Rethinking Aleatoric and Epistemic Uncertainty

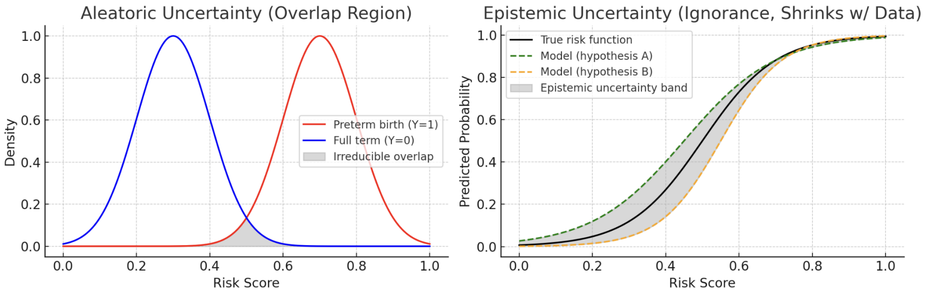
Freddie Bickford Smith, Jannik Kossen, Eleanor Trollope, Mark van der Wilk, Adam Foster, Tom Rainforth

August 29, 2025

Presented by Mian Wei

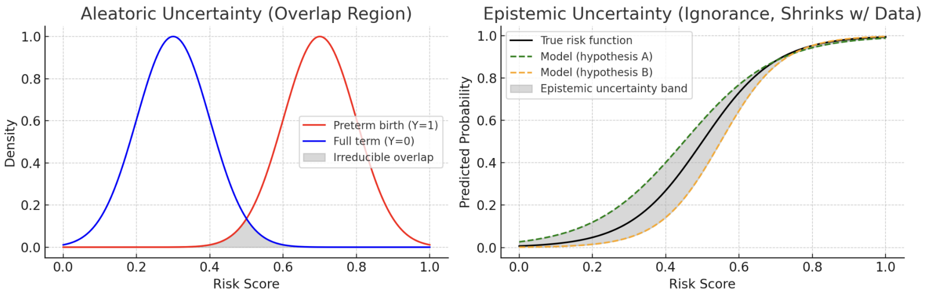
Aleatoric vs. Epistemic Uncertainty

We have a model to predict preterm birth, but there is predictive uncertainty. Should we use it cautiously, or should we collect more data and/or change the model?



- Aleatoric: Even with perfect features, some patients with identical profiles will deliver early, while others won't.
- Epistemic: Your dataset might be missing key variables or need better modeling.

Aleatoric vs. Epistemic Uncertainty



- Aleatoric: unavoidable risk → best to hedge.
- Epistemic: knowledge gap → best to reduce it.

Uncertainty Decomposition

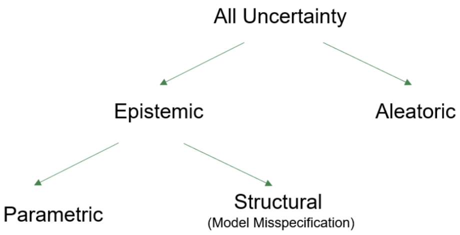


Figure: Adapted from [Che22]

- **Parametric**: uncertainty related to model parameter estimations under current model specification.
- **Structural**: the discrepancy between the assumed model specification and the true, unknown data-generating process

An area of ongoing debate

- Model-based prediction vs. the true data-generating process
- Uncertainty on unseen data
- Uncertainty vs. prediction accuracy
- Different approaches to epistemic uncertainty: density-based, information-based, variance-based, ...

