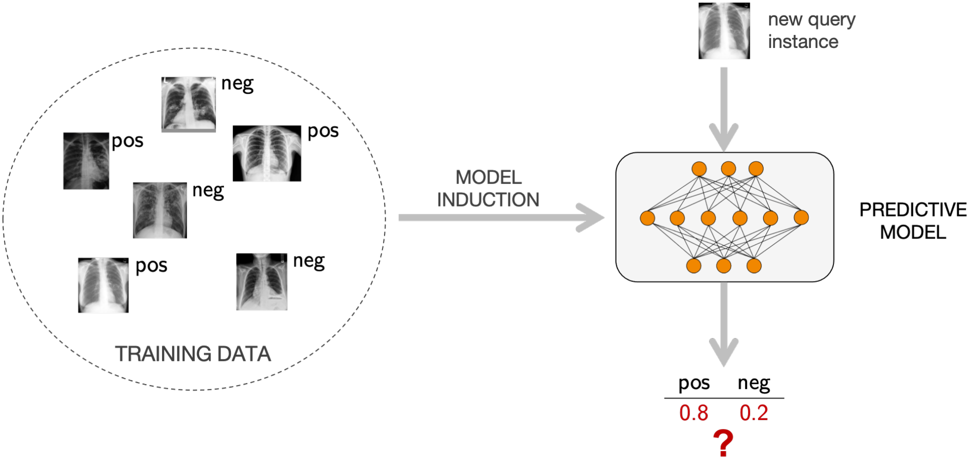
Representation of Quantification of Uncertainty in Machine Learning

Eyke Hüllermeier

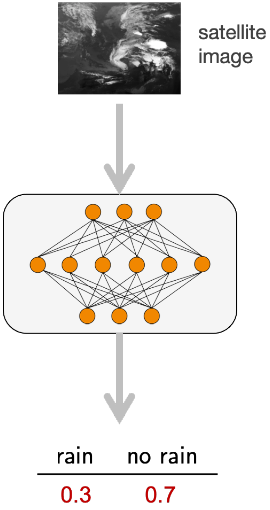
Artificial Intelligence and Machine Learning
Institute of Informatics, LMU Munich
Munich Center for Machine Learning (MCML)

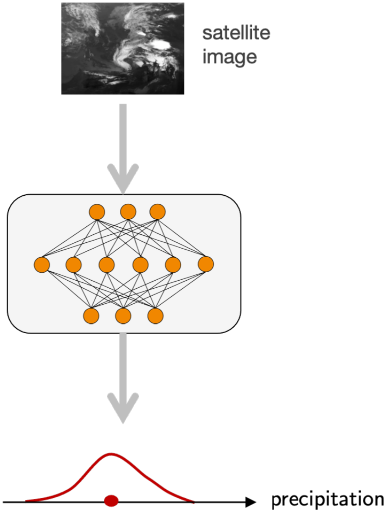
Joint TRR 165/181 Conference, Ingolstadt, March 2023

Need for uncertainty-awareness of ML systems



Classification versus regression





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 %.

typewriter keyboard

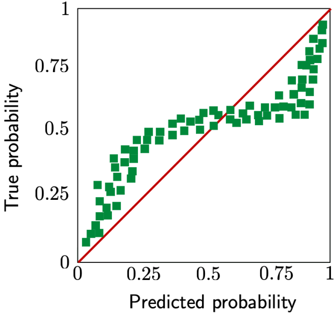
stone wall

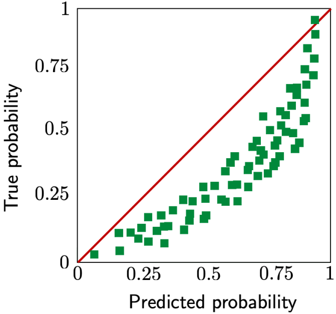
Agenda

1. Introduction
2. **Calibrating probabilistic predictors**
3. Epistemic uncertainty
4. Learning uncertainty-aware predictors
5. Uncertainty quantification
6. Summary and outlook

Calibration: improving probability estimates

- Examples: bias toward extreme probabilities (left), systematic overestimation (right)



