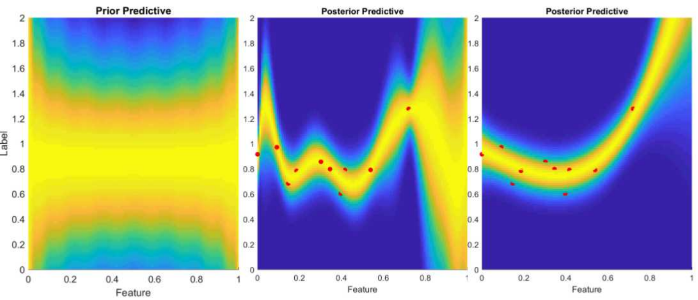
SAND2020-7119PE



A Bayesian Perspective on Machine Learning and UQ

Thomas A. Catanach and Jed A. Duersch

CSRI Summer Seminar

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and 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. SAND NO. 2019-8684 C

Overview

- Why Machine Learning?
- Key Questions for ML
- Bayesian Probability Theory
- Applying Bayesian Inference to ML
  - Challenges
  - Variational Inference
  - Priors

Why ML

- Machine Learning automates the process of learning predictive simplifications from data

- Enabling science and engineering for highly complex and evolving systems with masses of heterogenous data will require Machine Learning.

- For these tasks ML must be flexible, adaptable, and trustworthy.

Key Questions to Enable ML by UQ

- Representable: How do we represent our beliefs?
- Flexible: How to learn from data?
- Adaptable: How to gather new data for learning?
- Trustworthy: How to quantify the degree of trust in ML?

- Answers to these questions can be rigorously formulated within the Bayesian paradigm

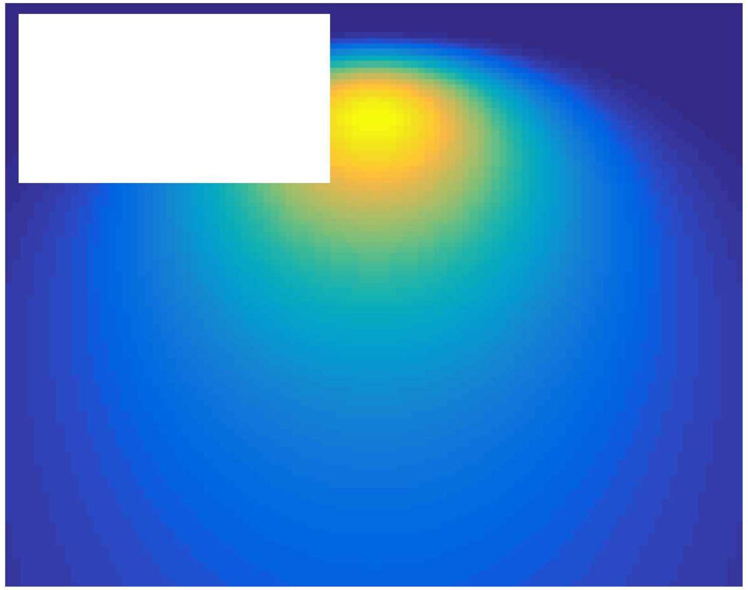
Representing Beliefs

- Within Bayesian theory, assumptions (φ) that express states of belief are represented using probability distributions.

- Prior $p(\theta \,|\mathcal{D}, \varphi)$: Initial belief about the universe
- Likelihood $p(\mathcal{D} \,|\theta, \varphi)$: Conditional beliefs about data
- Posterior $p(\theta \,|\mathcal{D}, \varphi)$: Updated beleif

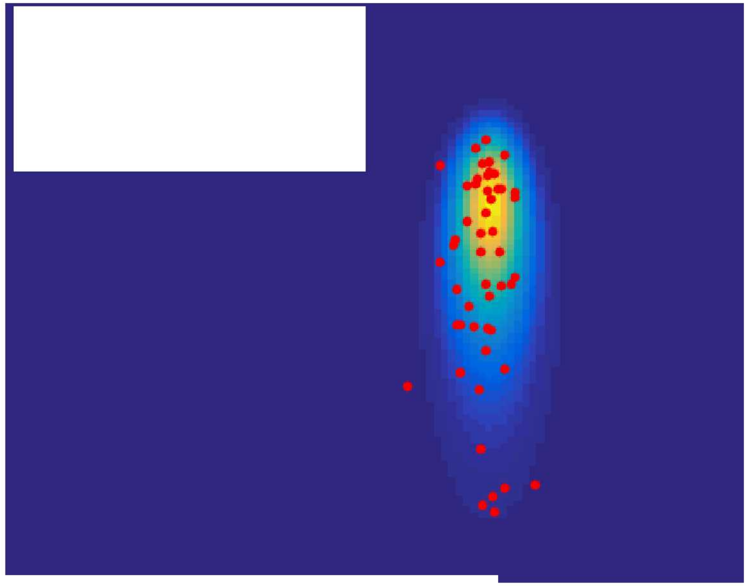
Updating Beliefs using Bayesian Inference

Observations: $\mathcal{D}$



Bayes' Theorem

$$p\left(\theta \mid \mathcal{D}, \mathcal{M}\right) = \frac{p(\mathcal{D}|\theta,\mathcal{M})p(\theta|\mathcal{M})}{p(\mathcal{D}|\mathcal{M})}$$



Evidence: $$p\left(\mathcal{D} \mid \mathcal{M}\right) = \int p\left(\mathcal{D} \mid \theta, \mathcal{M}\right) p\left(\theta \mid \mathcal{M}\right) d\theta$$

Posterior Estimation: $$\mathbb{E}\left[g\left(\theta\right) \mid \mathcal{D}, \mathcal{M}\right] = \int g\left(\theta\right) p\left(\theta \mid \mathcal{D}, \mathcal{M}\right) d\theta \approx \frac{1}{N} \sum_{i=1}^{N} g\left(\theta_i\right)$$

Quantifying Change in Belief

- Information quantifies how belief $q(\theta)$ change to $p(\theta)$ with respect to a state of belief $r(\theta)$:

$$\mathcal{I}_{r(\theta)}\left[p(\theta) \,||\, q(\theta)\right] = \int r(\theta) \log \frac{p(\theta)}{q(\theta)} d\theta$$

- Quantifying changes in belief due to inference:

$$\mathcal{I}_{p(\theta \mid \mathcal{D}, \psi, \mathcal{M})}\left[p(\theta \mid \mathcal{D}, \psi, \mathcal{M}) \,||\, p(\theta \mid \psi, \mathcal{M})\right] =$$

$$\mathrm{KL}\left[p(\theta \mid \mathcal{D}, \psi, \mathcal{M}) \,||\, p(\theta \mid \psi, \mathcal{M})\right] = \int p(\theta \mid \mathcal{D}, \psi, \mathcal{M}) \log \frac{p(\theta \mid \mathcal{D}, \psi, \mathcal{M})}{p(\theta \mid \psi, \mathcal{M})} d\theta$$