Contribution of this paper:

- A critique of the popular uncertainty decomposition view
- A new decision-based framework for uncertainty

Notations

- an action, $a \in \mathcal{A}$
- a ground-truth variable, $z \in \mathcal{Z}$
- a loss function, $\ell : \mathcal{A} \times \mathcal{Z} \to \mathbb{R}$
- a policy $\pi \in \Pi$ that controls data generation
- some training data $y_{1:n} \sim p_{train}(y_{1:n}|\pi)$
- predictive model $p_n(z) = p(z; y_{1:n})$ or the predictive distribution $p_n(z) = \mathbb{E}_{p_n(\theta)}\left[p_n(z|\theta)\right]$, where $\theta \sim p_n(\theta) = p(\theta; y_{1:n})$ is a set of stochastic model parameters
- data generating process, $p_{train}(y_i|\pi(y_{<i}, y_{<i})$
- a ground-truth realization of z or a reference distribution, $p_{eval}(z)$

A Popular Decomposition View

Usual formula

$$\underbrace{\mathrm{EIG}_\theta}_{\text{“epistemic”}} = \underbrace{\mathrm{H}\left[p_n(z)\right]}_{\text{“total”}} - \underbrace{\mathbb{E}_{p_n(\theta)}\left[\mathrm{H}(p_n\left[z \mid \theta)\right]\right]}_{\text{“aleatoric”}}$$

- H: Shannon entropy
- EIG_θ: the expected information gain about θ from observing z

For finite n, epistemic and aleatoric uncertainty are only estimators of the true quantities and can be highly inaccurate.

Aleatoric uncertainty

"captures noise inherent in the observations" $\mathrm{H}[p_{\text{train}}(y_{1:n}\|\pi)]$ or $\mathrm{H}[p_{\text{eval}}(z)]$	$\neq$	"cannot be reduced even if more data were to be collected" $\mathrm{H}[p_\infty(z)]$	$\neq$	Expected parameter-conditional predictive entropy $\mathbb{E}_{p_n(\theta)}[\mathrm{H}[p_n(z\|\theta)]]$

Epistemic uncertainty

"uncertainty in the model parameters" $\mathrm{H}[p_n(\theta)]$	$\neq$	"can be explained away given enough data" $\mathrm{H}[p_n(z)] - \mathrm{H}[p_\infty(z)]$	$\neq$	Expected information gain in the model parameters $\mathrm{H}[p_n(z)] - \mathbb{E}_{p_n(\theta)}[\mathrm{H}[p_n(z\|\theta)]]$

Model world $\neq$ Real world

Decision-Based Fr

- Bayes-optimal action:

$$a_n^* = \arg\min_a \mathbb{E}_{p_n(z)}[\ell(a, z)]$$

- Loss-grounded uncertainty measure:

$$h[p_n(z)] = \mathbb{E}_{p_n(z)}[\ell(a_n^*, z)]$$

Takeaway: The choice of loss (and thus the uncertainty measure) depends on the decision problem at hand.

Expected Uncertainty Reduction (EUR)

Definition:

$$UR_z(y^+_{1:m}) = h[p_n(z)] - h[p_{n+m}(z)].$$

$$EUR^{\text{true}}_z(\pi, m) = \mathbb{E}_{p_{\text{train}}(y^+_{1:m}|\pi)}[\, UR_z(y^+_{1:m})\,].$$

Expected Uncertainty Reduction (EUR)

Definition:

$$UR_z(y^+_{1:m}) = h[p_n(z)] - h[p_{n+m}(z)].$$

$$EUR^{\text{true}}_z(\pi, m) = \mathbb{E}_{p_{\text{train}}(y^+_{1:m}|\pi)}[\, UR_z(y^+_{1:m})\,].$$

As $m \to \infty$, the decomposition:

$$h[p_n(z)] = \underbrace{EUR^{\text{true}}_z(\pi, \infty)}_{\text{Reducible}} + \underbrace{\mathbb{E}_{p_{\text{train}}(y^+_{1:m}|\pi)}[h[p_\infty(z)]]}_{\text{Irreducible}}.$$

Expected Uncertainty Reduction (EUR)

Definition:

$$UR_z(y^+_{1:m}) = h[p_n(z)] - h[p_{n+m}(z)].$$

$$EUR^{\text{true}}_z(\pi, m) = \mathbb{E}_{p_{\text{train}}(y^+_{1:m}|\pi)}\left[\, UR_z(y^+_{1:m}) \,\right].$$

