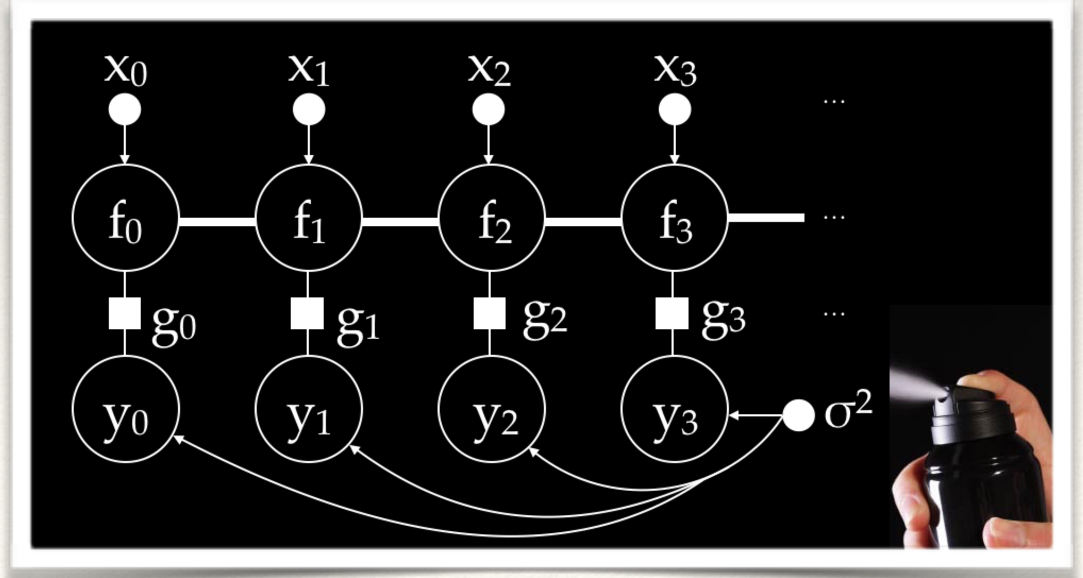
International Conference on Machine Learning, June 22nd, 2016

