Sandia National Laboratories

SAND2020-1804PE

A Gentle Introduction to Bayesian Neural Networks

PRESENTED BY

Daniel Ries

Statistical Sciences Department

Sandia National Laboratories

oratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, neywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract

tional Nuclear Security Administration under contract DE-NA-0003525. SAND NO.

Outline

1. Simple introduction to BNN
2. Simple introduction to variational inference
3. Hyperspectral image target dection example
4. Simulation study assessing BNN performance

BNN Overview

- Traditional neural networks (NN) give no uncertainty quantification (UQ)
- Bayesian neural networks (BNN) bring UQ to deep learning

BNN Use Example

BNN are useful when automated systems are tasked with high risk decision making

When used for target detection using imagery, BNNs can:

- Reduce false alarm rate
- Provide heatmap with degree of uncertainty

An Example BNN

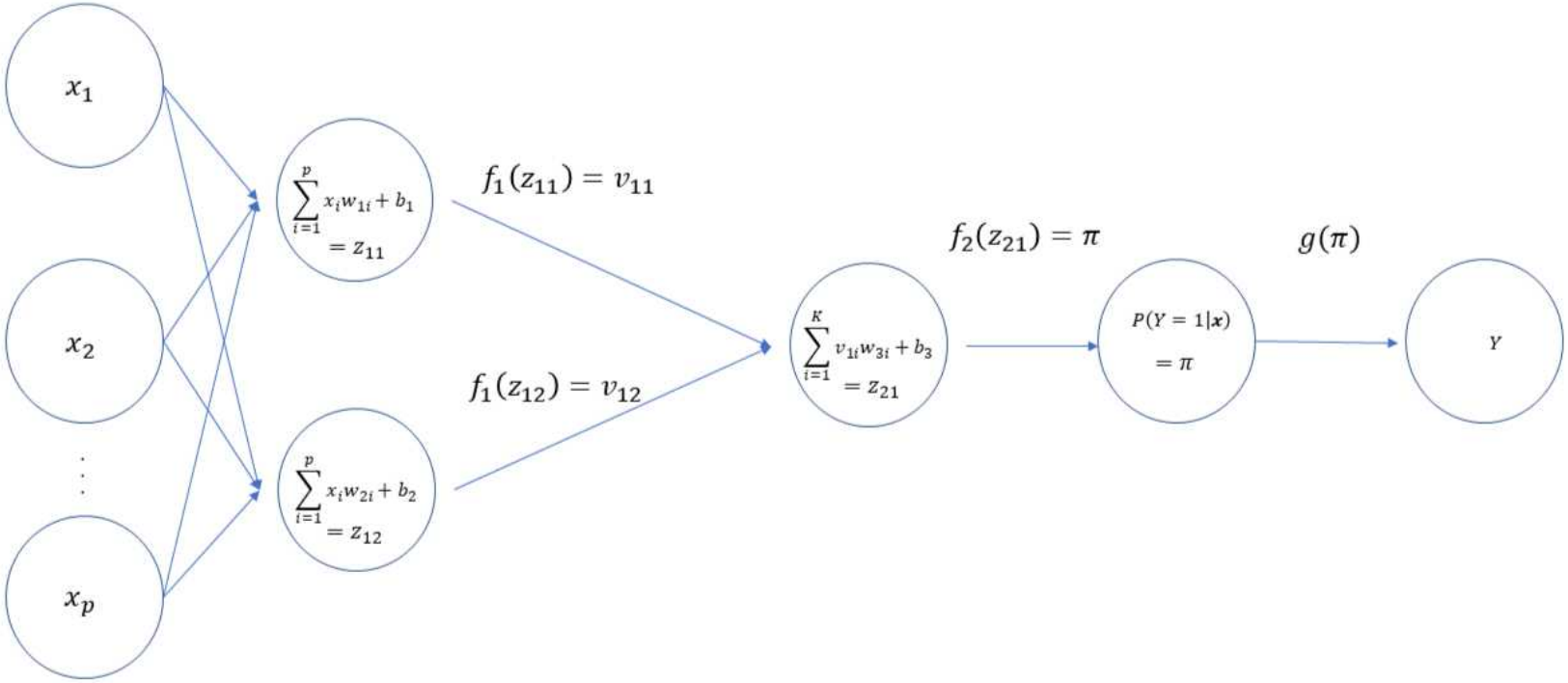
$$Y_i \sim \text{Bernoulli}(\pi_i)$$

$$\pi_i = f_2(z_{21})$$

$$= f_2\left(\sum_{j=1}^{2} f_1\left(\sum_{k=1}^{p} x_k w_{jk} + b_j\right) w_{2j} + b_3\right)$$

$$w_{jk} \overset{iid}{\sim} N(0,1), \forall j,k$$

$$b_l \overset{iid}{\sim} N(0,1), l = 1,2,3$$



Estimating (or Training) BNNs

Bayesians love to sample their posterior distributions via MCMC
But this can be slowwww

Enter **Variational Inference**:

- Instead approximate posterior with $q^*(\mathbf{w})$ by solving:

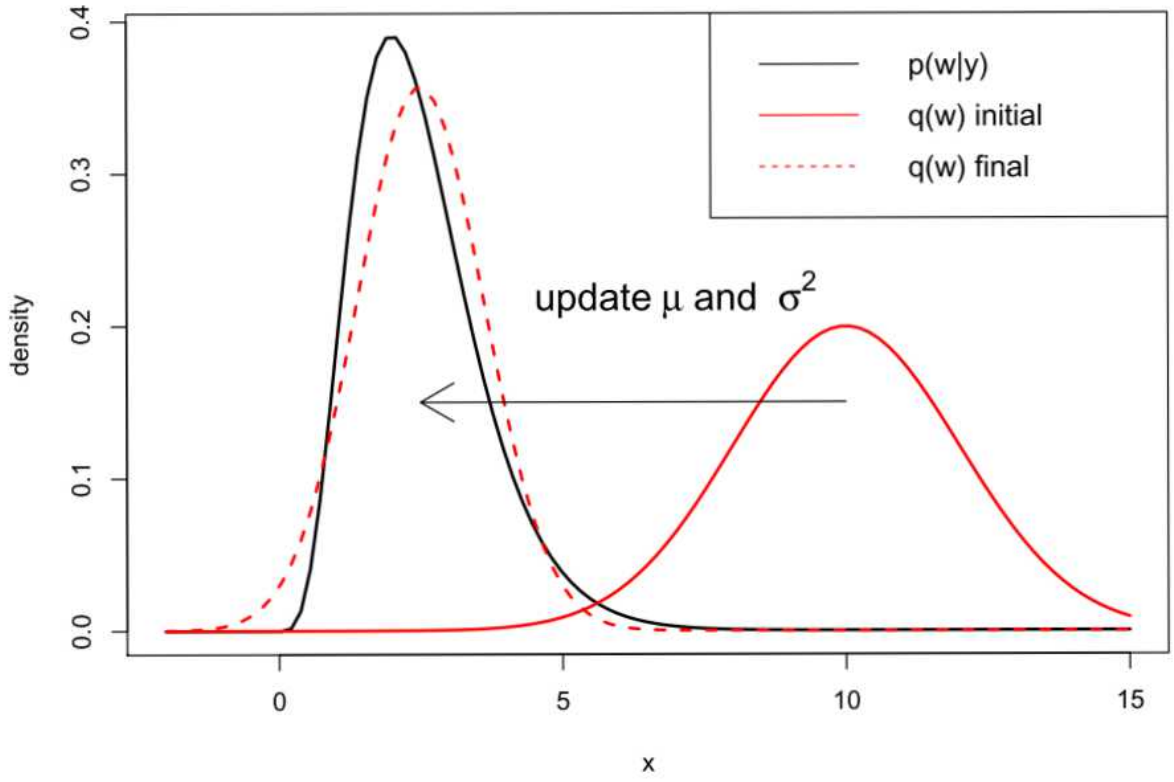
$$\underset{q^*}{\operatorname{argmin}}\, KL(q^*(\mathbf{w})||p(\mathbf{w}, \mathbf{b}|y))$$