As $m \to \infty$, the decomposition:

$$h[p_n(z)] = \underbrace{EUR^{\text{true}}_z(\pi, \infty)}_{\text{Reducible}} + \underbrace{\mathbb{E}_{p_{\text{train}}(y^+_{1:m}|\pi)}[h[p_\infty(z)]]}_{\text{Irreducible}}.$$

Compared with popular split:

$$H[p_n(z)] = \underbrace{\mathbb{E}_{p_n(\theta)}[H(p_n(z \mid \theta))]}_{\text{aleatoric}} + \underbrace{H[p_n(z)] - \mathbb{E}_{p_n(\theta)}[H(p_n(z \mid \theta))]}_{\text{BALD/epistemic}}$$

Expected Uncertainty Reduction (EUR)

Definition:

$$UR_z(y^+_{1:m}) = h[p_n(z)] - h[p_{n+m}(z)].$$

$$EUR^{\text{true}}_z(\pi, m) = \mathbb{E}_{p_{\text{train}}(y^+_{1:m}|\pi)}[\, UR_z(y^+_{1:m})\,].$$

As $m \to \infty$, the decomposition:

$$h[p_n(z)] = \underbrace{EUR^{\text{true}}_z(\pi, \infty)}_{\text{Reducible}} + \underbrace{\mathbb{E}_{p_{\text{train}}(y^+_{1:m}|\pi)}[h[p_\infty(z)]]}_{\text{Irreducible}}.$$

Compared with popular split:

$$H[p_n(z)] = \underbrace{\mathbb{E}_{p_n(\theta)}[H(p_n(z \mid \theta))]}_{\text{aleatoric}} + \underbrace{H[p_n(z)] - \mathbb{E}_{p_n(\theta)}[H(p_n(z \mid \theta))]}_{\text{BALD/epistemic}}$$

- Paper's: decision/loss grounded, any learner, depends on data process.
- BALD: an *estimator*, not a universal decomposition.

EUR in Practice

Problem: We cannot access the true data-generating process or infinite data.

So we approximate with:

- Use model-based simulator $p_n(y^+_{1:m} \mid \pi')$ instead of true p_{train}.
- Use approximate update $q_{n+m}(z)$ instead of true $p_{n+m}(z)$.

EUR in Practice

Problem: We cannot access the true data-generating process or infinite data.

So we approximate with:

- Use model-based simulator $p_n(y^+_{1:m} \mid \pi')$ instead of true p_{train}.
- Use approximate update $q_{n+m}(z)$ instead of true $p_{n+m}(z)$.

Estimator:

$$EUR^{\text{est}}_z(\pi', m) = h[p_n(z)] - \mathbb{E}_{p_n(y^+_{1:m} \mid \pi')}\big[h[q_{n+m}(z)]\big].$$

EUR in Practice

Problem: We cannot access the true data-generating process or infinite data.

So we approximate with:

- Use model-based simulator $p_n(y^+_{1:m} \mid \pi')$ instead of true p_{train}.
- Use approximate update $q_{n+m}(z)$ instead of true $p_{n+m}(z)$.

Estimator:

$$EUR_z^{\text{est}}(\pi', m) = h[p_n(z)] - \mathbb{E}_{p_n(y^+_{1:m} \mid \pi')}\big[h[q_{n+m}(z)]\big].$$

Sources of error:

1. Simulator mismatch (p_n vs p_{train}).
2. Update approximation (q_{n+m} vs p_{n+m}).

Some Concepts

Predictive Uncertainty

- $h[p_n(z)] = \mathbb{E}_{p_n(z)}[\ell(a_n^*, z)]$
- How uncertain *my model* thinks the future is
- Subjective, depends on $p_n(z)$

Some Concepts

Predictive Uncertainty

- $h[p_n(z)] = \mathbb{E}_{p_n(z)}[\ell(a_n^*, z)]$
- How uncertain *my model* thinks the future is
- Subjective, depends on $p_n(z)$

