **loss-based**, namely, based on the decomposition of **proper scoring rules** ℓ into a **divergence** (from ground-truth q) and an **entropy** term:

$$L_\ell(p, q) = \mathbf{E}_{Y\sim q}\, \ell(p, Y) = \underbrace{D_\ell(p, q)}_{\text{epistemic}} + \underbrace{H_\ell(q, q)}_{\text{aleatoric}}$$

- For a **second-order distribution** Q, it is sensible to define

$$\begin{aligned}\mathrm{EU}(Q) &= \mathbf{E}_{q\sim Q}[D_\ell(\bar{q}, q)] = \mathbf{E}_{q\sim Q}[L_\ell(\bar{q}, q) - L_\ell(q, q)] \\ &= \underbrace{\mathbf{E}_{q\sim Q}[L_\ell(\bar{q}, q)]}_{\mathrm{TU}(Q)} - \underbrace{\mathbf{E}_{q\sim Q}[L_\ell(q, q)]}_{\mathrm{AU}(Q)}\,.\end{aligned}$$

- EU is the **gain** (loss reduction) to be expect when predicting, not on the basis of the uncertain Q, but only after being revealed the true q.
- ℓ = log-loss recovers entropy-based measures

Measures for credal sets: axiomatics

- Basic properties that a measure of uncertainty for credal sets should satisfy:

  A1 **Non-negativity, range**: U is non-negative and upper-bounded by some value $r \in \mathbb{R}$, for example $r = \log(K)$, which is assumed for $Q = \Delta_K$ (the case of complete ignorance).

  A2 **Continuity**: U is a continuous functional.

  A3 **Monotonicity**: If $Q \subseteq Q'$ for credal sets Q, Q', then $U(Q) \leq U(Q')$.

  A4 **Probability consistency**: U reduces to standard Shannon entropy in the case where Q reduces to a single probability distribution.

  A5 **Sub-additivity**: For a (joint) credal set Q on a product space $\mathcal{Y}' \times \mathcal{Y}''$ with marginals Q' resp. Q'',

  $$\mathcal{Y}eleq : subad\, U(Q) \leq U(Q') + U(Q'')\,. \qquad (\star)$$

  A6 **Additivity**: The inequality (**??**) is an equality in the case where Q' and Q'' are independent.

Measures for credal sets

- A well-founded generalisation of entropy and natural measure of **total uncertainty** represented by a credal set is the **upper entropy**:

$$H^*(Q) := \max_{q \in Q} H(q)$$

- A well-founded measure of **epistemic uncertainty** is the **generalised Hartley measure**

$$\mathrm{GH}(Q) := \sum_{A \subseteq \mathcal{Y}} \mathrm{m}_Q(A) \log(|A|)\,,$$

  which extends the Hartley measure ($\log(|A|)$) from sets to graded sets.

- Although an equally well-justified measure of **aleatoric uncertainty** (conflict) in the form of an extension of Shannon entropy has not been found so far (Klir, 2005), the **lower entropy** is a natural measure of **irreducible uncertainty**:

$$H_*(Q) := \min_{q \in Q} H(q)$$

Disaggregation

- There is no **additive decomposition**

$$\mathrm{TU}(Q) = \mathrm{AU}(Q) + \mathrm{EU}(Q)$$

  such that all three measures behave well.

- Idea: Fix two "good" measures and **derive** the third one in terms of the **difference**

$$H^*(Q) = \underbrace{(H^*(Q) - \mathrm{GH}(Q))}_{\text{generalised entropy}} + \mathrm{GH}(Q)$$

$$H^*(Q) = H_*(Q) + (H^*(Q) - H_*(Q))$$

- Empirically, the derived measures show relatively poor performance

Agenda

1. Introduction and first-order uncertainty representation (Willem)
2. Conformal prediction (Willem)
3. Second-order uncertainty representation
4. **Uncertainty quantification and evaluation**
   - Quantification
   - **Evaluation**
5. Summary and outlook

Evaluation of uncertainty quantification

- Evaluation of uncertainty representation/quantification is difficult, mainly due to a **lack of a ground truth**.
- Formally, uncertainty measures can be justified in terms of suitable **axioms**.
- Empirically, one often considers how they help improve the performance of a machine learning pipeline in **downstream tasks**:
  - Selective classification
  - Out-of-distribution (OoD) detection
  - Active learning