$$= \underset{q^*}{\operatorname{argmin}}\, KL(q^*(\mathbf{w})||p(\mathbf{w}, \mathbf{b}, y))$$

- Restrict set of q^*, to family $q_\theta(\cdot)$
- Mean-field VI: $q_\theta(\mathbf{w}) = \prod_{i=1}^{K} q_\theta(w_i)$

Goal of VI

Suppose $q_\theta(w) = N(w; \mu, \sigma^2)$



$$q_\theta(w) = \underset{\theta^*}{\operatorname{argmin}}\, KL(q^*_\theta(w)||p(w|y)) = \ldots \propto -E_{q^*_\theta} log\left(\frac{p(w,y)}{q^*_\theta(w)}\right)$$

Stochastic VI (SVI) in Practice

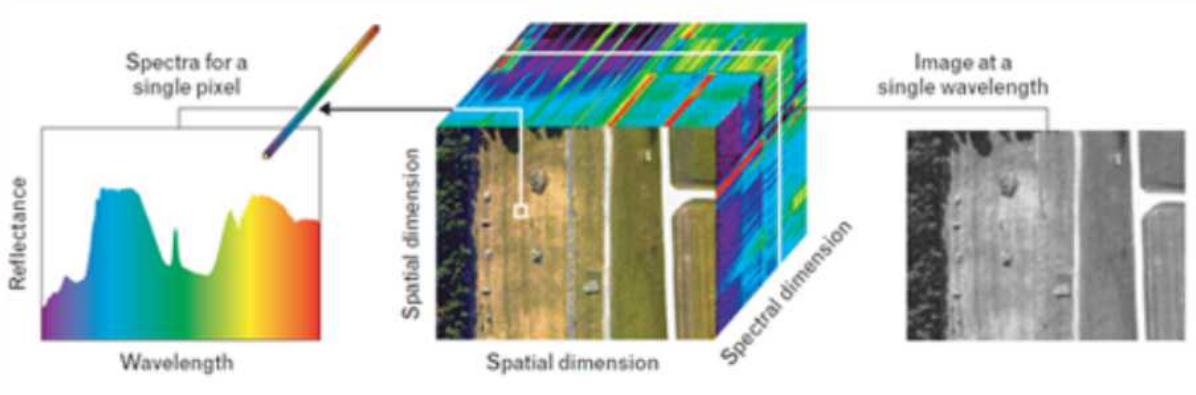
SVI uses backpropagation and looks like SGD

Suppose we use the mean-field assumption with Normal variational distributions, then we are optimizing the values of μ and σ^2 for each weight. From Blundell (2015):

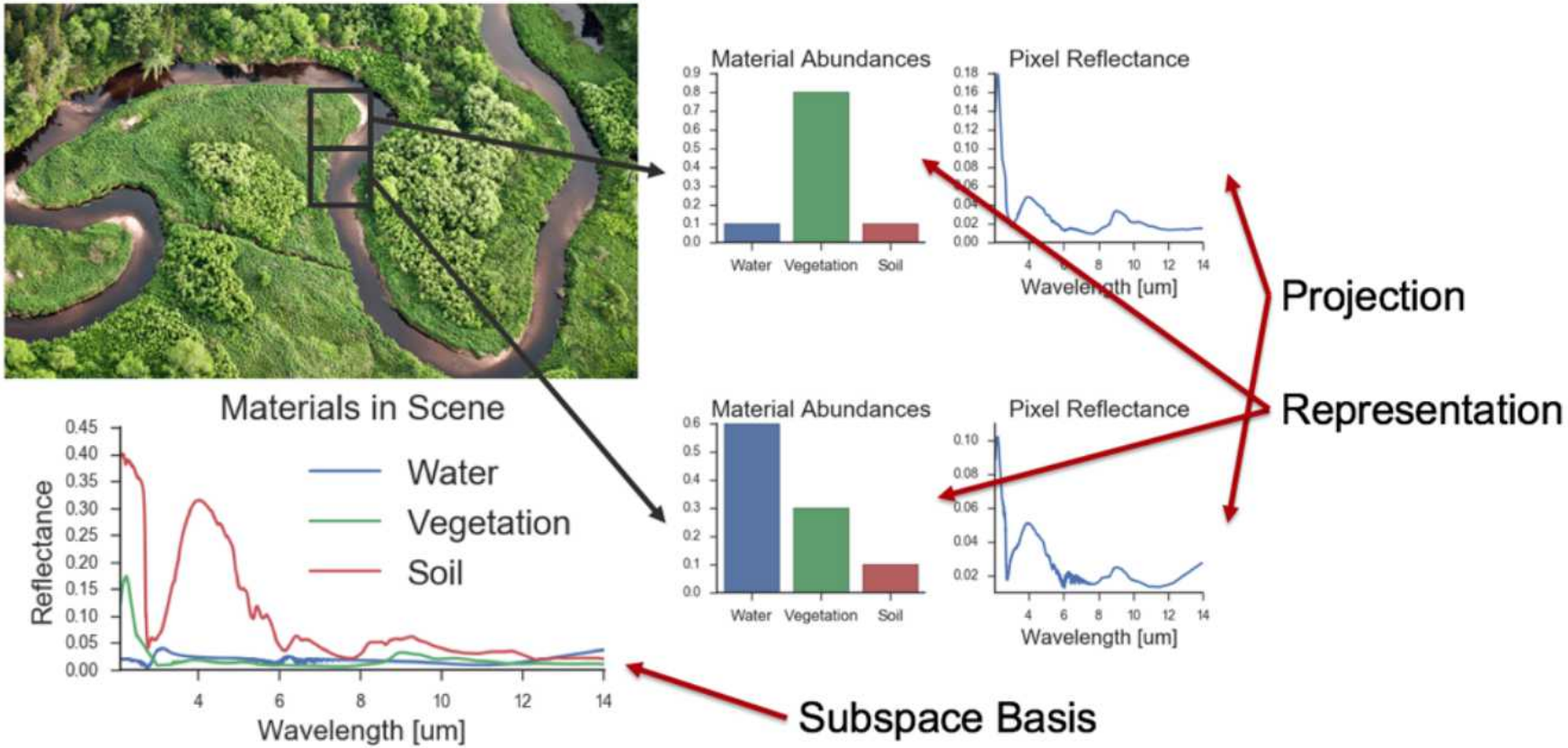
1. Sample $\epsilon \sim N(0, I)$
2. $w = \mu + \sigma\epsilon$
3. $\theta = (\mu, \sigma^2)$
4. Let $f(w, \theta) = \log q_\theta(w) - \log p(w, y)$
5. Calculate gradient of $f(w, \theta)$ wrt to μ
6. Calculate gradient of $f(w, \theta)$ wrt to σ^2
7. Update $\mu = \mu - \alpha\nabla_\mu, \sigma^2 = \sigma^2 - \alpha\nabla_{\sigma^2}$

