# Uncertainty Quantification in Machine Learning

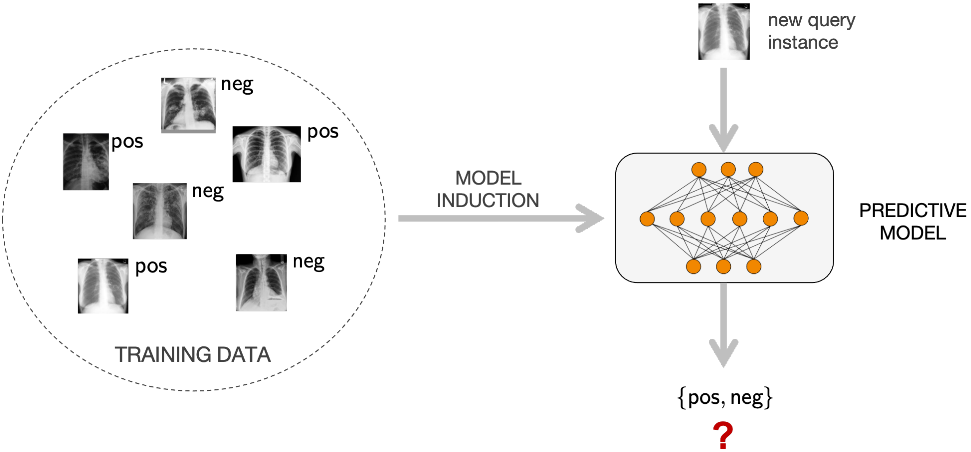
## From Aleatoric to Epistemic

Eyke Hüllermeier

Artificial Intelligence and Machine Learning
Institute of Informatics
University of Munich (LMU)

COPA 2022, Brighton, UK, August 24, 2022

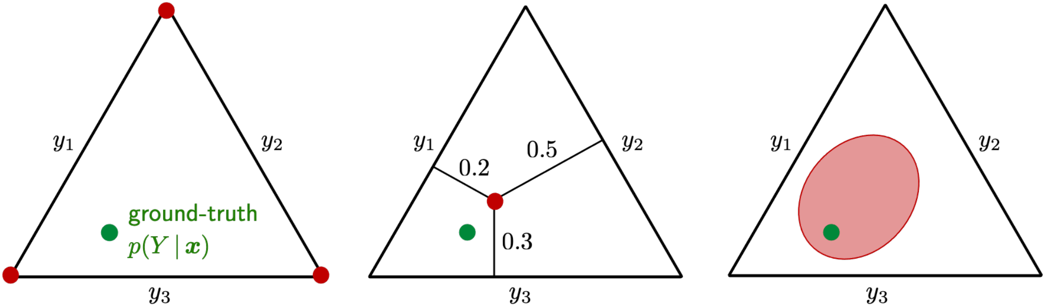
# Need for uncertainty-awareness of ML systems



# Lack of uncertainty-awareness of ML systems

- Predictions by EfficientNet on test images from ImageNet: For the left image, the neural network predicts "typewriter keyboard" with certainty 83.14 %, for the right image "stone wall" with certainty 87.63 %.

# Uncertainty representation and levels of uncertainty-awareness

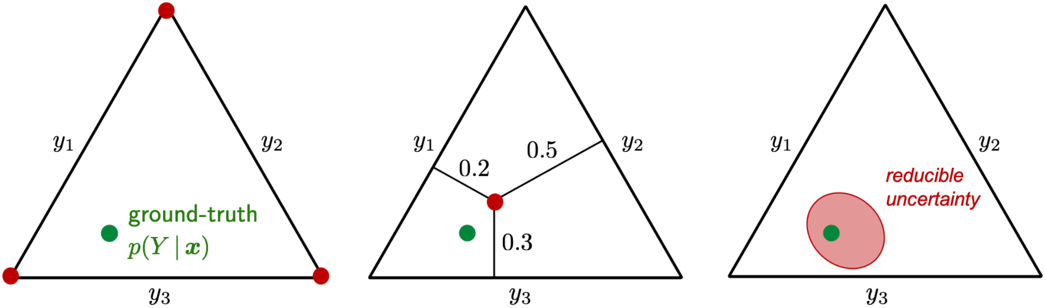



Deterministic predictor
$h : \mathcal{X} \longrightarrow \mathcal{Y}$

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

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

# Uncertainty representation and levels of uncertainty-awareness




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

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

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

# Aleatoric versus epistemic uncertainty

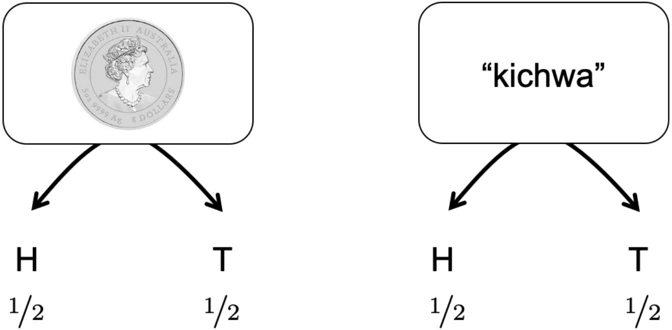
- **Aleatoric** (aka statistical) uncertainty
  - refers to the notion of **randomness**, that is, the variability in the outcome which is due to inherently random effects,
  - is a property of the **data-generating process**,
  - and as such **irreducible**.

# Aleatoric versus epistemic uncertainty

- **Aleatoric** (aka 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** (aka 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 (e.g., training data).

# Aleatoric versus epistemic uncertainty

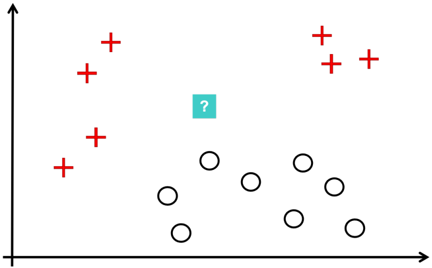



*"Not knowing the chance of mutually exclusive events and knowing the chance to be equal are two quite different states of knowledge"*

Ronald Fisher (1890-1962)

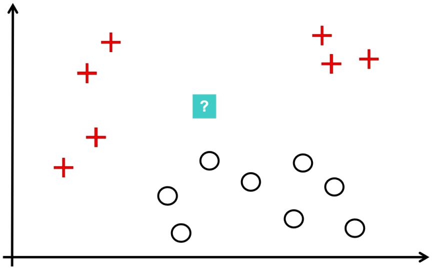
# Aleatoric versus epistemic uncertainty in ML

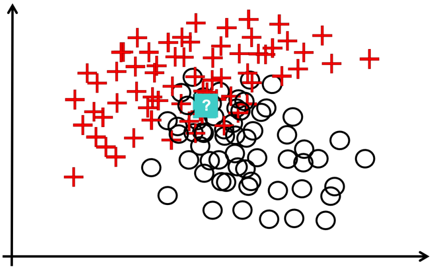
- Both types of uncertainty also play an important role in ML, where the learner's state of knowledge strongly depends on the amount of data seen so far ...



# Aleatoric versus epistemic uncertainty in ML

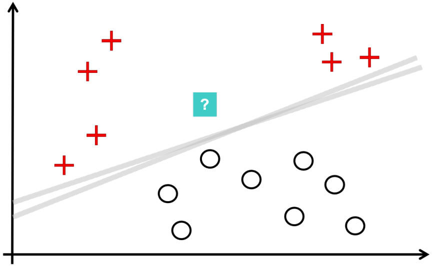
- Both types of uncertainty also play an important role in ML, where the learner's state of knowledge strongly depends on the amount of data seen so far ...



