SAND2019-9206PE

MLDL

Machine Learning and Deep Learning Conference 2019

Fusing Machine Learning and Uncertainty Quantification Methods

A Bayesian Approach to Neural Networks

Ahmad A. Rushdi (1463), *Optimization and Uncertainty Quantification*

Collaborators: Laura P. Swiler (1463), Aubrey C. Eckert (1544), Gabriel Huerta (9136), Brian Freno (1544)

Funding Source: ASC AML, Dakota

Abstract

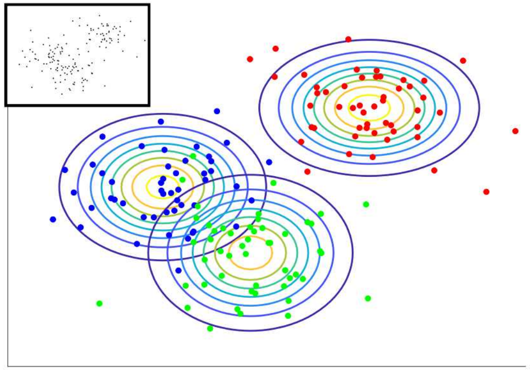
Machine and deep learning methods have evolved as core technologies in many scientific and engineering problems (e.g., image analysis). However, they lack robust methods to quantify and control their uncertainties due to overfitting among other technical challenges.

In this talk, we introduce a Bayesian approach to neural networks that creates probabilistic ML predictions using Variational Inference approximations, in contrast to classical softmax point estimates.

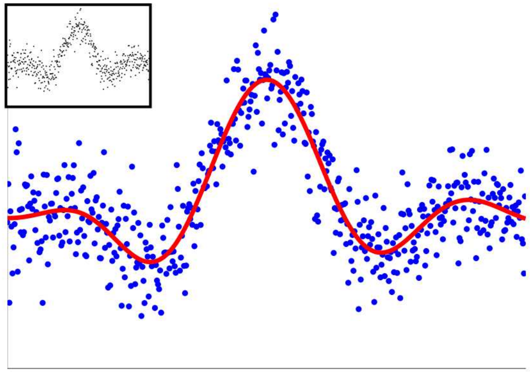
Background

⇞ Machine Learning (ML)

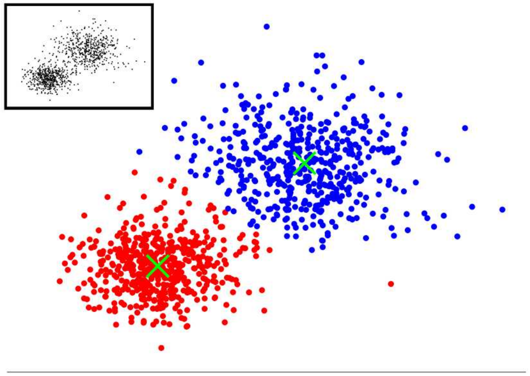
is a set of methods to solve **classification**, **regression** and **clustering** problems, enabling fast analysis of massive data volumes. Advancing AI tech., ML is empowered by efficient **algorithms**, training **data**, and powerful **computing** tools.