Assessing Learning Opportunities

- Bayesian Optimal Experimental Design
  1. Predict change in belief due to inference, the Expected Information Gain (EIG):

$$\mathrm{EIG}(d) = \int_{\mathcal{D}} p(\mathcal{D} \mid d) \int_{\theta} p(\theta \mid \mathcal{D}, d) \log \frac{p(\theta \mid \mathcal{D}, d)}{p(\theta)} d\theta d\mathcal{D}$$

  2. Optimize experiment or data collection to maximize expected information gain

$$d^* = \mathrm{argmax}_{d \in \mathfrak{D}} \mathrm{EIG}(d)$$

- EIG can also be assessed with respect to a QoI (Y)

$$\mathbb{E}_{\mathcal{D}|d} \{\mathrm{KL}[p(Y \mid \mathcal{D}, d) \,||\, p(Y)]\} = \int_{\mathcal{D}} p(\mathcal{D} \mid d) \int_{Y} p(Y \mid \mathcal{D}, d) \log \left[\frac{p(Y \mid \mathcal{D}, d)}{p(Y)}\right] dY\, d\mathcal{D}$$

Assessing Prediction Trustworthiness

Sandia National Laboratories

- Bayesian UQ requires estimating prediction uncertainty

Prediction with UQ

$$p(y|x,\mathcal{D},\varphi,\mathcal{M}) = \sum_{i=1}^{N} \int p(y|x,\theta,M_i)\, \frac{p(\mathcal{D} \mid \theta, M_i)\, p(\theta \mid \varphi, M_i)\, P(M_i)}{p(\mathcal{D} \mid \varphi, \mathcal{M})}\, d\theta$$

Assumptions: $\mathcal{D},\varphi,\mathcal{M}$

Prediction: y

Input: x

Prediction from specific model: $p(y|x,\theta,M_i)$

Posterior probability of specific model: $\frac{p(\mathcal{D} \mid \theta, M_i)\, p(\theta \mid \varphi, M_i)\, P(M_i)}{p(\mathcal{D} \mid \varphi, \mathcal{M})}$

Assessing Prediction Trustworthiness

Sandia National Laboratories

- Bayesian UQ requires estimating prediction uncertainty
- Bayesian Sensitivity Analysis or Robust Bayesian Inference quantifies the importance of assumptions on predictions. This predicts extrapolation or lack of generalization.

Prediction Perturbation

Assumption Perturbations: $\beta_{\mathcal{D}}, \beta_{\varphi}, \beta_{\mathcal{M}}$

Perturbed Likelihood: $p(\mathcal{D} \mid \theta, M_i, \beta_{\mathcal{D}})$

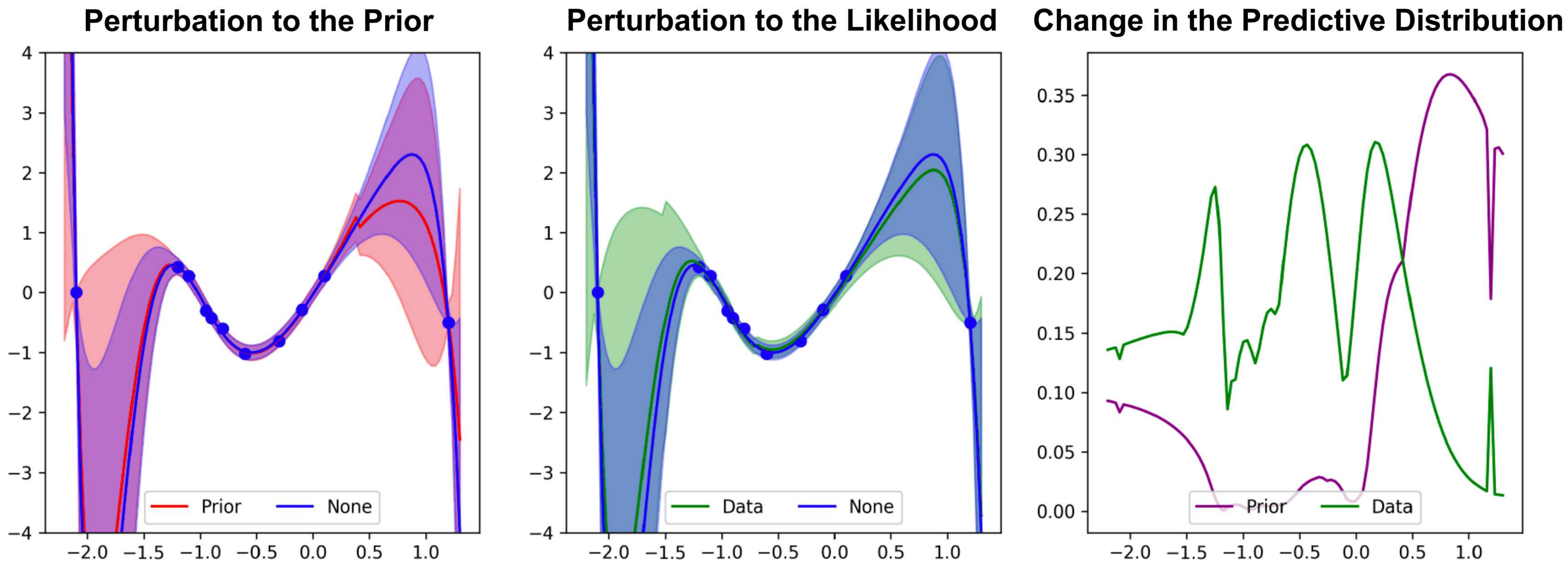
Perturbed Prior: $p(\theta \mid \varphi, M_i, \beta_{\varphi})$

Perturbed Model: $P(M_i|\beta_{\mathcal{M}})$

$$p(y|x,\mathcal{D},\varphi,\mathcal{M},\beta_{\mathcal{D}},\beta_{\varphi},\beta_{\mathcal{M}}) = \sum_{i=1}^{N} \int p(y|x,\theta,M_i) \frac{p(\mathcal{D} \mid \theta, M_i, \beta_{\mathcal{D}})\, p(\theta \mid \varphi, M_i, \beta_{\varphi})\, P(M_i|\beta_{\mathcal{M}})}{p(\mathcal{D} \mid \varphi,\mathcal{M},\beta_{\mathcal{D}},\beta_{\varphi},\beta_{\mathcal{M}})} d\theta$$

Assessing Prediction Trustworthiness

Linear regression example: Assessing prediction sensitivity to changes in the assumed prior and noise in the data



Bayesian Theory

- Representation -> Probability Distribution
- Learning -> Bayesian Inference
- Quantify change -> Information
- Gathering data -> Bayesian Optimal Experimental Design
- Trustworthiness -> Robust Bayesian Inference

- How do we apply Bayesian theory to ML?

Differences between ML and Standard UQ

- Parameters in physical systems have intrinsic meaning that makes defining priors easier and more interpretable
- Model structure often comes from first principles. When competing models exist, they often have more intuitive relationships i.e. fidelity.
- ML models often have too high capacity (many parameters), Physical systems often have too low capacity
- ML tries to build a prediction model while UQ for physical systems often trying to infer some unobservable states/parameters from data.

