IPML

IMPRECISE PROBABILISTIC MACHINE LEARNING

Lecture 8: Uncertainty Quantification

Krikamol Muandet

19 December 2025

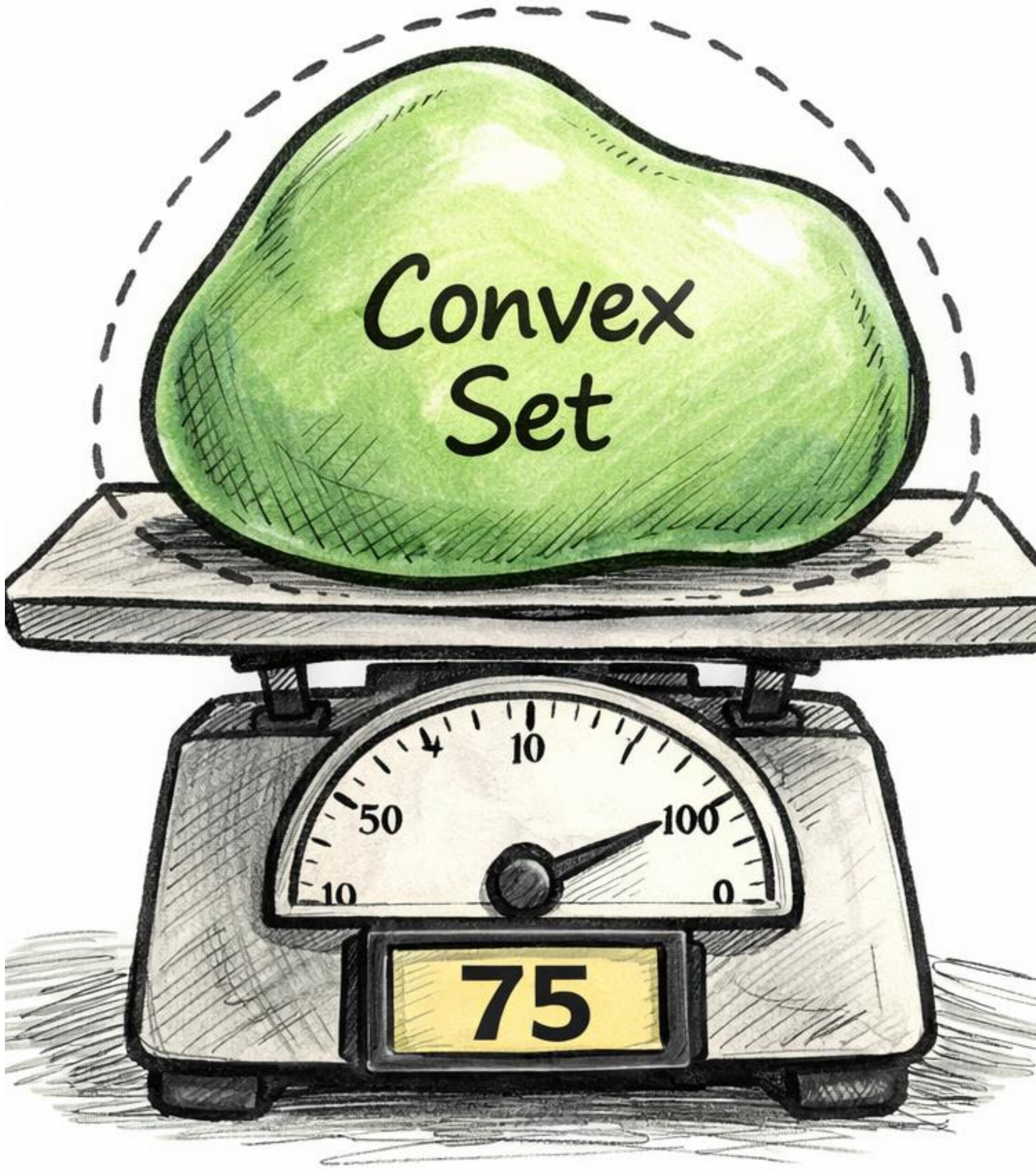
Outline

1. Uncertainty Quantification (UQ)
2. Axiomatic Characterisation
3. UQ in Machine Learning

Uncertainty Quantification

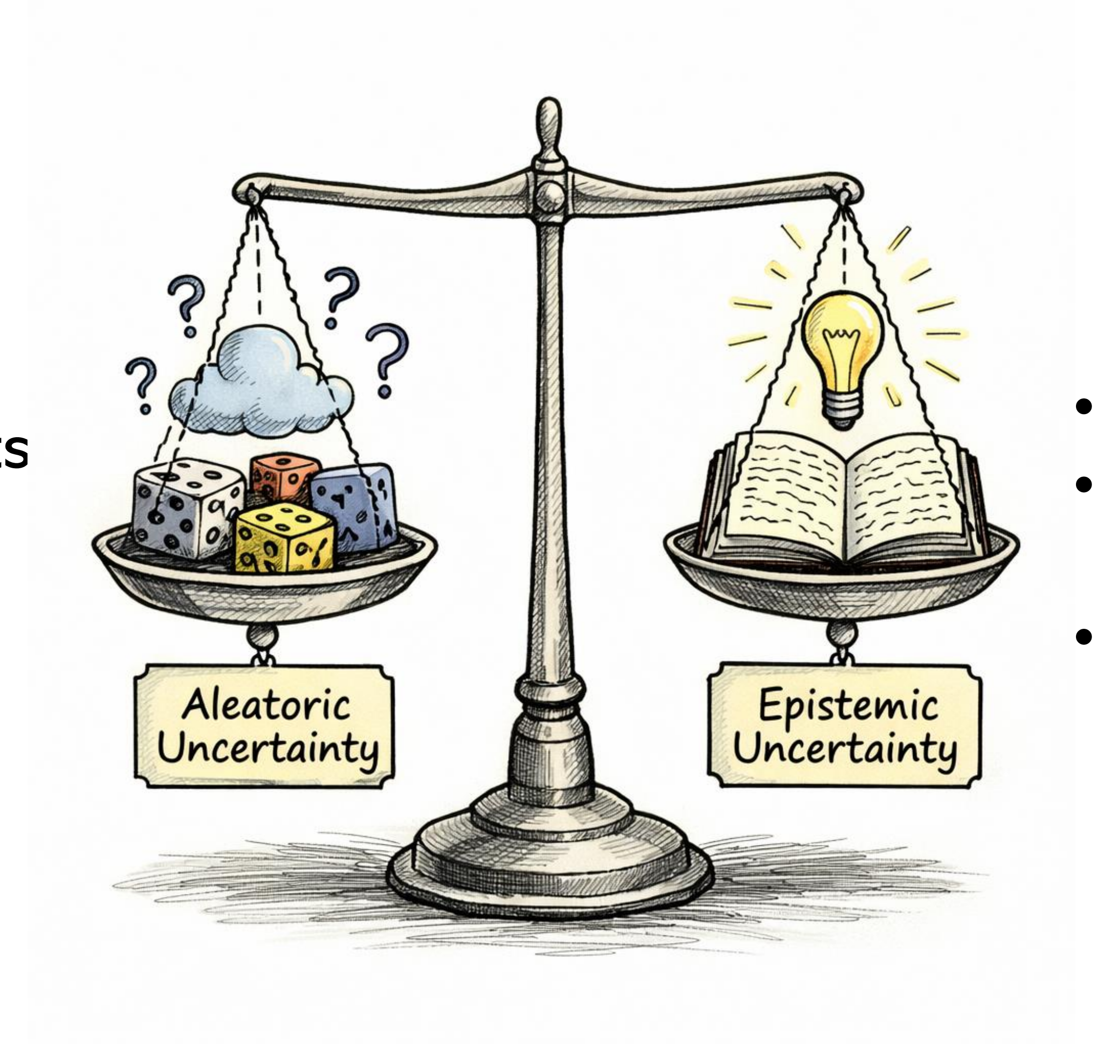
Uncertainty Quantification

- **Uncertainty quantification (UQ)** is the problem of quantifying the amount of uncertainty associated with a representation with a **single number**.
- Uncertainty quantification enables:
    - Prediction with (partial) abstention
    - Active learning
- This lecture focuses on **uncertainty quantification for credal sets**.



