Is Epistemic Uncertainty Faithfully Represented by Evidential Deep Learning Methods?

WUML 2024

Mira Jürgens[1], Nis Meinert[2], Viktor Bengs[3], Eyke Hüllermeier[3] and Willem Waegeman[1]

[1]Gent University [2]German Aerospace Center (DLR) [3]LMU Munich

Deutsches Zentrum für Luft- und Raumfahrt

Setting

Notation	Meaning
$\mathcal{X} = \mathbb{R}^D$	instance space
$\mathcal{Y} = \begin{cases} \mathbb{R}^{D'}, & \text{regression} \\ \{1, \dots, K\}, & \text{classification} \end{cases}$	label space
$\mathcal{D}_N = \{(\boldsymbol{x_i}, y_i)\}_{i=1}^N$	training data
P	joint prob. distribution on $\mathcal{X} \times \mathcal{Y}$
$p(\cdot \mid \boldsymbol{x})$	cond. density for $\boldsymbol{x} \in \mathcal{X}$ on $\mathcal{Y}$
$\mathbb{P}(A)$	set of all prob. distributions on A

What is Evidential Deep Learning

Evidential Deep Learning for Classification

Idea of EDL: Learn a second-order distribution directly by which both aleatoric and epistemic uncertainty can be predicted.

Dirichlet-Categorical Model

- level 1: $y \sim Cat(\boldsymbol{\theta})$ with $\boldsymbol{\theta} \in \Delta_K$
- level 2: $\boldsymbol{\theta} \sim Dir(\boldsymbol{m})$ with $\boldsymbol{m} \in \mathbb{R}_+^K$

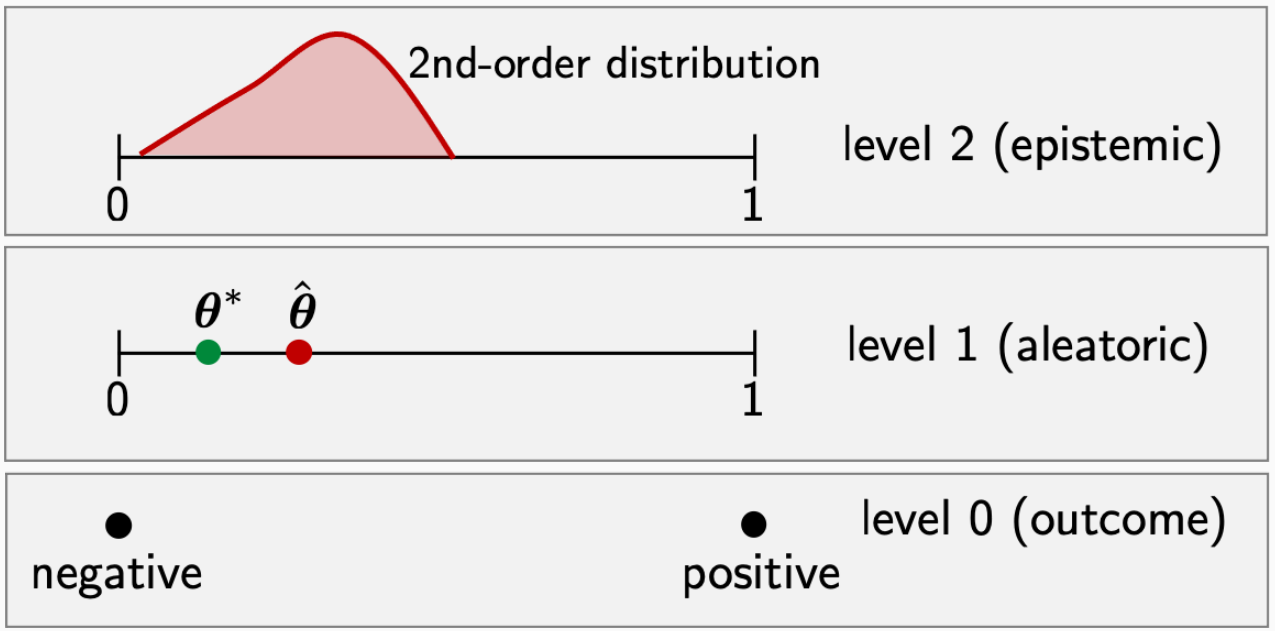


Figure 1: Different levels of predictions[1], exemplary for the binary classification case.

[1]Wimmer et al., *Quantifying Aleatoric and Epistemic Uncertainty in Machine Learning: Are Conditional Entropy and Mutual Information Appropriate Measures?*

Evidential Deep Learning for Regression

Normal-Normal-Inverse-Gamma Model

- level 1: $y \sim \mathcal{N}(\mu, \sigma^2)$ with $\mu \in \mathbb{R}$, $\sigma^2 \in \mathbb{R}_+$
- level 2: $(\mu, \sigma^2) \sim N\Gamma^{-1}(\boldsymbol{m})$ with $\boldsymbol{m} = (\gamma, \nu, \alpha, \beta) \in \mathbb{R} \times \mathbb{R}^3_+$