Hyperspectral Images (HSI)

- Airborne spectrometers construct a (x, y, z) tensor
  - x and y describe the spatial dimension
  - z describes the spectrum at a single pixel (x, y)
- Specific materials are identified by their reflectance spectrum
- The target object might be smaller than the projection of a pixel on the ground



Mixing in HSI



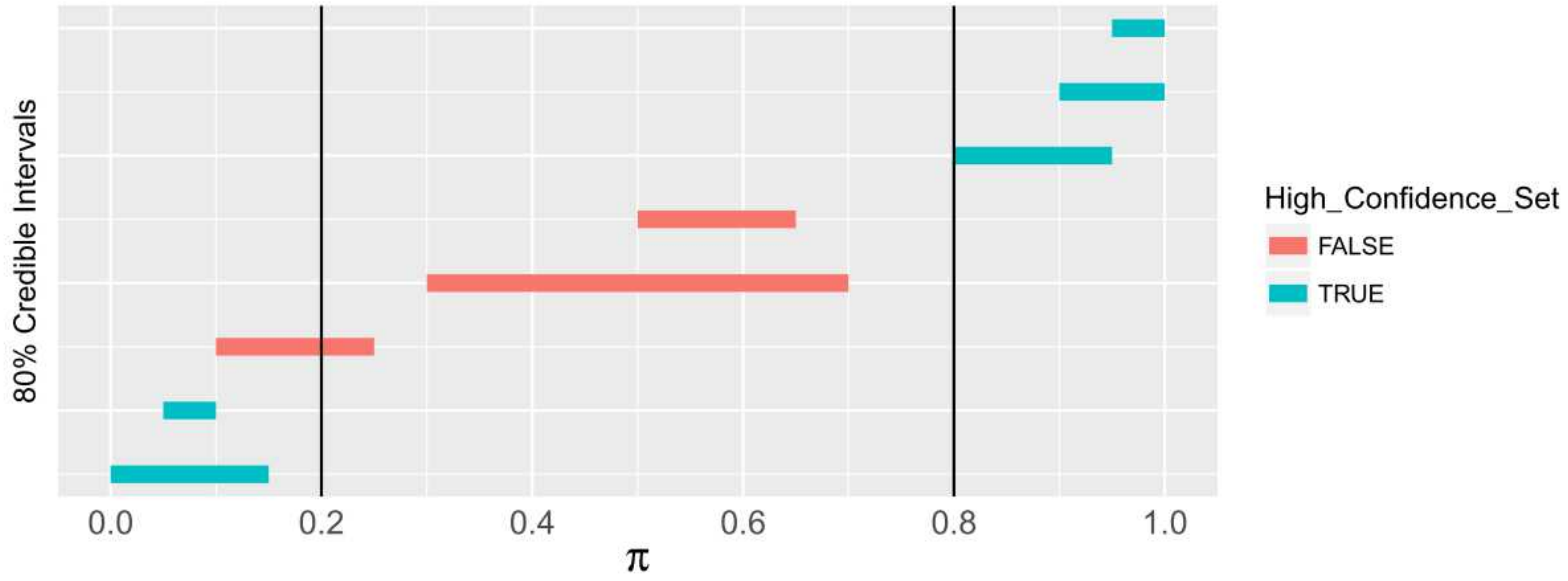
Megascene

Leveraging UQ When Assessing Performance

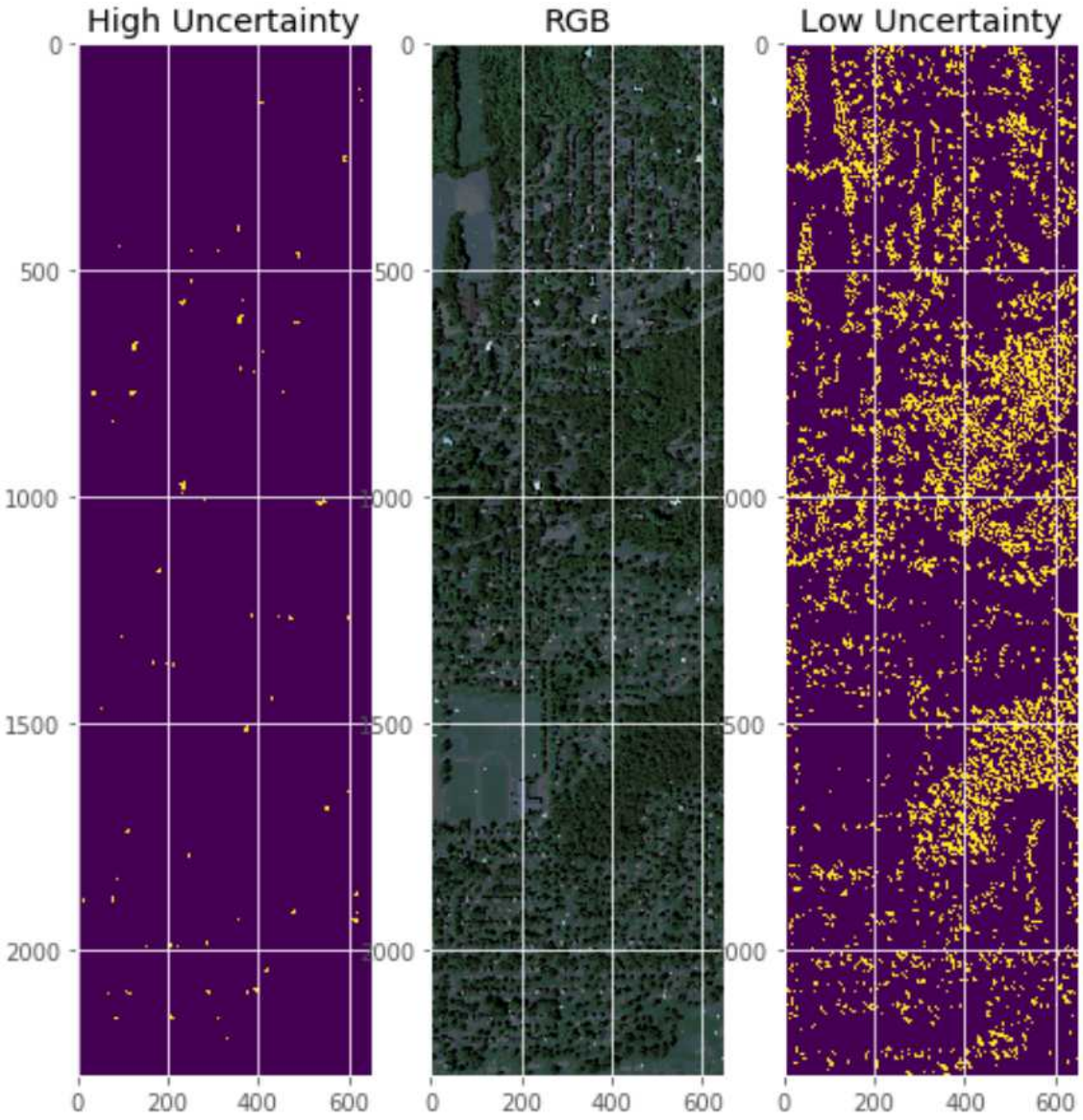
We will evaluate test set performance in two cases:

1. Using the full test set
2. Using a "high confidence" set

The "high confidence" set contains all pixels such that their 80% credible interval for $\pi \in (0, 0.2) \cup (0.8, 1)$

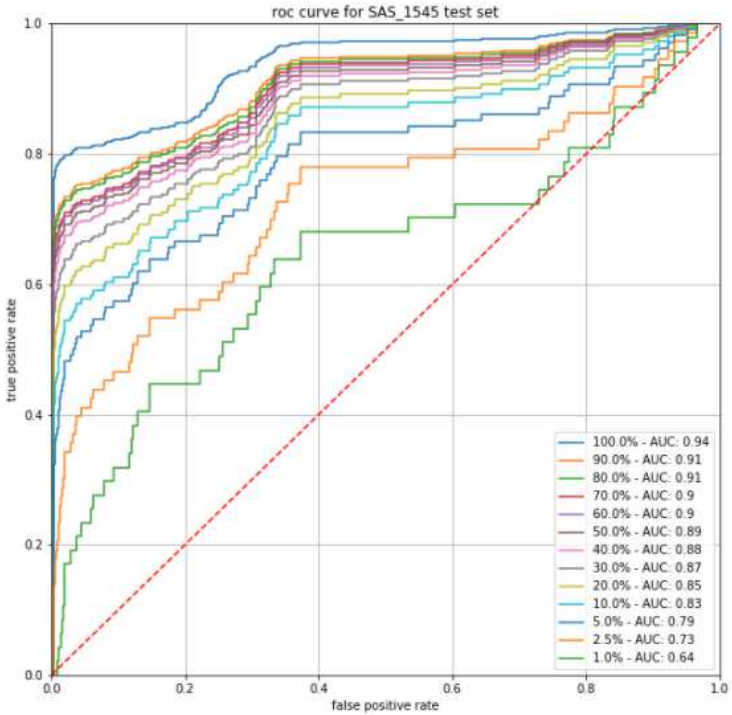


Test Set Performance

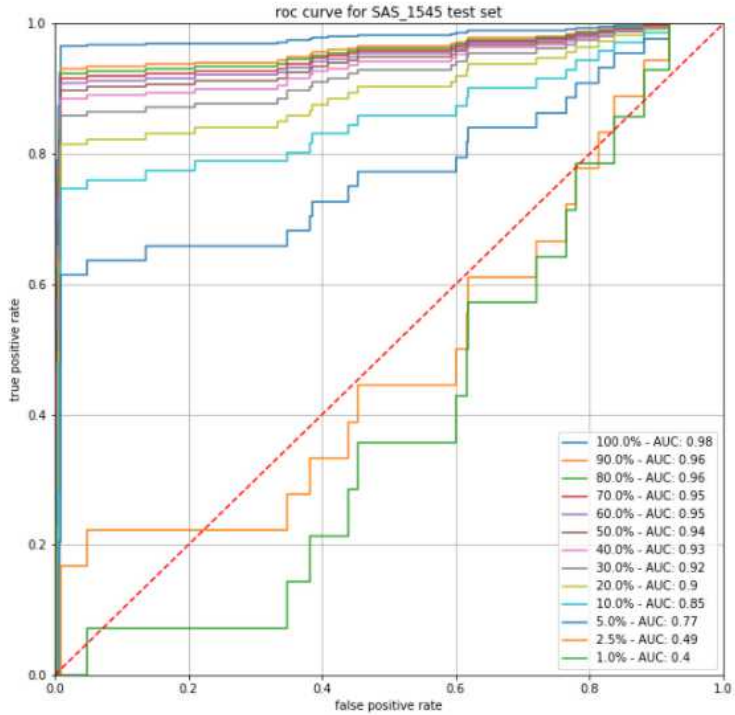


Test Set Performance

Full test set.



High confidence set.



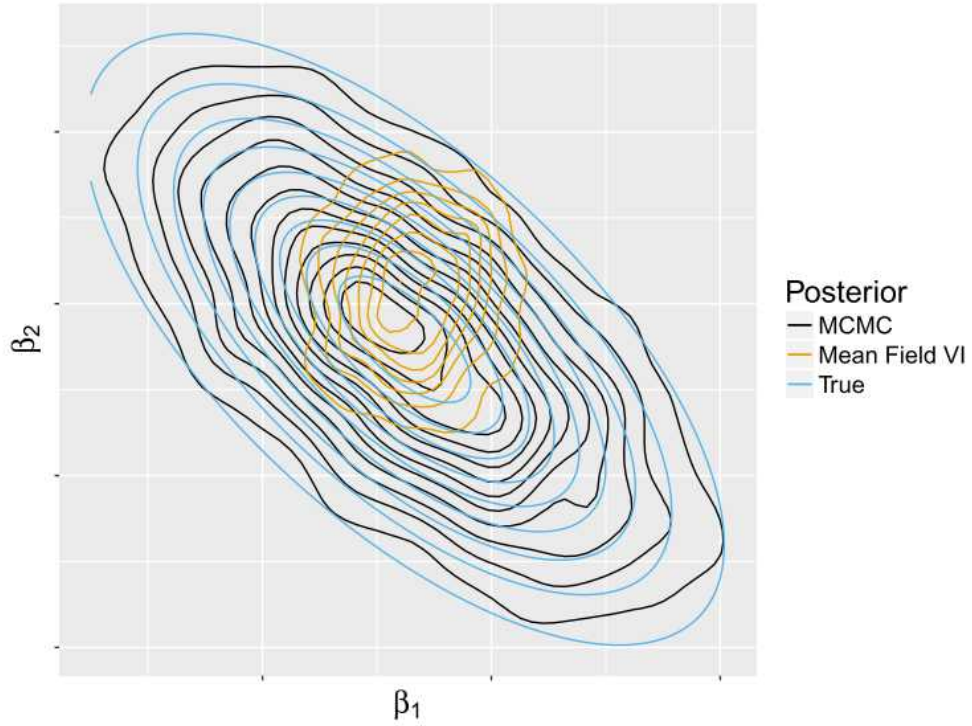
But does VI accurately account for uncertainty?