Two Types of Uncertainty

- Random effects
- Irreducible
- **Conflict**

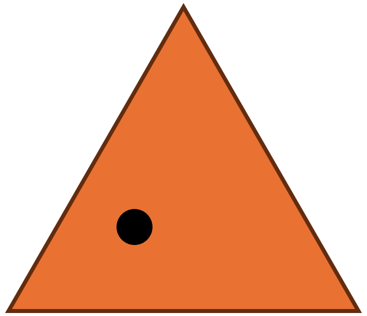


- Lack of knowledge
- Reducible with additional information
- **Non-specificity**

Uncertainty in Machine Learning

- Let $\mathcal{X}$ be an **instance space** and $\mathcal{Y}$ the set of **outcomes**.
- A classification scenario: $\mathcal{Y} = \{y_1, \dots, y_\mathrm{K}\}$ where $\triangle_\mathrm{K} = \mathbb{P}(\mathcal{Y})$ denotes the set of all probability measures on $\mathcal{Y}$.
- Given a hypothesis space $\mathcal{H}$, a **hypothesis** h is a mapping $\mathcal{X} \to \Delta_K$:

$$h^* := \mathrm{argmin}_{h \in \mathcal{H}} R(h) := \int_{\mathcal{X} \times \mathcal{Y}} \ell(h(\boldsymbol{x}), y) dP(\boldsymbol{x}, y)$$



- Given the prediction $\hat{y}(\boldsymbol{x}^\mathrm{t}) = \hat{h}(\boldsymbol{x}^\mathrm{t})$ for some test instance $\boldsymbol{x}^\mathrm{t} \in \mathcal{X}$, we are often interested in the **predictive uncertainty:**

$$\boldsymbol{p}(y \mid \boldsymbol{x}^\mathrm{t}) = \frac{\boldsymbol{p}(y \mid \boldsymbol{x}^\mathrm{t})}{\boldsymbol{p}(\boldsymbol{x}^\mathrm{t})}, \qquad \hat{h}(\boldsymbol{x}^\mathrm{t}) \approx h^*(\boldsymbol{x}^\mathrm{t}) \approx \boldsymbol{p}(y \mid \boldsymbol{x}^\mathrm{t})$$

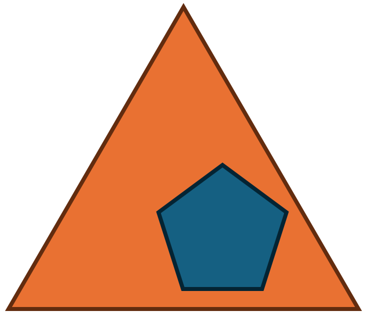
Uncertainty in Machine Learning

- A predictor $h : \mathcal{X} \to \Delta_K$ captures aleatoric but no epistemic uncertainty.
- To account for epistemic uncertainty, consider an uncertain-aware predictor:

$$h : \mathcal{X} \to [\![\Delta_K]\!]$$

- $[\![\Delta_K]\!]$ is a second-order formalism of **uncertainty about uncertainty**.
  - **Second-order probabilities** in Bayesian learning
  - **Credal sets** – (convex) sets of probability distributions

$$h(\boldsymbol{x}^{\mathrm{t}}) = Q \subseteq \Delta_K$$



Classical Measures of Uncertainty

- **Hartley Measure [Hartley, 1928]:**

$$H(A) \coloneqq \log(|A|)$$

  - Set theory: $A \subseteq \mathcal{Y}$
  - Minimal: $H(\{y\})$ (***precise information***)
  - Maximal: $H(\mathcal{Y})$ (***complete ignorance***)