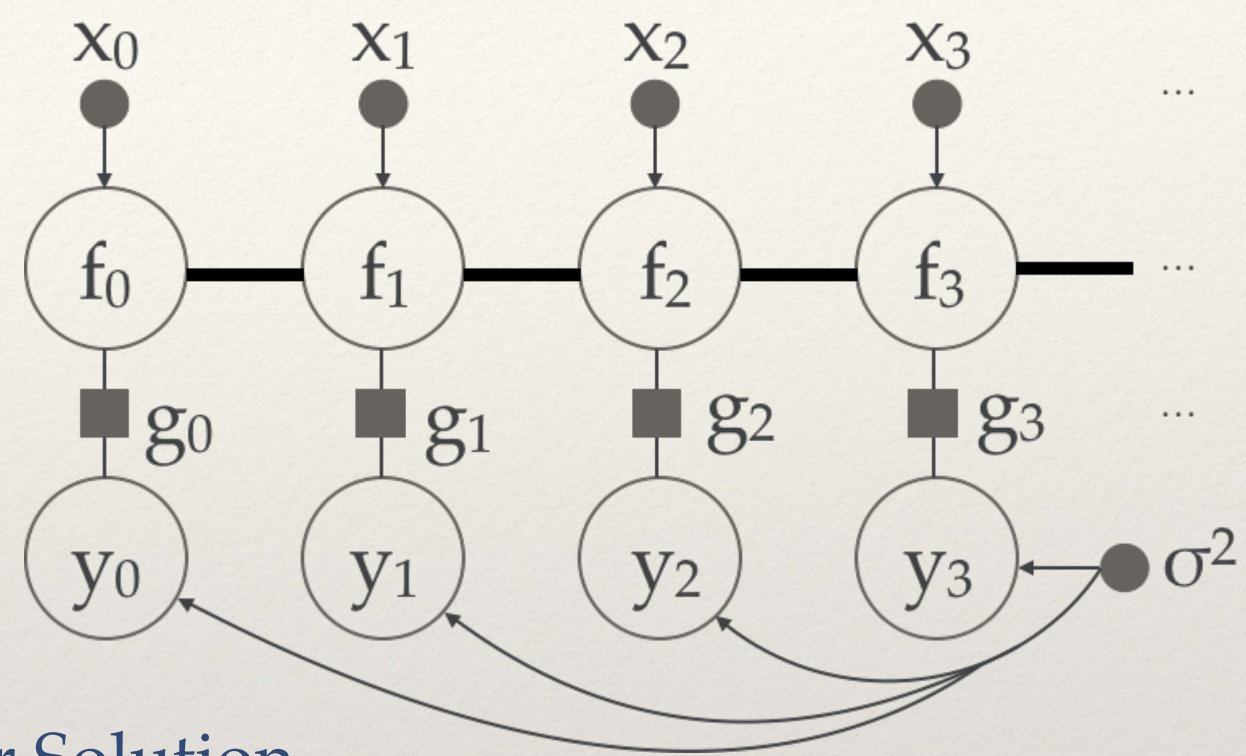
Extended and Unscented Kitchen Sinks

Edwin V. Bonilla, UNSW
Daniel Steinberg, NICTA
Alistair Reid, NICTA



Gaussian Process (GP) Models

- Models of the form $y = g(f) + \varepsilon$, where f is drawn from a GP
  - Standard supervised learning
  - Inversion problems



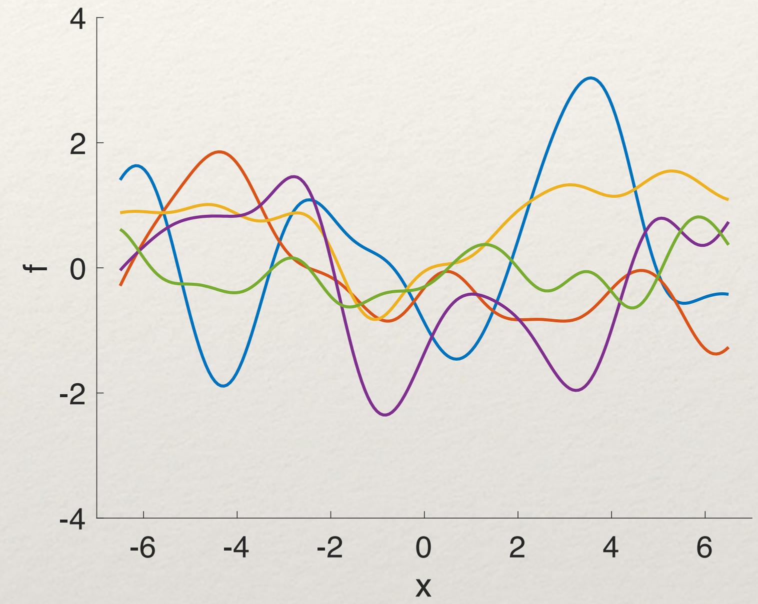
- Key Challenges
  1. Scalability on the number of observations
  2. Multi-task settings
  3. Nonlinear likelihoods

- Our Solution
  - Random feature approximations to the covariance function
  - Affine transformations of latent processes
  - Local and adaptive linearizations

All within a single variational inference framework

Multi-output Setting

- Supervised learning: $\mathcal{D} = \{\mathbf{x}_n, \mathbf{y}_n\}_{n=1}^{N} = \{\mathbf{X}, \mathbf{Y}\}$
  $\mathbf{x}_n$: d-dimensional, $\mathbf{y}_n$: P-dimensional