Predictive Performance

- $\mathrm{Perf}(p_n) = \mathbb{E}_{p_{\mathrm{eval}}}[\ell(a_n^*, z)]$
- How good the predictions are compared with reality
- Requires $p_{\mathrm{eval}}(z)$

Some Concepts

Predictive Uncertainty

- $h[p_n(z)] = \mathbb{E}_{p_n(z)}[\ell(a_n^*, z)]$
- How uncertain *my model* thinks the future is
- Subjective, depends on $p_n(z)$

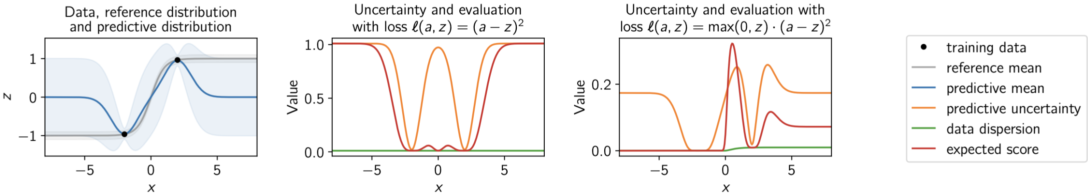
Predictive Performance

- $\mathrm{Perf}(p_n) = \mathbb{E}_{p_{\mathrm{eval}}}[\ell(a_n^*, z)]$
- How good the predictions are compared with reality
- Requires $p_{\mathrm{eval}}(z)$

Data Dispersion

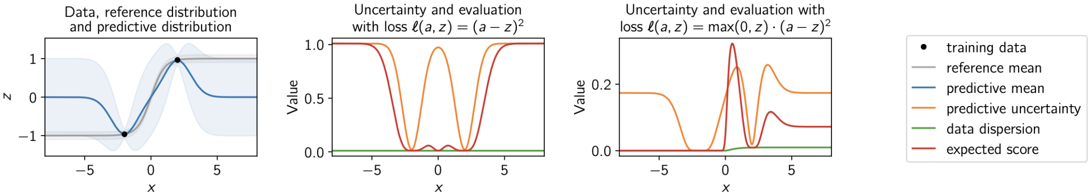
- Dispersion = entropy/variance of $p_{\mathrm{eval}}(z)$
- How random the world really is, regardless of the model
- World-based, not model-based

Some Concepts



Prop 1: Bayes estimator under quadratic loss = the posterior mean.
Prop 2: $h[p_n(z)]$ is the Bayes estimator of expected performance under p_{eval}.
Prop 3: $\mathbb{E}_{p_n(\theta)}[h[p_n(z \mid \theta))]$ is the Bayes estimator of data dispersion $h[p_{\text{eval}}(z)]$.