Challenges in applying Bayesian methods to ML

- Representing a probability distribution over models
- Choosing a prior and space of model architectures
- Solving the inference problem and computing information in high dimensions

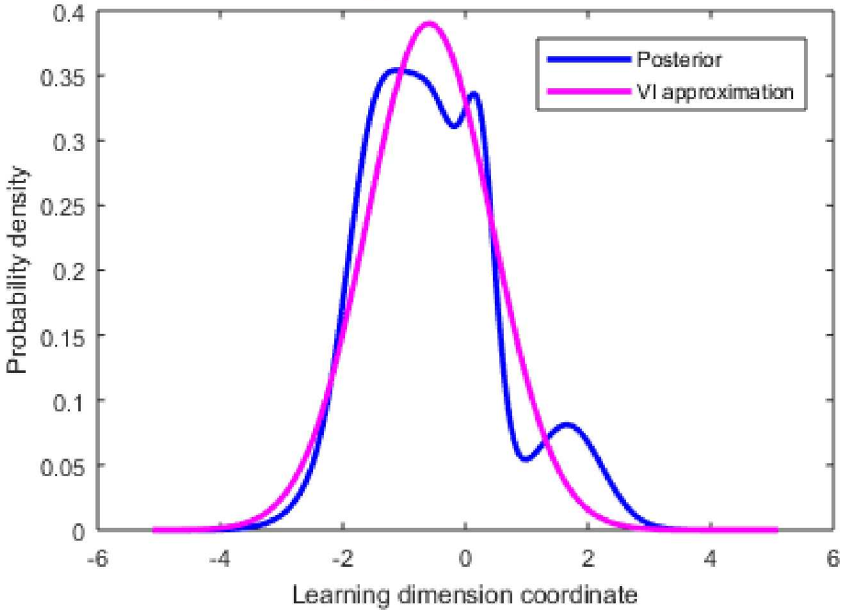
Variational inference

The **posterior distribution is virtually impossible to represent and solve in high dimensional problems** like over-parameterized deep learning.

Variational inference approximates the posterior distribution using more tractable methods:

- Local Gaussian approximation
- Stochastic gradient descent sampling
- Mean-field distribution
- Dropout sampling
- Variational Tempering

Illustration of Variational Approximation



Maximizing the ELBO amounts minimizing the following Kullback-Leibler divergences:

$$\mathcal{L}(\lambda) = -D_{\mathrm{KL}}\left[q(\theta|\lambda)||p(D,\theta)\right] = \log\left(p(D)\right) - D_{\mathrm{KL}}\left[q(\theta|\lambda)||p(\theta|D)\right]$$

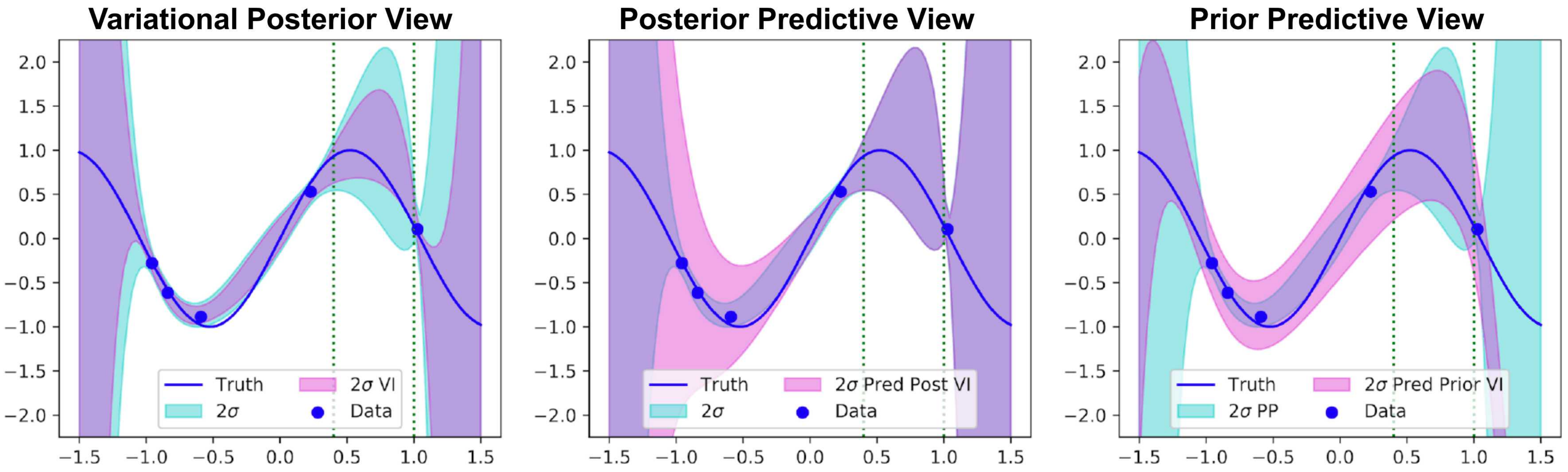
Variational Inference

- Different views of VI may be appropriate for different ML tasks

View	Expression
Posterior Distribution	$\mathrm{KL}\left[p(\theta \mid \mathcal{D}) \mid\mid Q(\theta \mid \phi)\right] = \int_\theta p(\theta \mid \mathcal{D}) \log \frac{p(\theta \mid \mathcal{D})}{Q(\theta \mid \phi)} d\theta$
Variational Distribution	$\mathrm{KL}\left[Q(\theta \mid \phi) \mid\mid p(\theta \mid \mathcal{D})\right] = \int_\theta Q(\theta \mid \phi) \log \frac{Q(\theta \mid \phi)}{p(\theta \mid \mathcal{D})} d\theta$
Variational Predictive	$\mathrm{KL}\left[Q(Y \mid \phi) \mid\mid p(Y \mid \mathcal{D})\right] = \int_Y Q(Y \mid \phi) \log \frac{Q(Y \mid \phi)}{p(Y \mid \mathcal{D})} dY$
Prior Predictive	$\mathrm{KL}\left[p(Y) \mid\mid Q(Y \mid \phi)\right] = \int_Y p(Y) \log \frac{p(Y)}{Q(Y \mid \phi)} dY$

Variational Inference

Linear regression example: Assessing different VI formulations for prediction



Monte Carlo Estimators for Variational Inference

- Evidence Lower Bound:

Posterior $$\mathrm{KL}\left[Q\left(\theta \mid \phi\right) \,||\, p\left(\theta \mid \mathcal{D}\right)\right] - \log p\left(\mathcal{D}\right) \approx \frac{1}{N}\sum_{i=1}^{N}\left[\log Q\left(\theta_i \mid \phi\right) - \log p\left(\mathcal{D} \mid \theta_i\right) p\left(\theta_i\right)\right], \theta_i \sim Q\left(\theta \mid \phi\right)$$

