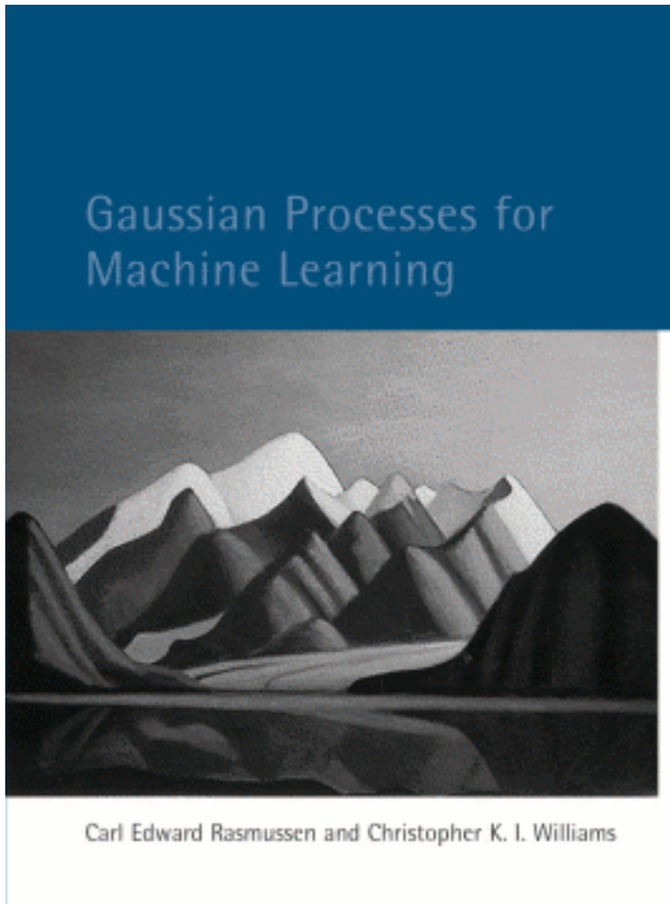
Gaussian Process Regression

PREDICTIVE
SCIENCE LABORATORY

Objectives

- to do regression using a GP
- to find the hyperparameters of the GP by maximizing the (marginal) likelihood
- to use GP regression for uncertainty propagation

The Best Book on the Subject



Gaussian Processes for Machine Learning
Carl Edward Rasmussen and Christopher K. I. Williams
The MIT Press, 2006. ISBN 0-262-18253-X.

Free online at www.gaussianprocess.org.
With Matlab code.

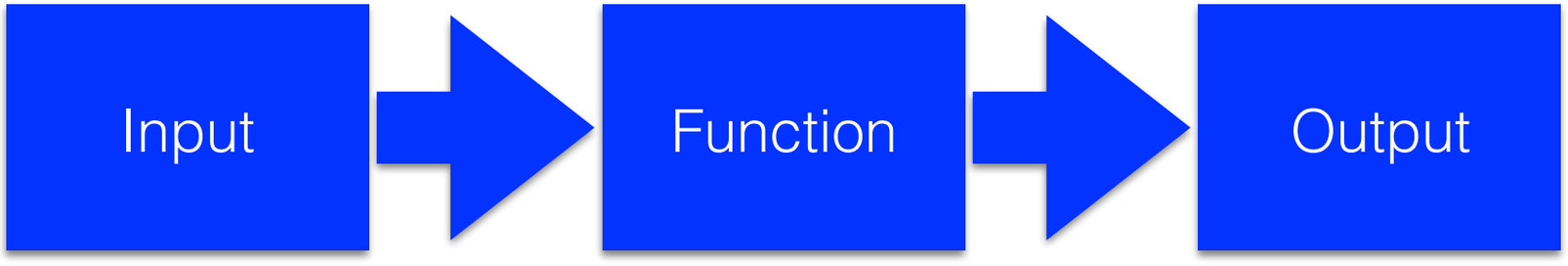
PREDICTIVE
SCIENCE LABORATORY

The Best Code on the Subject

GPy (in Python) from the group of N. Lawrence @ University of Sheffield
https://github.com/SheffieldML/GPy

PREDICTIVE
SCIENCE LABORATORY

Definition of a Gaussian process



- Treat *f* as unknown
- Unknown = uncertain = “random”, i.e., described with probabilities
- Let us denote our beliefs about *f* as follows:

$$f(\cdot) \sim p(f(\cdot))$$

Definition of a Gaussian process

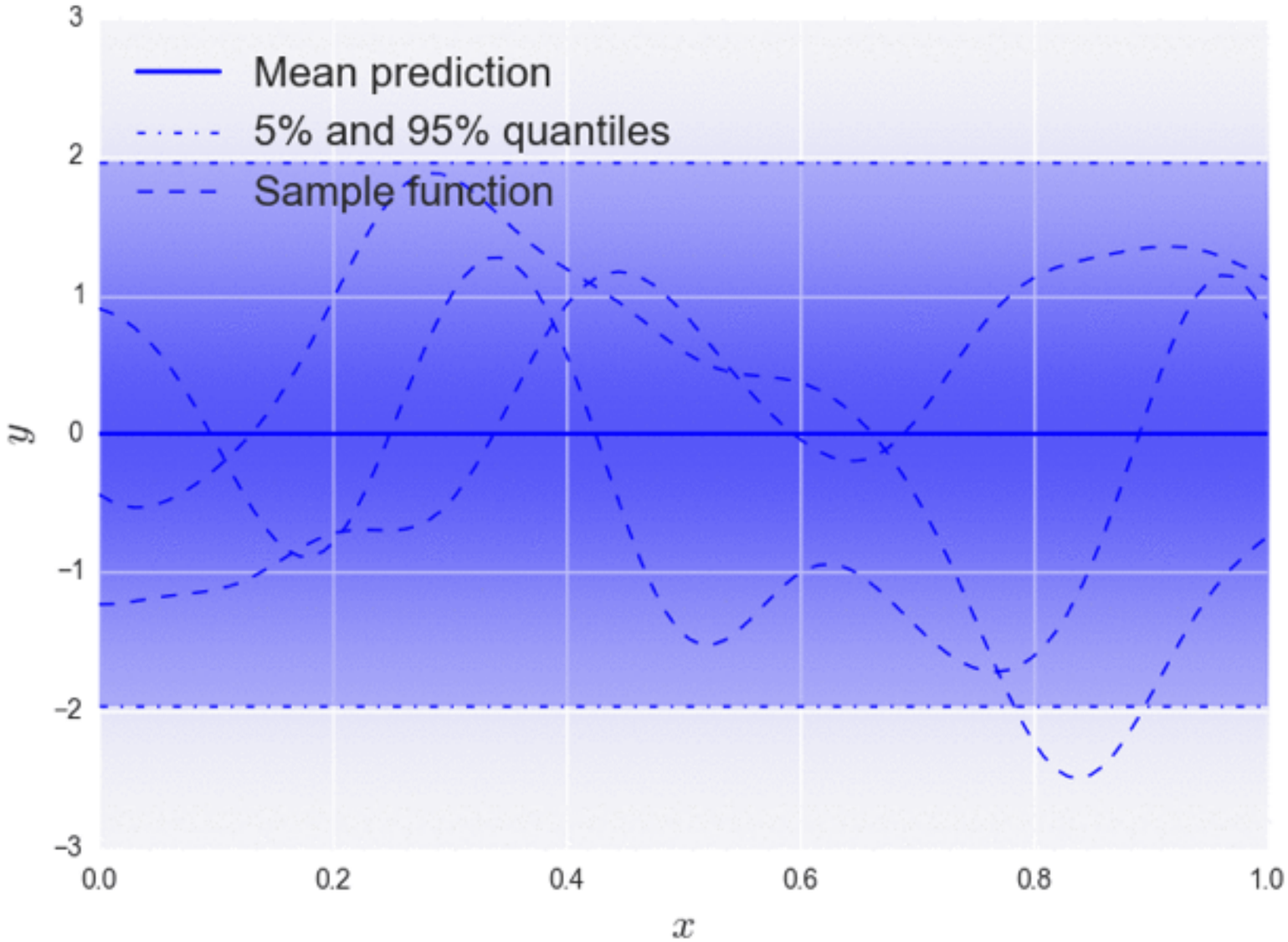
A Gaussian process needs two ingredients:

- a mean function
- a covariance function

It uses them to define a probability measure on the space of functions.

We write: $$f(\cdot) \sim p(f(\cdot)) = \mathrm{GP}\left(f(\cdot) \mid m(\cdot), k(\cdot,\cdot)\right)$$

Bayesian surrogate



PREDICTIVE
SCIENCE LABORATORY

GP Regression 1: No Noise

- You have some input/output data:

$$\mathbf{X} = \{x_1, \ldots, x_N\},$$

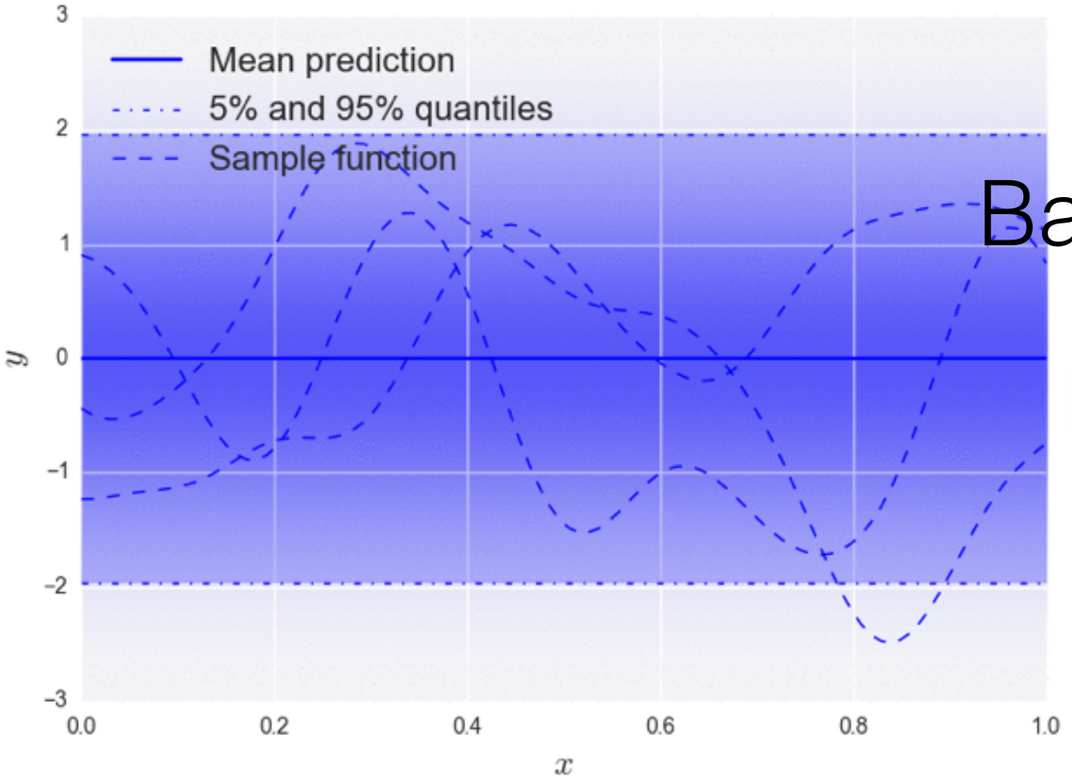
$$\mathbf{f} = \{f(x_1), \ldots, f(x_N)\}$$

- You wish to learn f.

