On the Challenge of Quantifying Epistemic Uncertainty in Machine Learning

Eyke Hüllermeier

Artificial Intelligence and Machine Learning
Institute of Informatics, LMU Munich
Munich Center for Machine Learning (MCML)

SIPTA Talk, June 19, 2024

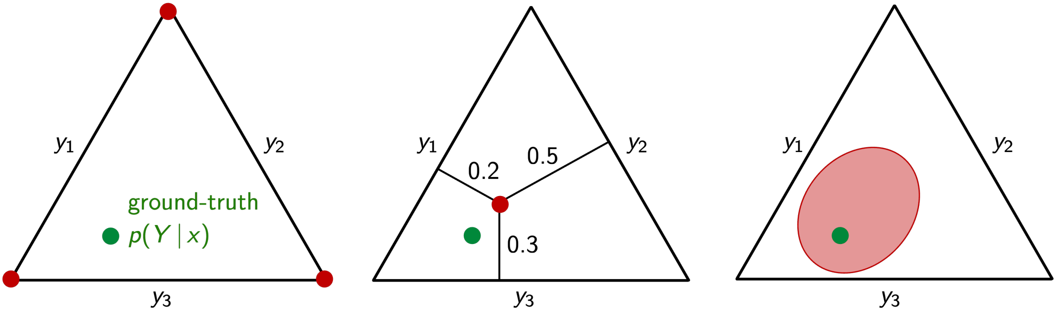
joint work with
Viktor Bengs, Paul Hofman, Mira Jürgens, Nis Meinert, Yusuf Sale, Willem Waegeman

Lack of uncertainty-awareness of ML systems

- Predictions by state-of-the-art neural network (Jia *et al.*, 2020): For the left image, the network predicts "umbrella" with confidence 97 %, for the right image "skirt" with confidence 96 %.

Uncertainty representation and levels of uncertainty-awareness

$$\mathcal{Y} = \{y_1, y_2, y_3\}, \text{ e.g. } \{\text{win, loss, tie}\}$$



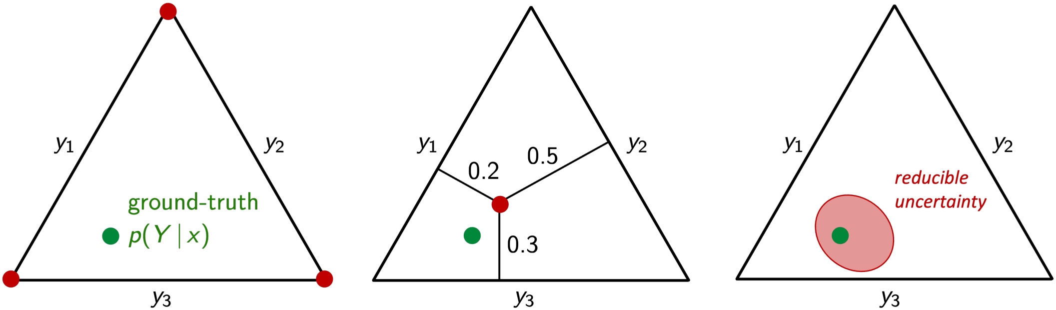
Deterministic predictor
$h : \mathcal{X} \longrightarrow \mathcal{Y}$

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

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

Uncertainty representation and levels of uncertainty-awareness

$$\mathcal{Y} = \{y_1, y_2, y_3\}, \text{ e.g. } \{\text{win}, \text{loss}, \text{tie}\}$$



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

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

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

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**.

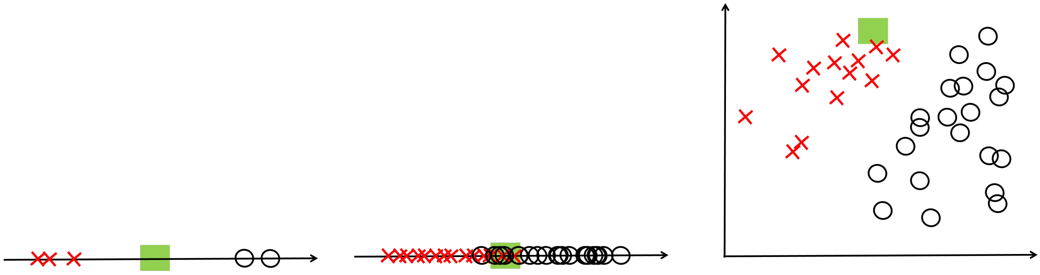
- **Epistemic** (systematic) uncertainty
  - refers to uncertainty caused by a **lack of knowledge**, i.e.,
  - to the epistemic state of the **agent** (e.g., learning algorithm),
  - can in principle be **reduced** on the basis of additional information.