Posterior Predictive $$\mathrm{KL}\left[Q\left(Y \mid \phi\right) \,||\, p\left(Y \mid \mathcal{D}\right)\right] - \log p\left(\mathcal{D}\right) \approx \frac{1}{N}\sum_{i=1}^{N}\left[\log Q\left(Y_i \mid \phi\right) - \log p\left(\mathcal{D}, Y_i\right)\right], Y_i \sim Q\left(\theta \mid \phi\right)$$

- Assuming $Q\left(Y \mid \phi\right) = \int_\theta p\left(Y \mid \theta\right) Q\left(\theta \mid \phi\right) d\theta$

$$\mathrm{KL}\left[Q\left(Y \mid \phi\right) \,||\, p\left(Y \mid \mathcal{D}\right)\right] - \log p\left(\mathcal{D}\right)$$
$$\approx \frac{1}{N}\sum_{i=1}^{N}\left[\log \frac{1}{K}\sum_{l=1}^{K} p\left(Y_i \mid \theta'_l\right) - \log \frac{1}{M}\sum_{j=1}^{M} p\left(\mathcal{D}, Y_i \mid \theta_j\right)\right], Y_i \sim Q\left(Y \mid \phi\right), \theta_j \sim p\left(\theta\right), \theta'_l \sim Q\left(\theta \mid \phi\right)$$

Variational Inference: Bayes By Backprop

Algorithm from Blundell et al. 2015

1. Sample $\epsilon \sim \mathcal{N}(0, I)$.
2. Let $\mathbf{w} = \mu + \log(1 + \exp(\rho)) \circ \epsilon$.
3. Let $\theta = (\mu, \rho)$.
4. Let $f(\mathbf{w}, \theta) = \log q(\mathbf{w}|\theta) - \log P(\mathbf{w})P(\mathcal{D}|\mathbf{w})$.
5. Calculate the gradient with respect to the mean

$$\Delta_\mu = \frac{\partial f(\mathbf{w}, \theta)}{\partial \mathbf{w}} + \frac{\partial f(\mathbf{w}, \theta)}{\partial \mu}. \quad (3)$$

6. Calculate the gradient with respect to the standard deviation parameter ρ

$$\Delta_\rho = \frac{\partial f(\mathbf{w}, \theta)}{\partial \mathbf{w}} \frac{\epsilon}{1 + \exp(-\rho)} + \frac{\partial f(\mathbf{w}, \theta)}{\partial \rho}. \quad (4)$$

7. Update the variational parameters:

$$\mu \leftarrow \mu - \alpha \Delta_\mu \quad (5)$$

$$\rho \leftarrow \rho - \alpha \Delta_\rho. \quad (6)$$

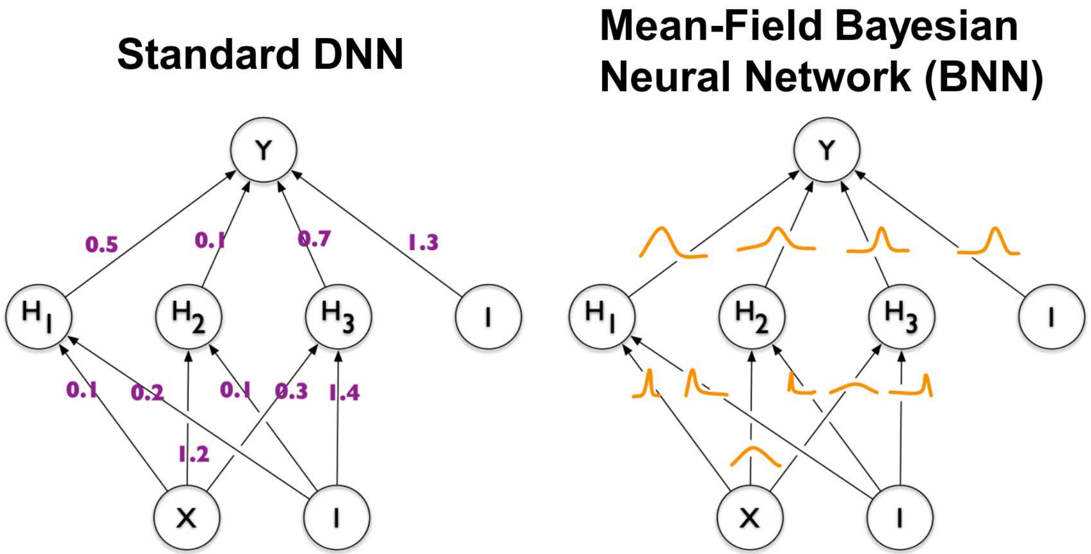


Illustration from Blundell et al. 2015

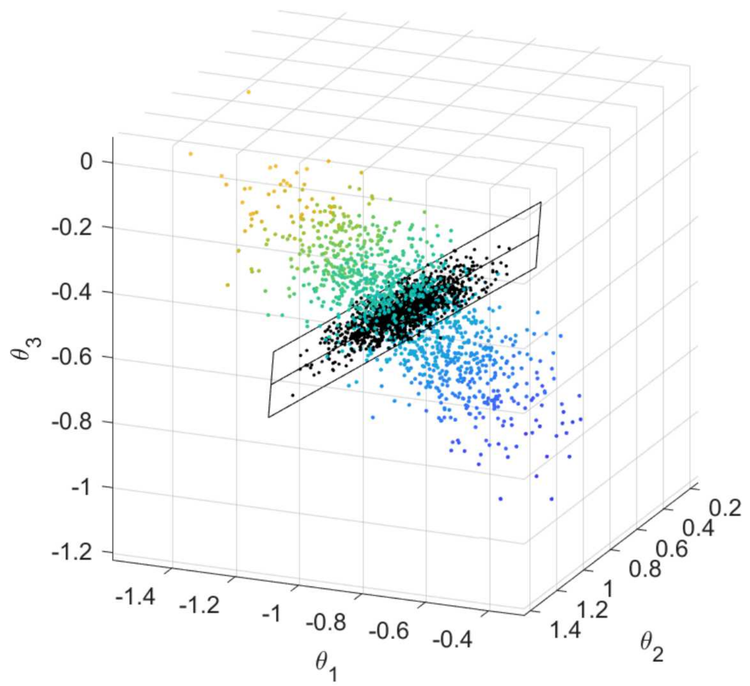
Variational Inference: Subspace Restriction

- Many ML models like DNNs are high over parameterized so uncertainty may only need to be captured in a small subspace in order to capture predictive uncertainty

- VI means finding the parameterization (ψ) in the basis (U) such that $\theta = U\psi$

- Subspace variational inference is exact when we reinterpret the subspace as just expressing a new model representation