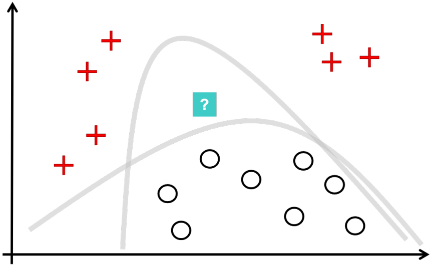



# Aleatoric versus epistemic uncertainty in ML

- … but also on the underlying model assumptions:




strong prior




weaker prior

# Uncertainty about (aleatoric) uncertainty

- The **distinction between aleatoric and epistemic uncertainty** can be difficult: Is the data-generating process completely random or only very complicated?

# Uncertainty about (aleatoric) uncertainty

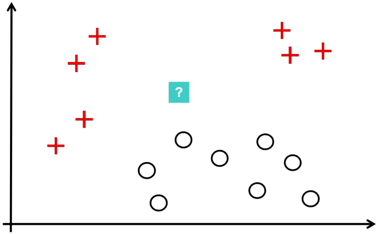
- The **distinction between aleatoric and epistemic uncertainty** can be difficult: Is the data-generating process completely random or only very complicated?
- **Predict the next number**: 116, 304, 194, 341, 224, 654, 609, 625, 533, 91, 205, 35, 527, 611, 128, 235, 348, 912, 582, 52, 672, 20, 856,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

# Uncertainty about (aleatoric) uncertainty

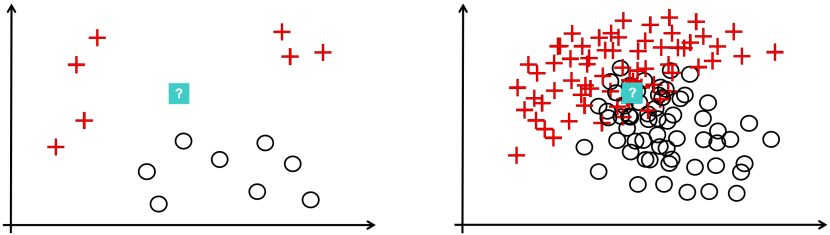
- The **distinction between aleatoric and epistemic uncertainty** can be difficult: Is the data-generating process completely random or only very complicated?
- **Predict the next number**: 116, 304, 194, 341, 224, 654, 609, 625, 533, 91, 205, 35, 527, 611, 128, 235, 348, 912, 582, 52, 672, 20, 856,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

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

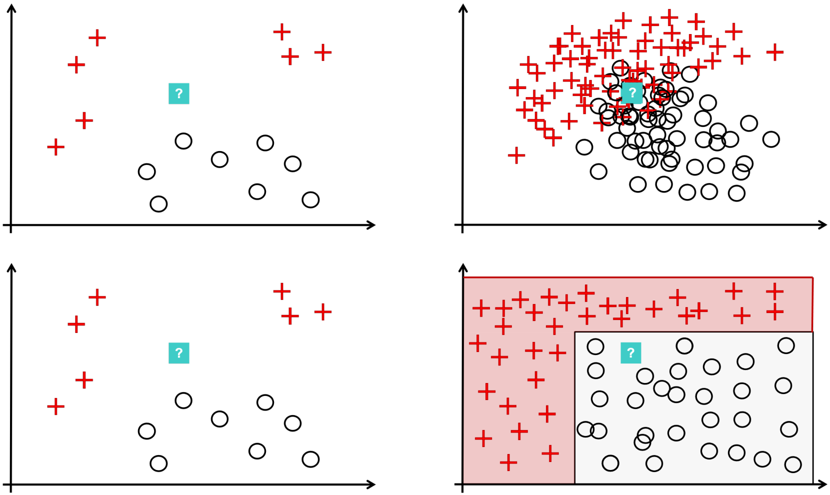
# Uncertainty about (aleatoric) uncertainty



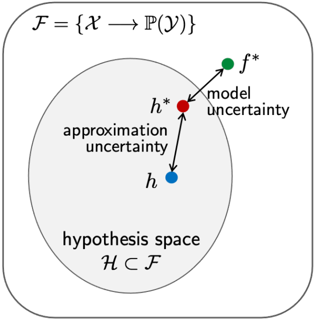
# Uncertainty about (aleatoric) uncertainty

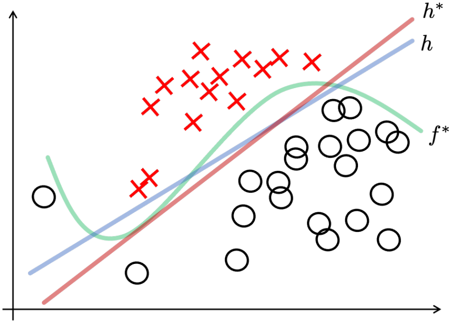


# Uncertainty about (aleatoric) uncertainty



# Sources of uncertainty




$$f^* = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{(\boldsymbol{x},y)\sim P}\, \ell(y, f(\boldsymbol{x}))$$

$$h^* = \arg\min_{h \in \mathcal{H}} \mathbb{E}_{(\boldsymbol{x},y)\sim P}\, \ell(y, h(\boldsymbol{x}))$$

$$h^* = \arg\min_{h \in \mathcal{H}} \sum_{i=1}^{N} \ell(y_i, h(\boldsymbol{x}_i))$$

# Agenda

1. Aleatoric and epistemic uncertainty
2. **Learning uncertainty-aware predictors**
3. Uncertainty quantification
4. Summary and outlook

# Predictive uncertainty

# Predictive uncertainty

- We assume a standard setting of **supervised learning** and are mainly interested in (per-instance) **predictive uncertainty**, i.e., the uncertainty in a prediction

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

produced for a query instance $\boldsymbol{x}$, where $h$ has been learned on training data $\mathcal{D}$.

# Predictive uncertainty

- We assume a standard setting of **supervised learning** and are mainly interested in (per-instance) **predictive uncertainty**, i.e., the uncertainty in a prediction

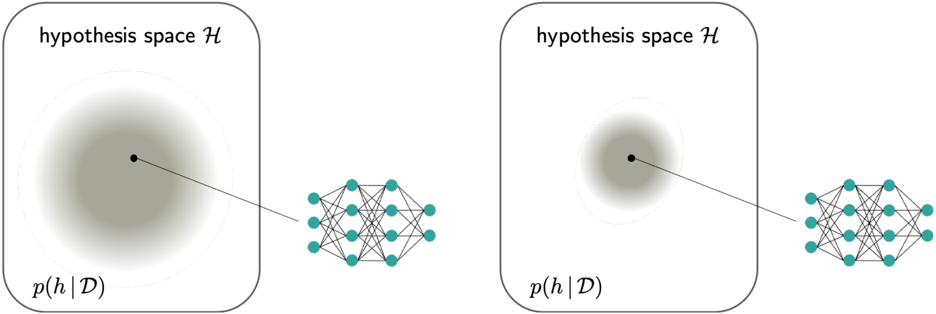
$$\hat{y} = h(\boldsymbol{x})$$

produced for a query instance $\boldsymbol{x}$, where $h$ has been learned on training data $\mathcal{D}$.