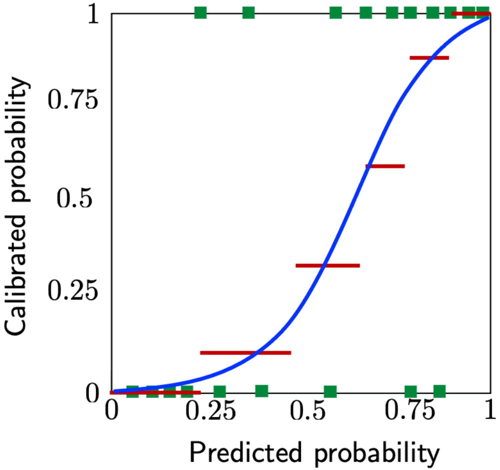


- A (binary) classifier is **calibrated** if

$$P\big(y \mid \hat{p}(y) = \alpha\big) = \alpha\,.$$

Calibration: improving probability estimates

- Example: calibration through isotonic regression or beta calibration (Kull *et al.*, 2017)

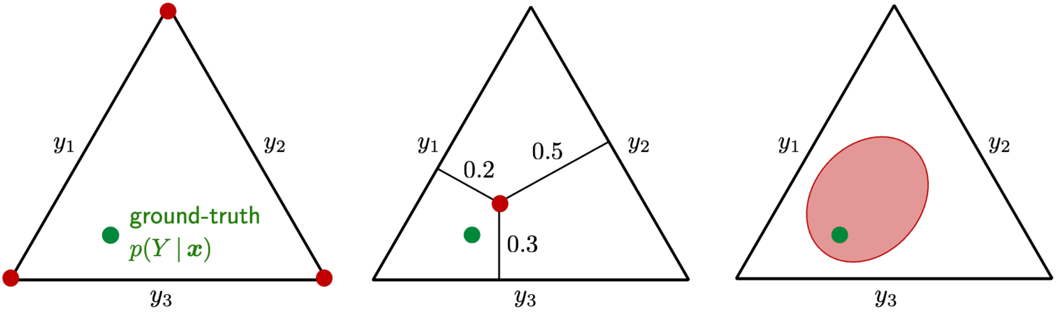


Agenda

1. Introduction
2. Calibrating probabilistic predictors
3. **Epistemic uncertainty**
4. Learning uncertainty-aware predictors
5. Uncertainty quantification
6. Summary and outlook

Uncertainty representation and levels of uncertainty-awareness

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



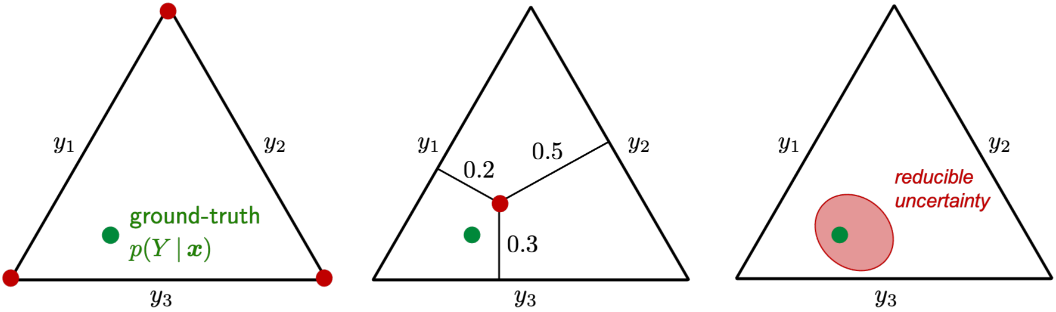
Deterministic predictor
$h : \mathcal{X} \longrightarrow \mathcal{Y}$

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

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

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

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

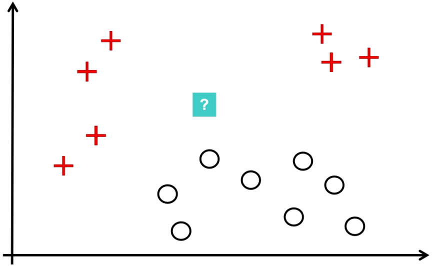
Aleatoric versus epistemic uncertainty

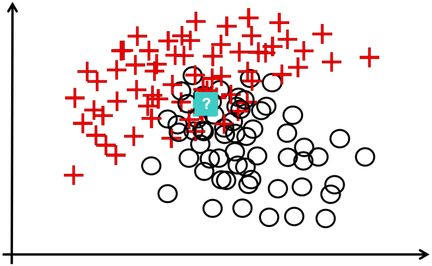
- **Aleatoric** (statistical) uncertainty
  - refers to the notion of **randomness**, that is, the variability in the outcome which is due to inherently random effects,
  - is a property of the **data-generating process**,
  - and as such **irreducible**.