Naïve Bayes classification



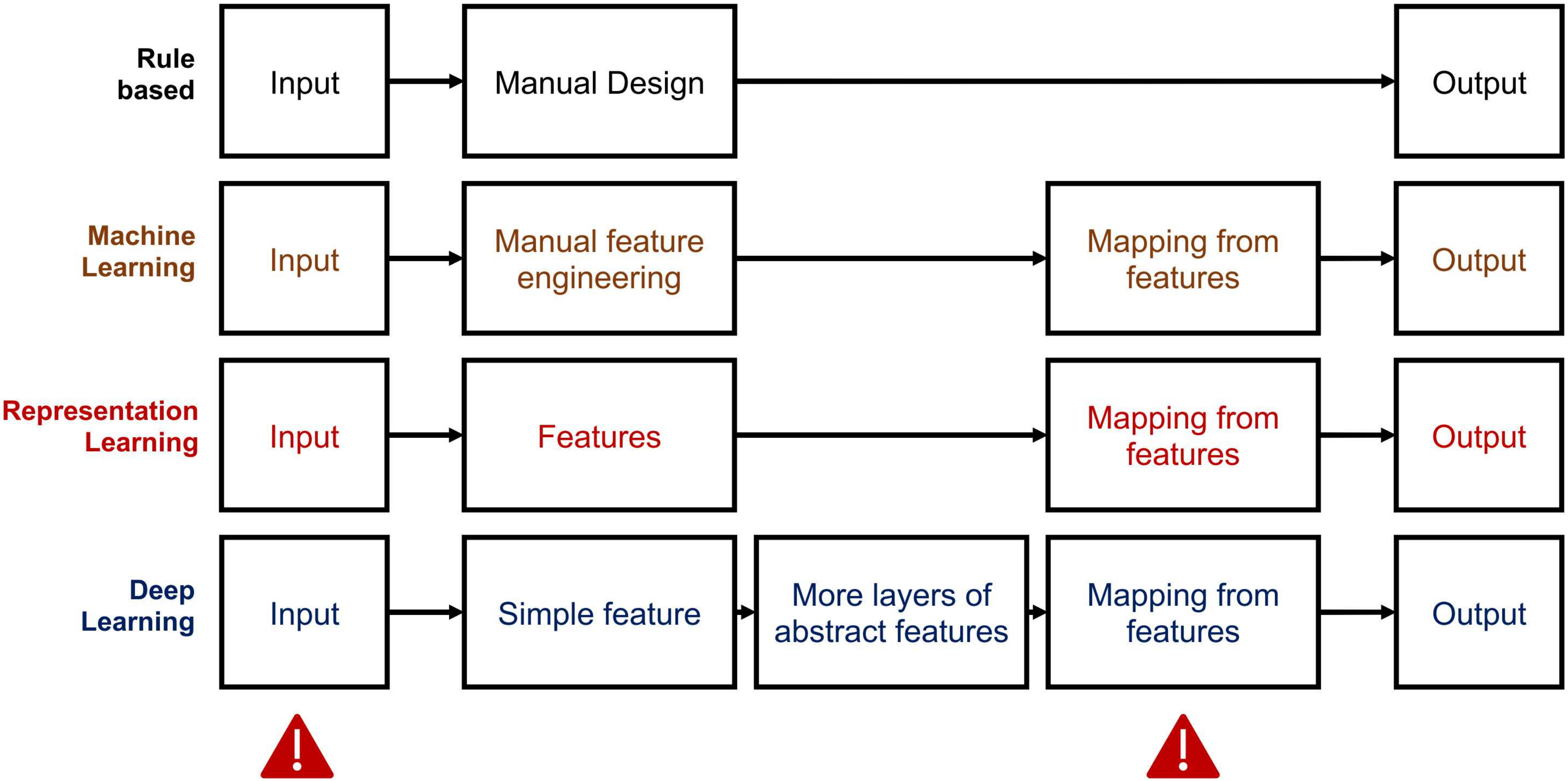
Gaussian process regression



k-means clustering

Background

⤉ Machine/Deep Learning Generic Workflow

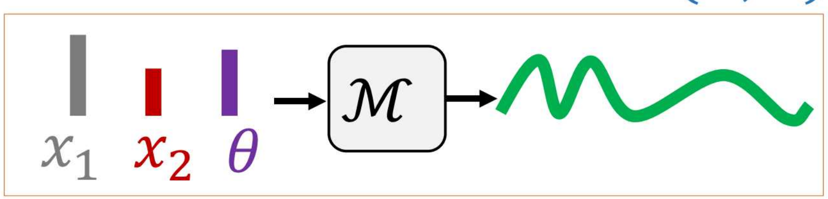


Background

⇞ Uncertainty Quantification (UQ)

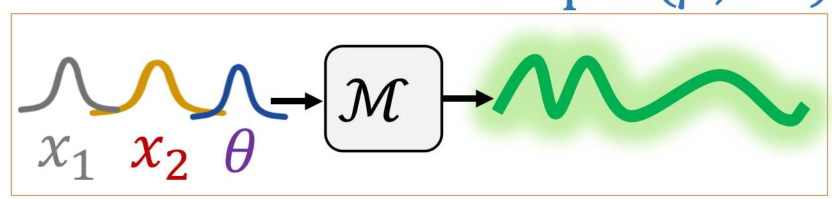
- Characterization of **aleatoric/epistemic** uncertainties
- Rigorous **statistical, V&V,** and **estimation** procedures
- Sensitivity analysis, Bayesian calibration, dimensionality reduction, surrogate modeling, etc.