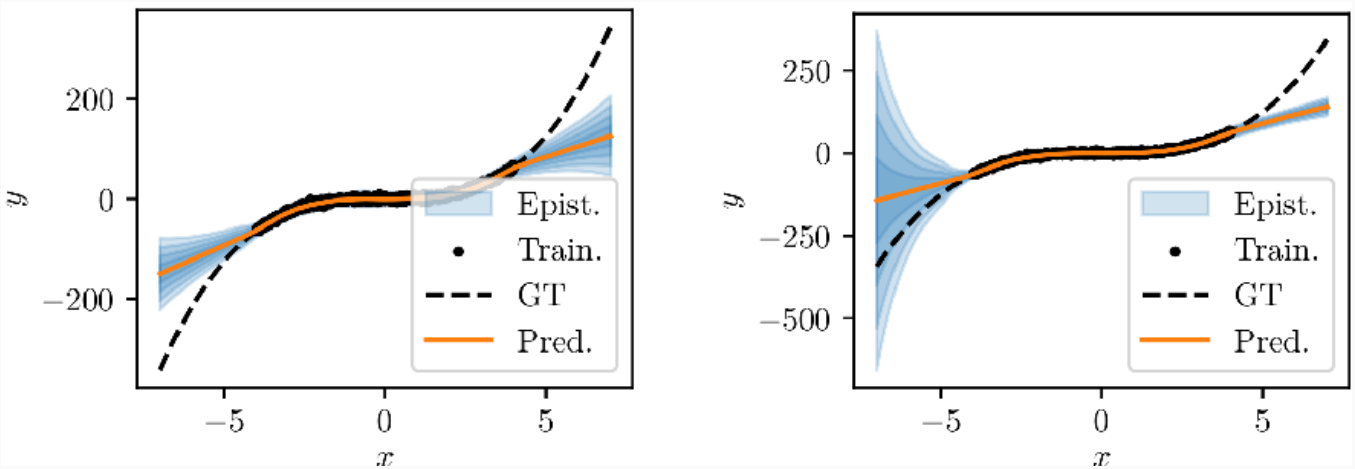


Figure 2: Results of Deep Evidential Regression[2] reproduced for two different runs[3].

[2] Amini et al., "Deep evidential regression".

[3] Meinert, Gawlikowski, and Lavin, "The Unreasonable Effectiveness of Deep Evidential Regression".

Different Problems, Different Distributions

For the following sections, let

- $\boldsymbol{\theta} \equiv \boldsymbol{\theta}(\boldsymbol{x}, \phi) \in \boldsymbol{\Theta}$, with $\boldsymbol{\theta} : \mathcal{X} \to \boldsymbol{\Theta}$ be the parameters of the first-order predictor
- $\boldsymbol{m} \equiv \boldsymbol{m}(\boldsymbol{x}, \phi) \in \mathcal{M}$ with $\boldsymbol{m} : \mathcal{X} \to \mathcal{M}$ the parameters of the second-order model

problem	likelihood	conjugate prior
classification[4]	$y \sim Cat(\boldsymbol{\theta})$ $\boldsymbol{\theta} \in \Delta_k$	$\boldsymbol{\theta} \sim Dir(\boldsymbol{m})$ $\boldsymbol{m} \in \mathbb{R}_+^K$
regression (univ.)[5]	$y \sim \mathcal{N}(\mu, \sigma^2)$ $\mu \in \mathbb{R}$, $\sigma^2 \in \mathbb{R}_+$	$\boldsymbol{\theta} \sim N\Gamma^{-1}(\boldsymbol{m})$ $\boldsymbol{m} = (\gamma, \nu, \alpha, \beta) \in \mathbb{R} \times \mathbb{R}_+^3$
regression (multiv.)[6]	$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ $\boldsymbol{\mu} \in \mathbb{R}^D$, $\boldsymbol{\Sigma} \in \mathbb{R}^{D \times D}$	$\boldsymbol{\theta} \sim NIW(\boldsymbol{m})$ $\boldsymbol{m} = (\boldsymbol{\mu}_0, \kappa, \nu, \boldsymbol{\Phi})$
point processes	$y \sim Pois(\boldsymbol{\theta})$ $\boldsymbol{\theta} \in \mathbb{R}_+$	$\boldsymbol{\theta} \sim \Gamma(\boldsymbol{m})$ $\boldsymbol{m} = (\alpha, \beta) \in \mathbb{R}_+^2$

[4] Sensoy, Kaplan, and Kandemir, "Evidential Deep Learning to Quantify Classification Uncertainty".

[5] Amini et al., "Deep evidential regression".

[6] Meinert and Lavin, "Multivariate deep evidential regression".

Flaws of EDL

While showing good results in downstream tasks, several critical analyses show theoretical flaws of EDL:

- Non-Convergence in the Classification case[7]
- Non-Convergence in the Regression Case[8]
- Non-Properness of its Loss functions[9]

$\longrightarrow$ **Our approach:** Does the learned second-order distribution represent the (epistemic) uncertainty of the first-order parameters in a faithful way?

---

[7] Bengs, Hüllermeier, and Waegeman, "Pitfalls of Epistemic Uncertainty Quantification through Loss Minimisation".

[8] Meinert, Gawlikowski, and Lavin, "The Unreasonable Effectiveness of Deep Evidential Regression".

[9] Bengs, Hüllermeier, and Waegeman, *On Second-Order Scoring Rules for Epistemic Uncertainty Quantification*.

Comparing First and Second-Order Risk Minimization

First-Order vs Second-Order Risk Minimization

First-Order Risk Minimization

- Hypothesis space

$$\mathcal{H}_1 \subset \mathbb{P}(\mathcal{Y}) = \{h_1 : \Theta \to \mathbb{P}(\mathcal{Y})\}$$

- Loss function (e.g. NLL)

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

- **Risk minimization** over parameters Φ

$$\min_{\mathbf{\Phi}} \sum_{i=1}^{N} L_1(y_i, p(y \mid \theta(\boldsymbol{x}_i; \mathbf{\Phi})))$$

First-Order vs Second-Order Risk Minimization

First-Order Risk Minimization

- Hypothesis space

$$\mathcal{H}_1 \subset \mathbb{P}(\mathcal{Y}) = \{h_1 : \Theta \to \mathbb{P}(\mathcal{Y})\}$$

- Loss function (e.g. NLL)

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

- **Risk minimization** over parameters Φ

$$\min_{\boldsymbol{\Phi}} \sum_{i=1}^{N} L_1(y_i, p(y \mid \theta(\boldsymbol{x}_i; \boldsymbol{\Phi})))$$