- **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

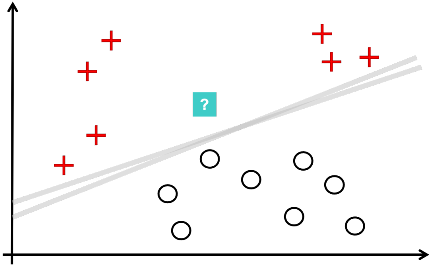
- Both types of uncertainty also play an important role in ML, where the learner's state of knowledge strongly depends on the amount of data seen so far ...



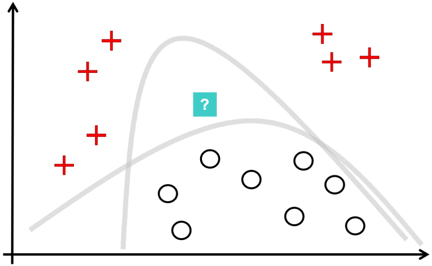


Aleatoric versus epistemic uncertainty in ML

- … but also on the underlying model assumptions:



strong prior (linear model)



weaker prior (nonlinear model)

Agenda

1. Introduction
2. Calibrating probabilistic predictors
3. Epistemic uncertainty
4. **Learning uncertainty-aware predictors**
    - Bayesian inference
    - Direct uncertainty prediction
    - Validation and self-assessment
5. Uncertainty quantification
6. Summary and outlook

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.

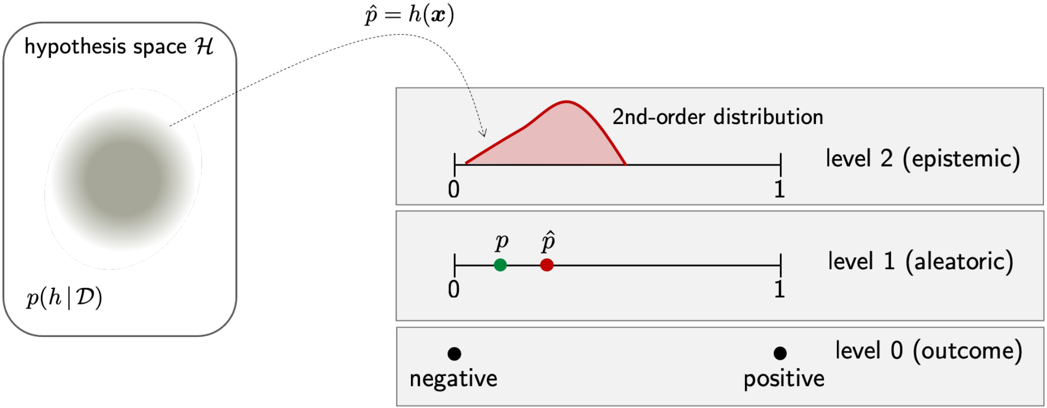
- Various **approaches** have been proposed in the literature:
  - Bayesian inference
  - Validation and self-assessment
  - Direct uncertainty prediction

Agenda

1. Introduction
2. Calibrating probabilistic predictors
3. Epistemic uncertainty
4. **Learning uncertainty-aware predictors**
   - **Bayesian inference**
   - Direct uncertainty prediction
   - Validation and self-assessment
5. Uncertainty quantification
6. Summary and outlook

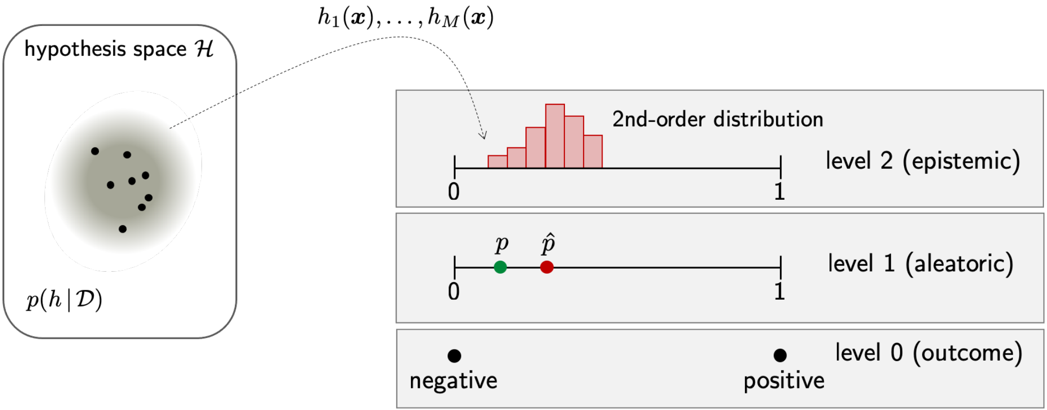
The Bayesian approach: posterior predictive distribution