- **Shannon Entropy**

$$S(q) = -\sum_{\boldsymbol{y}\in\boldsymbol{\mathcal{Y}}} q(y) \log_2 q(y)$$

  - Probability theory: $q \in \Delta_K$
  - Minimal: guaranteed outcome
  - Maximal: uniform distribution

Axiomatic Characterisation

Axiomatic Characterisation

A *measure of uncertainty* U over credal sets should obey [Abellan and Klir, 2005, Jiroušek and Shenoy, 2018]:

- **A1 (Non-negativity)**: U is non-negative and upper-bounded by $r \in \mathbb{R}$.
- **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 inequality in A5 becomes an equality when Q' and Q'' are *independent*.

Generalised Measures of Uncertainty

- **Maximal Entropy**
  [Abellan and Moral, 2003]:

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

  - Satisfies A1-A6
  - Maximal as soon as Q contains the uniform distribution
  - $S^*(Q_{\mathrm{uniform}}) = S^*(\Delta_K)$
  - **Total uncertainty**

- **Generalised Hartley**
  [Abellan and Moral, 2000]:

  $$\mathrm{GH}(\boldsymbol{Q}) = \sum_{A \subseteq \mathcal{Y}} m_Q(A) \log(|A|)$$

  - m_Q is the Möbius inverse of $v_Q(A) \coloneqq \inf_{q \in Q} q(A), \mathrm{A} \subseteq \mathcal{Y}$
  - Violates A4: $\mathrm{GH}(\{q\}) = 0$ for all $q \in \Delta_K$
  - **Epistemic uncertainty**

Disaggregation

- What about the measure of **aleatoric uncertainty** (conflict, randomness)?

$$\underset{\text{total}}{\mathrm{TU}(Q)} = \underset{\text{aleatoric}}{\mathrm{AU}(Q)} + \underset{\text{epistemic}}{\mathrm{EU}(Q)}$$

$$S^*(Q) = \underbrace{(S^*(Q) \; - \; \mathrm{GH}(Q))}_{\text{Generalised Shannon entropy, } \mathrm{GS}(Q)} + \mathrm{GH}(Q)$$

Disaggregation

- Disaggregation of total uncertainty:

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

total aleatoric epistemic

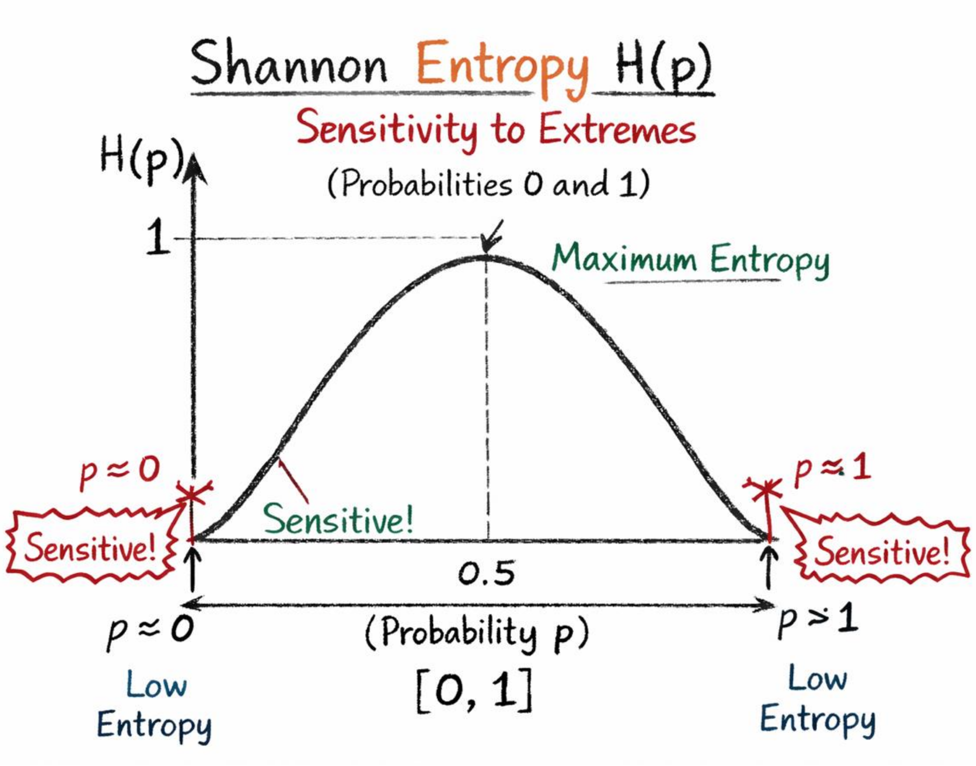
- Here, $S_*(Q) \coloneqq \min_{q \in Q} S(q)$ is that **lower Shannon entropy** which doesn't satisfy A3 (monotonicity)
- We can view $S_*(Q)$ as irreducible uncertainty: $S_*(Q)$ remains even when all epistemic uncertainty is removed.