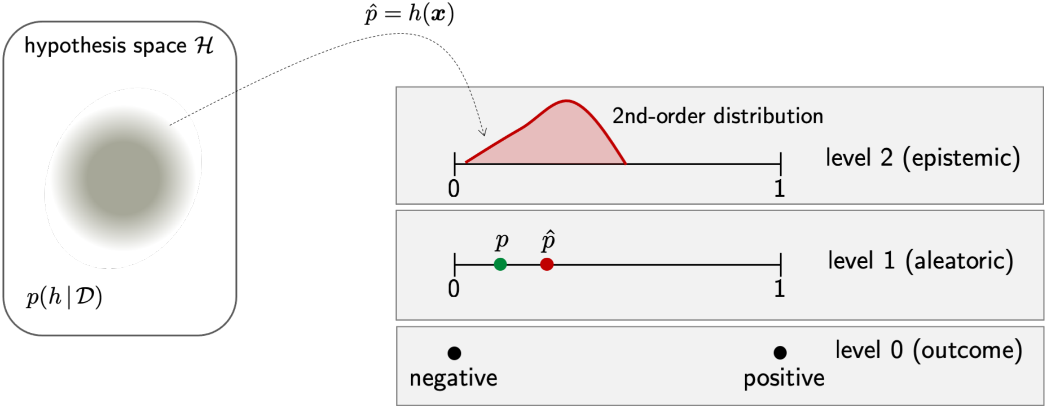
- Various **approaches** have been proposed in the literature:
  - Capture model uncertainty, translate into predictive uncertainty
  - Validation and self-assessment
  - Direct uncertainty prediction

# The Bayesian approach

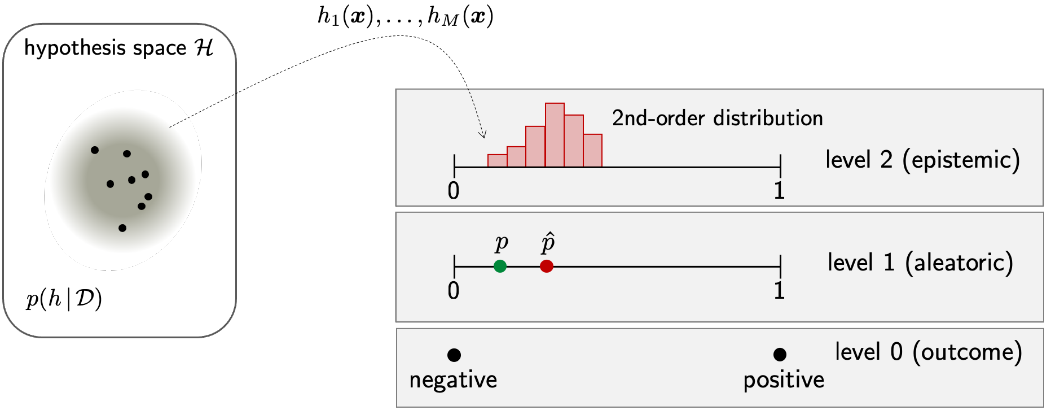
- A Bayesian learner maintains a probability distribution over the hypothesis space.
- The less concentrated that distribution, the higher the learner's epistemic uncertainty.



# Posterior predictive distribution



# Ensemble methods



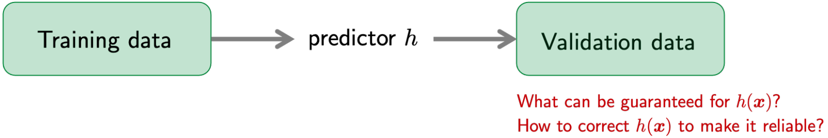
# Agenda

1. Aleatoric and epistemic uncertainty
2. **Learning uncertainty-aware predictors**
   - Model uncertainty and ensembling
   - **Validation and self-assessment**
   - Direct uncertainty prediction
3. Uncertainty quantification
4. Summary and outlook

# Validation and self-assessment

# Validation and self-assessment

- In addition to learning a predictor $h$ on $\mathcal{X}$, the learner also figures out how that predictor performs on out-of-sample data.



# Validation and self-assessment

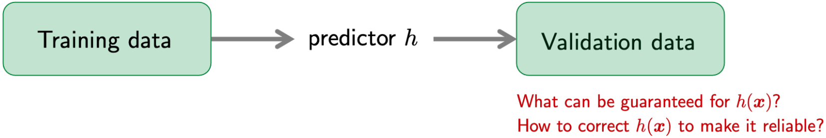
- In addition to learning a predictor $h$ on $\mathcal{X}$, the learner also figures out how that predictor performs on out-of-sample data.


What can be guaranteed for $h(\boldsymbol{x})$?
How to correct $h(\boldsymbol{x})$ to make it reliable?

- Example: Estimation of **error rate** via (cross-)validation.

# Validation and self-assessment

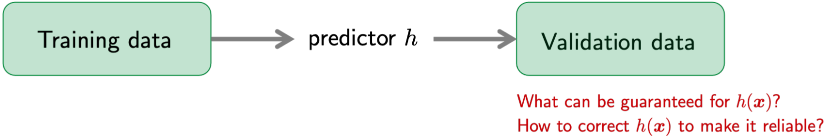
- In addition to learning a predictor $h$ on $\mathcal{X}$, the learner also figures out how that predictor performs on out-of-sample data.




- Example: Estimation of **error rate** via (cross-)validation.
- Yet, this is a **global** performance measure, not **per-instance**.

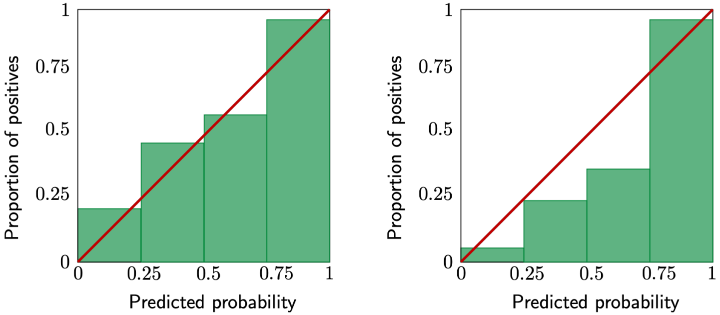
# Validation and self-assessment

- In addition to learning a predictor *h* on $\mathcal{X}$, the learner also figures out how that predictor performs on out-of-sample data.

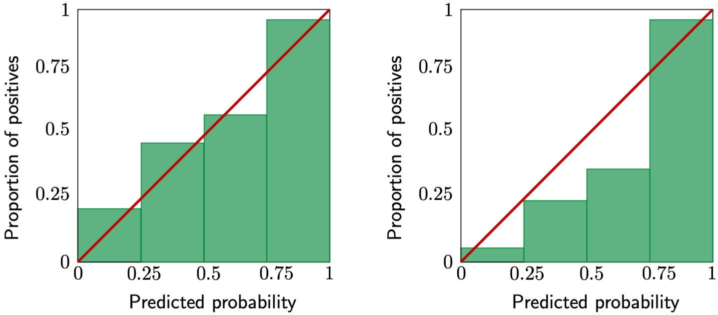



- Example: Estimation of **error rate** via (cross-)validation.
- Yet, this is a **global** performance measure, not **per-instance**.
- Per-instance uncertainty estimation appears to be difficult and indeed has theoretical limits (Barber *et al.*, 2021).