Aleatoric versus epistemic uncertainty in ML

- Both types of uncertainty also play an important role in ML ...

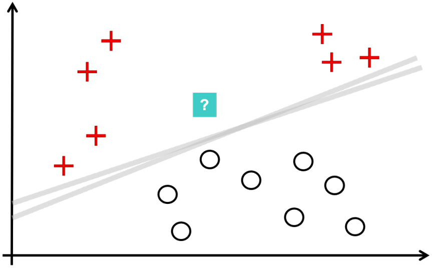
Aleatoric versus epistemic uncertainty in ML

- Both types of uncertainty also play an important role in ML ...

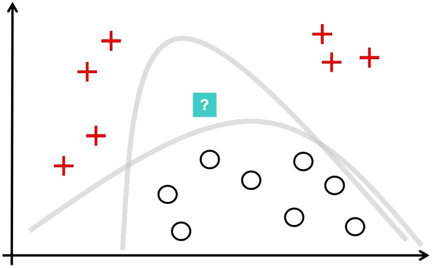


Aleatoric versus epistemic uncertainty in ML

- Uncertainty also depends on the underlying model assumptions:



strong prior (linear model)



weaker prior (nonlinear model)

Predictive uncertainty

- In the standard setting of supervised learning, we are mainly interested in (per-instance) **predictive uncertainty**: Instead of a **deterministic prediction**

$$\hat{y} = h(\boldsymbol{x}) \in \mathcal{Y}$$

of the outcome for a query instance $\boldsymbol{x}$, and also going beyond a (first-order) **probabilistic prediction**

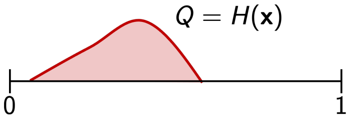
$$\hat{p} = h(\boldsymbol{x}) \in \mathfrak{P}(\mathcal{Y}),$$

we seek a **second-order prediction**

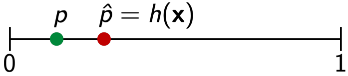
$$Q = H(\boldsymbol{x}) \in \mathfrak{Q}(\mathcal{Y})$$

adequately representing the learner's **epistemic uncertainty** about the prediction.

Uncertainty representation and levels of uncertainty-awareness



second-order predictor $H : \mathcal{X} \longrightarrow \mathfrak{Q}(\mathcal{Y})$

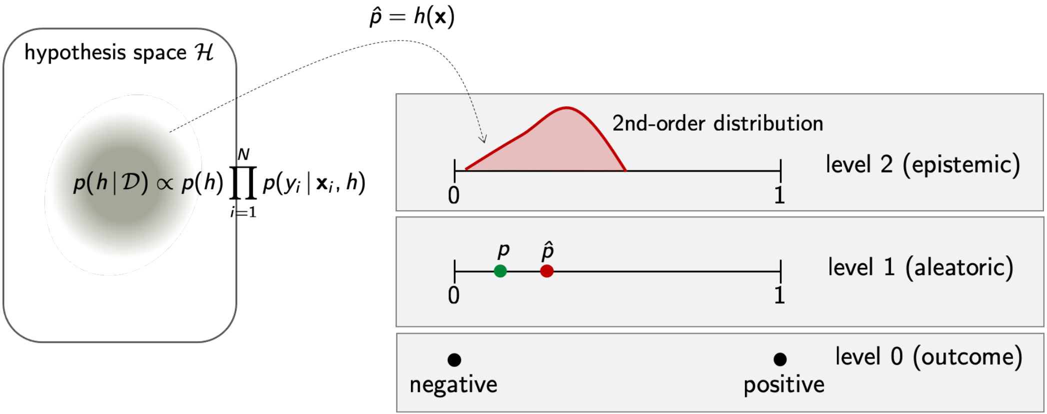


probabilistic predictor $h : \mathcal{X} \longrightarrow \mathfrak{P}(\mathcal{Y})$