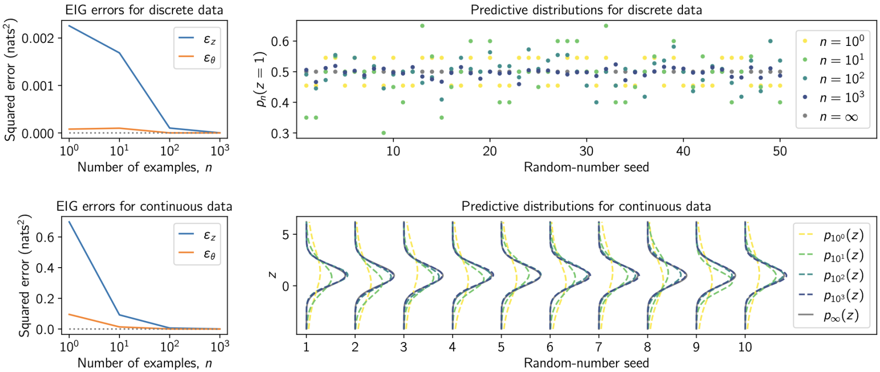
Some Concepts



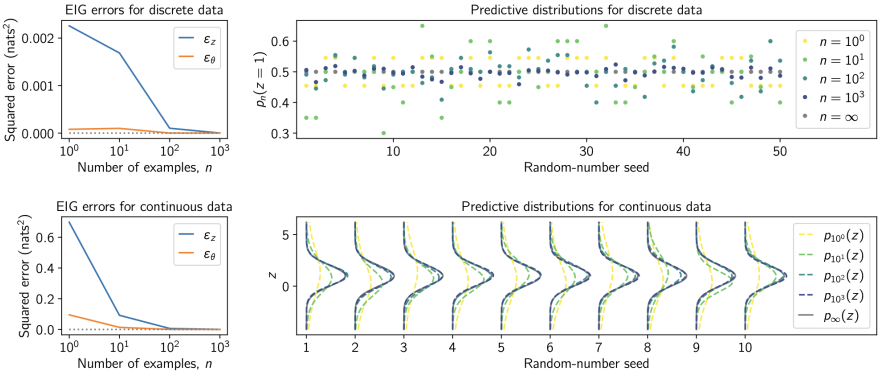
Prop 1: Bayes estimator under quadratic loss = the posterior mean.
Prop 2: $h[p_n(z)]$ is the Bayes estimator of expected performance under p_{eval}.
Prop 3: $\mathbb{E}_{p_n(\theta)}[h[p_n(z \mid \theta))]$ is the Bayes estimator of data dispersion $h[p_{\text{eval}}(z)]$.

Takeaway: Model-based uncertainty $\neq$ truth; only *estimators* of performance/dispersion.

Re-reading BALD



Re-reading BALD



Takeaway:

- BALD is not the "epistemic uncertainty" truth. It's an estimator, and often poor for long-run reducibility.
- But good proxy for *short-run parameter IG* $\rightarrow$ explains success in active learning.

Decision-Theoretic Framework

Learning $\rightarrow$ Reasoning $\rightarrow$

Decision problem
$\ell : \mathcal{A} \times \mathcal{Z} \to \mathbb{R}$

Training data
$y_{1:n} \sim p_{\text{train}}(y_{1:n}|\pi)$

Predictive model
$p_n(z) = p(z; y_{1:n})$

Bayes-optimal action
$a_n^* = \arg\min_{a \in \mathcal{A}} \mathbb{E}_{p_n(z)}[\ell(a, z)]$

Uncertainty quantification

Predictive uncertainty
$h[p_n(z)] = \mathbb{E}_{p_n(z)}[\ell(a_n^*, z)]$

Expected uncertainty reduction

$$\begin{aligned} \text{EUR}_z^{\text{true}}(\pi, m) &= h[p_n(z)] - \mathbb{E}_{p_{\text{train}}(y_{1:m}^+|\pi)}[h[p_{n+m}(z)]] \\ &\approx h[p_n(z)] - \mathbb{E}_{p_n(y_{1:m}^+|\pi')}[h[q_{n+m}(z)]] \end{aligned}$$

Model evaluation

Proper scoring rule
$s(p_n, z) = \ell(a_n^*, z)$

Discrepancy function
$d(p_n, p_{\text{eval}}) = \mathbb{E}_{p_{\text{eval}}(z)}[s(p_n, z)] - h[p_{\text{eval}}(z)]$

Accommodation

Is it worth reading? Maybe.

1. Uncertainty is decision-specific, not one-size-fits-all.
2. Decomposition: reducible vs. irreducible (depends on DGS, not just model).
3. Model-based quantities are *estimators*, not ground truths.
4. BALD works in practice by estimating short-run parameter IG.

Cons:

1. The authors claim their decomposition is better, but the argument is unconvincing, as there are no experiments, no rigorous proof, and no empirical validation.
2. More like a conceptual critique + framework clarification paper.

Thank you! Any questions?

Shuo Chen.
Introduction and exemplars of uncertainty decomposition.
arXiv preprint arXiv:2211.15475, 2022.