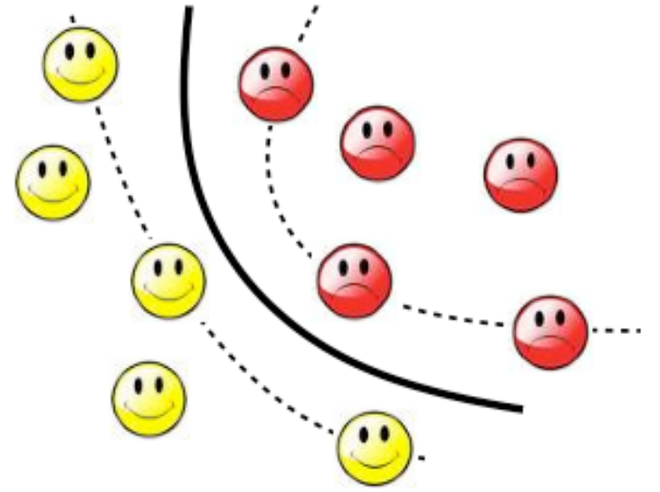
THE UNIVERSITY OF SYDNEY



Machine Learning and Data Mining (COMP 5318)

Bayesian Linear Regression And Gaussian Processes

Roman Marchant

THE UNIVERSITY OF SYDNEY

Recap

Linear Regression

Modelling Noisy Observations

Lets assume observations from a deterministic function with added Gaussian noise.

$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon \qquad \text{with} \qquad \epsilon \sim \mathcal{N}(0, \beta^{-1})$$

equivalently,

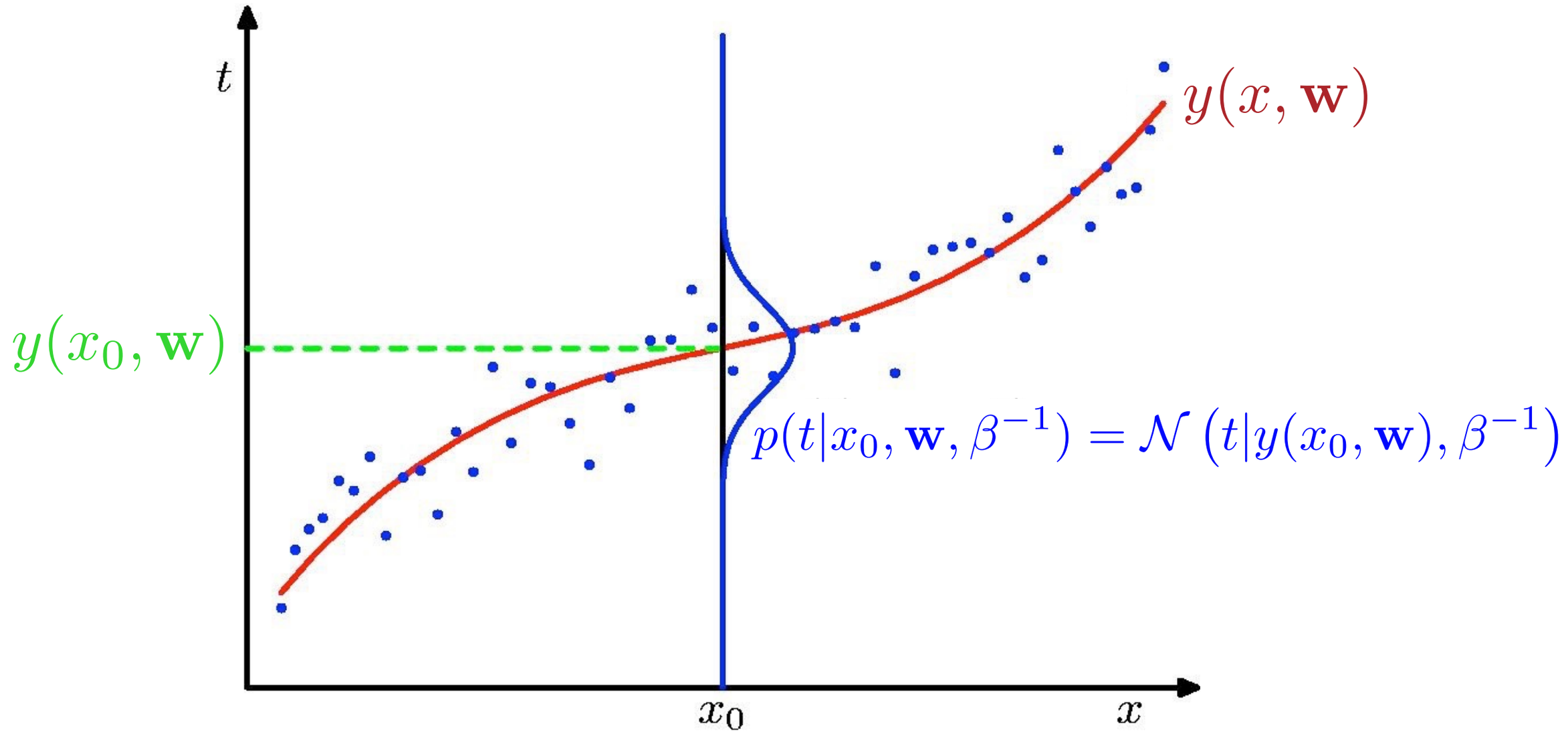
$$p(t|\mathbf{x}, \mathbf{w}, \beta^{-1}) = \mathcal{N}\left(t|y(\mathbf{x}, \mathbf{w}), \beta^{-1}\right)$$

Given the training data: $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ $\boldsymbol{\tau} = \{t_1, ..., t_N\}$

The expression of the likelihood of the iid data given the model is:

$$p(\boldsymbol{\tau}|\mathbf{X}, \mathbf{w}, \beta^{-1}) = \prod_{n=1}^{N} \mathcal{N}\left(t_n|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}\right)$$

Modelling Noisy Observations



Maximum Likelihood

THE UNIVERSITY OF SYDNEY

$$\mathbf{w}_{\mathrm{ML}} = \underset{\mathbf{w}}{\operatorname{argmax}} \ \ln p(\boldsymbol{\tau}|\mathbf{X}, \mathbf{w}, \beta^{-1})$$

$$\Rightarrow \qquad \left.\frac{\partial \ln p(\boldsymbol{\tau}|\mathbf{w}, \beta^{-1})}{\partial \mathbf{w}}\right|_{\mathbf{w}_{\mathrm{ML}}} = 0$$

$$\sum_{n=1}^{N} \left(t_n - \mathbf{w}_{\mathrm{ML}}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n)\right) \boldsymbol{\phi}(\mathbf{x}_n)^{\mathrm{T}} = 0$$

Solving for $\mathbf{w}_{\mathrm{ML}}$

$$\mathbf{w}_{\mathrm{ML}} = \left(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1} \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\tau}$$

where

$$\boldsymbol{\Phi} = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \cdots & \phi_{M-1}(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N) \end{pmatrix}.$$

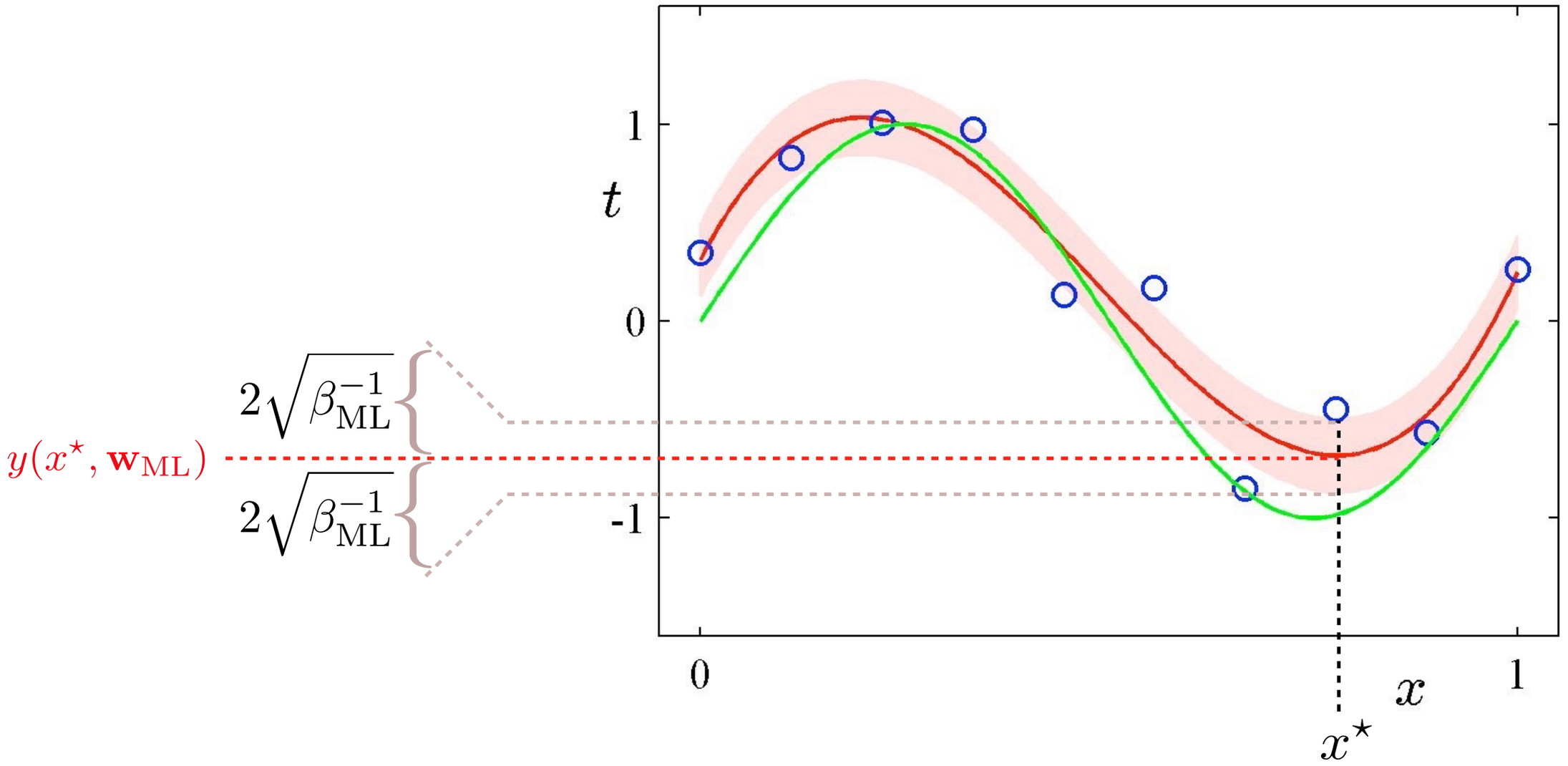
Maximum Likelihood

We can also maximise the log likelihood with respect to the noise precision parameter β .

$$\frac{1}{\beta_{\mathrm{ML}}} = \frac{1}{N}\sum_{n=1}^{N}\left(t_n - \mathbf{w}_{\mathrm{ML}}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\right)^2$$

Predictive Distribution

$$p(t|x, \mathbf{w}_{\mathrm{ML}}, \beta_{\mathrm{ML}}) = \mathcal{N}(t|y(x, \mathbf{w}_{\mathrm{ML}}), \beta_{\mathrm{ML}}^{-1})$$



THE UNIVERSITY OF SYDNEY

Bayesian Linear Regression

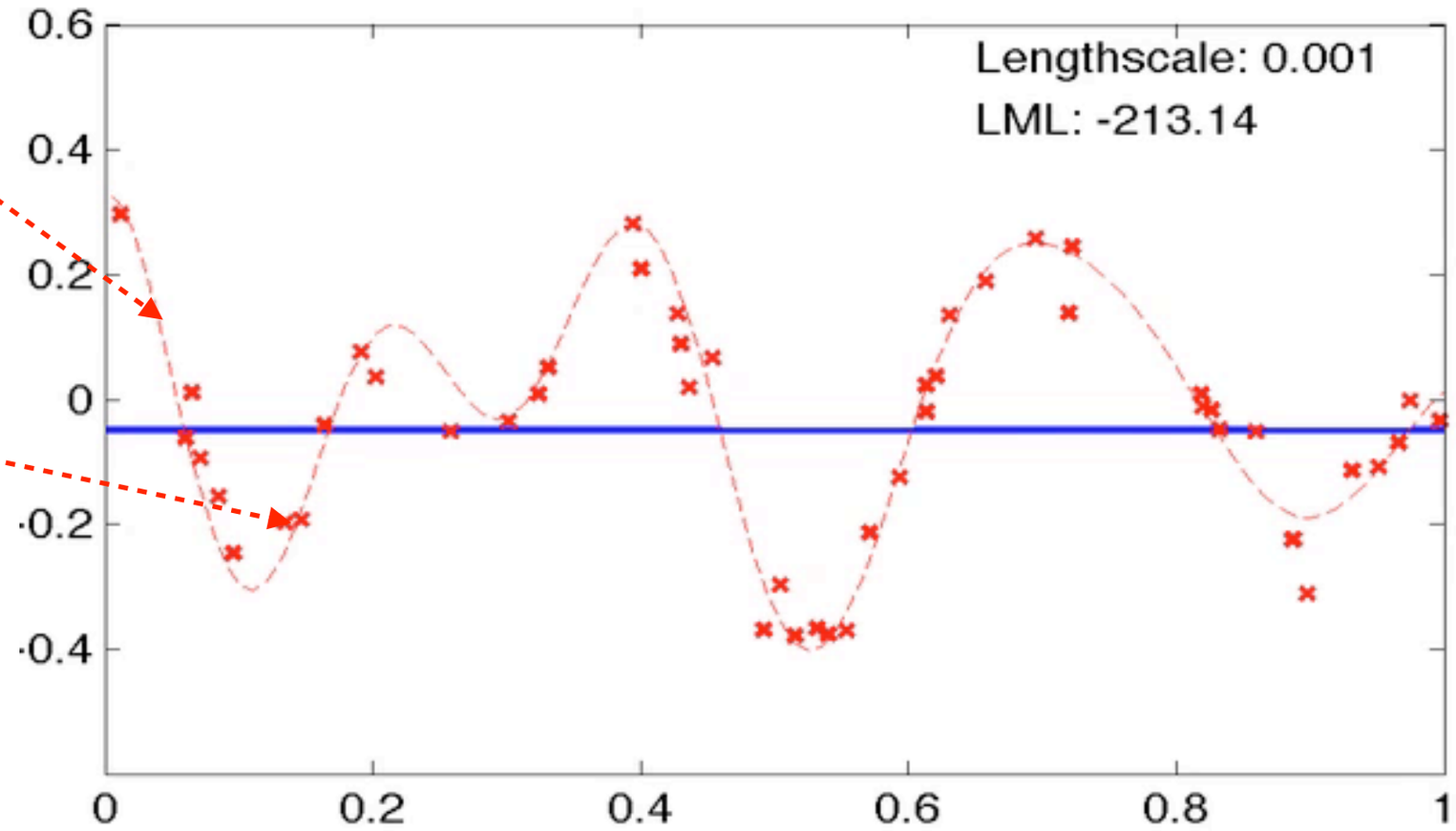
C. Bishop, *Pattern Recognition and Machine Learning,* Chapter 3: Linear Models for Regression
Springer New York, 2006