- Before you start you need to say what you know about f:

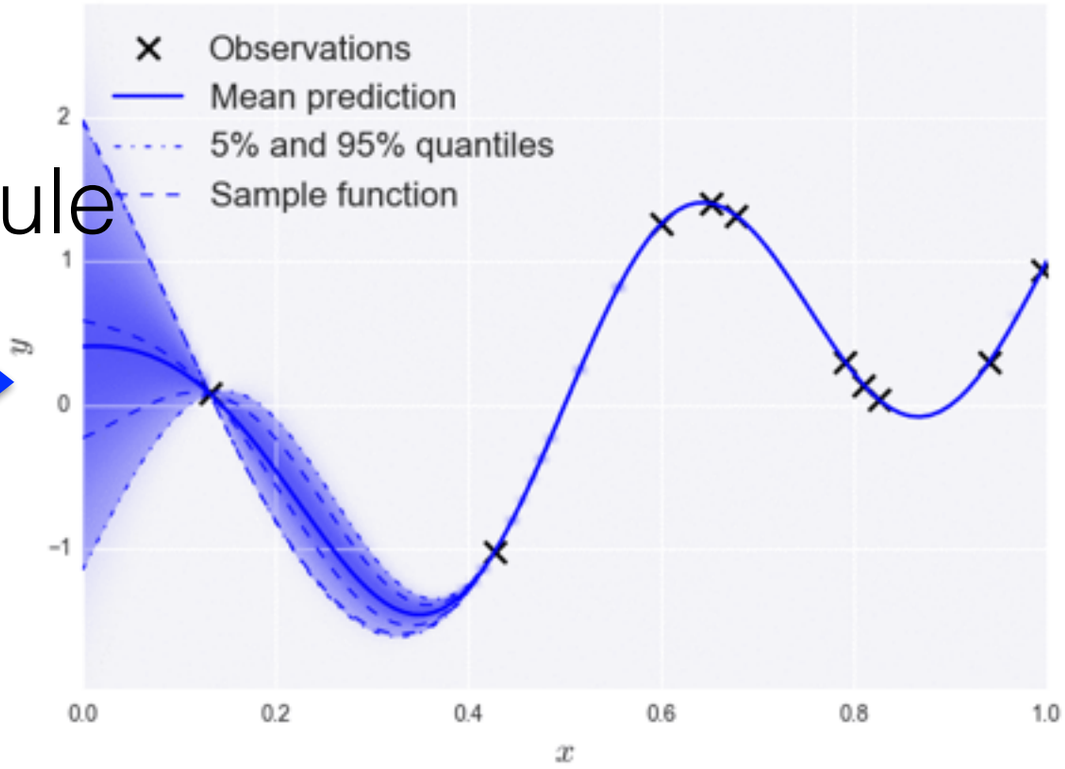
$$f(\cdot) \sim \mathrm{GP}\left(f(\cdot) \mid m(\cdot), k(\cdot,\cdot)\right)$$

Gaussian process regression



Prior GP

Bayes rule



Posterior GP

PREDICTIVE
SCIENCE LABORATORY

Notebook example

Read description

PREDICTIVE
SCIENCE LABORATORY

Gaussian process regression

- Assume that we have observed:

$$\mathbf{X} = \{x_1, \ldots, x_N\},$$

$$\mathbf{f} = \{f(x_1), \ldots, f(x_N)\}$$

- and that we want to make predictions at an arbitrary set of *test* inputs:

$$\mathbf{X}^* = \{x_1^*, \ldots, x_{N^*}^*\}$$

$$\mathbf{f}^* = \{f(x_1^*), \ldots, f(x_{N^*}^*)\}$$

Gaussian process regression

- Since, we have assumed a priori that:

$$p(f(\cdot)) \sim \mathrm{GP}\left(f(\cdot) \mid m(\cdot), k(\cdot,\cdot)\right)$$

- then by definition:

$$p\begin{pmatrix}\mathbf{f}\\ \mathbf{f}^*\end{pmatrix} = \mathcal{N}\left(\begin{pmatrix}\mathbf{f}\\ \mathbf{f}^*\end{pmatrix} \mid \begin{pmatrix}\mathbf{m}\\ \mathbf{m}^*\end{pmatrix}, \begin{pmatrix}\mathbf{K}(\mathbf{X},\mathbf{X}) & \mathbf{K}(\mathbf{X},\mathbf{X}^*)\\ \mathbf{K}(\mathbf{X}^*,\mathbf{X}) & \mathbf{K}(\mathbf{X}^*,\mathbf{X}^*)\end{pmatrix}\right)$$

Gaussian process regression

Mean on observations

Covariance matrix of observations

$$p\begin{pmatrix}\mathbf{f}\\ \mathbf{f}^*\end{pmatrix} = \mathcal{N}\left(\begin{pmatrix}\mathbf{f}\\ \mathbf{f}^*\end{pmatrix} \middle| \begin{pmatrix}\mathbf{m}\\ \mathbf{m}^*\end{pmatrix}, \begin{pmatrix}\mathbf{K}(\mathbf{X},\mathbf{X}) & \mathbf{K}(\mathbf{X},\mathbf{X}^*)\\ \mathbf{K}(\mathbf{X}^*,\mathbf{X}) & \mathbf{K}(\mathbf{X}^*,\mathbf{X}^*)\end{pmatrix}\right)$$

Mean on test inputs

Cross covariance matrix (test-observed)

Covariance matrix of test inputs

Gaussian process regression

$$p\begin{pmatrix}\mathbf{f}\\ \mathbf{f}^*\end{pmatrix} = \mathcal{N}\left(\begin{pmatrix}\mathbf{f}\\ \mathbf{f}^*\end{pmatrix} | \begin{pmatrix}\mathbf{m}\\ \mathbf{m}^*\end{pmatrix}, \begin{pmatrix}\mathbf{K}(\mathbf{X},\mathbf{X}) & \mathbf{K}(\mathbf{X},\mathbf{X}^*)\\ \mathbf{K}(\mathbf{X}^*,\mathbf{X}) & \mathbf{K}(\mathbf{X}^*,\mathbf{X}^*)\end{pmatrix}\right)$$

