Quantifying Local Model Validity using Active Learning

Sven Lämmle[a, b] Can Bogoclu[c] Robert Voßhall[d] Anselm Haselhoff[e] Dirk Roos[b]

[a]CoE, ZF Friedrichshafen AG [b]IMH, Niederrhein UAS [c]Zalando SE [d]auxmoney GmbH [e]Ruhr West UAS

Hochschule Niederrhein University of Applied Sciences

IMH Institut für Modellbildung und Hochleistungsrechnen Institute of Modelling and High-Performance Computing

Motivation

- Safety of deployed machine learning models is highly important in many applications.
- Data is costly to obtain in small sample settings such as in engineering or medical applications.
- Identify valid subdomains of input space with a model error smaller than some required tolerance.

a) Background & Problem

- Validate a model $f_{\mathrm{M}}\colon \mathbb{X} \to \mathbb{Y}$ over $\mathbb{X} \subset \mathbb{R}^d$
- Expensive observations $Y_{\mathbf{x}} = f_{\mathrm{E}}(\mathbf{x}) + \epsilon$ subject to homoscedastic Gaussian noise ϵ
- **Validation Metric.**

$$P\left(-\xi < f_{\mathrm{D}}(\mathbf{x}) < \xi\right),$$

  with tolerance $\xi \in \mathbb{R}_{>0}$ and model discrepancy $f_{\mathrm{D}}(\mathbf{x}) := f_{\mathrm{M}}(\mathbf{x}) - Y_{\mathbf{x}}$.
- Reformulation: $P\left(g(\mathbf{x}) > 0\right)$ with *limit state function* $g(\mathbf{x}) := \xi - |f_{\mathrm{D}}|$

Definitions

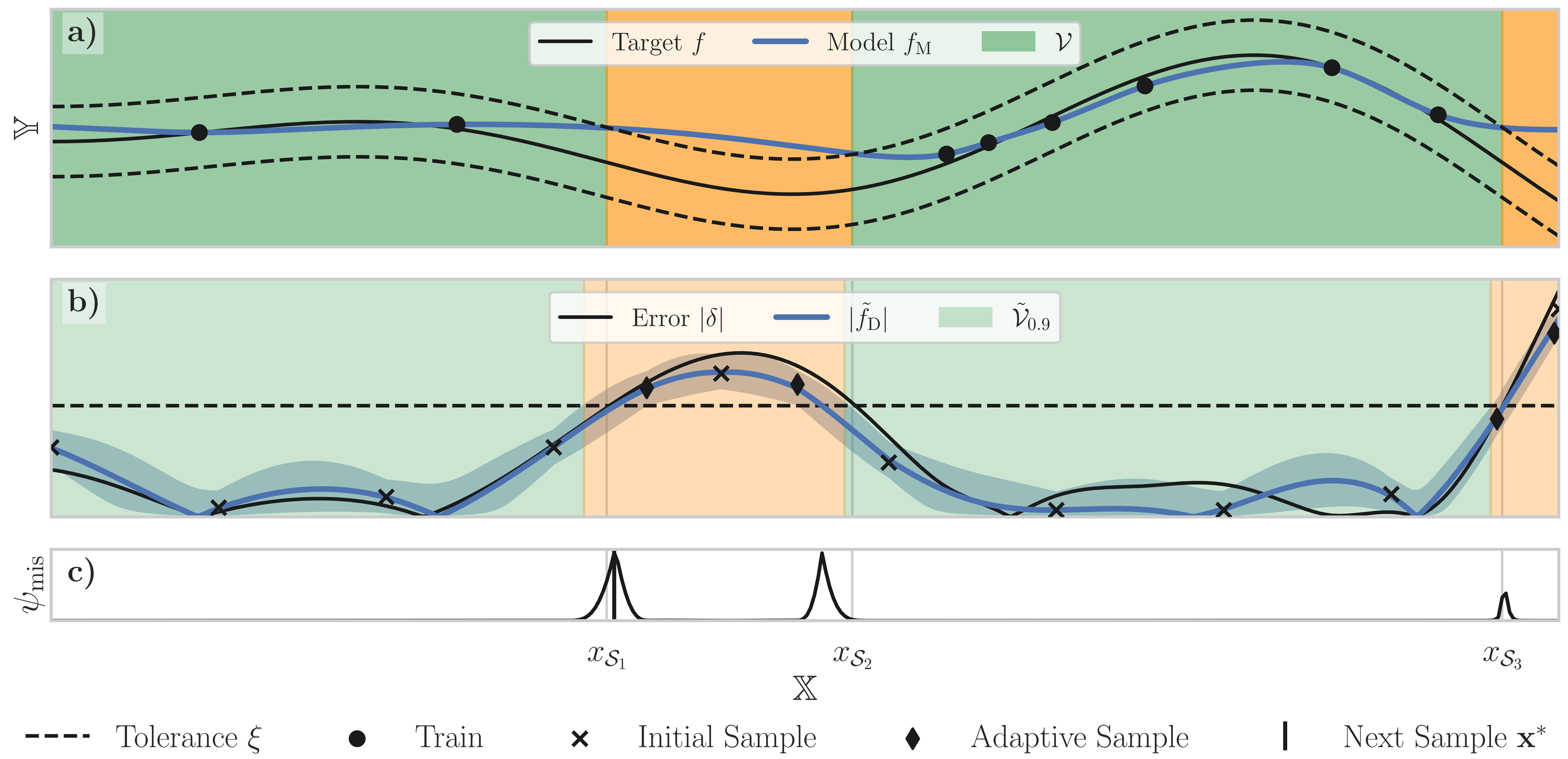
Local Validity. A model f_{M} is locally valid at $\mathbf{x}$, given a tolerance level ξ, if $\xi - |\delta(\mathbf{x})| \geq 0$. Then, the valid region of f_{M} is

$$\mathcal{V} = \{\mathbf{x} \in \mathbb{X}\colon \xi - |\delta(\mathbf{x})| \geq 0\},$$

with noiseless discrepancy $\delta(\mathbf{x}) = f_{\mathrm{M}}(\mathbf{x}) - f_{\mathrm{E}}(\mathbf{x})$.

Limit State. The limit state of f_{M} is given by

$$\mathcal{S} = \{\mathbf{x} \in \mathbb{X}\colon \xi - |\delta(\mathbf{x})| = 0\}.$$



b) Gaussian Process (GP) Error Model

Use a Gaussian process (GP) to learn the limit state

$$\hat{g} = \xi - |\tilde{f}_{\mathrm{D}}| \qquad \tilde{f}_{\mathrm{D}} \sim \mathcal{GP}\left(\mu, k\right).$$

The prediction is a folded Gaussian posterior, available in closed-form.

c) Learning the Limit State with MC-Prob.

Bayesian Active Learning. A new query $\mathbf{x}^*$ for evaluation is obtained as

$$\mathbf{x}^* = \arg\max_{\mathbf{x} \in \mathcal{C}} \psi_{\mathrm{mis}}(\mathbf{x}),$$

with candidates $\mathcal{C}$.

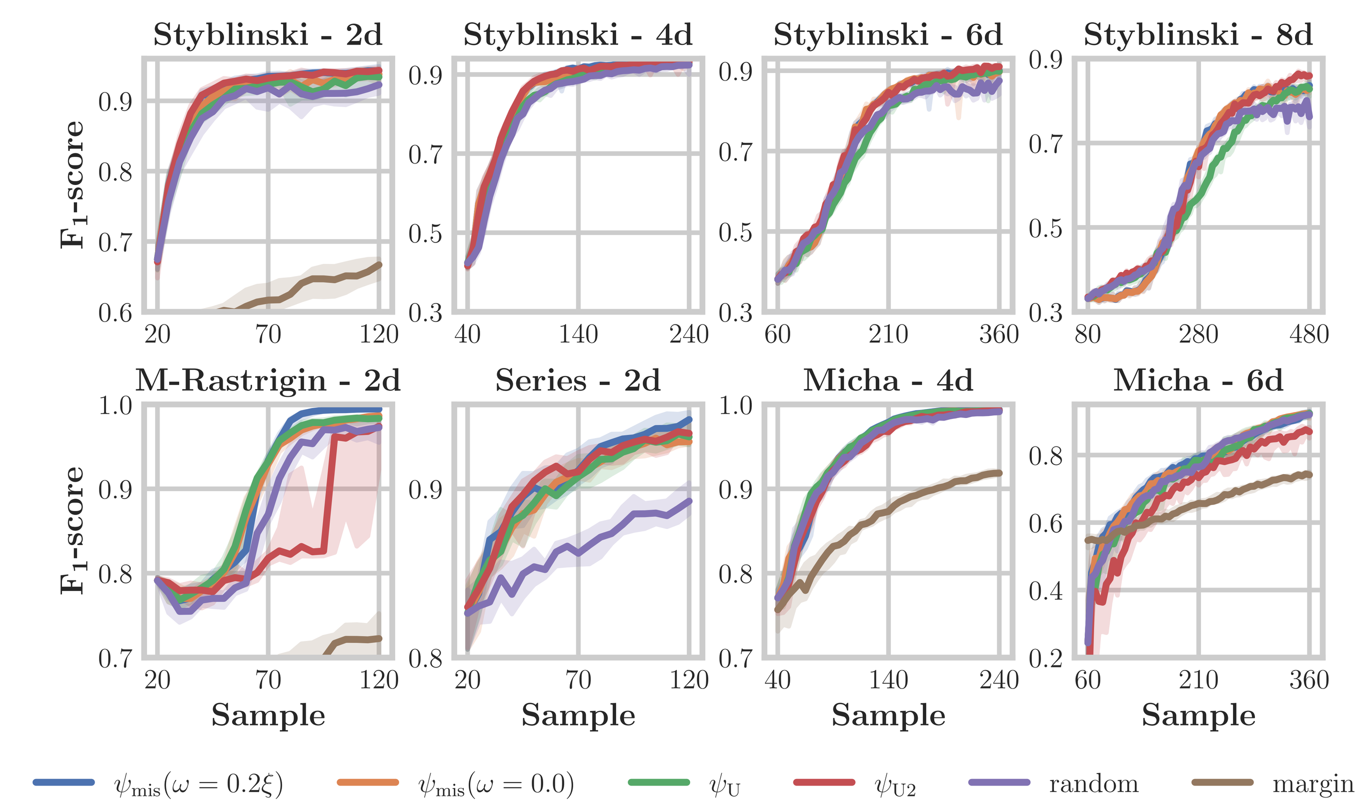
Acquisition Function. We use the misclassification probability (MC-Prob.) as acquisition function

$$\psi_{\mathrm{mis}}(\mathbf{x}; \omega) = \begin{cases} P\left(\hat{G}_{\mathbf{x}} \leq -\omega\right), & \text{for } |\mu_{y|\mathcal{D}}(\mathbf{x})| \leq \xi \\ 1 - P\left(\hat{G}_{\mathbf{x}} \leq \omega\right), & \text{for } |\mu_{y|\mathcal{D}}(\mathbf{x})| > \xi, \end{cases}$$

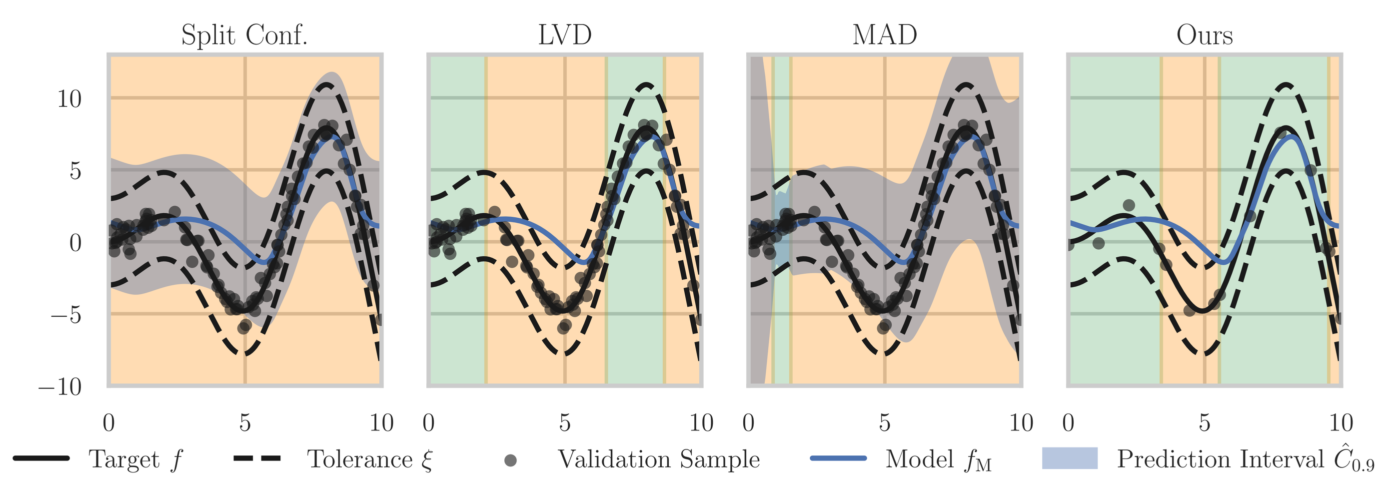
with hyperparmeter $\omega \in \mathbb{R}_+$ to control the exploration-exploitation trade-off.

Experiments

Benchmark results



Comparison with conformal prediction



Conclusion

- Novel formulation for local validation, inspired by active learning reliability
- Misclassification probability (MC-Prob) based on epistemic uncertainty is used
- Higher sample efficiency and thus accuracy in limited sample settings compared to previous work

Paper + Code