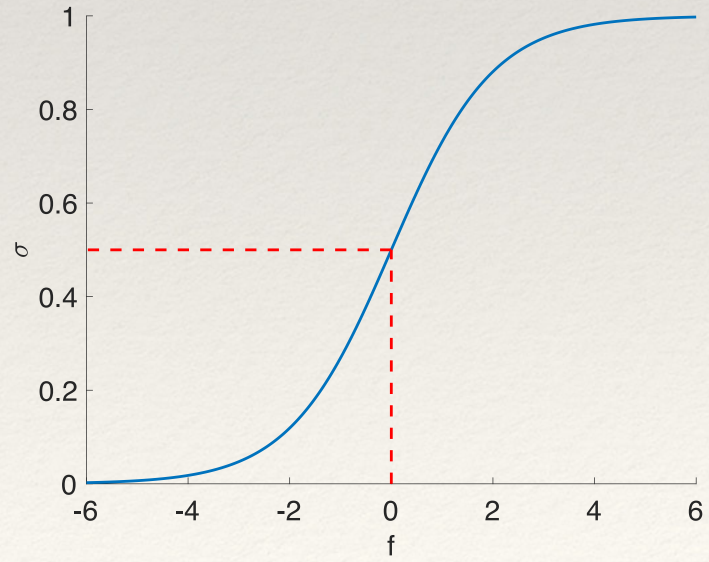
- Prior: Q latent functions

$$f_q \sim \mathcal{GP}(0, k_q(\cdot,\cdot))$$

$$p(\mathbf{F}) = \prod_{q=1}^{Q} \mathcal{N}(\mathbf{f}_{\cdot q}; \mathbf{0}, \mathbf{K}_q)$$

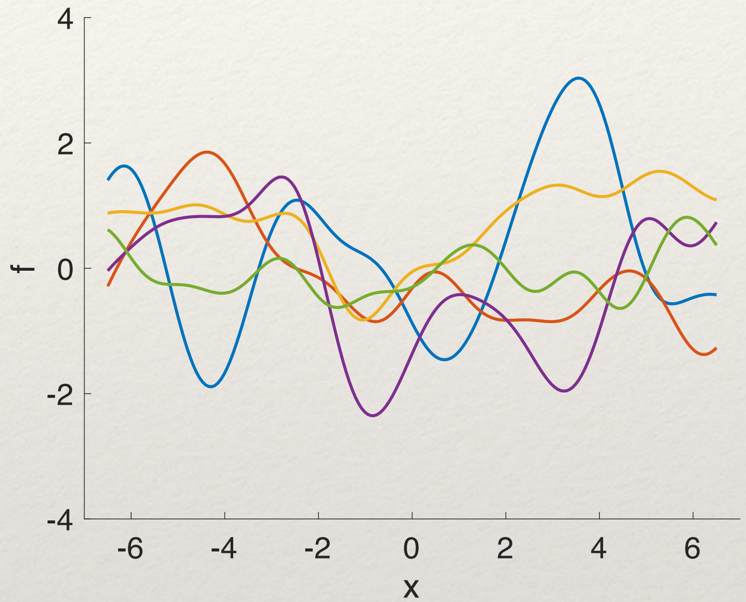
- Likelihood: For a given nonlinear forward model **g**(.)

$$p(\mathbf{Y}|\mathbf{F}) = \prod_{n=1}^{N} p(\mathbf{y}_n|\mathbf{g}(\mathbf{f}_{n\cdot}))$$



Multi-output Setting

- Supervised learning: $\mathcal{D} = \{\mathbf{x}_n, \mathbf{y}_n\}_{n=1}^{N} = \{\mathbf{X}, \mathbf{Y}\}$
  $\mathbf{x}_n$: d-dimensional, $\mathbf{y}_n$: P-dimensional