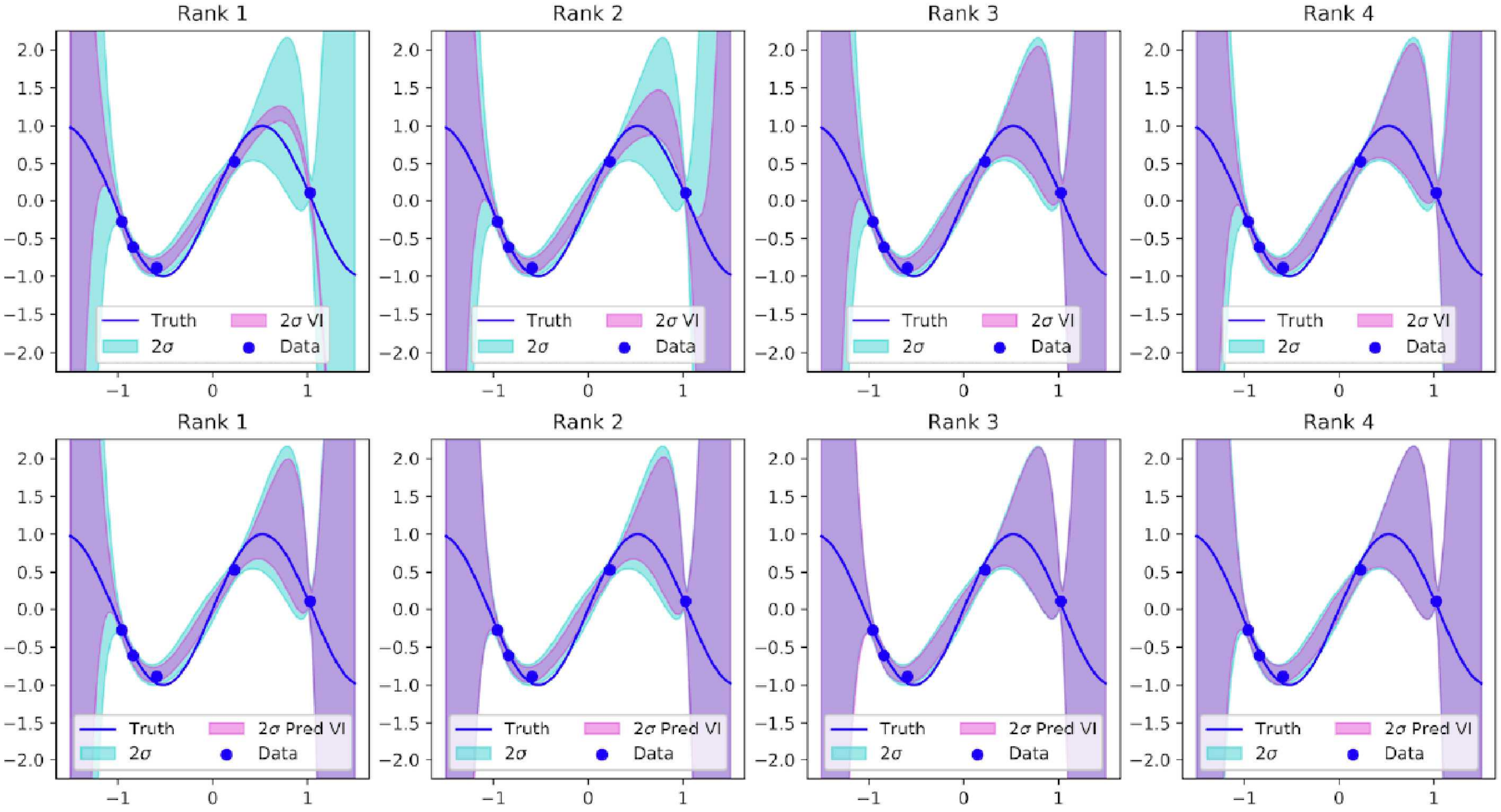
Example subspace restriction



Variational Inference: Subspace Restriction

Variational Posterior View

Variational Posterior Predictive View



Priors for ML models

- Priors are critical for accurate UQ and model selection
- Because ML models are abstract, prior assumptions are difficult to quantify
- ML model structure encode priors that we do not quite understand but is useful
- Examples of priors
  - Strict assumptions about model architectures to include symmetries, invariances, and hierarchical structure i.e. CNNs
  - Tasks being shared from one model to another i.e. transfer learning
  - Principle of maximum entropy
  - Explicitly developing priors to encode beliefs about the underlying predictions
  - Model complexity (Algorithmic Probablity)

Maximum Entropy Prior with Prediction Properties

Principle of maximum entropy:

- The prior that best represents parameter uncertainty would maximize the parameter posterior entropy while being consistent with our beliefs.
- This maximizes our uncertainty about parameters not the predictions

$$\omega^* = \text{argmax}_\omega - \int p(\theta \mid \omega) \log p(\theta \mid \omega) \, d\theta$$

$$\text{s.t.} \int g(\theta) \, p(\theta \mid \omega) \, d\theta = g_{\text{belief}}$$

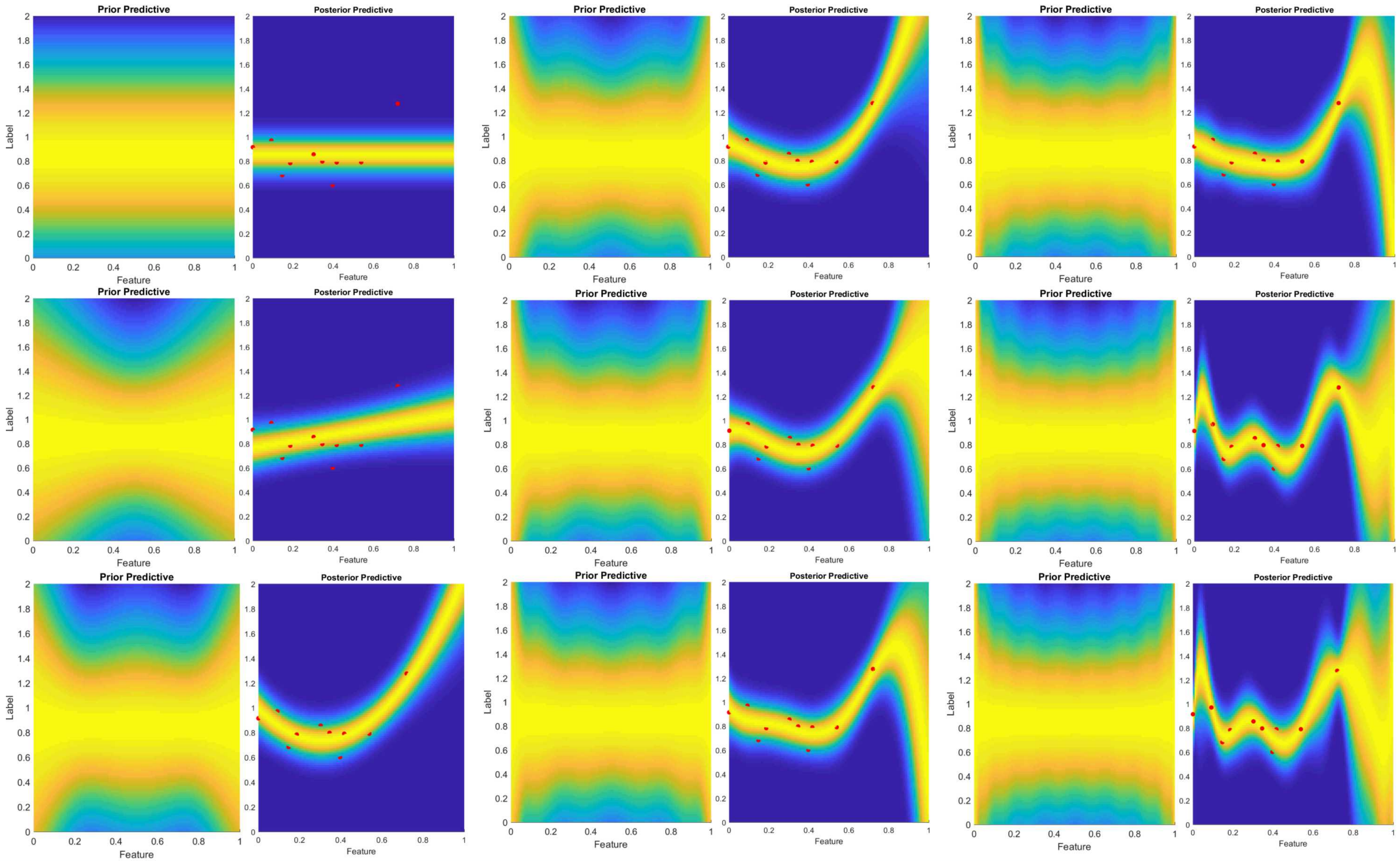
Example: Belief about expected mean and variance of predictions with linear regression