Bayes rule

$$\mathbf{f}^* \mid \mathbf{X}^*, \mathbf{X}, \mathbf{f} \sim ?$$

Gaussian process regression

$$p\begin{pmatrix}\mathbf{f}\\ \mathbf{f}^*\end{pmatrix} = \mathcal{N}\left(\begin{pmatrix}\mathbf{f}\\ \mathbf{f}^*\end{pmatrix}\middle|\begin{pmatrix}\mathbf{m}\\ \mathbf{m}^*\end{pmatrix},\begin{pmatrix}\mathbf{K}(\mathbf{X},\mathbf{X}) & \mathbf{K}(\mathbf{X},\mathbf{X}^*)\\ \mathbf{K}(\mathbf{X}^*,\mathbf{X}) & \mathbf{K}(\mathbf{X}^*,\mathbf{X}^*)\end{pmatrix}\right)$$

Bayes rule

$$\mathbf{f}^* \mid \mathbf{X}^*,\mathbf{X},\mathbf{f} \sim \mathcal{N}\left(\mathbf{f}^* \mid \widetilde{\mathbf{m}},\widetilde{\mathbf{K}}\right),$$

$$\widetilde{\mathbf{m}} = \mathbf{m}^* + \mathbf{K}(\mathbf{X}^*,\mathbf{X})\mathbf{K}^{-1}(\mathbf{f}-\mathbf{m}),$$

$$\widetilde{\mathbf{K}} = \mathbf{K}^* - \mathbf{K}(\mathbf{X}^*,\mathbf{X})\mathbf{K}^{-1}\mathbf{K}(\mathbf{X},\mathbf{X}^*)$$

Proof in Ch. 2.3 Bishop (2006)

The posterior Gaussian process

- Since the choice of test points was arbitrary, the procedure actually defines a *posterior* Gaussian process:

$$p(f(\cdot)\,|\,\mathbf{X},\mathbf{f}) = \mathcal{GP}\Big(f(\cdot)\,|\,\tilde{m}(\cdot),\tilde{k}(\cdot,\cdot)\Big),$$

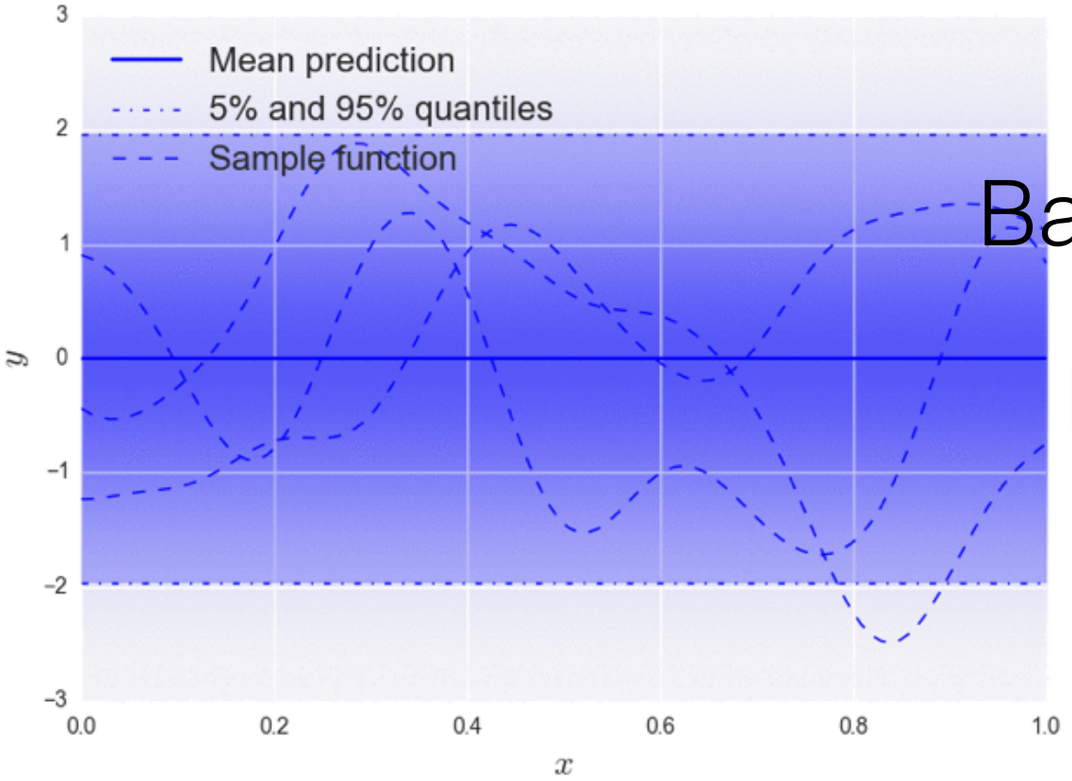
$$\tilde{m}(\mathbf{x}) = m(\mathbf{x}) + \mathbf{K}(\mathbf{x},\mathbf{X})\mathbf{K}^{-1}(\mathbf{f}-\mathbf{m}),$$

$$\tilde{k}(\mathbf{x},\mathbf{x}') = k(\mathbf{x},\mathbf{x}') - \mathbf{K}(\mathbf{x},\mathbf{X})\mathbf{K}^{-1}\mathbf{K}(\mathbf{X},\mathbf{x}')$$