Second-Order Risk Minimization

- Hypothesis space

$$\mathcal{H}_2 \subset \mathbb{P}(\boldsymbol{\Theta}) = \{h_2 : \mathcal{M} \to \mathbb{P}(\boldsymbol{\Theta})\}$$

- Loss function (e.g. NLL)

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

- **Inner expectation minimization**

$$\min_{\boldsymbol{\Phi}} \sum_{i=1}^{N} L_1 \left(y_i, \mathbb{E}_{\boldsymbol{\theta} \sim p(\boldsymbol{\theta} \mid \boldsymbol{m}(\boldsymbol{x}_i; \boldsymbol{\Phi}))} \left[p(y \mid \boldsymbol{\theta}) \right] \right) + \lambda R(\boldsymbol{\Phi}).$$

- **Outer expectation minimisation**

$$\min_{\boldsymbol{\Phi}} \sum_{i=1}^{N} \mathbb{E}_{\boldsymbol{\theta} \sim p(\boldsymbol{\theta} \mid \boldsymbol{m}(\boldsymbol{x}_i; \boldsymbol{\Phi}))} \left[L_1 \left(y_i, p(y \mid \boldsymbol{\theta}) \right) \right] + \lambda R(\boldsymbol{\Phi}) \,.$$

Regularization

Objective: Enforce $p(\boldsymbol{\theta} \mid \boldsymbol{m}(\boldsymbol{x}_i; \boldsymbol{\Phi}))$ to look similar to a predefined distribution $p(\boldsymbol{\theta}|\boldsymbol{m}_0)$ by minimising per-instance KL-divergence:

$$R(\boldsymbol{\Phi}) = \sum_{i=1}^{N} \mathrm{d}_{\mathrm{KL}}(p(\boldsymbol{\theta} \mid \boldsymbol{m}(\boldsymbol{x}_i; \boldsymbol{\Phi})), p(\boldsymbol{\theta} \mid \boldsymbol{m}_0)).$$

- choose $p(\boldsymbol{\theta}|\boldsymbol{m}_0)$ to parametrise uniform distribution (on first-order distributions)
- minimizing **KL-divergence** $\leftrightarrow$ maximizing **entropy**
- other regularization losses exist, e.g. evidential loss[10]

$$R(\boldsymbol{\Phi}) = \sum_{i=1}^{N} |y_i - \mathbb{E}[\mu_i]|\Psi = \sum_{i=1}^{N} |y_i - \gamma_i|(2\nu + \alpha)$$

[10] Amini et al., "Deep evidential regression".

Example: First vs Second Order Risk Minimization

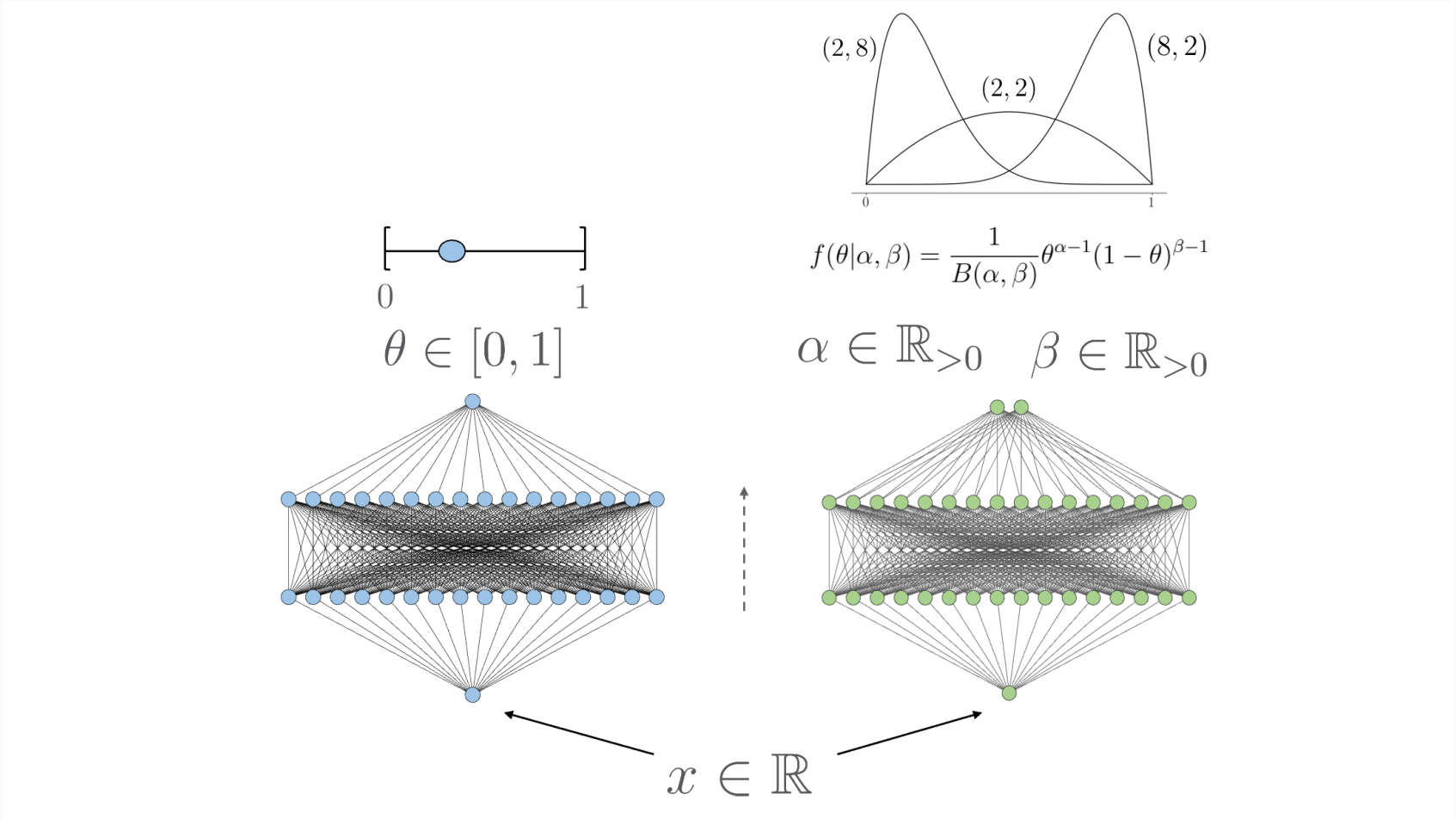
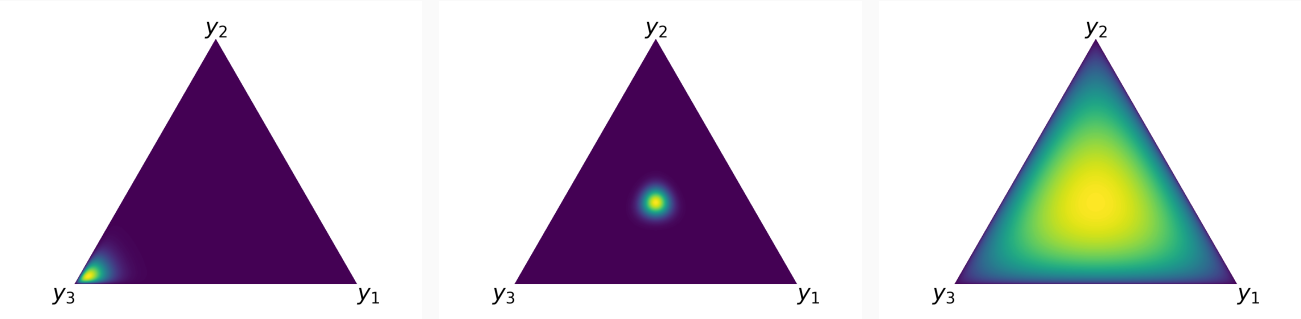


