An Introduction to Gaussian Process Regression

Eric Schearer
Intelligent Control Systems

Overview

1. What is regression?
2. Why are some regression problems hard?
3. What is a Gaussian process?
4. How does Gaussian process regression work?
5. How does Gaussian process regression compare to other nonlinear regression methods?

Overview

1. What is regression?
2. Why are some regression problems hard?
3. What is a Gaussian process?
4. How does Gaussian process regression work?
5. How does Gaussian process regression compare to other nonlinear regression methods?

General Regression Problem

$$y = f(\mathbf{x}) + \epsilon$$

$f(\mathbf{x})$	scalar function that we are trying to find
$\mathbf{x} \in \mathbb{R}^{D \times 1}$	vector of inputs to the function
D	dimension of the input space
y	scalar measured output of the function, a continuous variable
$\epsilon \sim \mathcal{N}(0, \sigma^2)$	normally distributed measurement noise

General Regression Problem

$$y = f(\mathbf{x}) + \epsilon$$

$$\mathcal{D} = \{(\mathbf{x}_i, y_i) | i = 1, \ldots, n\} = (X, \mathbf{y})$$ training set

n number of samples

$\mathbf{y} \in \mathbb{R}^{n \times 1}$ column vector of the series of measured outputs

$X \in \mathbb{R}^{D \times n}$ design matrix with columns that are the individual input vectors

General Regression Problem

Given:

$$\mathcal{D} = \{(\mathbf{x}_i, y_i) | i = 1, \ldots, n\} = (X, \mathbf{y})$$ training set

$$\mathbf{x}_*$$ new input

Predict:

$$y_*$$ new output

$$p(y_*)$$ distribution of the new output

Linear Model

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{w}$$

$\mathbf{w} \in \mathbb{R}^{D\times 1}$ vector of parameters

$$\mathbf{x} = [1\ x]^\top$$

Overview

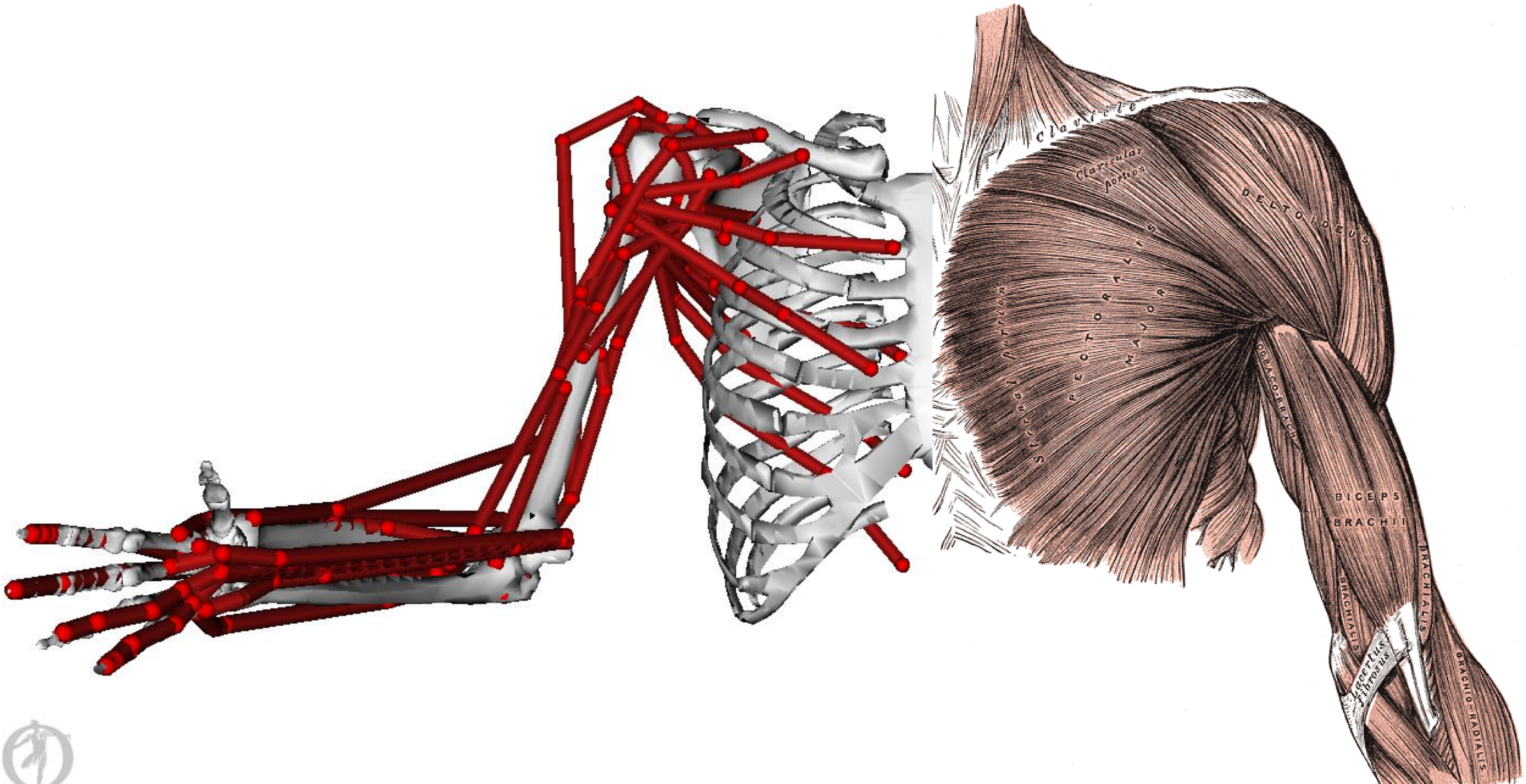
1. What is regression?
2. Why are some regression problems hard?
3. What is a Gaussian process?
4. How does Gaussian process regression work?
5. How does Gaussian process regression compare to other nonlinear regression methods?

Inverse Dynamics of the Human Arm

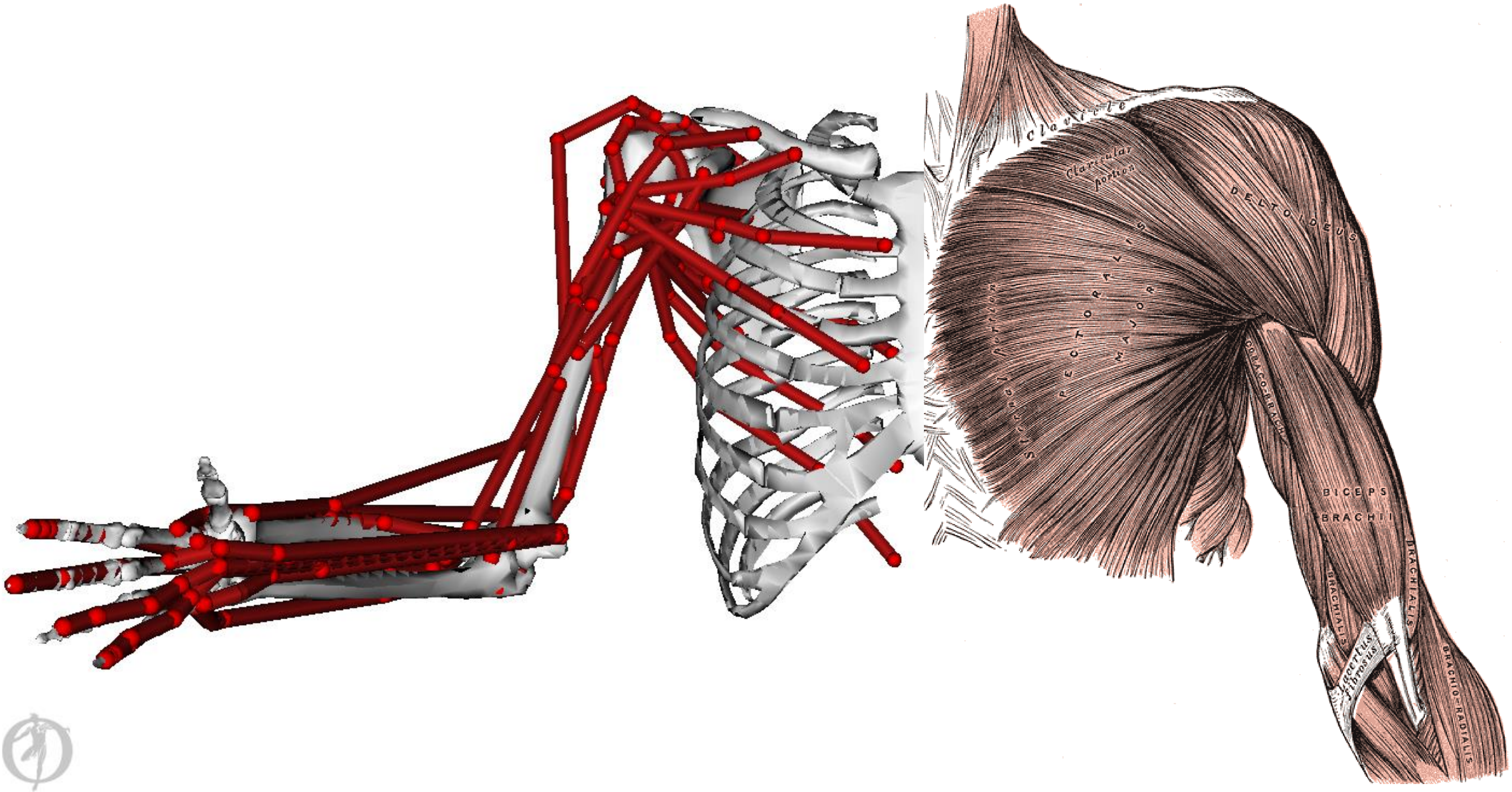
Given: shoulder and elbow joint angles, velocities, and accelerations