# Calibration

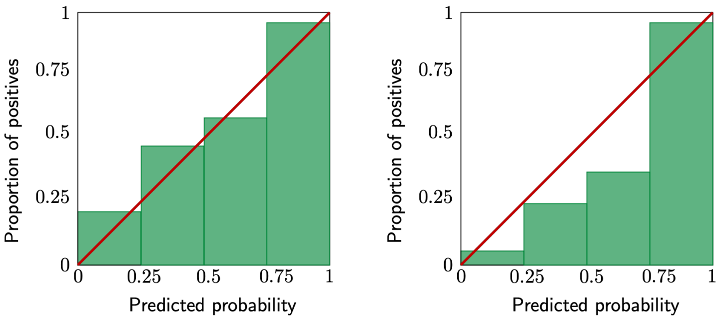


# Calibration



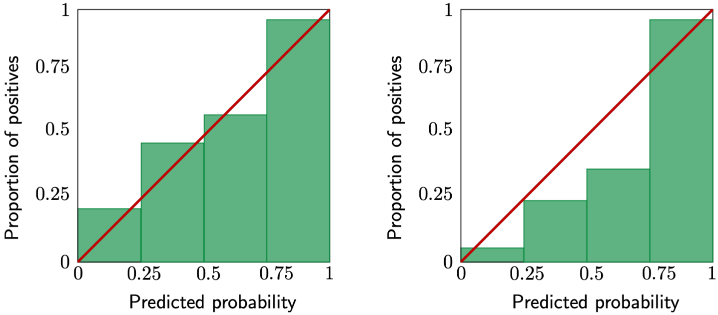

- On **calibration** data, the learner extracts information such as: A predicted probability of $\approx 0.6$ actually means a true probability of $\approx 0.3$.

# Calibration



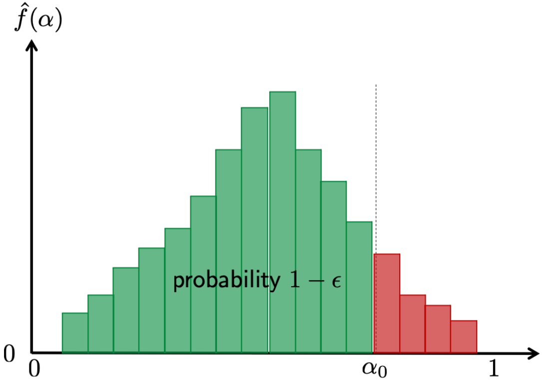

- On **calibration** data, the learner extracts information such as: A predicted probability of $\approx 0.6$ actually means a true probability of $\approx 0.3$.
- **Grouping** of instances with same score (predicted probability), needed to construct frequentist corrections of level-1 predictions based on level-0 data.

# Calibration



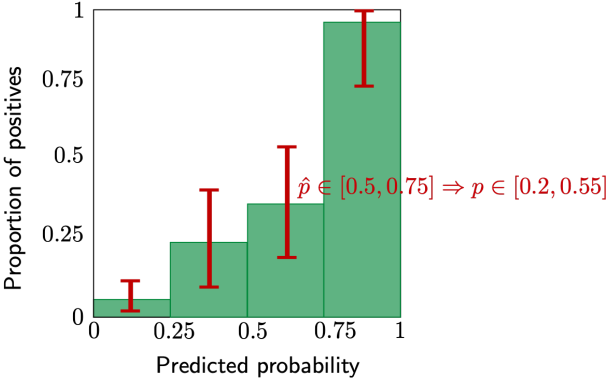

- On **calibration** data, the learner extracts information such as: A predicted probability of $\approx 0.6$ actually means a true probability of $\approx 0.3$.
- **Grouping** of instances with same score (predicted probability), needed to construct frequentist corrections of level-1 predictions based on level-0 data.
- A calibrator is a **one-dimensional** function, hence easier to learn.

# Conformal prediction




- A **conformal predictor** uses calibration data to learn rules such as: With high probability, true outcomes have a nonconformity of at most $\alpha_0$.
- This allows for constructing non-trivial yet valid **prediction sets**.

# Level-2 predictions




- Previous approaches refer to level-1 uncertainty, though level-2 estimation is in principle also possible (e.g., Venn predictors)

# Per-instance assessment

# Per-instance assessment

- Previous approaches require grouping of instances, though attempts at **per-instance assessment** have also been made.

# Per-instance assessment

- Previous approaches require grouping of instances, though attempts at **per-instance assessment** have also been made.
- For example, Lahlou *et al.* (2021) train an **error predictor** on validation data, which can be used to estimate epistemic uncertainty in terms of pointwise (excess) prediction error
$$\mathcal{E}\left(\hat{h}, \boldsymbol{x}\right) = \left(\hat{h}(\boldsymbol{x}) - f^*(\boldsymbol{x})\right)^2 .$$

# Per-instance assessment

- Previous approaches require grouping of instances, though attempts at **per-instance assessment** have also been made.
- For example, Lahlou *et al.* (2021) train an **error predictor** on validation data, which can be used to estimate epistemic uncertainty in terms of pointwise (excess) prediction error

$$\mathcal{E}\left(\hat{h}, \boldsymbol{x}\right) = \left(\hat{h}(\boldsymbol{x}) - f^*(\boldsymbol{x})\right)^2 .$$

- Yet, learning such a predictor appears to be difficult (and also includes learning of $f^*(\boldsymbol{x})$ or knowledge thereof).

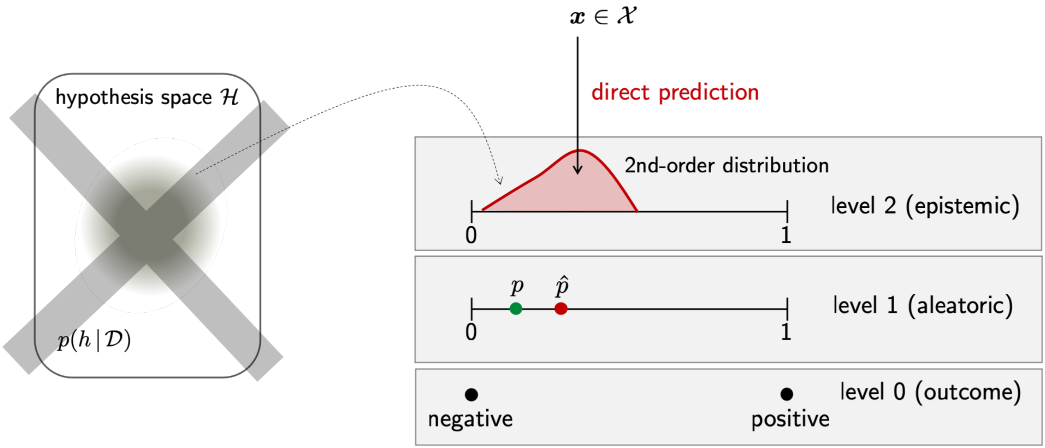
# Per-instance assessment

- Previous approaches require grouping of instances, though attempts at **per-instance assessment** have also been made.
- For example, Lahlou *et al.* (2021) train an **error predictor** on validation data, which can be used to estimate epistemic uncertainty in terms of pointwise (excess) prediction error
$$\mathcal{E}\left(\hat{h}, \boldsymbol{x}\right) = \left(\hat{h}(\boldsymbol{x}) - f^*(\boldsymbol{x})\right)^2 .$$
- Yet, learning such a predictor appears to be difficult (and also includes learning of $f^*(\boldsymbol{x})$ or knowledge thereof).
- Besides, one may question the definition of **uncertainty** in terms of **loss**.