Figure 3: Overview of an exemplary neural network architecture for first (left) and second (right) order risk minimization for the case of binary classification.

Do the resulting second-order distributions represent the underlying epistemic uncertainty of the first-order predictor?

What is a Faithful Representation of Epistemic Uncertainty?

Spread of the distribution should yield a valid estimate of EU:



(a) low uncertainty **(b)** high aleatoric uncertainty **(c)** high epistemic uncertainty

Figure 4: Second-order predictions for the classification case with $K = 3$. See also Posterior network[11], evidential DNN[12]

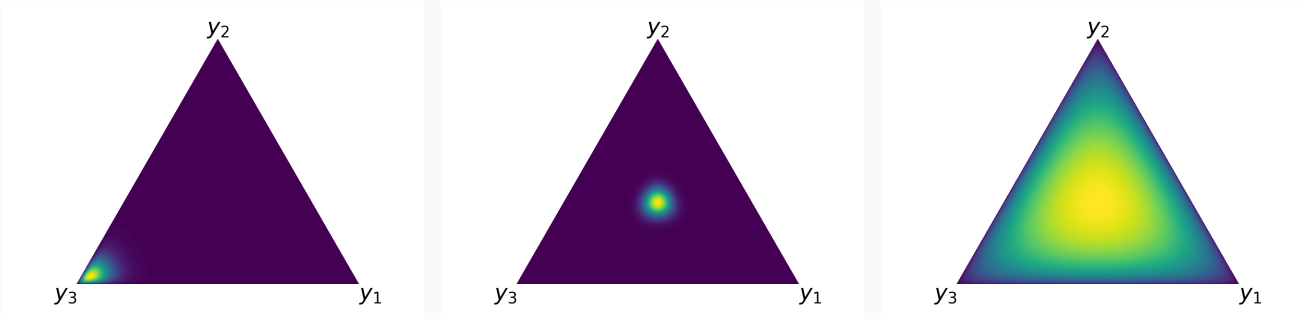
[11] Charpentier, Zügner, and Günnemann, "Posterior network: Uncertainty estimation without OOD samples via density-based pseudo-counts".

[12] Sensoy, Kaplan, and Kandemir, "Evidential Deep Learning to Quantify Classification Uncertainty".

[13] Bengs, Hüllermeier, and Waegeman, "Pitfalls of Epistemic Uncertainty Quantification through Loss Minimisation".

What is a Faithful Representation of Epistemic Uncertainty?

Spread of the distribution should yield a valid estimate of EU:



(a) low uncertainty **(b)** high aleatoric uncertainty **(c)** high epistemic uncertainty

Figure 4: Second-order predictions for the classification case with $K = 3$. See also Posterior network[11], evidential DNN[12]

Desirable convergence properties[13]:

1. Monotonicity: decreasing uncertainty with increasing sample size N
2. Convergence to Dirac delta distribution when $N \to \infty$

---

[11] Charpentier, Zügner, and Günnemann, "Posterior network: Uncertainty estimation without OOD samples via density-based pseudo-counts".

[12] Sensoy, Kaplan, and Kandemir, "Evidential Deep Learning to Quantify Classification Uncertainty".

[13] Bengs, Hüllermeier, and Waegeman, "Pitfalls of Epistemic Uncertainty Quantification through Loss Minimisation".

What is a Faithful Representation of Epistemic Uncertainty?

How can we directly evaluate the faithfulness of the resulting second-order distribution?

What is a Faithful Representation of Epistemic Uncertainty?

How can we directly evaluate the faithfulness of the resulting second-order distribution?