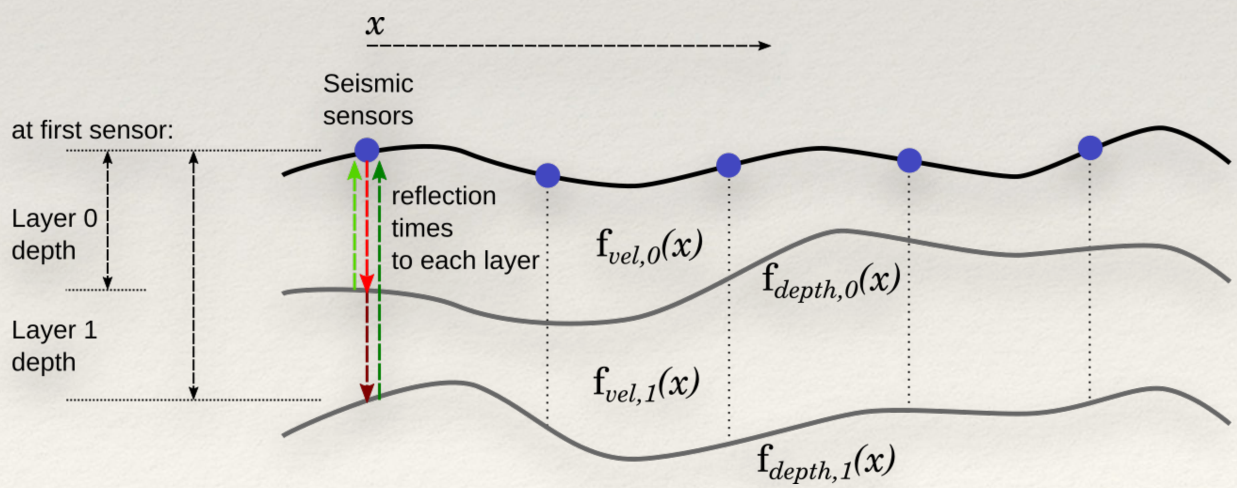
- Prior: Q latent functions

$$f_q \sim \mathcal{GP}(0, k_q(\cdot,\cdot))$$

$$p(\mathbf{F}) = \prod_{q=1}^{Q} \mathcal{N}(\mathbf{f}_{\cdot q}; \mathbf{0}, \mathbf{K}_q)$$

- Likelihood: For a given nonlinear forward model $\mathbf{g}(.)$

$$p(\mathbf{Y}|\mathbf{F}) = \prod_{n=1}^{N} p(\mathbf{y}_n|\mathbf{g}(\mathbf{f}_{n\cdot}))$$



- Goal: Probabilistic predictions and posterior estimation $p(\mathbf{F} \mid \mathbf{Y})$

Multi-output Setting

- Supervised learning: $\mathcal{D} = \{\mathbf{x}_n, \mathbf{y}_n\}_{n=1}^{N} = \{\mathbf{X}, \mathbf{Y}\}$
  $\mathbf{x}_n$: d-dimensional, $\mathbf{y}_n$: P-dimensional
- Prior: Q latent functions

$$f_q \sim \mathcal{GP}(0, k_q(\cdot, \cdot))$$

$$p(\mathbf{F}) = \prod_{q=1}^{Q} \mathcal{N}(\mathbf{f}_{\cdot q}; \mathbf{0}, \mathbf{K}_q)$$

- Likelihood: For a given nonlinear forward model **g**(.)

$$p(\mathbf{Y}|\mathbf{F}) = \prod_{n=1}^{N} p(\mathbf{y}_n|\mathbf{g}(\mathbf{f}_{n\cdot}))$$

Extended and Unscented GPs
(Steinberg & Bonilla, NIPS 2014)

Variational inference based on linearization of $g(f_n)$

- [x] Linearization is *local* and *adaptive*
- [ ] Multi-output?, Q=1
- [ ] Scalable inference?, $O(N^3)$ in time

Random Kitchen Sinks

- Goal: Probabilistic predictions and posterior estimation $p(\mathbf{F} \mid \mathbf{Y})$

Random Kitchen Sinks (RKS)

(Rahimi and B. Recht , NIPS 2008)

- Fourier duality of the covariance function of a stationary process and its spectral density:

$$k(\boldsymbol{\tau}) = \int S(\mathbf{s}) e^{2\pi i \mathbf{s}^T \boldsymbol{\tau}} d\mathbf{s} \longleftrightarrow S(\mathbf{s}) = \int k(\boldsymbol{\tau}) e^{-2\pi i \mathbf{s}^T \boldsymbol{\tau}} d\boldsymbol{\tau}$$

  - Approximate *k(τ)* by explicitly constructing "suitable" random features and (Monte Carlo) averaging over samples

$$k(\mathbf{x} - \mathbf{x}') = k(\boldsymbol{\tau}) \approx \frac{1}{D} \sum_{i=1}^{D} \phi_i(\mathbf{x})\, \phi_i(\mathbf{x}')$$

Use RKS bases to approximate GP model

Example:

$$\mathbf{s}_i \sim \mathcal{N}\left(\mathbf{s}_i \middle| \mathbf{0}, \sigma_\phi^2 \mathbf{I}_d\right)$$

$$[\phi_i(\mathbf{x})\,, \phi_{D+i}(\mathbf{x})] = \frac{1}{\sqrt{D}} [\cos(2\pi \mathbf{s}_i^T \mathbf{x}), \sin(2\pi \mathbf{s}_i^T \mathbf{x})]$$