Predict: shoulder and elbow joint torques to drive arm along a trajectory

Complexity

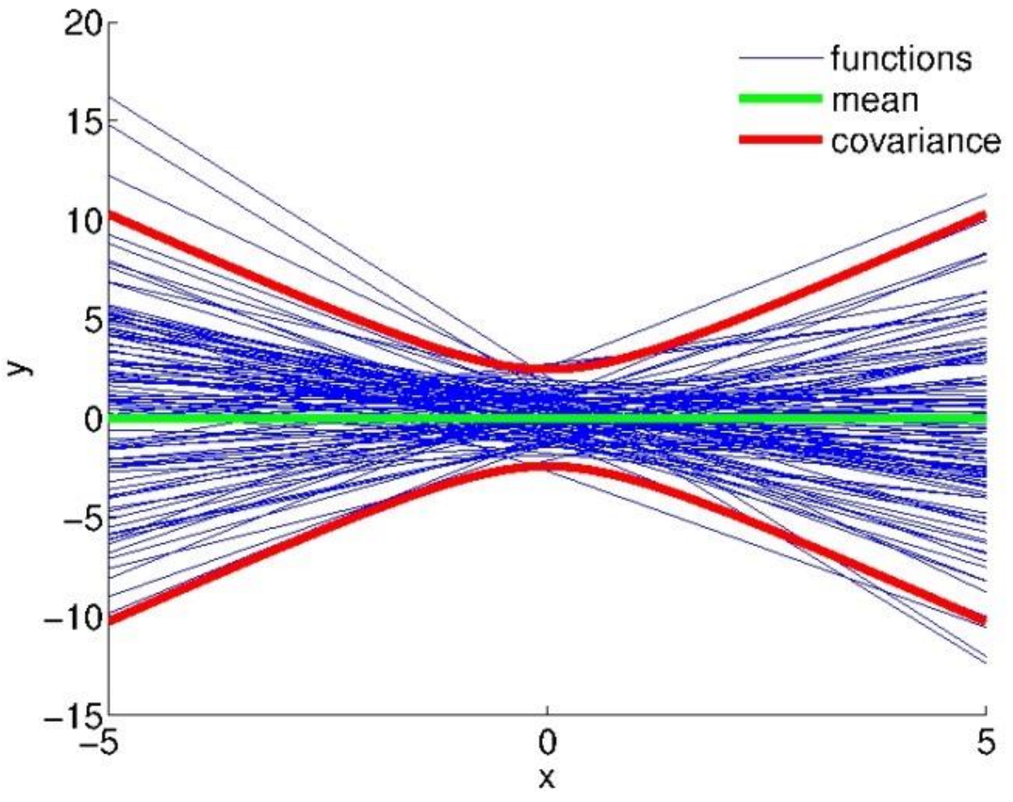

Clavicle
Clavicular portion
DELTOIDEUS
PECTORALIS MAJOR
BICEPS BRACHII
BRACHIALIS
BRACHIO-RADIALIS
Lacertus fibrosus

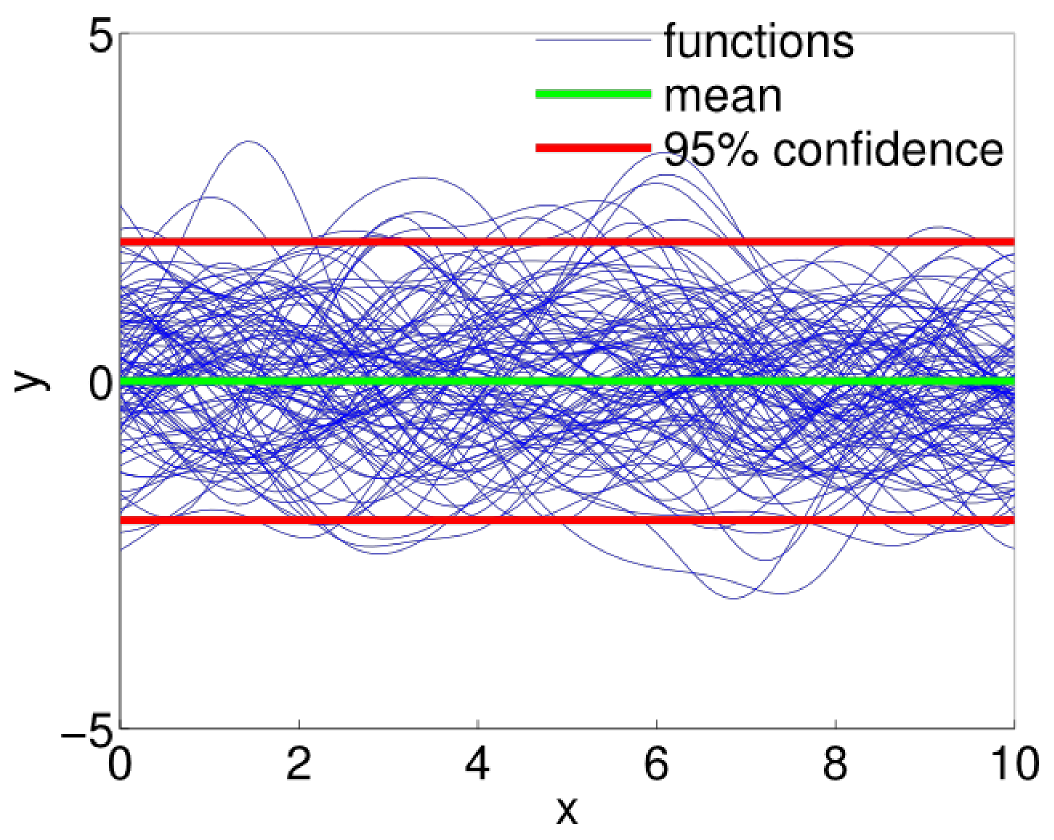
Identifiability?



Two Approaches

1. Restrict the possible functions
   - maybe not flexible enough
   - maybe overfit
2. Consider all functions with some more likely than others
   - How in the world are you going to compute this?
   - Gaussian processes to the rescue

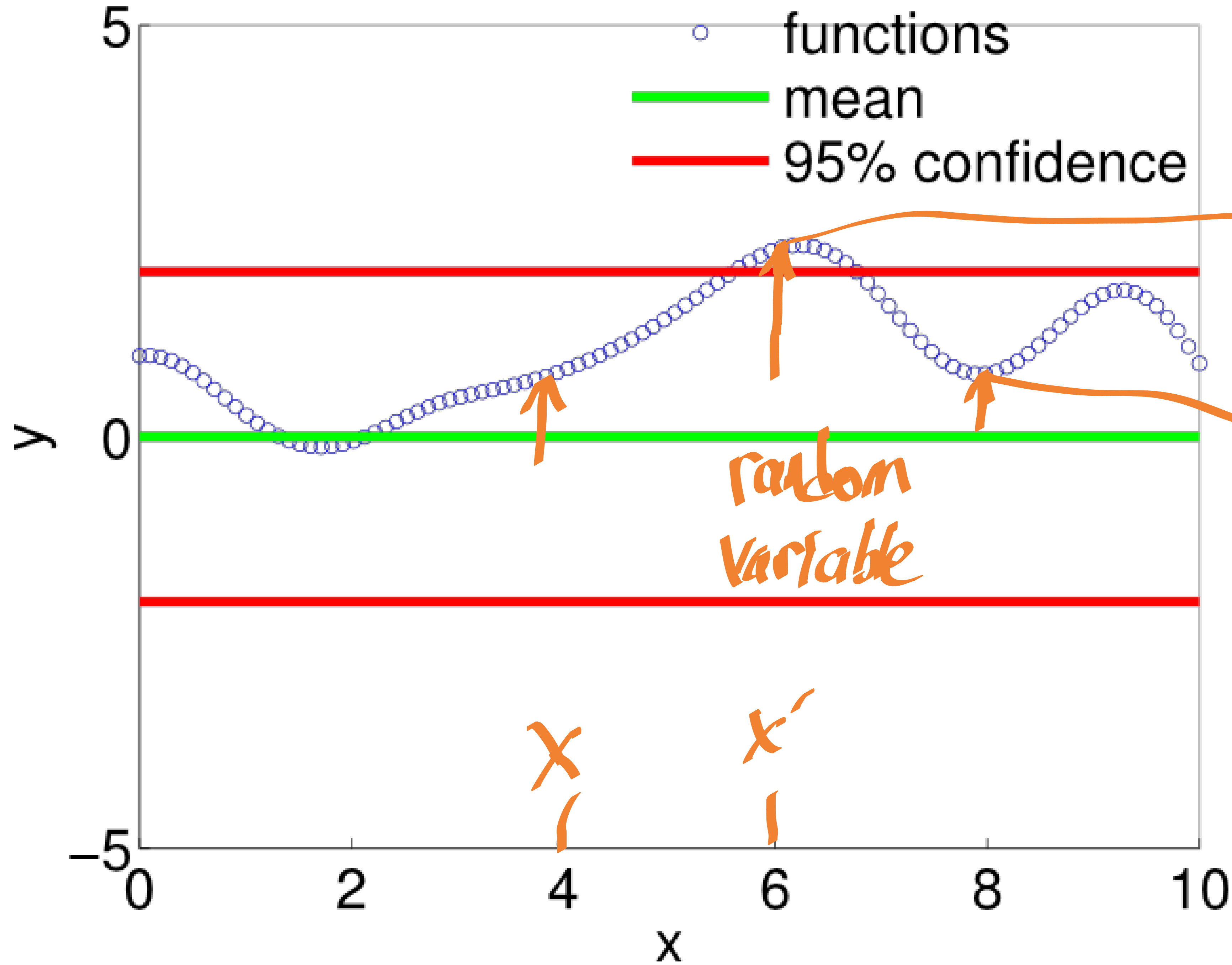




Overview

1. What is regression?
2. Why are some regression problems hard?
3. What is a Gaussian process?
4. How does Gaussian process regression work?
5. How does Gaussian process regression compare to other nonlinear regression methods?