$Y = \mathcal{M}(X, \theta)$



Deterministic model

$Y \sim \mathrm{pdf}(\mu, \sigma^2)$

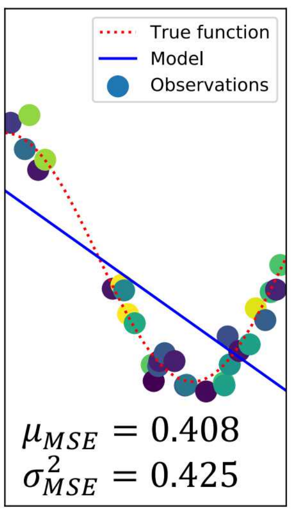


Model with uncertain inputs/parameters

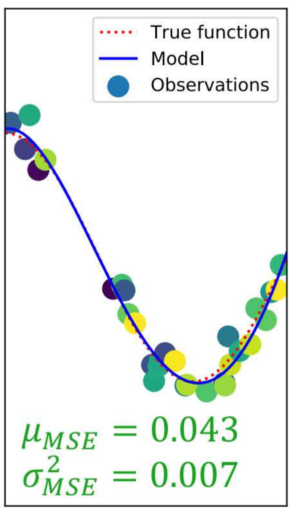
Problem Statement

⇞ ML/UQ Technical Challenges

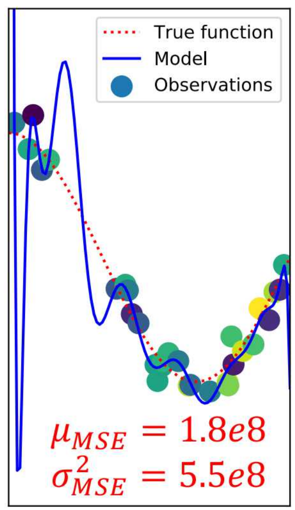
- **Data:** Limited obs./exp., costly sims, noise, errors, gaps
- **Models:** Unknown parameters and boundary conditions
- **Curses:** Dimensionality, nonlinearity, multi-physics