Discussion

- A fully satisfactory disaggregation $\mathrm{TU}(Q) = \mathrm{AU}(Q) + \mathrm{EU}(Q)$ where all three measures have nice theoretical properties *has not been found*.
- S^* and GH appear to be well justified, but not for S_*.
- The decomposition $\mathrm{TU}(Q) = \mathrm{AU}(Q) + \mathrm{EU}(Q)$ may not be **semantically** meaningful.
  - GH measures imprecision regarding $q \in \Delta_K$, while Shannon entropy captures *randomness* on the level of outcomes $\mathcal{Y}$.
  - For complete ignorance ($Q = \Delta_K$), the decomposition forces aleatoric uncertainty to be zero.
- The two types of uncertainty should better be kept separate.

Discussion

- Is the set of axioms reasonable? A1-A3 may appear indisputable, but this is less the case for A4-A6.
- Most measures in the literature were proposed without regard to any specific application domain.
- The Shannon entropy itself has some undesirable properties for prediction problems (sensitivity to the extremes).
- $\mathrm{EU}(Q) := S^*(Q) - S_*(Q)$ is not **shift-invariant**.



Uncertainty Quantification in Machine Learning

Credal Uncertainty Score

- Assume the setting of binary classification: $\mathcal{Y} = \{-1, +1\}$
- Treat uncertainty as a **lack of class dominance:** *A class y dominates another class y' if y is more probable than y' for each $q \in Q$:*

$$\gamma(y, y') \coloneqq \inf_{q \in Q} \frac{q(y)}{q(y')} > 1$$

- Then, consider the **maximum degree of dominance** over all classes:

$$u \coloneqq \max(\gamma(+1, -1), \gamma(-1, +1))$$

- This is a **measure of certainty**!

Credal Uncertainty Score

- For interval-representations where we specify Q by $q(+1) \in [a, b]$:

$$\gamma(+1,-1) \coloneqq \inf_{q \in Q} \frac{q(+1)}{q(-1)} = \inf_{q \in Q} \frac{q(+1)}{1-q(+1)} = \frac{a}{1-a}$$

$$\gamma(-1,+1) \coloneqq \inf_{q \in Q} \frac{q(-1)}{q(+1)} = \inf_{q \in Q} \frac{1-q(+1)}{q(+1)} = \frac{1-b}{b}$$

- The **maximum degree of dominance** can be expressed as:

$$u(a,b) \coloneqq \max\left(\frac{a}{1-a}, \frac{1-b}{b}\right)$$

Total Measure of Predictive Uncertainty

$$\mathrm{TP}(a,b) := \frac{1}{1+u(a,b)} = \overset{\text{total}}{\min(1-a,b)} = \overset{\text{aleatoric}}{\min(a,a-b)} + \overset{\text{epistemic}}{(b-a)}$$

- This measure takes values between 0 and 1.
- Aleatoric uncertainty is upper-bounded by $1/2$.
- Full (total) uncertainty is only assumed for the interval $[0,1]$, whereas $[1/2\,,1/2]$ has a total uncertainty of only $1/2$.
- This measure avoids the problem of (partial) insensitivity of measures.