Converges in expectation to the (isotropic) squared exponential kernel

Approximate Model

- Using RKS bases, we approximate our GP model

Prior variance over weights

$$p(\mathbf{W}) = \prod_{q=1}^{Q} \mathcal{N}\left(\mathbf{w}_q | \mathbf{0}, \omega_q^2 \mathbf{I}_D\right)$$

QxD weight matrix

$$p(\mathbf{Y}|\mathbf{W}) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{y}_n | \mathbf{g}(\mathbf{W}\boldsymbol{\phi}_n), \boldsymbol{\Sigma})$$

D-dimensional feature vector

Noise variance

  - Effectively, $\mathbf{f}_q \simeq \mathbf{\Phi}\mathbf{w}_q$, where $\mathbf{\Phi}$ is the NxD feature matrix

- Approximate inference due to nonlinear $\mathbf{g}(.)$

$$\tilde{q}_{\mathbf{W}} \stackrel{\text{def}}{=} \prod_{q=1}^{Q} \mathcal{N}(\mathbf{w}_q | \mathbf{m}_q, \mathbf{C}_q)$$

Variational posterior

- The evidence lower bound (ELBO) involves:

$$\left\langle (\mathbf{y}_n - \mathbf{g}(\mathbf{W}\boldsymbol{\phi}_n))^{\top} \boldsymbol{\Sigma}^{-1} (\mathbf{y}_n - \mathbf{g}(\mathbf{W}\boldsymbol{\phi}_n)) \right\rangle_{\tilde{q}_{\mathbf{W}}}$$

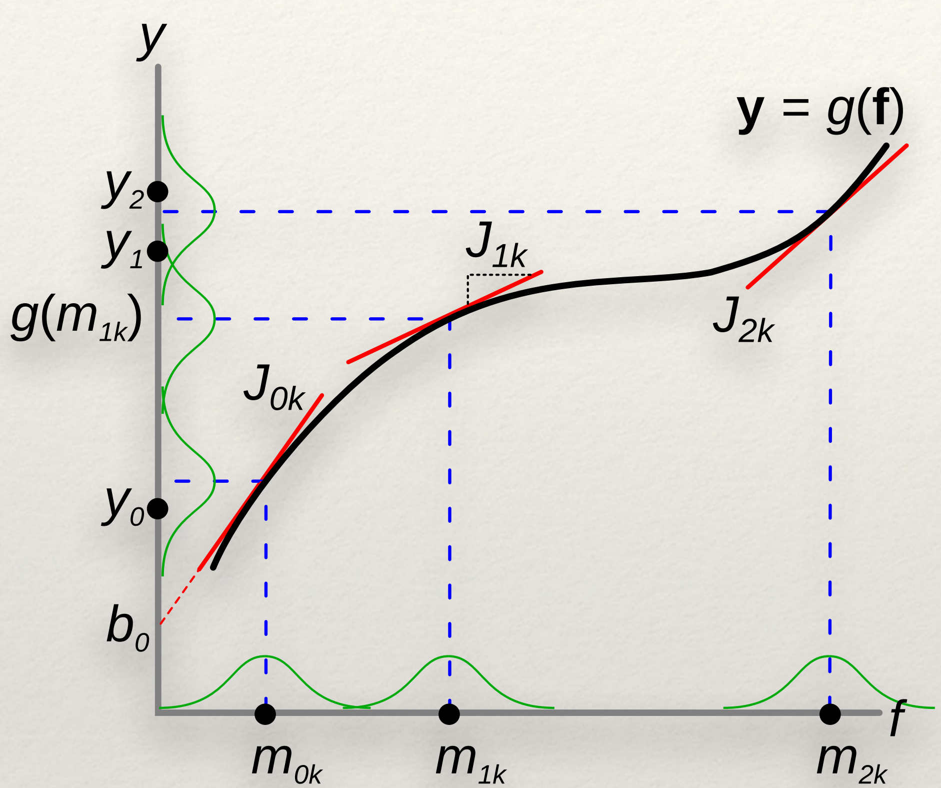
  - For which we make:
$\mathbf{g}(\mathbf{W}\boldsymbol{\phi}_n) \approx \mathbf{A}_n \mathbf{W}\boldsymbol{\phi}_n + \mathbf{b}_n$