# Agenda

1. Aleatoric and epistemic uncertainty
2. **Learning uncertainty-aware predictors**
   - Model uncertainty and ensembling
   - Validation and self-assessment
   - **Direct uncertainty prediction**
3. Uncertainty quantification
4. Summary and outlook

# Direct prediction



# Direct (epistemic) uncertainty prediction

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

$$\hat{h} : \mathcal{X} \longrightarrow \mathbb{P}\big(\mathbb{P}(\mathcal{Y})\big)$$

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

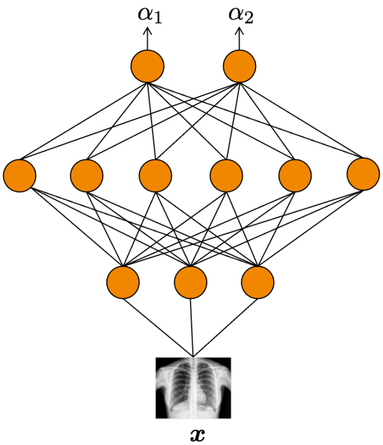
$$\hat{h} = \arg\min_h \sum_{i=1}^N \ell_2\left(\hat{h}(\boldsymbol{x}_i), y_i\right) \ ,$$

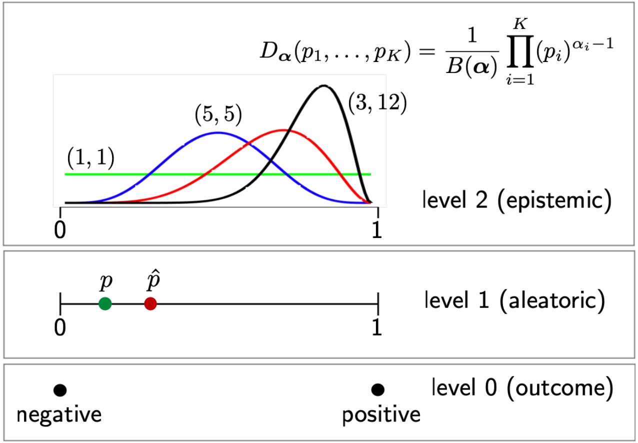
with a suitable **level-2 loss function**

$$\ell_2 : \mathbb{P}\big(\mathbb{P}(\mathcal{Y})\big) \times \mathcal{Y} \longrightarrow \mathbb{R} \, ,$$

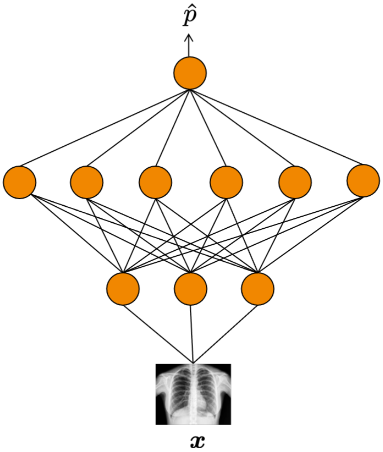
such that the predictor represents its epistemic uncertainty in a faithful way?

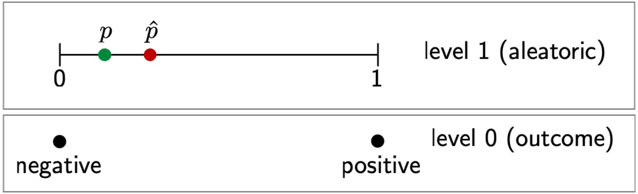
# Example: predicting a Dirichlet distribution



# The case of level-1 predictions



# Direct (epistemic) uncertainty prediction

- Training a probabilistic predictor via **empirical risk minimisation**, i.e.,

$$\hat{h} = \arg\min_h \sum_{i=1}^{N} \ell_1\left(\hat{h}(\boldsymbol{x}_i), y_i\right) ,$$

yields good (unbiased) predictors if $\ell_1$ is a (strictly) **proper scoring rule**, which incentivises the learner to predict the true $p(y \mid \boldsymbol{x})$.

# Direct (epistemic) uncertainty prediction

- Training a probabilistic predictor via **empirical risk minimisation**, i.e.,

$$\hat{h} = \arg\min_{h} \sum_{i=1}^{N} \ell_1 \left( \hat{h}(\boldsymbol{x}_i), y_i \right) ,$$

  yields good (unbiased) predictors if $\ell_1$ is a (strictly) **proper scoring rule**, which incentivises the learner to predict the true $p(y \mid \boldsymbol{x})$.

- A loss function $\ell_1 : \mathbb{P}(\mathcal{Y}) \times \mathcal{Y} \longrightarrow \mathbb{R}$ is a proper scoring rule if the expected loss minimiser coincides with the true probability $\boldsymbol{p}$:

$$\boldsymbol{p} = \arg\min_{\hat{\boldsymbol{p}}} \mathbb{E}_{Y \sim \boldsymbol{p}}\, \ell_1(\hat{\boldsymbol{p}}, Y)$$

  A scoring rule is **strictly proper** if the minimiser is unique.

# Direct epistemic uncertainty prediction

- Several authors proposed a **level-2 loss** of the form

$$\ell_2\big(Q, y\big) = \mathbb{E}_{P \sim Q}\, \ell_1\left(P, y\right) \, ,$$

where $Q$ is the level-2 prediction for a query instance $\boldsymbol{x}$.

# Direct epistemic uncertainty prediction

- Several authors proposed a **level-2 loss** of the form

$$\ell_2\big(Q, y\big) = \mathbb{E}_{P \sim Q}\, \ell_1\,(P, y)\ ,$$

where $Q$ is the level-2 prediction for a query instance $\boldsymbol{x}$.

- Thus, an individual prediction $Q$ is penalised in terms of the **expected level-1 loss**, with the expectation taken over the realisations of $P$.

# Direct epistemic uncertainty prediction

- Several authors proposed a **level-2 loss** of the form

$$\ell_2\big(Q, y\big) = \mathbb{E}_{P\sim Q}\,\ell_1\left(P, y\right) \ ,$$

where $Q$ is the level-2 prediction for a query instance $\boldsymbol{x}$.

- Thus, an individual prediction $Q$ is penalised in terms of the **expected level-1 loss**, with the expectation taken over the realisations of $P$.
- Examples of level-1 losses include cross entropy (Charpentier *et al.*, 2020) and Brier score (Sensoy *et al.*, 2018).

# Direct epistemic uncertainty prediction

- Several authors proposed a **level-2 loss** of the form

$$\ell_2\big(Q, y\big) = \mathbb{E}_{P \sim Q}\, \ell_1\,(P, y)\ ,$$

where $Q$ is the level-2 prediction for a query instance $\boldsymbol{x}$.

- Thus, an individual prediction $Q$ is penalised in terms of the **expected level-1 loss**, with the expectation taken over the realisations of $P$.
- Examples of level-1 losses include cross entropy (Charpentier *et al.*, 2020) and Brier score (Sensoy *et al.*, 2018).
- Besides, a **regularised version** has been proposed:

$$\ell_2\big(Q, y\big) = \mathbb{E}_{P \sim Q}\, \ell_1\,(P, y) + \lambda\, d_{KL}\,(Q, Q_0)$$

# Appropriate level-2 losses

- Informally, we define a level-2 loss function $\ell_2$ as **appropriate** if the following holds for the empirical loss minimiser

$$Q^{(N)} = \arg\min_{Q} \frac{1}{N} \sum_{n=1}^{N} \ell_2\left(Q, y^{(n)}\right)$$

on any i.i.d. observational data sequence $y^{(1)}, y^{(2)}, \dots$ with $y^{(i)} \sim P^*$:

(A1) The learner's **uncertainty gradually decreases** (in expectation) with increasing sample size $N$, in terms of a suitable uncertainty measure $U$.

(A2) In the limit $N \to \infty$, all **epistemic uncertainty disappears** and $Q^{(N)} \to \delta_{P^*}$.

# A negative result

- We formally proved that a loss minimisation approach using a level-2 loss as specified above does not lead to an appropriate level-2 loss (Bengs *et al.*, 2022).