Def. 3.1: Reference Distribution

Let $\boldsymbol{\theta}_{\mathcal{D}_N}(\boldsymbol{x}; \boldsymbol{\Phi}_{\mathcal{D}_N})$ denote the minimizer of the first order minimization problem for a training set $\mathcal{D}_N$ of size N, where $\mathcal{D}_N \sim P^N$. For a given $\boldsymbol{x} \in \mathcal{X}$, we define the *reference second-order distribution* as

$$q_N(\boldsymbol{\theta} \mid \boldsymbol{x}) := \int_{(\mathcal{X}\times\mathcal{Y})^N} \mathbb{I}\left[\boldsymbol{\theta}_{\mathcal{D}_N}(\boldsymbol{x}; \boldsymbol{\Phi}_{\mathcal{D}_N}) = \boldsymbol{\theta}\right] dP^N \,, \tag{1}$$

where $\mathbb{I}[\cdot]$ is the indicator function.

What is a Faithful Representation of Epistemic Uncertainty?

How can we directly evaluate the faithfulness of the resulting second-order distribution?

Def. 3.1: Reference Distribution

Let $\boldsymbol{\theta}_{\mathcal{D}_N}(\boldsymbol{x}; \boldsymbol{\Phi}_{\mathcal{D}_N})$ denote the minimizer of the first order minimization problem for a training set $\mathcal{D}_N$ of size N, where $\mathcal{D}_N \sim P^N$. For a given $\boldsymbol{x} \in \mathcal{X}$, we define the *reference second-order distribution* as

$$q_N(\boldsymbol{\theta} \mid \boldsymbol{x}) := \int_{(\mathcal{X}\times\mathcal{Y})^N} \mathbb{I}\left[\boldsymbol{\theta}_{\mathcal{D}_N}(\boldsymbol{x}; \boldsymbol{\Phi}_{\mathcal{D}_N}) = \boldsymbol{\theta}\right] dP^N \,, \tag{1}$$

where $\mathbb{I}[\cdot]$ is the indicator function.

$\Longrightarrow$ **Can be approximated empirically** by resampling datasets $\mathcal{D}_k = \{(x_{i,k}, y_{i.k})\}_{i=1}^N$ for $k = 1, \ldots, d \in \mathbb{N}$, resulting in risk-minimizers $(\hat{\theta}_1, \ldots, \hat{\theta}_d)$ and

$$\hat{Q}_{N,d}(\boldsymbol{\theta}|\boldsymbol{x}) := \frac{1}{d} \sum_{i=1}^{d} \mathbb{I}_{\hat{\theta}_i(\boldsymbol{x},\boldsymbol{\phi}) \leq \boldsymbol{\theta}}$$

Reference Distribution: Example

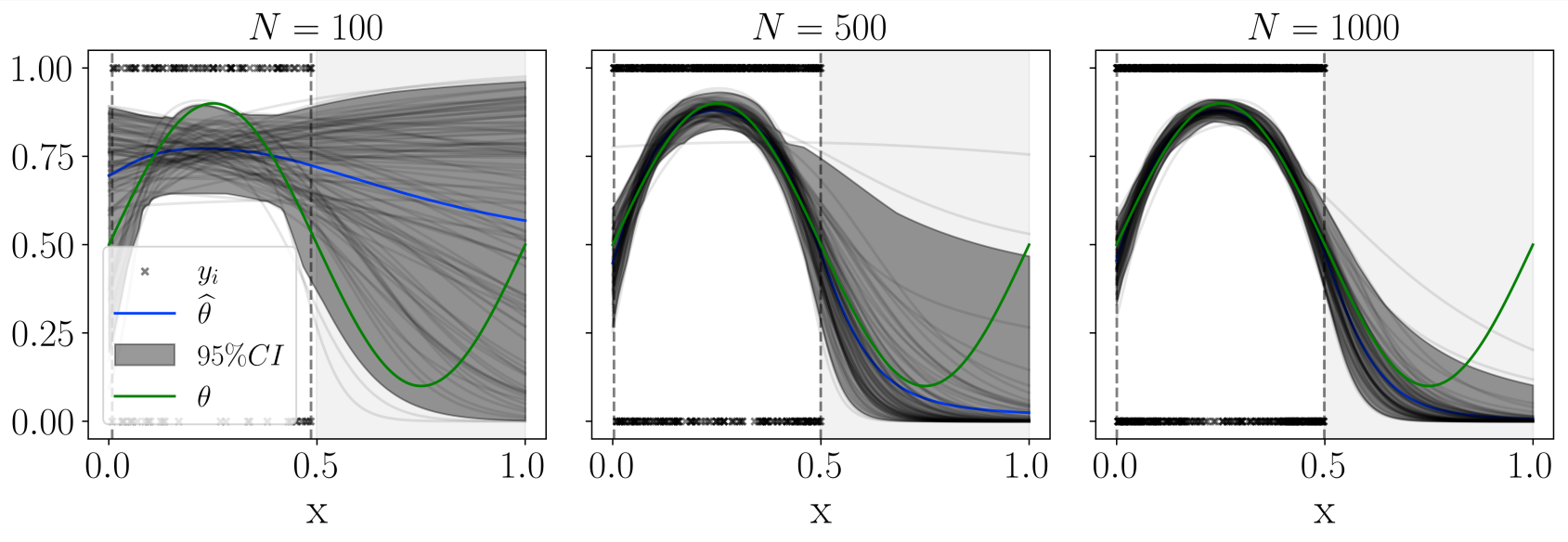


Figure 5: Example of empirically obtaining the reference distribution of the parameter $\theta \in [0, 1]$ of a Bernoulli distribution, where $\theta = 0.5 \cdot (0.8 \cdot \sin(2\pi x) + 1)$. Empirical 95% confidence bounds are obtained by resampling the training data 100 times.