Gaussian Process



$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{21} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$$

A function is a collection of ∞ random variables
A vector might be a finite representation of a function.

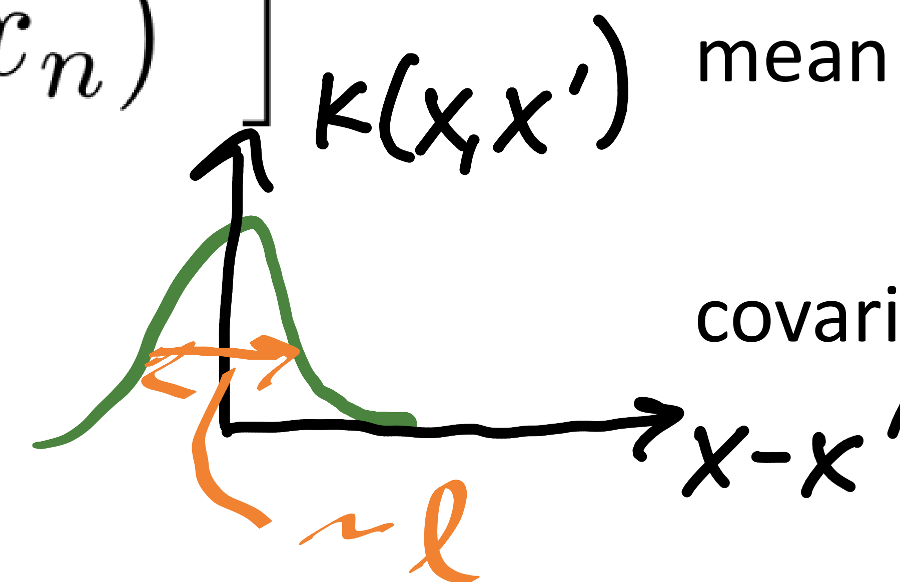
Gaussian process: a collection of random variables, any finite number of which have a joint Gaussian distribution

Draw a Function from a Gaussian Process

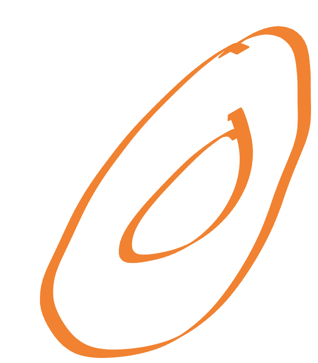
$$m(x) = 0$$ mean function

$$\mathbf{m} = \begin{bmatrix} m(x_1) & \dots & m(x_n) \end{bmatrix}$$ mean vector

$$k(x, x') = \exp\left(-\frac{(x - x')^2}{2l^2}\right)$$ covariance function

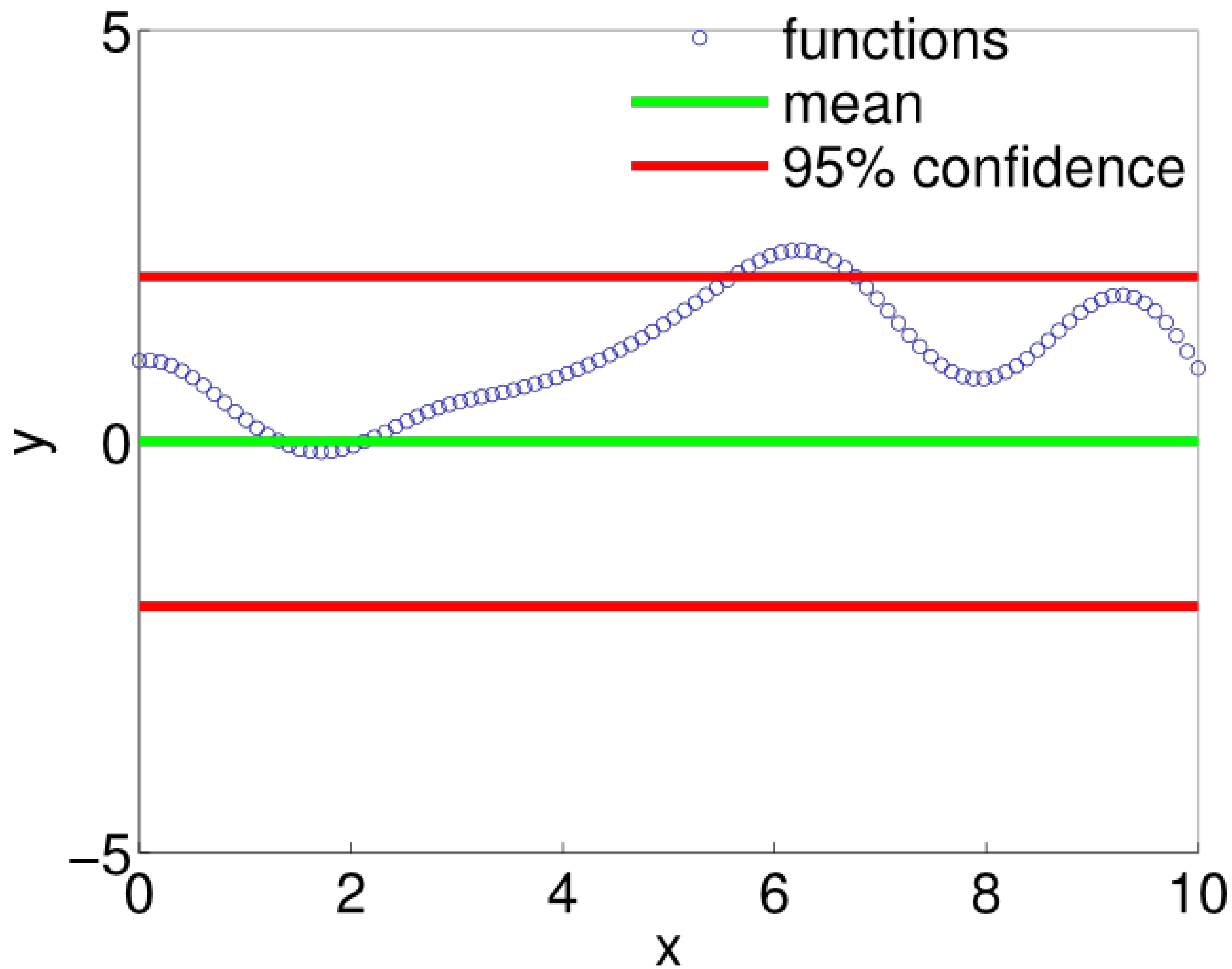


$$K = \begin{bmatrix} k(x_1, x_1) & \dots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \dots & k(x_n, x_n) \end{bmatrix}$$ covariance matrix



```
f=mvnrnd(m,K)
```