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

The Bayesian approach: posterior predictive distribution

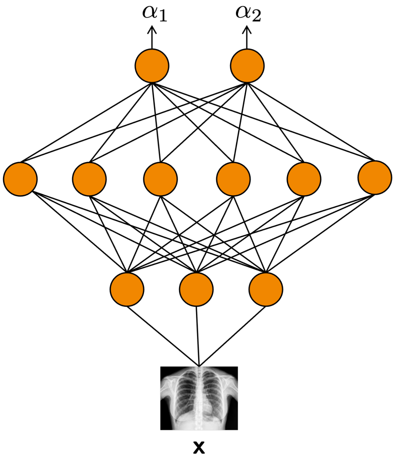
- Model uncertainty translates into predictive uncertainty:



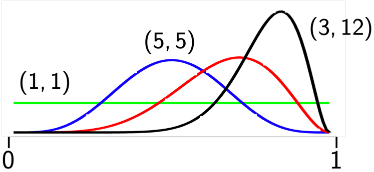
Agenda

1. Introduction
2. **Second-order loss minimisation**
3. Uncertainty quantification
4. Conclusion

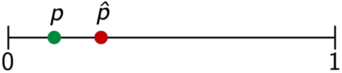
Direct second-order prediction (evidential deep learning)



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



level 2 (epistemic)



level 1 (aleatoric)

level 0 (outcome)

Direct second-order prediction

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

$$H : \mathcal{X} \longrightarrow \mathfrak{Q}(\mathcal{Y})$$

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

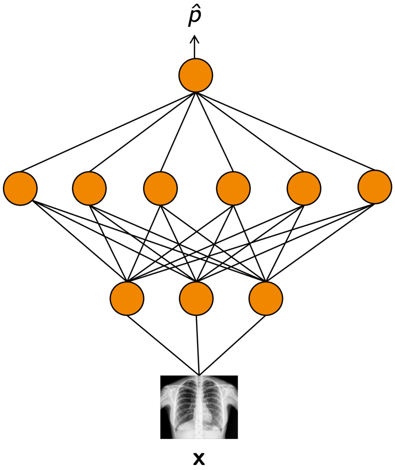
$$R_{emp}(H) = \sum_{i=1}^{N} L_E\big(H(\boldsymbol{x}_i), y_i\big) \, ,$$

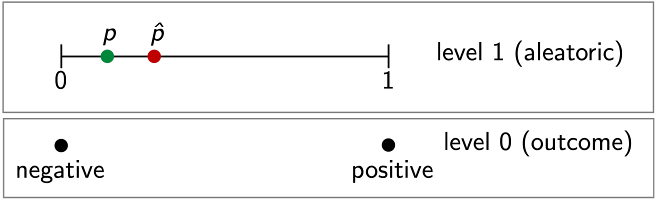
with a suitable **second-order (epistemic) loss function**

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

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

The case of first-order predictions





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 good (unbiased) predictors if L_A is a (strictly) **proper scoring rule**, which incentivises the learner to predict the true $p(y \mid \boldsymbol{x})$.

- Loss $L_A : \mathfrak{P}(\mathcal{Y}) \times \mathcal{Y} \longrightarrow \mathbb{R}$ is a **strictly proper scoring rule** if the expected loss minimiser is unique and coincides with the true probability p:

$$p = \arg\min_{\hat{p}} \mathbb{E}_{Y \sim p} L_A(\hat{p}, Y)$$

- Examples include cross-entropy and Brier score.