- Model uncertainty translates into predictive uncertainty:



Ensemble methods

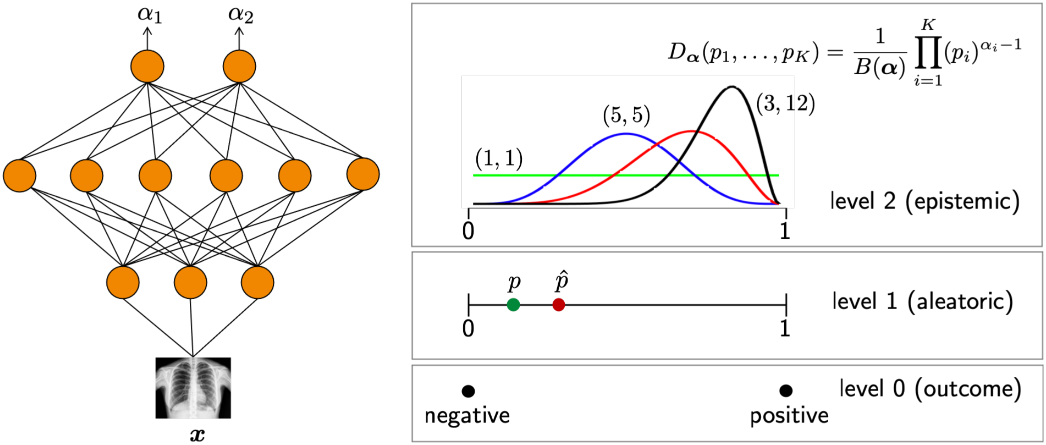
- Model uncertainty translates into predictive uncertainty:



Agenda

1. Introduction
2. Calibrating probabilistic predictors
3. Epistemic uncertainty
4. **Learning uncertainty-aware predictors**
   - Bayesian inference
   - **Direct uncertainty prediction**
   - Validation and self-assessment
5. Uncertainty quantification
6. Summary and outlook

Example of second-order prediction with Dirichlet distributions



Direct (epistemic) uncertainty prediction through loss minimisation

- 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 predictor

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

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

$$g = \arg\min_{h} \sum_{i=1}^{N} L_2\left(h(\boldsymbol{x}_i), y_i\right) ,$$

with a suitable **second-order loss function**

$$L_2 : \mathbb{P}\big(\mathbb{P}(\mathcal{Y})\big) \times \mathcal{Y} \longrightarrow \mathbb{R} \, ,$$

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

Direct (epistemic) uncertainty prediction through loss minimisation

- Negative results by Bengs *et al.* (2022), Meinert *et al.* (2022) ...
- For **first-order predictions** $\hat{p} \in \mathbb{P}(\mathcal{Y})$, there are loss functions (proper scoring rules)

$$L_1 : \mathbb{P}(\mathcal{Y}) \times \mathcal{Y} \longrightarrow \mathbb{R}$$

  that incentivise the learner to predict ground-truth probabilities $P(y \,|\, \boldsymbol{x})$.

- For **second-order predictions** $\hat{Q} \in \mathbb{P}(\mathbb{P}(\mathcal{Y}))$, corresponding losses

$$L_2 : \mathbb{P}\big(\mathbb{P}(\mathcal{Y})\big) \times \mathcal{Y} \longrightarrow \mathbb{R}$$

  do not seem to exist.

Agenda

Validation and self-assessment

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

Training data $\longrightarrow$ predictor h $\longrightarrow$ Validation 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 (e.g., make mistake in $\approx 20\%$ of the cases).
- Yet, this is a **global** performance measure, not **per-instance** (e.g., per-patient).
- Truly per-instance uncertainty estimation appears to be difficult and indeed has theoretical limits (Barber *et al.*, 2021).