# A negative result

- We formally proved that a loss minimisation approach using a level-2 loss as specified above does not lead to an appropriate level-2 loss (Bengs *et al.*, 2022).
- The results are general in the sense that $Q$ can be any level-2 distribution, not necessarily restricted to Dirichlet distributions.

# A negative result

- We formally proved that a loss minimisation approach using a level-2 loss as specified above does not lead to an appropriate level-2 loss (Bengs *et al.*, 2022).
- The results are general in the sense that $Q$ can be any level-2 distribution, not necessarily restricted to Dirichlet distributions.
- Moreover, the results do not depend on the underlying uncertainty measure $U$, as long as $U$ is not constant, maximal for the uniform distribution and minimal for Dirac measures.

# A negative result

- We formally proved that a loss minimisation approach using a level-2 loss as specified above does not lead to an appropriate level-2 loss (Bengs *et al.*, 2022).
- The results are general in the sense that $Q$ can be any level-2 distribution, not necessarily restricted to Dirichlet distributions.
- Moreover, the results do not depend on the underlying uncertainty measure $U$, as long as $U$ is not constant, maximal for the uniform distribution and minimal for Dirac measures.
- The results reveal that the quality of a (level-2) prediction $Q$ cannot be judged solely in the context of (level-0) observations $y$.

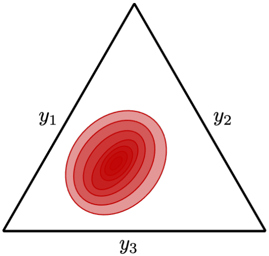
# Agenda

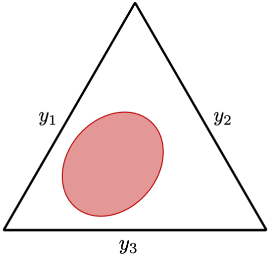
1. Aleatoric and epistemic uncertainty
2. Learning uncertainty-aware predictors
3. **Uncertainty quantification**
4. Summary and outlook

# Uncertainty quantification

# Uncertainty quantification

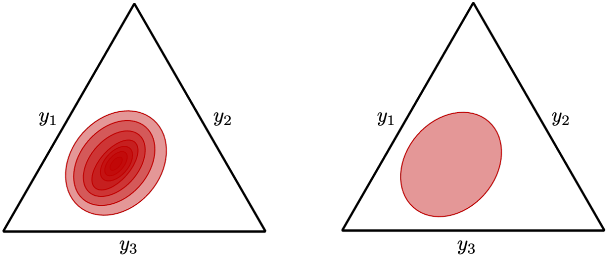
- Given a prediction $h(\boldsymbol{x})$ in the form of a second-order distribution or a credal set, how to quantify the **total uncertainty** in that prediction in terms of a single number?



# Uncertainty quantification

- Given a prediction $h(\boldsymbol{x})$ in the form of a second-order distribution or a credal set, how to quantify the **total uncertainty** in that prediction in terms of a single number?




- We may also seek a **decomposition** into an aleatoric and an epistemic part:

$$\mathrm{TU} = \mathrm{AU} + \mathrm{EU}$$

# Uncertainty quantification

# Uncertainty quantification

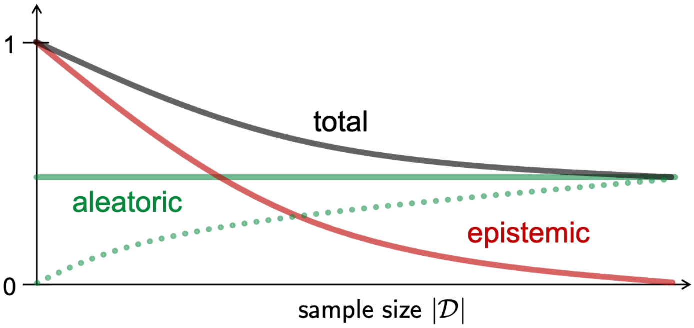
- One idea is to quantify the different types of uncertainty in terms of
  - **Shannon entropy**
$$H[Y] = -\sum_{y \in \mathcal{Y}} \mathbf{p}(y) \log_2 \mathbf{p}(y) \, ,$$
,
  - **conditional entropy**
$$H[Y \mid P] = -\int Q(p \mid \mathcal{D}) \left( \sum_{y \in \mathcal{Y}} \mathbf{p}(y \mid p) \log_2 \mathbf{p}(y \mid p) \right) \, d\, p \, ,$$
  - and **mutual information**

  between outcome $Y$ and (level-1) distribution $P$ (Malinin and Gales, 2018), respectively:

$$\underbrace{H[Y]}_{\text{total uncertainty}} = \underbrace{H[Y \mid P]}_{\text{aleatoric uncertainty}} + \underbrace{I(Y; P)}_{\text{epistemic uncertainty}}$$

# Remarks

- MI actually measures the (average) **divergence** between the candidate (level-1) distributions, so it is rather a measure of **conflict** than **ignorance** (which is difficult to capture in terms of probabilities anyway).
- One may also question the **additive decomposition** $\text{TU} = \text{AU} + \text{EU}$ itself.



# Uncertainty of credal sets

# Uncertainty of credal sets

- Uncertainty measures $U$ for credal sets have been studied **axiomatically**:

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. $Q''$,

$$U(Q) \leq U(Q') + U(Q'') .$$

A6 **Additivity**: The last inequality is an equality in the case where $Q'$ and $Q''$ are independent (assuming a suitably defined notion of independence).

# Measures of total, aleatoric, and epistemic uncertainty

# Measures of total, aleatoric, and epistemic uncertainty

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

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

# 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.

# 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

# Disaggregation

- There is no **additive decomposition**

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

such that all three measures behave well.

# 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**.

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

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

# 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 and isolate **potential deficiencies**.

# A new measure

# A new measure

- We proposed and axiomatically justified a new measure of **total predictive uncertainty**, more tailored to the ML setting, as well as its decomposition into aleatoric and epistemic uncertainty.

# A new measure

- We proposed and axiomatically justified a new measure of **total predictive uncertainty**, more tailored to the ML setting, as well as its decomposition into aleatoric and epistemic uncertainty.
- In the case of **binary classification**, where a credal prediction is of the form

$$Q_{\alpha,\beta} = \left\{ \mathrm{Bern}(p) \,\middle|\, \alpha \leq p \leq \beta \right\},$$

the measure is given as follows:

$$\mathrm{TP}(\alpha, \beta) = \underbrace{\min(1-\alpha, \beta)}_{\text{total}} = \underbrace{\min(\alpha, 1-\beta)}_{\text{aleatoric}} + \underbrace{(\beta - \alpha)}_{\text{epistemic}}$$

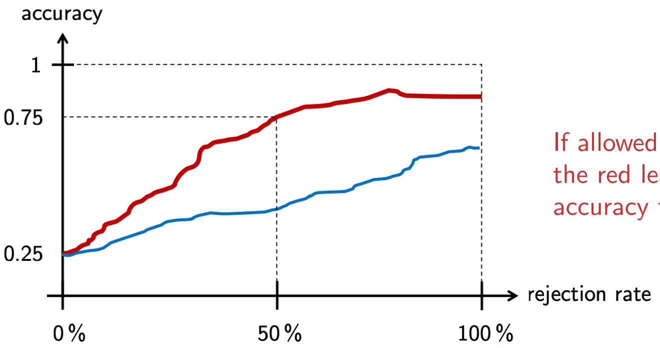
# Empirical evaluation

# Empirical evaluation

- **Ensemble-based construction** of credal predictions.