Theoretical Results

Theoretical Results 1

Let $\mathcal{H}_J$ be the joint hypothesis space of $\mathcal{H}_1$ and $\mathcal{H}_2$ containing operators $\mathcal{M} \to \mathbb{P}(\mathcal{Y})$ defined by

$$p(y \mid \boldsymbol{m}(\boldsymbol{x})) = \mathbb{E}_{\boldsymbol{\theta} \sim p(\boldsymbol{\theta} \mid \boldsymbol{m}(\boldsymbol{x}; \boldsymbol{\Phi}))} \, p(y \mid \boldsymbol{\theta}).$$

Theorem 3.2: Unregularized Inner Loss Minimisation

Let $\mathcal{C}_1$ and $\mathcal{C}_J$ be the co-domains of $\mathcal{H}_1$ and $\mathcal{H}_J$. The following properties hold when $\mathcal{H}_2$ consists of

- *Dirichlet distributions: $\mathcal{H}_J$ is not injective, and $\mathcal{C}_1 = \mathcal{C}_J$;*
- *NIG distributions: $\mathcal{H}_J$ is not injective, and $\mathcal{C}_1 \subset \mathcal{C}_J$;*
- *Gamma distributions: $\mathcal{H}_J$ is injective, and $\mathcal{C}_1 \subset \mathcal{C}_J$.*

Theoretical Results 1

Let $\mathcal{H}_J$ be the joint hypothesis space of $\mathcal{H}_1$ and $\mathcal{H}_2$ containing operators $\mathcal{M} \to \mathbb{P}(\mathcal{Y})$ defined by

$$p(y \mid \boldsymbol{m}(\boldsymbol{x})) = \mathbb{E}_{\boldsymbol{\theta} \sim p(\boldsymbol{\theta} \mid \boldsymbol{m}(\boldsymbol{x}; \boldsymbol{\Phi}))}\, p(y \mid \boldsymbol{\theta}).$$

Theorem 3.2: Unregularized Inner Loss Minimisation

Let $\mathcal{C}_1$ and $\mathcal{C}_J$ be the co-domains of $\mathcal{H}_1$ and $\mathcal{H}_J$. The following properties hold when $\mathcal{H}_2$ consists of

- *Dirichlet distributions: $\mathcal{H}_J$ is not injective, and $\mathcal{C}_1 = \mathcal{C}_J$;*
- *NIG distributions: $\mathcal{H}_J$ is not injective, and $\mathcal{C}_1 \subset \mathcal{C}_J$;*
- *Gamma distributions: $\mathcal{H}_J$ is injective, and $\mathcal{C}_1 \subset \mathcal{C}_J$.*

Implications: **non-identifiability** for Dirichlet, NIG priors, leading to a wide spectrum of possible values for entropy of $p(\boldsymbol{\theta} | \boldsymbol{m}(\boldsymbol{x_i}, \phi)$

Theoretical Results 2

Generalization of Theorem 1 of Bengs et al[14] for all exponential family members:

Theorem 3.3: Unregularized Outer Expectation Minimisation

*Let $L_1 : \mathcal{Y} \times \mathbb{P}(\mathcal{Y})$ be convex in its second argument, and let $\mathcal{H}_2$ be a hypothesis space that contains universal approximators. A parameter set $\boldsymbol{\Phi}$ that yields $p(\boldsymbol{\theta} \mid \boldsymbol{m}(\boldsymbol{x}_i; \hat{\boldsymbol{\Phi}})) = \delta(\boldsymbol{\theta}(\boldsymbol{x}_i))$ for all $i \in \{1, \ldots, N\}$ is a minimizer of the outer expectation minimization problem, where δ is the **Dirac delta function** on $\boldsymbol{\theta}(\boldsymbol{x}_i) \in \Theta$.*

[14] Bengs, Hüllermeier, and Waegeman, "Pitfalls of Epistemic Uncertainty Quantification through Loss Minimisation".

Theoretical Results 2

Generalization of Theorem 1 of Bengs et al[14] for all exponential family members:

Theorem 3.3: Unregularized Outer Expectation Minimisation

*Let $L_1 : \mathcal{Y} \times \mathbb{P}(\mathcal{Y})$ be convex in its second argument, and let $\mathcal{H}_2$ be a hypothesis space that contains universal approximators. A parameter set $\boldsymbol{\Phi}$ that yields $p(\boldsymbol{\theta} \mid \boldsymbol{m}(\boldsymbol{x}_i; \hat{\boldsymbol{\Phi}})) = \delta(\boldsymbol{\theta}(\boldsymbol{x}_i))$ for all $i \in \{1, \ldots, N\}$ is a minimizer of the outer expectation minimization problem, where δ is the **Dirac delta function** on $\boldsymbol{\theta}(\boldsymbol{x}_i) \in \Theta$.*

Implications: one of the risk minimizers is a Dirac distribution..

[14] Bengs, Hüllermeier, and Waegeman, "Pitfalls of Epistemic Uncertainty Quantification through Loss Minimisation".

Theoretical Results 3

Theorem 3.3: Regularized Loss Minimisation

Consider the optimization problems for inner and outer expectation minimization with the entropy regularization of the second-order distribution. Then, for any data-generating distribution P, there exists $\lambda \geq 0$ and $\boldsymbol{x} \in \mathcal{X}$ for which the second-order distribution differs from the reference distribution in Def 3.1.

Theoretical Results 3

Theorem 3.3: Regularized Loss Minimisation

Consider the optimization problems for inner and outer expectation minimization with the entropy regularization of the second-order distribution. Then, for any data-generating distribution P, there exists $\lambda \geq 0$ and $\boldsymbol{x} \in \mathcal{X}$ for which the second-order distribution differs from the reference distribution in Def 3.1.

Implications: $\lambda \in \mathbb{R}_{\geq 0}$ defines an *uncertainty budget* that cannot be exceeded for the given amount of training data points!