Learning as Optimisation

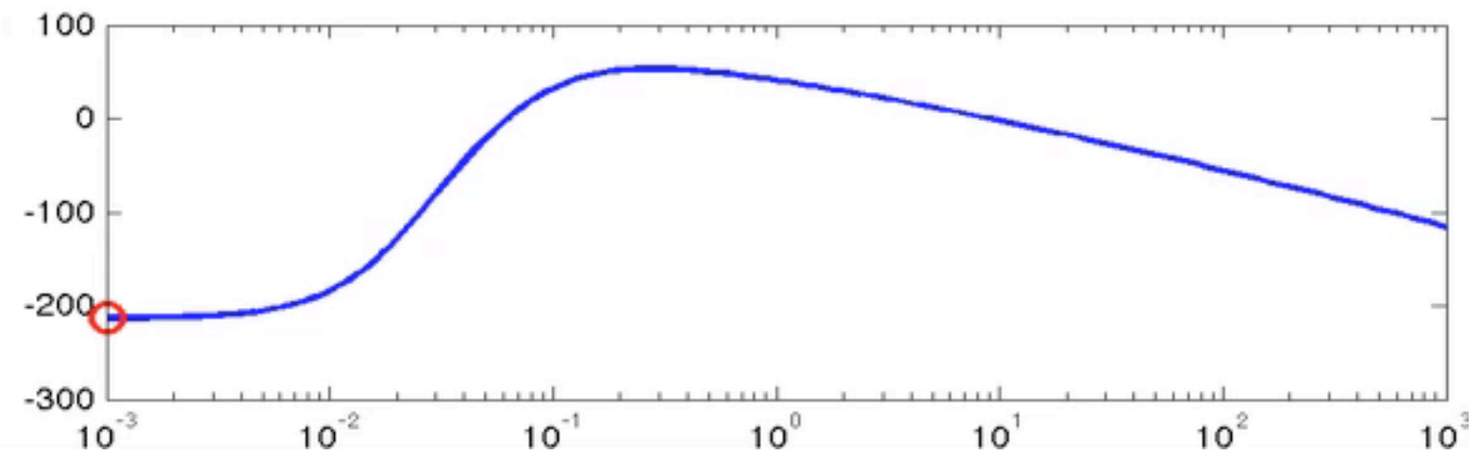
THE UNIVERSITY OF SYDNEY

Unknown Function
$f(\mathbf{x})$

Noisy Samples
from f



Log Marginal Likelihood



Deterministic vs Probabilistic Modelling

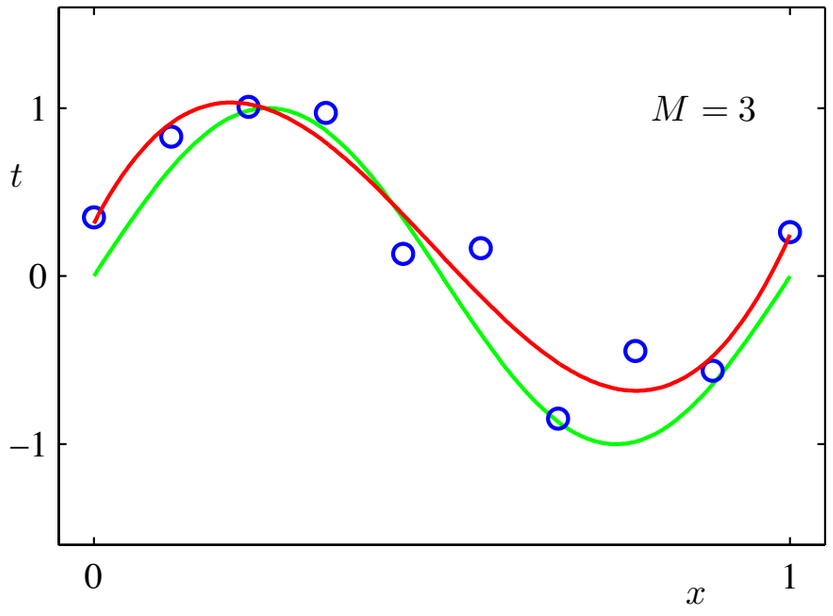
THE UNIVERSITY OF SYDNEY

Deterministic

Noisy Data

Ground Truth

Model Fit

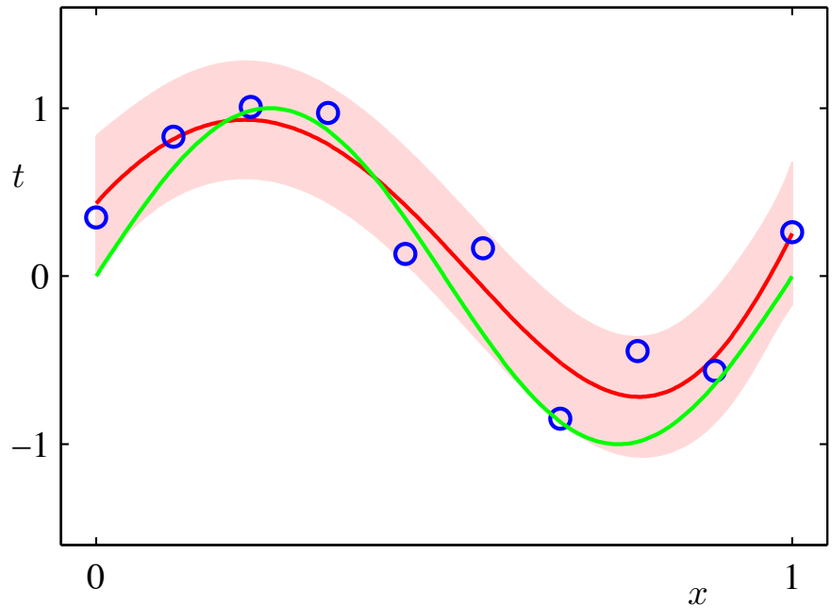


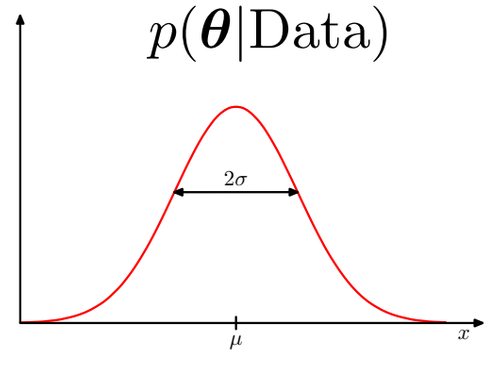
$$y(x, \boldsymbol{\theta}) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$$

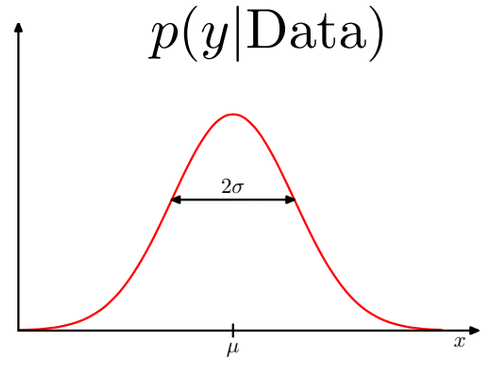
$$\boldsymbol{\theta}^\star = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\ E(\boldsymbol{\theta})$$

Probabilistic

Prediction Uncertainty







Bayes Teorem

Thomas
Bayes
1701–1761

THE UNIVERSITY OF SYDNEY

Likelihood **Prior**

$$\underbrace{p(\boldsymbol{\theta}|\mathcal{D})}_{\textbf{Posterior Distribution (Inference)}} \propto \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})}$$

Normalising Constant

*The data does not speak by itself.

Predictive Distribution

$$p(y|\mathcal{D}) = \int p(y|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D})\,\mathrm{d}\boldsymbol{\theta}$$

Summing over models.
(integrating over all possible model space)

Bayesian Linear Regression

Apply Bayes theorem to find posterior over parameters:

$$p(\mathbf{w}|\mathbf{X}, \boldsymbol{\tau}, \boldsymbol{\theta}, \beta) \propto p(\boldsymbol{\tau}|\mathbf{X}, \mathbf{w}, \beta)p(\mathbf{w}|\boldsymbol{\theta})$$

Define a conjugate prior (assuming gaussian likelihood):

$$p(\mathbf{w}|\boldsymbol{\theta}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0(\boldsymbol{\theta}), \mathbf{S}_0(\boldsymbol{\theta}))$$

$\boldsymbol{\theta}$ prior parameters (hyper-parameters).

Combining the likelihood and the prior provides the following analytical solution:

$$p(\mathbf{w}|\boldsymbol{\tau}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$$

where,

$$\mathbf{m}_N = \mathbf{S}_N \left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\tau}\right)$$
$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}$$

*Applying theorem from Bishop Section 2.3.3

Bayesian Linear Regression

THE UNIVERSITY OF SYDNEY

Example:

Prior: $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$

Previous Result

$$p(\mathbf{w}|\boldsymbol{\tau}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$$

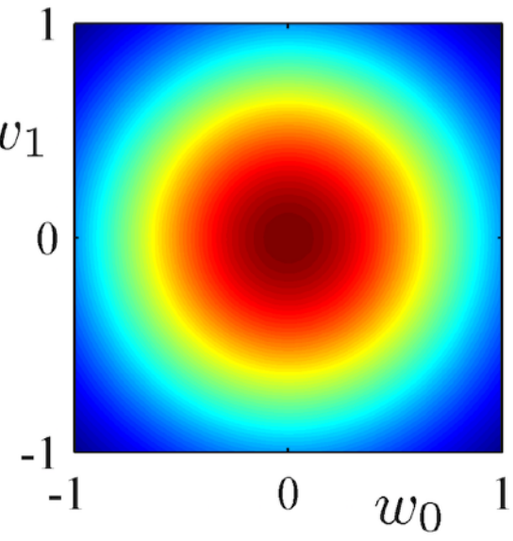
$$\mathbf{m}_N = \mathbf{S}_N\left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\tau}\right)$$

$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}$$

$$\mathbf{m}_0 = 0$$

$$\mathbf{S}_0 = \alpha^{-1}\mathbf{I}$$

2D Zero-mean Isotropic Gaussian



Posterior: $p(\mathbf{w}|\boldsymbol{\tau}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$

$$\mathbf{m}_N = \beta\mathbf{S}_N\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\tau}$$

$$\mathbf{S}_N^{-1} = \alpha\mathbf{I} + \beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}$$

$$\ln p(\mathbf{w}|\mathbf{t}) = -\frac{\beta}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2 - \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + \text{const.}$$

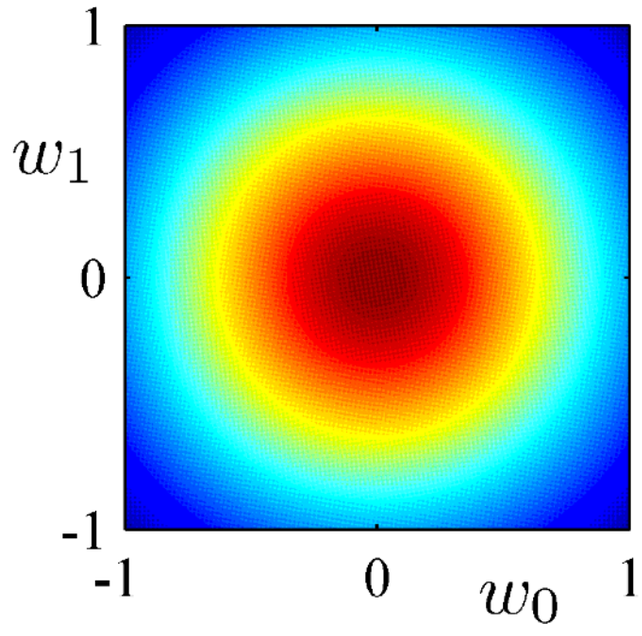
Bayesian Linear Regression

THE UNIVERSITY OF SYDNEY

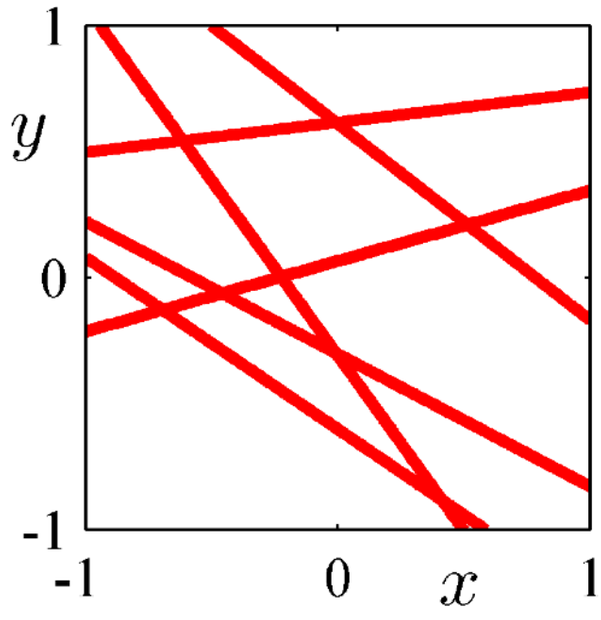
Example:

0 Points Observed

Prior



Data Space



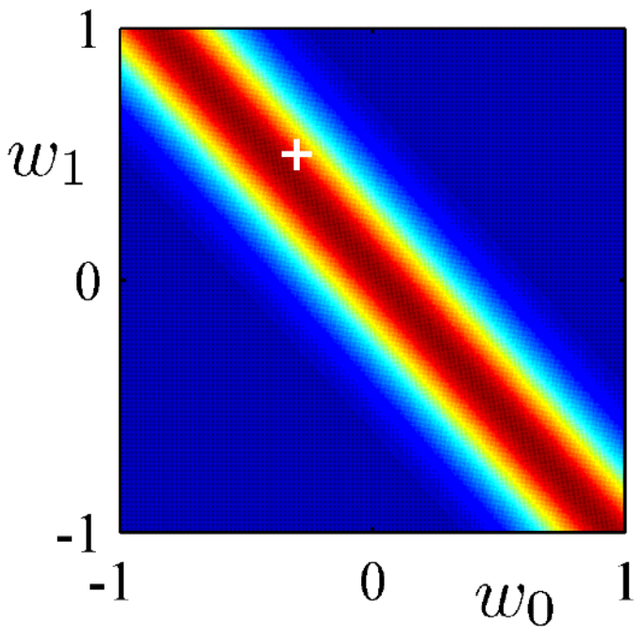
Bayesian Linear Regression

THE UNIVERSITY OF SYDNEY

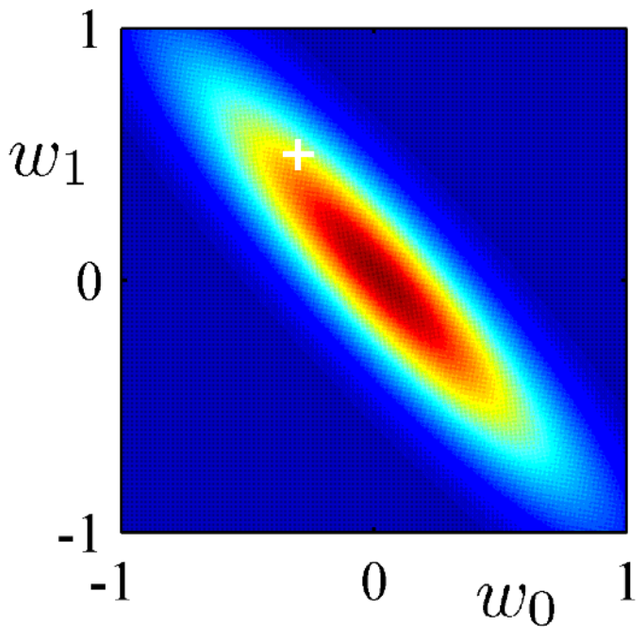
Example:

1 Point Observed

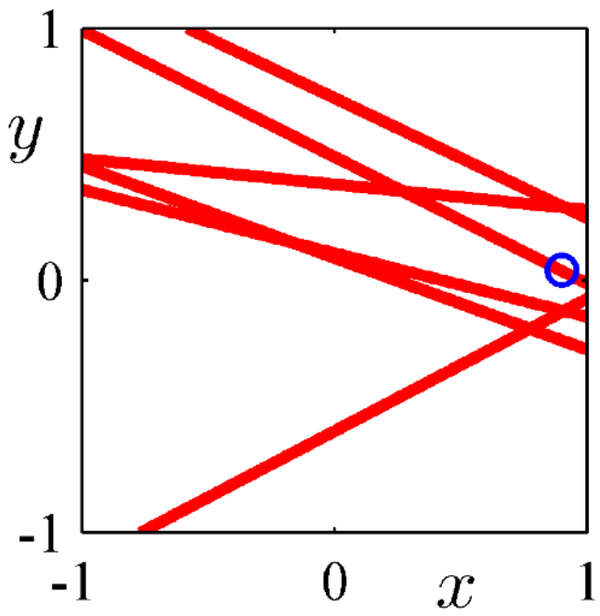
Likelihood



Posterior



Data Space

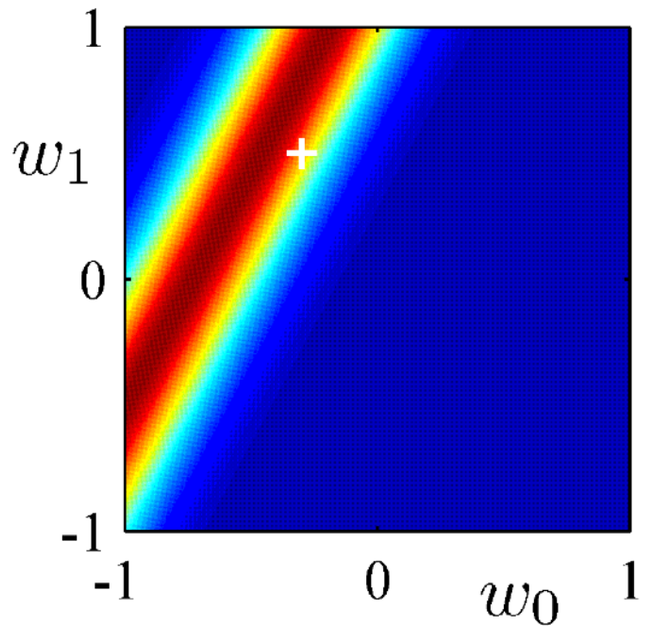


Bayesian Linear Regression

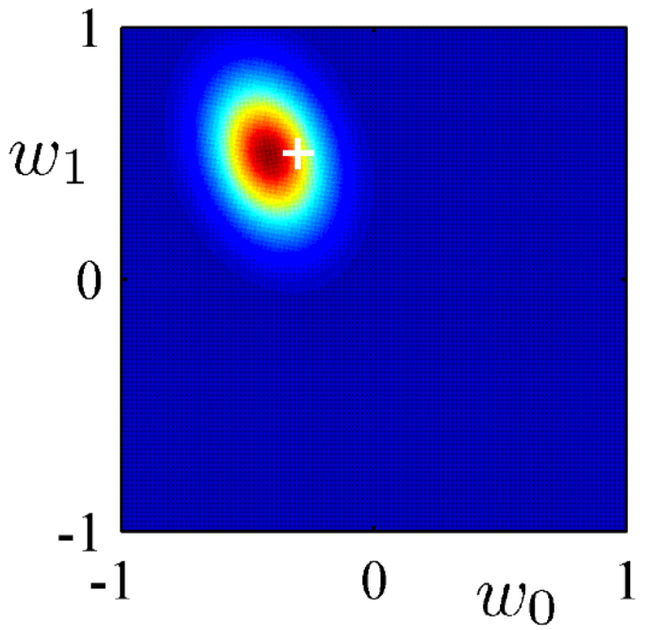
Example:

2 Points Observed

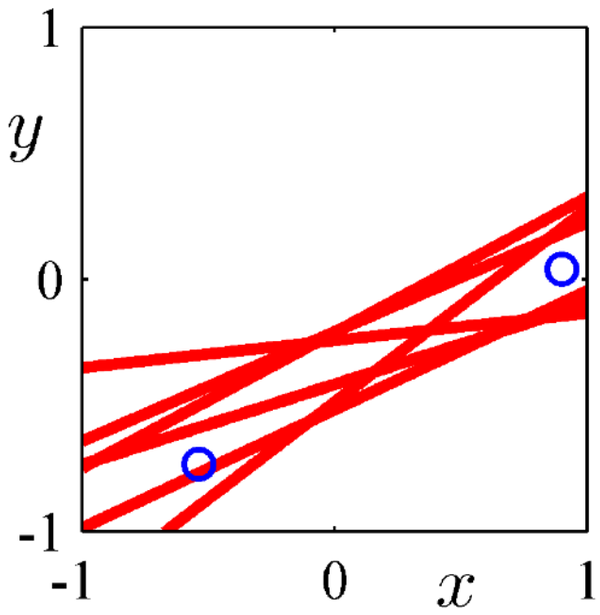
Likelihood



Posterior



Data Space

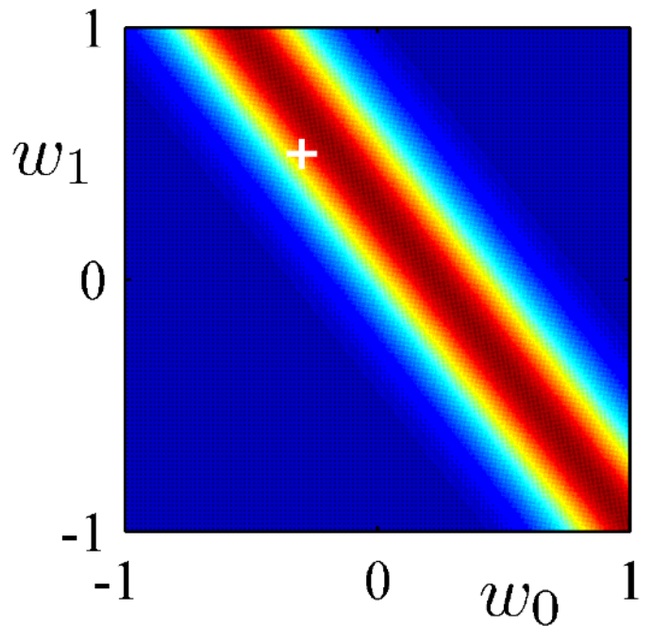


Bayesian Linear Regression

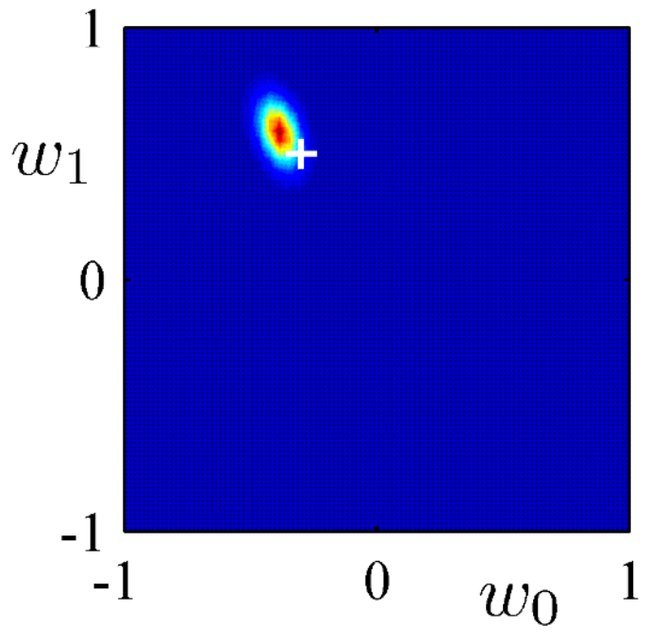
Example:

20 Points Observed

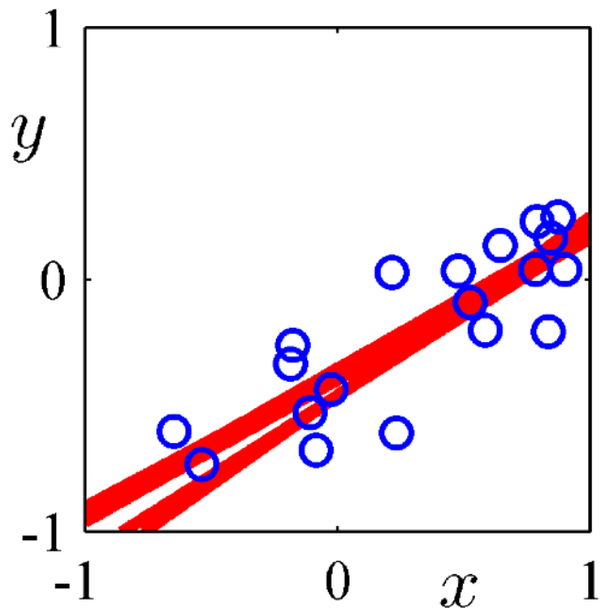
Likelihood



Posterior



Data Space



Predictive Distribution

Predict the value of t for new values of $\mathbf{x}$.

$$p(t|\boldsymbol{\tau}, \alpha, \beta) = \int p(t|\mathbf{w}, \beta) p(\mathbf{w}|\boldsymbol{\tau}, \alpha, \beta) \mathrm{d}\mathbf{w}$$

$$\mathcal{N}\left(t|y(\mathbf{x}, \mathbf{w}), \beta^{-1}\right) \qquad \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$$

$$\mathbf{m}_N = \mathbf{S}_N \left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta \boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\tau}\right)$$

$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta \boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi}$$

$$p(t|\boldsymbol{\tau}, \alpha, \beta) = \mathcal{N}\left(t|\mathbf{m}_N^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}), \sigma_N^2(\mathbf{x})\right)$$

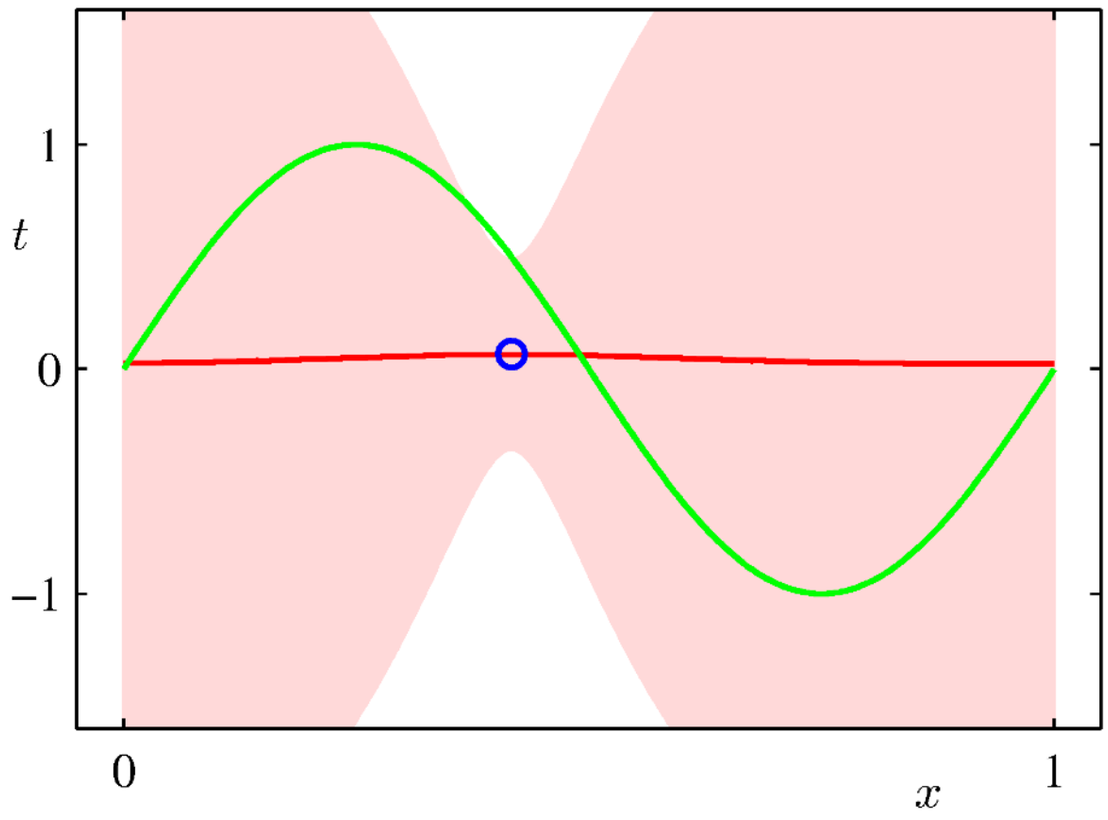
Where,

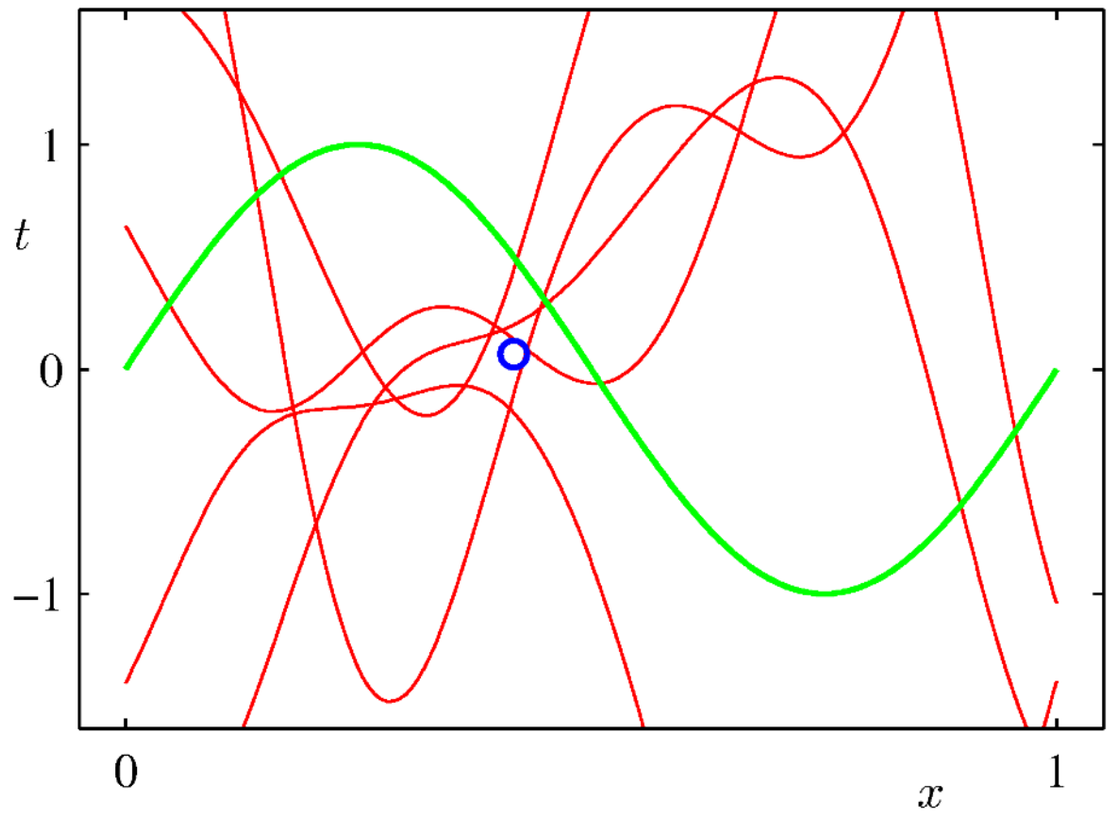
$$\sigma_N^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}^{\mathrm{T}}(\mathbf{x}) \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x})$$

Predictive Distribution

THE UNIVERSITY OF SYDNEY

Example: Sinusoidal Data, 9 Gaussian basis functions

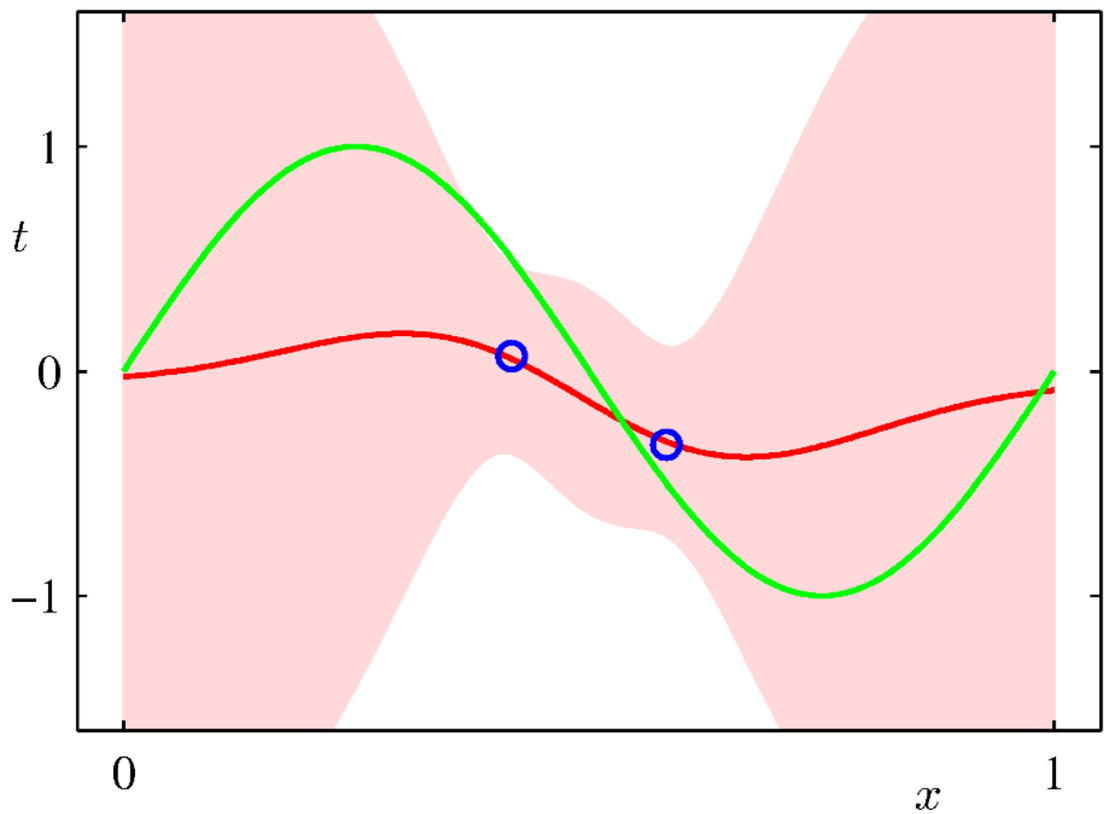


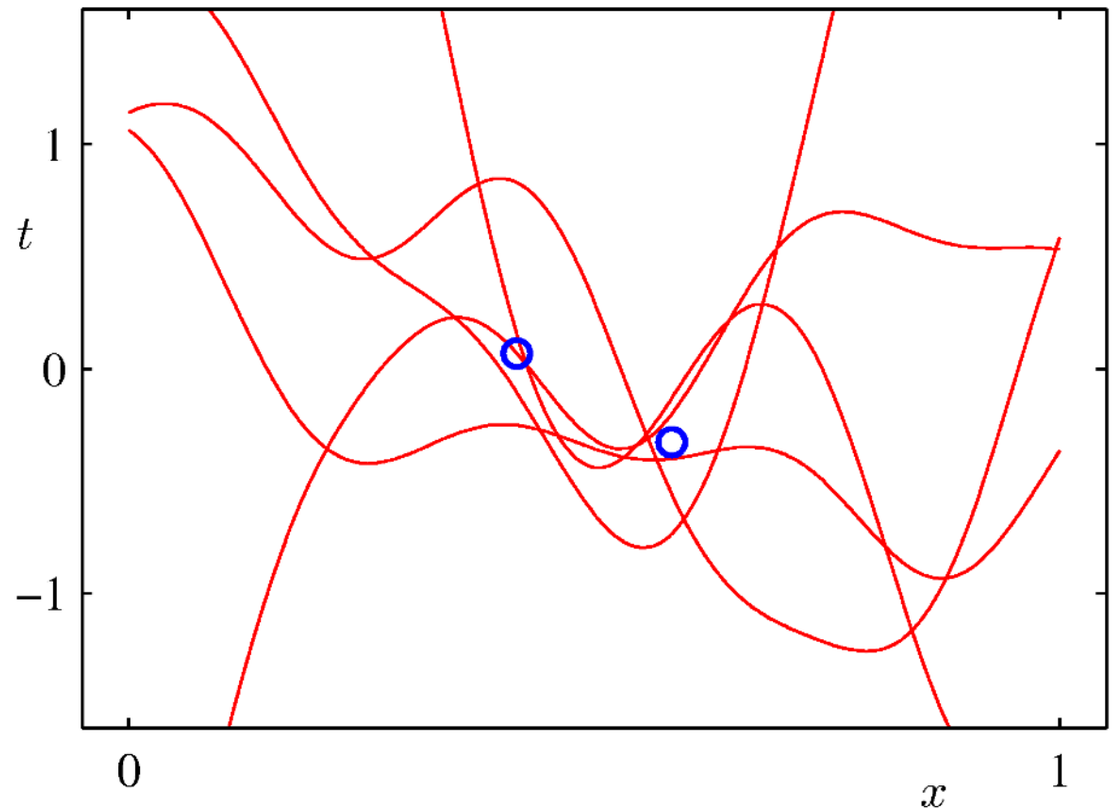


Predictive Distribution

THE UNIVERSITY OF SYDNEY

Example: Sinusoidal Data, 9 Gaussian basis functions

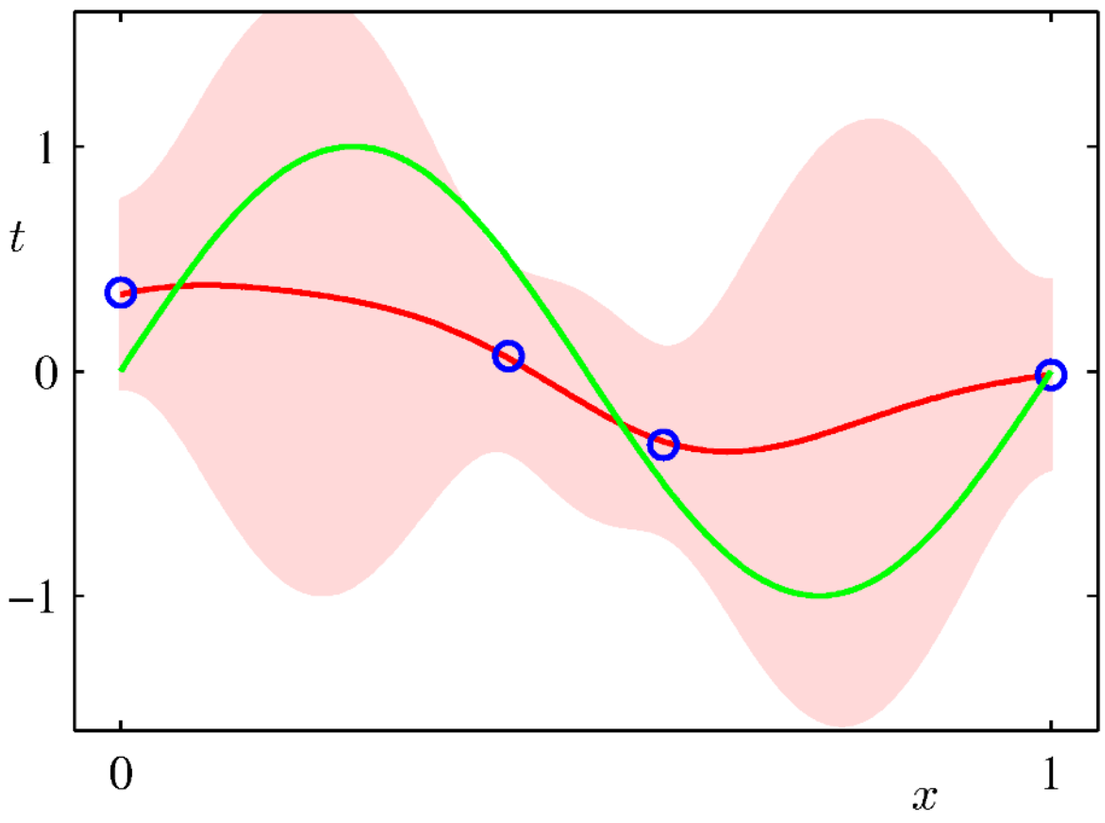


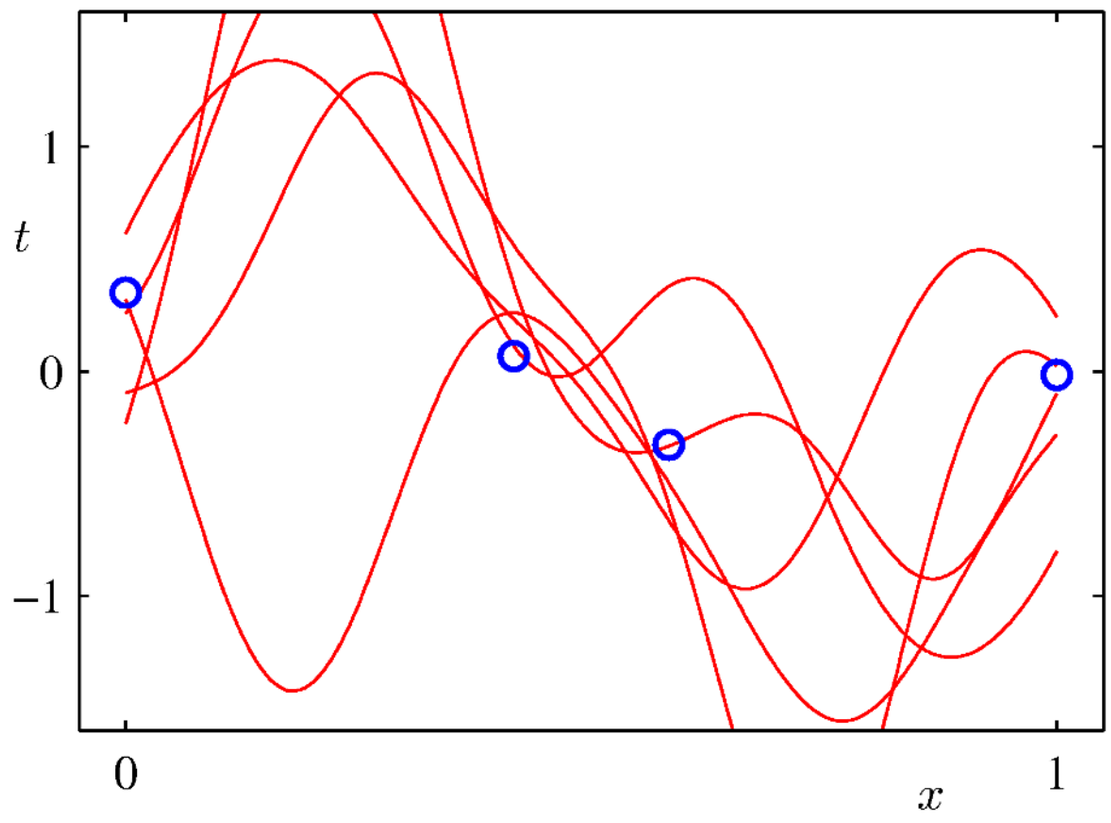


Predictive Distribution

THE UNIVERSITY OF SYDNEY

Example: Sinusoidal Data, 9 Gaussian basis functions

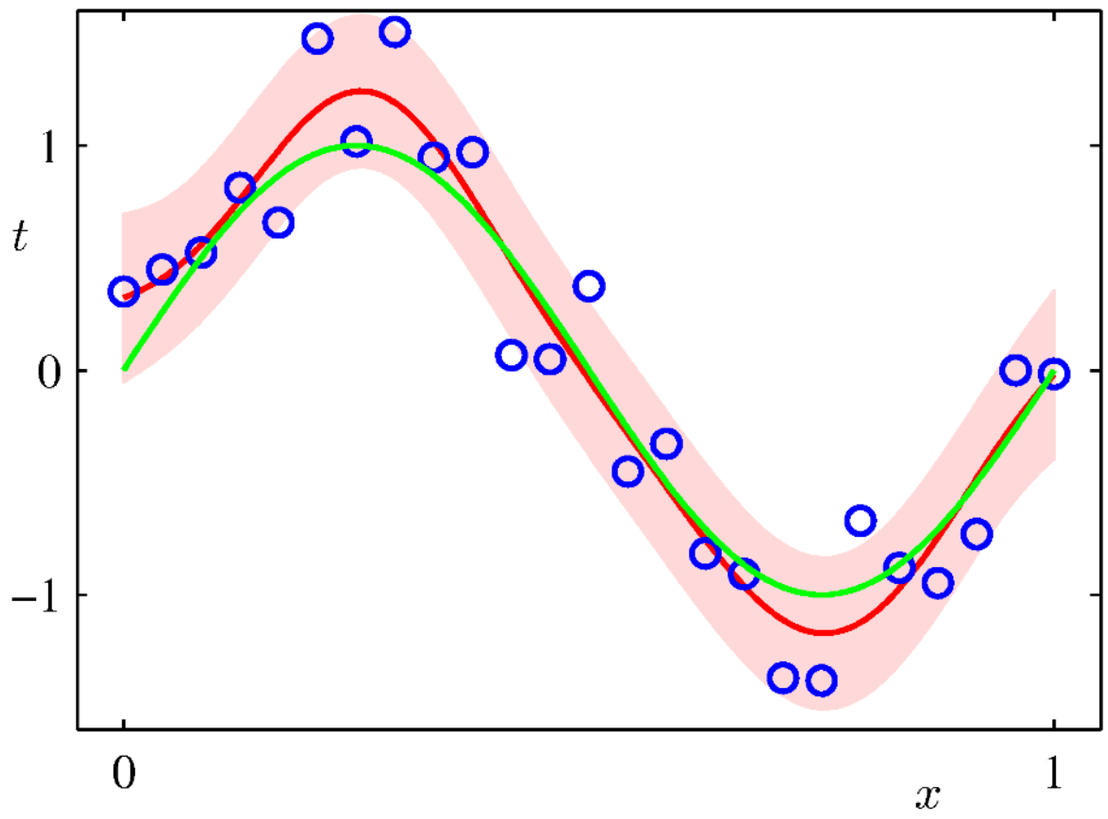


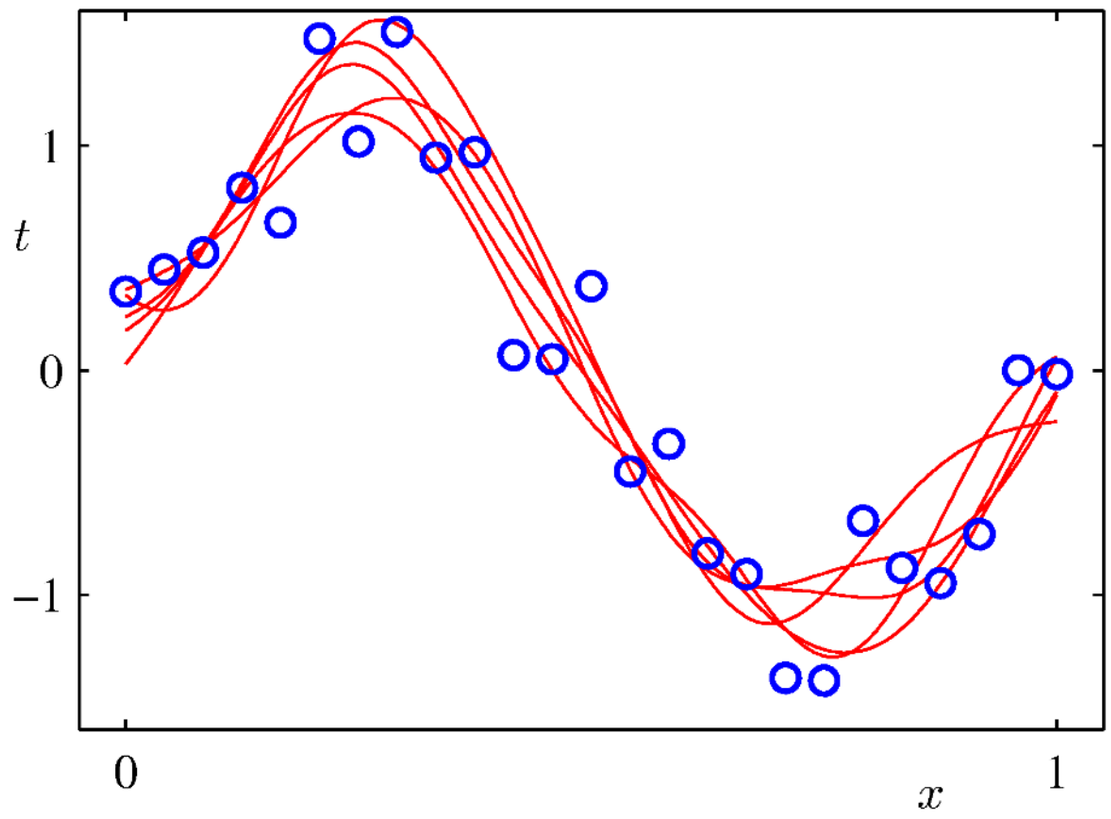


Predictive Distribution

THE UNIVERSITY OF SYDNEY

Example: Sinusoidal Data, 9 Gaussian basis functions





Equivalent Kernel

The predictive distribution:

$$p(t|\boldsymbol{\tau}, \alpha, \beta) = \mathcal{N}\left(t|\mathbf{m}_N^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}), \sigma_N^2(\mathbf{x})\right)$$

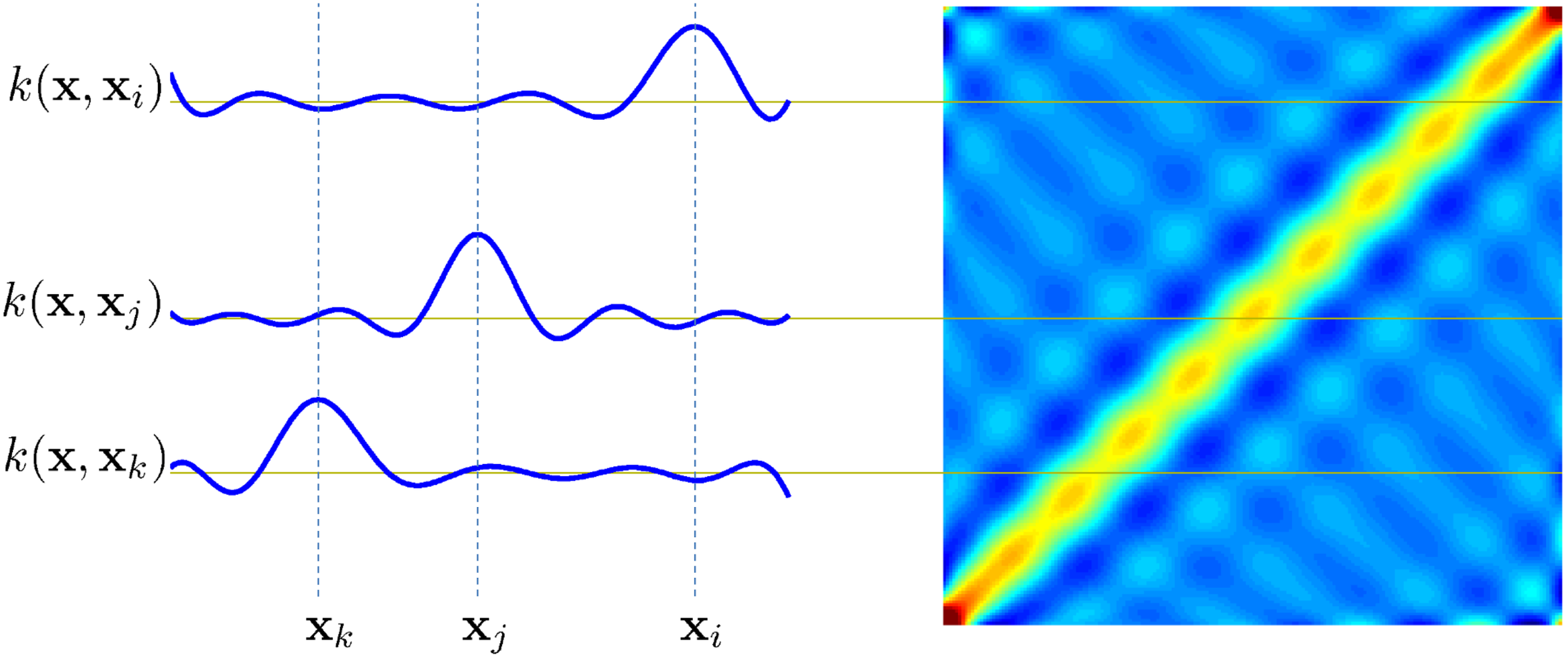
Rewriting the predictive mean:

$$\begin{aligned}\mathbf{m}_N^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}) &= \beta\boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\tau} \\ &= \sum_{n=1}^{N}\beta\boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}_n)t_n \\ &= \sum_{n=1}^{N}k(\mathbf{x}, \mathbf{x}_n)t_n.\end{aligned}$$

Equivalent kernel or *smoother matrix*.

This is a weighted sum of the training data target values.

Equivalent Kernel

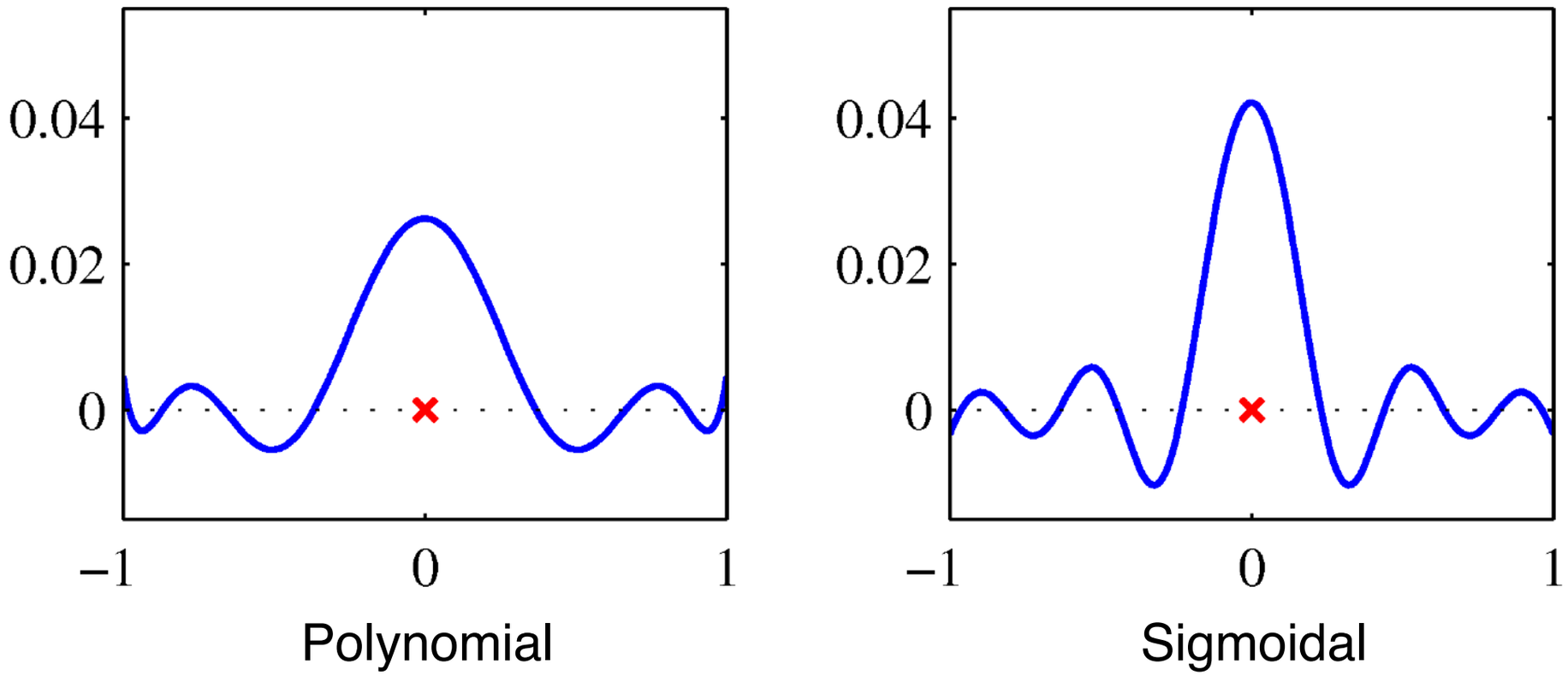


Weight of t_n depends on distance between x and x_n; nearby x_n carry more weight.

Equivalent Kernel

THE UNIVERSITY OF SYDNEY

Non-local basis functions have local equivalent



Equivalent Kernel

THE UNIVERSITY OF SYDNEY

The kernel as a covariance function: consider

$$\begin{aligned}\mathrm{cov}[y(\mathbf{x}), y(\mathbf{x}')] &= \mathrm{cov}[\boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{w}, \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}')] \\ &= \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}') = \beta^{-1}k(\mathbf{x}, \mathbf{x}').\end{aligned}$$

We can avoid the use of basis functions and define the kernel function directly, leading to Gaussian Processes.

Gaussian Processes

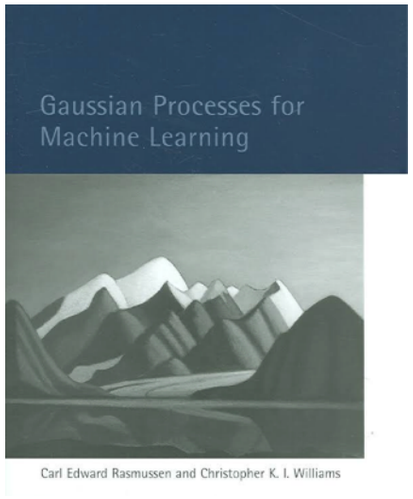
THE UNIVERSITY OF SYDNEY

Definition: A Gaussian process is a collection of random variables, any number of which have (consistent) Gaussian distributions.

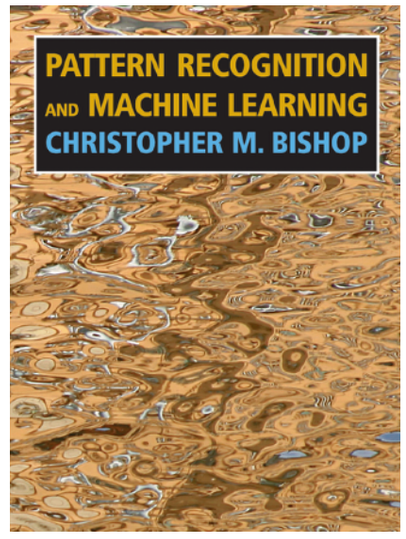
A GP is fully specified by a mean function $m(\mathbf{x})$ and covariance function $k(\mathbf{x}, \mathbf{x}')$:

$$f(\mathbf{x}) \sim \mathrm{GP}\left(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')\right)$$

Resources



C. E. Rasmussen and C. Williams, Gaussian processes for machine learning. The MIT Press, Cambridge, Massachuset, 2006.



Chapter 6.4: Kernel Methods
MB Christopher, Pattern Recognition and Machine Learning, Springer-Verlag New York, 2016.

Online resources:
www.gaussianprocess.org

GPy, Stan, gpml, TacoPig

Gaussian Process Regression

Given a set of samples

$$X = [\mathbf{x}_0, \ldots, \mathbf{x}_{N-1}]^T$$
$$\mathbf{y} = [y_0, \ldots, y_{N-1}]^T$$

Choose a covariance function

$$k(\mathbf{x}_i, \mathbf{x}_j \mid \theta)$$

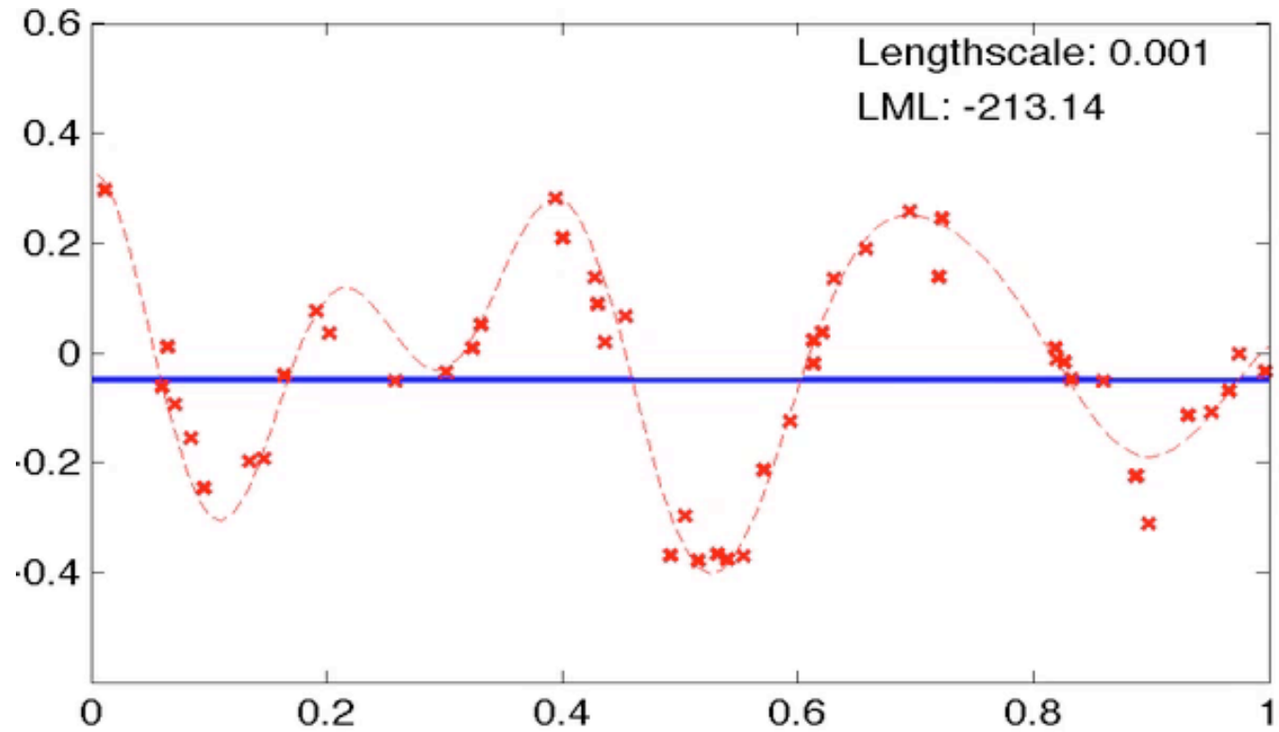
Predict the value of $f(\mathbf{x}_*)$

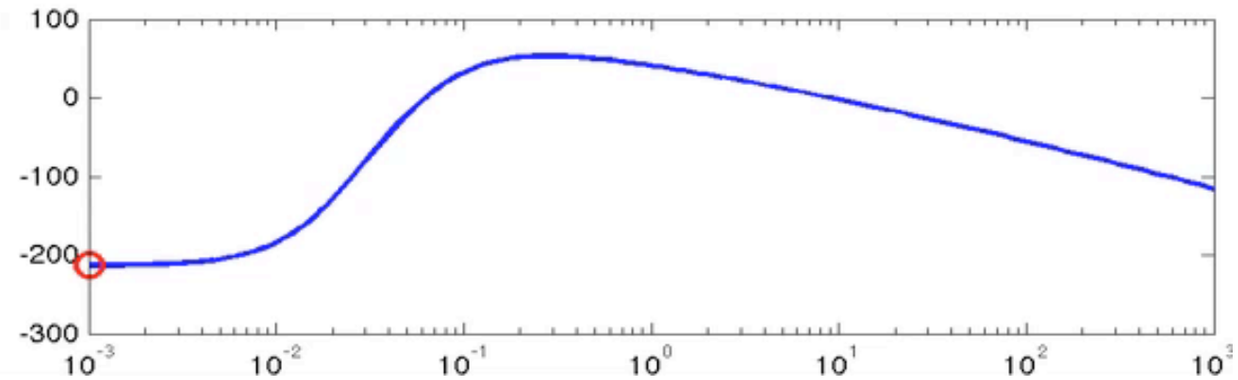
Predictive Posterior Distribution

Mean $\mu(f(\mathbf{x}_*))$

Variance $\sigma(f(\mathbf{x}_*))$

Regression





Log Marginal Likelihood

Gaussian Process Regression

$$y = f(\mathbf{x}) + \epsilon \qquad \text{, with} \qquad \epsilon \sim \mathcal{N}(0, \sigma_n^2)$$

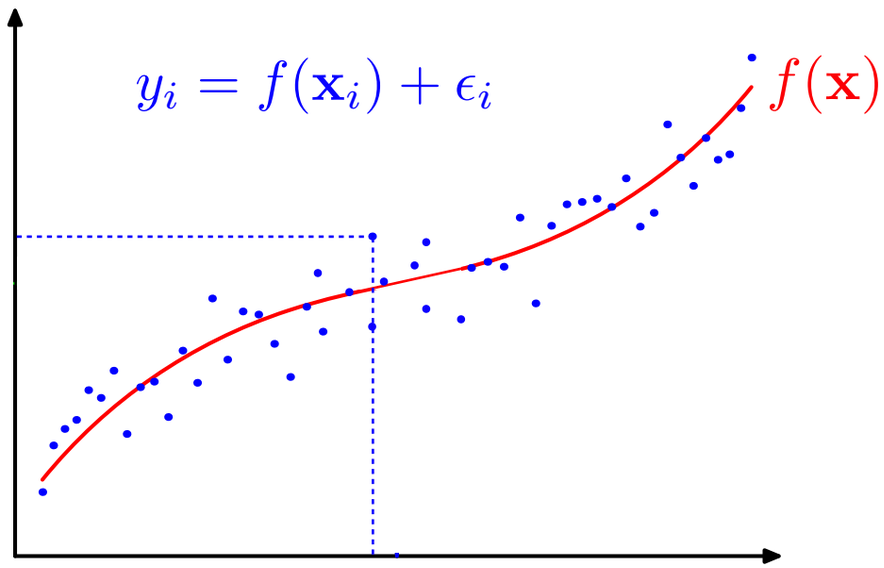
Noisy observations from the function

$$\mathbf{y} = \{y_i\}_{i=1}^N$$

$$X = \{\mathbf{x}_i\}_{i=1}^N$$

Observations Dataset

$$\mathcal{D} = (X, \mathbf{y})$$



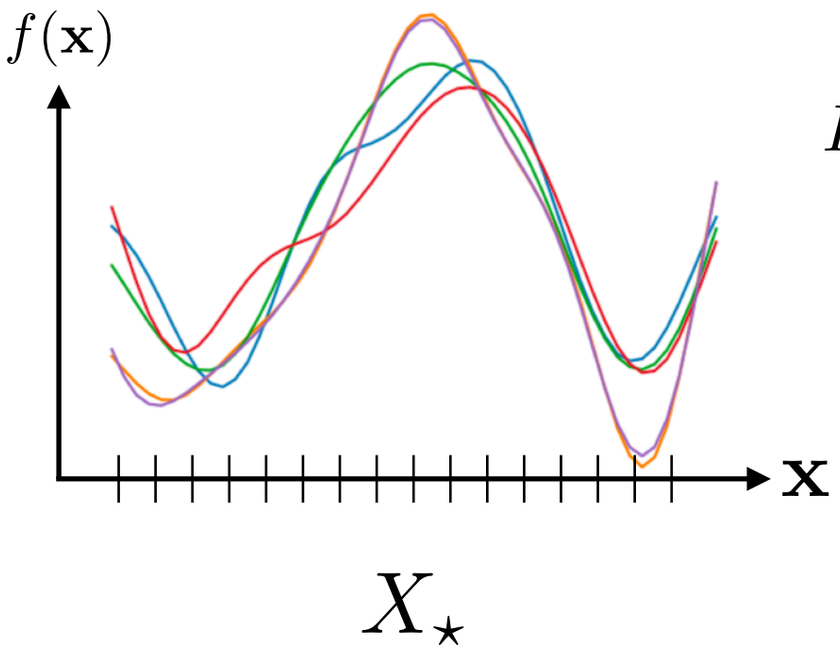
$$f(\mathbf{x}) \sim \mathcal{GP}\left(m(\mathbf{x}|\boldsymbol{\theta}_m), k(\mathbf{x}, \mathbf{x}'|\boldsymbol{\theta}_c)\right)$$

Samples from a Gaussian Process

$$m(\mathbf{x}) = 0$$

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\frac{(\mathbf{x}-\mathbf{x}')^{\mathrm{T}}(\mathbf{x}-\mathbf{x}')}{2\,\ell^2}\right)$$

$$K(X_\star, X_\star)_{(i,j)} = k(\mathbf{x}_i, \mathbf{x}_j)$$



$$K(X_\star, X_\star) = \begin{bmatrix} k(\mathbf{x}_0,\mathbf{x}_0) & k(\mathbf{x}_0,\mathbf{x}_1) & \cdots & k(\mathbf{x}_0,\mathbf{x}_M) \\ k(\mathbf{x}_1,\mathbf{x}_0) & k(\mathbf{x}_1,\mathbf{x}_1) & & \\ \vdots & & \ddots & \\ k(\mathbf{x}_M,\mathbf{x}_0) & & & k(\mathbf{x}_M,\mathbf{x}_M) \end{bmatrix}$$

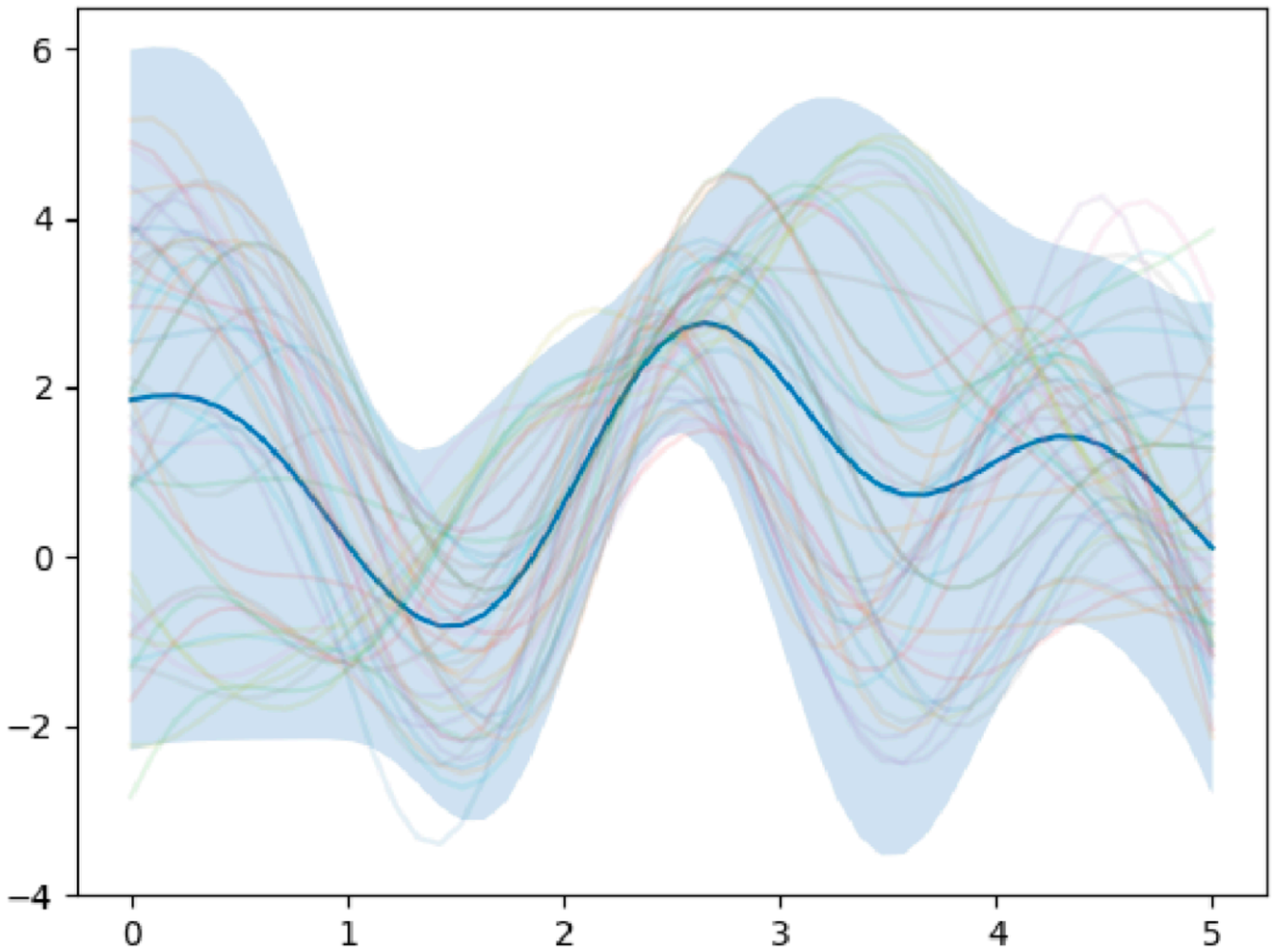
$$\mathbf{f}_\star \sim \mathcal{N}(\mathbf{0}, K(X_\star, X_\star))$$

How to draw samples?

$$K = LL^{\mathrm{T}} \quad \mathbf{u} \sim \mathcal{N}(\mathbf{0}, I) \quad \mathbf{f}_\star = L\mathbf{u}$$

Cholesky decomposition

Stan GP Samples

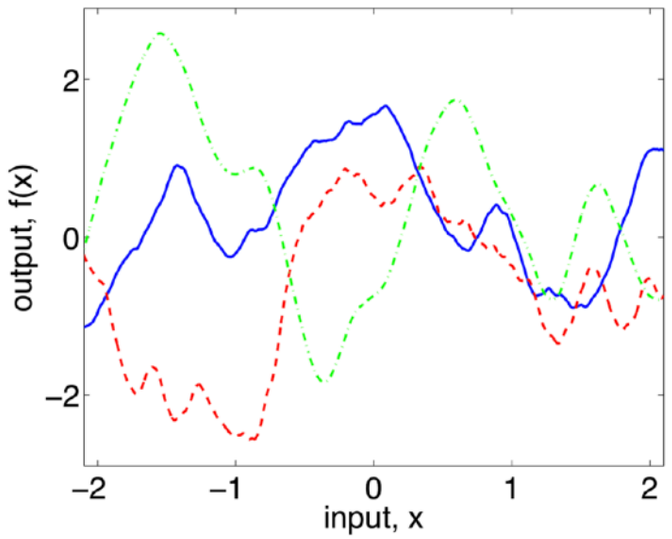


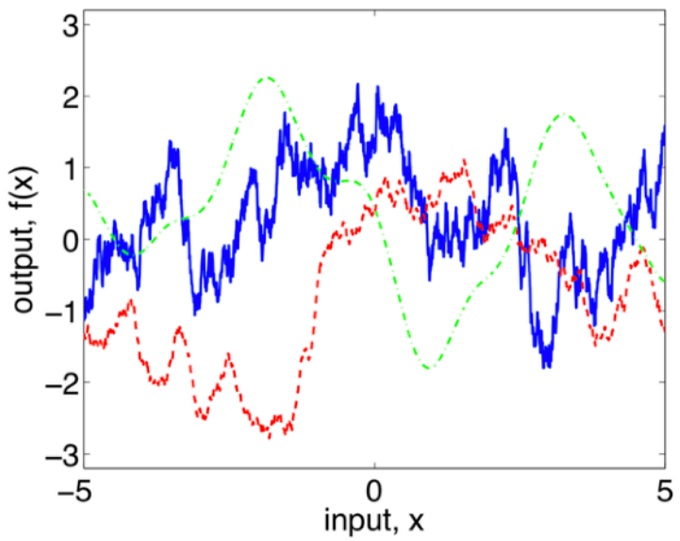
Covariance Functions

THE UNIVERSITY OF SYDNEY

Name	Expression $k(\mathbf{x}', \mathbf{x}'')$	Hyper-Parameters
Linear	$\sigma_f^2(\sigma_0^2 + \mathbf{x}'^T L \mathbf{x}'')$	$\boldsymbol{\theta}_c = \{\sigma_f, \sigma_0, L\}$
Matern 3	$\sigma_f^2 \left(1 + \sqrt{3\mathbf{d}}\right) \exp\left(-\sqrt{3\mathbf{d}}\right)$	$\boldsymbol{\theta}_c = \{\sigma_f, L\}$
Matern 5	$\sigma_f^2 \left(1 + \sqrt{5\mathbf{d}} + \frac{5\mathbf{d}}{3}\right) \exp\left(-\sqrt{5\mathbf{d}}\right)$	$\boldsymbol{\theta}_c = \{\sigma_f, L\}$
Polynomial	$\sigma_f^2(\sigma_0^2 + \mathbf{x}'^T L \mathbf{x}'')^p$	$\boldsymbol{\theta}_c = \{\sigma_f, \sigma_0, L, p\}$
Rational Quadratic	$\sigma_f^2 \left(1 + \frac{\mathbf{d}}{2\alpha}\right)^{-\alpha}$	$\boldsymbol{\theta}_c = \{\sigma_f, L, \alpha\}$
Squared Exponential	$\sigma_f^2 \exp\left(-\frac{\mathbf{d}}{2}\right)$	$\boldsymbol{\theta}_c = \{\sigma_f, L\}$
Periodic Exponential	$\sigma_f^2 \exp\left(-\frac{2\sin^2\left(\pi T \sqrt{\mathbf{d}}\right)}{\rho^2}\right)$	$\boldsymbol{\theta}_c = \{\sigma_f, L, T, \rho\}$

$$\mathbf{d} = (\mathbf{x}' - \mathbf{x}'')^\top L (\mathbf{x}' - \mathbf{x}'') \qquad L_{(i,i)} = \frac{1}{\ell_i^2}$$





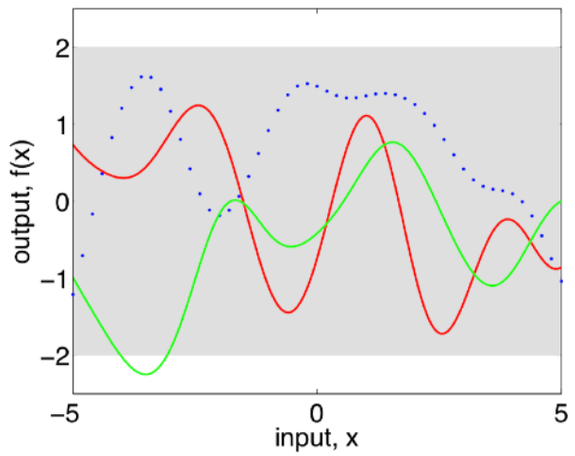
Gaussian Process Regression

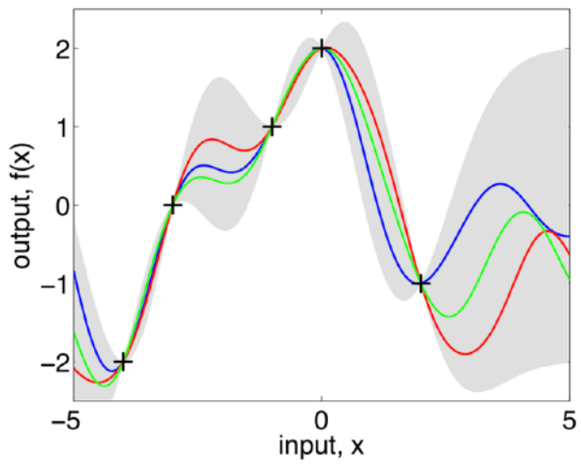
$$y = f(\mathbf{x}) + \epsilon \qquad \text{, with} \quad \epsilon \sim \mathcal{N}(0, \sigma_n^2)$$

$$\mathcal{D} = (X, \mathbf{y})$$

Query locations: $X^\star$

$$\mathbf{f}^\star = \{f_i^\star\}_{i=1}^N$$





Predictive distribution over test locations

$$\mathbf{f}^\star | X, \mathbf{y}, X^\star \sim \mathcal{N}(\bar{\mathbf{f}^\star}, \text{cov}(\mathbf{f}^\star))$$

$$\bar{\mathbf{f}^\star} = \mathbb{E}[\mathbf{f}^\star | X, \mathbf{y}, X^\star] = K(X^\star, X)\left[K(X, X) + \sigma_n^2 I\right]^{-1}(\mathbf{y} - M(X))$$

$$\text{cov}(\mathbf{f}^\star) = K(X^\star, X^\star) - K(X^\star, X)\left[K(X, X) + \sigma_n^2 I\right]^{-1} K(X, X^\star) \ ,$$

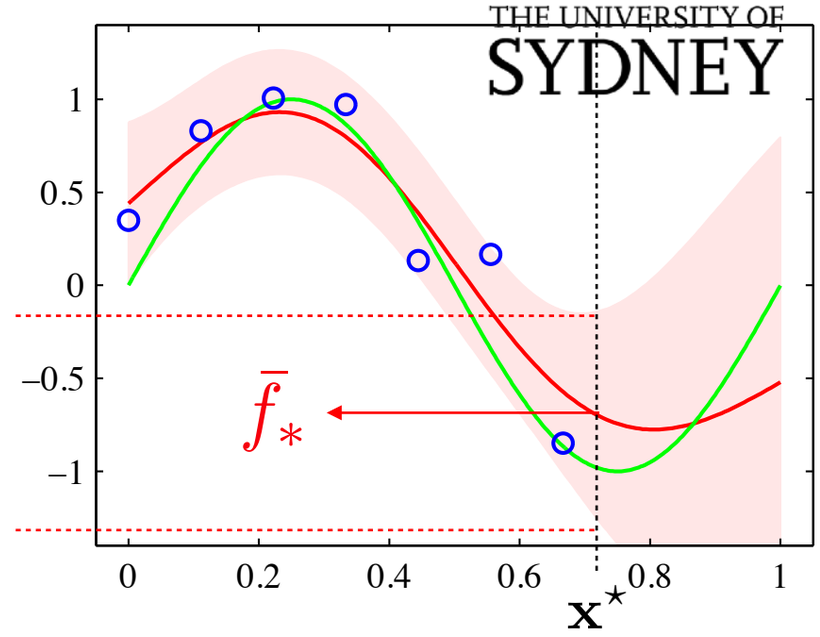
Gaussian Process Regression

One query location: $\mathbf{x}^\star$

$$\bar{f}_* = \mathbf{k}_*^\top (K + \sigma_n^2 I)^{-1} \mathbf{y},$$
$$\mathbb{V}[f_*] = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^\top (K + \sigma_n^2 I)^{-1} \mathbf{k}_*$$

$\bar{f}_* + 2\sqrt{\mathbb{V}[f_*]}$

$\bar{f}_* - 2\sqrt{\mathbb{V}[f_*]}$



input: X (inputs), $\mathbf{y}$ (targets), k (covariance function), σ_n^2 (noise level), $\mathbf{x}_*$ (test input)

$L := \text{cholesky}(K + \sigma_n^2 I)$

$\boldsymbol{\alpha} := L^\top \backslash (L \backslash \mathbf{y})$

$\bar{f}_* := \mathbf{k}_*^\top \boldsymbol{\alpha}$

$\mathbf{v} := L \backslash \mathbf{k}_*$

$\mathbb{V}[f_*] := k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{v}^\top \mathbf{v}$

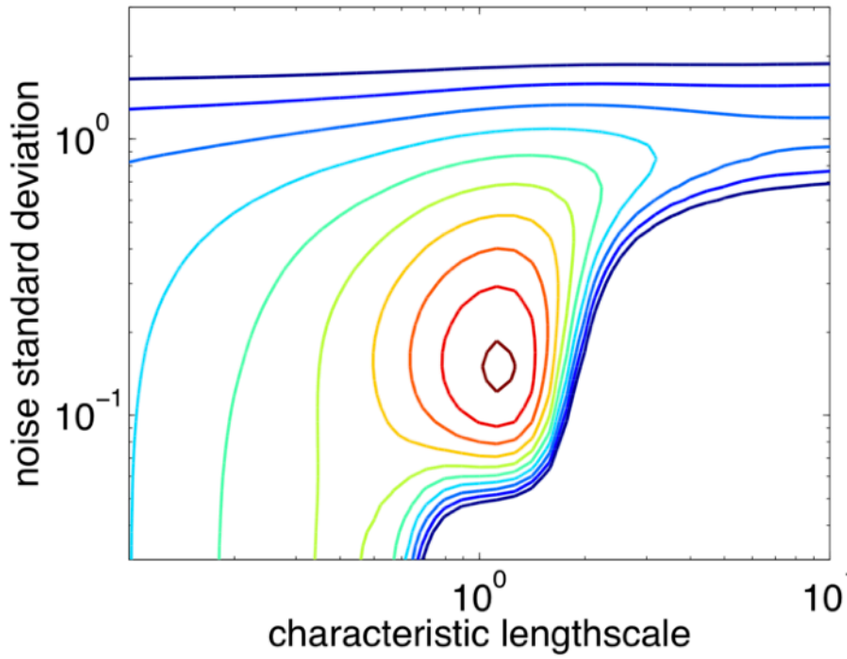
$\log p(\mathbf{y}|X) := -\frac{1}{2}\mathbf{y}^\top \boldsymbol{\alpha} - \sum_i \log L_{ii} - \frac{n}{2} \log 2\pi$

return: $\bar{f}_*$ (mean), $\mathbb{V}[f_*]$ (variance), $\log p(\mathbf{y}|X)$ (log marginal likelihood)

Model Selection - Optimisation

Hyper-parameter Set $\boldsymbol{\theta} = \{\boldsymbol{\theta}_m, \boldsymbol{\theta}_c, \sigma_n\}$

Optimal Hyper-Parameters $\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \ \mathrm{GF}(\mathbf{y}, X, \boldsymbol{\theta})$



Log Marginal Likelihood

$$\mathrm{LML}(\mathbf{y}, X, \boldsymbol{\theta}) = -\frac{1}{2}(\mathbf{y} - M(X))^T K_X^{-1}(\mathbf{y} - M(X)) - \frac{1}{2}\log|K_X| - \frac{n}{2}\log 2\pi$$

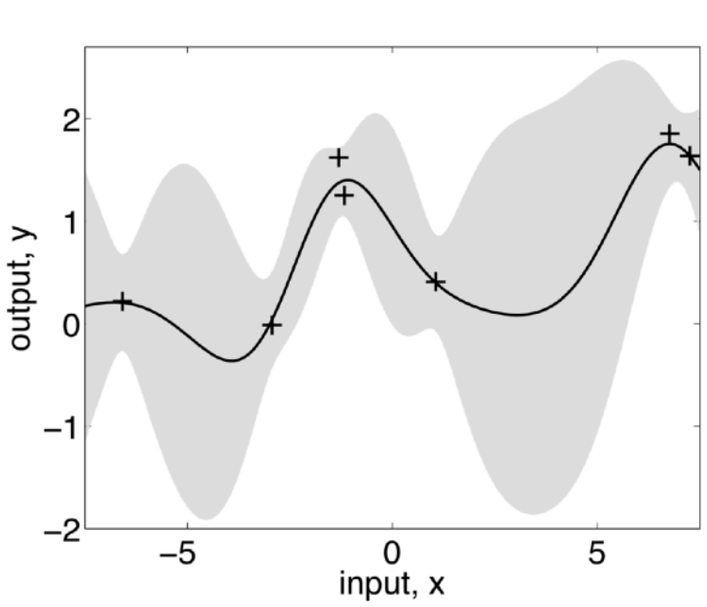
$$K_X = K(X, X) + \sigma_n I \qquad M(X)_{(i)} = m(x_i)$$

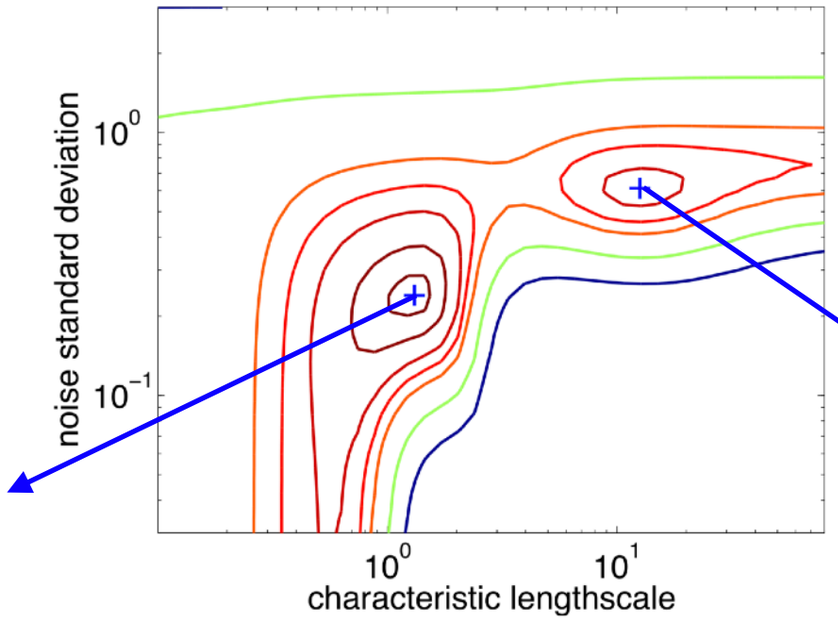
$$\begin{aligned}\frac{\partial}{\partial\theta_j}\log p(\mathbf{y}|X, \boldsymbol{\theta}) &= \frac{1}{2}\mathbf{y}^\top K^{-1}\frac{\partial K}{\partial\theta_j}K^{-1}\mathbf{y} - \frac{1}{2}\,\mathrm{tr}\left(K^{-1}\frac{\partial K}{\partial\theta_j}\right) \\ &= \frac{1}{2}\,\mathrm{tr}\left((\boldsymbol{\alpha}\boldsymbol{\alpha}^\top - K^{-1})\frac{\partial K}{\partial\theta_j}\right) \quad \text{where} \quad \boldsymbol{\alpha} = K^{-1}\mathbf{y}\end{aligned}$$

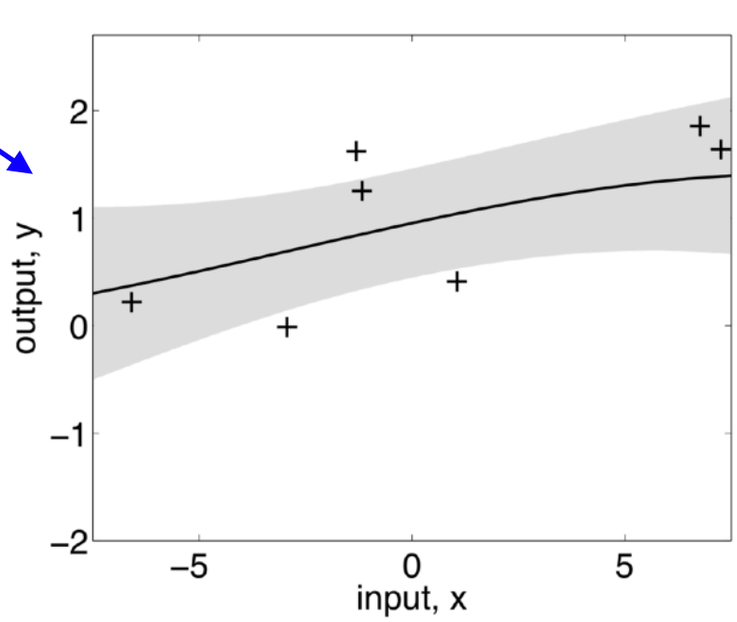
Model Selection (Optimisation)

There is no guarantee that the **marginal likelihood** does not suffer from **multiple local optima**.

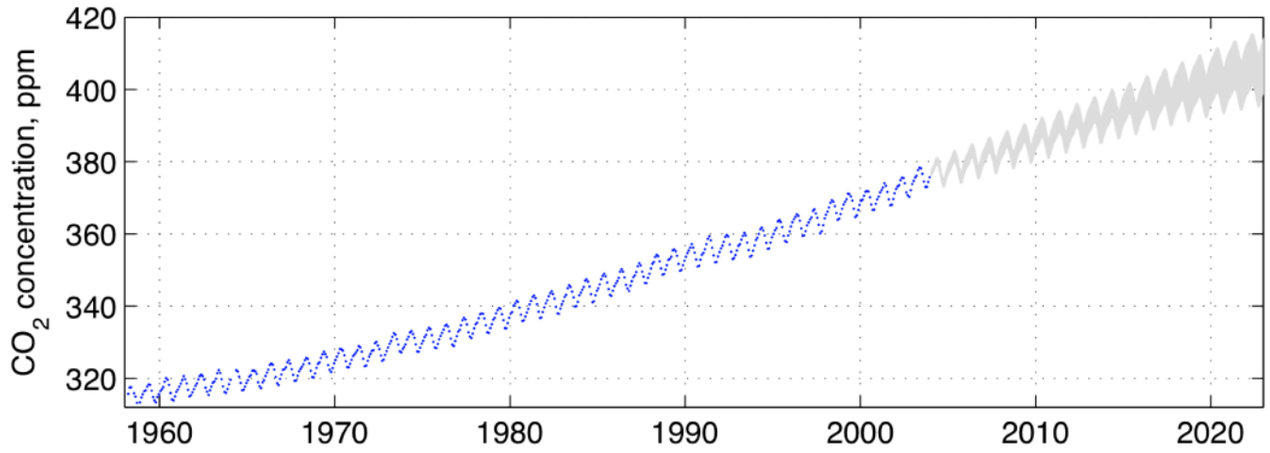
Every local maximum corresponds to a particular interpretation of the data.







Covariance Function Combination

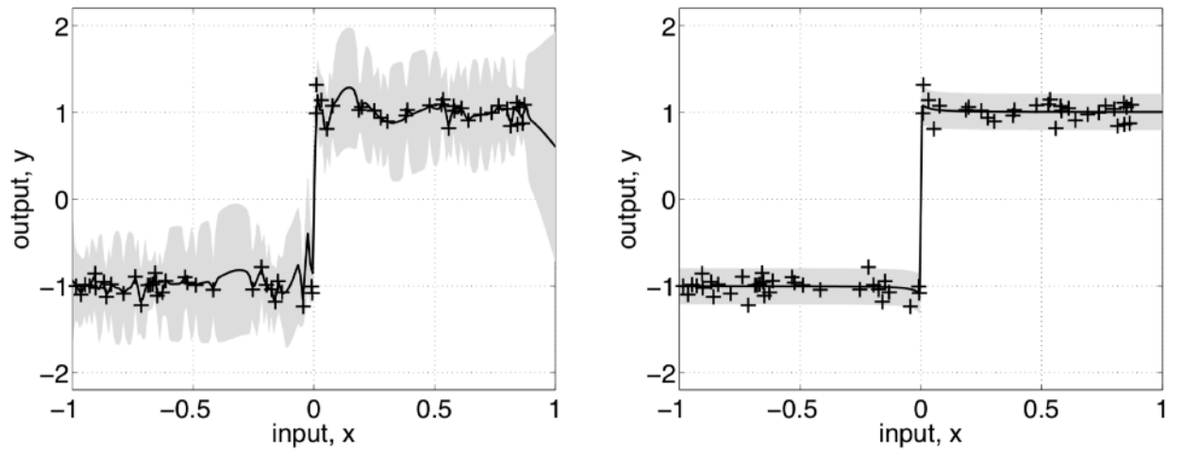


$$k(x, x') = k_1(x, x') + k_2(x, x') + k_3(x, x') + k_4(x, x')$$

$$k_2(x, x') = \theta_3^2 \exp\left(-\frac{(x - x')^2}{2\theta_4^2} - \frac{2\sin^2(\pi(x - x'))}{\theta_5^2}\right)$$

$$k_4(x_p, x_q) = \theta_9^2 \exp\left(-\frac{(x_p - x_q)^2}{2\theta_{10}^2}\right) + \theta_{11}^2 \delta_{pq}$$

Covariance Function Combination



$$k_{\mathrm{NN}}(\mathbf{x}, \mathbf{x}') = \frac{2}{\pi} \sin^{-1}\left(\frac{2\tilde{\mathbf{x}}^\top \Sigma \tilde{\mathbf{x}}'}{\sqrt{(1+2\tilde{\mathbf{x}}^\top \Sigma \tilde{\mathbf{x}})(1+2\tilde{\mathbf{x}}'^\top \Sigma \tilde{\mathbf{x}}')}}\right)$$

Large Datasets

$$\bar{f}_* = \mathbf{k}_*^\top (K + \sigma_n^2 I)^{-1} \mathbf{y},$$
$$\mathbb{V}[f_*] = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^\top (K + \sigma_n^2 I)^{-1} \mathbf{k}_*$$

K is a $N \times N$ matrix.

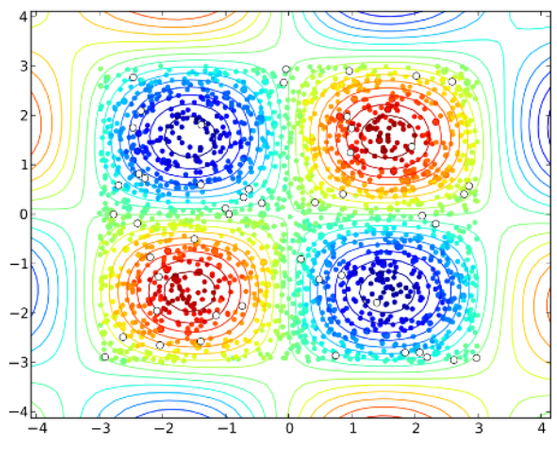
What happens when N is really big?

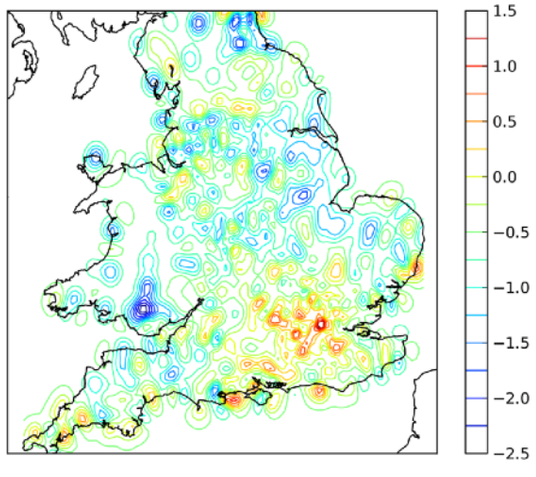
Calculating this inverse $(K + \sigma_n I)^{-1}$ is of order $O(N^3)$

Stochastic Variational Inference.

Current Solutions:

James Hensman, Nicolo Fusi, Neil D. Lawrence, Gaussian Processes for Big Data, arXiv:1309.6835





THE UNIVERSITY OF SYDNEY

Research Example 1

Spatial Temporal Monitoring

Space Time Gaussian Processes

THE UNIVERSITY OF SYDNEY

The domain of the function now contains spatial and temporal components.

$$f\left(\mathbf{s};t\right) \sim \mathcal{GP}\left(m\left(\mathbf{s};t\right), k\left(\left(\mathbf{s};t\right),\left(\mathbf{s};t\right)'\right)\right).$$

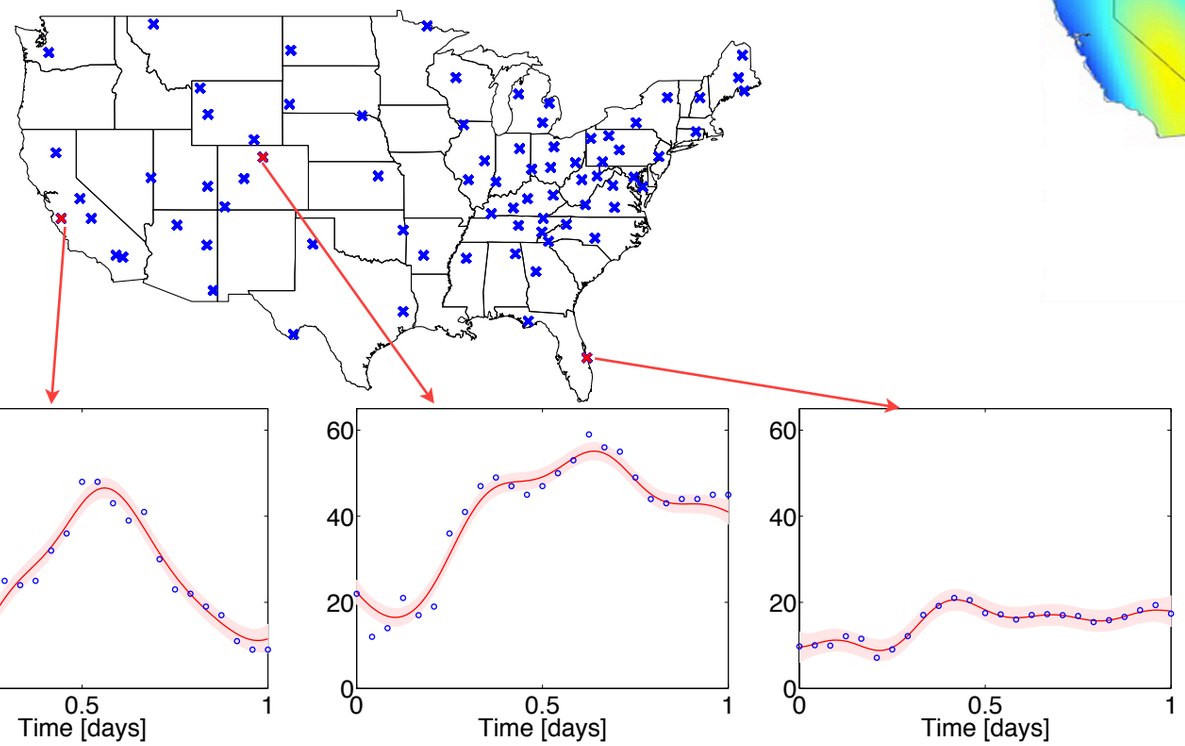
Predictive distribution over test locations

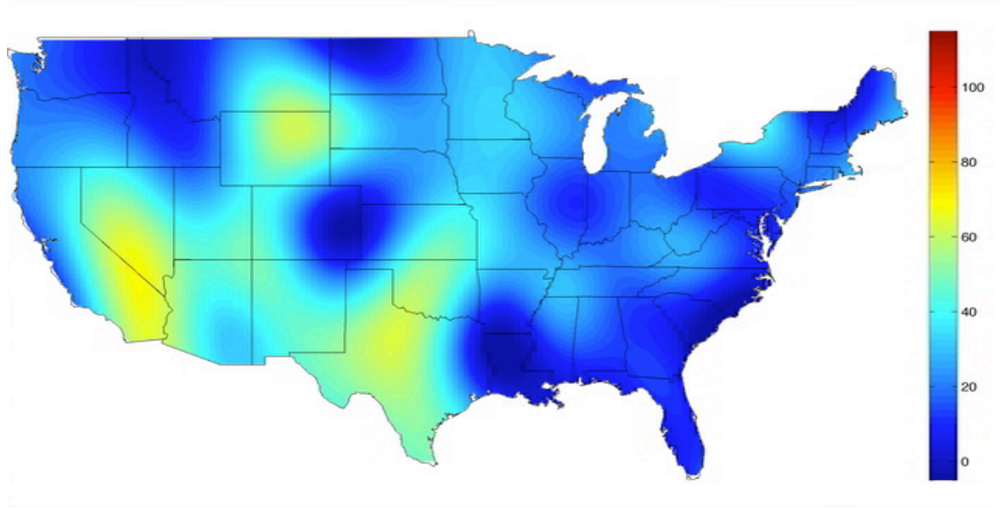
$$f((\mathbf{s};t)^*) \sim \mathcal{N}\left(\mu, \sigma^2\right)$$

$$\begin{aligned}\mu((\mathbf{s};t)^*) = &\quad K((\mathbf{s};t)^*, X)K_X^{-1}(\mathbf{y} - m(X)) \,, \\ \sigma^2((\mathbf{s};t)^*) = &\quad K((\mathbf{s};\mathbf{t})^*, (\mathbf{s};t)^*) - K((\mathbf{s};\mathbf{t})^*, X)K_X^{-1}K(X, (\mathbf{s};\mathbf{t})^*) \,.\end{aligned}$$

Spatial Temporal Process
Ozone Concentration

THE UNIVERSITY OF SYDNEY





Space Time Covariance Functions

Able to capture spatial and temporal specific behaviour.

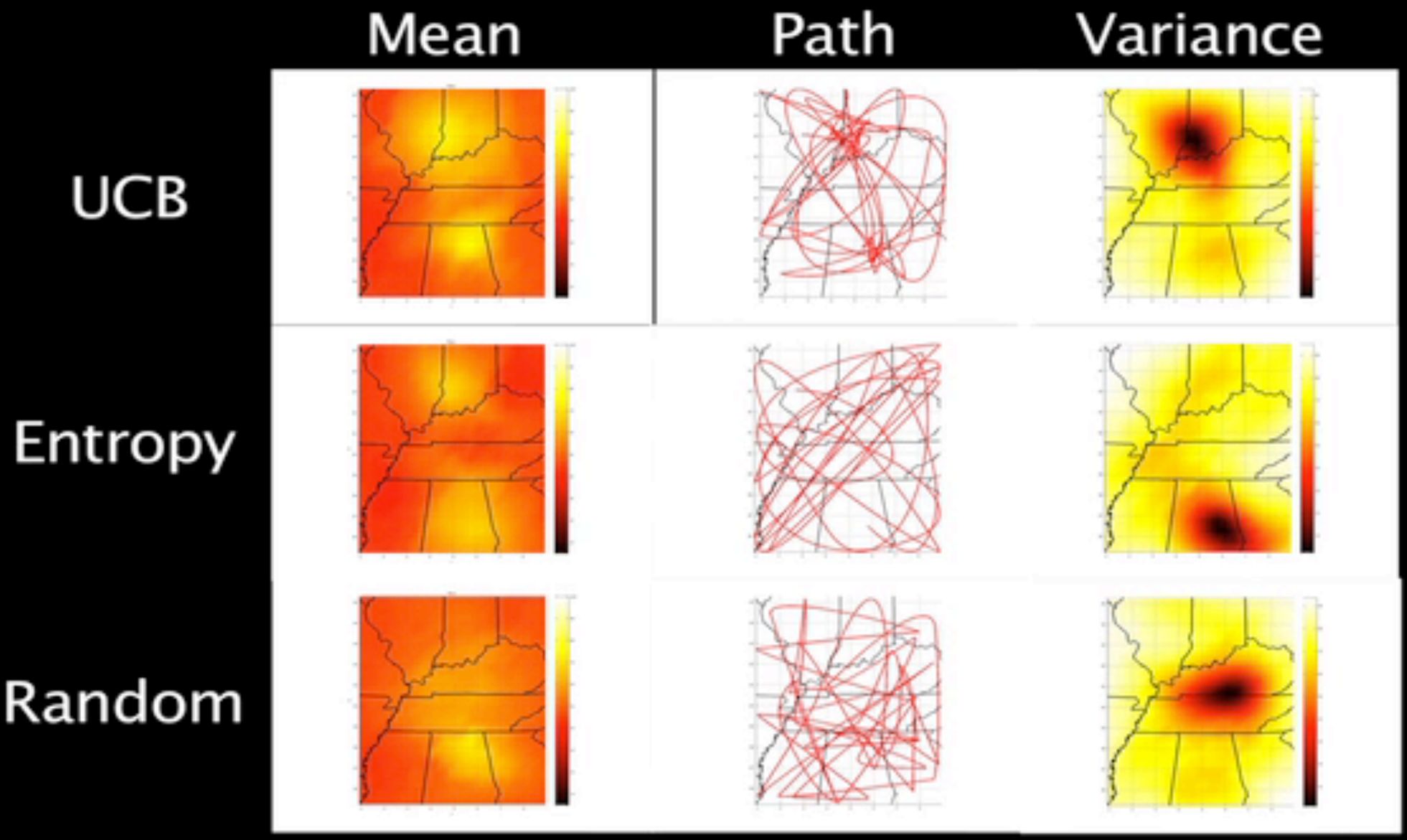
When spatial and temporal patterns are not coupled:

$$k_{sep}\left((\mathbf{s};t),(\mathbf{s};t)'|\boldsymbol{\theta}_c\right) = k_{\text{space}}(\mathbf{s},\mathbf{s}'|\boldsymbol{\theta}_{\text{space}})\, k_{\text{time}}(t,t'|\boldsymbol{\theta}_{\text{time}})$$

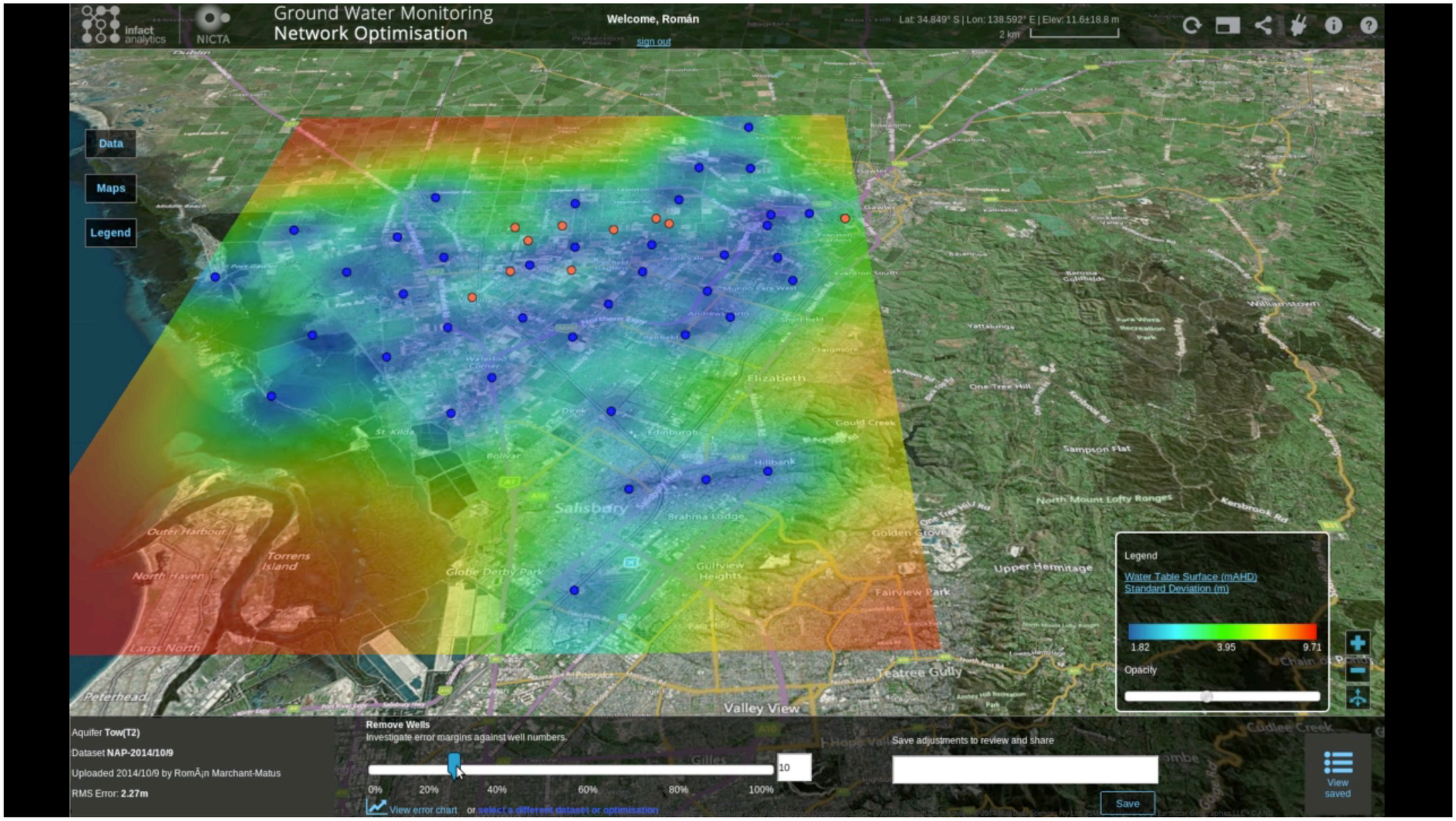
For example: Separable Non-Stationary

$$k(\mathbf{s},\mathbf{s}';t,t' \mid \boldsymbol{\theta}) = \sigma^2 \exp\{-b^2\|\mathbf{s}+\mathbf{s}'\|^2 - a^2(t+t')^2\} + \sigma^2 \exp\{-b^2\|\mathbf{s}-\mathbf{s}'\|^2 - a^2(t-t')^2\} \quad (13)$$