Selective prediction

- In **selective prediction**, the learner can abstain from making a prediction on some inputs of its choice
- Consider a test set $\mathcal{D}_{\text{test}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$, with Q_i the prediction for $\boldsymbol{x}_i$
- For $\alpha \in [0, 1]$, let $k = \lfloor \alpha N \rfloor$ be a (fixed) **rejection level**
- Given measure U, define the permutation π of $\{1, \ldots, N\}$ through

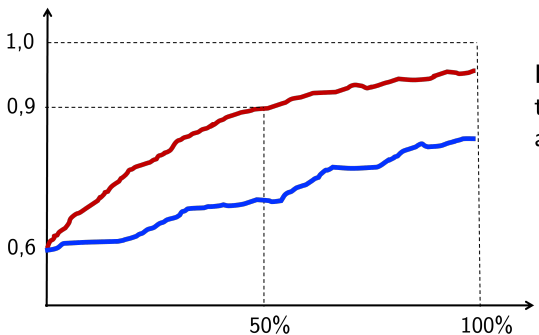
$$U(Q_{\pi(1)}) \leq U(Q_{\pi(2)}) \leq \cdots \leq U(Q_{\pi(N)})$$

- The **accuracy-rejection curve** is then defined as

$$\alpha \mapsto \frac{1}{\lfloor (1-\alpha)N \rfloor} \sum_{j=1}^{\lfloor (1-\alpha)n \rfloor} \text{acc}(q_{\pi(j)}, y_{\pi(j)})$$

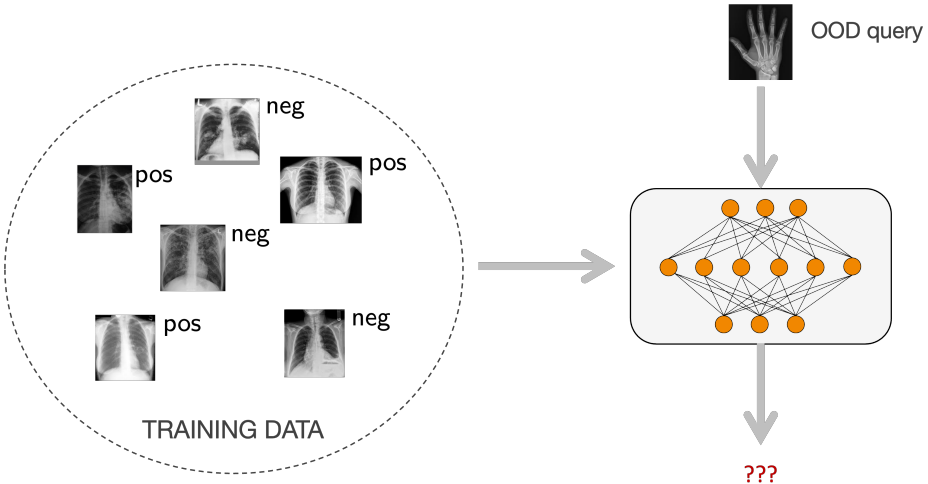
Empirical evaluation

- **Accuracy-rejection curves**: Allow the learner to reject the $\alpha\,\%$ presumably most uncertain test cases and measure accuracy only on the remaining ones



If allowed to reject 50% of the cases, the red learner manages to increase accuracy from 0,6 to 0,9

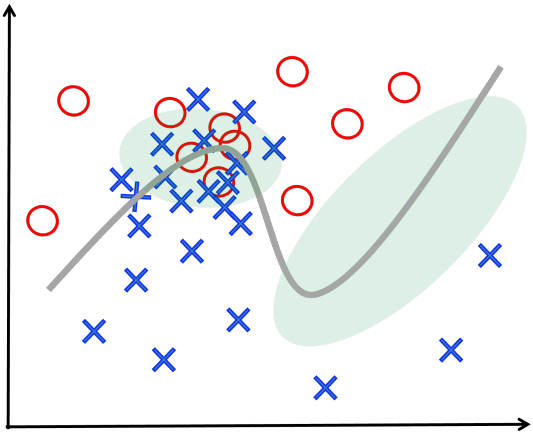
Out-of-distribution data: "I don't know"