☑ Unlike original EGP/UGP, inference scales up to large N
- Objective amenable to parallel / stochastic optimization

How to linearize (estimate A_n, b_n)? —> Extended vs Unscented

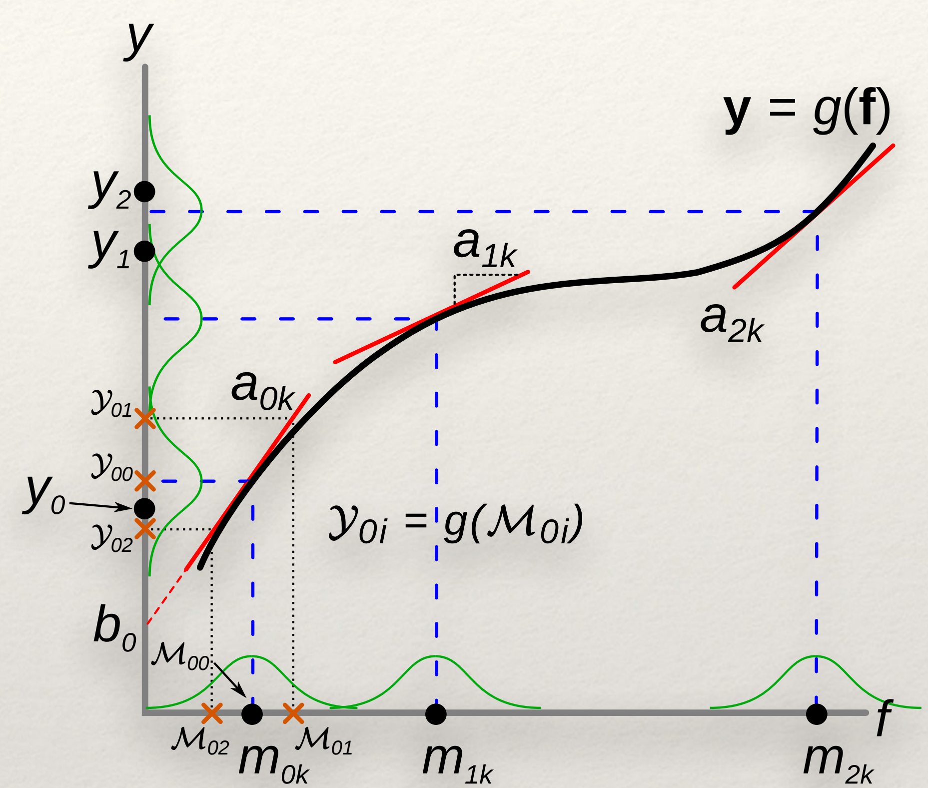
Extended or Unscented?

- Extended Kitchen Sinks (EKS)



- First-order Taylor expansion around the posterior mean $\bar{\mathbf{f}}_{n\cdot} = \mathbf{M}\phi_n$
  - Requires Jacobian estimation

- Unscented Kitchen Sinks (UKS)



- Fits a linear model using deterministic samples given by the Unscented Transform
  - Exploits structure of the posterior
  - 'black-box' method

- Both methods are *local* (datapoint-dependent) and *adaptive* (updated according to the current posterior estimate)