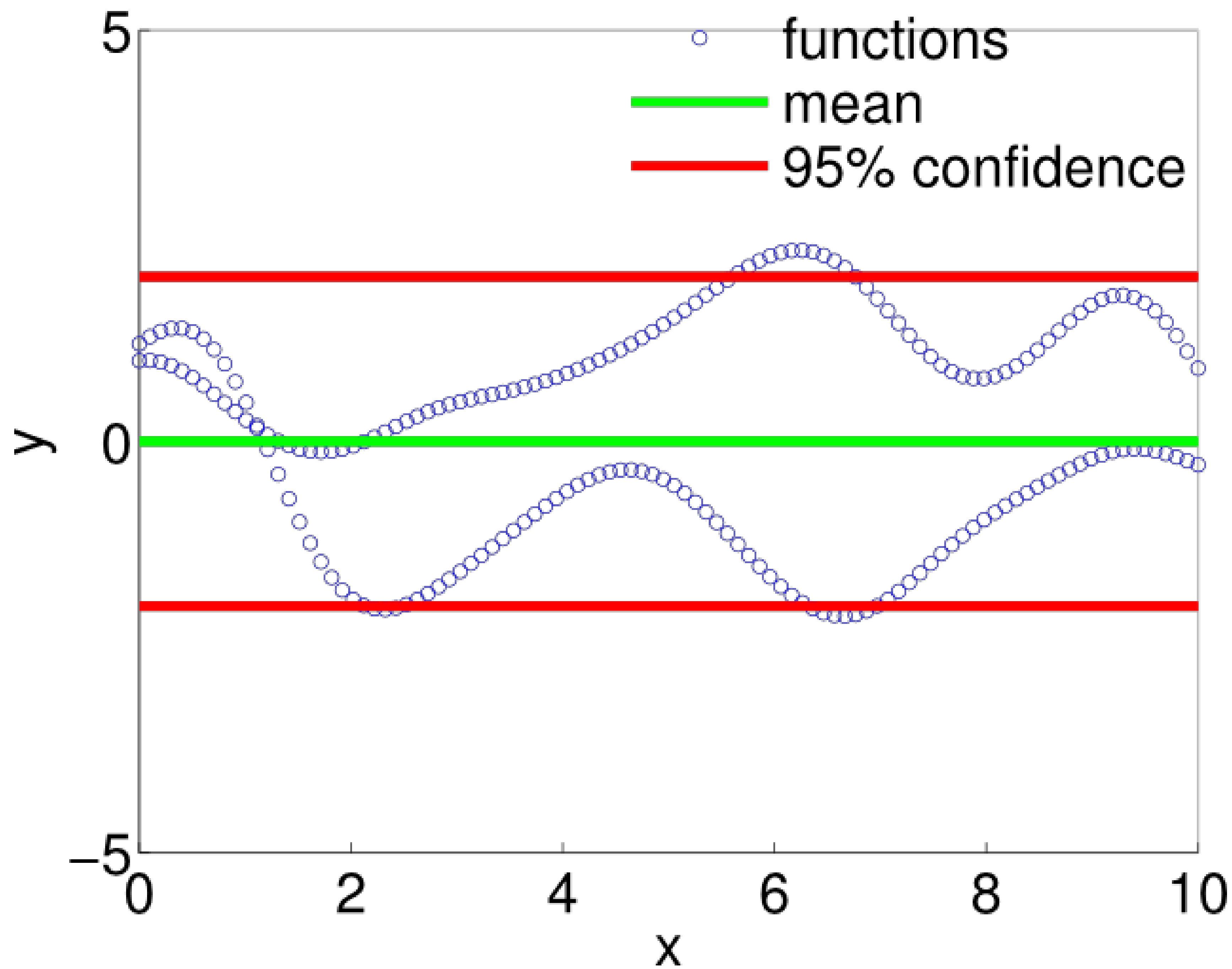
Gaussian Process



$$m(x) = 0 \qquad k(x, x') = \exp\left(-\frac{(x - x')^2}{2l^2}\right)$$

$$\mathbf{m} \in \mathbb{R}^n \qquad K \in \mathbb{R}^{n \times n}$$

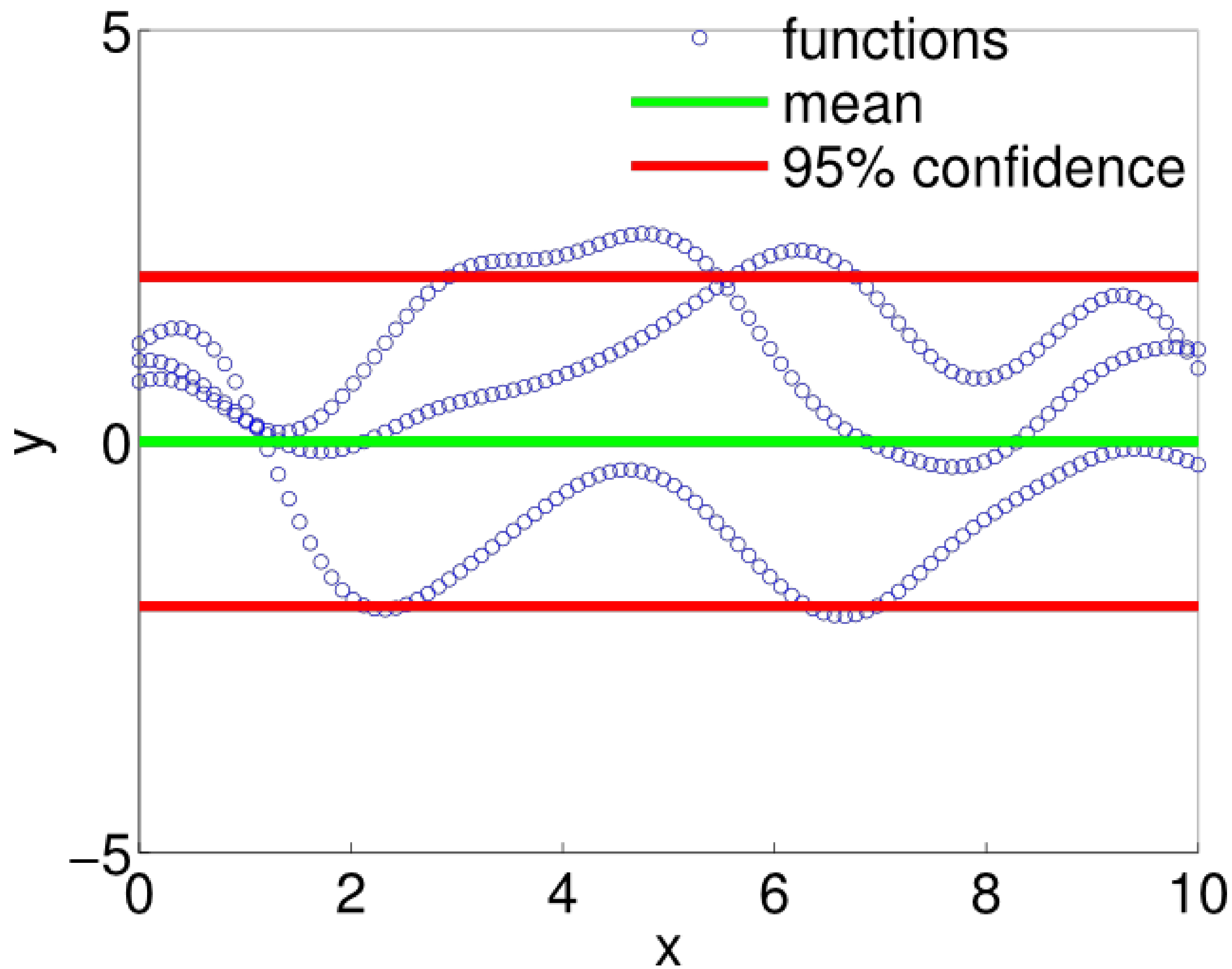
Gaussian Process



$$m(x) = 0 \qquad k(x, x') = \exp\left(-\frac{(x - x')^2}{2l^2}\right)$$

$$\mathbf{m} \in \mathbb{R}^n \qquad K \in \mathbb{R}^{n \times n}$$

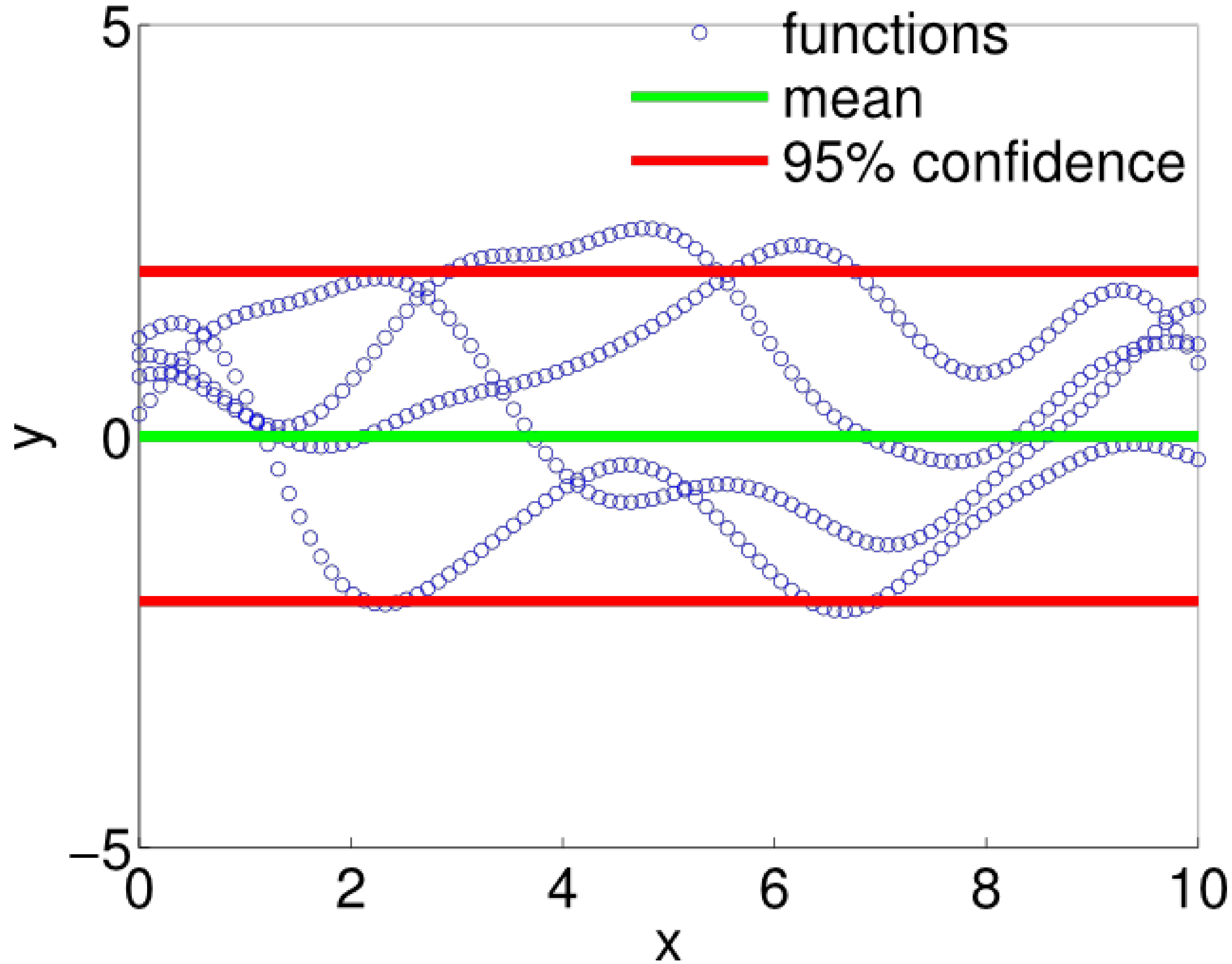
Gaussian Process



$$m(x) = 0 \qquad k(x, x') = \exp\left(-\frac{(x - x')^2}{2l^2}\right)$$

$$\mathbf{m} \in \mathbb{R}^n \qquad K \in \mathbb{R}^{n \times n}$$

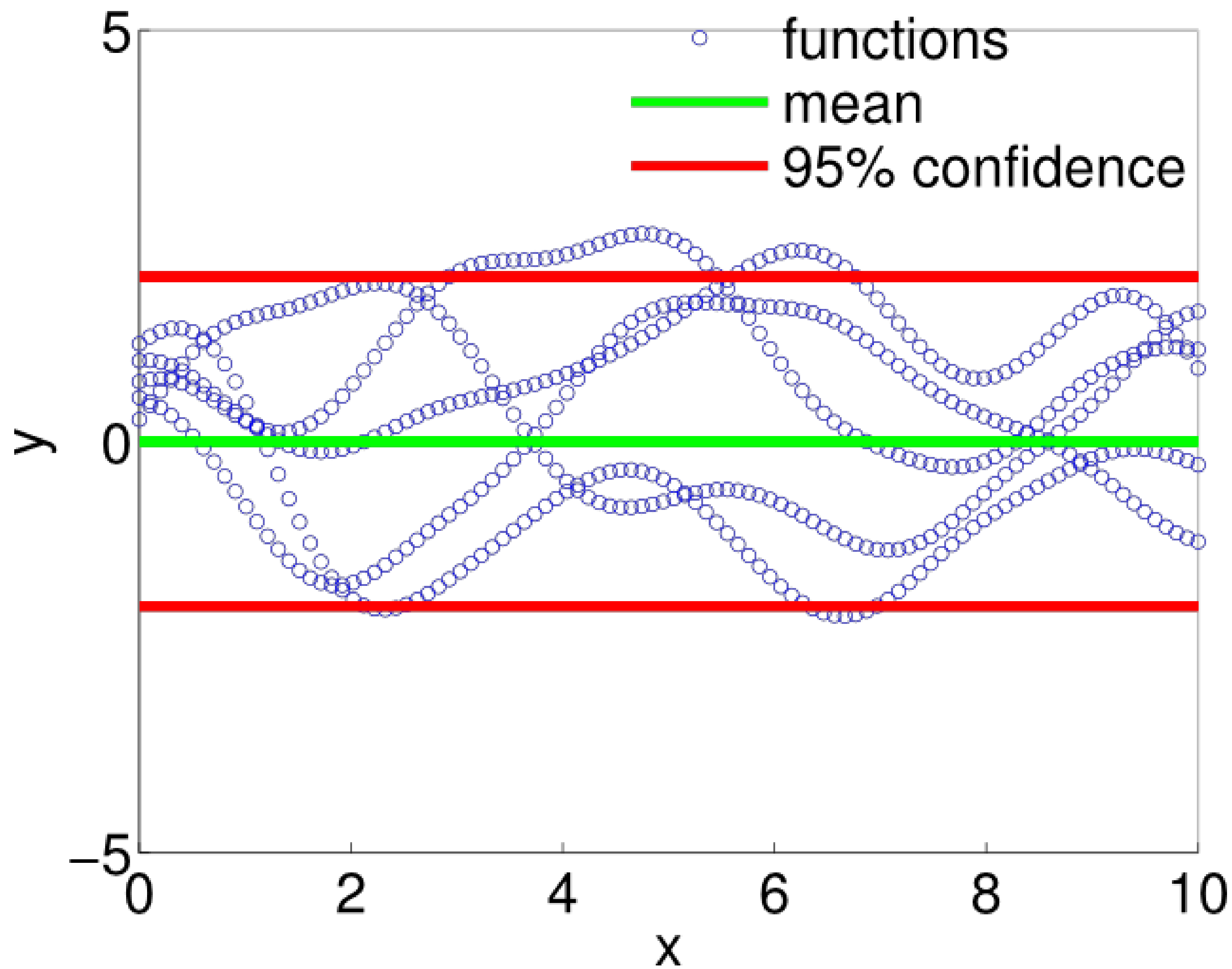
Gaussian Process



$$m(x) = 0 \qquad k(x, x') = \exp\left(-\frac{(x - x')^2}{2l^2}\right)$$

$$\mathbf{m} \in \mathbb{R}^n \qquad K \in \mathbb{R}^{n \times n}$$

Gaussian Process



$$m(x) = 0 \qquad k(x, x') = \exp\left(-\frac{(x - x')^2}{2l^2}\right)$$

$$\mathbf{m} \in \mathbb{R}^n \qquad K \in \mathbb{R}^{n \times n}$$