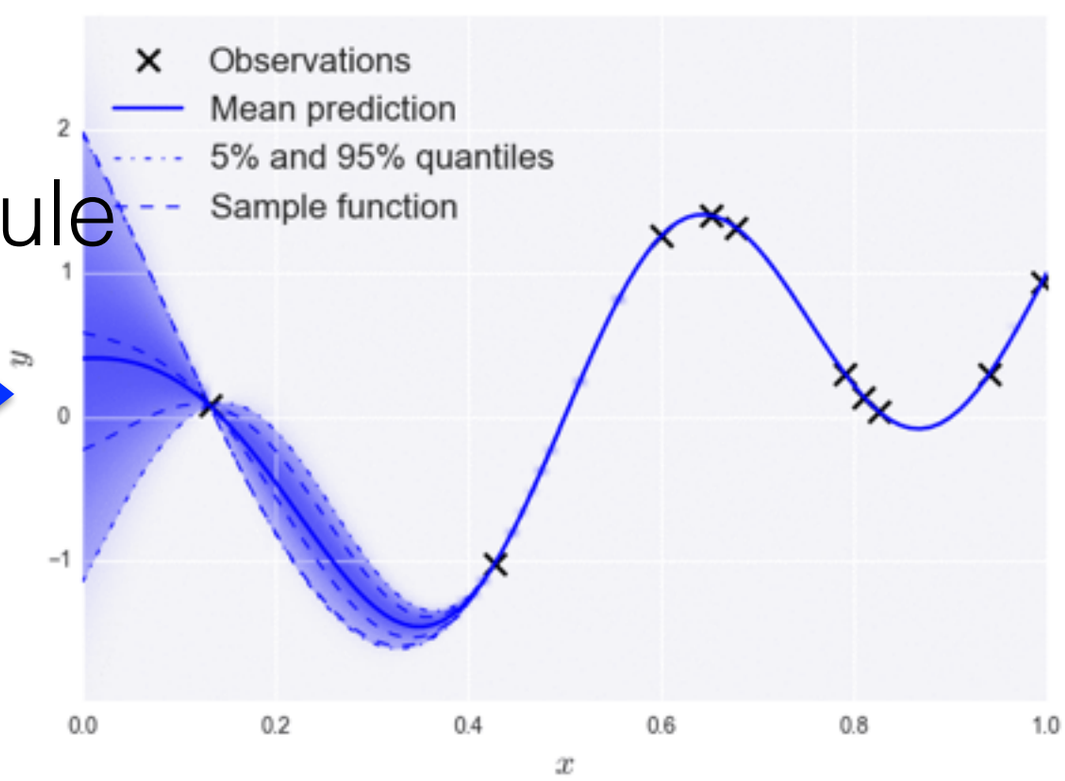
- This encodes are beliefs about the model output after seeing the data.

- Predictions require a Cholesky decomposition.

Gaussian process regression



Bayes rule



Prior GP

Posterior GP

PREDICTIVE
SCIENCE LABORATORY

The point predictive distribution

- Posterior GP:

$$p(f(\cdot) \mid \mathbf{X}, \mathbf{f}) = \mathrm{GP}\Big(f(\cdot) \mid \tilde{m}(\cdot), \tilde{k}(\cdot,\cdot)\Big),$$

- Looking at just one point, we get the *point predictive distribution:*

$$p(y \mid \mathbf{x}, \mathbf{X}, \mathbf{f}) = \mathcal{N}\left(y \mid \tilde{m}(\mathbf{x}), \tilde{\sigma}^2(\mathbf{x})\right),$$

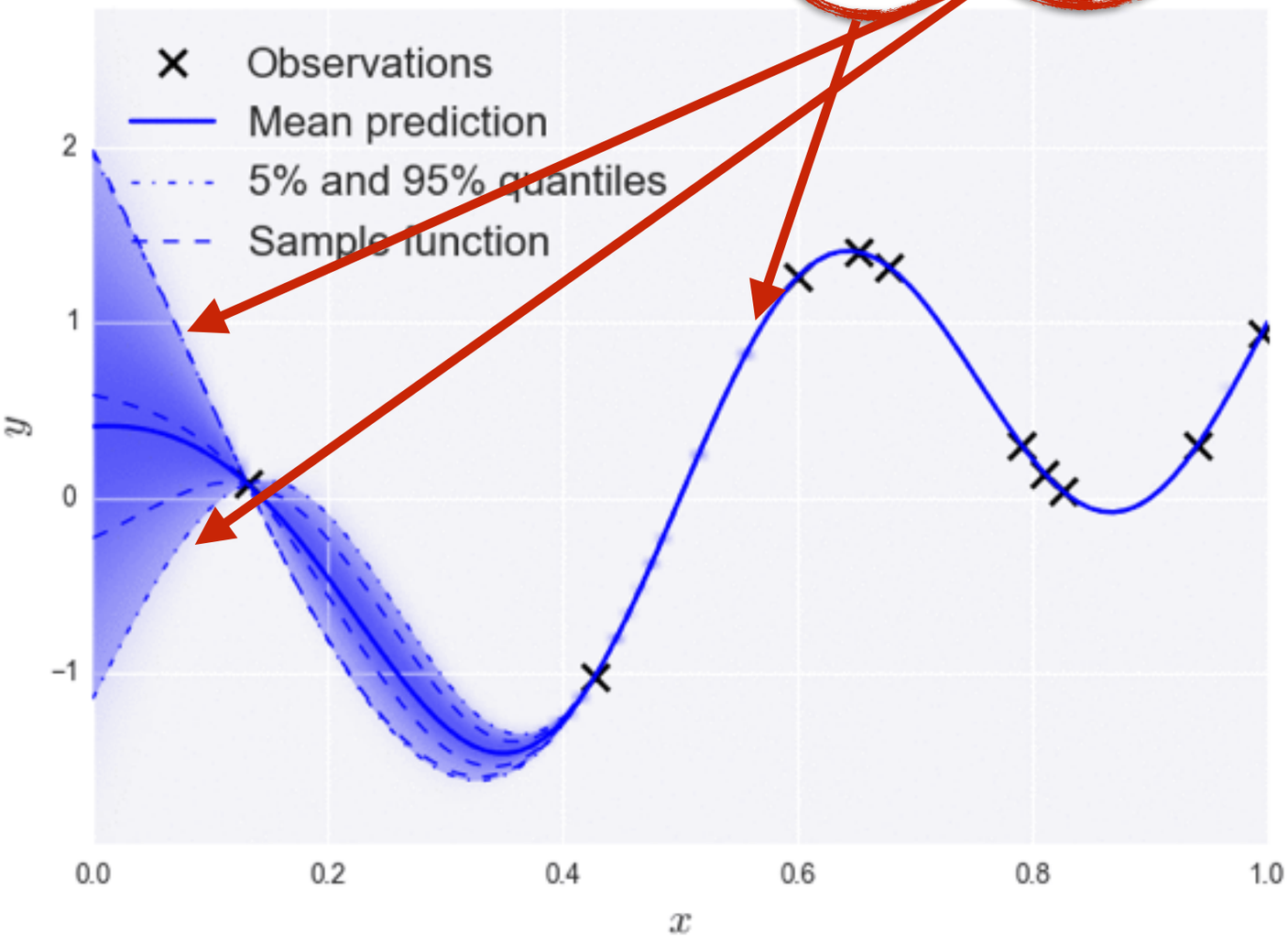
$$\tilde{\sigma}^2(\mathbf{x}) = \tilde{k}(\mathbf{x}, \mathbf{x}).$$

- You may use the mean as a surrogate.

Gaussian process regression

$$p(y \mid \mathbf{x}, \mathbf{X}, \mathbf{f}) = \mathcal{N}\left(y \mid \tilde{m}(\mathbf{x}), \tilde{\sigma}^2(\mathbf{x})\right),$$

$$f(\mathbf{x}) = \tilde{m}(\mathbf{x}) \pm 2\tilde{\sigma}(\mathbf{x})$$



GP Regression 2: With Noise

- Assume that we have observed:

$$\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\},$$

$$\mathbf{y} = \{y_1, \ldots, y_N\}$$

- where *y* is a noisy measurement of the ideal *f(x)* (MD simulation).

- We need to model the measurement process using a likelihood (typically Gaussian):

$$p(y_i \mid f(\mathbf{x}_i)) = \mathcal{N}\left(y_i \mid f(\mathbf{x}_i), \sigma^2\right)$$

Noise (likelihood) variance

Gaussian process regression - Noisy observations

- The posterior GP, changes to:

$$p(f(\cdot)\,|\,\mathbf{X},\mathbf{f},\sigma^2) = \mathrm{GP}\Big(f(\cdot)\,|\,\tilde{m}(\cdot),\tilde{k}(\cdot,\cdot)\Big),$$

$$\tilde{m}(\mathbf{x}) = m(\mathbf{x}) + \mathbf{K}(\mathbf{x},\mathbf{X})(\mathbf{K}+\sigma^2\mathbf{I}_N)^{-1}(\mathbf{f}-\mathbf{m}),$$

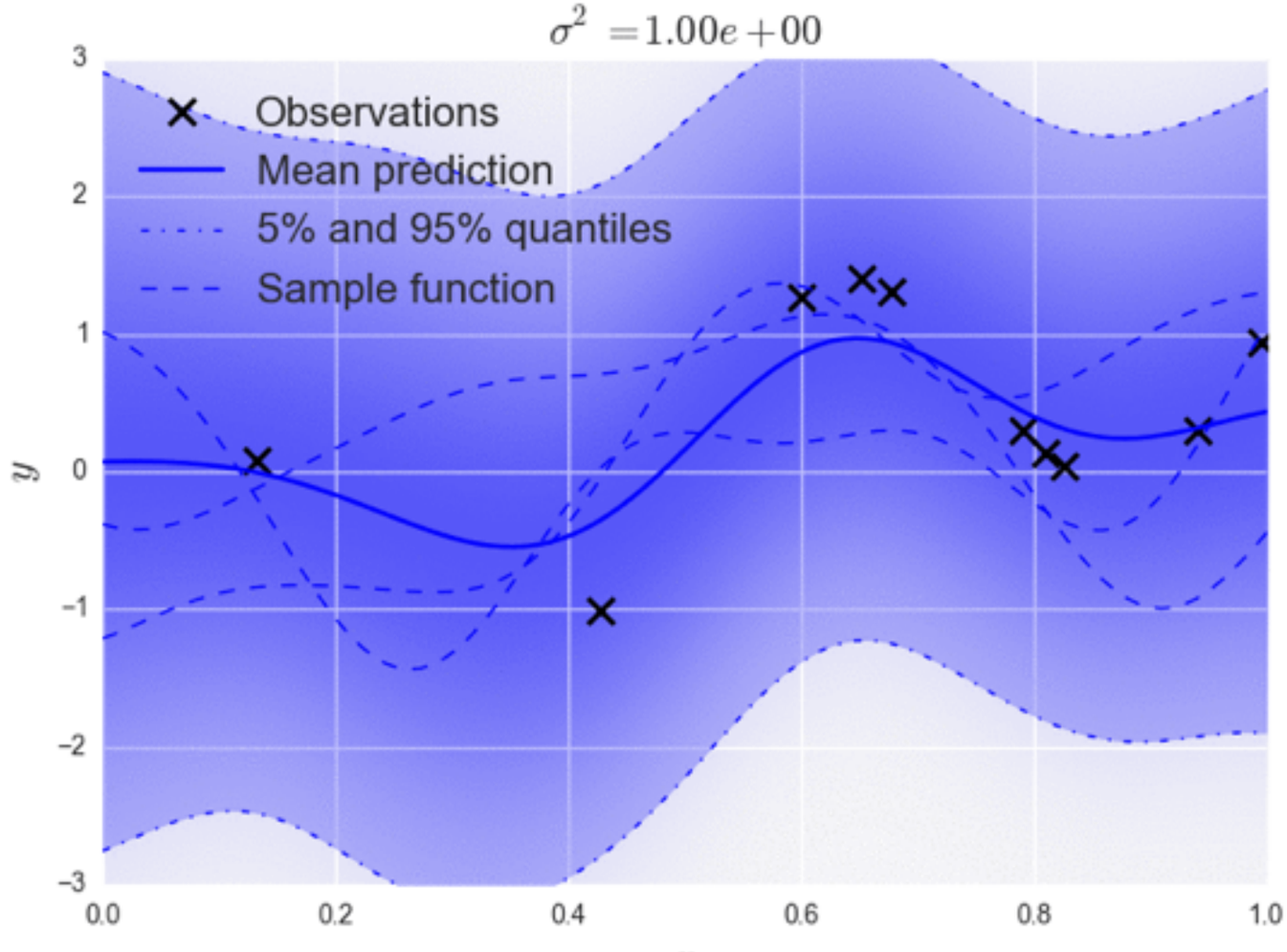
$$\tilde{k}(\mathbf{x},\mathbf{x}') = k(\mathbf{x},\mathbf{x}') - \mathbf{K}(\mathbf{x},\mathbf{X})(\mathbf{K}+\sigma^2\mathbf{I}_N)^{-1}\mathbf{K}(\mathbf{X},\mathbf{x}')$$

- and the point predictive distribution to:

$$p(y\,|\,x,\mathbf{X},\mathbf{f}) = \mathcal{N}\Big(y\,|\,\tilde{m}(\mathbf{x}),\tilde{\sigma}^2(\mathbf{x})\Big),$$

$$\tilde{\sigma}^2(\mathbf{x}) = \tilde{k}(\mathbf{x},\mathbf{x}) + \sigma^2$$

Gaussian process regression - Noisy observations



Each choice of the noise corresponds to a different interpretation of the data.

PREDICTIVE SCIENCE LABORATORY

Noise improves numerical stability

- It is common to use small noise even if there is not any in the data.

- Cholesky fails when covariance is close to being semi-positive definite.

- Adding a small noise improves numerical stability.

- It is known as the “jitter” or as the “nugget” in this case.

Example, Part I, Questions 1-5

Model Selection for GP regression

- Our prior assumptions were conditional mean and covariance parameters:

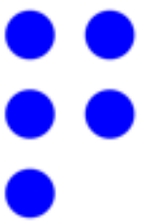
$$p(f(\cdot) \mid \theta) = \mathrm{GP}\left(f(\cdot) \mid m(\cdot;\theta), k(\cdot,\cdot;\theta)\right)$$

- Observations are conditional on the noise level:

$$p(y \mid f(x), \sigma^2) = \mathcal{N}\left(y \mid f(x), \sigma^2\right)$$

- Thus, the (marginal) *likelihood* of all the observations is:

$$p(\mathbf{y} \mid \mathbf{X}, \theta, \sigma^2) = p(\mathbf{y} \mid \mathbf{X}, \theta, \sigma^2) = \mathcal{N}\left(\mathbf{y} \mid \mathbf{m}, \mathbf{K} + \sigma^2 \mathbf{I}_N\right)$$



Model Selection for GP regression

- The (marginal) *likelihood* of all the observations is:

$$p(\mathbf{y} \mid \mathbf{X}, \theta, \sigma^2) = \mathcal{N}\left(\mathbf{y} \mid \mathbf{m}, \mathbf{K} + \sigma^2 \mathbf{I}_N\right)$$

- To complete the prior specification, we must give:

$$p(\theta, \sigma^2) \sim p(\theta, \sigma^2).$$

- Then, after seeing the data, our beliefs about the parameters should change to:

$$p(\theta, \sigma^2 \mid \mathbf{X}, \mathbf{y}) \propto p(\mathbf{y} \mid \mathbf{X}, \sigma^2) p(\theta, \sigma^2)$$

Model Selection for GP regression

- After seeing the data, our beliefs about the parameters are:

$$p(\theta,\sigma^2 \mid \mathbf{X},\mathbf{y}) \propto p(\mathbf{y} \mid \mathbf{X},\sigma^2)p(\theta,\sigma^2)$$

- Ideally, we would sample from this posterior with MCMC.

- Alternatively, we can find the MAP estimate of the parameters:

$$\theta^*,(\sigma^*)^2 = \text{argmax}_{\theta,\sigma} \left\{ \log p(\mathbf{y} \mid \mathbf{X},\sigma^2) + \log p(\theta,\sigma^2) \right\}$$

Model Selection for GP regression

- MAP estimate of the parameters:

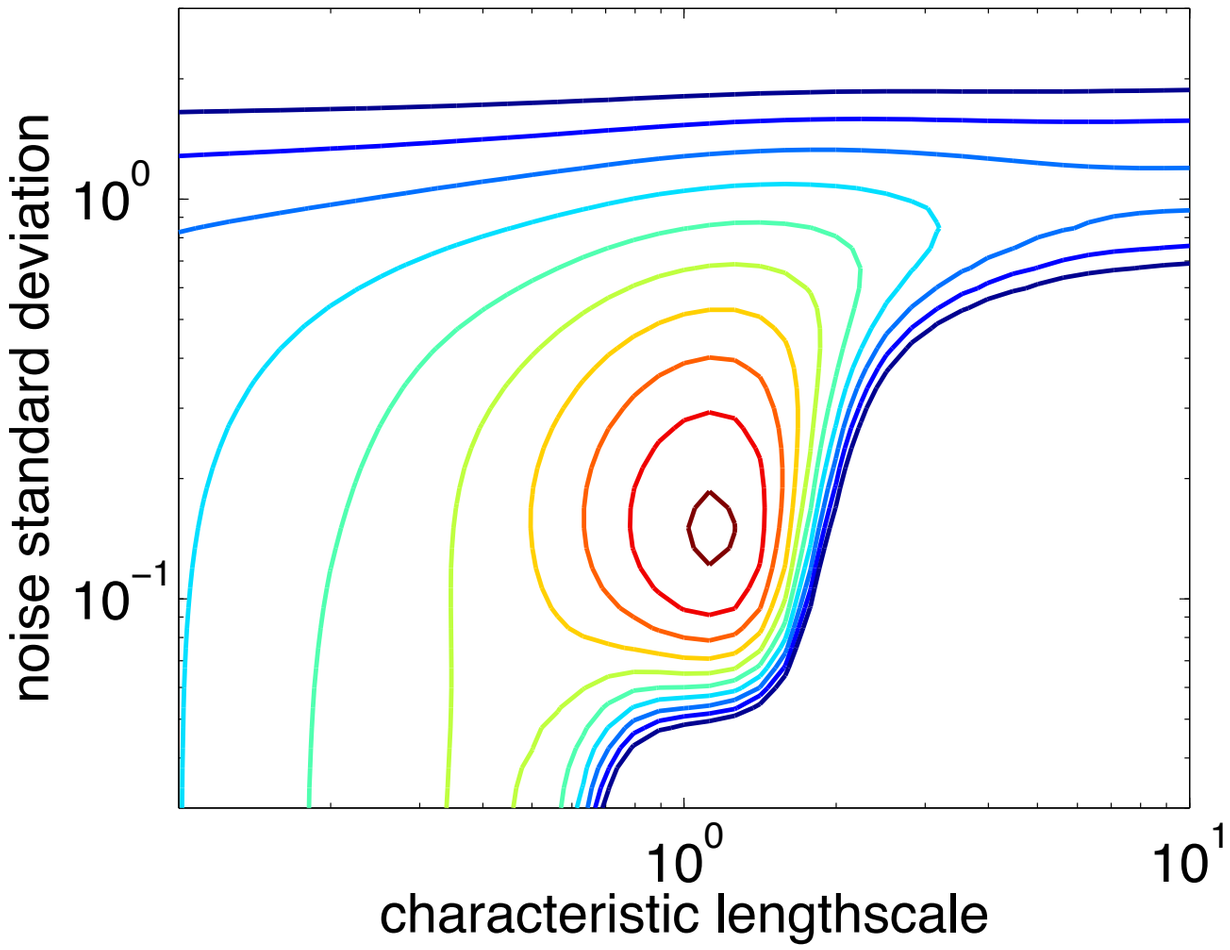
$$\theta^*,(\sigma^*)^2 = \text{argmax}_{\theta,\sigma}\left\{\log p(\mathbf{y} \mid \mathbf{X},\sigma^2) + \log p(\theta,\sigma^2)\right\}$$