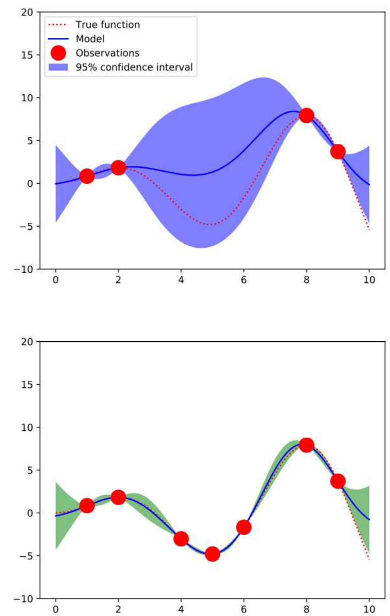
Poly deg 1 underfitting



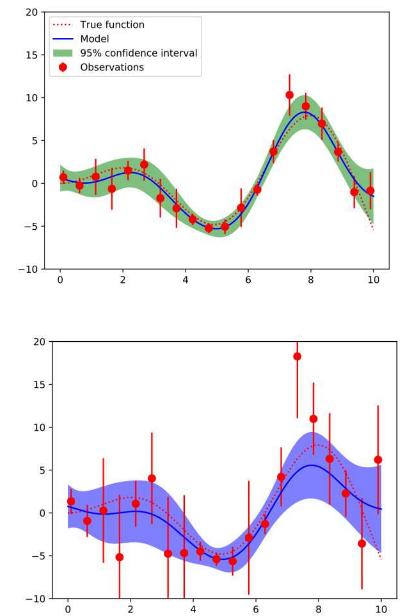
Poly deg 4 fitting



Poly deg 15 overfitting



Sampling highest uncertainty regions



Regression with UQ ranges, noisy data

Significance

⇞ Informed Decision Making

- High-stakes and cost-sensitive national security applications
- Unintended and harmful AI behavior [Amodei 2016]
- Evaluate trust/risk and iteratively improve the ML model **reliability**
- Numerical sims of physical systems are rife with uncertainties
- Improve quality of **learning** in reinforcement sequential choices

“Predictions without UQ are neither predictions nor actionable.”

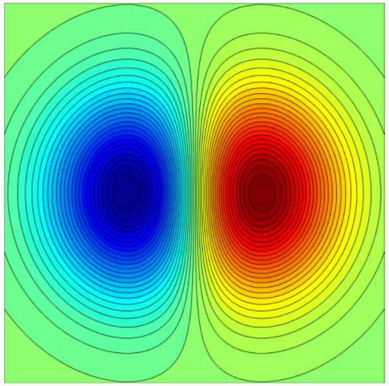
Begoli E, bhattacharya T, kusnezov D., Nature Machine Intelligence 2019

Variational inference (VI) approximations to upgrade machine learning predictions

from point-estimates to approximate probability distributions

Significance

‡ Sandia Mission-Relevant Applications



NUMERICAL SIMULATIONS
Turbulence Modeling
Subsurface Modeling
Hypersonics

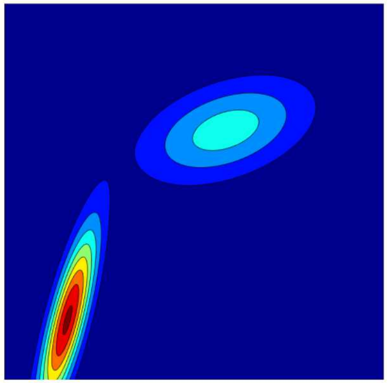
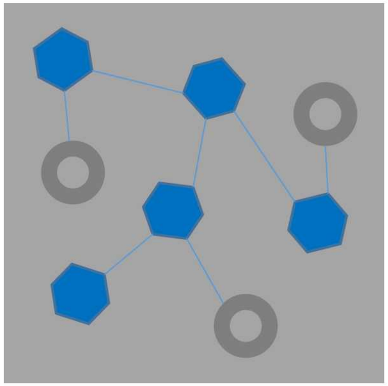
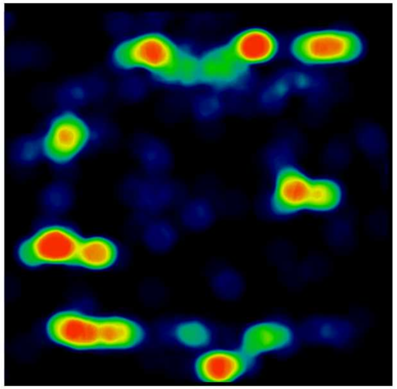


IMAGE PROCESSING
Additive Manufacturing
Radar/SAR/X-ray Imagery
Image Segmentation



SENSOR NETWORKS
Graph Neural Networks
Rare Event Prediction
Genomic Analysis



HUMAN FACTOR
Performance Evaluation
Eye Tracking
Cognitive Load Estimation

Solution Approach

⇞ NN Parametric Model

[Murphy 2012, Bishop 2006]