Direct epistemic uncertainty prediction

- Several authors proposed to penalise a **second-order prediction** Q in terms of the **expected first-order loss**, lifting L_A to the epistemic level as follows:

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

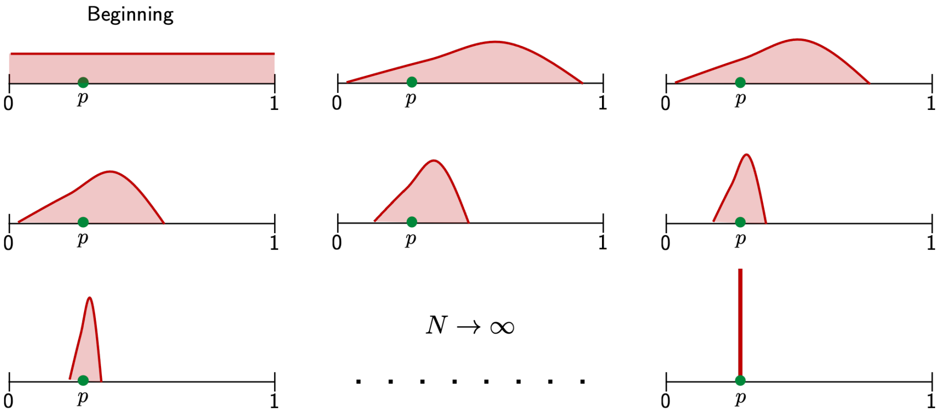
- Besides, a **regularised version** has been proposed:

$$L_E\big(Q, y\big) = \mathbb{E}_{p \sim Q}\, L_A\,(p, y) + \lambda\, \underbrace{d_{KL}\,(Q, Q_0)}_{R(Q)}$$

- This leads to the empirical risk

$$R_{emp}(H) = \sum_{i=1}^{N} \underbrace{\mathbb{E}_{p \sim H(\boldsymbol{x}_i)}\, L_A(p, y_i) + \lambda\, R(H(\boldsymbol{x}_i))}_{L_E(H(\boldsymbol{x}_i), y_i)}\,. \qquad (\star)$$

Appropriate second-order losses: Uncertainty gradually decreases



Non-existence of second-order scoring rules

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

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

  If the learner holds "second-order believe" 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).

Degenerate second-order predictions

- **Theorem** (Jürgens *et al.*, 2024): Let $L_A : \mathcal{Y} \times \mathfrak{P}(\mathcal{Y}) \longrightarrow \mathbb{R}$ be convex in its first argument, and let the second-order hypothesis space have a universal approximation property. Then the minimiser of the empirical risk $(\star)$ with $\lambda = 0$ is a second-order predictor H such that

$$H(\boldsymbol{x}_i) = \delta(p(\boldsymbol{x}_i))$$

  for all $i = 1, \ldots, N$, where δ is the Dirac delta function.

- In other words, the second-order predictor pretends **zero (epistemic) uncertainty** and effectively reduces to a first-order predictor.

- Deviation from this behavior requires $\lambda > 0$ and is then solely due to the **regulariser**, which makes the representation of (epistemic) uncertainty largely **arbitrary**.

Inner loss minimisation

- An alternative approach, more commonly used in (count) regression analysis, is to take the expectation inside (rather than outside) the first-order loss:

$$R_{emp}(H) = \sum_{i=1}^{N} L_A\Big(\underbrace{\mathbb{E}_{p\sim H(\boldsymbol{x}_i)}\, p(\boldsymbol{x}_i)}_{\bar{p}(\boldsymbol{x}_i)}, y_i\Big) + \lambda\, R\big(H(\boldsymbol{x}_i)\big) \qquad (\star\star)$$

- Essentially, first- and second-order spaces are now combined into a single hypothesis space: each combination of $p \in \mathcal{P} \subset \mathfrak{P}(\mathcal{Y})$ and $Q \in \mathcal{Q} \subset \mathfrak{Q}(\mathcal{Y})$ gives rise to a (predictive) distribution $\bar{p} \in \bar{\mathcal{P}} \subset \mathfrak{P}(\mathcal{Y})$ defined as

$$\bar{p} : \boldsymbol{x} \mapsto \mathbb{E}_{p\sim Q}\, p\,.$$

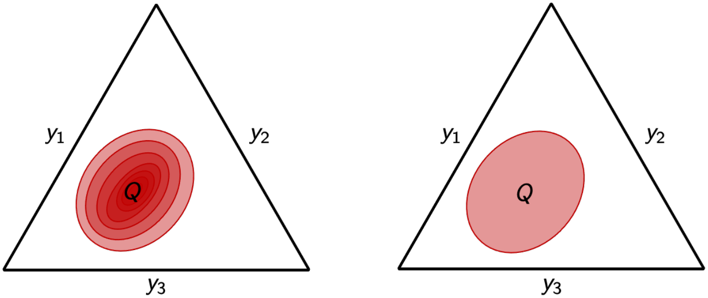
- Leads to similar problems of **non-identifiability** and/or **non-uniqueness** (Jürgens *et al.*, 2024).