Active learning

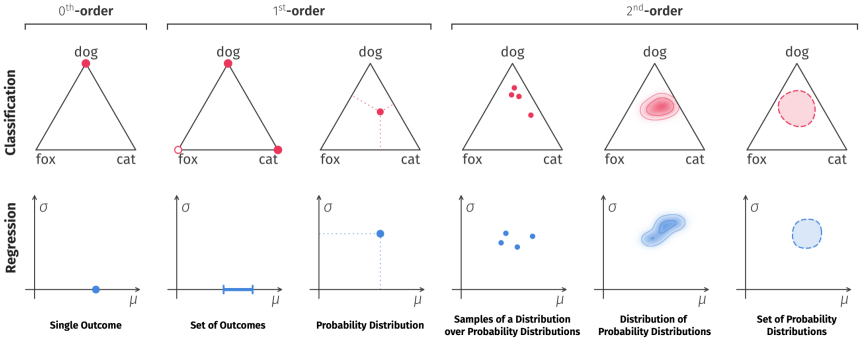
- **Active learning** is another popular downstream task
- The goal is to identify examples for which **supervision is most beneficial**, thereby reaching strong performance with as few labels as possible
- Starting with a (small) labeled seed set, the learner iteratively selects unlabeled examples to be annotated by an oracle (and retrains on the augmented data)
- **Epistemic uncertainty** should arguably be a key criterion (Nguyen *et al.*, 2022) — note the connection to bandits (unlike epistemic uncertainty, aleatoric uncertainty cannot be reduced by more supervision)

Active learning



probly

Uncertainty Representation and Quantification for Machine Learning



Any model: point predictions

Conformal: LAC, APS, RAPS, SAPS, …

Calibration: temp., vec., label relaxation, …

Sampling: dropout, ensembles, BNNs, …

Distributions: evidential, Gauss. heads, Laplace, …, ensembles, BNNs, …

Credal: credal ensembling, wrapper, relative likelihood, …

probly Uncertainty Representation and Quantification for Machine Learning

```python
import probly.transformation

# any model -> uncertainty
model = credal_ensembling(model)
train(model, data) # your logic
```

```python
import probly.representer

# build representations
rep = representer(model)
out = rep.represent(data)
```

```python
from probly.quantification

# TU / EU / AU
d = decompose(out)
d['TU'] == d['EU'] + d['AU']
```

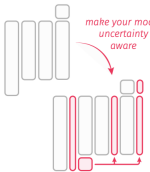
```python
import probly.decider

# bayes-optimal label
y = categorical_from_mean(out)
y = categorical_from_maximin(out)
```

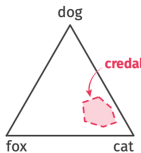
```python
import probly.datasets

# easily get first-order data
data = CIFAR10H()
data = ImageNetReaL()
```

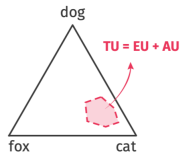
1 **transform** → 2 **represent** → 3 **quantify** **decide** → 4 **evaluate**



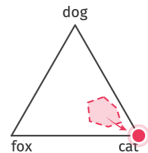
library-agnostic:
works natively with *torch*, *flax*, *hugginface*, *torch-uncertainty*, *sklearn*, *river*, ...



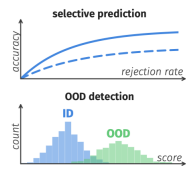
composable representations:
utilize *DirichletDistribution*, *ConvexCredalSet*, *Sample*, *OneHotConformalSet*, ...



common decompositions:
quantify the uncertainty and decompose it into *total*, *epistemic*, and *aleatoric*



canonical decisions:
use representations to make decisions or use your *own* logic



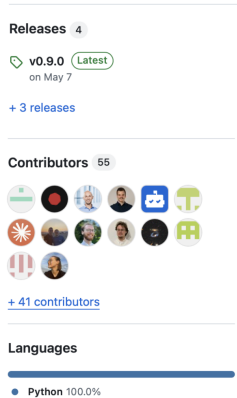
evaluation pipelines:
run experiments on *selective prediction*, *active learning*, *1st order data*, ...