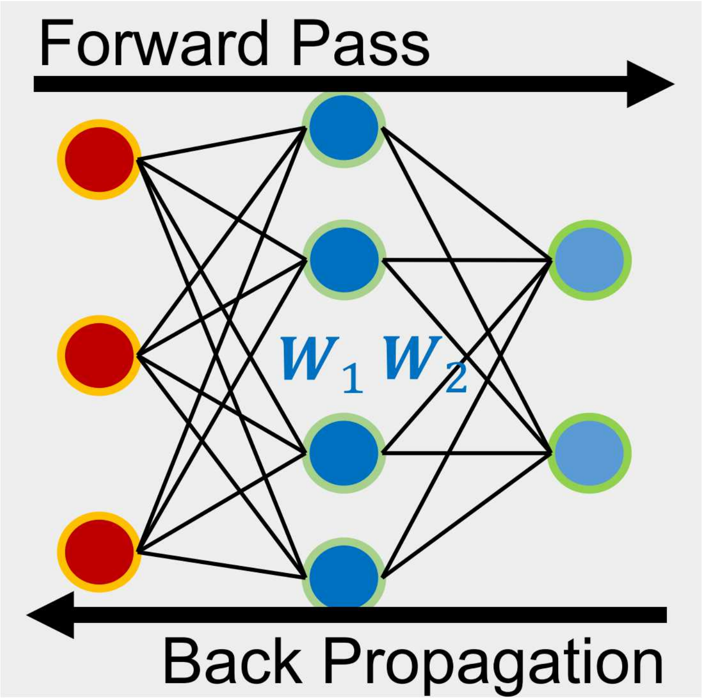
$$y = f_2\big(W_2\big(f_1(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1)\big) + \boldsymbol{b}_2\big)$$

Training data $\mathcal{D} = \{\boldsymbol{X}, \boldsymbol{Y}\}_{i=1}^{N}$

Observations $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^{N} \in \mathbb{R}^d$

Hidden layer(s) $\boldsymbol{W} = [\boldsymbol{W}_1 \quad \cdots \quad \boldsymbol{W}_L]$

Output variables $\boldsymbol{Y} = \{\boldsymbol{y}_i\}_{i=1}^{N} \in \mathbb{R}^k$



Solution Approach

⇞ Softmax for classification

$$\hat{p} = p(\hat{\boldsymbol{y}} = k|\hat{\boldsymbol{x}}, \omega) = e^{f(\hat{\boldsymbol{x}},\omega_k)} / \sum_j e^{f(\hat{\boldsymbol{x}},\omega_j)}$$

⇝ Max Likelihood **single-estimate** minimizes cross-entropy between the true class distribution and the softmax output.

y	$f^{\omega}(x)$
a	5
b	4
c	2

⇒

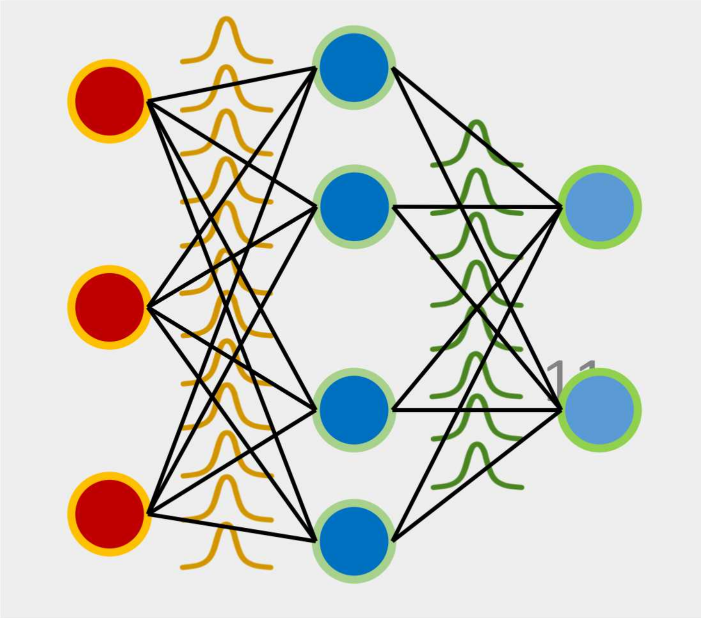
$p(y \vert f^{\omega}(x))$
0.71
0.26
0.04

⇝ class a

Solution Approach

⇞ Bayesian Neural Networks [Polson 2017, Ghahramani 2016. Seeger 2009]