Integral Imprecise Probability Metric

- The **integral imprecise probability metric (IIPM):**

For *additive* distributions P, Q:

$$\mathrm{IPM}(P,Q) \coloneqq \sup_{f\in\mathcal{F}}\left|\int f\,dP - \int f\,dQ\right|$$

$$\mathrm{IIPM}(\mu,\nu) \coloneqq \sup_{f\in\mathcal{F}}\left|\oint f(\boldsymbol{x})d\mu(\boldsymbol{x}) - \oint f(\boldsymbol{x})d\nu(\boldsymbol{x})\right|$$

- Here, μ, ν are **Choquet capacities** and $\oint$ is a **Choquet integral**.
- For a lower probability $\underline{P}$, define the **maximum mean imprecision (MMI):**

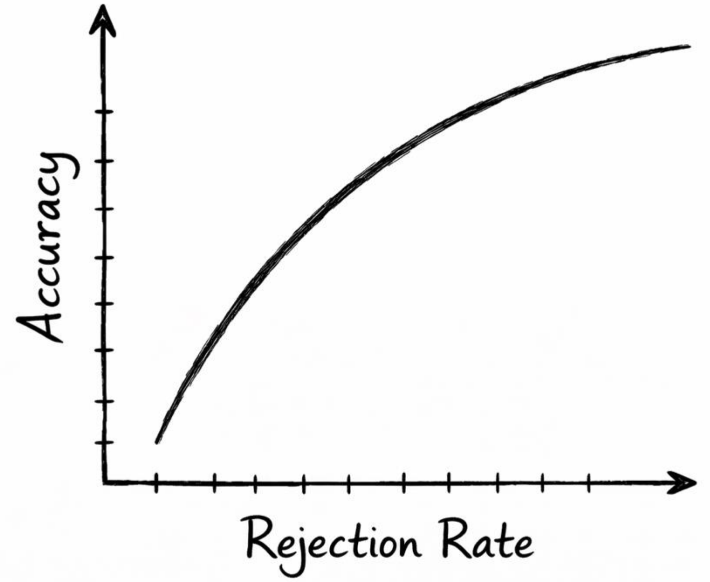
$$\begin{aligned}\mathrm{MMI}(\underline{P}) \coloneqq \mathrm{IIPM}(\underline{P},\overline{P}) &= \sup_{f\in\mathcal{F}}\left|\oint f(\boldsymbol{x})d\overline{P}(\boldsymbol{x}) - \oint f(\boldsymbol{x})d\underline{P}(\boldsymbol{x})\right| \\ &= \sup_{f\in\mathcal{F}} \int_{\underline{f}}^{\overline{f}} 1 - \left(\underline{\mathrm{P}}(\{f < t\}) + \underline{\mathrm{P}}(\{f \geq t\})\right) dt\end{aligned}$$

- We use $\mathrm{MMI}(\underline{P})$ as a measure of epistemic uncertainty.

Selective Classification

- An **accuracy-rejection curve (ARC)** represents the accuracy of a predictor as a function of the percentage of rejections.
- A predictor only makes prediction on the top $(1-p)\%$ instances that have been ranked using measure of uncertainty, abstaining on the rest.

$$\underbrace{x_{[\sigma(1)]}, x_{[\sigma(2)]}, \dots,}_{\text{predict}}\ \underbrace{x_{[\sigma(n-1)]}, x_{[\sigma(n)]}}_{\text{abstain}}$$