Conformal prediction

- **Conformal prediction** (Balasubramanian *et al.*, 2014) is a framework for reliable prediction that is rooted in classical frequentist statistics and **hypothesis testing**.
- Instead of point predictions, CP makes **set-valued predictions** covering the true outcome with high probability.

$$\longrightarrow \qquad P\Big(y \in Y = \{2, 3, 9\}\Big) \text{ w.h.p.}$$

- **Guaranteed validity**: probability of an invalid prediction ($y \notin Y$) is (asymptotically) bounded by $\epsilon > 0$.

Conformal prediction

- CP uses a **scoring function** that assigns a degree of **nonconformity** to tuples consisting of query $\boldsymbol{x}$ and hypothetical outcome $\hat{y}$:

$$s = f(\boldsymbol{x}, \hat{y})$$

- On **calibration data**, CP finds a threshold α_0, such that

$$P\big(f(\boldsymbol{x}, y) \leq \alpha_0\big) \geq 1 - \epsilon$$

  if $(\boldsymbol{x}, y)$ is a **real observation**.

- This allows for constructing (valid) **prediction sets**:

$$\hat{Y}(\boldsymbol{x}) = \Big\{\hat{y} \in \mathcal{Y} \,|\, f(\boldsymbol{x}, \hat{y}) \leq \alpha_0\Big\}$$

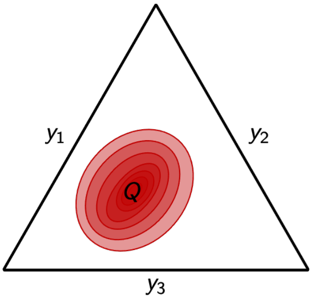
Agenda

1. Introduction
2. Calibrating probabilistic predictors
3. Epistemic uncertainty
4. Learning uncertainty-aware predictors
   - Bayesian inference
   - Direct uncertainty prediction
   - Validation and self-assessment
5. **Uncertainty quantification**
6. Summary and outlook

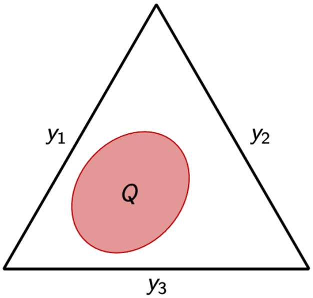
Uncertainty quantification

- **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

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



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



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

Uncertainty quantification

- 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**.

Uncertainty quantification

- Common approach for second-order probabilities $Q \in \mathbb{P}(\mathbb{P}(\mathcal{Y}))$, where each model θ induces a distribution $p_{\theta,\boldsymbol{x}} \in \mathbb{P}(\mathcal{Y})$, and the model itself is a RV $\Theta \sim Q$:

$$\Theta \sim Q \quad \longrightarrow \quad Y|\boldsymbol{x} \sim P_{\theta,\boldsymbol{x}}$$

  - TU = **Shannon entropy** $H(Y \,|\, \boldsymbol{x})$ of the probabilistic prediction $Y \,|\, \boldsymbol{x} \sim P_{\boldsymbol{x}}$, where $P_{\boldsymbol{x}}$ is the predictive distribution (averaged over models)

$$Y \,|\, \boldsymbol{x} \sim P_{\boldsymbol{x}} = \int p_{\theta} \, dQ(\theta) \in \mathbb{P}(\mathcal{Y}) \,.$$

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

$$H(Y \,|\, \boldsymbol{x}, \Theta) = \int H(Y \,|\, \boldsymbol{x}, \theta) \, dQ(\theta)$$

  - EU = **mutual information** $I(Y, \Theta)$ of prediction Y and model Θ.

- Recently criticised by H. (2022) ...

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), evaluation, other forms of uncertainty, applications, etc.

References

V. Balasubramanian, S.S. Ho, and V. Vovk, editors. *Conformal Prediction for Reliable Machine Learning: Theory, Adaptations and Applications*. Morgan Kaufmann, 2014.

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. H., and W. Waegeman. Pitfalls of epistemic uncertainty quantification through loss minimisation. In A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Proc. NeurIPS, Advances in Neural Information Processing Systems 35*, 2022.

E. H. Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures?, 2022.

M. Kull, T. Silva Filho, and P. Flach. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Proc. AISTATS, 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *PMLR*, pages 623–631, 2017.

Nis Meinert, Jakob Gawlikowski, and Alexander Lavin. The unreasonable effectiveness of deep evidential regression. *arXiv preprint arXiv:2205.10060*, 2022.