To span the parameter space of θ



Prior

$$\omega \sim \mathcal{N}(0, I)$$

Posterior

$$p(\omega|\boldsymbol{X},\boldsymbol{Y}) = \frac{p(\boldsymbol{Y}|\boldsymbol{X},\omega)p(\omega)}{p(\boldsymbol{Y}|\boldsymbol{X})}$$

$$p(\hat{\boldsymbol{y}}|\hat{\boldsymbol{x}},\boldsymbol{X},\boldsymbol{Y}) = \int p(\hat{\boldsymbol{y}}|\hat{\boldsymbol{x}},\omega)p(\omega|\boldsymbol{X},\boldsymbol{Y})d\theta$$ ⇝ Intractable integration

Variational Inference
Hamiltonian Monte Carlo (HMC) [Neal 2012], Laplace approximation [MacKay 1992], Dropout [Ghahramani 2016], Bayes by Backprop [Blundell 2015].

Solution Approach

⤉ Variational Inference

[Kwon 2018, Shridhar 2019, Gal 2017]

Approximation using a variational distribution $q_\theta(\omega)$ which minimizes the Kullback-Leibler divergence

$$\mathcal{KL}_{VI} = \int p(\boldsymbol{Y}|\boldsymbol{X},\omega) q_\theta(\omega) d\omega - KL\big(q_\theta(\omega)||p(\omega|\boldsymbol{X},\boldsymbol{Y})\big)$$

$$\theta^* = \arg\min_\theta KL\big(q_\theta(\omega)||p(\omega|\boldsymbol{X},\boldsymbol{Y})\big)$$

⇝ From integration to optimization

We sample $\omega^{(i)}$ from $q_\theta(\omega|\boldsymbol{X},\boldsymbol{Y})$

⇝ variational method

$$\mathcal{KL}_{VI} \approx \sum_{i=1}^{N} \log q(\omega|\boldsymbol{X},\boldsymbol{Y}) - \log p\big(\omega^{(i)}\big) - \log p\big(\boldsymbol{X},\boldsymbol{Y}|\omega^{(i)}\big)$$

⇝ Tractable

The predictive distribution is

$$p(\hat{\boldsymbol{y}}|\hat{\boldsymbol{x}},\boldsymbol{X},\boldsymbol{Y}) \approx \int p(\hat{\boldsymbol{y}}|\hat{\boldsymbol{x}},\omega) q(\omega) d\omega$$

Solution Approach

⇞ Variational Inference Approximation

Mean $\mathbb{E}_q\{p(\hat{\boldsymbol{y}}|\hat{\boldsymbol{x}})\} \approx \frac{1}{T}\sum_{t=1}^{T} p(\hat{\boldsymbol{y}}|\hat{\boldsymbol{x}}, \theta_t^*)$

T: MC samples

$\bar{p} = \frac{1}{T}\sum_{t=1}^{T} \hat{p}_t$

Variance $\sigma_q^2\{p(\hat{\boldsymbol{y}}|\hat{\boldsymbol{x}})\} \approx$

$$\overbrace{\frac{1}{T}\sum_{t=1}^{T} \mathrm{diag}(\hat{p}_t) - \hat{p}_t\hat{p}_t^T}^{\rightsquigarrow \text{Aleatoric}} + \overbrace{\frac{1}{T}\sum_{t=1}^{T} (\hat{p}_t - \bar{p})(\hat{p}_t - \bar{p})^T}^{\rightsquigarrow \text{Epistemic}}$$

Aleatoric Uncertainty (irreducible)

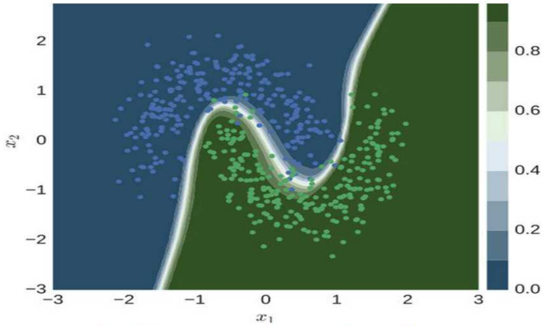
- persists even in the limit of infinite data, e.g., measurement noise.