Selective Classification

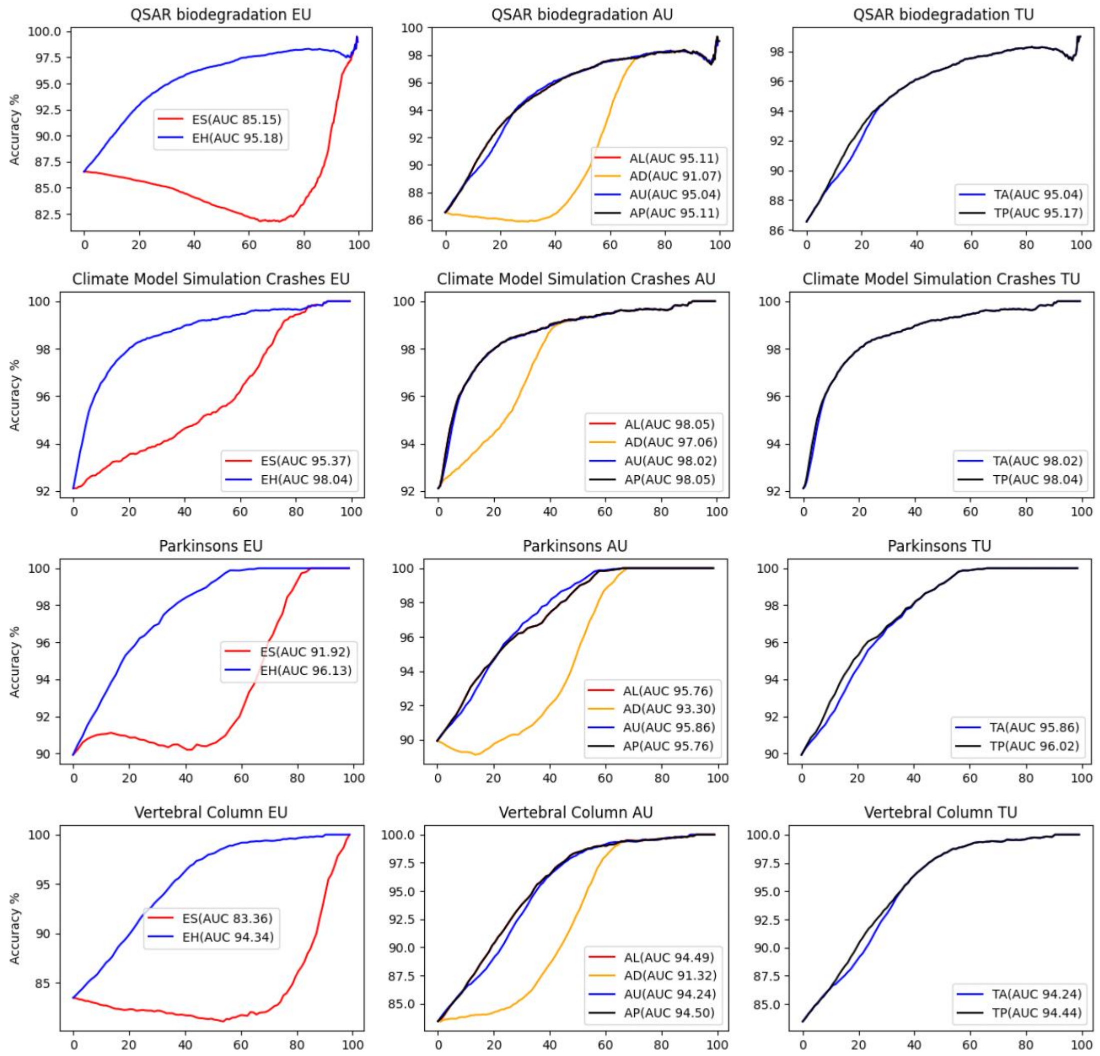


Figure 1: Accuracy-rejection curves for four data sets and different uncertainty measures (epistemic on the left, aleatoric in the middle, total on the right).