Agenda

Uncertainty quantification

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

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



$\mathrm{TU} = 0.6$, $\mathrm{AU} = 0.2$, $\mathrm{EU} = 0.4$ $\quad$ $\mathrm{TU} = 0.5$, $\mathrm{AU} = 0.2$, $\mathrm{EU} = 0.3$

Uncertainty quantification

- Common approach for second-order predictions $Q = H(\boldsymbol{x}) \in \mathfrak{P}(\mathfrak{P}(\mathcal{Y}))$, treating first-order predictions $p \in \mathfrak{P}(\mathcal{Y})$ 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):

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

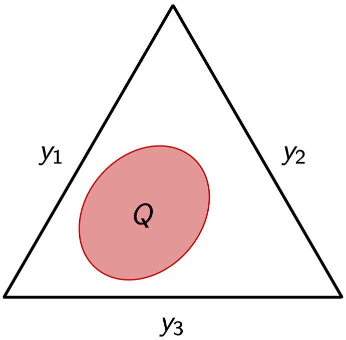
  - EU = **mutual information** $I(Y, P) = \mathrm{ENT}(Y) - \mathrm{ENT}(Y \,|\, P)$.

Second-order uncertainty quantification

- Recently criticised by Wimmer *et al.* (2023) ...
- 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?
- Another proposal by Hofman *et al.* (2024) is **loss-based**, namely, based on the decomposition of **proper scoring rules** ϕ into a **divergence** (from ground-truth q) and an **entropy** term:

$$s(p,q) = \mathbb{E}_{Y\sim q}\,\phi(p,Y) = \underbrace{\mathrm{DIV}_{\phi}(p,q)}_{\text{epistemic}} + \underbrace{\mathrm{ENT}_{\phi}(q,q)}_{\text{aleatoric}}$$

Alternative formalisms: credal sets



- A credal set Q captures both **aleatoric** and **epistemic uncertainty**, also called **conflict** and **nonspecificity**, respectively.
- How to learn credal predictors (producing credal sets as predictions)?
- How to quantify the **total uncertainty** represented by a credal set, and how to disaggregate it into **aleatoric** and **epistemic uncertainty**?

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

See e.g. Klir (2005) or more recent work by H. *et al.* (2022) and Sale *et al.* (2023).

Measures of total, aleatoric, and epistemic uncertainty

- A well-founded generalisation of entropy and natural measure of **total uncertainty** is the **upper entropy**:

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

- 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 $H(A) := \log(|A|)$ 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**:

$$S_*(Q) := \min_{q \in Q} S(q)$$

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**.

$$S^*(Q) = \big(\underbrace{S^*(Q) - \mathrm{GH}(Q)}_{\mathrm{GS}(Q)}\big) + \mathrm{GH}(Q)$$

$$S^*(Q) = S_*(Q) + \big(S^*(Q) - S_*(Q)\big)$$

- H. *et al.* (2022) provide a **critical discussion** of such decompositions (in an ML context) and show that derived measures show poor **empirical performance**.

Disaggregation

- In the case of second-order distributions, EU = entropy − conditional entropy can be interpreted as (expected) **gain in terms of log-loss reduction** (because $\mathrm{ENT}(p) = \mathbb{E}_{Y \sim p} L(p, Y)$ with $L =$ log-loss).
- We generalise this principle to **credal sets** and define EU as **maximal gain**:

$$\mathrm{EU}(Q) = \max_{p, p' \in Q} D_L(p, p') \, ,$$

  with the L-divergence

$$D_L(p, p') = \mathbb{E}_{Y \sim p} \left\{ L(p', Y) - L(p, Y) \right\} \, .$$

- We further define **aleatoric uncertainty** in terms of lower and upper bounds on the L-entropy $H_L(p) = \mathbb{E}_{Y \sim p} L(p, Y)$:

$$\underline{\mathrm{AU}}(Q) = \inf_{p \in Q} H_L(p) \, , \quad \overline{\mathrm{AU}}(Q) = \sup_{p \in Q} H_L(p) \, .$$