Epistemic Uncertainty (reducible)

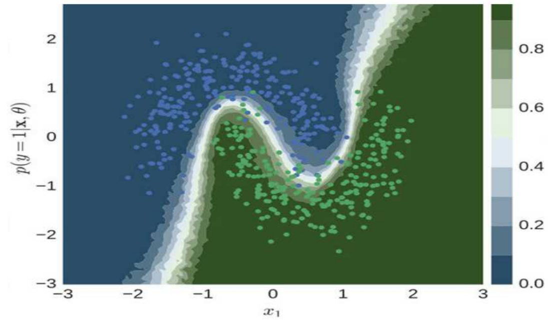
- Vanishes in the limit of infinite data, e.g., model parameters.

Preliminary Results

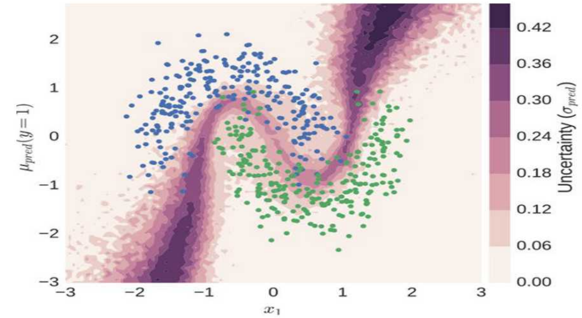
Toy problem – 2D binary classification



Softmax classification boundary with no uncertainty



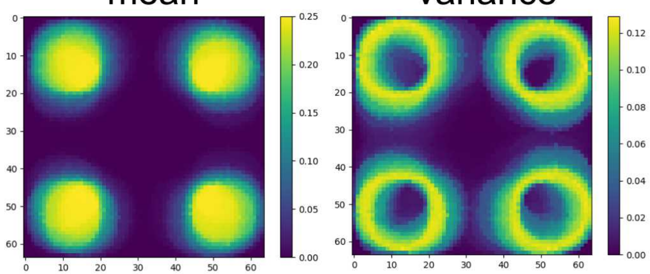
Mean of approximate predictive posterior using VI



Variance of approximate predictive posterior using VI

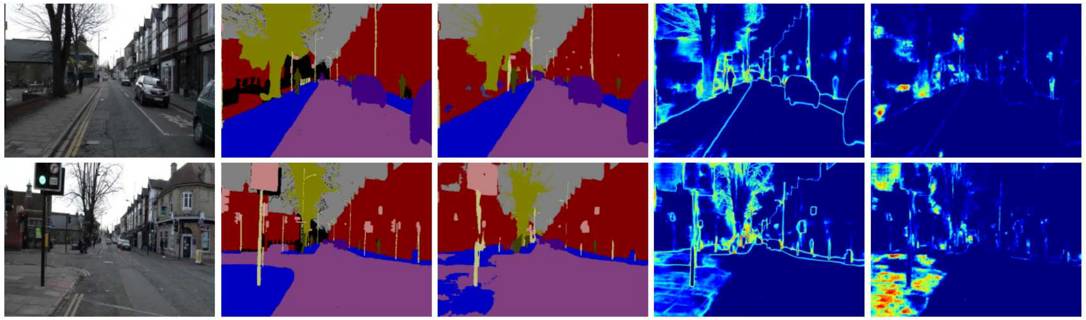
Heat Conduction

mean variance



[Oberai 2019]

Image Segmentation



(a) Input Image (b) Ground Truth (c) Semantic Segmentation (d) Aleatoric Uncertainty (e) Epistemic Uncertainty

[Kendall and Gal 2017]

Conclusions

- New Bayesian techniques to approximate the ML posteriors.
- ML predictions without UQ are *not actionable*.

Future Work:

- Quality, computational complexity, and scalability
  - e.g., stochastic Variational Inference [Hoffman 2013].
- Non-Bayesian approaches
  - e.g., deep ensembles [Balaji L. 2017].

Thank You!