Mean Field VI (MFVI) with Normal variational distributions is common approach

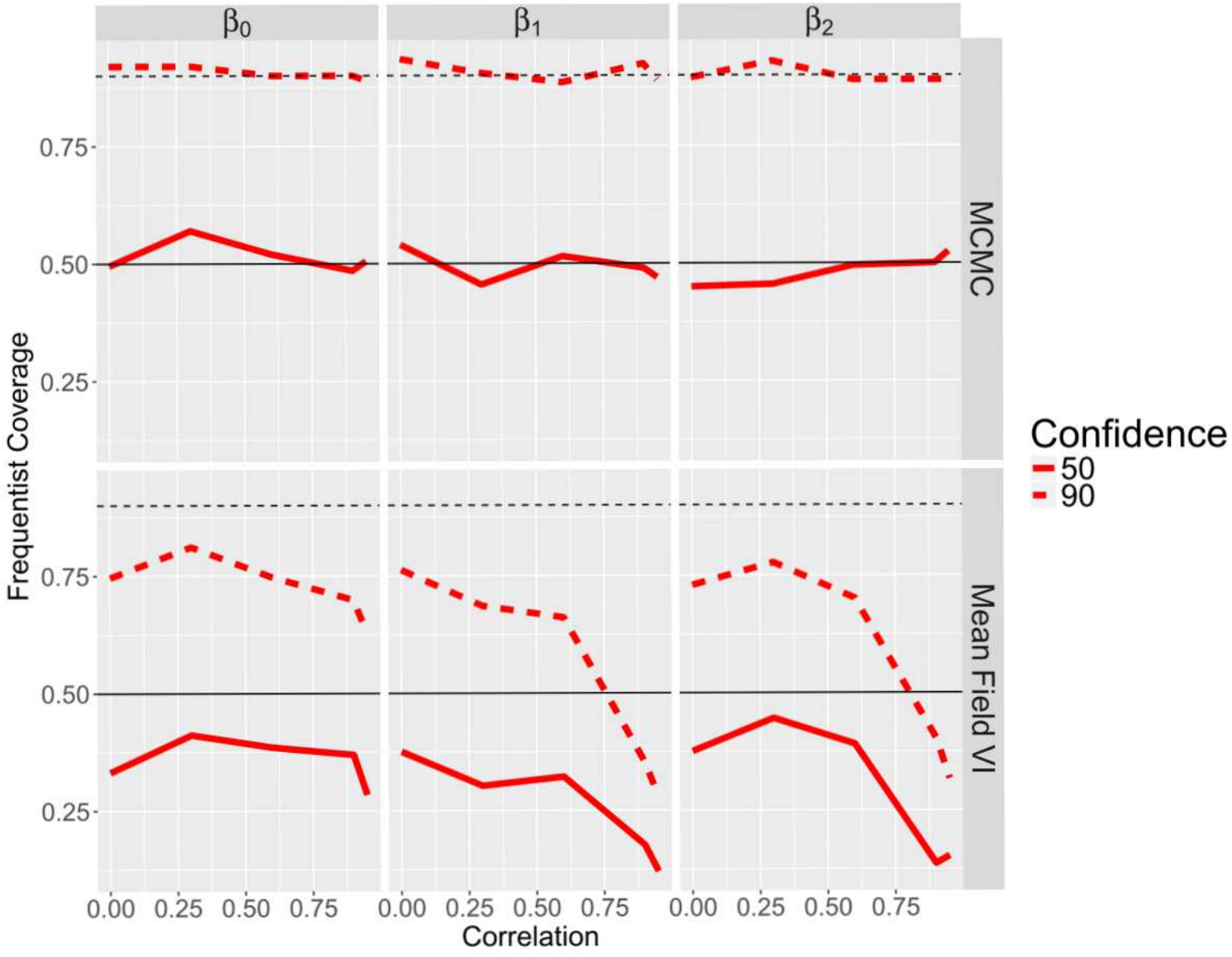
- Often much faster than MCMC or VI with multivariate variational distribution
- Relatively easy to implement
- Asymptotic convergence of *marginal* posterior distributions

Assessing MFVI on Linear Regression

$$y_i|\beta, \sigma^2, x \sim N(\beta_1 x_{i1} + \beta_2 x_{i2}, \sigma^2), i = 1, ..., 50$$
$$\sigma^2 \sim IG(a, b)$$
$$\beta|\sigma^2 \sim N(m, \sigma^2 V)$$



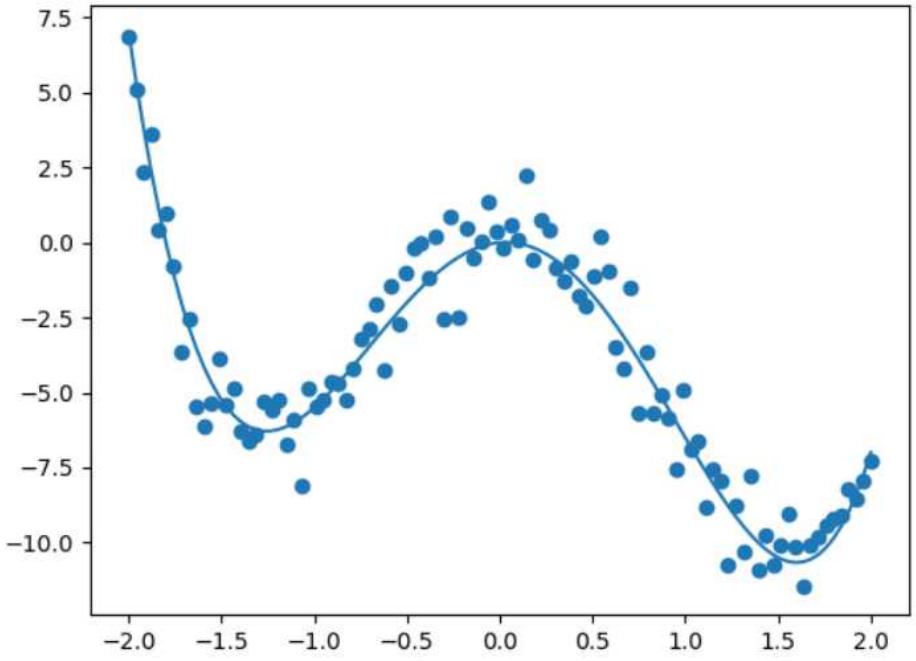
Frequentist Coverage of MCMC vs MFVI



Simulation Study

Simulate data from:

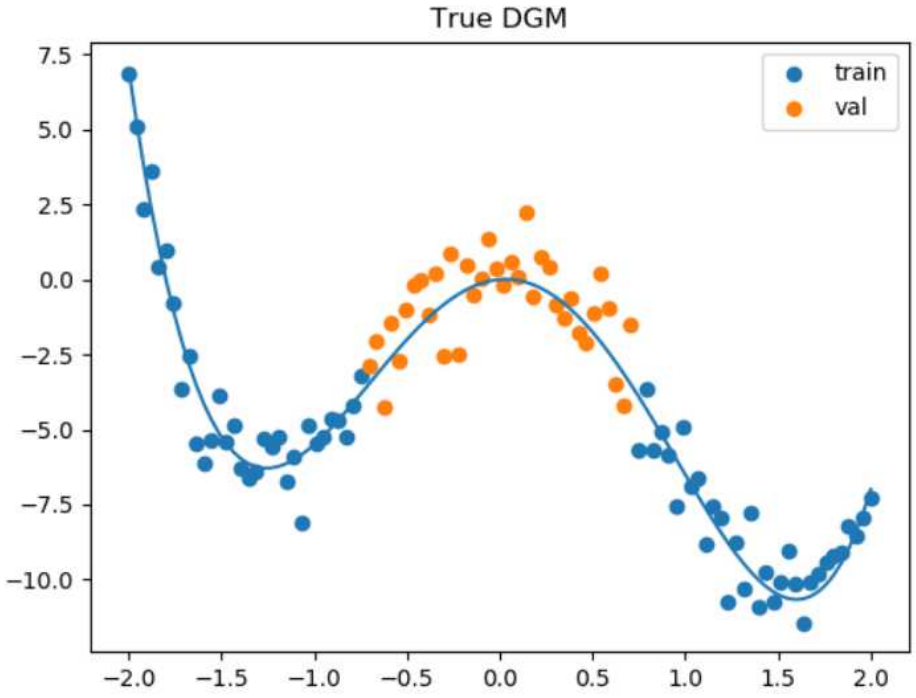
$$Y|x \sim N(0.5x - 8x^2 - x^3 + 2x^4, 1)$$