Gaussian Process

e.g $x \sim N(m, \sigma^2)$

Definition:

A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution.

gaussian process

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$$

drawn from

mean function

Covariance function

mean function

$$m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})]$$

covariance function

$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))]$$

Squared Exponential Covariance Function

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$$

p_2 is the length scale

p_1 is the magnitude or vertical scale

mean function

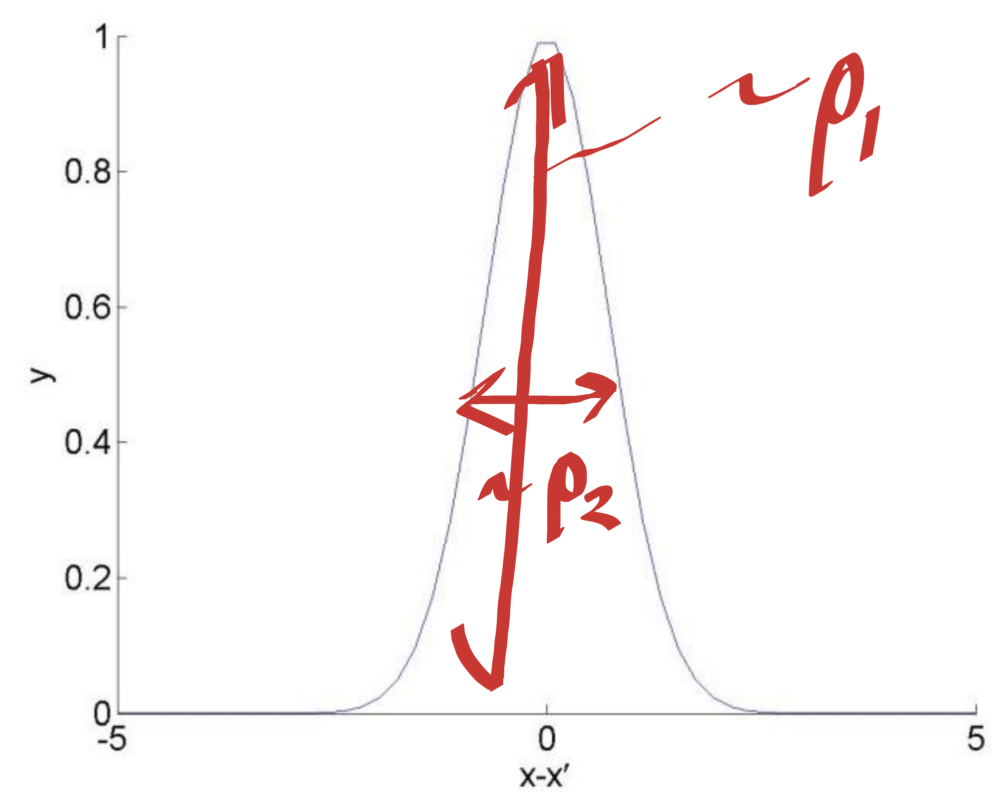
$$m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})] = 0$$

covariance function

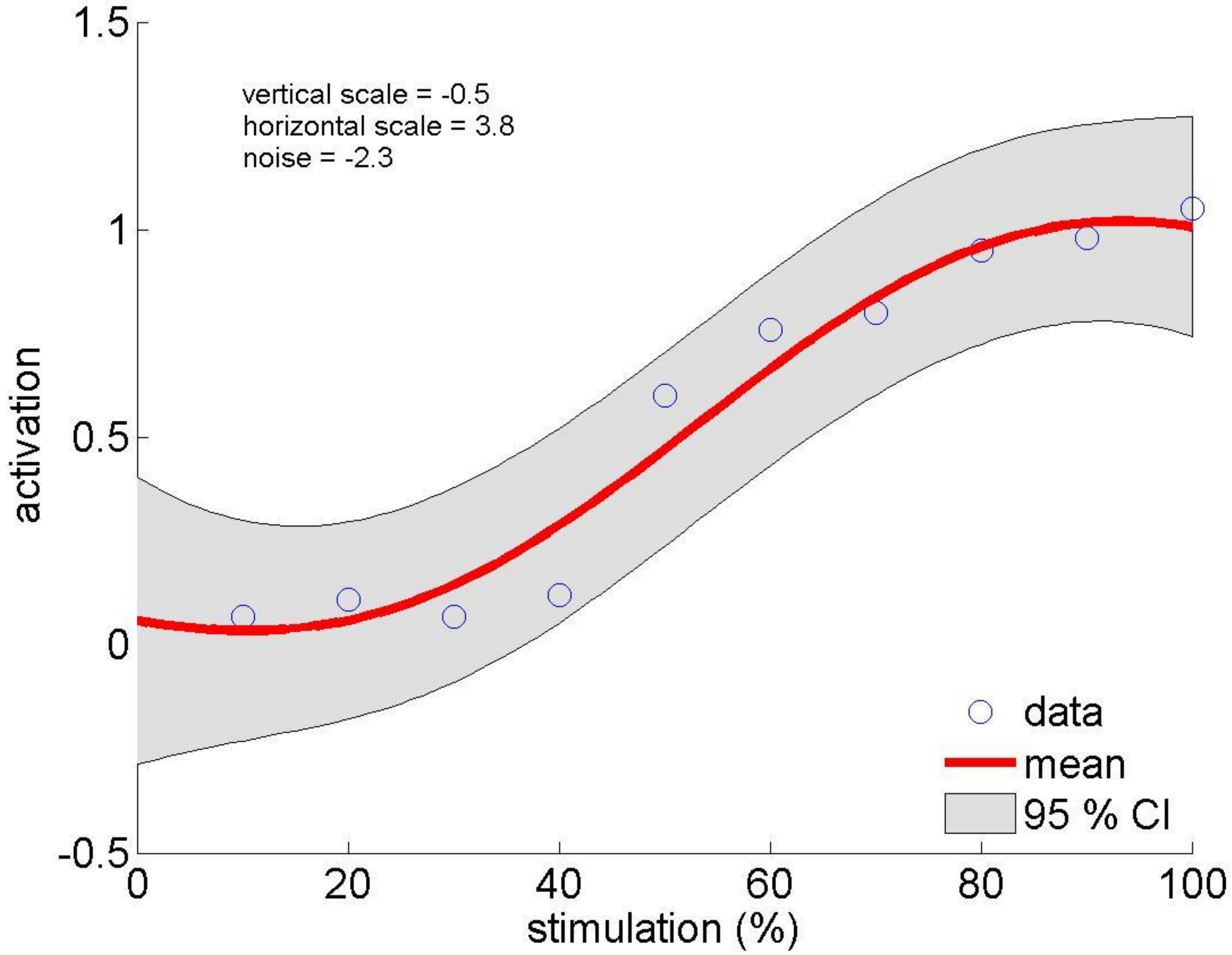
$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))]$$

$$k(\mathbf{x}, \mathbf{x}') = p_1 e^{-\frac{(\mathbf{x}-\mathbf{x}')^\top (\mathbf{x}-\mathbf{x}')}{2p_2^2}}$$