$$\mu^*, \Sigma^* = \text{argmax}_{\mu,\Sigma} - \int p(\theta \mid \mu, \Sigma) \log p(\theta \mid \mu, \Sigma) \, d\theta$$

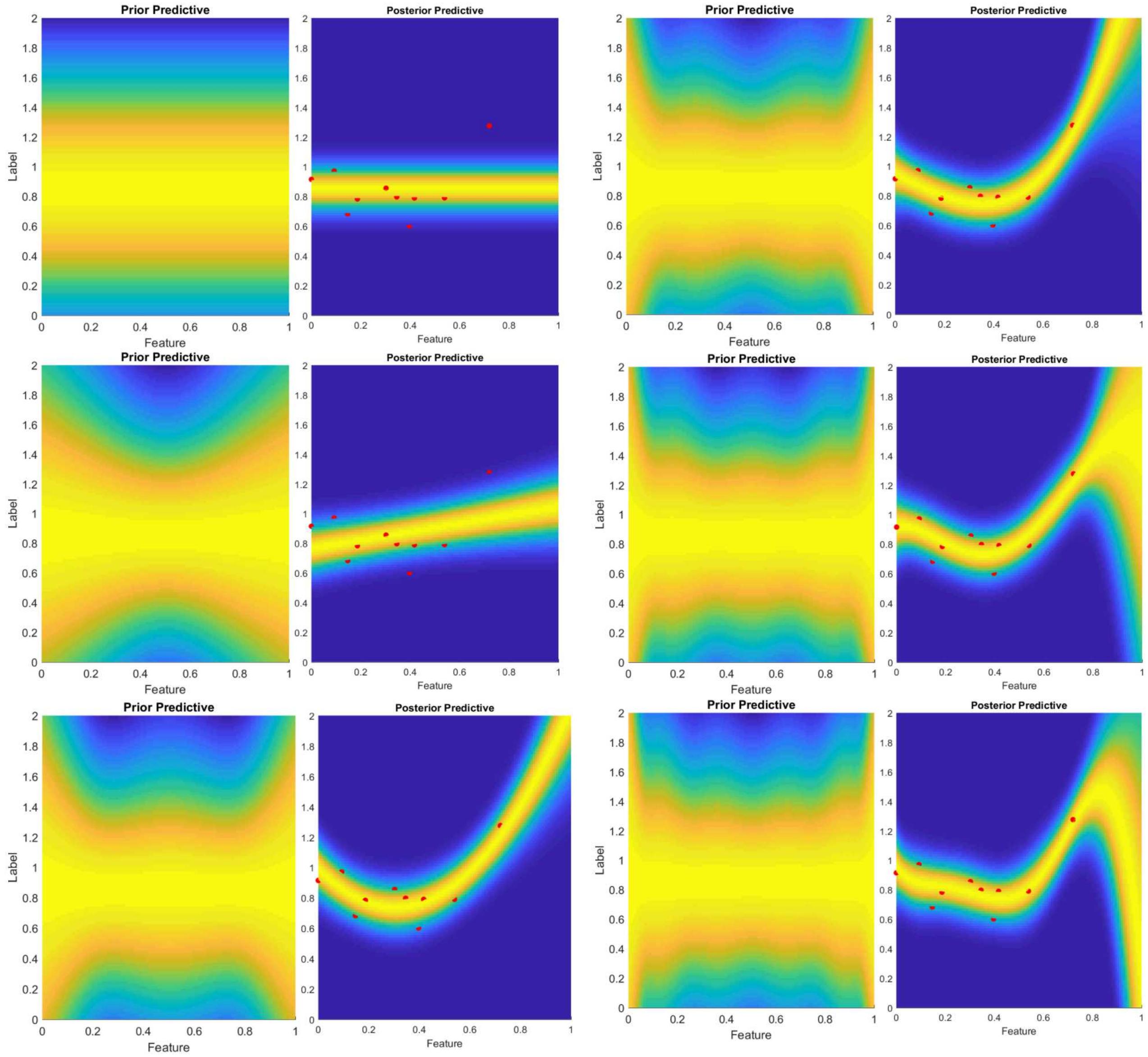
$$\text{s.t.} \int \int y(\theta, x) \, p(\theta \mid \mu, \Sigma) \, p(x) \, d\theta dx = \bar{\mu}$$