# Empirical evaluation

- **Ensemble-based construction** of credal predictions.
- **Accuracy-rejection curves**: Allow the learner to reject the $r$ % presumably most uncertain test cases and measure accuracy on the remaining ones.




If allowed to reject 50 % of the cases, the red learner manages to increase accuracy from 0.25 to 0.75.

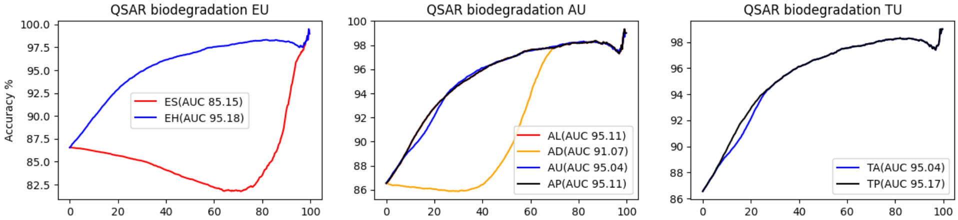
# Results

# Results

- Empirical results match with theory: Formally justified measures show strong performance, whereas the “derived” measures perform very poorly.

# Results

- Empirical results match with theory: Formally justified measures show strong performance, whereas the "derived" measures perform very poorly.
- Newly proposed measure yields the only decomposition of total into aleatoric and epistemic uncertainty, such that all three measures produce meaningful results.



# Summary and Outlook

# Summary and Outlook

- **Learning reliable predictors** that represent their uncertainty in a faithful way is a challenging task, both conceptually and computationally.

# Summary and Outlook

- **Learning reliable predictors** that represent their uncertainty in a faithful way is a challenging task, both conceptually and computationally.
- **Distinguishing different sources and types of uncertainty** is useful, though it seems that epistemic uncertainty hard to represent in an objective way (depends on prior, regularisation, incentive, etc.).

# Summary and Outlook

- **Learning reliable predictors** that represent their uncertainty in a faithful way is a challenging task, both conceptually and computationally.
- **Distinguishing different sources and types of uncertainty** is useful, though it seems that epistemic uncertainty hard to represent in an objective way (depends on prior, regularisation, incentive, etc.).
- **Quantifying predictive uncertainty** in a theoretically sound manner, and disentangling total into aleatoric and epistemic uncertainty, is difficult, too.

# Summary and Outlook

- **Learning reliable predictors** that represent their uncertainty in a faithful way is a challenging task, both conceptually and computationally.
- **Distinguishing different sources and types of uncertainty** is useful, though it seems that epistemic uncertainty hard to represent in an objective way (depends on prior, regularisation, incentive, etc.).
- **Quantifying predictive uncertainty** in a theoretically sound manner, and disentangling total into aleatoric and epistemic uncertainty, is difficult, too.
- Usefulness of **generalized uncertainty calculi**?

# References


R. Foygel Barber, J. Candes, J. Emmanuel, A. Ramdas, and R.J. Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference*, 10(2):455–482, 2021.

V. Bengs, E. Hüllermeier, and W. Waegeman. On the difficulty of epistemic uncertainty quantification in machine learning: The case of direct uncertainty estimation through loss minimisation. *arXiv:2203.06102*, 2022.

B. Charpentier, D. Zügner, and S. Günnemann. Posterior network: Uncertainty estimation without OOD samples via density-based pseudo-counts. In *Proc. NeurIPS, Neural Information Processing Systems*, 2020.

E. Hüllermeier, 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.

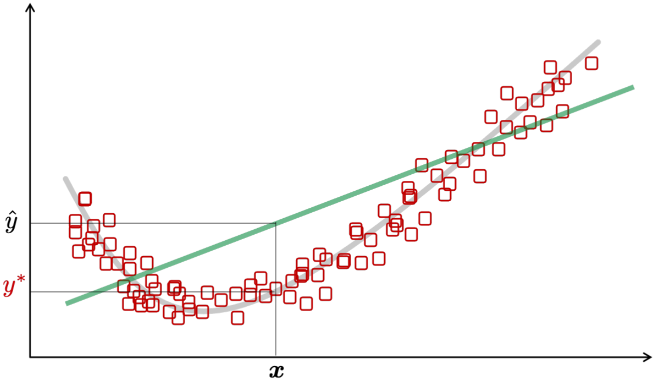
G.J. Klir. *Uncertainty and Information: Foundations of Generalized Information Theory*. Wiley, 2005.

S. Lahlou, M. Jain, H. Nekoei, V. Butoi, P. Bertin, J. Rector-Brooks, M. Korablyov, and Y. Bengio. DEUP: Direct epistemic uncertainty prediction, 2021.

A. Malinin and M. Gales. Predictive uncertainty estimation via prior networks. In *Proc. NeurIPS, 32nd Conf. on Neural Information Processing Systems*. Montreal, Canada, 2018.

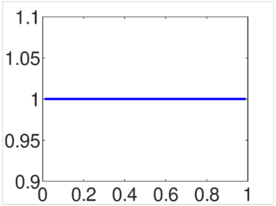
M. Sensoy, L. Kaplan, and M. Kandemir. Evidential deep learning to quantify classification uncertainty. In *Proc. NeurIPS, 32nd Conf. on Neural Information Processing Systems*, Montreal, Canada, 2018.

# Model misspecification

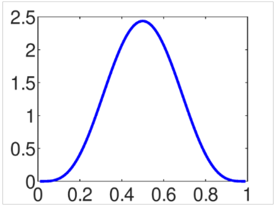



What uncertainty should the learner report at $\boldsymbol{x}$?

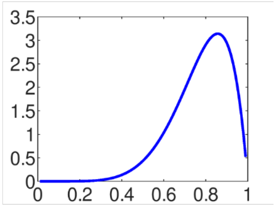
# Example: level-2 distributions over Bernoulli



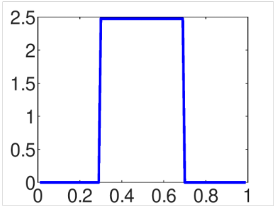

TU = 1.00, AU = 0.73, EU = 0.27



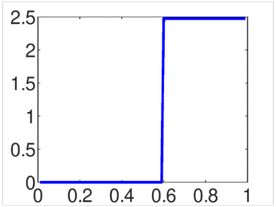

TU = 1.00, AU = 0.93, EU = 0.07



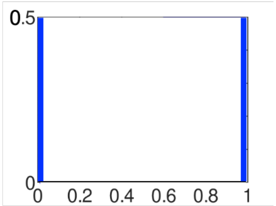

TU = 0.76, AU = 0.69, EU = 0.07




TU = 1.00, AU = 0.96, EU = 0.04



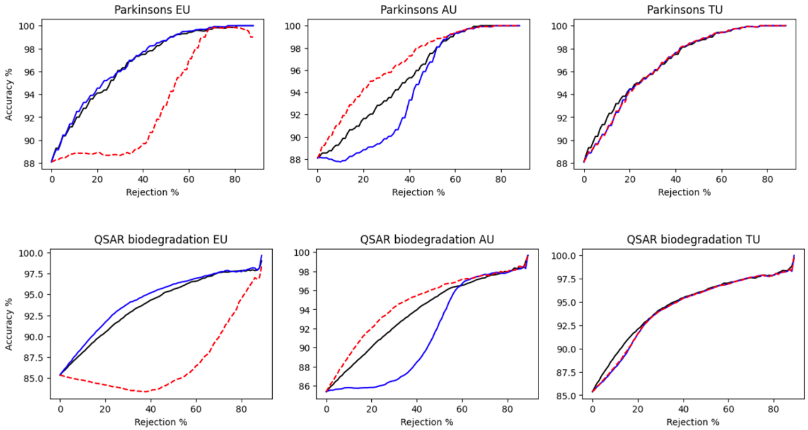

TU = 0.73, AU = 0.67, EU = 0.07




TU = 1.00, AU = 0.00, EU = 1.00

# Evaluation: accuracy-rejection curves

- Reject test instances for which (total, aleatoric, epistemic) uncertainty exceeds a certain threshold, measure accuracy on the remaining ones.