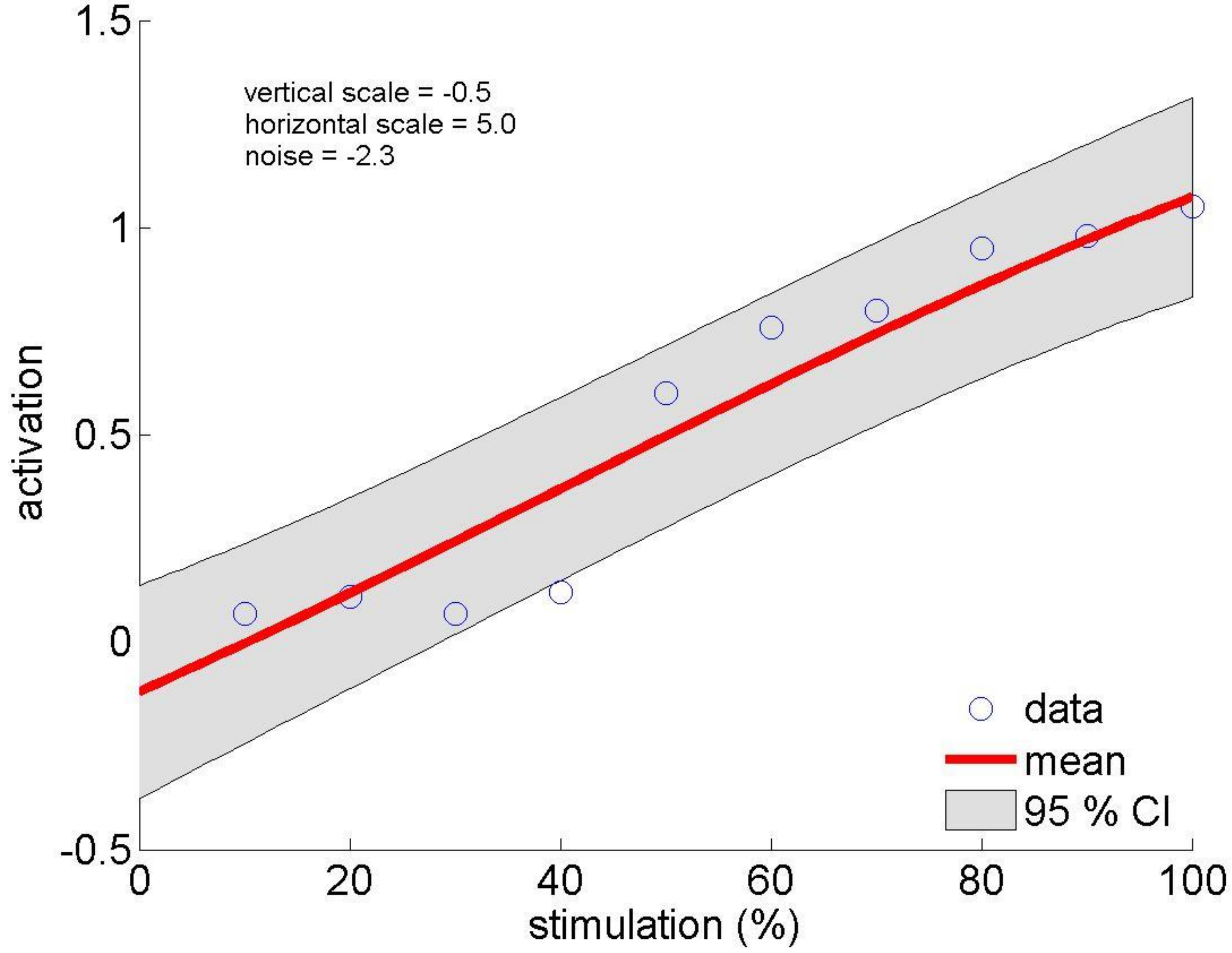
hyperparameters



Changing the Length Scale



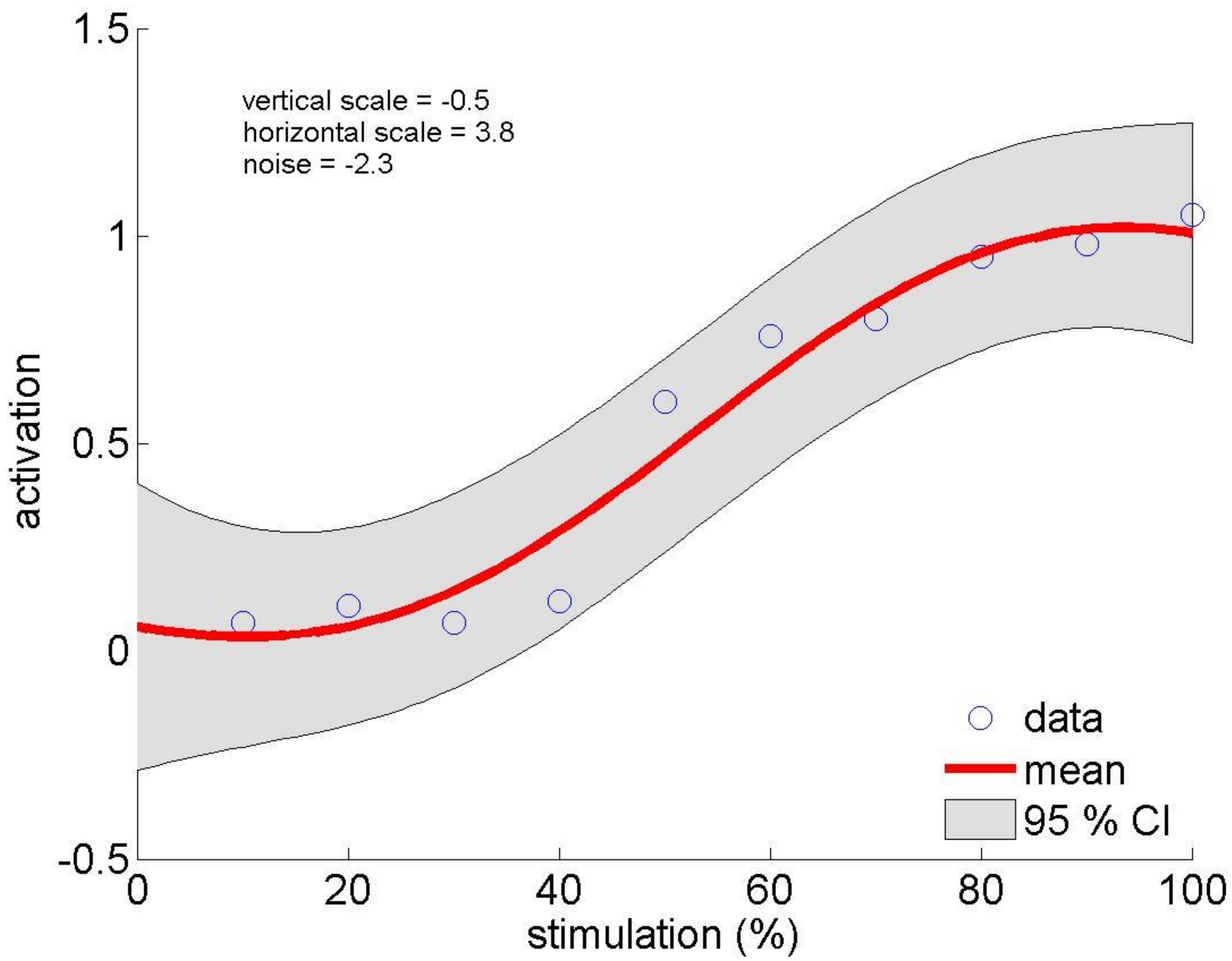
nice length scale



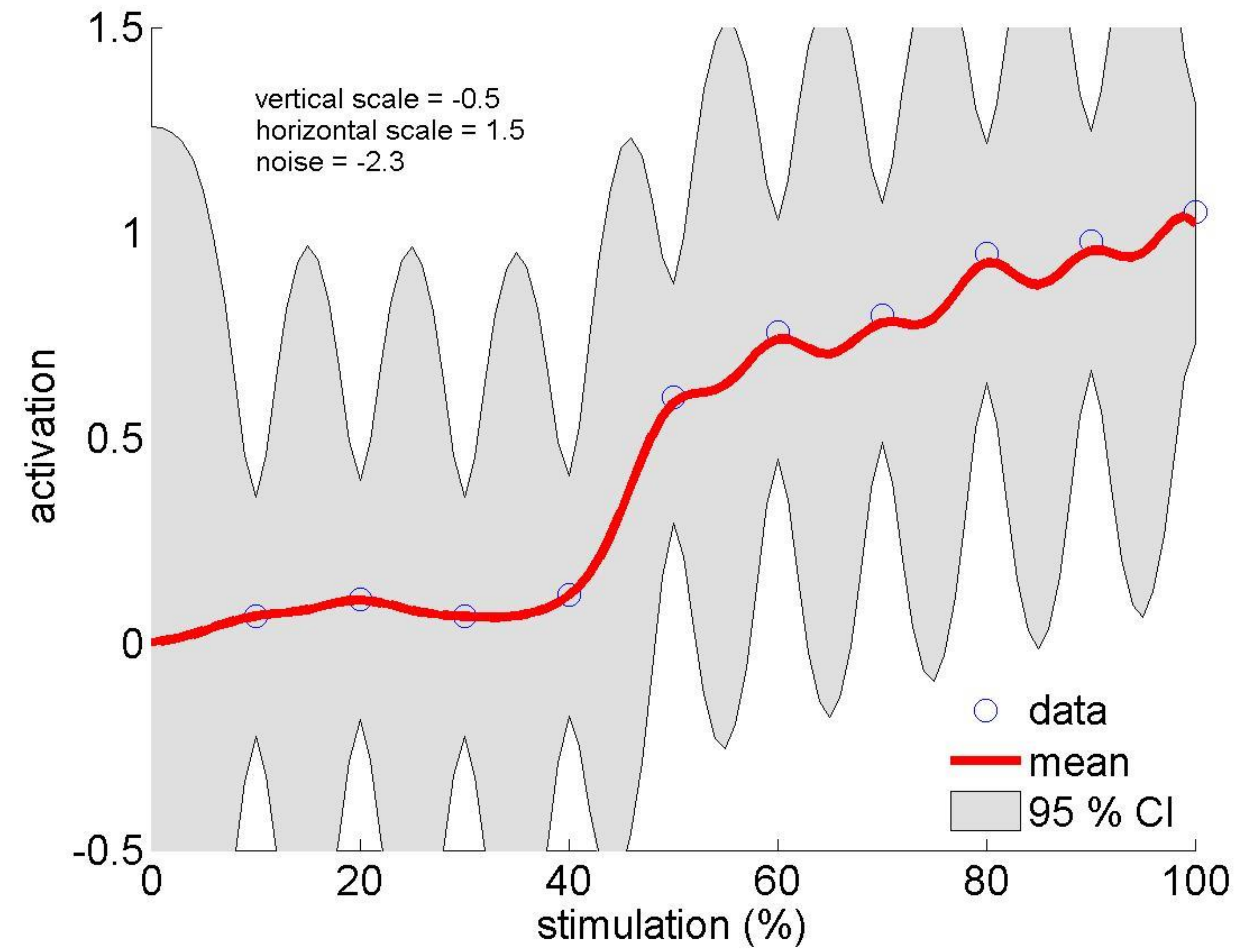
length scale too long

$$k(\mathbf{x}, \mathbf{x}') = p_1 e^{-\frac{(\mathbf{x}-\mathbf{x}')^\top (\mathbf{x}-\mathbf{x}')}{2p_2^2}}$$

