Is Epistemic Uncertainty Faithfully Represented by Evidential Deep Learning Methods?

ICML 2024

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

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

What is Evidential Deep Learning (EDL)?

Idea of EDL

Learn a **second-order distribution** directly by which both *aleatoric* and *epistemic* uncertainty can be disentangled predicted.

Evidential Deep Learning for Classification

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 $\boldsymbol{m} \in \mathbb{R}_+^K$

Evidential Deep Learning for Classification

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 $\boldsymbol{m} \in \mathbb{R}_+^K$

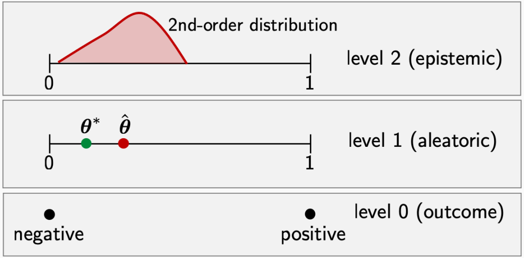


Fig. 1: Different levels of predictions[a] (binary classification)

[a] 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-Inverse-Gamma Model

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

[1] Amini et al., "Deep evidential regression".
[2] Meinert, Gawlikowski, and Lavin, "The unreasonable effectiveness of deep evidential regression".

Evidential Deep Learning for Regression

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

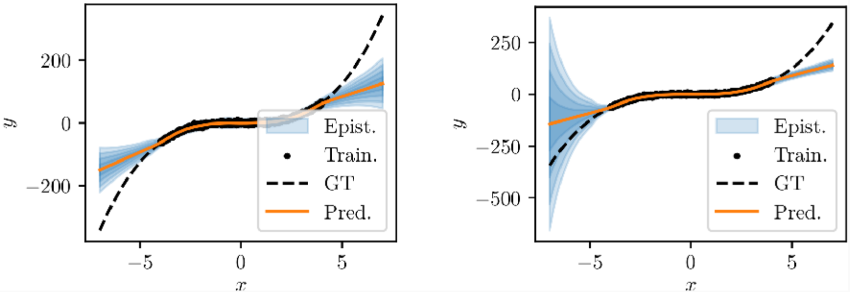


Fig. 2: Results of Deep Evidential Regression[1] reproduced for two different runs[2]

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

[2] Meinert, Gawlikowski, and Lavin, "The unreasonable effectiveness of deep evidential regression".

Choose your Distribution!

Problem	Likelihood	Conjugate prior
classification[3]	$y \sim Cat(\boldsymbol{\theta})$ $\boldsymbol{\theta} \in \Delta_K$	$\boldsymbol{\theta} \sim Dir(\boldsymbol{m})$ $\boldsymbol{m} \in \mathbb{R}_+^K$
regression (univariate)[4]	$y \sim \mathcal{N}(\mu, \sigma^2)$ $\mu \in \mathbb{R},\ \sigma^2 \in \mathbb{R}_+$	$\boldsymbol{\theta} \sim N\text{-}\Gamma^{-1}(\boldsymbol{m})$ $\boldsymbol{m} = (\gamma, \nu, \alpha, \beta) \in \mathbb{R} \times \mathbb{R}_+^3$
regression (multivariate)[5]	$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ $\boldsymbol{\mu} \in \mathbb{R}^D,\ \boldsymbol{\Sigma} \in \mathbb{R}^{D \times D}$	$\boldsymbol{\theta} \sim NIW(\boldsymbol{m})$ $\boldsymbol{m} = (\boldsymbol{\mu}_0, \kappa, \nu, \boldsymbol{\Phi})$
point processes	$y \sim Pois(\boldsymbol{\theta})$ $\boldsymbol{\theta} \in \mathbb{R}_+$	$\boldsymbol{\theta} \sim \Gamma(\boldsymbol{m})$ $\boldsymbol{m} = (\alpha, \beta) \in \mathbb{R}_+^2$

$\boldsymbol{\theta}(\boldsymbol{x}, \phi) : \mathcal{X} \to \boldsymbol{\Theta}$ (1st-order predictor), $\boldsymbol{m}(\boldsymbol{x}, \phi) : \mathcal{X} \to \mathcal{M}$ (2nd-order predictor)

[3] Sensoy, Kaplan, and Kandemir, "Evidential deep learning to quantify classification uncertainty".
[4] Amini et al., "Deep evidential regression".
[5] 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[6]
- Non-Convergence in the Regression Case[7]
- Non-Properness of its Loss functions[8]

$\longrightarrow$ **Our approach:** Does the learned **second order distribution** represent the (epistemic) uncertainty of the 1st-order parameters in a faithful way?