- If our prior assumptions are vague, then

$$\log p(\theta,\sigma^2) = \text{const}$$

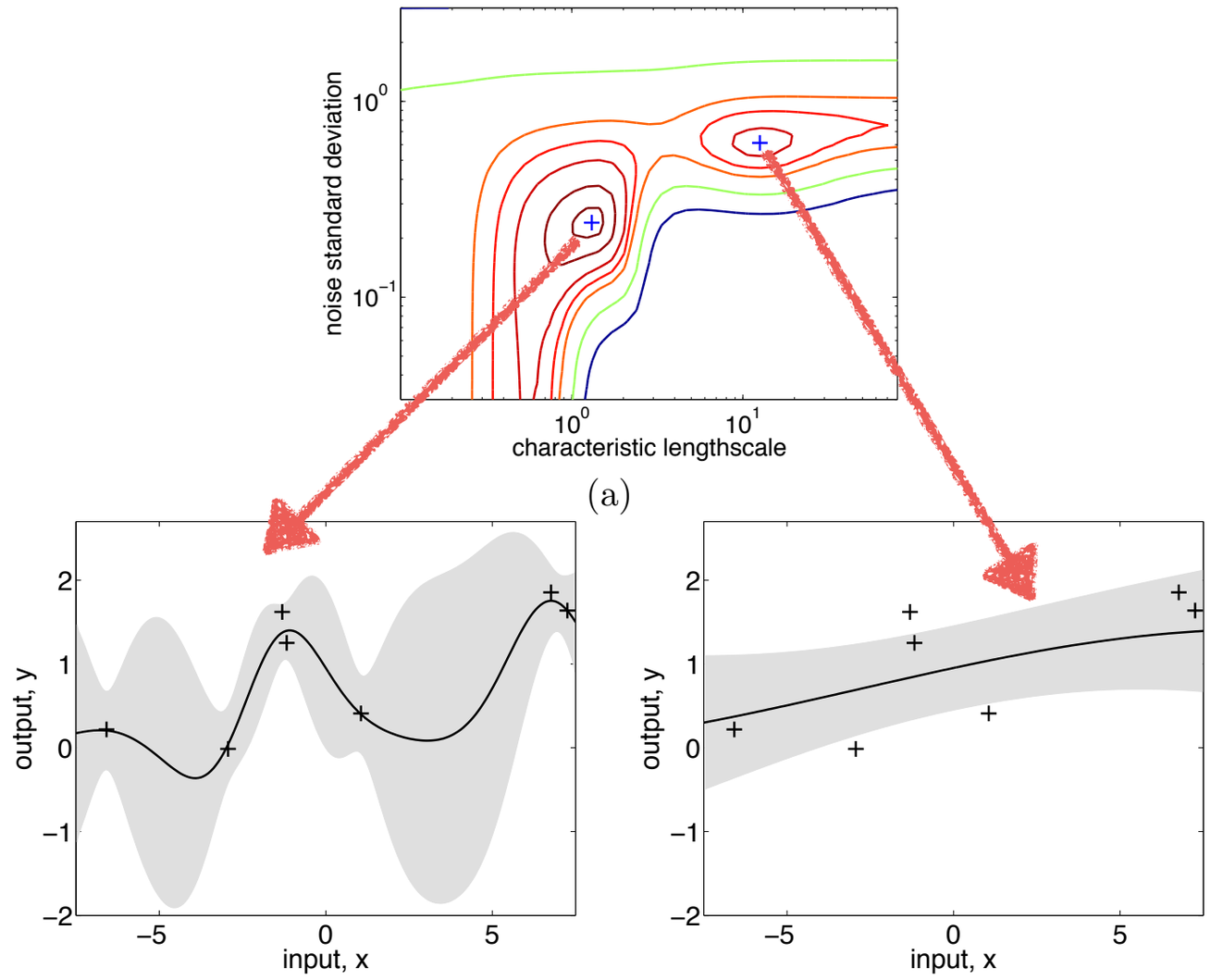
- and we are effectively just maximizing the likelihood.

Model Selection for GP regression



Contour plot of marginal likelihood for specific example in Rasmussen (2006)

Careful: Different optima correspond to different interpretations



Contour plot of marginal likelihood for specific example in Rasmussen (2006)

Example, Part I,
Questions 6-8
Example, Part II&III