The limits of aleatoric and epistemic uncertainty representations

Seminar on Modern Numerical Methods in Physics 28/01/2026

Mira Jürgens

PhD student
Department of Data Science and Mathematical Modeling
Ghent University
Promotor: Willem Waegeman

BioML

GHENT UNIVERSITY

Contributors

Mira Jürgens

Ghent University

Sebastian Jimènez

Ghent University

Willem Waegeman

Ghent University

Uncertainty estimation in machine learning

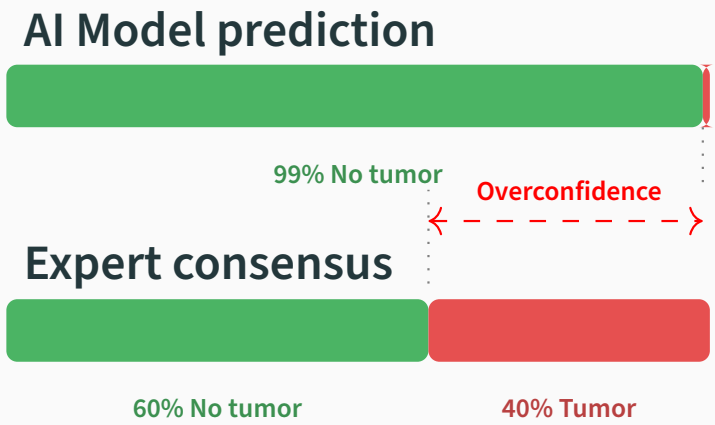
Common machine learning models are often overconfident.

Common machine learning models are often overconfident.

AI Model prediction

99% No tumor

Common machine learning models are often overconfident.



Aleatoric vs epistemic uncertainty

Aleatoric vs epistemic uncertainty

Assume we have a probabilistic model $f : \mathcal{X} \to \mathbb{P}(\mathcal{Y})$ that estimates the conditional distribution $\mathbb{P}_{Y|X}$ of a r.v. Y given X. We can distinguish between **two types of uncertainty**:

Aleatoric vs epistemic uncertainty

Assume we have a probabilistic model $f : \mathcal{X} \to \mathbb{P}(\mathcal{Y})$ that estimates the conditional distribution $\mathbb{P}_{Y|X}$ of a r.v. Y given X. We can distinguish between **two types of uncertainty**:

Aleatoric uncertainty
irreducible randomness in $\mathbb{P}_{Y|X}$

Aleatoric vs epistemic uncertainty

Assume we have a probabilistic model $f : \mathcal{X} \to \mathbb{P}(\mathcal{Y})$ that estimates the conditional distribution $\mathbb{P}_{Y|X}$ of a r.v. Y given X. We can distinguish between **two types of uncertainty**:

Aleatoric uncertainty
irreducible randomness in $\mathbb{P}_{Y|X}$

Epistemic uncertainty
reducible ignorance about the true $\mathbb{P}_{Y|X}$

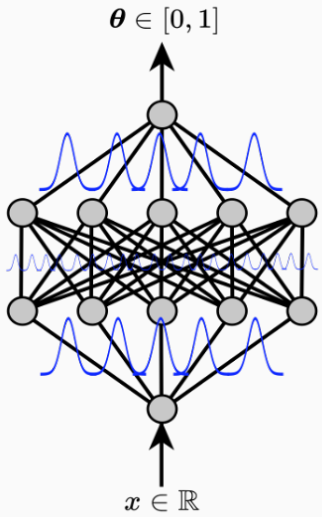
Different representations of epistemic uncertainty

Different representations of epistemic uncertainty

Distribution-based

second-order distribution $q(\boldsymbol{\theta}|\boldsymbol{x})$

- e.g. Bayesian neural networks, evidential models, ensembles

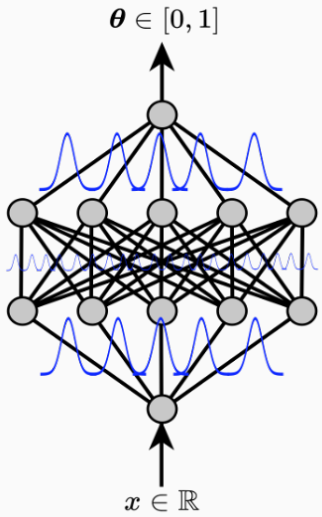


Different representations of epistemic uncertainty

Distribution-based

second-order distribution $q(\boldsymbol{\theta}|\boldsymbol{x})$

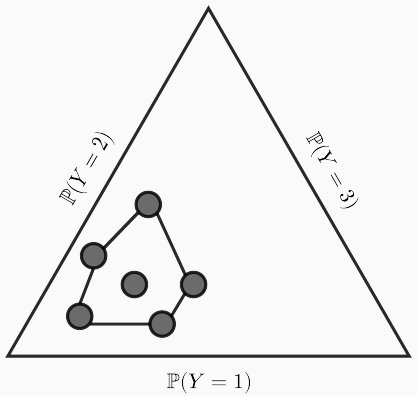
- e.g. Bayesian neural networks, evidential models, ensembles



Set-based

(convex) set $\mathcal{S}$ *of probability distributions*

- e.g. ensemble models, credal/interval networks

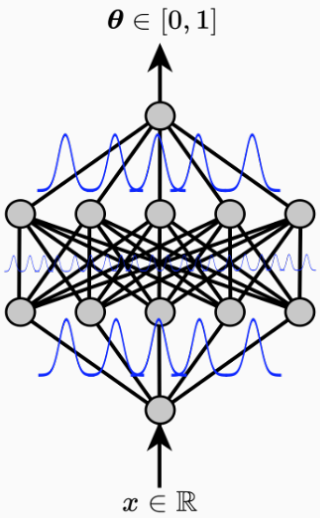


Different representations of epistemic uncertainty

Distribution-based

second-order distribution $q(\boldsymbol{\theta}|\boldsymbol{x})$

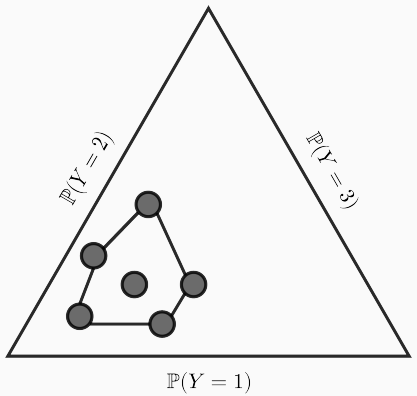
- e.g. Bayesian neural networks, evidential models, ensembles



Set-based

(convex) set $\mathcal{S}$ of probability distributions

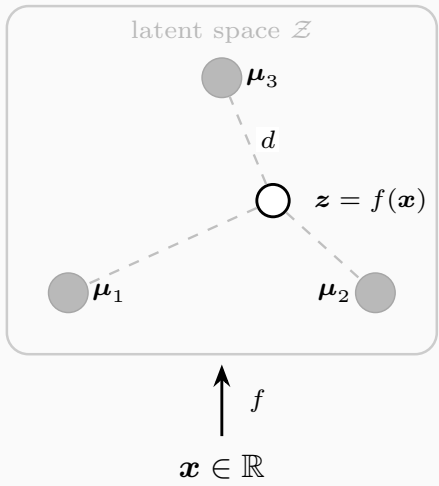
- e.g. ensemble models, credal/interval networks



Distance-based

scalar uncertainty value based on distance to latent embeddings

- e.g. Deterministic Uncertainty Quantification, Mahalanobis method

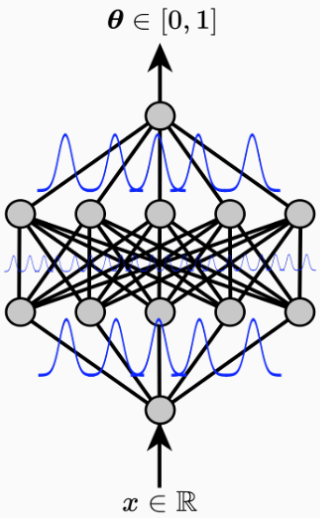


Different representations of epistemic uncertainty

Distribution-based

second-order distribution $q(\boldsymbol{\theta}|\boldsymbol{x})$

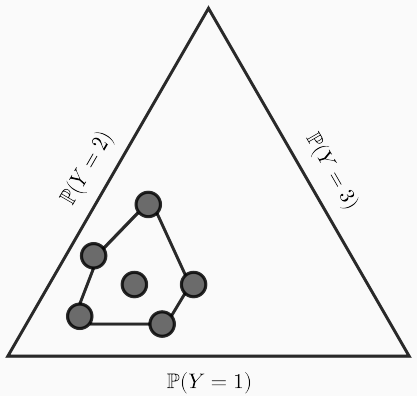
- e.g. Bayesian neural networks, evidential models, ensembles



Set-based

(convex) set $\mathcal{S}$ of probability distributions

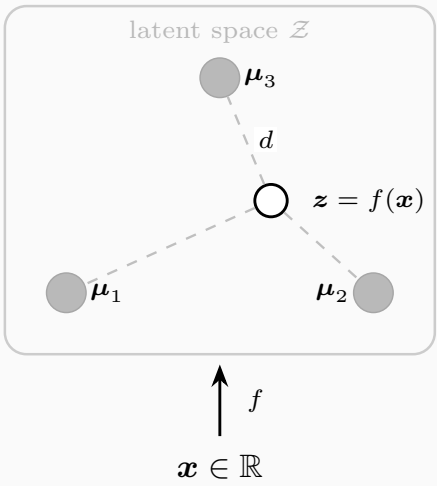
- e.g. ensemble models, credal/interval networks



Distance-based

scalar uncertainty value based on distance to latent embeddings

- e.g. Deterministic Uncertainty Quantification, Mahalanobis method



When do we know that a given representation is a *valid* representation of EU?