Selective Classification

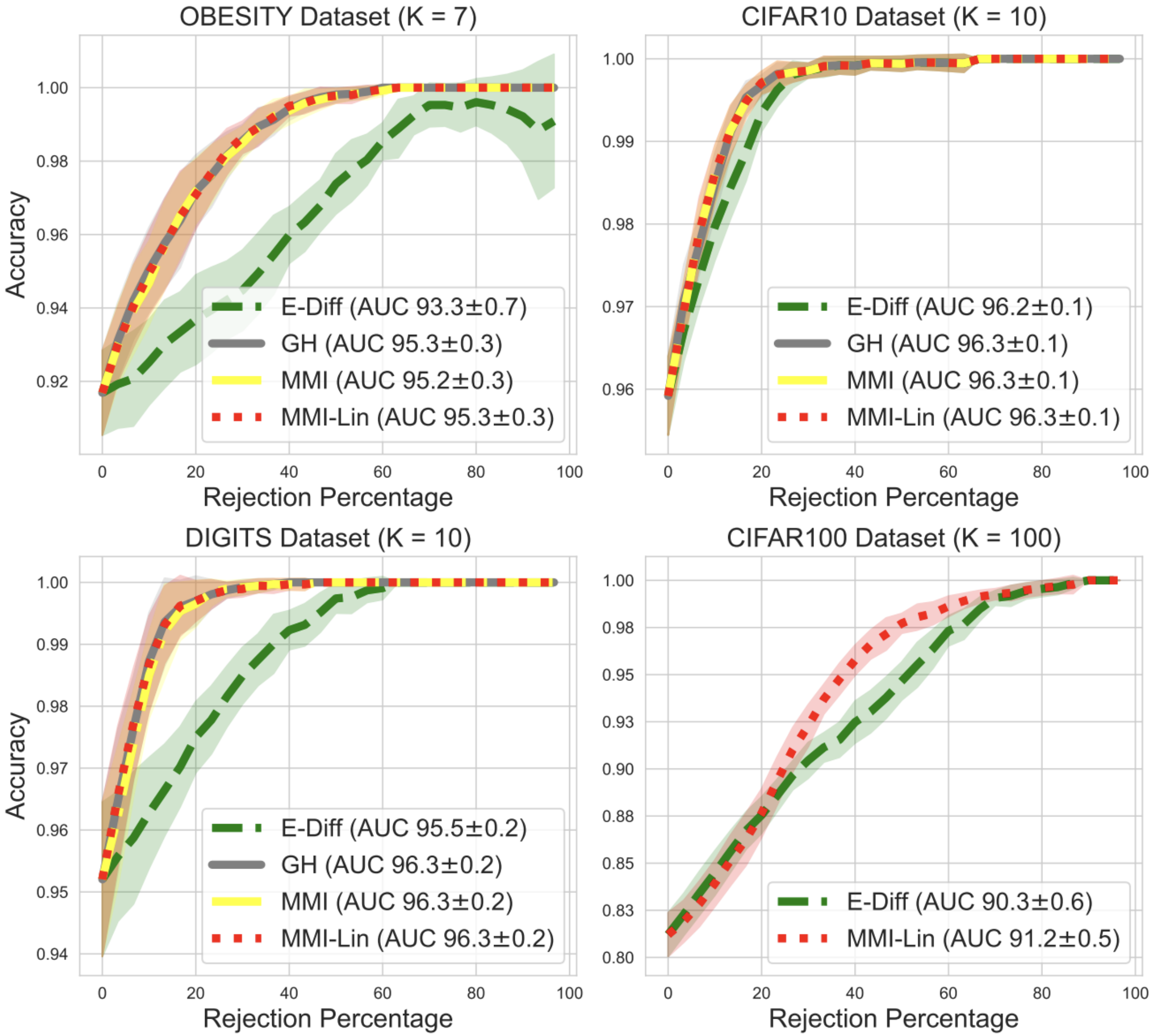


Figure 1: Accuracy-Rejection (AR) curves on four classification tasks. The area under the curve (AUC) is reported for numerical comparison. We consistently outperform entropy difference (`E-Diff`) and match the performance of Generalised Hartley (`GH`). On large-scale problems, our efficient upper bound (`MMI-Lin`) remains tractable and continues to outperform `E-Diff`.