Experiments – Classification

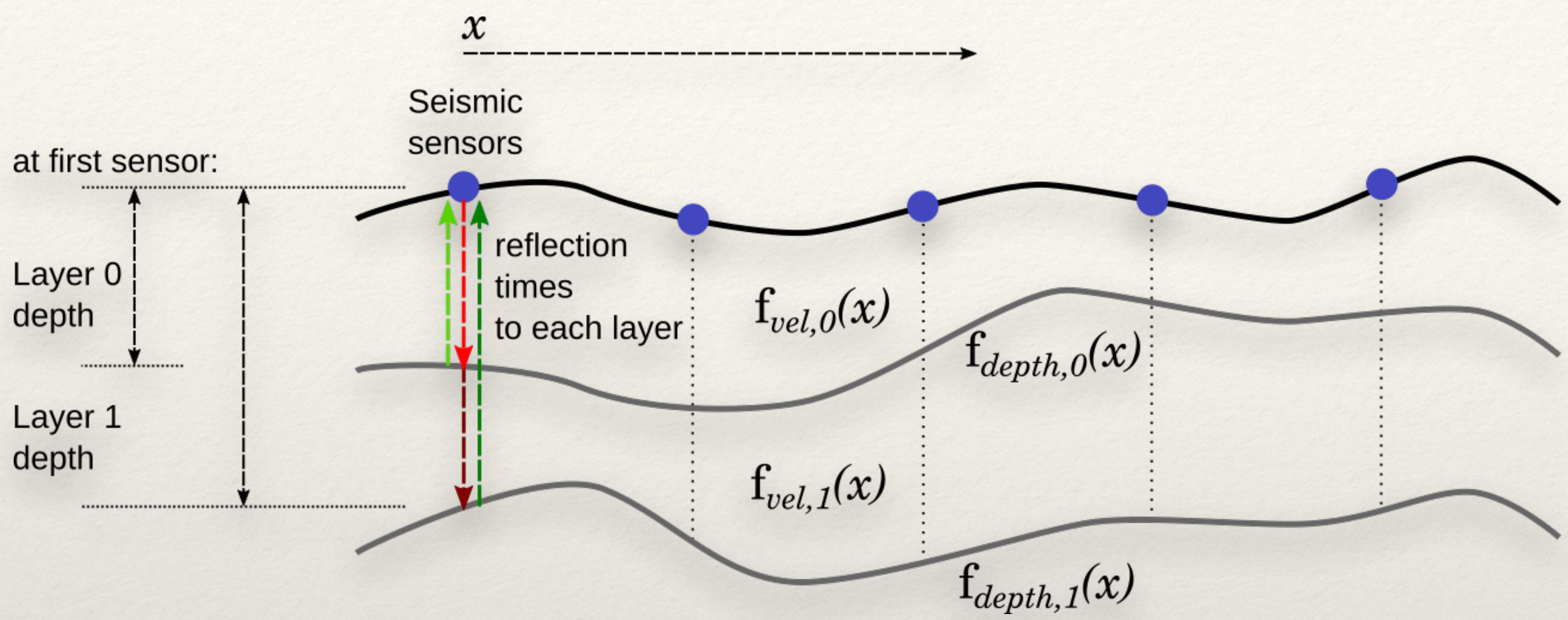
Odd Digits vs Even Digits on MNIST

	NLP		Error Rate	
	D=1000	D=2000	D=1000	D=2000
EKS	0.129	0.088	0.043	0.026
UKS	0.129	0.088	0.043	0.026
HMG [3]	0.069		0.022	
DB [4]	0.068		0.022	

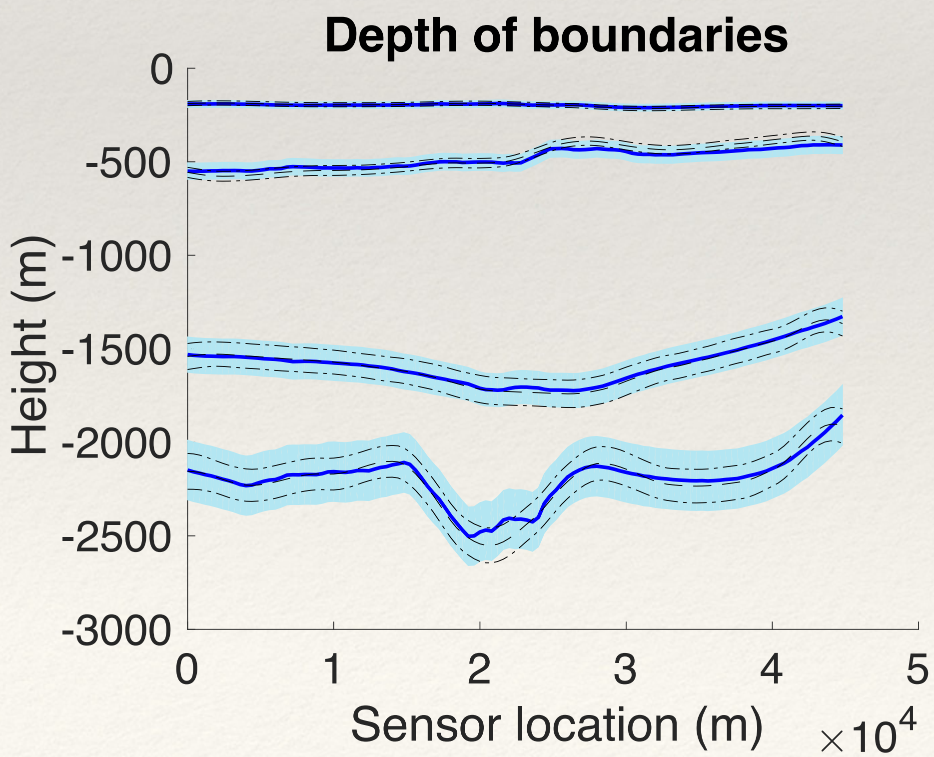
Similar performance to recently developed inducing-point approximations