2. Limitations of uncertainty disentanglement

Epistemic uncertainty estimation is fundamentally incomplete.

Why machine learning models fail to fully capture epistemic uncertainty

Sebastián Jimenez[1*†], Mira Jürgens[1†] and Willem Waegeman[1]

[1*]Department of Data Analysis and Mathematical Modeling, Ghent University, Coupure Links 653, Ghent, 9000, Belgium.

Epistemic uncertainty estimation is fundamentally incomplete.

Why machine learning models fail to fully capture epistemic uncertainty

Sebastián Jimenez[1*†], Mira Jürgens[1†] and Willem Waegeman[1]

[1*]Department of Data Analysis and Mathematical Modeling, Ghent University, Coupure Links 653, Ghent, 9000, Belgium.

We address **2 main limitations** of epistemic uncertainty representations:

Epistemic uncertainty estimation is fundamentally incomplete.

Why machine learning models fail to fully capture epistemic uncertainty

Sebastián Jimenez[1*†], Mira Jürgens[1†] and Willem Waegeman[1]

[1*]Department of Data Analysis and Mathematical Modeling, Ghent University, Coupure Links 653, Ghent, 9000, Belgium.

We address **2 main limitations** of epistemic uncertainty representations:

1. The *bias* is unaccounted for and can lead to false trust in the uncertainty estimates.

Epistemic uncertainty estimation is fundamentally incomplete.

Why machine learning models fail to fully capture epistemic uncertainty

Sebastián Jimenez[1*†], Mira Jürgens[1†] and Willem Waegeman[1]

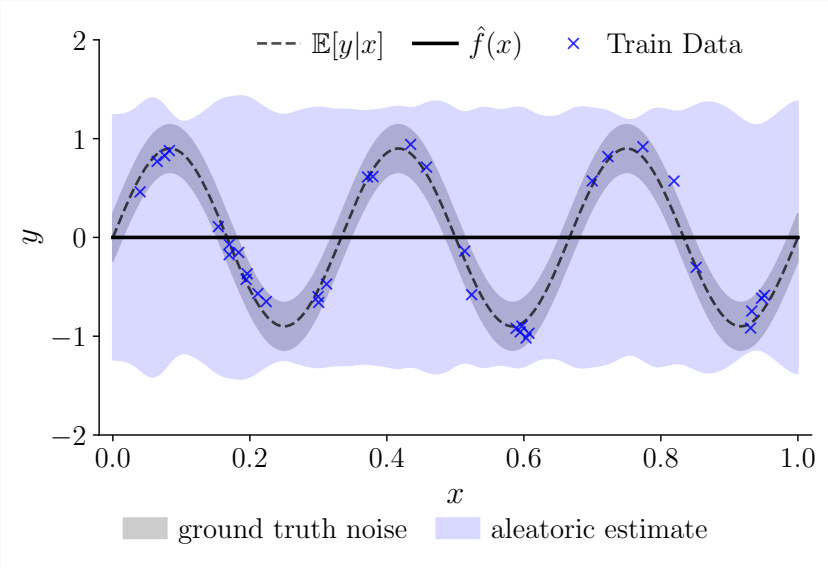
[1*]Department of Data Analysis and Mathematical Modeling, Ghent University, Coupure Links 653, Ghent, 9000, Belgium.

We address **2 main limitations** of epistemic uncertainty representations:

1. The *bias* is unaccounted for and can lead to false trust in the uncertainty estimates.
2. The *variance* is only captured partly, often rather capturing *procedural* uncertainty rather than *data uncertainty*.

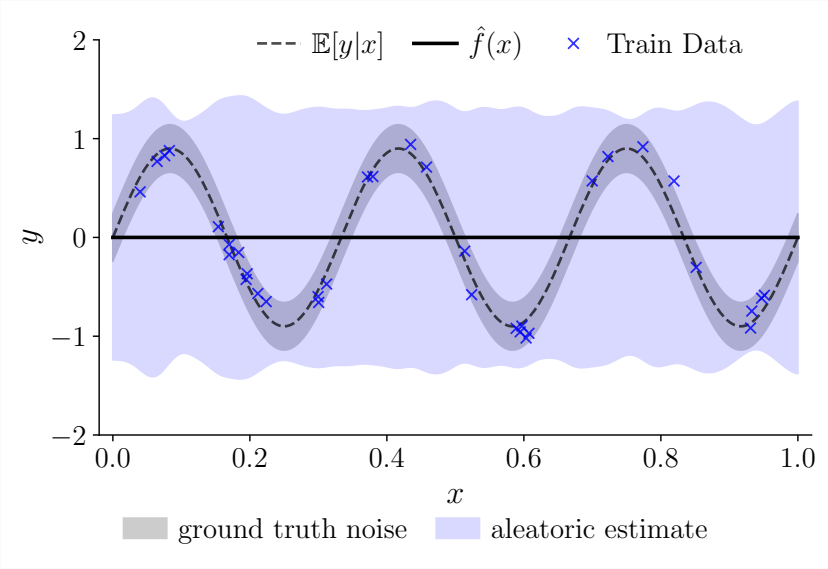
Limitations of uncertainty disentanglement

High model bias

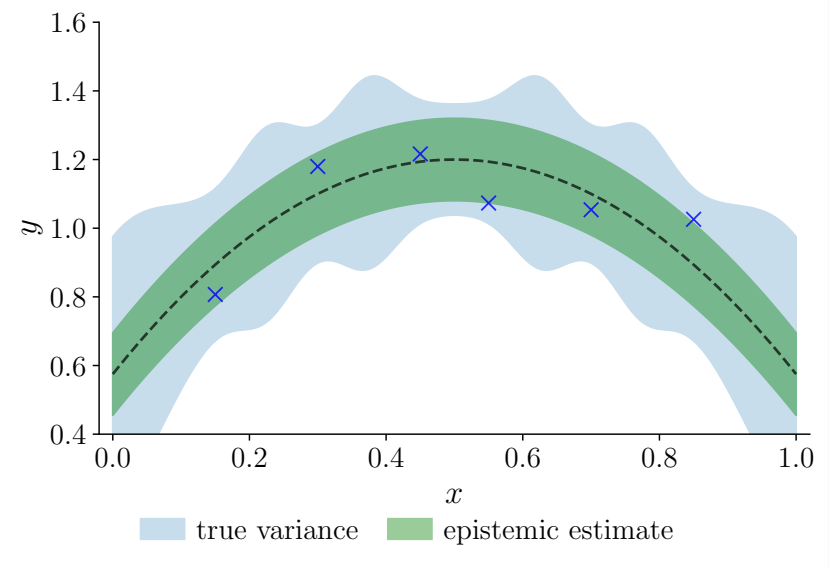


Limitations of uncertainty disentanglement

High model bias



Uncaptured parts of variance



Notation

Symbol	Meaning
Supervised Learning	
$\mathcal{X} \subseteq \mathbb{R}^D$	instance space
$\mathcal{Y} = \{1, \dots K\}$ or $\mathbb{R}$	target space
$\mathcal{D}_N = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$	dataset of N i.i.d. realizations
Probabilistic Modeling	
$p(y\|\boldsymbol{x})$	(true) conditional distribution
$\boldsymbol{\theta} \in \Theta$	distributional parameters
$\mathcal{H} := \{p_\theta : \mathcal{X} \to \mathbb{P}(\mathcal{Y})\}$	hypothesis space
$\hat{p}(y\|\boldsymbol{x}), \hat{p}^*(y\|\boldsymbol{x})$	estimate and Bayes-optimal model
Uncertainty Representation	
$q(\boldsymbol{\theta}\|\boldsymbol{x})$	second-order distribution over $\boldsymbol{\theta}$

Second-order distributions model uncertainty over first-order parameters.

Let $q : \mathcal{X} \to \mathcal{P}(\Theta)$ be a second-order distribution.

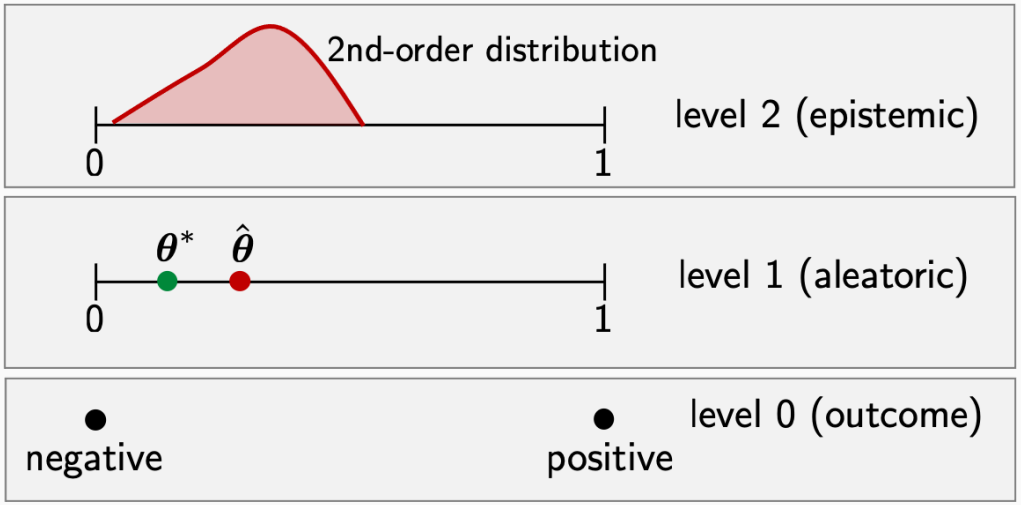


Figure 1: Levels of uncertainty[1]

[1] Wimmer et al., "Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures?"

Second-order distributions model uncertainty over first-order parameters.

Let $q : \mathcal{X} \to \mathcal{P}(\Theta)$ be a second-order distribution.

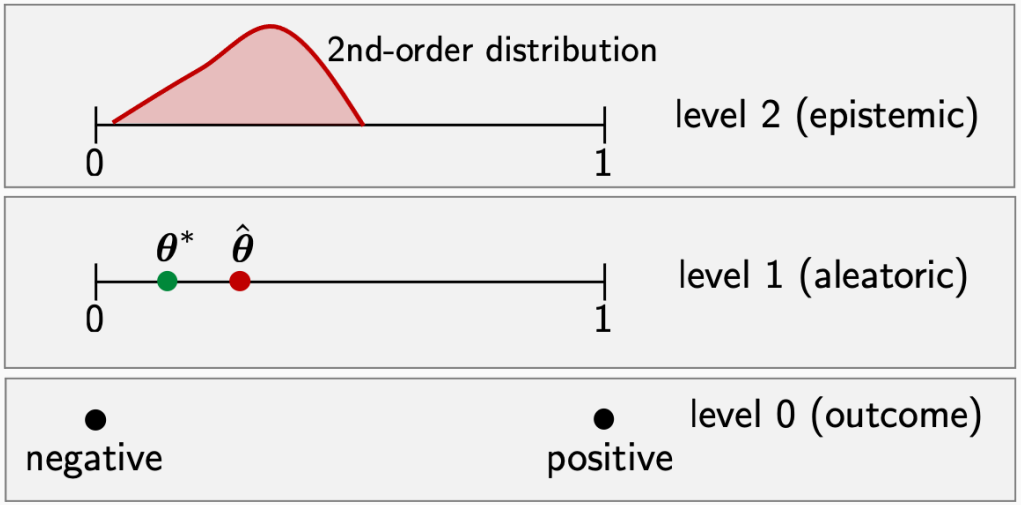


Figure 1: Levels of uncertainty[1]

The predictive distribution $\hat{p}(y|\boldsymbol{x})$ can be derived by model averaging:

$$\hat{p}(y|\boldsymbol{x}) = \int_{\Theta} p(y|\boldsymbol{\theta}, \boldsymbol{x})\, q(\boldsymbol{\theta}|\boldsymbol{x})\, d\boldsymbol{\theta}.$$

[1] Wimmer et al., "Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures?"

Uncertainty Disentanglement of second-order distributions

Let $q(\boldsymbol{\theta}|\boldsymbol{x})$ be a (predicted) second-order distribution and

$$\hat{y} \sim \hat{p}(y|\boldsymbol{x}) = \int_{\Theta} p(y|\boldsymbol{\theta}, \boldsymbol{x})\, q(\boldsymbol{\theta}|\boldsymbol{x})\, d\boldsymbol{\theta}.$$

How can we get separate estimates of aleatoric and epistemic uncertainty?

[2] Depeweg et al., "Decomposition of Uncertainty in Bayesian Deep Learning for Efficient and Risk-sensitive Learning".

Uncertainty Disentanglement of second-order distributions

Let $q(\boldsymbol{\theta}|\boldsymbol{x})$ be a (predicted) second-order distribution and

$$\hat{y} \sim \hat{p}(y|\boldsymbol{x}) = \int_{\Theta} p(y|\boldsymbol{\theta}, \boldsymbol{x})\, q(\boldsymbol{\theta}|\boldsymbol{x})\, d\boldsymbol{\theta}.$$

How can we get separate estimates of aleatoric and epistemic uncertainty?

One common method is the *variance-based decomposition*[2]:

$$\mathrm{Var}(\hat{y}|\boldsymbol{x}) = \underbrace{\mathbb{E}_{\boldsymbol{\theta}}\left[\mathrm{Var}(\hat{y}|\boldsymbol{x}, \boldsymbol{\theta})\right]}_{\text{aleatoric estimate}} + \underbrace{\mathrm{Var}_{\boldsymbol{\theta}}\left(\mathbb{E}[\hat{y}|\boldsymbol{x}, \boldsymbol{\theta}]\right)}_{\text{epistemic estimate}}. \tag{1}$$

[2] Depeweg et al., "Decomposition of Uncertainty in Bayesian Deep Learning for Efficient and Risk-sensitive Learning".

Sources of epistemic uncertainty

[3] Dawid, “The geometry of proper scoring rules”.

Sources of epistemic uncertainty

Define $\mathcal{D}_N \sim P^N$ $\gamma \sim P^\gamma$ as the random variables reflecting data- and procedural variability. For a hypothesis space $\mathcal{H}$, define

$$\hat{p}_{N,\gamma}(y|\boldsymbol{x})$$

as the prediction under data and procedural uncertainty.

[3] Dawid, "The geometry of proper scoring rules".

Sources of epistemic uncertainty

Define $\mathcal{D}_N \sim P^N$ $\gamma \sim P^\gamma$ as the random variables reflecting data- and procedural variability. For a hypothesis space $\mathcal{H}$, define

$$\hat{p}_{N,\gamma}(y|\boldsymbol{x})$$

as the prediction under data and procedural uncertainty.
For $x \in \mathcal{X}$, there are **3 different well-known sources of epistemic uncertainty**:

[3] Dawid, "The geometry of proper scoring rules".

Sources of epistemic uncertainty

Define $\mathcal{D}_N \sim P^N$ $\gamma \sim P^\gamma$ as the random variables reflecting data- and procedural variability. For a hypothesis space $\mathcal{H}$, define

$$\hat{p}_{N,\gamma}(y|\boldsymbol{x})$$

as the prediction under data and procedural uncertainty.
For $\boldsymbol{x} \in \mathcal{X}$, there are **3 different well-known sources of epistemic uncertainty**:

1. **Model uncertainty:** discrepancy between the *dual mean*[3] $\bar{p}^*(y|\boldsymbol{x})$ and the ground truth $p(y|\boldsymbol{x})$

[3] Dawid, "The geometry of proper scoring rules".

Sources of epistemic uncertainty

Define $\mathcal{D}_N \sim P^N$ $\gamma \sim P^\gamma$ as the random variables reflecting data- and procedural variability. For a hypothesis space $\mathcal{H}$, define

$$\hat{p}_{N,\gamma}(y|\boldsymbol{x})$$

as the prediction under data and procedural uncertainty.
For $\boldsymbol{x} \in \mathcal{X}$, there are **3 different well-known sources of epistemic uncertainty**:

1. **Model uncertainty:** discrepancy between the *dual mean*[3] $\bar{p}^*(y|\boldsymbol{x})$ and the ground truth $p(y|\boldsymbol{x})$
2. **Estimation uncertainty:** variance of $p_{N,\gamma}(y|\boldsymbol{x})$ around $\bar{p}^*(y|\boldsymbol{x})$

[3] Dawid, "The geometry of proper scoring rules".

Sources of epistemic uncertainty

Define $\mathcal{D}_N \sim P^N$ $\gamma \sim P^\gamma$ as the random variables reflecting data- and procedural variability. For a hypothesis space $\mathcal{H}$, define

$$\hat{p}_{N,\gamma}(y|\boldsymbol{x})$$

as the prediction under data and procedural uncertainty.
For $\boldsymbol{x} \in \mathcal{X}$, there are **3 different well-known sources of epistemic uncertainty**:

1. **Model uncertainty:** discrepancy between the *dual mean*[3] $\bar{p}^*(y|\boldsymbol{x})$ and the ground truth $p(y|\boldsymbol{x})$
2. **Estimation uncertainty:** variance of $p_{N,\gamma}(y|\boldsymbol{x})$ around $\bar{p}^*(y|\boldsymbol{x})$
3. **Distributional uncertainty:** distribution shift between training and test data (i.e. new ground truth $\tilde{p}(y|\boldsymbol{x})$)

[3] Dawid, "The geometry of proper scoring rules".

Sources of epistemic uncertainty

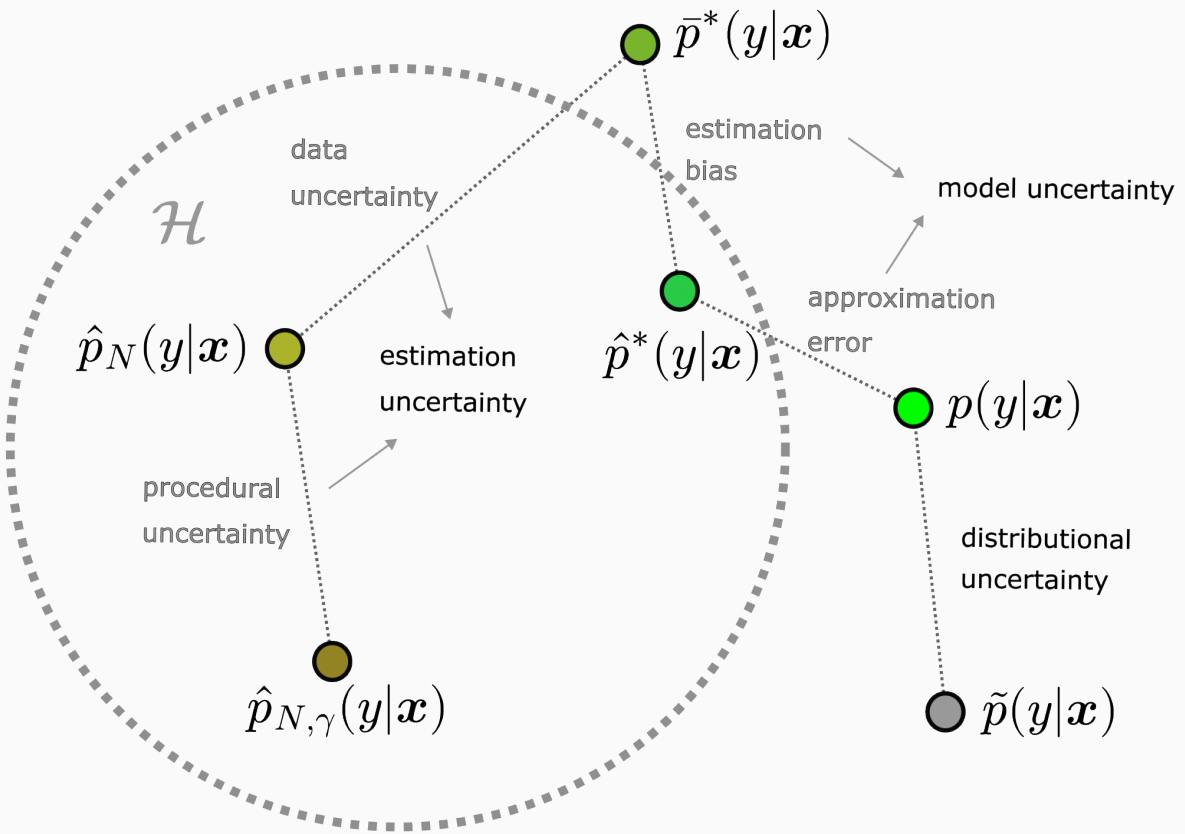


Figure 2: Sources of epistemic uncertainty. Figure adapted from Huang et al.[4]

[4] Huang et al., "Efficient uncertainty quantification and reduction for over-parameterized neural networks".

Uncertainty from a generalised bias-variance perspective

Preliminaries

[5] Gneiting et al., “Strictly proper scoring rules, prediction, and estimation”.
[6] Banerjee et al., “Clustering with Bregman Divergences”.

Preliminaries

Proper scoring rules

A loss function $\ell : \mathcal{P}(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}$ is a **strictly proper scoring rule**[5] if, for any $\boldsymbol{x} \in \mathcal{X}$, the expected loss is uniquely minimized when the estimate $\hat{p}(y|\boldsymbol{x})$ equals the ground truth $p(y|\boldsymbol{x})$.

[5] Gneiting et al., "Strictly proper scoring rules, prediction, and estimation".

[6] Banerjee et al., "Clustering with Bregman Divergences".

Preliminaries

Proper scoring rules

A loss function $\ell : \mathcal{P}(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}$ is a **strictly proper scoring rule**[5] if, for any $\boldsymbol{x} \in \mathcal{X}$, the expected loss is uniquely minimized when the estimate $\hat{p}(y|\boldsymbol{x})$ equals the ground truth $p(y|\boldsymbol{x})$.

Entropy and Bregman divergence

Strictly proper scoring rules induce an *entropy* $\mathbb{H}_\ell$ and a *divergence* D_ℓ[6] such that

$$D_\ell(p(\cdot|\boldsymbol{x}), \hat{p}(\cdot|\boldsymbol{x})) = \mathbb{E}_{y|\boldsymbol{x}}[\ell(\hat{p}(\cdot|\boldsymbol{x}), y)] - \mathbb{E}_{y|\boldsymbol{x}}[\ell(p(\cdot|\boldsymbol{x}), y)]$$

and $\mathbb{H}_\ell(p(\cdot|\boldsymbol{x})) := \mathbb{E}_{y|\boldsymbol{x}}[\ell(p(\cdot|\boldsymbol{x}), y)]$.

[5] Gneiting et al., "Strictly proper scoring rules, prediction, and estimation".
[6] Banerjee et al., "Clustering with Bregman Divergences".

Uncertainty from a generalized bias-variance perspective

Let ℓ be a be a strictly proper scoring rule (e.g., brier score, log loss). The pointwise expected loss of a predictive distribution $\hat{p}(y|\boldsymbol{x})$ decomposes as[7]

$$\mathbb{E}_{y|\boldsymbol{x}}[\ell(\hat{p}(\cdot|\boldsymbol{x}), y)] = \underbrace{\mathbb{H}_\ell[p(\cdot|\boldsymbol{x})]}_{\substack{\textbf{aleatoric} \\ \text{(generalized entropy)}}} + \underbrace{D_\ell(p(\cdot|\boldsymbol{x}), \hat{p}(\cdot|\boldsymbol{x}))}_{\substack{\textbf{epistemic} \\ \text{(divergence between } \hat{p} \text{ and } p\text{)}}}$$

[7] Gruber et al., "Sources of Uncertainty in Machine Learning–A Statisticians' View".
[8] Pfau, "A generalized bias-variance decomposition for bregman divergences".

Uncertainty from a generalized bias-variance perspective

Let ℓ be a be a strictly proper scoring rule (e.g., brier score, log loss). The pointwise expected loss of a predictive distribution $\hat{p}(y|\boldsymbol{x})$ decomposes as[7]

$$\mathbb{E}_{y|\boldsymbol{x}}[\ell(\hat{p}(\cdot|\boldsymbol{x}), y)] = \underbrace{\mathbb{H}_\ell[p(\cdot|\boldsymbol{x})]}_{\substack{\textbf{aleatoric} \\ \text{(generalized entropy)}}} + \underbrace{D_\ell(p(\cdot|\boldsymbol{x}), \hat{p}(\cdot|\boldsymbol{x}))}_{\substack{\textbf{epistemic} \\ \text{(divergence between } \hat{p} \text{ and } p\text{)}}}$$

For an estimate $\hat{p}_{N,\gamma}$ under data and procedural uncertainty, the expected epistemic term can be decomposed as[8]:

$$\mathbb{E}_{\mathcal{D}_N,\gamma}\left[D_\ell(p(\cdot|\boldsymbol{x}), \hat{p}_{N,\gamma}(\cdot|\boldsymbol{x})\right] = \underbrace{D_\ell\left(p(\cdot|\boldsymbol{x}), \bar{p}^*(\cdot|\boldsymbol{x})\right)}_{\substack{\textbf{bias} \\ \text{(systematic + search bias)}}} + \underbrace{\mathbb{E}_{\mathcal{D}_N,\gamma}\left[D_\ell(\bar{p}^*(\cdot|\boldsymbol{x}), \hat{p}_{N,\gamma}(\cdot|\boldsymbol{x})\right]}_{\substack{\textbf{variance} \\ \text{(procedural and data variability)}}}.$$

[7] Gruber et al., "Sources of Uncertainty in Machine Learning–A Statisticians' View".
[8] Pfau, "A generalized bias-variance decomposition for bregman divergences".

Illustration for this talk: univariate regression

Assume $y_i = f(\boldsymbol{x}) + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, \sigma^2(\boldsymbol{x}_i))$. For the loss

$$\ell_{MSE}(\hat{p}, y) = (\hat{y} - y)^2,$$

the decomposition attains a familiar form:

Illustration for this talk: univariate regression

Assume $y_i = f(\boldsymbol{x}) + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, \sigma^2(\boldsymbol{x}_i))$. For the loss

$$\ell_{MSE}(\hat{p}, y) = (\hat{y} - y)^2,$$

the decomposition attains a familiar form:

$$\mathbb{E}_{y|\boldsymbol{x}}\left[\mathbb{E}_{\mathcal{D}_N,\gamma}(\hat{y} - y)^2 | \boldsymbol{x}\right] = \underbrace{\sigma^2(\boldsymbol{x})}_{\text{aleatoric (intrinisc noise)}} + \underbrace{\text{Var}_{\mathcal{D}_N,\gamma}(\hat{y}|\boldsymbol{x}) + \text{bias}^2(\hat{y}|\boldsymbol{x})}_{\text{epistemic (bias \textbf{and} variance)}}. \quad (2)$$

1. Limitation: Second-order uncertainty representations do not account for bias.

Illustration for this talk: Synthetic regression setup

Univariate regression:

$$y_i = f(\boldsymbol{x}_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2(\boldsymbol{x}_i)).$$

Our setting:

$$\begin{aligned} x_i &\sim Beta(1.2, 0.5), \\ y_i &= \sin\left(\frac{1}{5(x_i + 0.16)^3}\right) + \epsilon_i \\ \epsilon_i &\sim \mathcal{N}(0, \sigma^2(x_i)) \end{aligned}$$

with $\sigma^2(x) = x^4$.