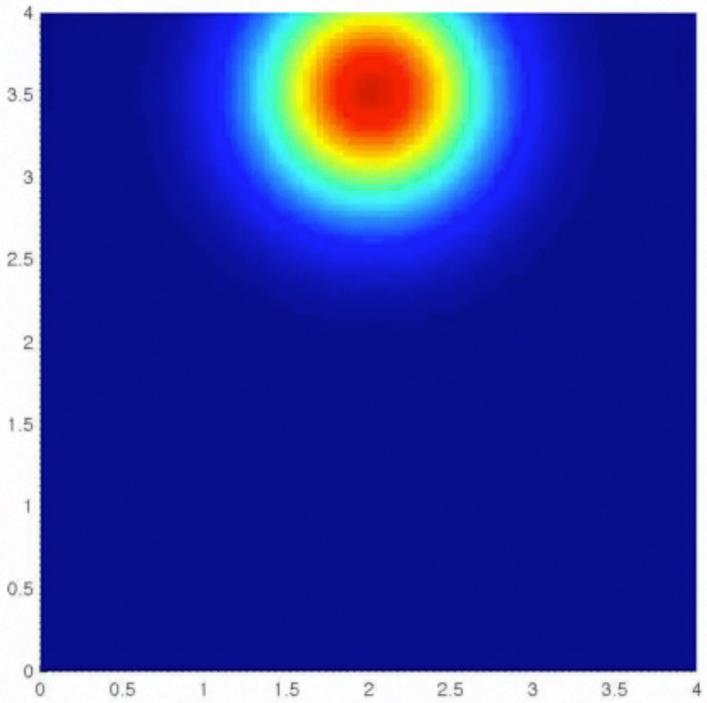
$$-2[\sigma^2\exp\{-b^2\|\mathbf{s}\|^2-a^2(t)^2\}+\sigma^2\exp\{-b^2\|\mathbf{s}'\|^2-a^2(t')^2\}-\sigma^2].$$


Mean
Path
Variance
UCB
Entropy
Random
explores the unknown environment and at the same time identif

Example: Ground Water Monitoring

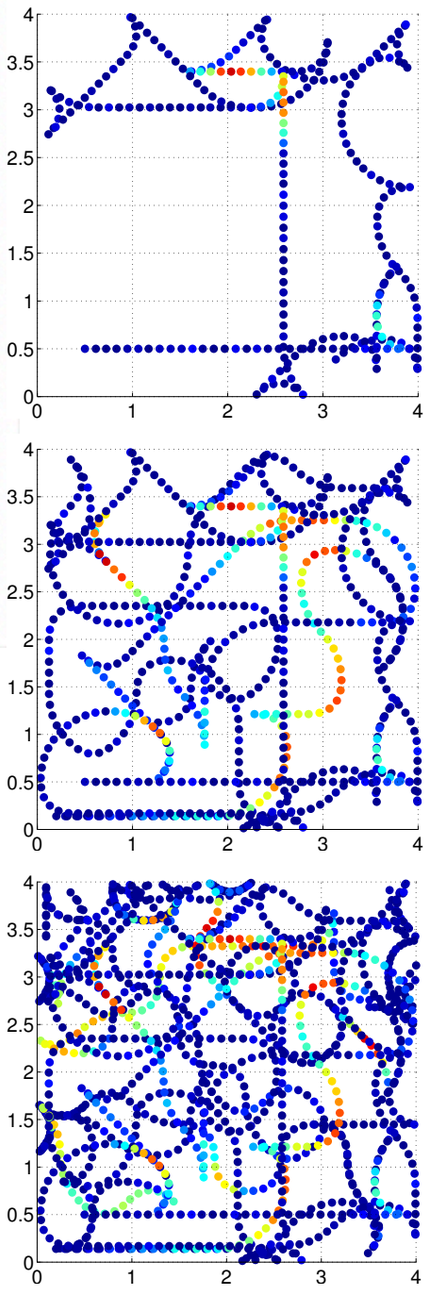


Path Planning

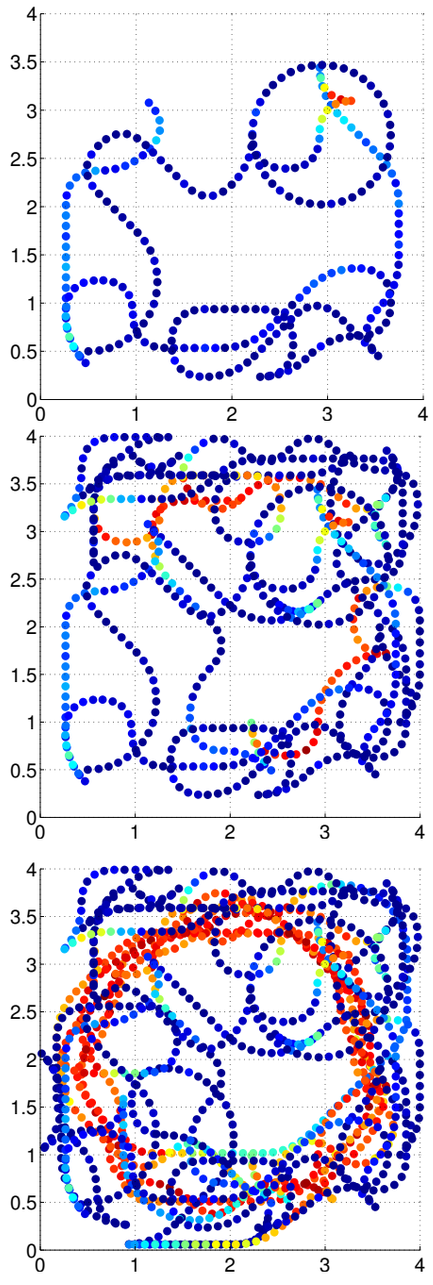


Space Time Function
Ground Truth

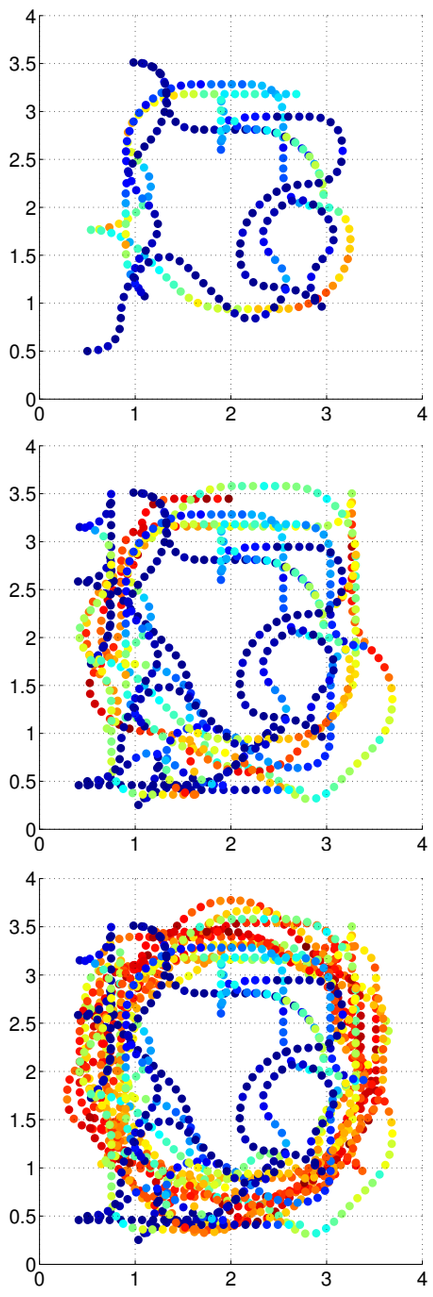
Depth 1



Depth 2



Depth 5



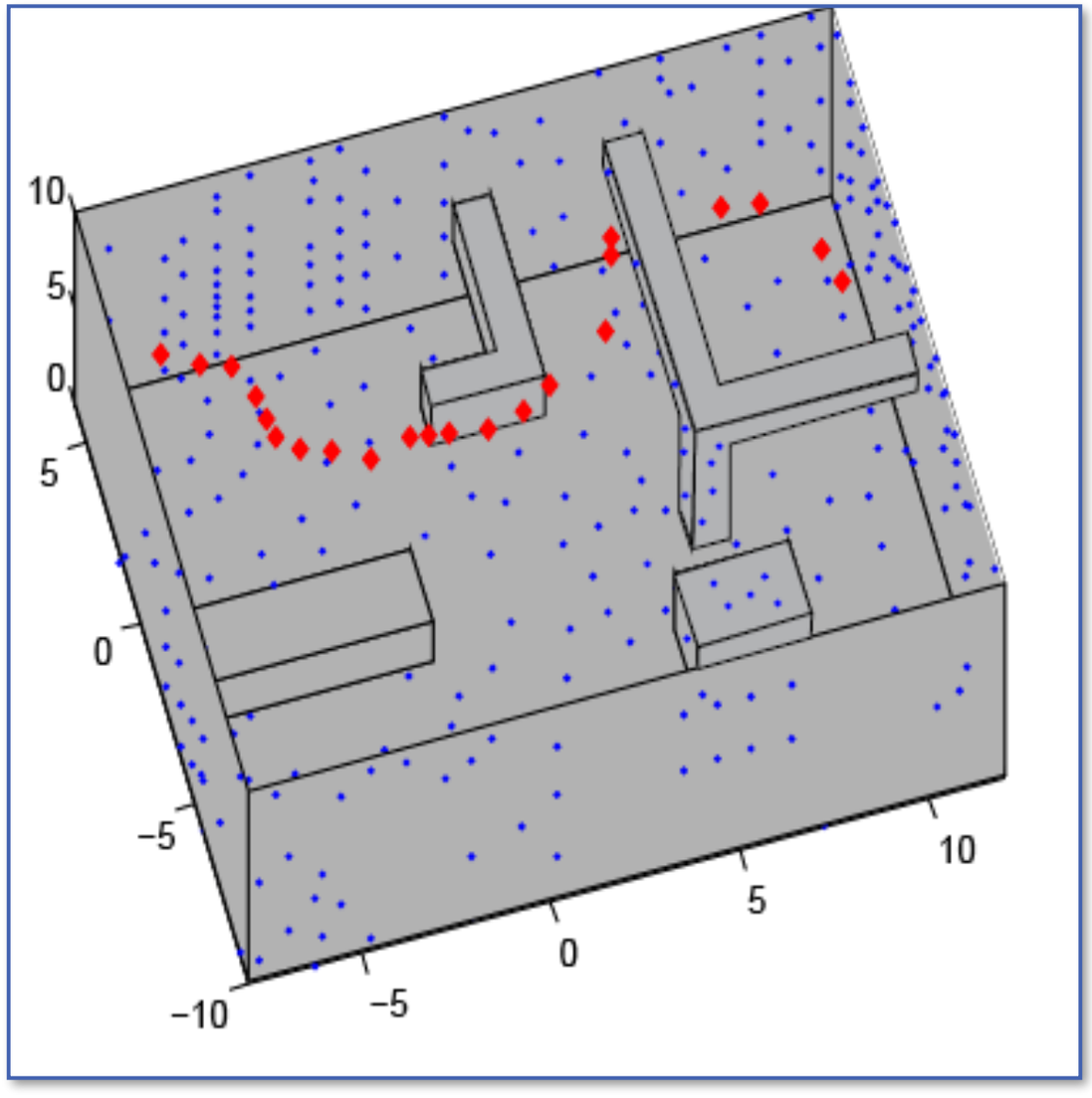
THE UNIVERSITY OF SYDNEY

Research Example 2

Mapping with Gaussian Processes

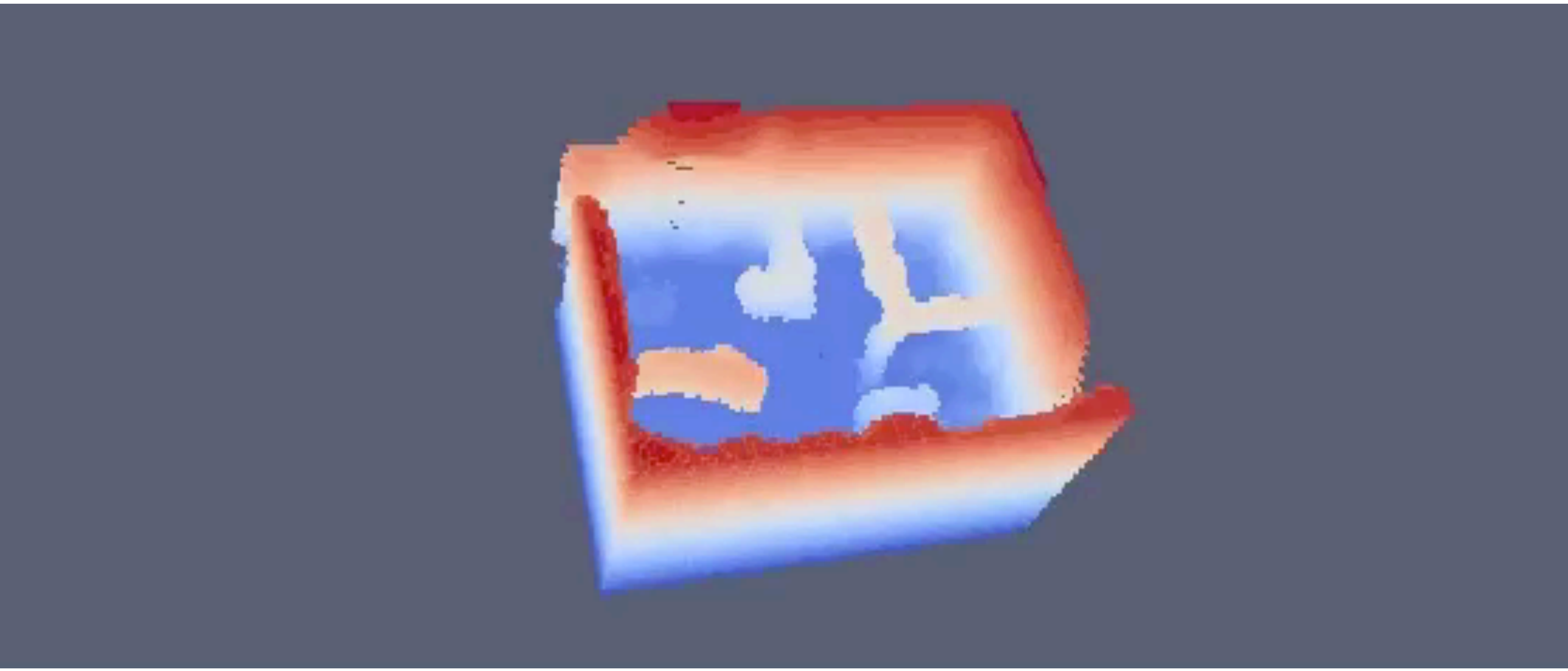
Simulated Dataset: Probability of Occupancy Versus Location

THE UNIVERSITY OF SYDNEY



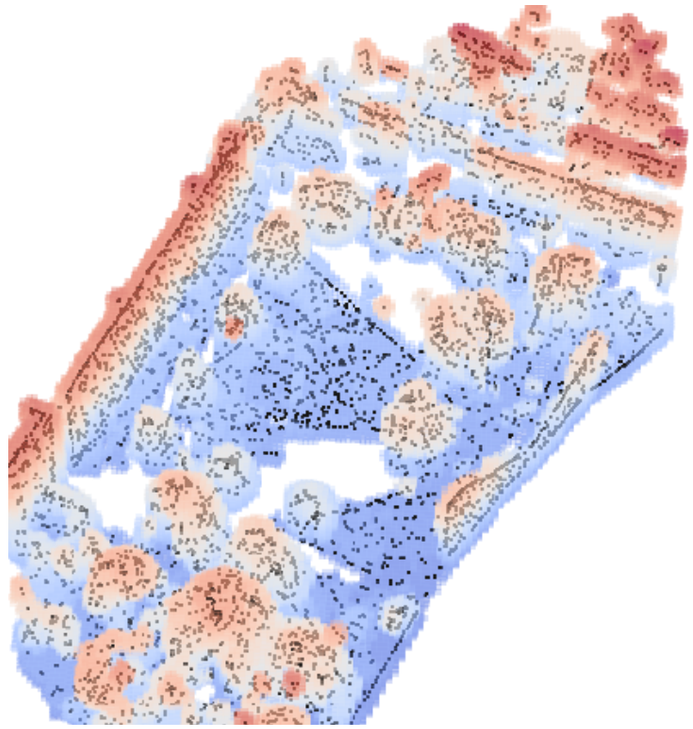
Simulated Dataset: Predicted Occupied Points

THE UNIVERSITY OF SYDNEY



Outdoor Dataset:
Subsampled Reigl Rangefinder Point Cloud

THE UNIVERSITY OF SYDNEY



Outdoor Dataset:
Subsampled Rangefinder Point Cloud

THE UNIVERSITY OF SYDNEY