[6] Bengs, Hüllermeier, and Waegeman, "Pitfalls of epistemic uncertainty quantification through loss minimisation".
[7] Meinert, Gawlikowski, and Lavin, "The unreasonable effectiveness of deep evidential regression".
[8] Bengs, Hüllermeier, and Waegeman, "On second-order scoring rules for epistemic uncertainty quantification".

Comparing 1st and 2nd-Order Risk Minimization

1st-Order vs 2nd-Order Risk Minimization

1st-Order Risk Minimization

- Loss function (e.g., NLL)

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

- **Risk minimization**

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

1st-Order vs 2nd-Order Risk Minimization

1st-Order Risk Minimization

- Loss function (e.g., NLL)

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

- **Risk minimization**

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

2nd-Order Risk Minimization

- Loss function (e.g., NLL)

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

- ***Outer* expectation minimization**

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

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 minimizing per-instance KL-divergence:

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

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 minimizing 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 $p(\boldsymbol{\theta} \mid \boldsymbol{m}_0)$ to parametrize uniform distribution (on 1st-order distributions)

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 minimizing 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 $p(\boldsymbol{\theta} \mid \boldsymbol{m}_0)$ to parametrize uniform distribution (on 1st-order distributions)
- minimizing **KL-divergence** $\leftrightarrow$ maximizing **entropy**

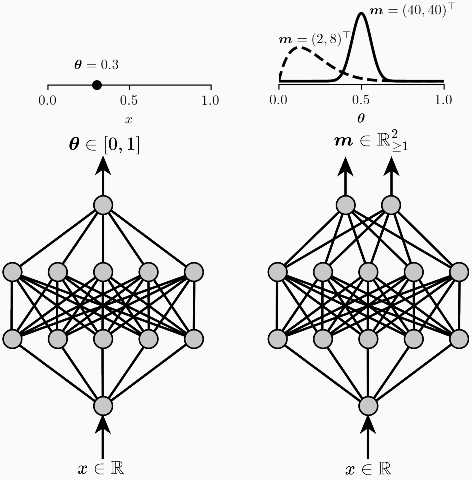
Regularization

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 minimizing 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 $p(\boldsymbol{\theta} \mid \boldsymbol{m}_0)$ to parametrize uniform distribution (on 1st-order distributions)
- minimizing **KL-divergence** $\leftrightarrow$ maximizing **entropy**

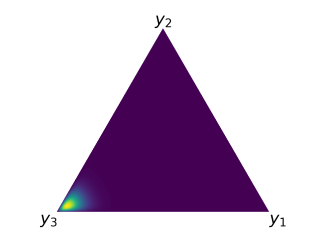
Example: 1st vs 2nd Order Risk Minimization



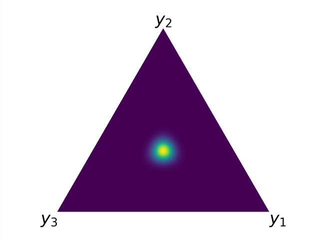
Do the resulting 2nd-order distributions represent the underlying epistemic uncertainty of the 1st-order predictor?

What is a Faithful Representation of Epistemic Uncertainty?

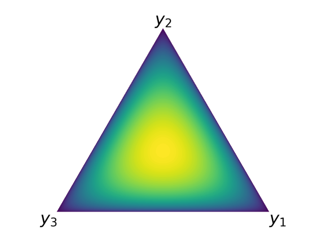
Spread of the distribution should yield a valid estimate of EU:



(a) low uncertainty



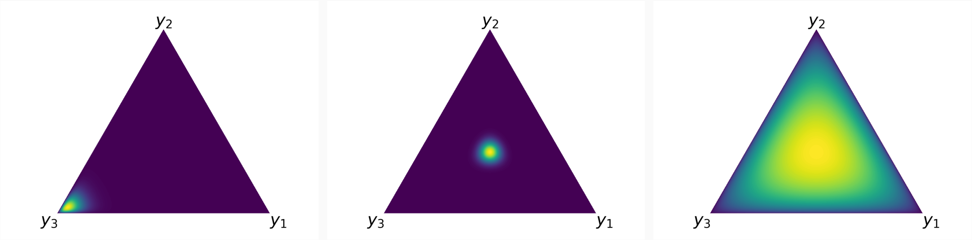
(b) high aleatoric uncertainty



(c) high epistemic uncertainty

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

Desirable **convergence properties**:[9]

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

[9]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 2nd-order distribution?

What is a Faithful Representation of Epistemic Uncertainty?

How can we directly evaluate the faithfulness of the resulting 2nd-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 1st-order minimization problem for a training set $\mathcal{D}_N$ of size N, where $\mathcal{D}_N \sim P^N$. Define the *reference 2nd-order distribution* as

$$q_N(\boldsymbol{\theta} \mid \boldsymbol{x}) := \mathbb{P}\big(\boldsymbol{\theta}_{\mathcal{D}_N}(\boldsymbol{x}; \boldsymbol{\Phi}_{\mathcal{D}_N}) = \boldsymbol{\theta}\big) \quad \text{for } x \in \mathcal{X}.$$

How can we directly evaluate the faithfulness of the resulting 2nd-order distribution?

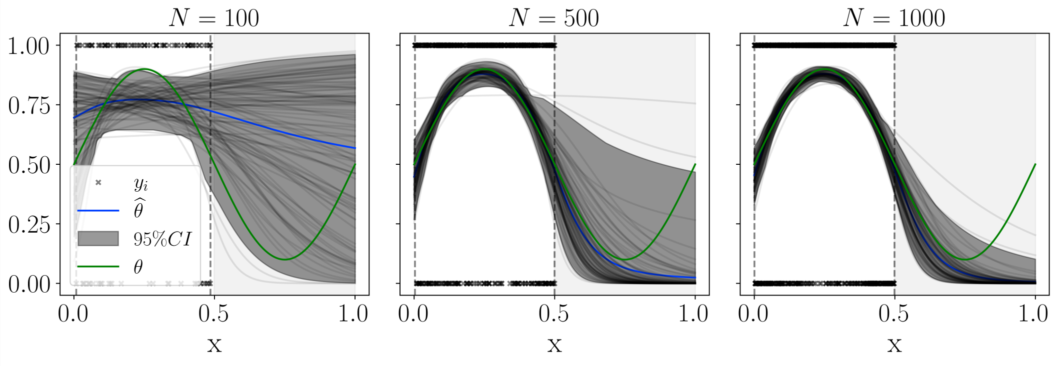
Def. 3.1: Reference Distribution

Let $\boldsymbol{\theta}_{\mathcal{D}_N}(\boldsymbol{x}; \boldsymbol{\Phi}_{\mathcal{D}_N})$ denote the minimizer of the 1st-order minimization problem for a training set $\mathcal{D}_N$ of size N, where $\mathcal{D}_N \sim P^N$. Define the *reference 2nd-order distribution* as

$$q_N(\boldsymbol{\theta} \mid \boldsymbol{x}) := \mathbb{P}\big(\boldsymbol{\theta}_{\mathcal{D}_N}(\boldsymbol{x}; \boldsymbol{\Phi}_{\mathcal{D}_N}) = \boldsymbol{\theta}\big) \quad \text{for } x \in \mathcal{X}.$$

$\Longrightarrow$ **Can be approximated empirically** by resampling and computing the empirical distribution function.

Reference Distribution: Example



Theoretical Results

Theoretical Results: Unregularized Case ($\lambda = 0$)

Theorem 3.2 and Theorem 3.1. in our paper show that

- **non-identifiability** issues arise for inner loss minimisation, leading to a wide range of possible values for the estimated uncertainty
- **convergence to a Dirac Delta distribution** in the case of outer loss minimisation (generalization of Theorem 1 of Bengs et al[10] to all exponential family members)

[10] Bengs, Hüllermeier, and Waegeman, "Pitfalls of epistemic uncertainty quantification through loss minimisation".

Theoretical Results: Regularized Case ($\lambda > 0$)

Theorem 3.3 shows that for inner and outer loss minimisation with entropy regularization

- there exists $\lambda \geq 0$, $x \in \mathcal{X}$ for which the 2nd-order distribution differs from the reference distribution.

Theoretical Results: Regularized Case ($\lambda > 0$)

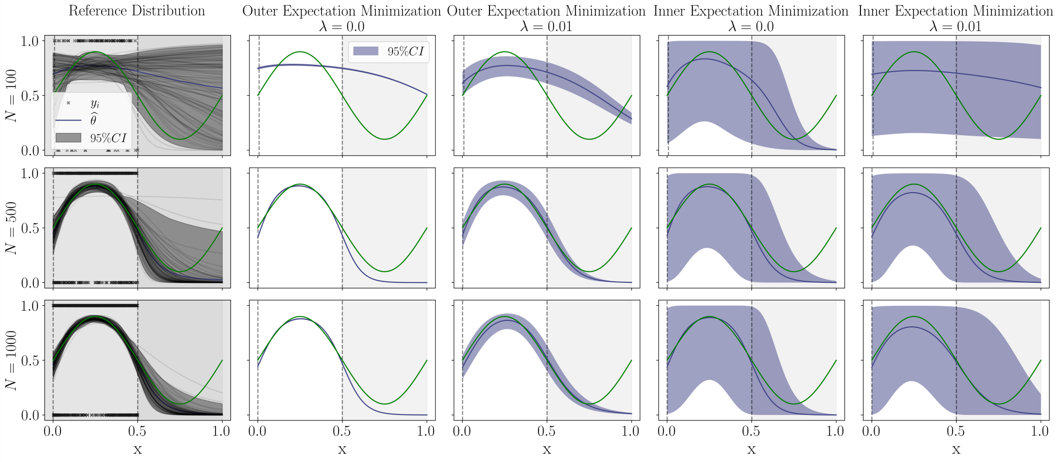
Theorem 3.3 shows that for inner and outer loss minimisation with entropy regularization

- there exists $\lambda \geq 0$, $x \in \mathcal{X}$ for which the 2nd-order distribution differs from the reference distribution.

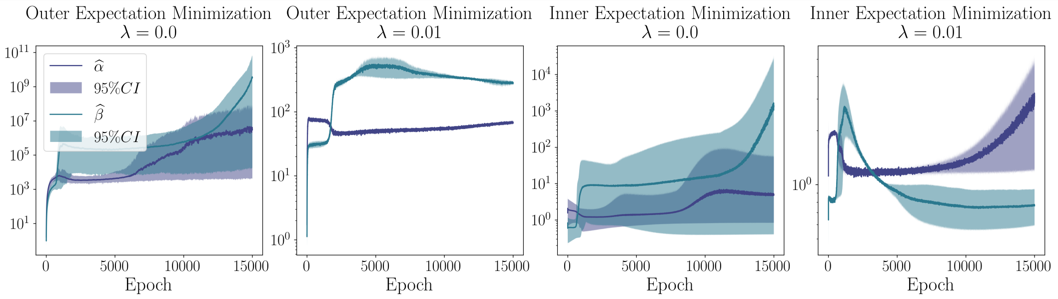
Common ways in EDL to optimize the parameter λ are mainly based on heuristics, defining an *uncertainty budget* that cannot be exceeded!

Empirical Results

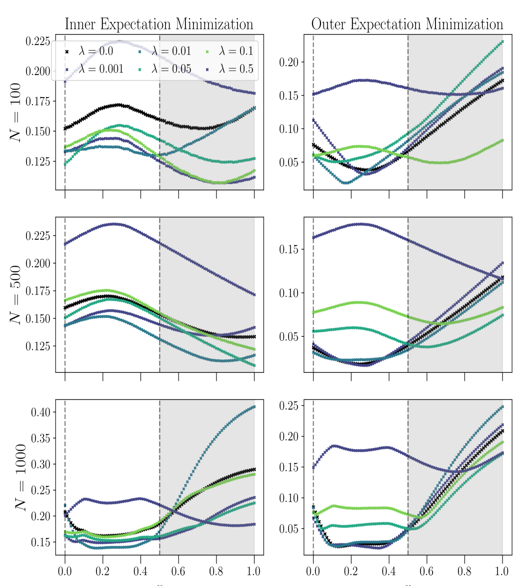
(Some) Experimental Results: Classification



(Some) Experimental Results: Convergence Analysis



(Some) Experimental Results: Distance Analysis



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

Outlook:

- further analysis on different regularization
  - different regularizers
  - more advanced second-order loss functions

  is needed

For other exp. results, proofs, and more theoretical analysis, check out our paper.

THANK YOU

Conclusion

For other exp. results, proofs, and more theoretical analysis, check out our paper:

`arXiv:2402.09056 [cs.AI]`

(*Is Epistemic Uncertainty Faithfully Represented by Evidential Deep Learning Methods?*)

Empirical Results: Regression ($N = 100$)

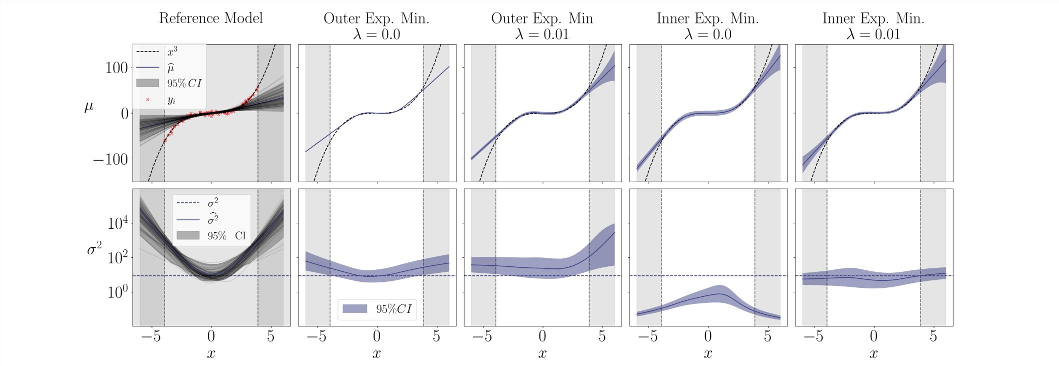
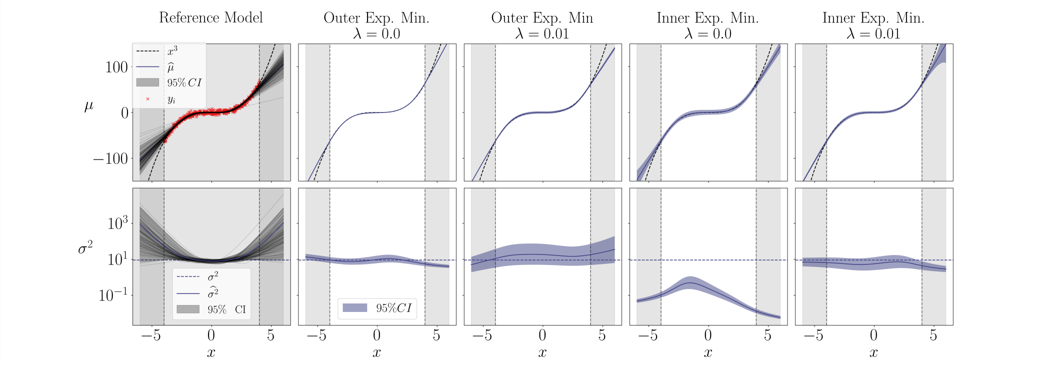


Fig. 4: Regression experiment with $\mathcal{D}_N = \{(x_i, x_i^3 + \epsilon)\}_{i=1}^N$ for $N \in \{100, 500, 1000\}$, where $x_i \in U([-4, 4])$, $\epsilon \sim \mathcal{N}(0, \sigma^2 = 9)$. The reference model learns the parameters $\theta = (\mu, \sigma)$ of the underlying normal distribution,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

Empirical Results: Regression ($N = 500$)



Empirical Results: Regression ($N = 1000$)