Illustration for this talk: Synthetic regression setup

Univariate regression:

$$y_i = f(\boldsymbol{x}_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2(\boldsymbol{x}_i)).$$

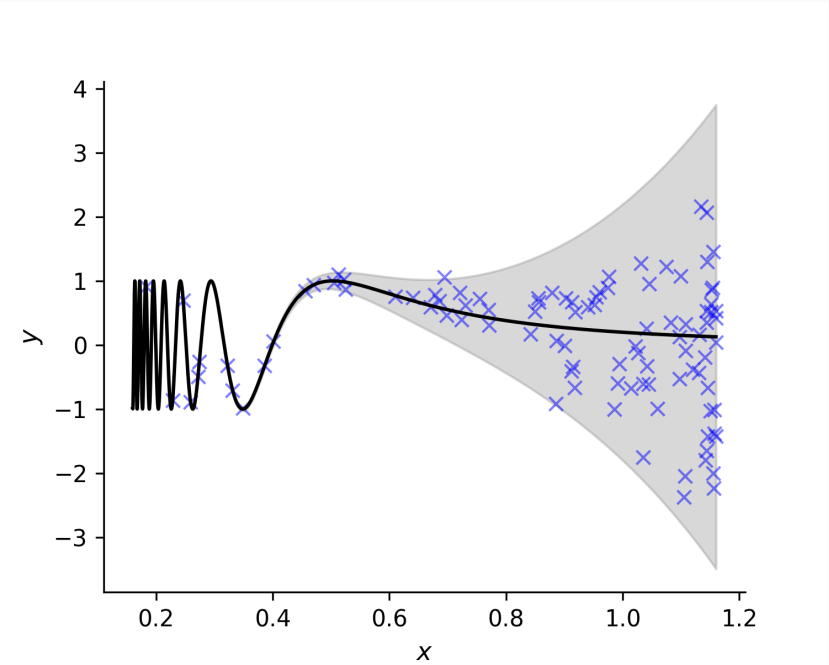
Our setting:

$$x_i \sim Beta(1.2, 0.5),$$

$$y_i = \sin\left(\frac{1}{5(x_i + 0.16)^3}\right) + \epsilon_i$$

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2(x_i))$$

with $\sigma^2(x) = x^4$.



Unaccounted bias can lead to false trust in uncertainty estimates.

Recall the uncertainty estimates of a second-order model (Eqn. 1):

$$\mathrm{Var}(\hat{y}|\boldsymbol{x}) = \underbrace{\mathbb{E}_{\boldsymbol{\theta}}\left[\mathrm{Var}(\hat{y}|\boldsymbol{x},\boldsymbol{\theta})\right]}_{\text{aleatoric estimate}} + \underbrace{\mathrm{Var}_{\boldsymbol{\theta}}\left(\mathbb{E}[\hat{y}|\boldsymbol{x},\boldsymbol{\theta}]\right)}_{\text{epistemic estimate}}.$$

Unaccounted bias can lead to false trust in uncertainty estimates.

Recall the uncertainty estimates of a second-order model (Eqn. 1):

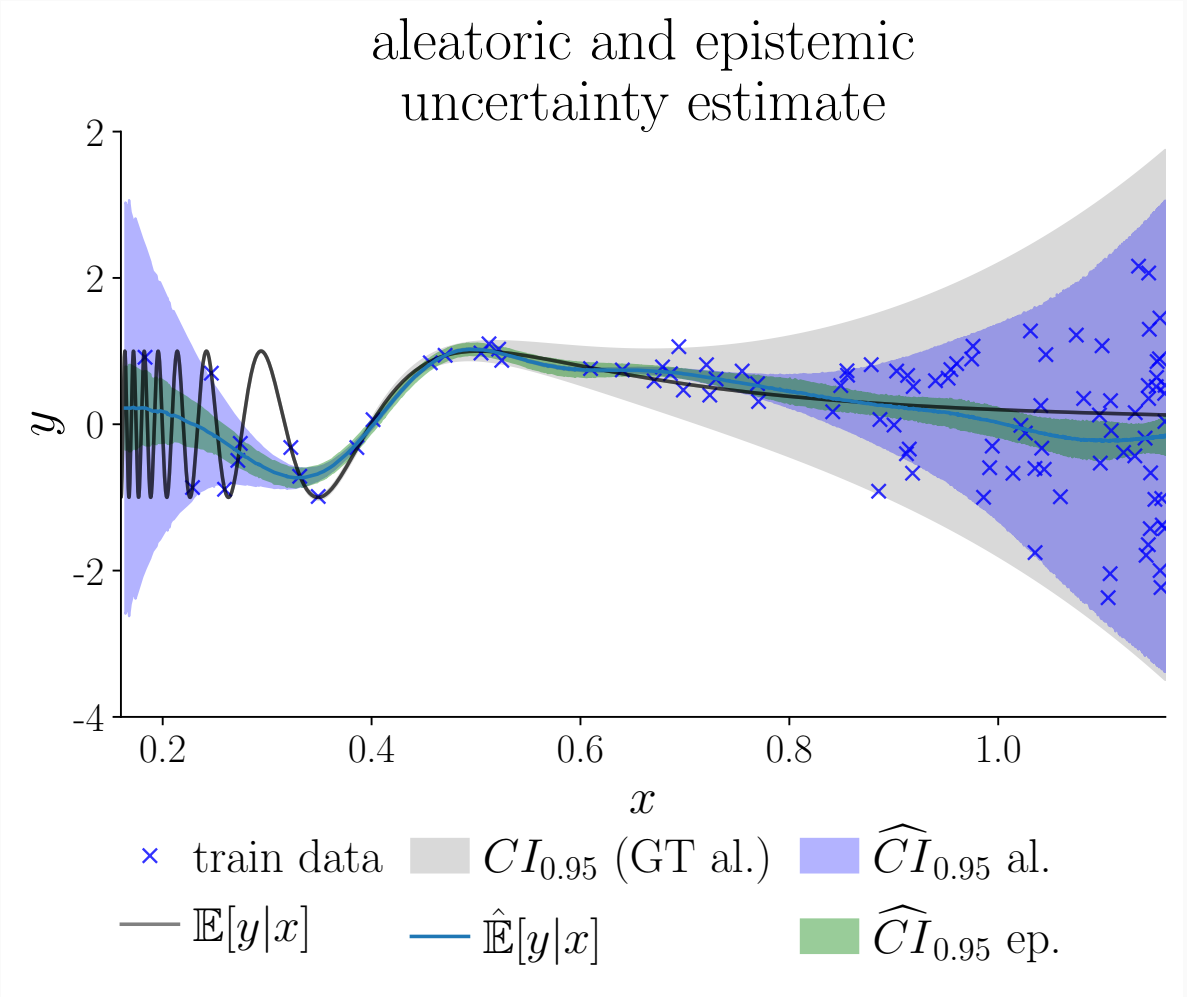
$$\mathrm{Var}(\hat{y}|\boldsymbol{x}) = \underbrace{\mathbb{E}_{\boldsymbol{\theta}}\left[\mathrm{Var}(\hat{y}|\boldsymbol{x},\boldsymbol{\theta})\right]}_{\text{aleatoric estimate}} + \underbrace{\mathrm{Var}_{\boldsymbol{\theta}}\left(\mathbb{E}[\hat{y}|\boldsymbol{x},\boldsymbol{\theta}]\right)}_{\text{epistemic estimate}}.$$

and the epistemic uncertainty part of the bias variance decomposition (Eqn. 2):

$$\mathbb{E}_{y|\boldsymbol{x}}\left[\mathbb{E}_{\mathcal{D}_N,\gamma}(\hat{y}-y)^2|\boldsymbol{x}\right] = \underbrace{\sigma^2(\boldsymbol{x})}_{\text{aleatoric (intrinisc noise)}} + \underbrace{\mathrm{Var}_{\mathcal{D}_N,\gamma}(\hat{y}|\boldsymbol{x}) + \mathrm{bias}^2(\hat{y}|\boldsymbol{x})}_{\text{epistemic (bias \textbf{and} variance)}}.$$

By design, the **bias cannot be captured** by the model!

High bias leads to false uncertainty disentanglement.



Predicted AU (blue) and EU (green) of a heteroscedastic GP model

2. Limitation: Common uncertainty representations often capture only *parts* of the variance.

Decompostion of the variance

Recall bias-variance decomposition:

$$\mathbb{E}_{y|\boldsymbol{x}}\left[\mathbb{E}_{\mathcal{D}_N,\gamma}(\hat{y}-y)^2|\boldsymbol{x}\right] = \sigma^2(\boldsymbol{x}) + \underbrace{\mathrm{Var}_{\mathcal{D}_N,\gamma}(\hat{y}|\boldsymbol{x})}_{\textbf{variance (data and procedural uncertainty)}} + \mathrm{bias}^2(\hat{y}|\boldsymbol{x}).$$

Decompostion of the variance

Recall bias-variance decomposition:

$$\mathbb{E}_{y|\boldsymbol{x}}\left[\mathbb{E}_{\mathcal{D}_N,\gamma}(\hat{y}-y)^2|\boldsymbol{x}\right] = \sigma^2(\boldsymbol{x}) + \underbrace{\mathrm{Var}_{\mathcal{D}_N,\gamma}(\hat{y}|\boldsymbol{x})}_{\textbf{variance (data and procedural uncertainty)}} + \mathrm{bias}^2(\hat{y}|\boldsymbol{x}).$$

It can be further decomposed into **data and procedural variance**:

$$\mathrm{Var}_{\mathcal{D}_N,\gamma}(\hat{y}|\boldsymbol{x}) = \underbrace{\mathbb{E}_{\mathcal{D}_N}\left[\mathrm{Var}_{\gamma}(\hat{y}|\boldsymbol{x})|\mathcal{D}_N\right]}_{\text{procedural uncertainty}} + \underbrace{\mathrm{Var}_{\mathcal{D}_N}(\mathbb{E}_{\gamma}[\hat{y}|\boldsymbol{x}]|\mathcal{D}_N)}_{\text{data uncertainty}}.$$

Decompostion of the variance

Recall bias-variance decomposition:

$$\mathbb{E}_{y|\boldsymbol{x}}\left[\mathbb{E}_{\mathcal{D}_N,\gamma}(\hat{y}-y)^2|\boldsymbol{x}\right] = \sigma^2(\boldsymbol{x}) + \underbrace{\mathrm{Var}_{\mathcal{D}_N,\gamma}(\hat{y}|\boldsymbol{x})}_{\textbf{variance (data and procedural uncertainty)}} + \mathrm{bias}^2(\hat{y}|\boldsymbol{x}).$$

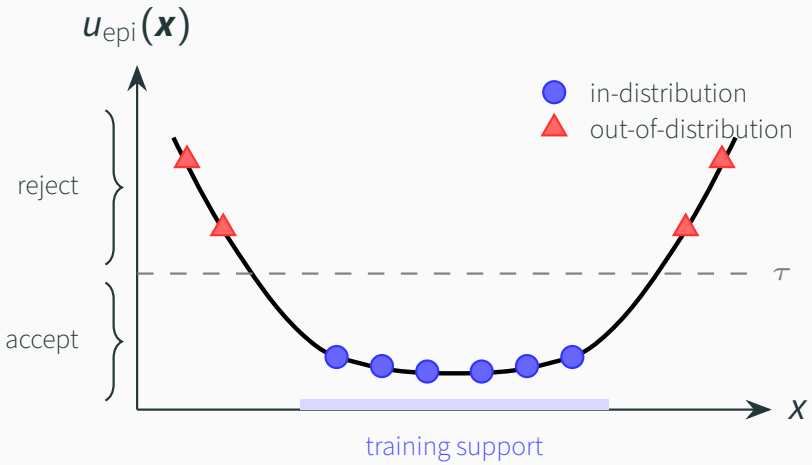
It can be further decomposed into **data and procedural variance**:

$$\mathrm{Var}_{\mathcal{D}_N,\gamma}(\hat{y}|\boldsymbol{x}) = \underbrace{\mathbb{E}_{\mathcal{D}_N}\left[\mathrm{Var}_{\gamma}(\hat{y}|\boldsymbol{x})|\mathcal{D}_N\right]}_{\text{procedural uncertainty}} + \underbrace{\mathrm{Var}_{\mathcal{D}_N}(\mathbb{E}_{\gamma}[\hat{y}|\boldsymbol{x}]|\mathcal{D}_N)}_{\text{data uncertainty}}.$$

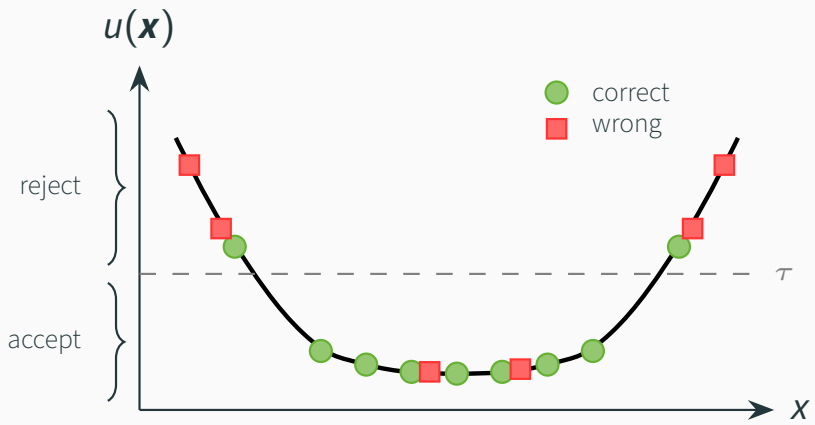
How do we know which parts the variance estimate of our model captures?

Common way of evaluation: Downstream tasks

Out-of-Distribution Detection



Selective Prediction



The reference distribution: a frequentist evaluation approach

Reference distribution for data and procedural uncertainty

The reference distribution[9] $q_{N,\gamma}$ describes the probability of obtaining parameters $\boldsymbol{\theta}$ given the randomness in **data sampling** (P^N) and **procedure** (P^γ):

$$q_{N,\gamma}(\boldsymbol{\theta}|\boldsymbol{x}) := \int_\Gamma \int_{\mathcal{X}\times\mathcal{Y}} \mathbb{I}\left[\hat{\boldsymbol{\theta}}_{N,\gamma}(\boldsymbol{x}) = \boldsymbol{\theta}\right] \, dP^N dP^\gamma .$$

[9] Jürgens et al., "Is Epistemic Uncertainty Faithfully Represented by Evidential Deep Learning Methods?"
[10] Jain et al., "Frequentist Validity of Epistemic Uncertainty Estimators".

The reference distribution: a frequentist evaluation approach

Reference distribution for data and procedural uncertainty

The reference distribution[9] $q_{N,\gamma}$ describes the probability of obtaining parameters $\boldsymbol{\theta}$ given the randomness in **data sampling** (P^N) and **procedure** (P^γ):

$$q_{N,\gamma}(\boldsymbol{\theta}|\boldsymbol{x}) := \int_\Gamma \int_{\mathcal{X}\times\mathcal{Y}} \mathbb{I}\left[\hat{\boldsymbol{\theta}}_{N,\gamma}(\boldsymbol{x}) = \boldsymbol{\theta}\right] \, dP^N dP^\gamma.$$

Empirical approximation: We estimate $q_{N,\gamma}$ by nested resampling of:

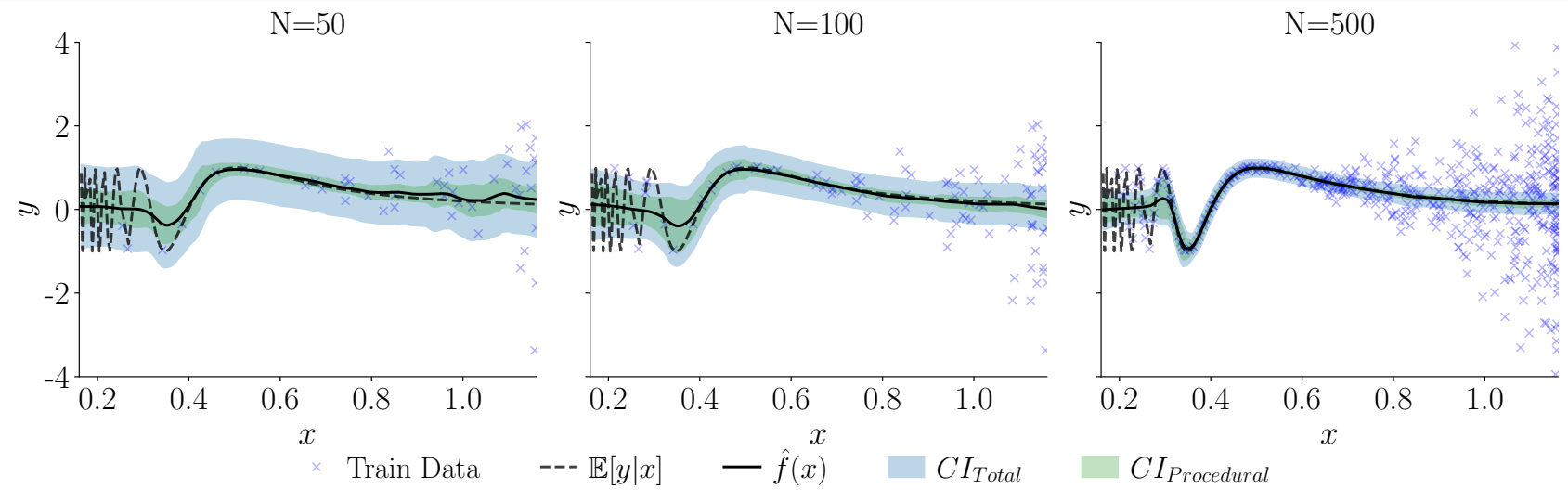
1. datasets $\mathcal{D}_N$ (data uncertainty)
2. different random seeds γ (procedural uncertainty).

Recent results[10] show estimation of $q_{N,\gamma}(\boldsymbol{\theta}|\boldsymbol{x})$ based on bootstrappping.

---

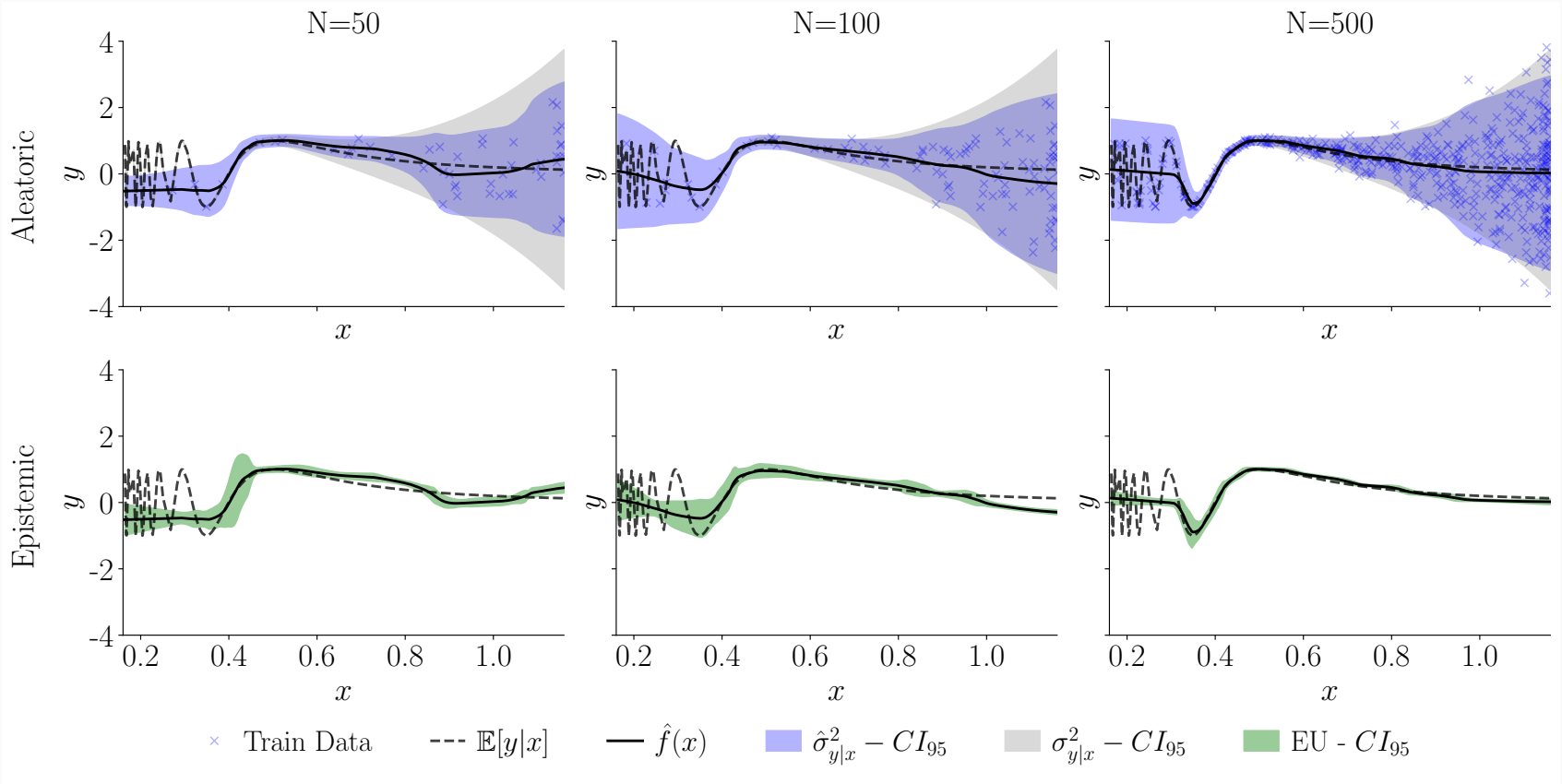
[9] Jürgens et al., "Is Epistemic Uncertainty Faithfully Represented by Evidential Deep Learning Methods?"
[10] Jain et al., "Frequentist Validity of Epistemic Uncertainty Estimators".