Empirical Results

Experimental Results: Classification

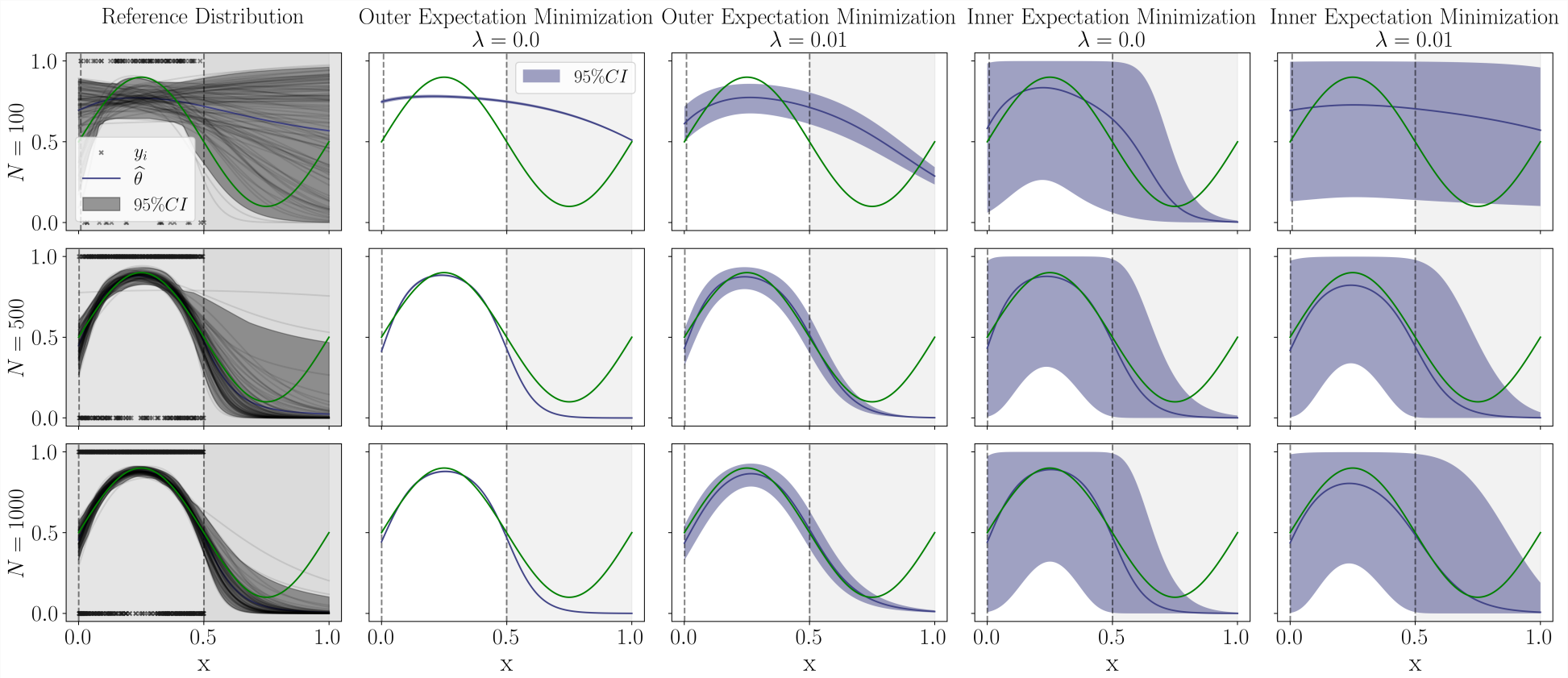


Figure 6: Binary classification experiments for training sample size $N \in \{100, 500, 1000\}$, using synthetic training data $\mathcal{D}_N = \{(x_i, y_i)\}_{i=1}^N$, where $x_i \sim U([0, 0.5])$, $y_i \sim Bern(\theta(x_i))$. The underlying Bernoulli parameter is defined by $\theta(x_i) = 0.5 \cdot (0.8 \cdot \sin(2\pi x_i) + 1)$.

$\rightarrow$ more experimental results in

`https://arxiv.org/pdf/2402.09056.pdf`

Conclusion

Discussion

Summary

- DER does not result in distributions that are faithfully representing EU
- The regularization parameter λ yields an "uncertainty budget" that cannot be exceeded for a given amount of training data points
- While outer loss minimisation can lead to a convergence to a Dirac delta distribution, inner loss minimisation leads to training instability

Future outlook:

- further analysis is needed, especially for different regularization losses
- distance-based approaches to quantify distance to reference distribution
- experimental analysis of heteroscedastic case

Conclusion

For regression results, proofs, more theoretical analysis, check out our paper:

THANK YOU!

Empirical Results: Regression

$N = 100$:

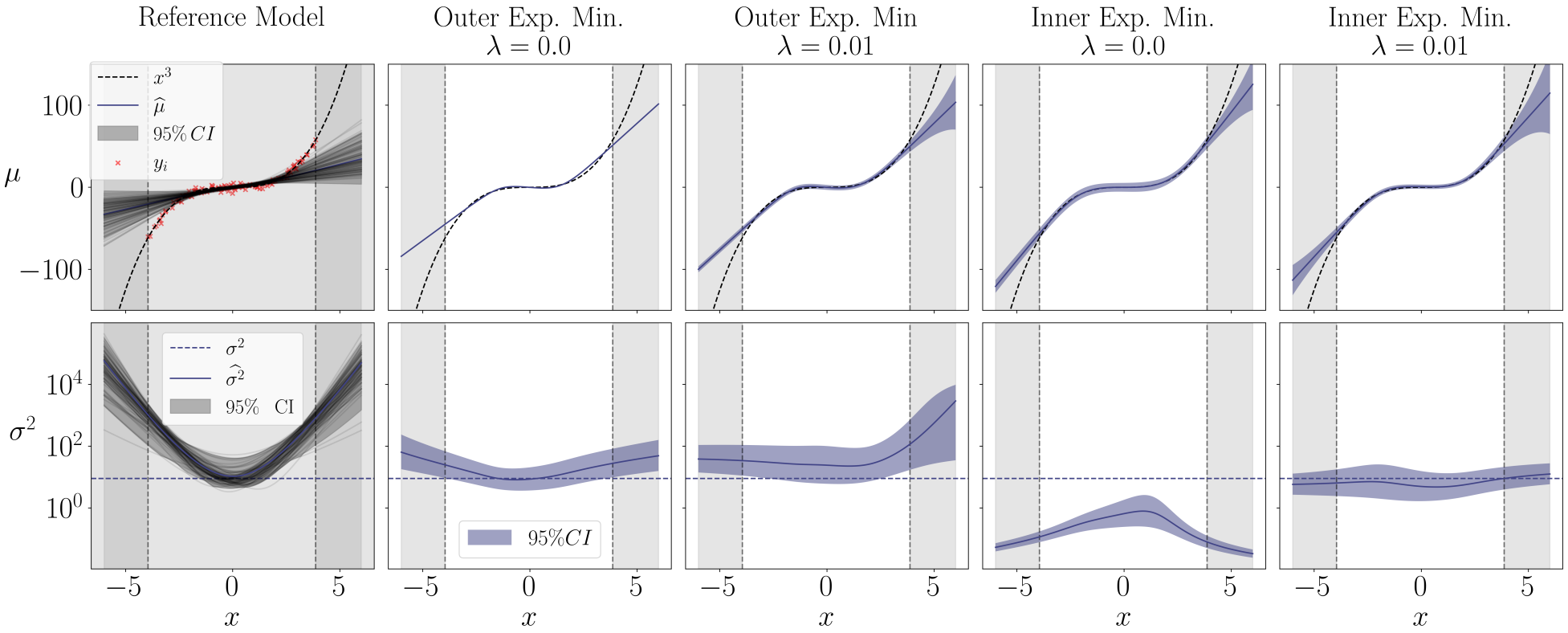
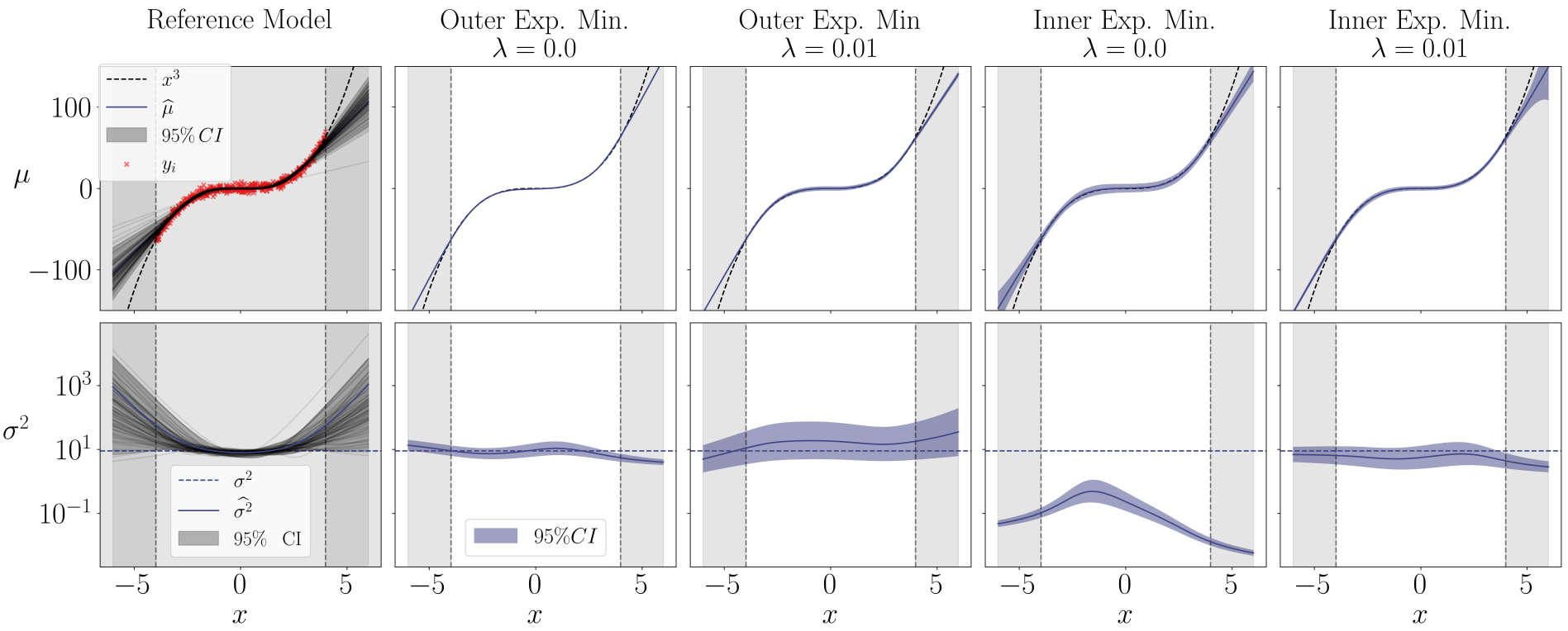


Figure 7: Regression experiment with $\mathcal{D}_N = \{(x_i, x_i^3 + \epsilon)\}_{i=1}^N$ for $N \in \{100, 500, 1000\}$, where $x_i \in U([-4, 4])$, $\epsilon \sim \mathcal{N}(0, \sigma^2 = 9)$. The reference model learns the parameters $\theta = (\mu, \sigma)$ of the underlying normal distribution, by optimizing the negative log-likelihood.

Empirical Results: Regression

$N = 500$:



Empirical Results: Regression

$N = 1000$: