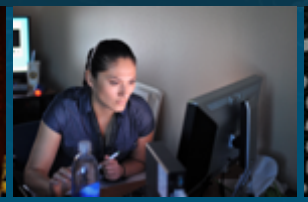
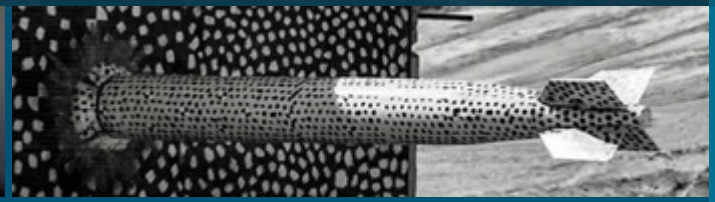
Sandia National Laboratories

SAND2020-12732C

Estimating Predictive Uncertainty in Scientific Machine Learning

A Library of Methods and Test Problems





Workshop
Uncertainty Management and Machine Learning in Engineering Applications
November 16-17, 2020

Ahmad A. Rushdi, Aubrey C. Eckert, Gabriel Huerta, Laura P. Swiler, Brian M. Adams

Sandia National Laboratories

U.S. DEPARTMENT OF ENERGY | NNSA National Nuclear Security Administration

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Table of Contents

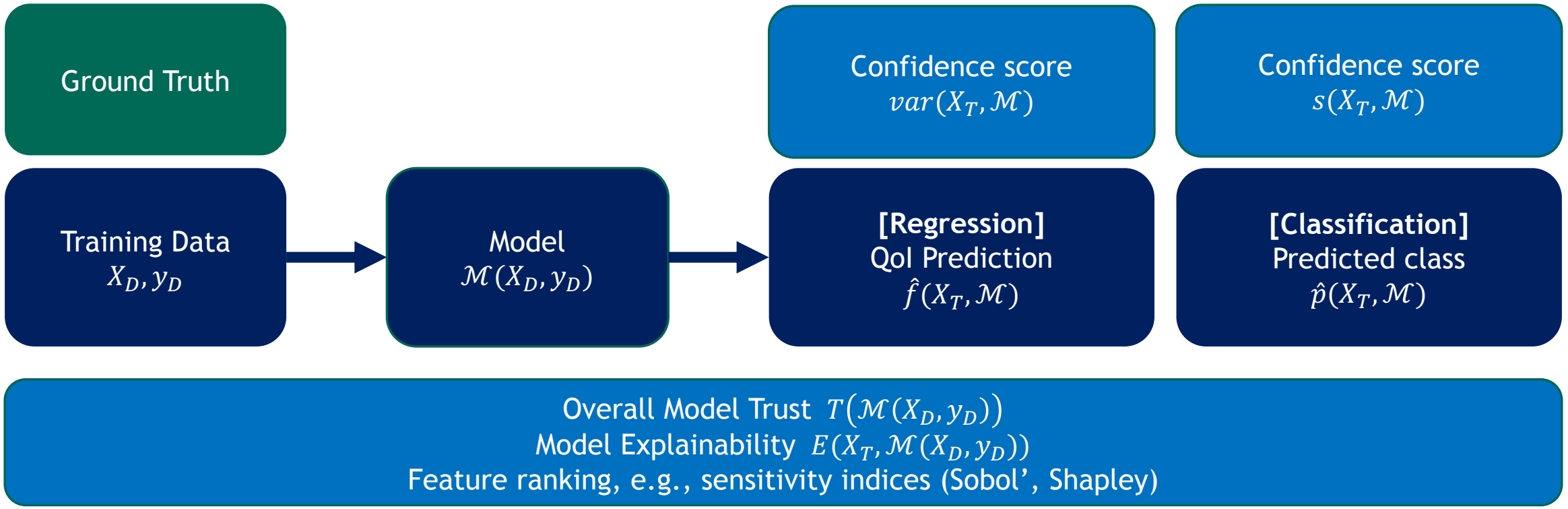
Problem Statement

Neural network models have attracted a lot of research attention in Scientific Machine Learning (SciML) problems. However, they tend to be overconfident when reporting typical point-estimate predictions in classification and regression problems. This could be very harmful when dealing with costly numerical simulations or high-stakes decisions in national security applications.

In this work, we assess uncertainty quantification techniques for neural network models. To understand their variability, we rely on different sources of randomness associated with training samples, weight initialization, dropout methods, and ensemble formations. Motivated by typical SciML situations, we assume a limited sample budget, noisy training data, and suggest approaches for reporting and possibly reducing uncertainty.

If an NN trained on some dataset is evaluated on a totally different dataset, it should be able to report higher predictive uncertainty. Inputs from a different dataset would be far away from the training data.

Problem Statement



Supervised Machine Learning: extract *models* from *data* and use them to make predictions.

Given data: $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\} \subset \mathcal{X} \times \mathcal{Y}$

minimize mean squared error of: $\frac{1}{N}\sum_i \|y_i - (\mathbf{x}_i\mathbf{W} + \mathbf{b})\|$

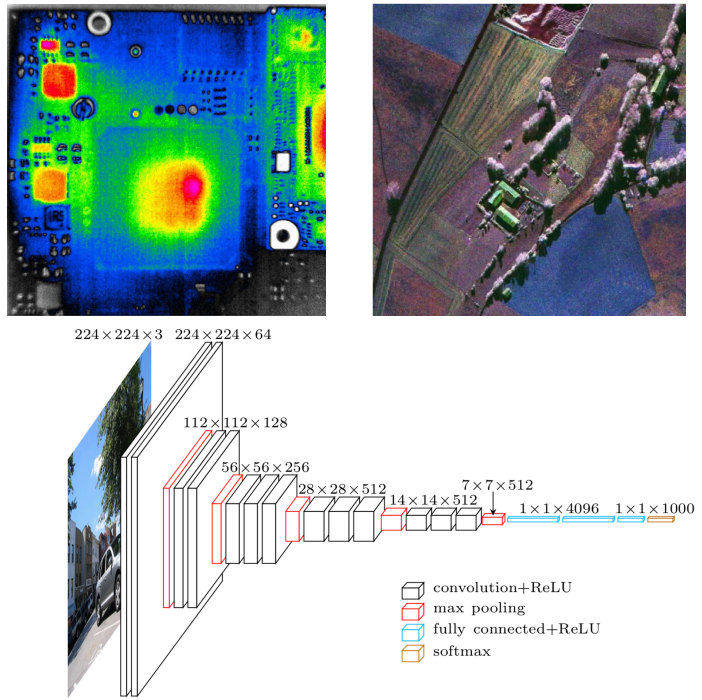
Problem Statement

Deep learning "black box" models are popular, yet are difficult to interpret and understand.

UQ today underpins many decision processes in nuclear security, our risk management and associated investments, which can be at the scale of billions of dollars. Predictions without UQ are neither predictions nor actionable. The data-rich world of ML, especially the powerful deep learning (DL) models, poses parallel challenges.

To develop consequential decision support from 'learned' models built on complex datasets, there is an important need to co-develop UQ for this domain.

Feature engineering in deep learning: rethinking architectures



- Begoli E, Tanmoy B, and Dimitri K "The need for uncertainty quantification in machine-assisted medical decision making." Nature Machine Intelligence, 2019
- Mehta, Pankaj, et al. "A high-bias, low-variance introduction to machine learning for physicists." Physics reports 810 (2019): 1-124

Predictive Uncertainty

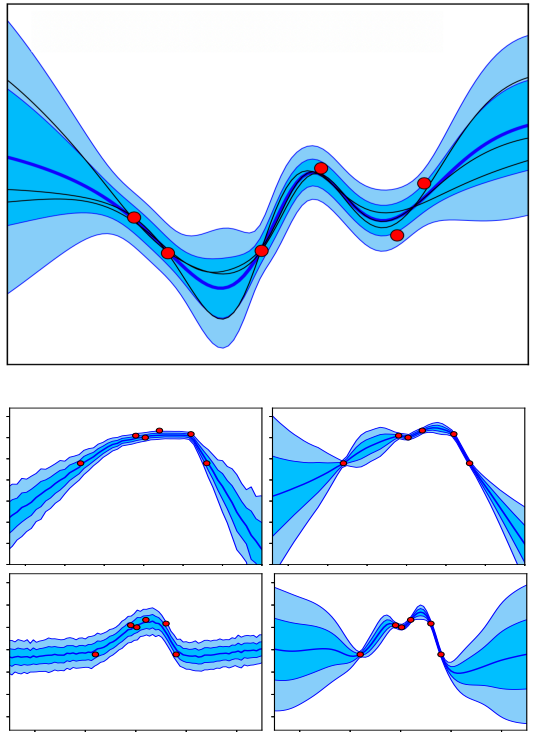
We represent DL predictions as **distributions** (μ, σ^2) rather than point estimates $\hat{p}$.

Scientific (vs. data rich) Machine Learning

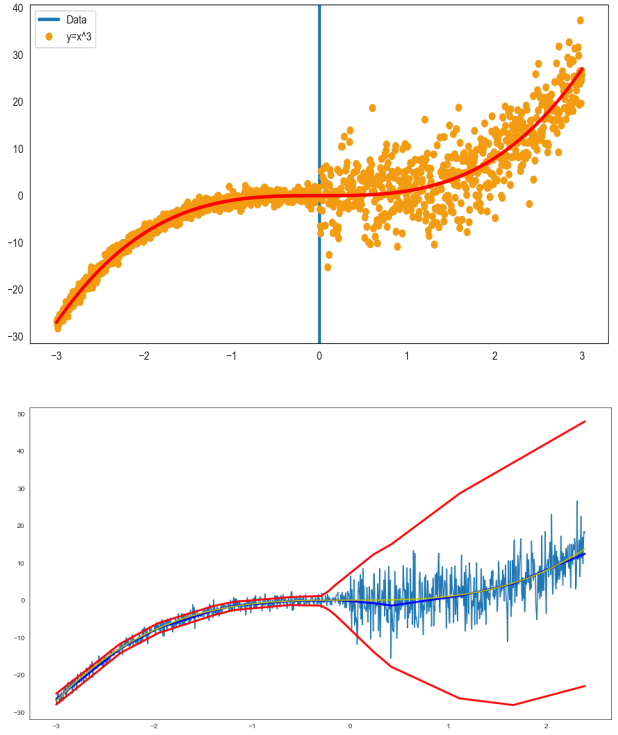
1. Training sample **budget is limited**, causing Out of Distribution (OoD) issues.
2. Different **types of noise**, imposed onto the input data and propagate that noise into an uncertainty metric in the resulting prediction.

Active Learning / Adaptive Sampling

- Model chooses which unlabeled data are most informative
- An uncertainty estimator (e.g., variance) suggests points of highest value for model accuracy improvement



Very few training samples



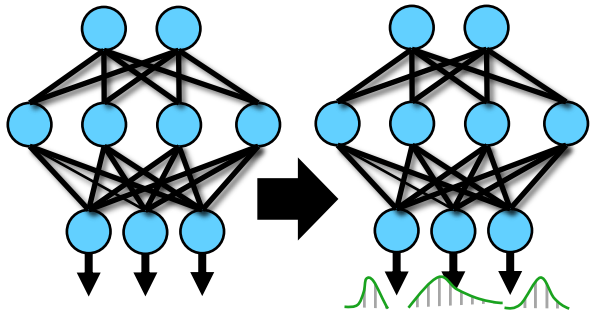
Noisy training data

Quality of prediction measures

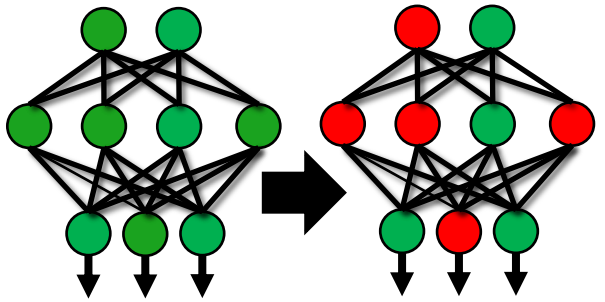
Our approach is to assign a numerical score to a prediction $p_\theta(y|\mathbf{x})$, rewarding *agreed upon predictions* over worse.

Approach – Methods

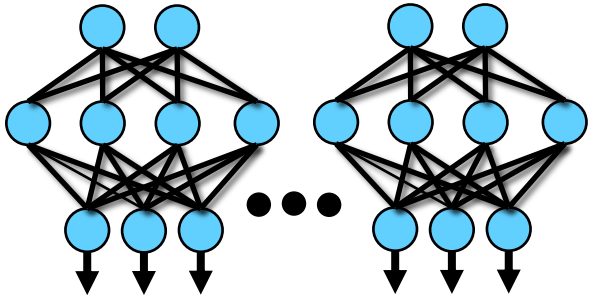
Bayesian NNs



MC Dropout

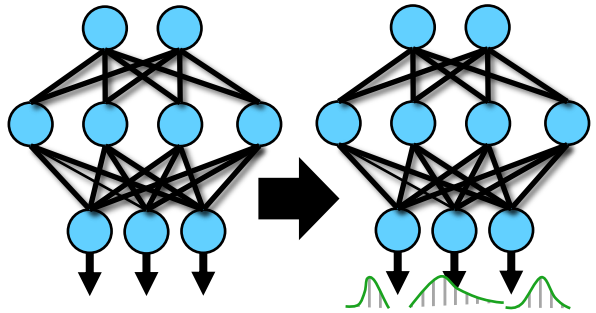


Deep Ensembles



Approach – Methods

Bayesian NNs



Bayesian Neural Networks (BNNs): place a prior distribution over the network weights and use data to learn a posterior distribution.

$$\begin{aligned} \boldsymbol{w}^{\mathrm{MLE}} &= \arg\max_{w} \log p(\mathcal{D}|\boldsymbol{w}) \\ &= \arg\max_{w} \textstyle\sum_i \log p(\boldsymbol{Y}_i|\boldsymbol{X}_i, \boldsymbol{w}) \\ \boldsymbol{w}^{\mathrm{MAP}} &= \arg\max_{w} \log p(\boldsymbol{w}|\mathcal{D}) \\ &= \arg\max_{w} \log p(\mathcal{D}|\boldsymbol{w}) + \log p(\boldsymbol{w}) \end{aligned}$$

Variational Inference (VI) approximation:

$$\theta^* = \arg\min KL\big(q_\theta(\boldsymbol{w}|\boldsymbol{D})||p(\boldsymbol{w}|\boldsymbol{D})\big)$$

Graves, Alex. "Practical variational inference for neural networks." *Advances in neural information processing systems*. 2011.

Exact Bayesian inference of posteriors:
computationally intractable.

Approximations:

- Markov chain Monte Carlo (MCMC)
- Laplace approximation
- Hamiltonian methods
- Variational Bayesian methods

Quality assessment:

- Choice of priors
- Degree of approximation

Hurdles:

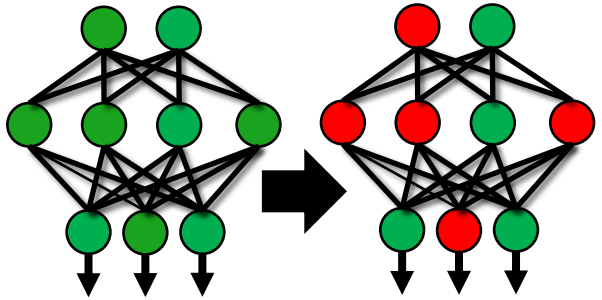
Computationally slower to train. Harder to implement.

Approach – Methods

Approximate Bayesian Interpretation

Popular in practice
Simplicity of implementation at test time.

MC Dropout



Neurons are randomly dropped in each iteration, with some probability (p), often fixed at empirical values (e.g., 0.3).

$$\mathbb{E}_{q(y^*|x^*)}(y^*) \approx \frac{1}{T}\sum_{t=1}^{T}\hat{y}^*(x^*, W_1^t, \cdots, W_L^t)$$

Average of T stochastic forward passes through the network.

Y. Gal and Z. Ghahramani. Dropout as a Bayesian approximation: Representing modeluncertainty in deep learning. InICML, 2016.

Dropout may be interpreted as an ensemble:
Predictions are averaged over an ensemble of NNs, when dropout rates are not tuned based on training.

Approach – Methods

Ensembles of NNs

have been successfully used to boost predictive performance in traditional ML problems (e.g. classification accuracy in ImageNet).

Variance

MSE captures only predictive mean, but not variance.

Adversarial training

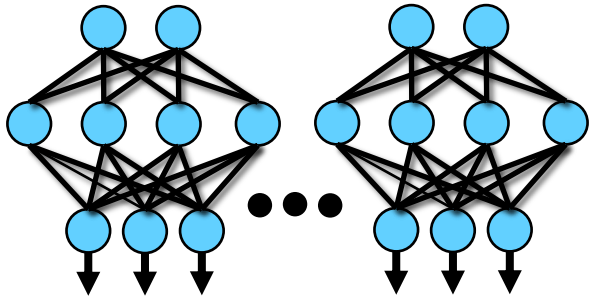
Typically: improves robustness to adversarial examples

Fast Gradient Sign Method (FGSM):

$$x' = x + \epsilon \operatorname{sign}\left(\nabla_x \ell(\theta, \boldsymbol{x}, y)\right)$$

Here: smooth predictive distributions.

Deep Ensembles



Estimate 2 outputs: mean and variance

Modified loss function + adversarial training

$$-\log p_\theta(y_n|\mathbf{x}_n) = \frac{\log \sigma_\theta^2(\mathbf{x})}{2} + \frac{(y-\mu_\theta(\mathbf{x}))^2}{2\sigma_\theta^2(\mathbf{x})}$$

Decomposed uncertainty

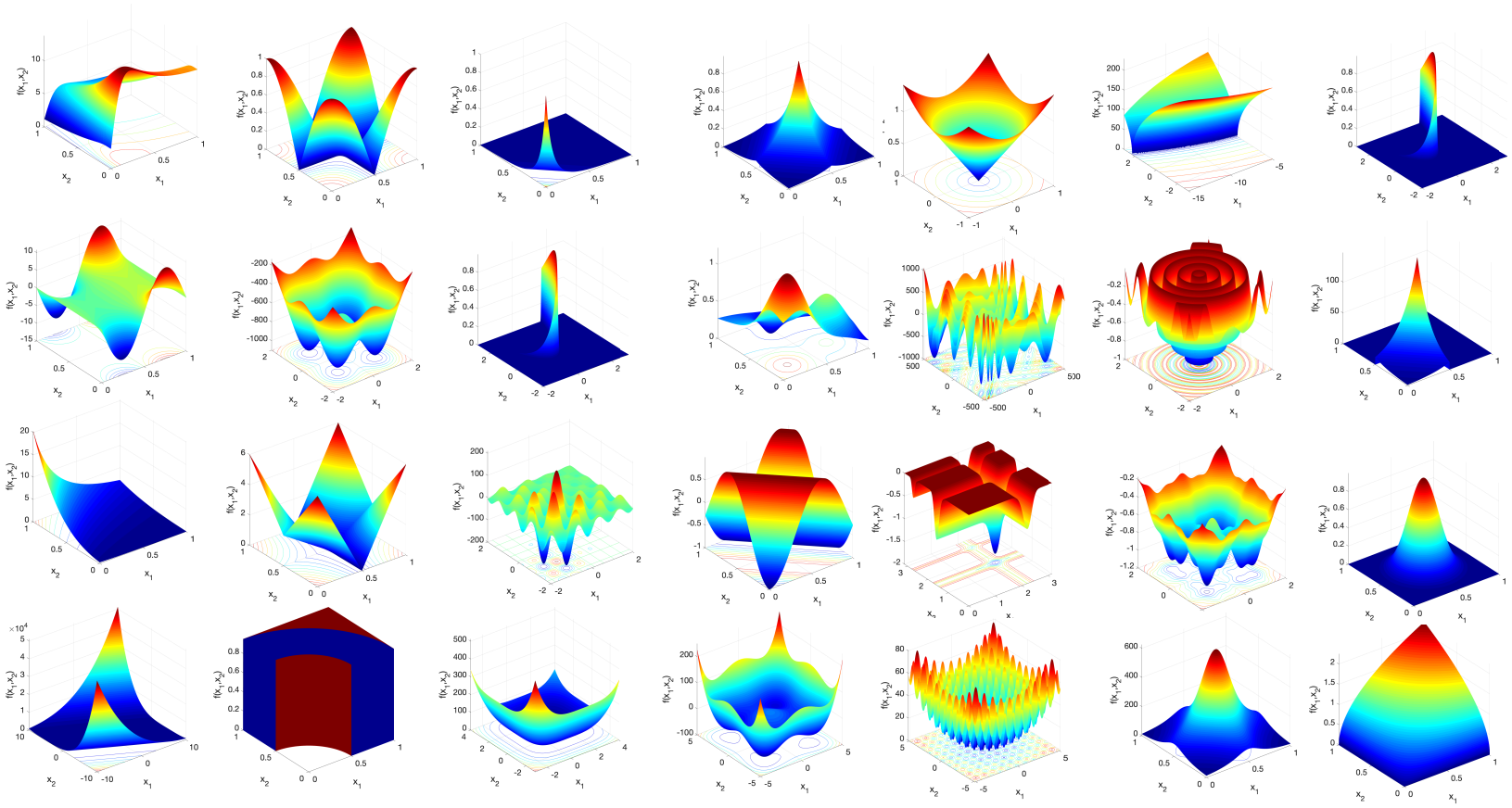
$$\mu_e = \frac{1}{M}\sum_{i=1}^{M} \mu_i \text{ and}$$

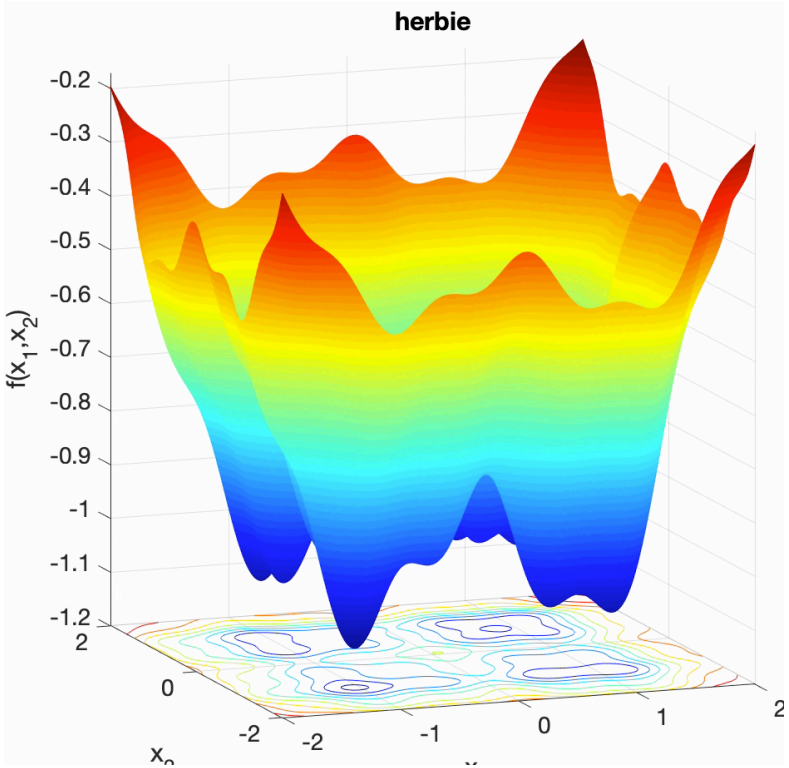
$$\sigma_e^2 = \underbrace{\frac{1}{M}\sum_{i=1}^{M} \sigma_i^2}_{\textit{Aleatoric}} + \underbrace{\left[\frac{1}{M}\sum_{i=1}^{M} \mu_i^2 - \mu_e^2\right]}_{\textit{Epistemic}}$$

Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell. "Simple and scalable predictive uncertainty estimation using deep ensembles." Advances in neural information processing systems. 2017.

Approach – Analytical Data

A library of **high-dimensional analytical test functions**, frequently used for QoI experimental studies (optimization, numerical integration, uncertainty quantification, and multi-fidelity analysis).





- Dakota www.Dakota.sandia.gov.
- Virtual Library of Simulation Experiments https://www.sfu.ca/~ssurjano/index.html

Numerical Experiments - Regression

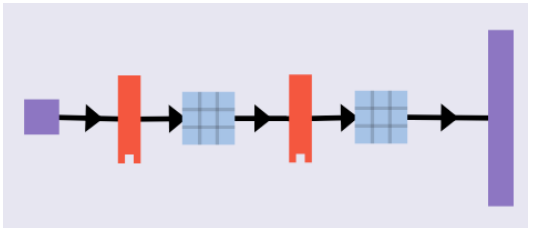
Data:

function samples + $\begin{cases} AWGN\,(0,\sigma_1) \\ AWGN\,(0,\sigma_2) \end{cases}$

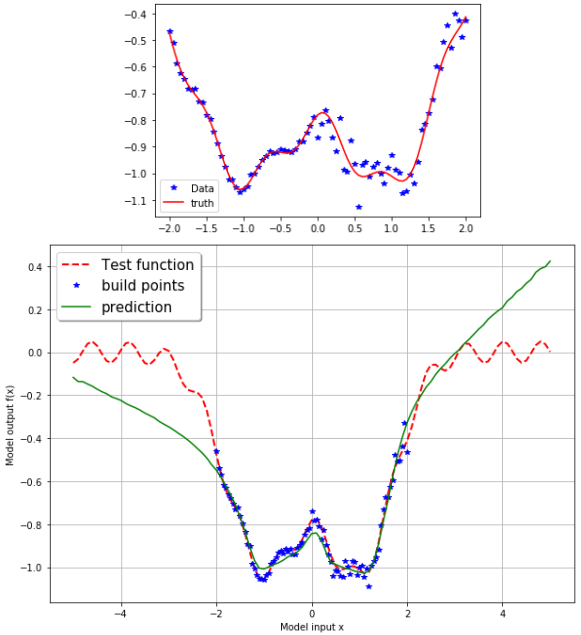
Model:

```
input = Input(shape=(1,))
x = Dense(512, activation="relu")(input)
x = Dropout(0.5)(x, training=True)
x = Dense(512, activation="relu")(x)
x = Dropout(0.5)(x, training=True)
output = Dense(1)(x)
```

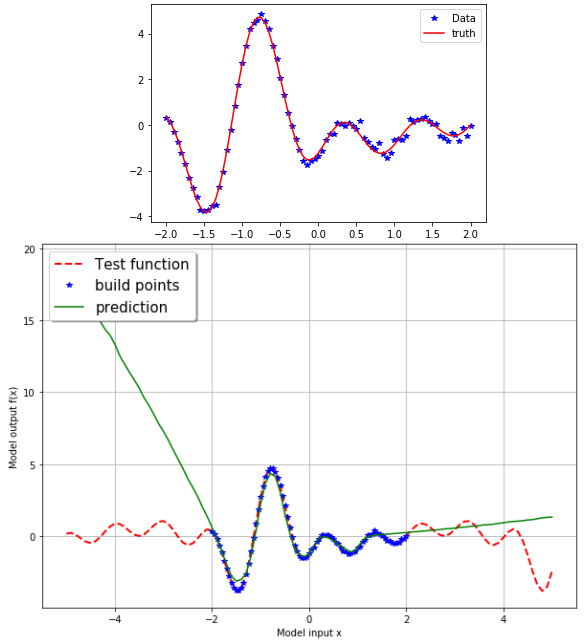
Total params: 264,193



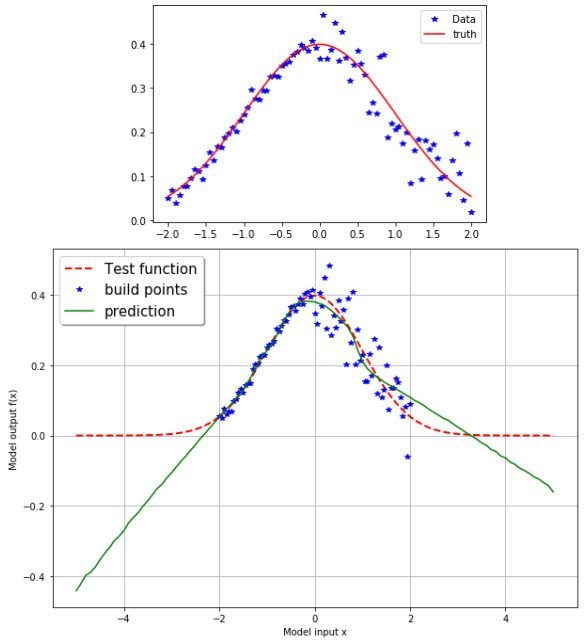
Herbie



Shubert



Gaussian

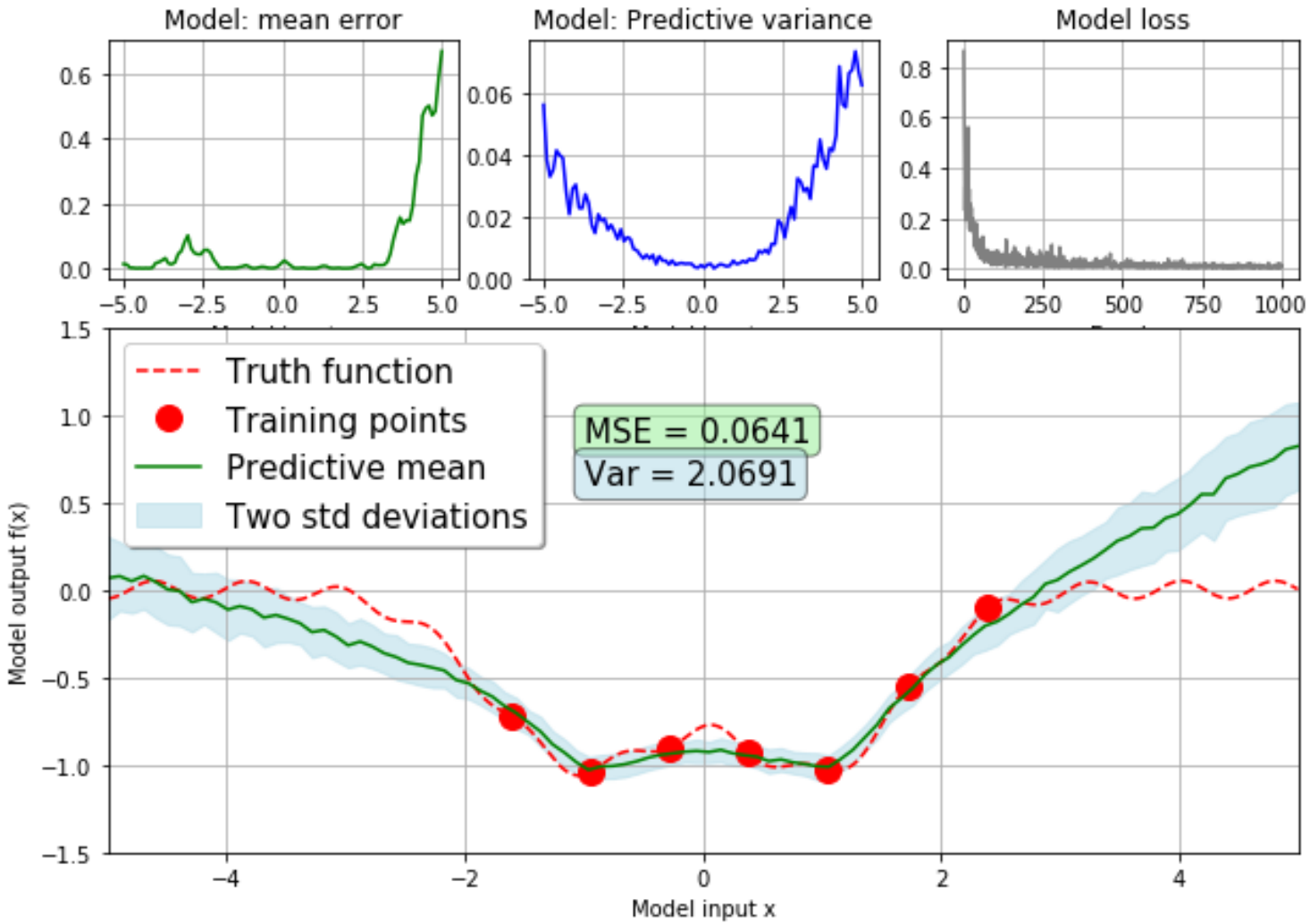


A model with 200K trainable parameters still needs an assessment of confidence/uncertainty.

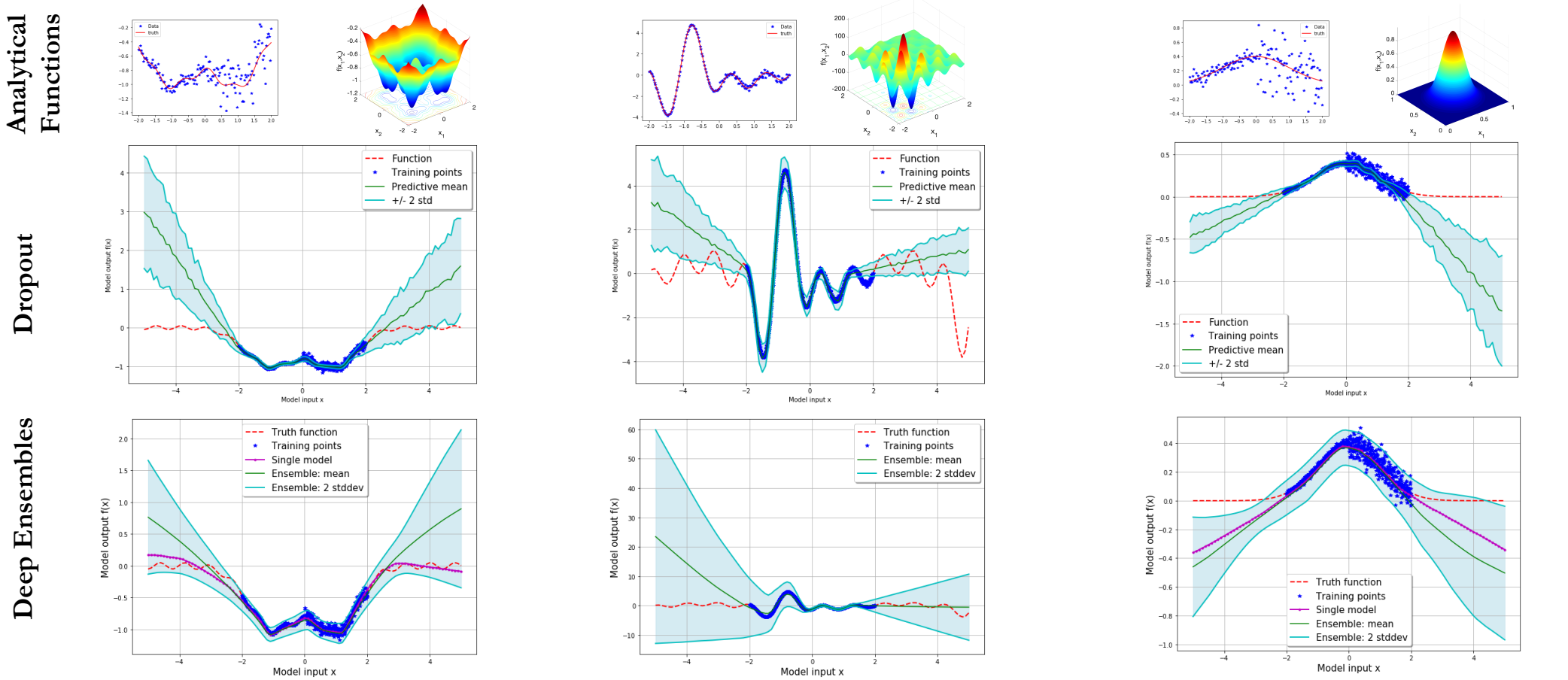
Numerical Experiments - Regression

Global Sensitivity Analysis for Model Explainability

- Uncertainty quantification at a new **test point**.
- Overall model uncertainty/confidence within a **test domain**.
- Guidance of **adaptive sampling** towards points/regions of high estimated variance.



Numerical Experiments - Regression

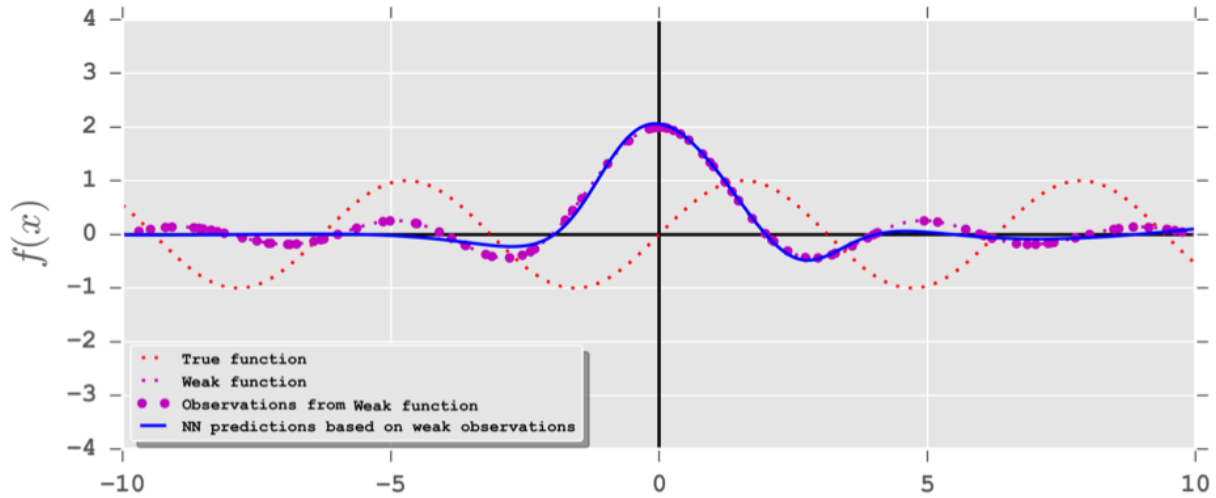


Methods vary in terms of complexity, accuracy, scalability, and computational cost.

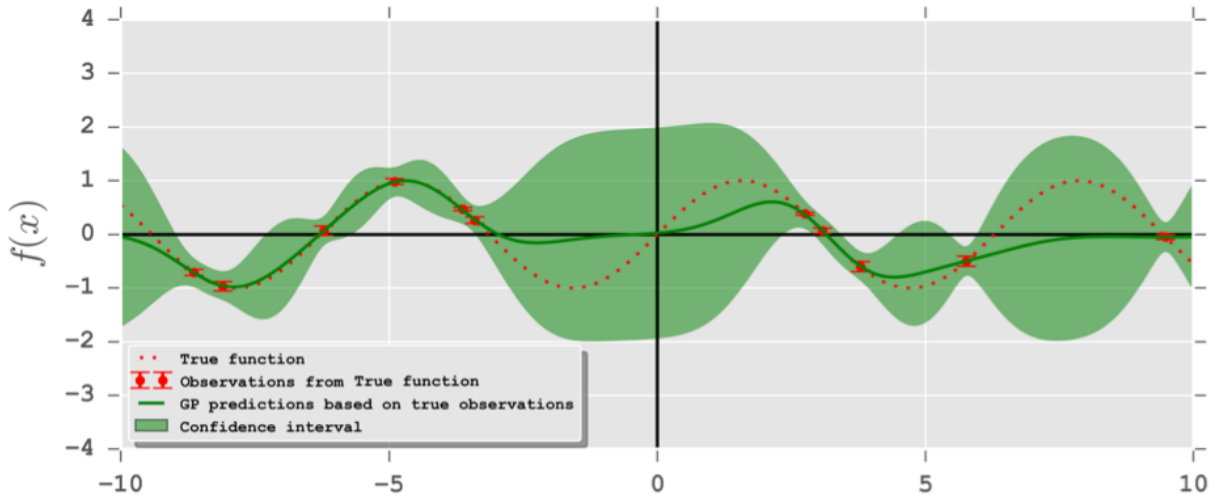
Weighted Fidelity Learning

True/HF: $y = \sin(x)$
Weak/LF: $y = 2\ sinc(x)$

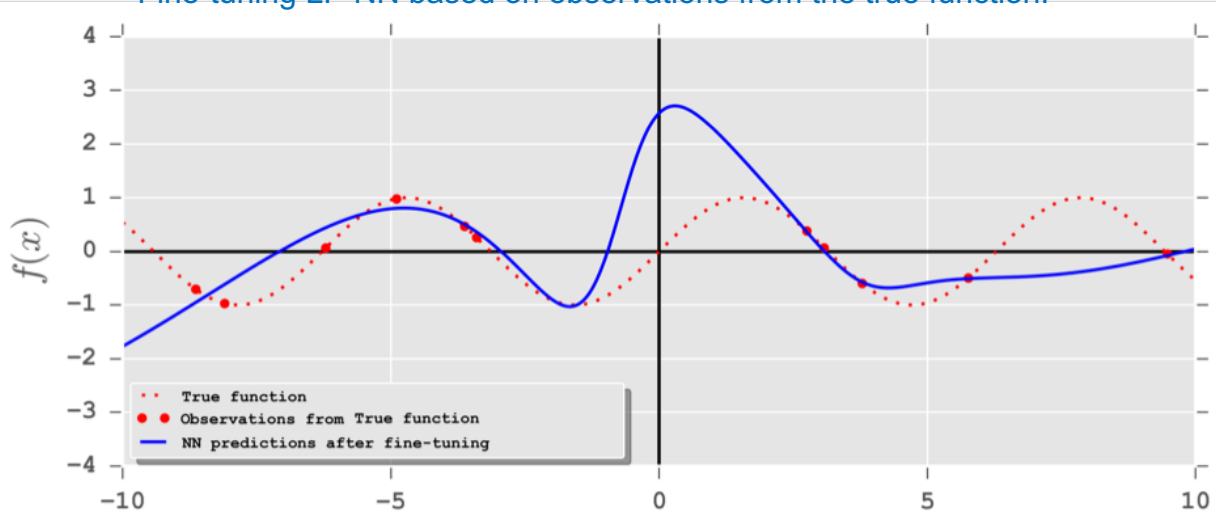
Training LF-NN on 100 examples from the weak function.



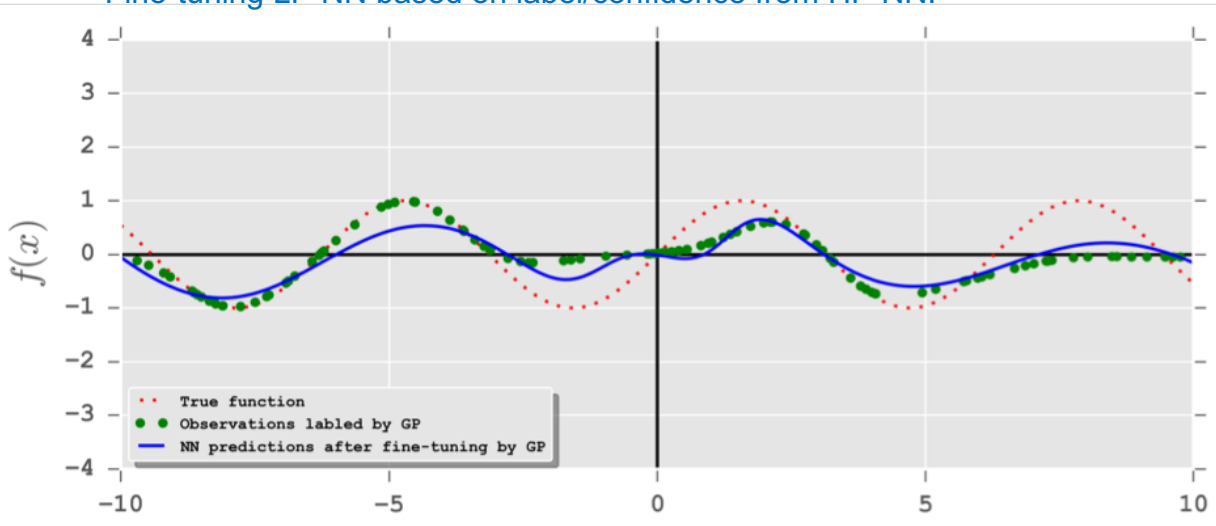
Fitting HF-NN based on 10 observations from the true function.



Fine-tuning LF-NN based on observations from the true function.

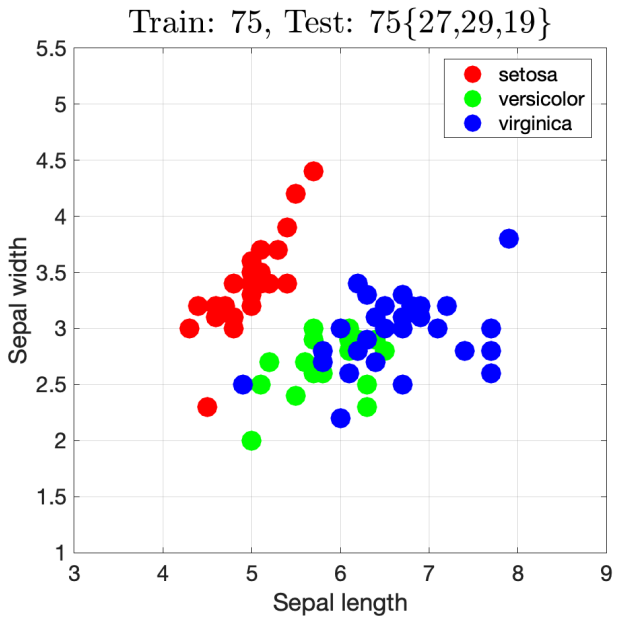


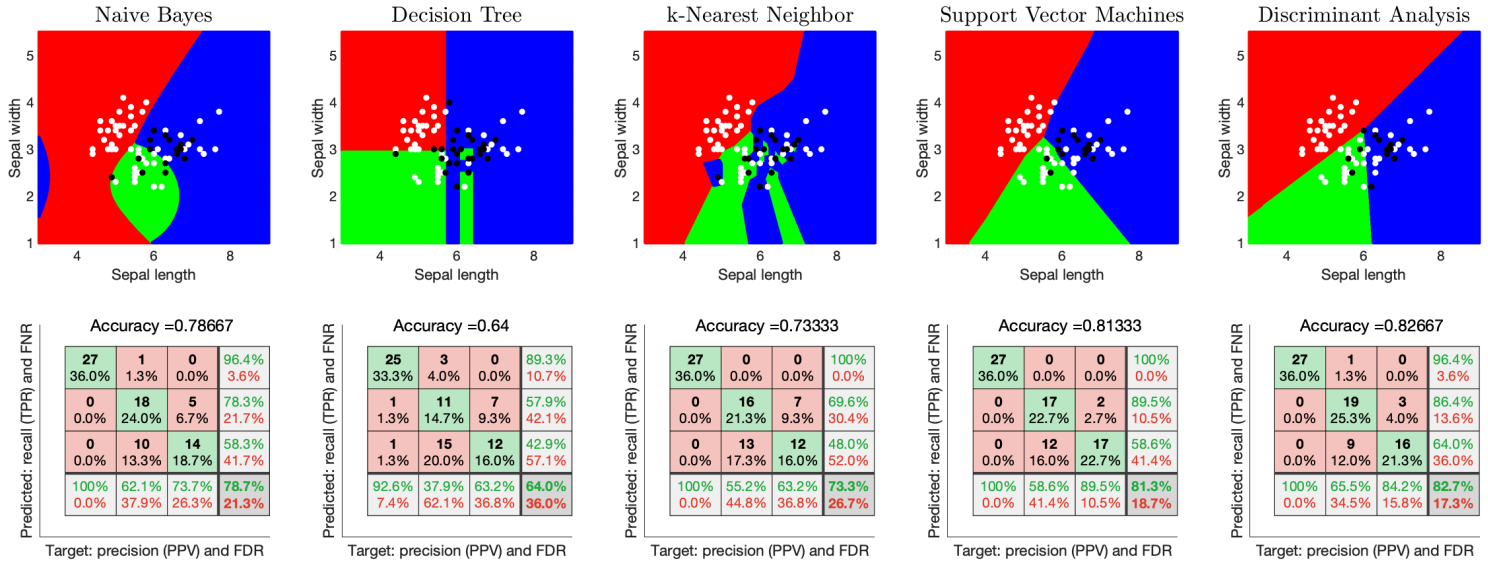
Fine-tuning LF-NN based on label/confidence from HF-NN.

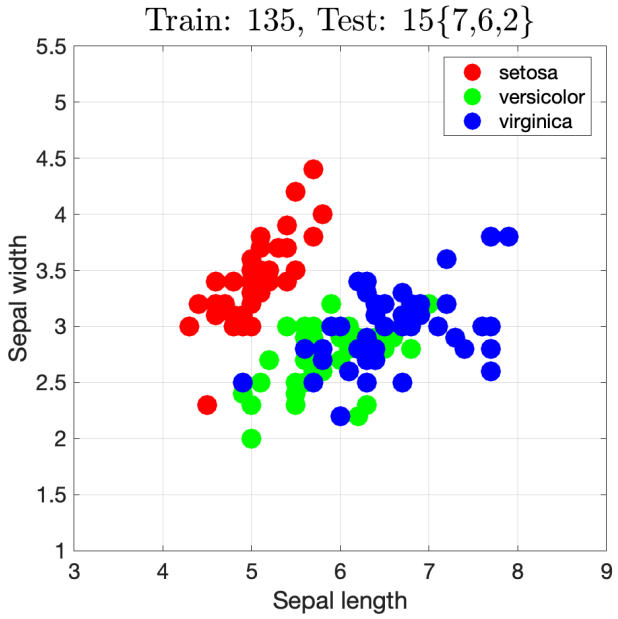


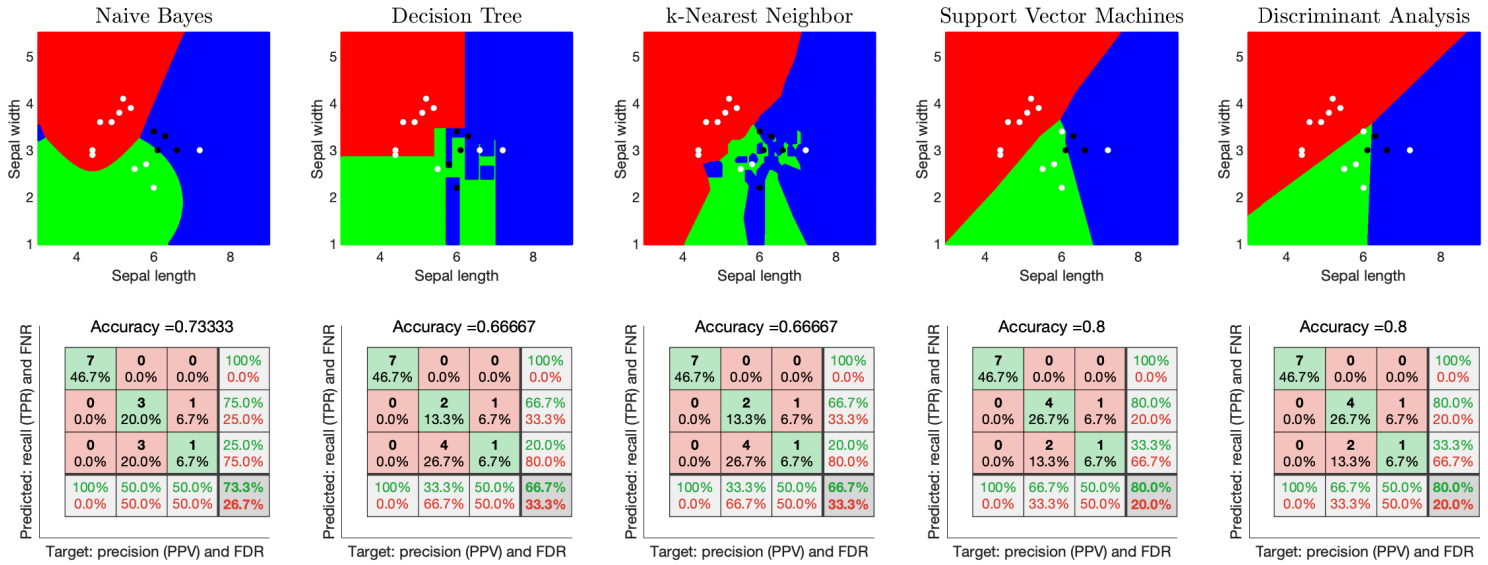
Dehghani, Mostafa, et al. "Fidelity-weighted learning." *arXiv preprint arXiv:1711.02799* (2017).

Numerical Experiments – IRIS Classification

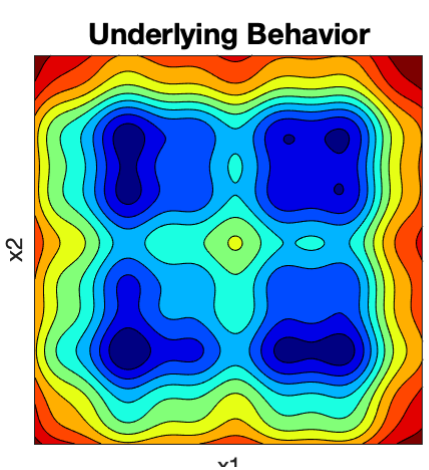


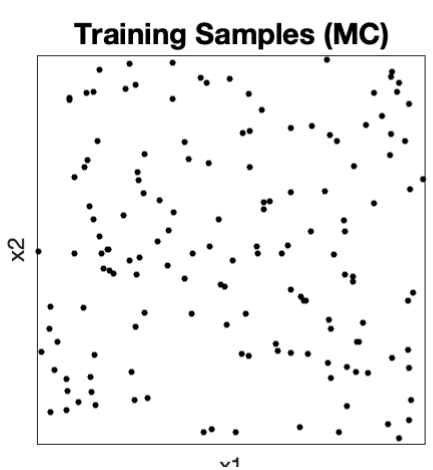


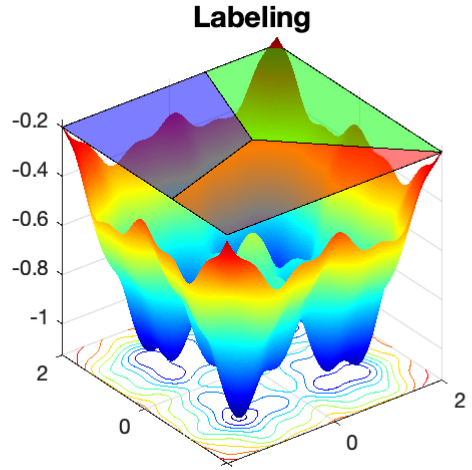


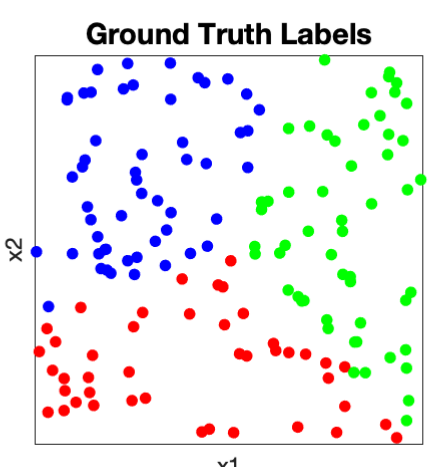


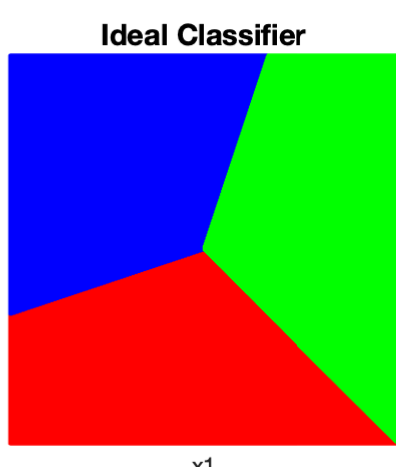
Numerical Experiments – Classification on $\mathbf{x}$

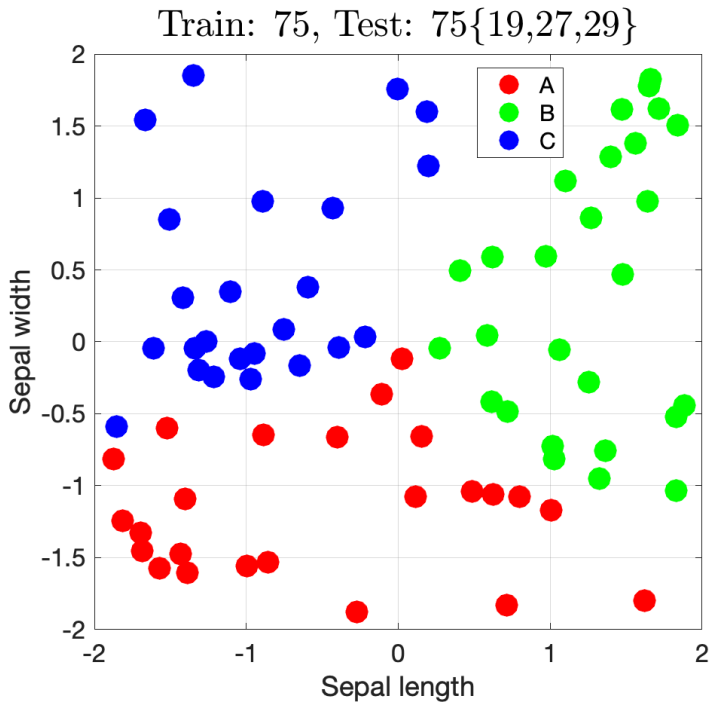


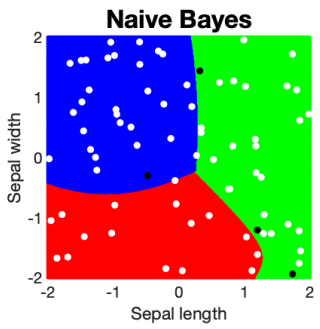


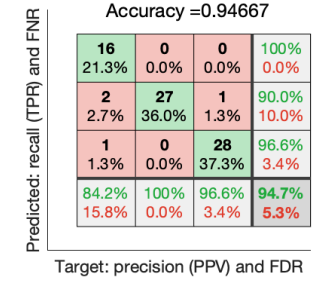


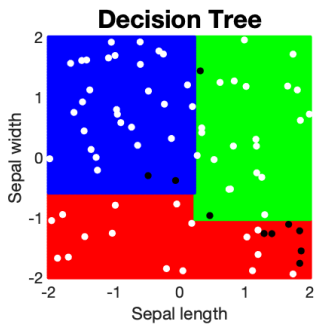


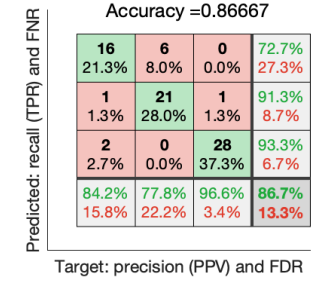


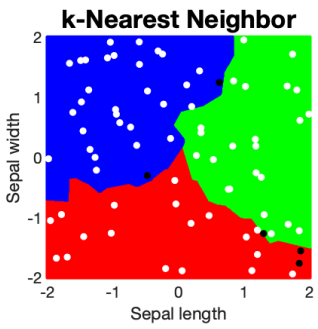


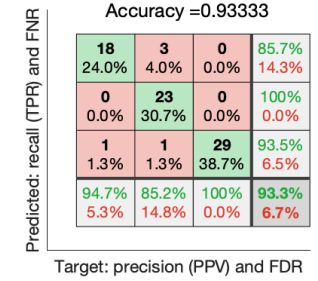


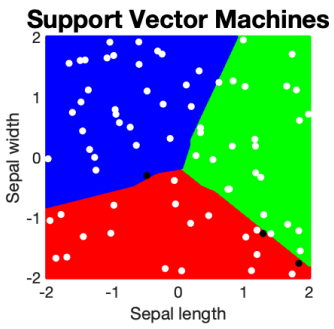


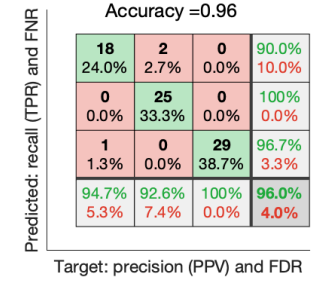


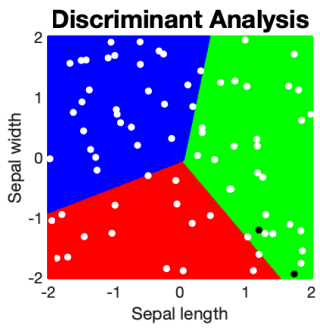


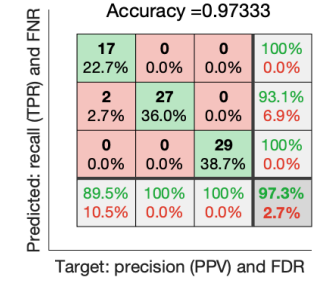




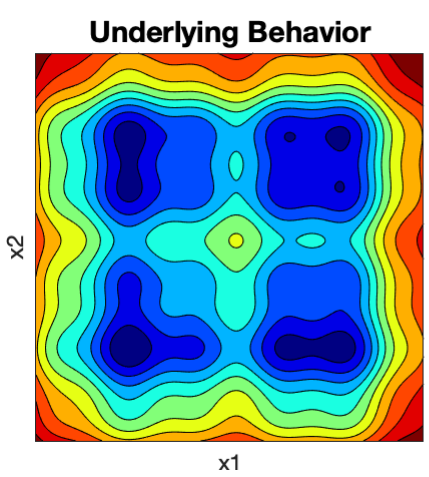


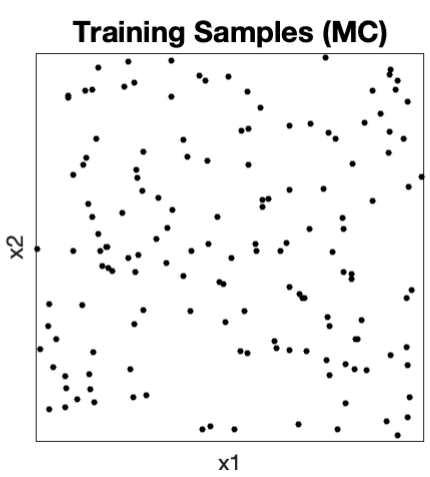


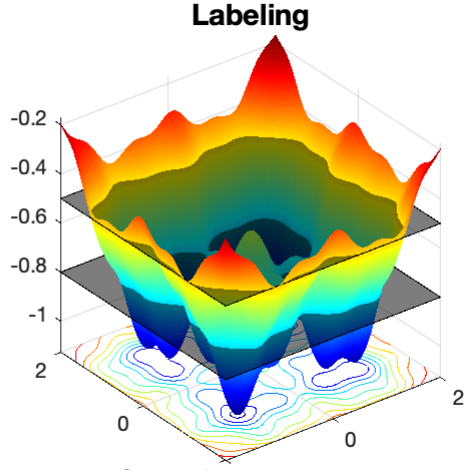


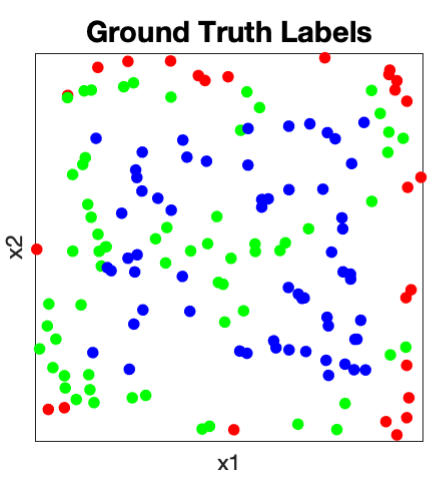


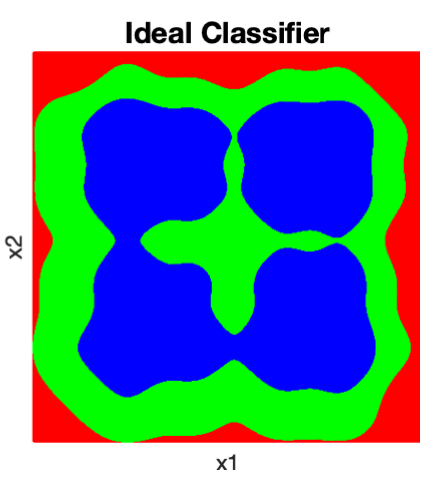
Numerical Experiments – Classification on $f(\mathbf{x})$

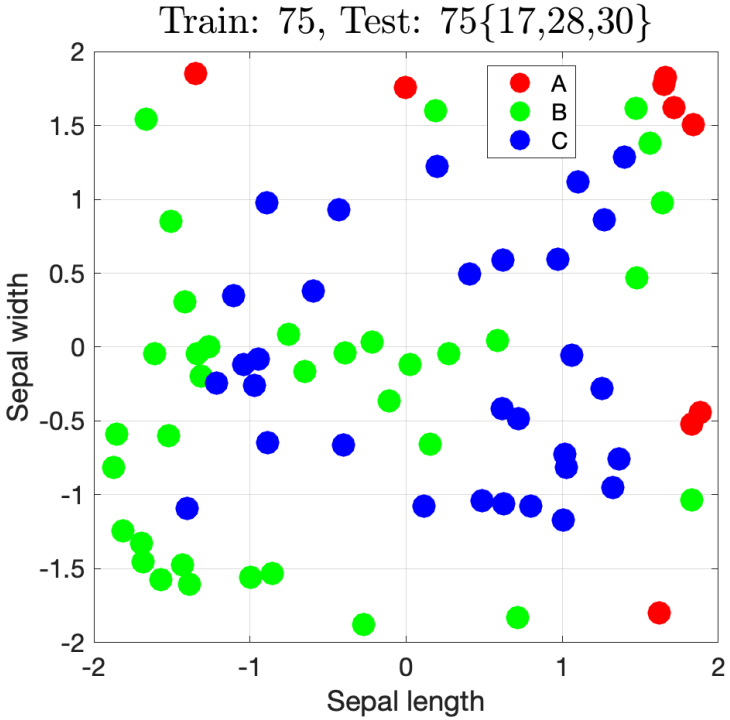


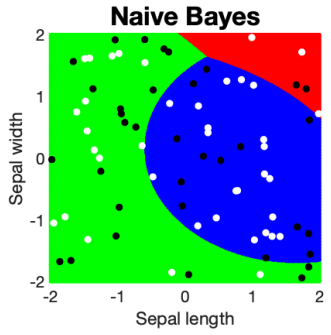


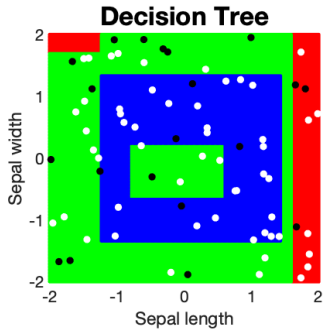


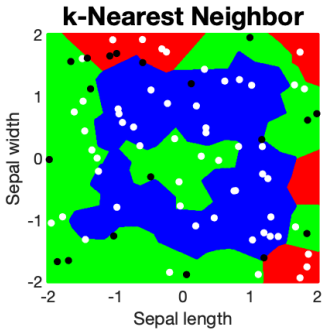


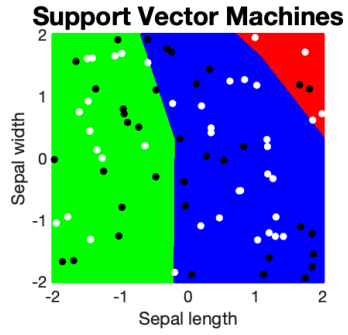


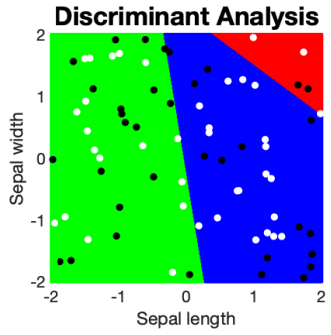


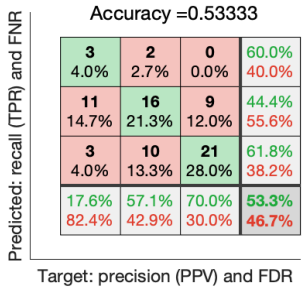


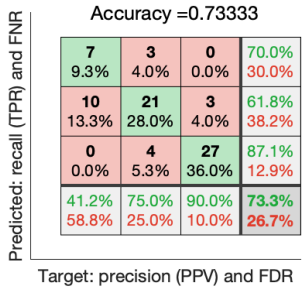


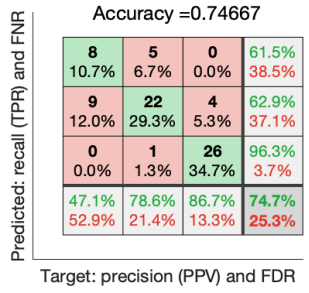


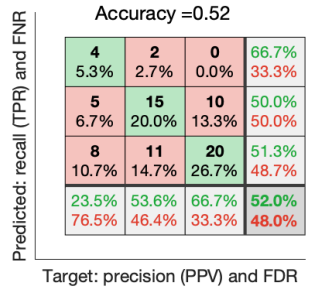


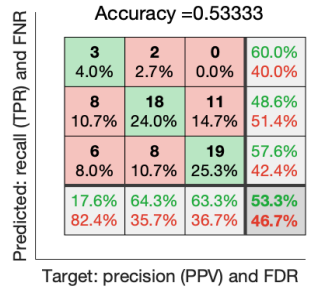












Observations & Conclusions

Ongoing work

- No one solution fits all. Classes of functions vary in terms of smoothness, oscillation, discontinuities ..
- Methods vary in terms of complexity, accuracy, scalability, and computational cost.

Preliminary Results

- Deep ensembles seem to offer multiple advantages for regression problems, including i) smoothed out performance by adversarial training, ii) conservative estimates, and iii) low training cost (with M=5 models)
- Comparisons to hybrid Bayesian/non-Bayesian methods

Thank you!