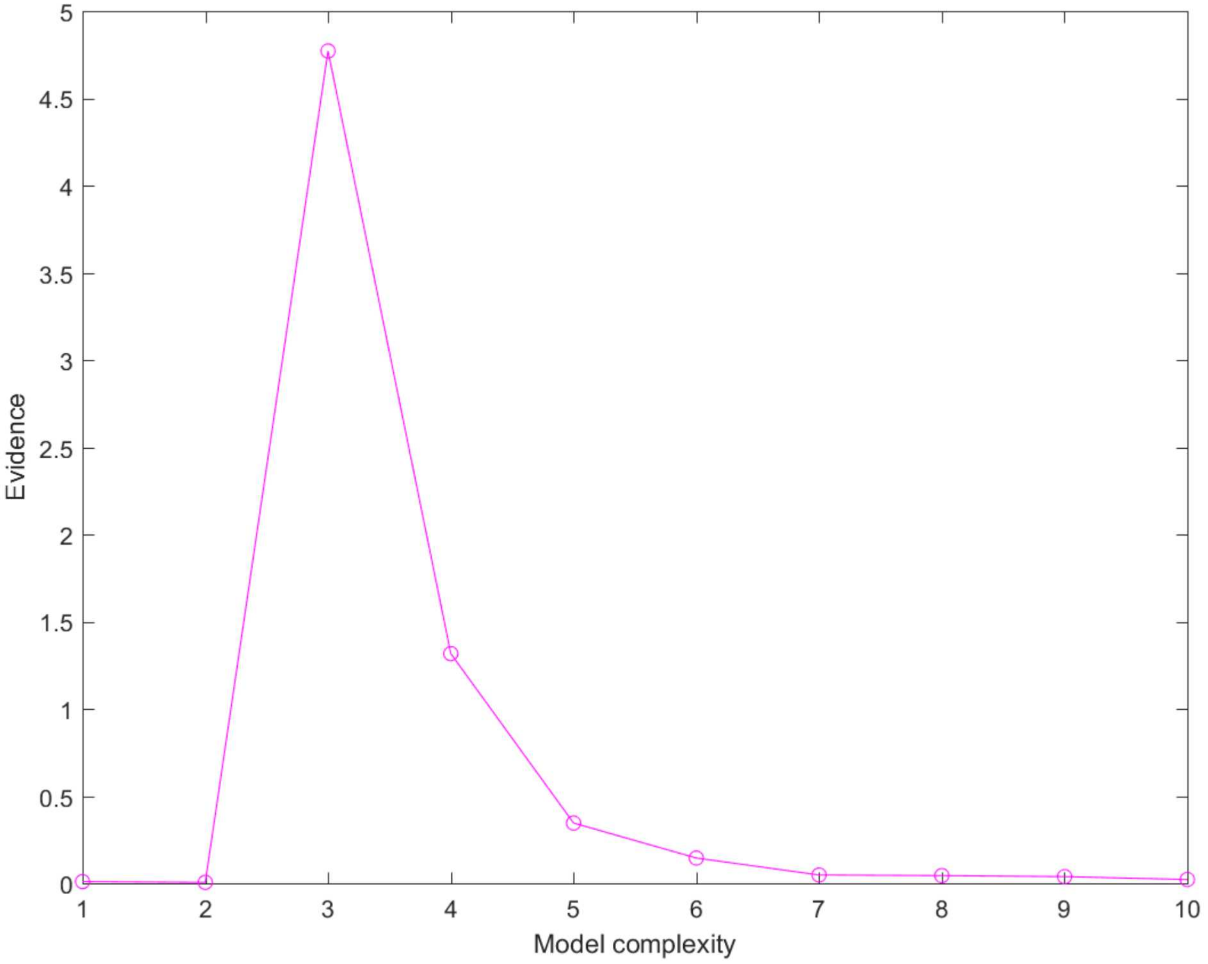
$$\int \int (y(\theta, x) - \bar{\mu})^2 \, p(\theta \mid \mu, \Sigma) \, p(x) \, d\theta dx = \bar{\sigma}^2$$

Linear Regression Model Example



Linear Regression Model Evidence





Prior on Model Complexity

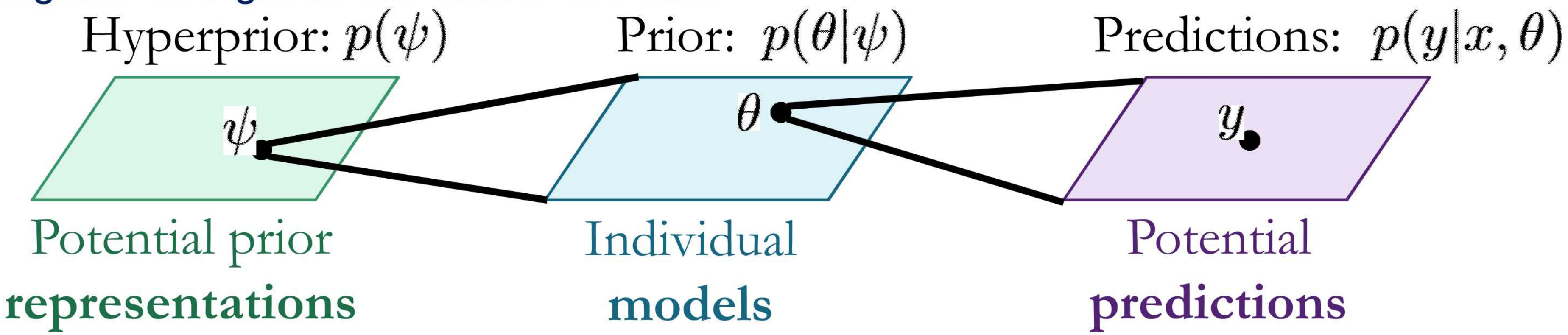
- Occam's Razor expresses a belief that similar models should be preferred over complex models.
- This is a type of prior. We a priori believe that predictions can be learned from the data because the relationship is relatively simple.
- Prior belief should be consistent with the minimum number of bits i.e. information needed to represent the model (Solomonov, 1960-1964; Rissanen, 1983).

$$p(\psi) \propto 2^{-\chi(\psi)}$$

- This leads to parsimonious inference: Bayesian inference with a prior that limits model complexity, including both source-coding information in the symbolic representations and information in inference

Parsimonious inference

Our theory regards changes in three kinds of belief:



We derive the following optimization objective from the total information, change in belief, that occurs when we observe the data and construct a predictive model.

$$\omega(\check{\psi}) = \mathbb{E}_{p(\theta|\check{y},\check{\psi})} \mathbb{I}_{r(y|\check{y})} \left[p(y|\theta) \,\|\, q_0(y) \right]$$

Expected info gained about data.

complexity terms:

$$- \mathbb{I}_{p(\theta|\check{y},\check{\psi})} \left[p(\theta|\check{y}, \check{\psi}) \,\|\, p(\theta|\check{\psi}) \right]$$

Model info due to inference.

$$- \mathbb{I}_{r(\psi|\check{\psi})} \left[r(\psi|\check{\psi}) \,\|\, p(\psi) \right].$$

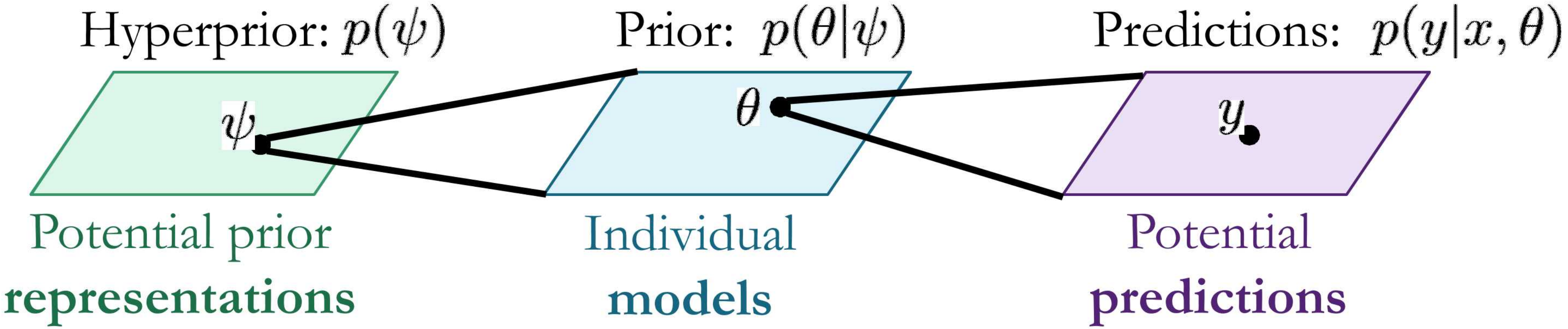
Representation info due to selection.

$$= \log \left(p(\check{\psi}|\check{y}) \right) + \mathbb{I}_{r(y|\check{y})} \left[p(y) \| q_0(y) \right].$$

Representation posterior.