# A negative result

- **Theorem 1.** If $\ell_1 : \mathbb{P}(\mathcal{Y}) \times \mathcal{Y} \longrightarrow \mathbb{R}$ is such that

$$\ell_1\left(\mathbb{E}_{\boldsymbol{\theta}\sim Q}[\boldsymbol{\theta}], y\right) \leq \mathbb{E}_{\boldsymbol{\theta}\sim Q}\left[\ell_1\left(\boldsymbol{\theta}, y\right)\right]$$

for all $y \in \mathcal{Y}$, then $\ell_2(Q, y) = \mathbb{E}_{\boldsymbol{\theta}\sim Q}\left[\ell_1(\boldsymbol{\theta}, y)\right]$ violates A1.

  - Condition on $\ell_1$ is fulfilled if $\ell_1$ is convex (in the first argument)
  - Includes Brier score and cross-entropy, which are (strictly) convex
  - Proof reveals that $\hat{Q}$ is always a point-mass on $\mathbb{P}(\mathcal{Y})$ (i.e., a level-1 prediction)

# A negative result

- **Theorem 2.** If $\ell_1 : \mathbb{P}(\mathcal{Y}) \times \mathcal{Y} \longrightarrow \mathbb{R}$ is strictly convex in its first argument, then there exists $\overline{\lambda} > 0$ such that

$$\ell_2(Q, y) = \mathbb{E}_{\boldsymbol{\theta} \sim Q}\left[\ell_1(\boldsymbol{\theta}, y)\right] + \lambda \cdot \mathrm{KL}\left[Q, \mathrm{Unif}(\mathbb{P}(\mathcal{Y}))\right]$$

violates A1 for all $\lambda < \overline{\lambda}$.

# A negative result

- **Theorem 3.** If $\ell_1 : \mathbb{P}(\mathcal{Y}) \times \mathcal{Y} \longrightarrow \mathbb{R}$ is
  (i) strictly proper,
  (ii) locally Lipschitz (in the first argument),

  then there exists $\underline{\lambda} > 0$ such that

$$\ell_2(Q, y) = \mathbb{E}_{\boldsymbol{\theta} \sim Q}\left[\ell_1(\boldsymbol{\theta}, y)\right] + \lambda \cdot \mathrm{KL}\left[Q, \mathrm{Unif}(\mathbb{P}(\mathcal{Y}))\right]$$

  violates A2 for all $\lambda > \underline{\lambda}$.

# A negative result

- **Theorem 3.** If $\ell_1 : \mathbb{P}(\mathcal{Y}) \times \mathcal{Y} \longrightarrow \mathbb{R}$ is
  (i) strictly proper,
  (ii) locally Lipschitz (in the first argument),

  then there exists $\underline{\lambda} > 0$ such that

$$\ell_2(Q, y) = \mathbb{E}_{\boldsymbol{\theta} \sim Q}\left[\ell_1(\boldsymbol{\theta}, y)\right] + \lambda \cdot \mathrm{KL}\left[Q, \mathrm{Unif}(\mathbb{P}(\mathcal{Y}))\right]$$

  violates A2 for all $\lambda > \underline{\lambda}$.

- Brier score and cross-entropy fulfill both (i) and (ii).

# The need for assumptions

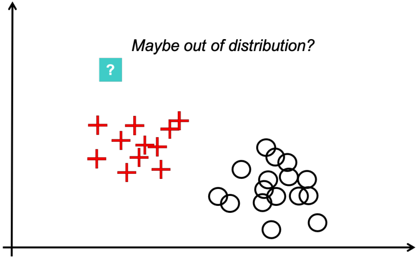
- A precise specification of the **problem setting** and **underlying assumptions** is an important prerequisite, not only for providing learning guarantees, but also for **uncertainty quantification**.

# The need for assumptions

- A precise specification of the **problem setting** and **underlying assumptions** is an important prerequisite, not only for providing learning guarantees, but also for **uncertainty quantification**.
- Quite obvious for assumptions such as **i.i.d. data generation** (future is like past).

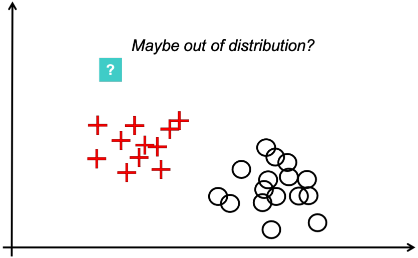
# The need for assumptions

- A precise specification of the **problem setting** and **underlying assumptions** is an important prerequisite, not only for providing learning guarantees, but also for **uncertainty quantification**.
- Quite obvious for assumptions such as **i.i.d. data generation** (future is like past).



# The need for assumptions

- A precise specification of the **problem setting** and **underlying assumptions** is an important prerequisite, not only for providing learning guarantees, but also for **uncertainty quantification**.
- Quite obvious for assumptions such as **i.i.d. data generation** (future is like past).




- Here, one might be quite sure about the class of the query under standard assumptions of binary classification, but much less so in a setting of **novelty detection**, where new classes may emerge.

# The Dirichlet distribution

- A **Dirichlet distribution** $\mathrm{Dir}(\alpha)$ is specified by means of $K \geq 2$ positive real-valued parameters, i.e., a vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K) \in \mathbb{R}_+^K$.

# The Dirichlet distribution

- A **Dirichlet distribution** $\text{Dir}(\alpha)$ is specified by means of $K \geq 2$ positive real-valued parameters, i.e., a vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K) \in \mathbb{R}_+^K$.
- The probability density function is defined on the $K$ simplex

$$\Delta_K = \left\{ \boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)^\top \,|\, \theta_1, \ldots, \theta_K \geq 0, \sum_{k=1}^{K} \theta_k = 1 \right\}$$

and given as follows:

$$p(\boldsymbol{\theta} \,|\, \boldsymbol{\alpha}) = p(\theta_1, \ldots, \theta_K \,|\, \boldsymbol{\alpha}) = \frac{1}{\mathbb{B}(\boldsymbol{\alpha})} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1},$$

where the normalisation constant is the multivariate Beta function.

# The Dirichlet distribution

- A **Dirichlet distribution** $\mathrm{Dir}(\alpha)$ is specified by means of $K \geq 2$ positive real-valued parameters, i.e., a vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K) \in \mathbb{R}_+^K$.
- The probability density function is defined on the $K$ simplex

$$\Delta_K = \left\{ \boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)^\top \,|\, \theta_1, \ldots, \theta_K \geq 0, \sum_{k=1}^{K} \theta_k = 1 \right\}$$

and given as follows:

$$p(\boldsymbol{\theta} \,|\, \boldsymbol{\alpha}) = p(\theta_1, \ldots, \theta_K \,|\, \boldsymbol{\alpha}) = \frac{1}{\mathbb{B}(\boldsymbol{\alpha})} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1},$$

where the normalisation constant is the multivariate Beta function.
- In Bayesian statistics, the Dirichlet distribution is commonly used as the conjugate prior of the **multinomial distribution**.