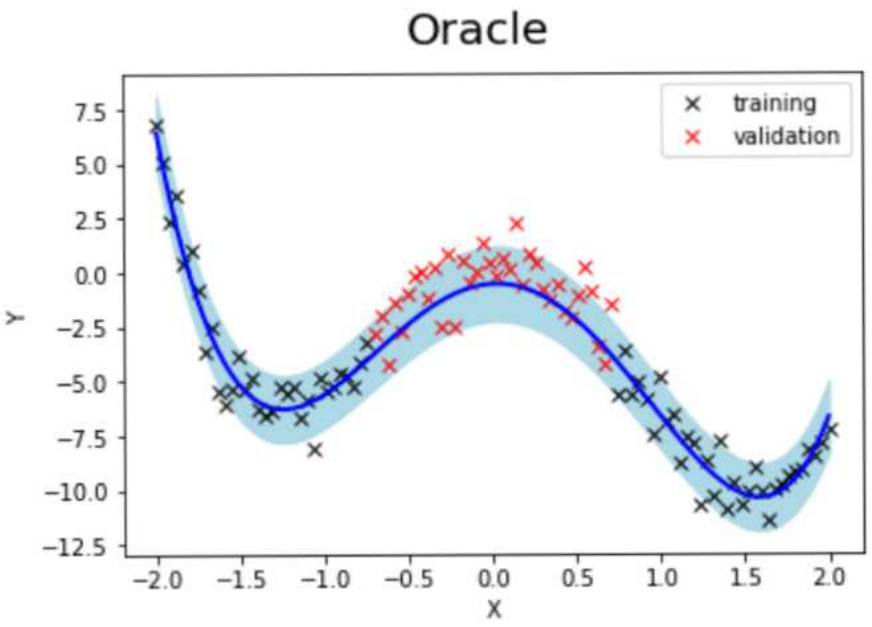
Simulation Study

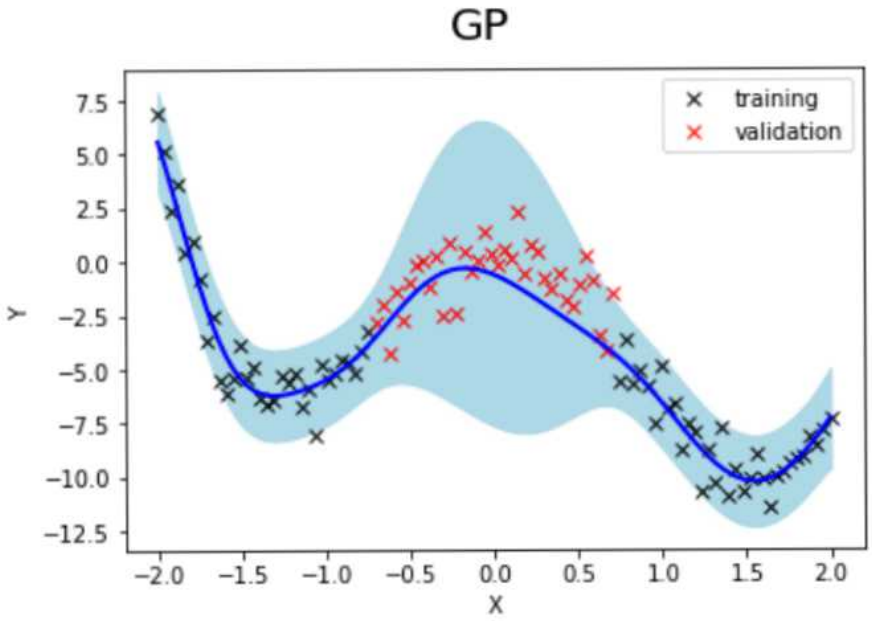
Simulate data from:

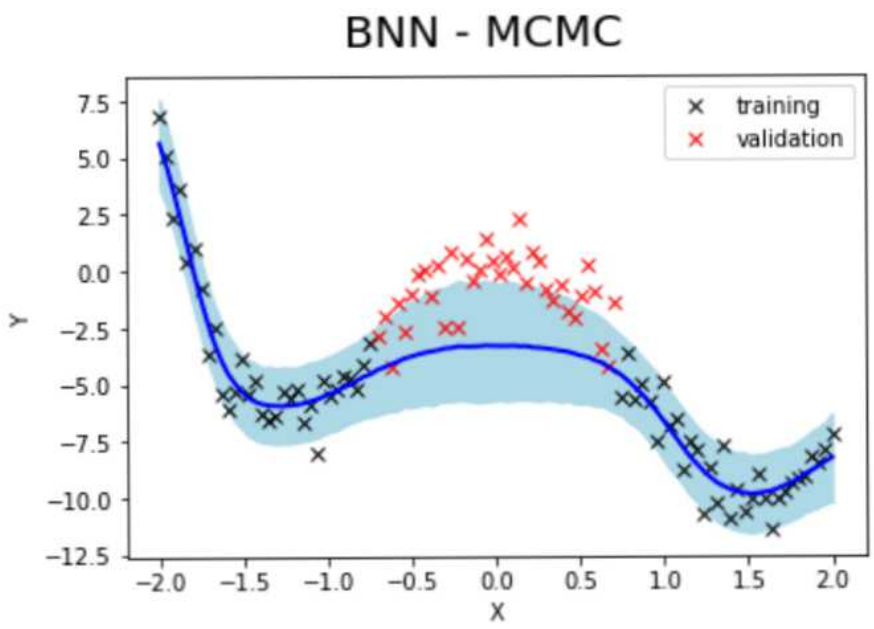
$$Y|x \sim N(0.5x - 8x^2 - x^3 + 2x^4, 1)$$

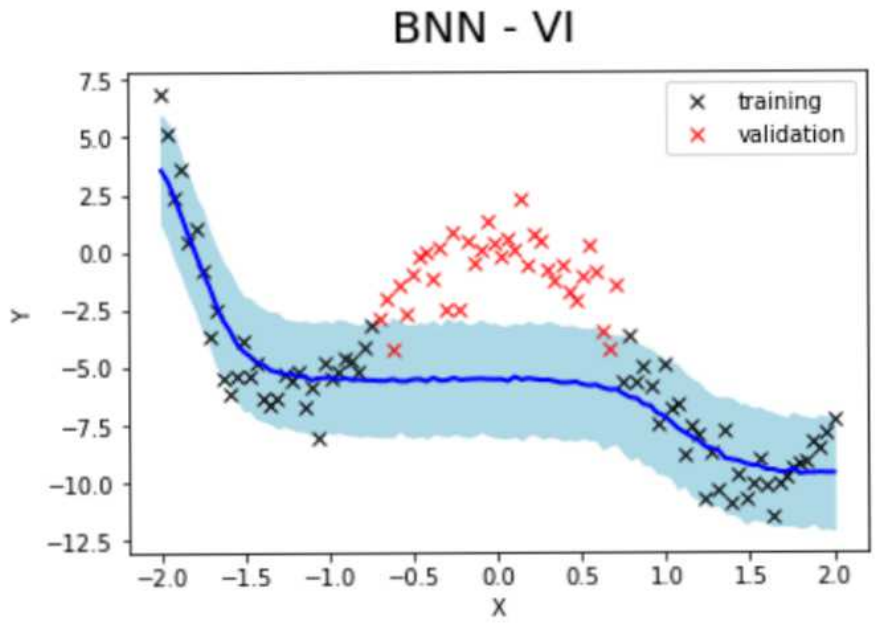


Compare Nonlinear Models







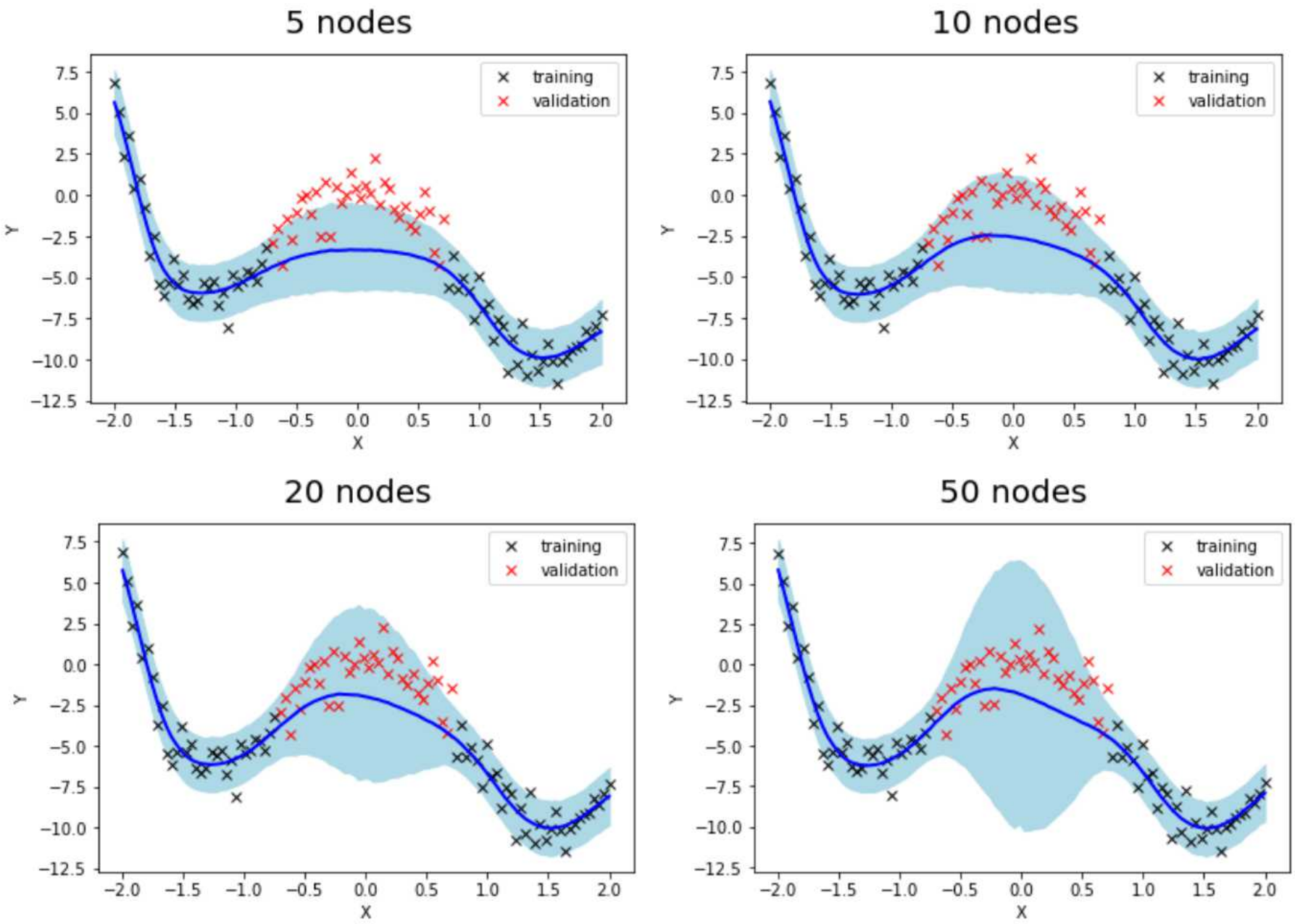


Model Metrics

Model	MSE	Coverage (0.9)	Width-Train	Width-Val
Oracle	1.2 (0.4)	0.83 (0.02)	2.6 (0.2)	2.8 (0.2)
GP	1.9 (1.1)	0.94 (0.02)	3.5 (0.3)	8.2 (0.5)
BNN- MCMC	6.3 (2.9)	0.93 (0.01)	3.6 (0.3)	4.8 (0.5)
BNN-VI	15.3 (6.1)	0.88 (0.02)	4.9 (1.0)	4.8 (1.0)

Table: Summary statistics from simulation study assessing predictive and UQ power of different models. Coverage are assessed on training set only.

BNN - MCMC with different number nodes



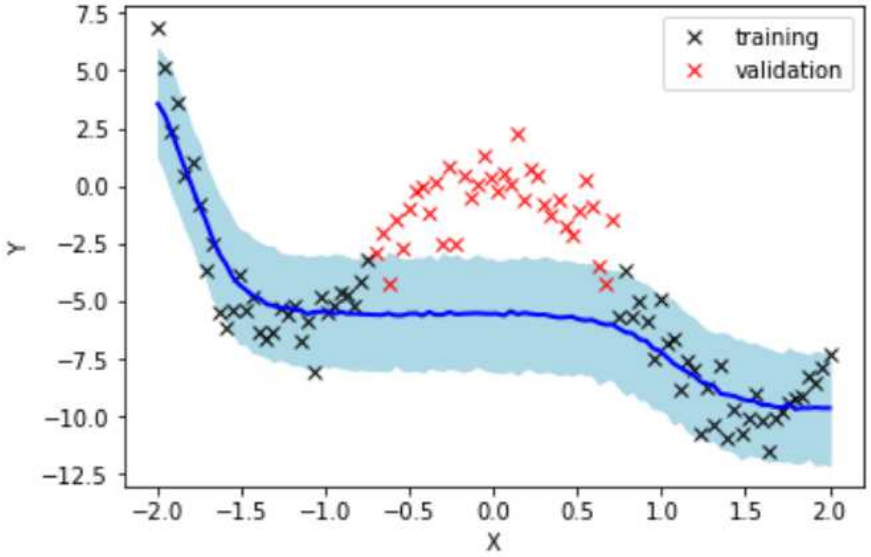
Model Metrics-MCMC

Model	MSE	Coverage (0.9)	Width-Train	Width-Val
5 nodes	6.3 (2.9)	0.93 (0.01)	3.6 (0.3)	4.8 (0.5)
10 nodes	4.0 (2.1)	0.93 (0.02)	3.5 (0.3)	5.7 (0.4)
20 nodes	2.8 (1.9)	0.93 (0.02)	3.5 (0.3)	7.5 (0.4)
50 nodes	3.2 (2.8)	0.88 (0.02)	3.4 (0.3)	10.9 (0.3)