Disaggregation

Losses and their decomposition into aleatoric and epistemic uncertainty:

Loss	Aleatoric (upper\lower)	Epistemic
log-loss	$\sup_{p \in Q} \backslash \inf_{p \in Q} \; S(p)$	$\max_{p, p' \in Q} \; D_{KL}(p' \,\|\, p)$
Brier	$\sup_{p \in Q} \backslash \inf_{p \in Q} \; 1 - \sum_{k=1}^{K} p_k^2$	$\max_{p, p' \in Q} \; \sum_{k=1}^{K} (p_k - p'_k)^2$
spherical	$\sup_{p \in Q} \backslash \inf_{p \in Q} \; 1 - \|\|p\|\|_2$	$\max_{p, p' \in Q} \; \|\|p'\|\|_2 - \sum_{k=1}^{K} p_k p'_k / \|\|p'\|\|_2$
0/1	$\sup_{p \in Q} \backslash \inf_{p \in Q} \; 1 - \max p_k$	$\max_{p, p' \in Q} \; \max p'_k - p'_{k = \arg\max p_k}$

Disaggregation

- **Theorem:** If $L : \Delta_K \times \mathcal{Y} \longrightarrow \mathbb{R}$ is continuous in $p \in \Delta_K$, the following holds:

  (i) **Continuity**: EU as well as lower and upper bounds for AU and TU $=$ EU $+$ AU are continuous functionals.
  (ii) **Monotonicity**: for all credal sets C, Q such that $C \subseteq Q$, we have $\mathrm{EU}(C) \leq \mathrm{EU}(Q)$; the same holds for $\overline{\mathrm{AU}}$ and $\overline{\mathrm{TU}}$, respectively.
  (iii) **Precise probabilities**: for all credal sets Q such that $Q = \{p\}$, we have $\mathrm{EU}(Q) = 0$; the lower and upper bounds for TU and AU, respectively, coincide.

  Additionally, if L is a proper scoring rule, lower and upper bounds for TU, AU, and EU are **non-negative**.

Disaggregation

- Different losses L allow one to distinguish between sorts of uncertainty:
  - Uncertainty about the **true probability** $p(\cdot \mid \boldsymbol{x})$
  - Uncertainty about the **true outcome** $Y \sim p(\cdot \mid \boldsymbol{x})$
  - Uncertainty about the **best prediction**
  - Uncertainty about the **loss of the prediction**
- Our measures show strong **performance in downstream tasks** (accuracy-rejection curves, out-of-distribution data, active learning).

Conclusion

- **Designing reliable, uncertainty-aware learners** is an important task, but also challenging, both conceptually and computationally.
- **Distinguishing aleatoric and epistemic uncertainty** is useful, and several methods have been proposed for that purpose — though it seems that second-order uncertainty is hard to tackle.
- In particular, we showed that direct epistemic uncertainty prediction (aka **evidential deep learning**) via minimisation of a **second-order loss function** is theoretically flawed — EU is controlled through regularisation in a rather arbitrary way.
- It is clear that **prior knowledge** has an influence, and that there is no ground-truth EU, but can one represent EU in a somewhat **more objective manner**?
- What is the role of **credal sets** and IP in this regard?

References

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.

E. H., S. Destercke, and M.H. Shaker. Quantification of credal uncertainty in machine learning: A critical analysis and empirical comparison. In *Proc. UAI, 38th Conference on Uncertainty in Artificial Intelligence*, Eindhoven, Netherlands, 2022.

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

M. Jia, M. Shi, M. Sirotenko, Y. Cui, C. Cardie, B. Hariharan, H. Adam, and S. Belongie. Fashionpedia: Ontology, segmentation, and an attribute localization dataset, 2020.

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.

Y. Sale, M. Caprio, and E. Hüllermeier. Is the volume of a credal set a good measure for epistemic uncertainty? In *Proc. UAI, 39th Conference on Uncertainty in Artificial Intelligence*, 2023.

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

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 *Proc. UAI*, 2023.