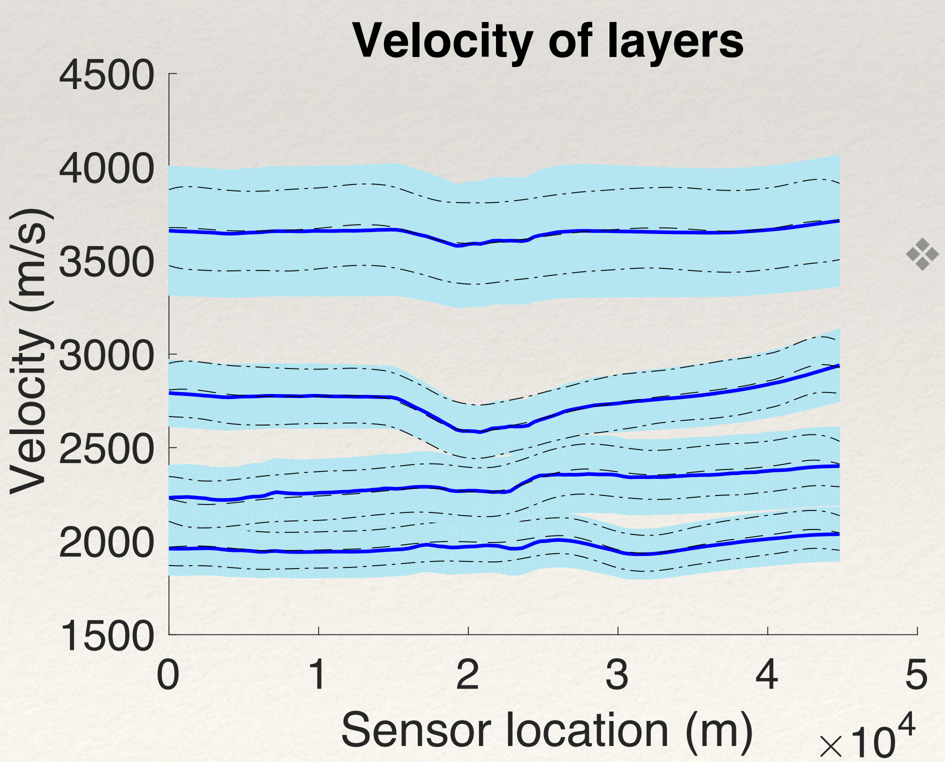
Experiments – Seismic Inversion

(Otway Basin, Australia)



- Goal: Infer geometry of layers and seismic velocity from sound reflection times





- Similar solution to long-running MCMC simulation

Conclusion & Discussion

- EKS and UKS: scalable methods for approximate inference in GP models with nonlinear likelihoods
  - UKS is a `black-box' method
- By using RKS-based approximations we can achieve similar performance to EGP and UGP but at a significantly lower computational cost
- Algorithms useful for inversion problems as fast and scalable alternatives to MCMC
  - Approximate models no longer GPs so can further investigate sampling approaches
  - More complex posteriors and stochastic optimizers

References

- [1] D. M. Steinberg and E. V. Bonilla, “Extended and unscented Gaussian processes”, in NIPS, 2014.
- [2] A. Rahimi and B. Recht, “Random features for large-scale kernel machines”, in NIPS, 2008.
- [3] J. Hensman, A. Matthews, and Z. Ghahramani, “Scalable variational Gaussian process classification”, in AISTATS, 2015.
- [4] A. Dezfouli and E. V. Bonilla, “Scalable inference for Gaussian process models with black-box likelihoods”, in NIPS, 2015.