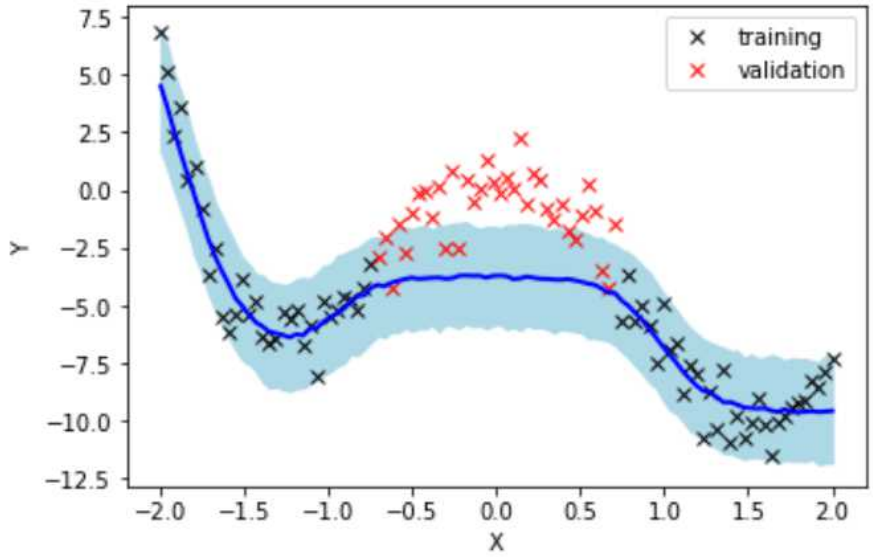
Table: Summary statistics from simulation study assessing predictive and UQ power of MCMC-trained BNN with different architectures. Coverage are assessed on training set only.

BNN - VI with different number nodes

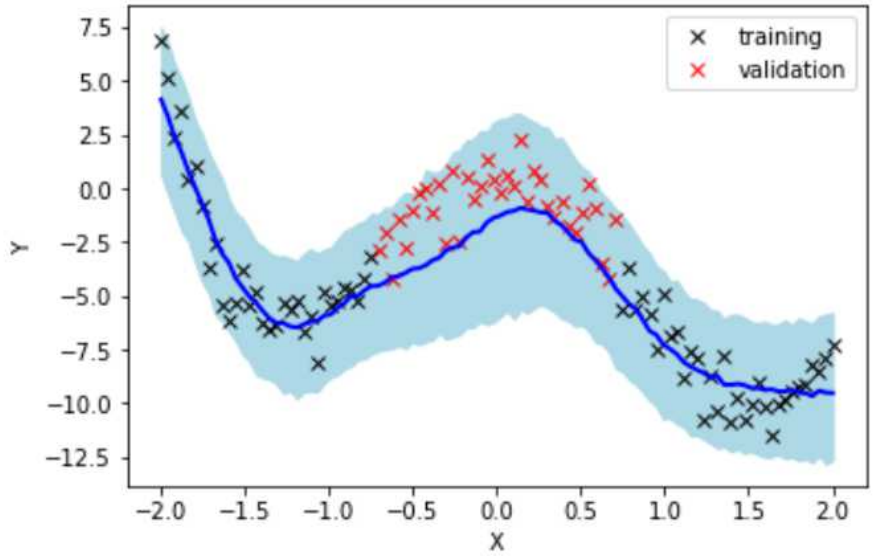
5 nodes



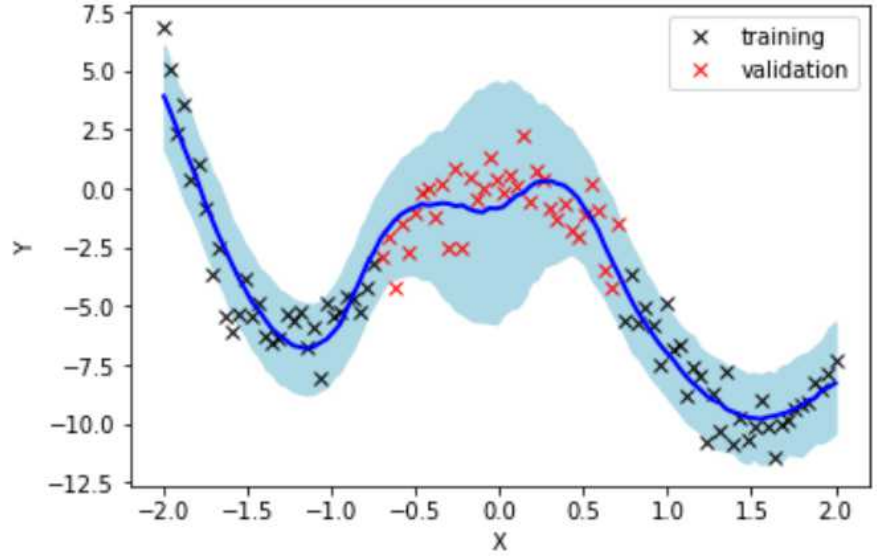
10 nodes



20 nodes



50 nodes



Model Metrics-VI

Model	MSE	Coverage (0.9)	Width-Train	Width-Val
5 nodes	15.3	0.88	4.9	4.8
10 nodes	8.7	0.91	5.0	5.1
20 nodes	6.8	0.98	6.2	6.9
50 nodes	13.8	1	27.3	28.7

Table: Summary statistics from simulation study assessing predictive and UQ power of VI-trained BNN with different architectures. Coverage are assessed on training set only.

Simulation Study Summary

- MCMC-fit BNN behave as expected
- VI-fit BNN did not behave as expected
- NN architecture will affect over/underfitting (obviously) of UQ
- Ad hoc NN architectures may present problems with UQ in practical problems

Discussion

Successes:

- BNN showed its potential in providing “high confidence sets”
- UQ from MCMC-fit BNN appears reliable

Questions:

- How do we effectively use VI for deep nets?
- How do we choose architecture with UQ in mind?
- What about the priors?

Thank you for listening!

dries@sandia.gov

Test Set Performance-Constant False Alarm Rate

A: Abundance of target in pixel

	$A < 0.25$	$0.25 \leq A < 0.75$	$A \geq 0.75$
Full Set	0.471	0.964	0.992
High Confidence	0.726	0.994	1.0

Table: Probability of detection for FAR = 0.05