Changing the Length Scale



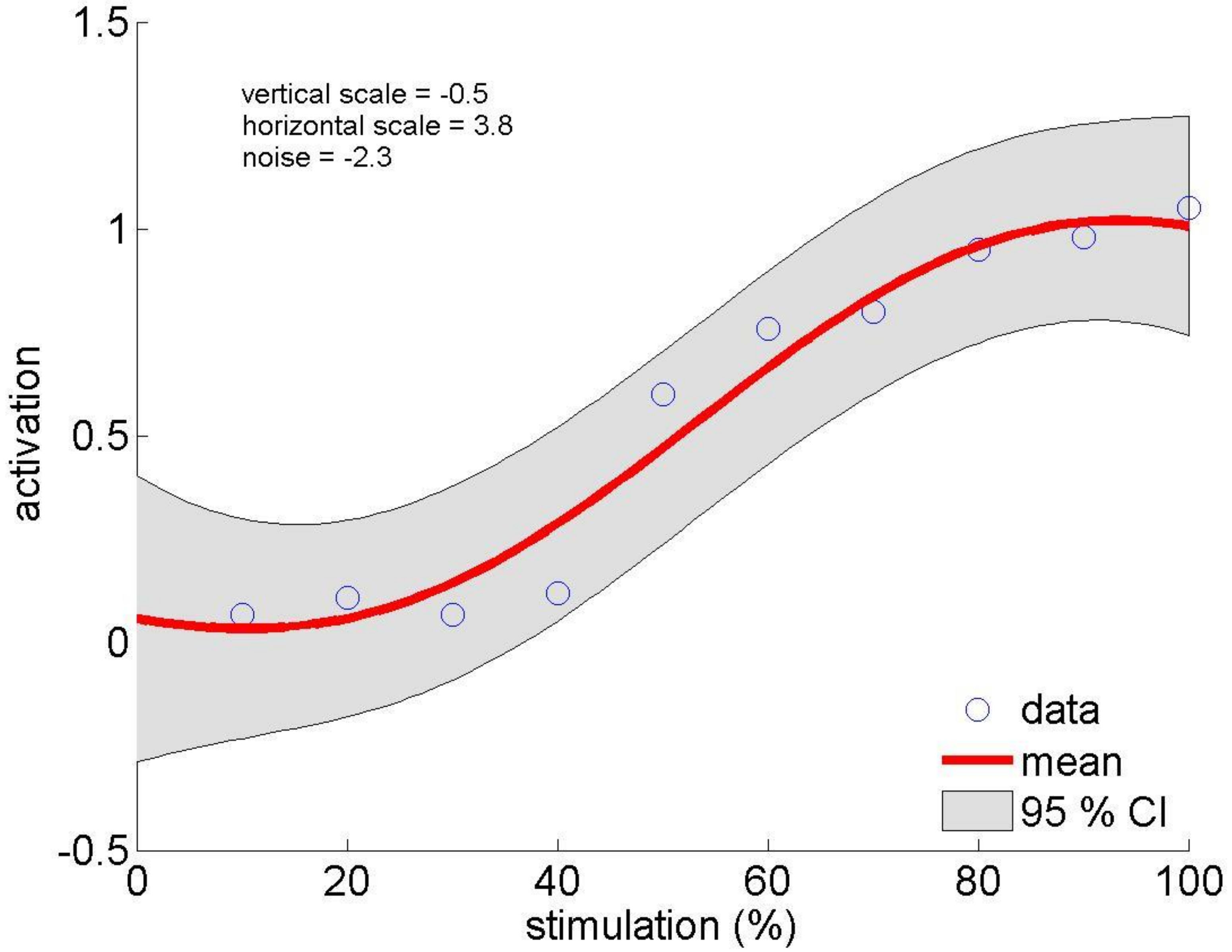
nice length scale



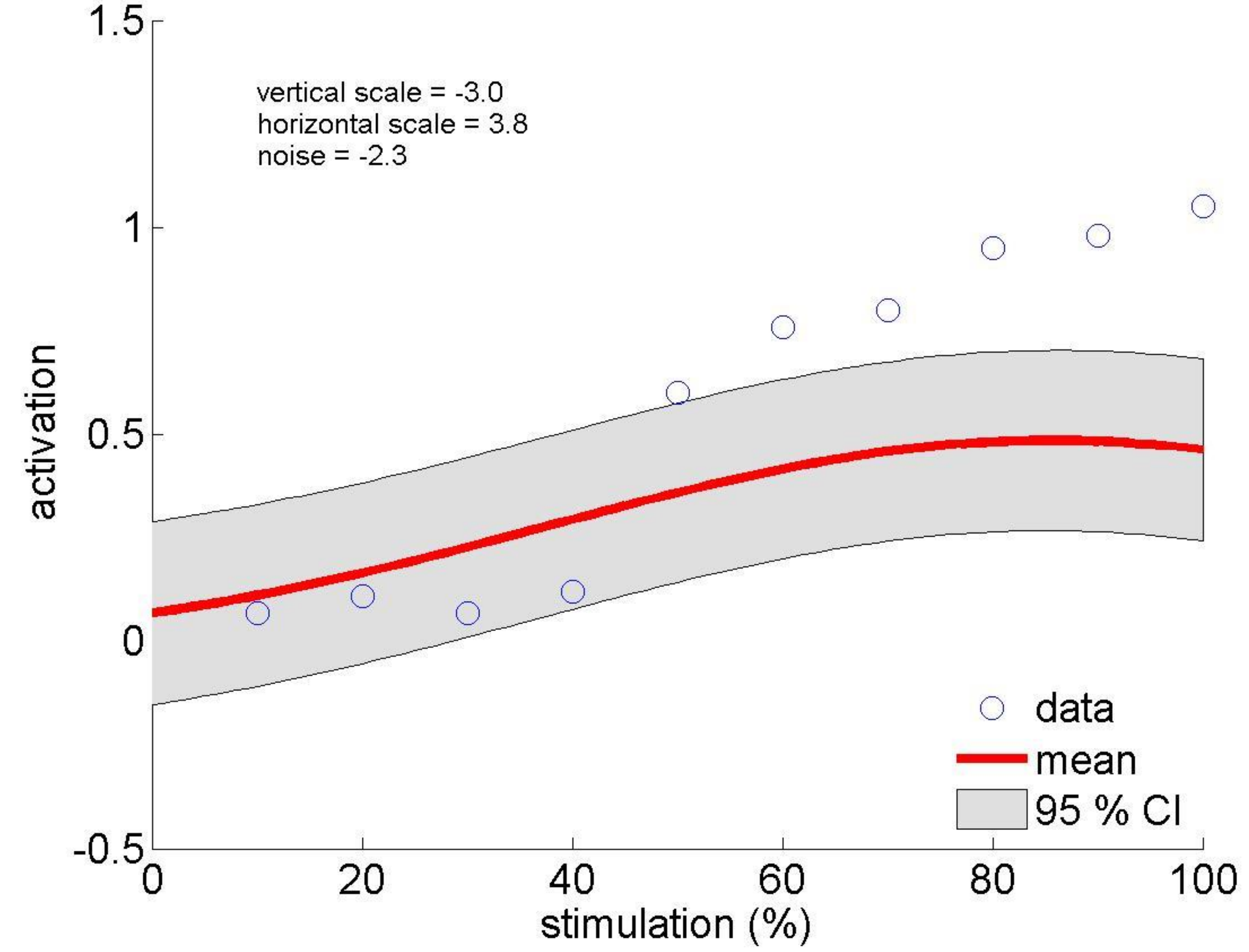
length scale too short

$$k(\mathbf{x}, \mathbf{x}') = p_1 e^{-\frac{(\mathbf{x}-\mathbf{x}')^\top (\mathbf{x}-\mathbf{x}')}{2p_2^2}}$$