Parsimonious inference

Examples:

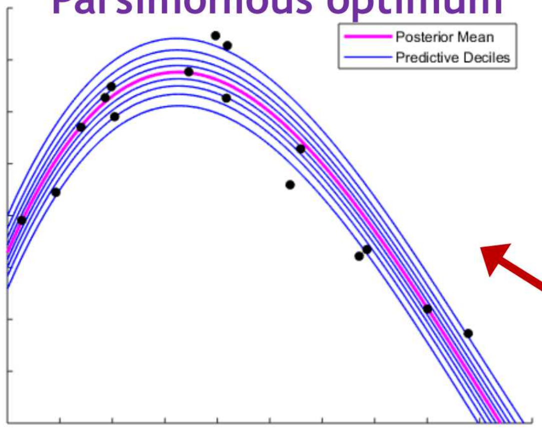


Polynomial Regression	Prior Mean and Variation	Polynomial Coefficient Distribution	Predicted Response
Decision Tree	Tree structure i.e. splitting decisions and location	Dirichlet Distribution for Class Probability	Predicted Class Probability

Parsimonious inference

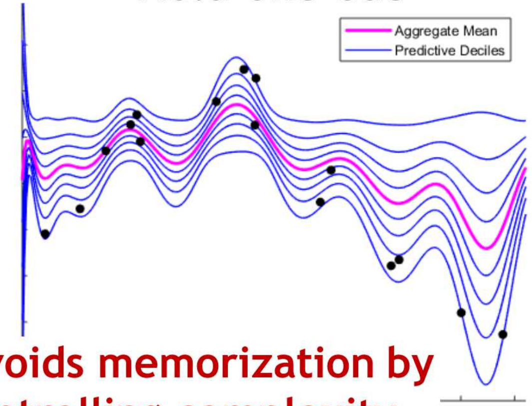
Parsimonious inference (PI) extracts simple regression models even if we train with 20th degree polynomials

Parsimonious optimum

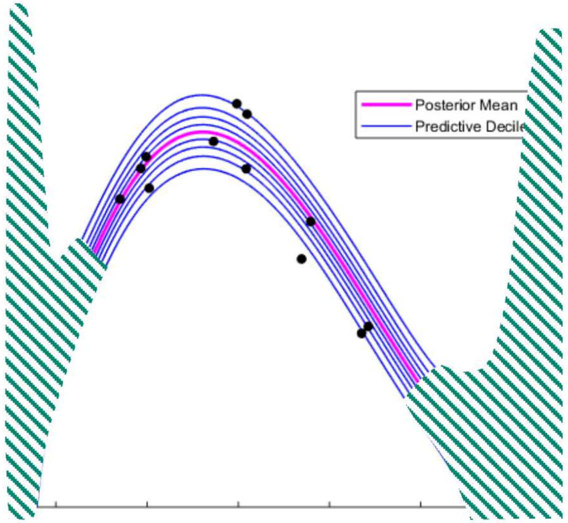


vs

Hold-one-out



PI avoids memorization by controlling complexity

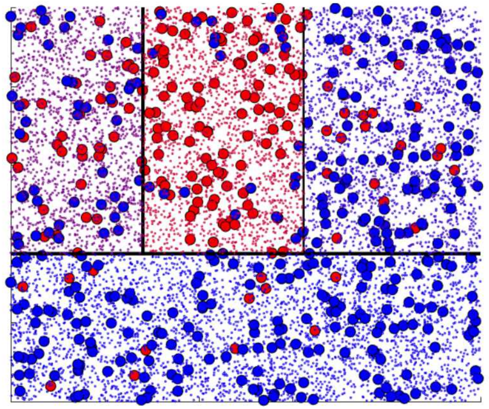


Natural extrapolation uncertainty

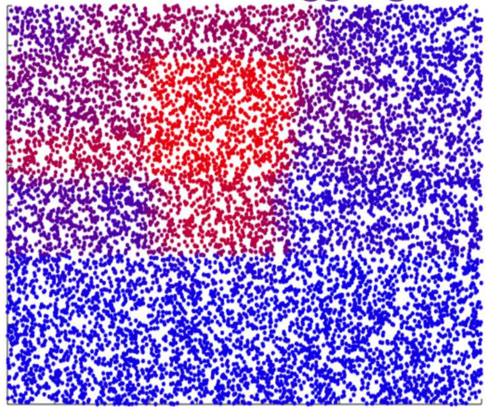
Aggregated predictions over 60 plausible models

PI also learns simpler decision trees than standard approaches

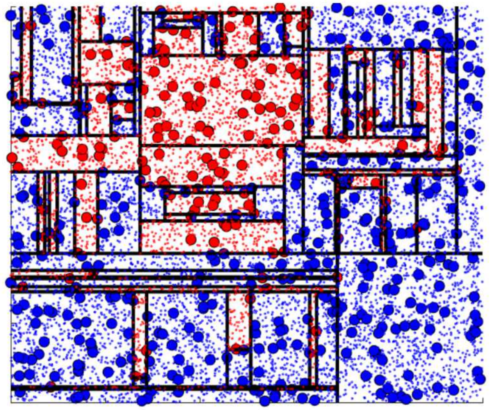
Parsimonious tree



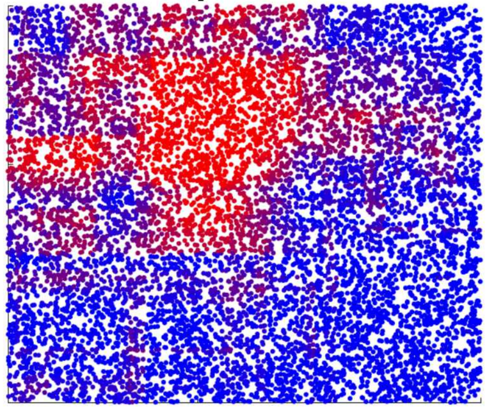
Parsimonious aggregate



Standard information-based decision tree



Bootstrap aggregate forest predictions



Why do we need UQ for ML?

- Applications where highly confident predictions are needed
- Layering ML models together which magnifies errors and requires a notion of a model’s operational envelope
- Designing for specification instead of what the data is capable of
- How to improve a prediction model by augmenting the model structure or gathering new data

Key Challenges for Bayesian UQ for ML

- Representing a probability distribution over models

- Choosing a prior and space of model architectures

- Solving the inference problem and computing information in high dimensions