Synthetic example - Decomposition of the reference distribution



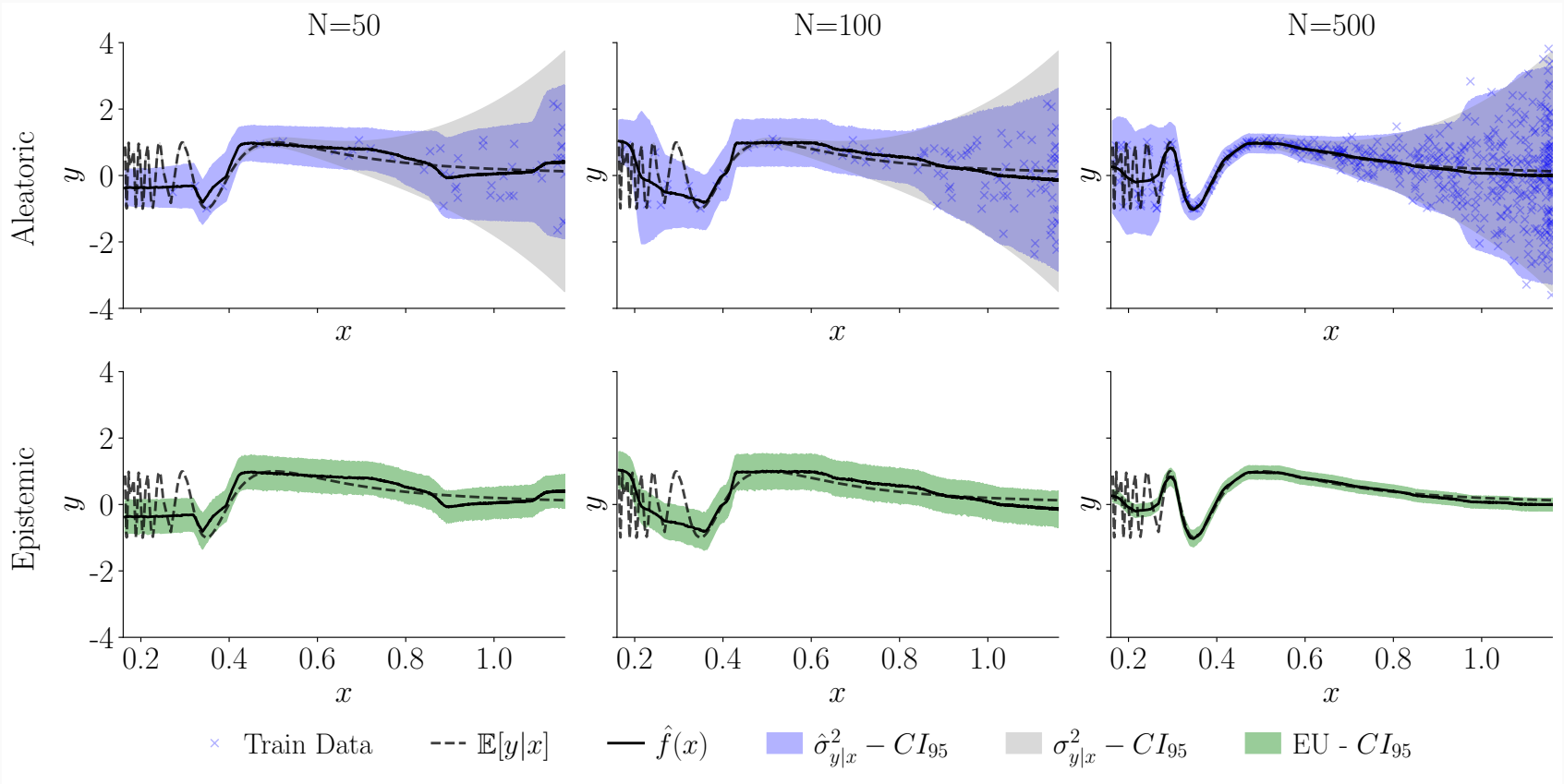
Experimental results - synthetic data

Deep Ensembles



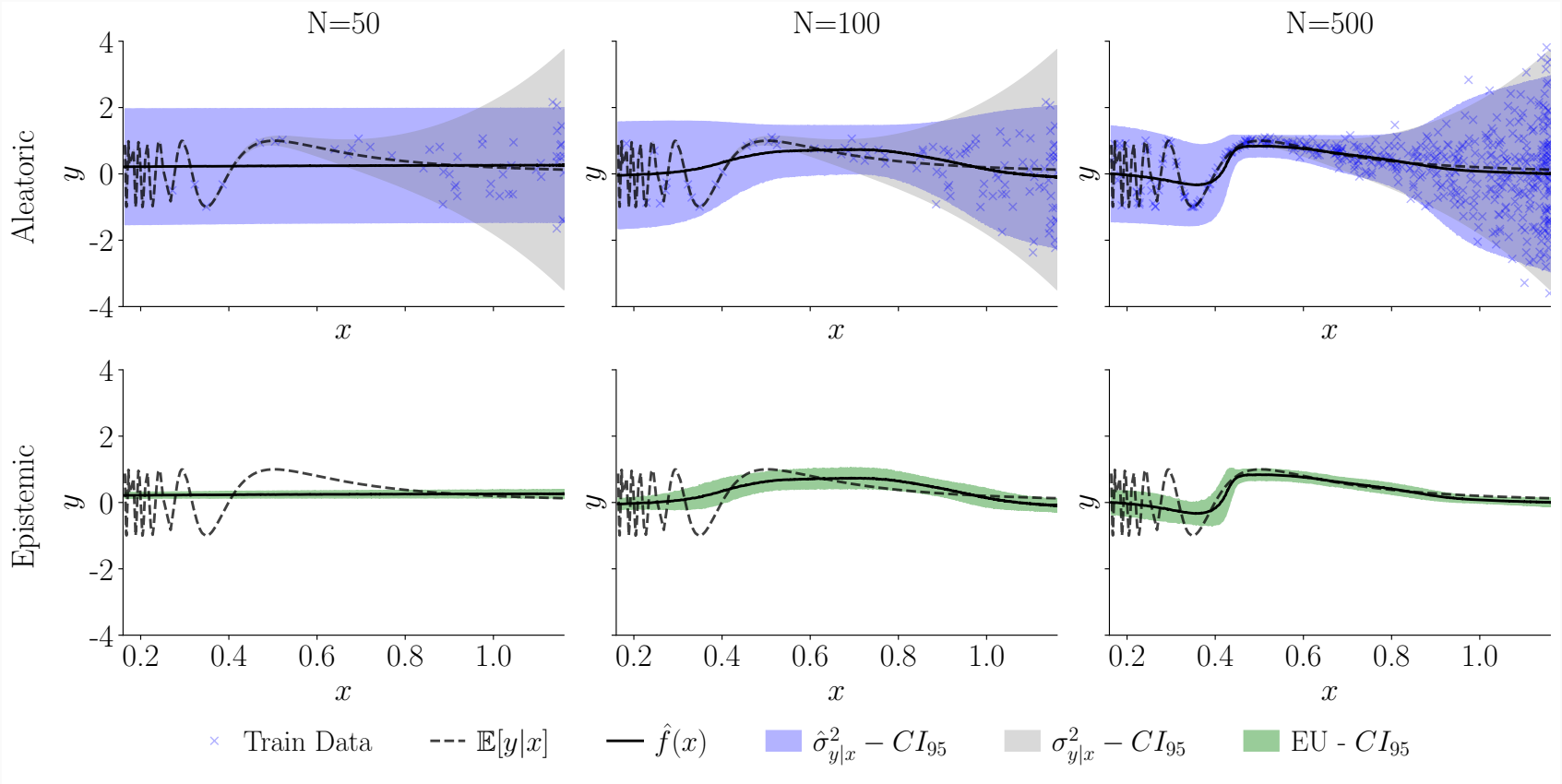
Experimental results - synthetic data

Variational Inference



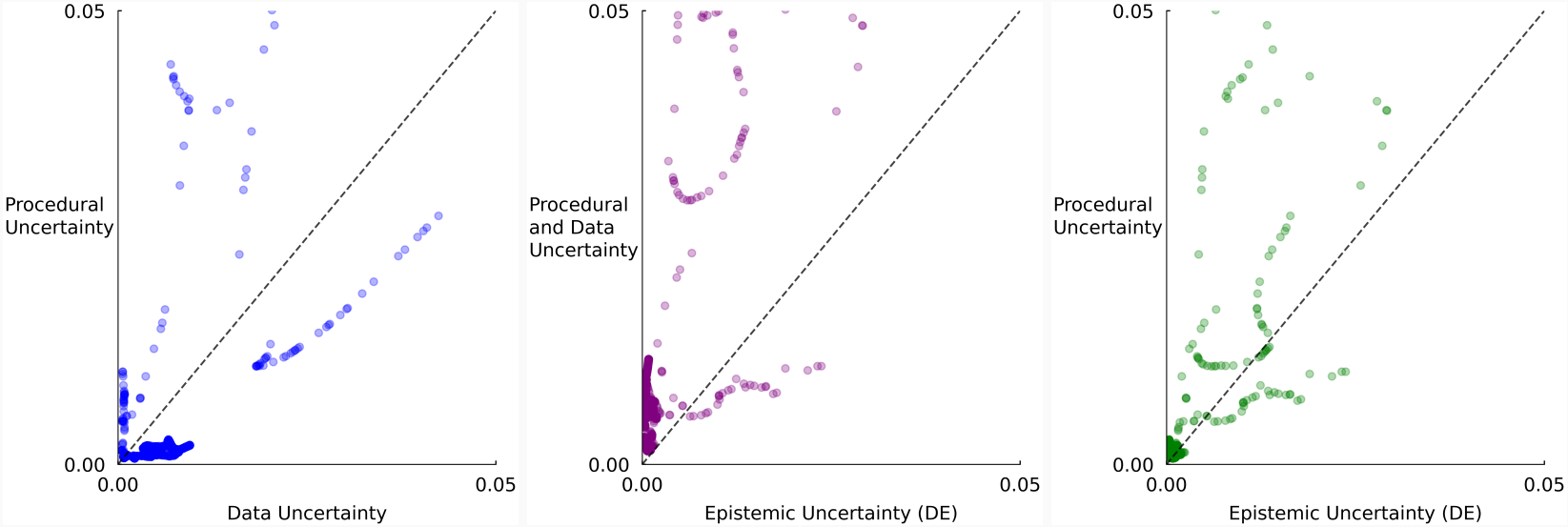
Experimental results - synthetic data

MC Dropout



Direct comparison of data, procedural and estimated uncertainty

Deep Ensemble estimates vs reference distribution estimates

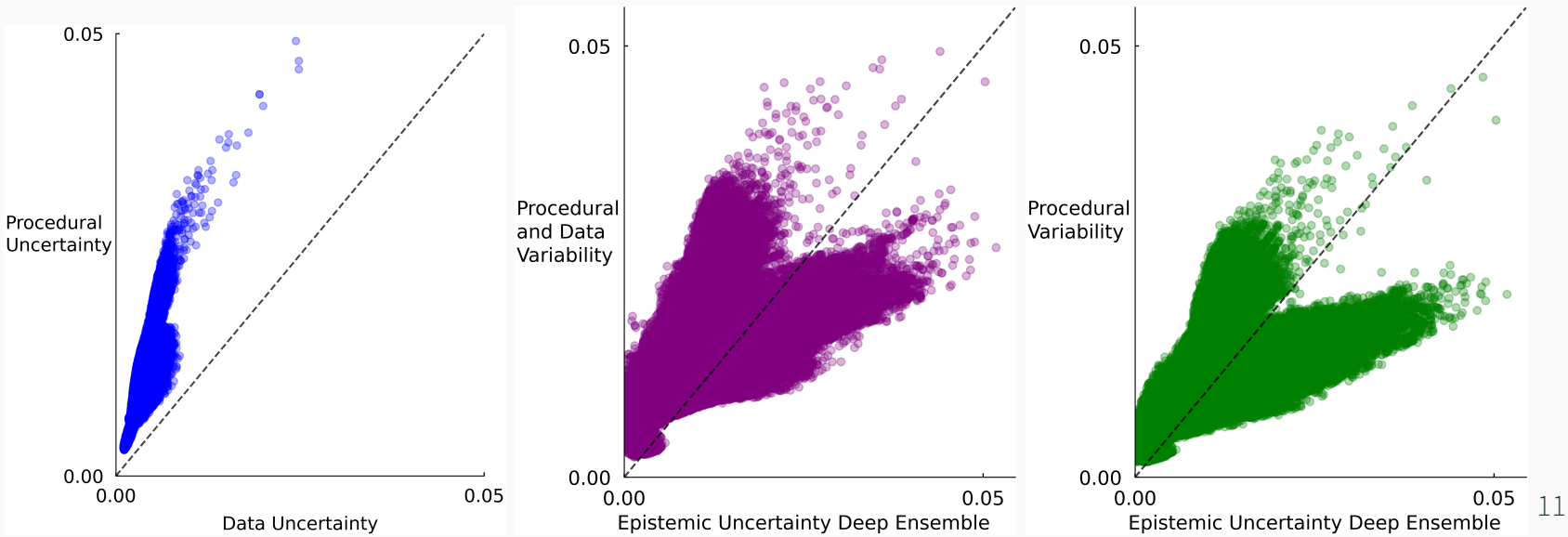


Experimental results - real data

Approximation of the reference distribution on real data

- New York taxi trip duration dataset published in Kaggle competition
- $1,206,737$ observations $\rightarrow$ divide into 100 disjoint datasets of $12,067$ observations to estimate reference distribution ($n_d = 100$)
- train $n_\gamma = 5$ multilayer perceptrons (MLPs) with different random weight initializations on each subset

Deep Ensemble estimates vs reference distribution



[11]

[11] https://www.kaggle.com/competitions/nyc-taxi-trip-duration/data

Conclusion and outlook

1. **Taxonomy matters:** A fine-grained taxonomy of epistemic uncertainty is useful for proper evaluation.

Conclusion and outlook

1. **Taxonomy matters:** A fine-grained taxonomy of epistemic uncertainty is useful for proper evaluation.
2. **High model bias can lead to confusion of aleatoric/epistemic uncertainty**.

Conclusion and outlook

1. **Taxonomy matters:** A fine-grained taxonomy of epistemic uncertainty is useful for proper evaluation.
2. **High model bias can lead to confusion of aleatoric/epistemic uncertainty**.
3. Second-order methods primarily capture **parts of epistemic the variance** uncertainty, ignoring data uncertainty and model bias.

Conclusion and outlook

1. **Taxonomy matters:** A fine-grained taxonomy of epistemic uncertainty is useful for proper evaluation.
2. **High model bias can lead to confusion of aleatoric/epistemic uncertainty**.
3. Second-order methods primarily capture **parts of epistemic the variance** uncertainty, ignoring data uncertainty and model bias.

Call to action

- "bias-aware" uncertainty quantification: explicitly model systematic bias.
- usage of simulation based protocols to validate which source a method captures.

arXiv Preprint

Thank you!