probly Uncertainty Representation and Quantification for Machine Learning

`pip install probly` — `uv add probly`

arXiv **Coming soon!**

Star on GitHub

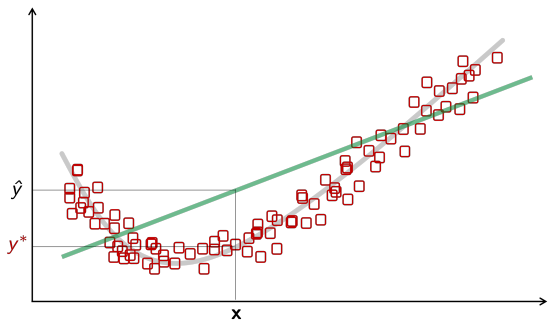


Summary and outlook

- **Learning reliable predictors** that represent their uncertainty in a faithful way is an important task, but also challenging, both conceptually and computationally
- **Distinguishing different types of uncertainty**, aleatoric and epistemic, is useful, though it seems that second-order uncertainty is hard to tackle
- **Quantifying predictive uncertainty** in a theoretically sound manner, and disentangling total into aleatoric and epistemic uncertainty, is difficult, too
- Various other **open problems**: model uncertainty, generalised settings (eg., OOD data), uncertainty in GenAI and LLMs, evaluation, other forms of uncertainty (e.g., in the data), applications, etc.

Model misspecification is an open issue

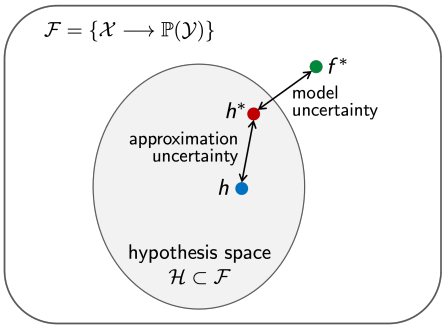
- Uncertainty due to **bias** can be both, aleatoric/irreducible (model class fixed) or epistemic/reducible (model class can be changed/extended)



What uncertainty should the learner report at $\mathbf{x}$?

Model misspecification is an open issue

- Existing methods (implicitly) assume that the model class covers the true data-generating process, handling **model misspecification** is an open problem



- A related problem is caused by an **implicit bias** due to **regularisation** (Jimenez *et al.*, 2025).

References

V. Bengs, E. H., and W. Waegeman. Pitfalls of epistemic uncertainty quantification through loss minimisation. In *Proc. NeurIPS*, 2022.

V. Bengs, E. H., and W. Waegeman. On second-order scoring rules for epistemic uncertainty quantification. In *Proc. ICML*, 2023.

E. H. and W. Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3), 2021.

P. Hofman, Y. Sake, and E. H. Quantifying aleatoric and epistemic uncertainty with proper scoring rules. *arXiv preprint arXiv:2404.12215*, 2024.

P. Hofman, T. Löhr, M. Muschalik, Y. Sale, and E. Hüllermeier. Efficient credal prediction through decalibration. In *Proc. ICLR*, 2025.

S. Jimenez, M. Jürgens, and W. Waegeman. Why machine learning models fail to fully capture epistemic uncertainty, 2025.

M. Jürgens, V. Bengs, N. Meinert, E. H., and W. Waegeman. Is epistemic uncertainty faithfully represented by evidential deep learning methods?, 2024. ICML 2024.

G.J. Klir. *Uncertainty and Information: Foundations of Generalized Information Theory*. Wiley, 2005.

B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proc. NeurIPS, 31st Conf. on Neural Information Processing Systems*, Long Beach, California, USA, 2017.

T. Löhr, P. Hofman, F. Mohr, and E. H. Credal prediction based on relative likelihood. In *Proc. NeurIPS*, volume 38, San Diego, CA, USA, 2025.

V.L. Nguyen, M.H. Shaker, and E. H. How to measure uncertainty in uncertainty sampling for active learning. *Machine Learning*, 111(1):89–122, 2022.

Y. Sale, V. Bengs, M. Caprio, and E. Hüllermeier. Second-order uncertainty quantification: A distance-based approach. In *ICML*, 2024.

O. Shorinwa, Z. Mei, J. Lidard, A.Z. Ren, and A. Majumdar. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges and future directions, 2024.

K. Wang, F. Cuzzolin, K. Shariatmadar, D. Moens, and H. Hallez. Credal wrapper of model averaging for uncertainty estimation in classification. In *Proc. ICLR, International Conference on Learning Representations*, Singapore, 2025.

L. Wimmer, Y. Sale, P. Hofmann, B. Bischl, and E. H. Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures? In *UAI*, 2023.