Changing the Vertical Scale



nice vertical scale



vertical scale too small

$$k(\mathbf{x}, \mathbf{x}') = p_1 e^{-\frac{(\mathbf{x}-\mathbf{x}')^\top(\mathbf{x}-\mathbf{x}')}{2p_2^2}}$$

Overview

1. What is regression?
2. Why are some regression problems hard?
3. What is a Gaussian process?
4. How does Gaussian process regression work?
5. How does Gaussian process regression compare to other nonlinear regression methods?

General Regression Problem

Given:

$$\mathcal{D} = \{(\mathbf{x}_i, y_i) | i = 1, \ldots, n\} = (X, \mathbf{y})$$ training set

$$\mathbf{x}_*$$ new input

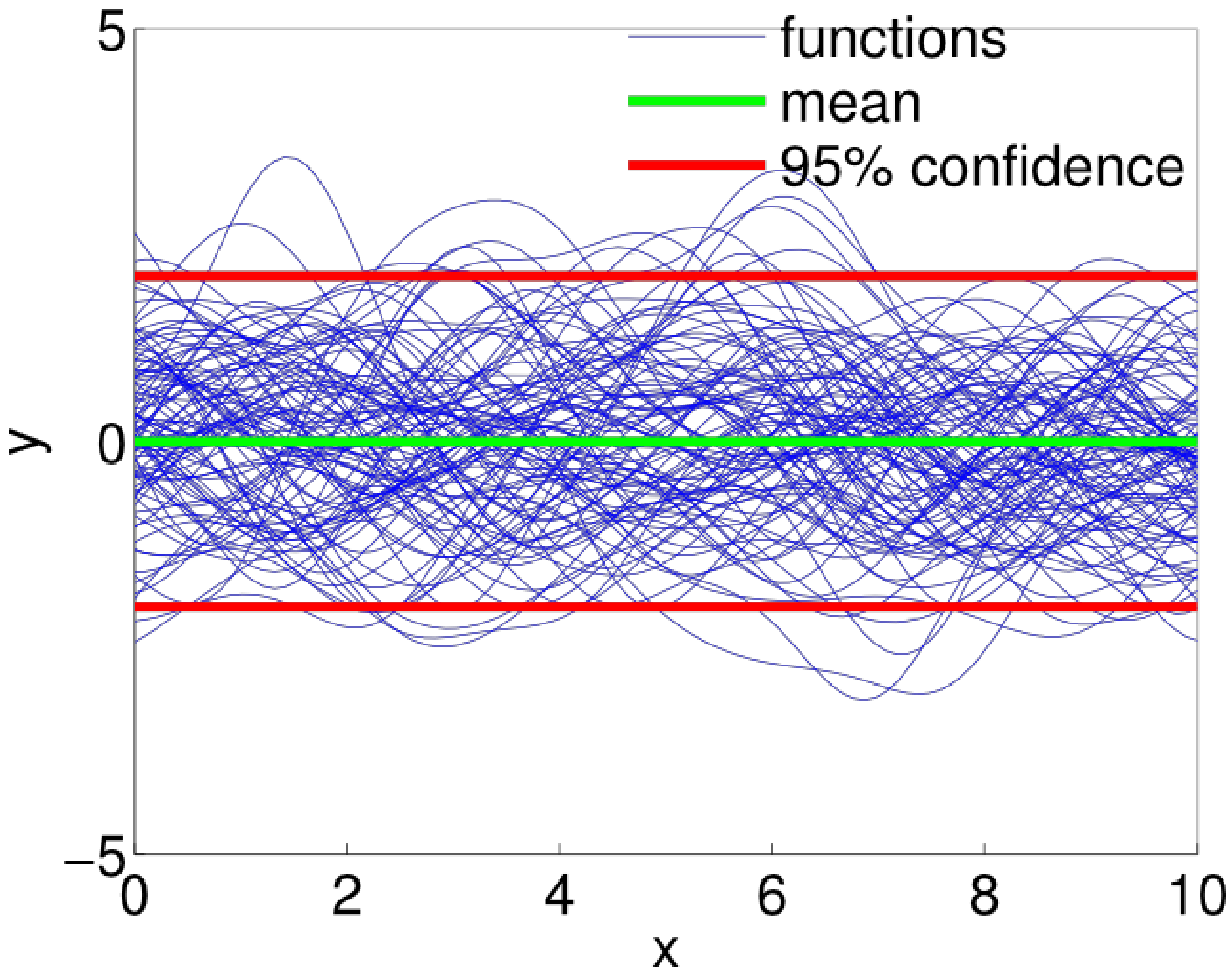
(query point)

Predict:

$$y_*$$ new output

$$p(y_*)$$ distribution of the new output

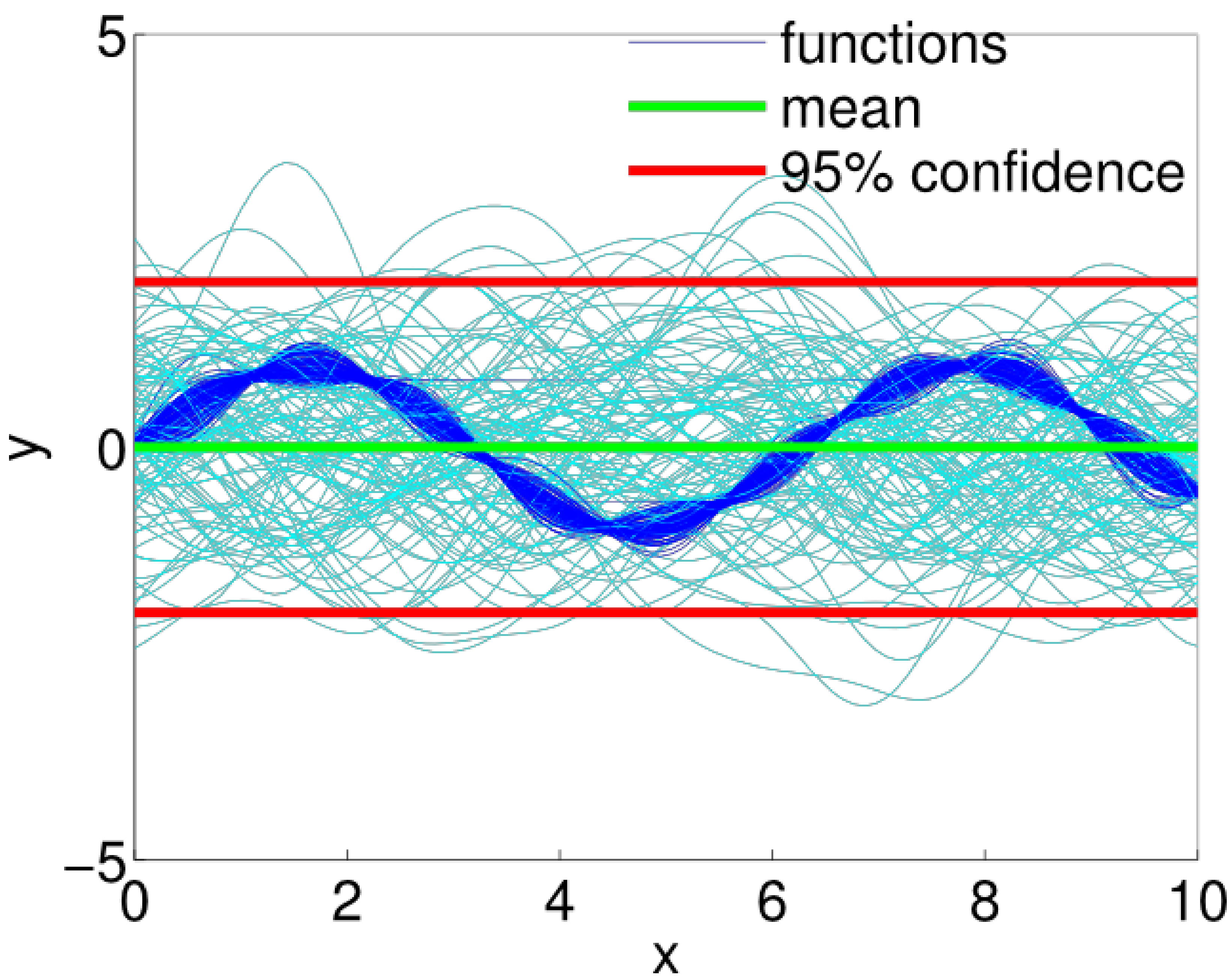
Gaussian Process Regression



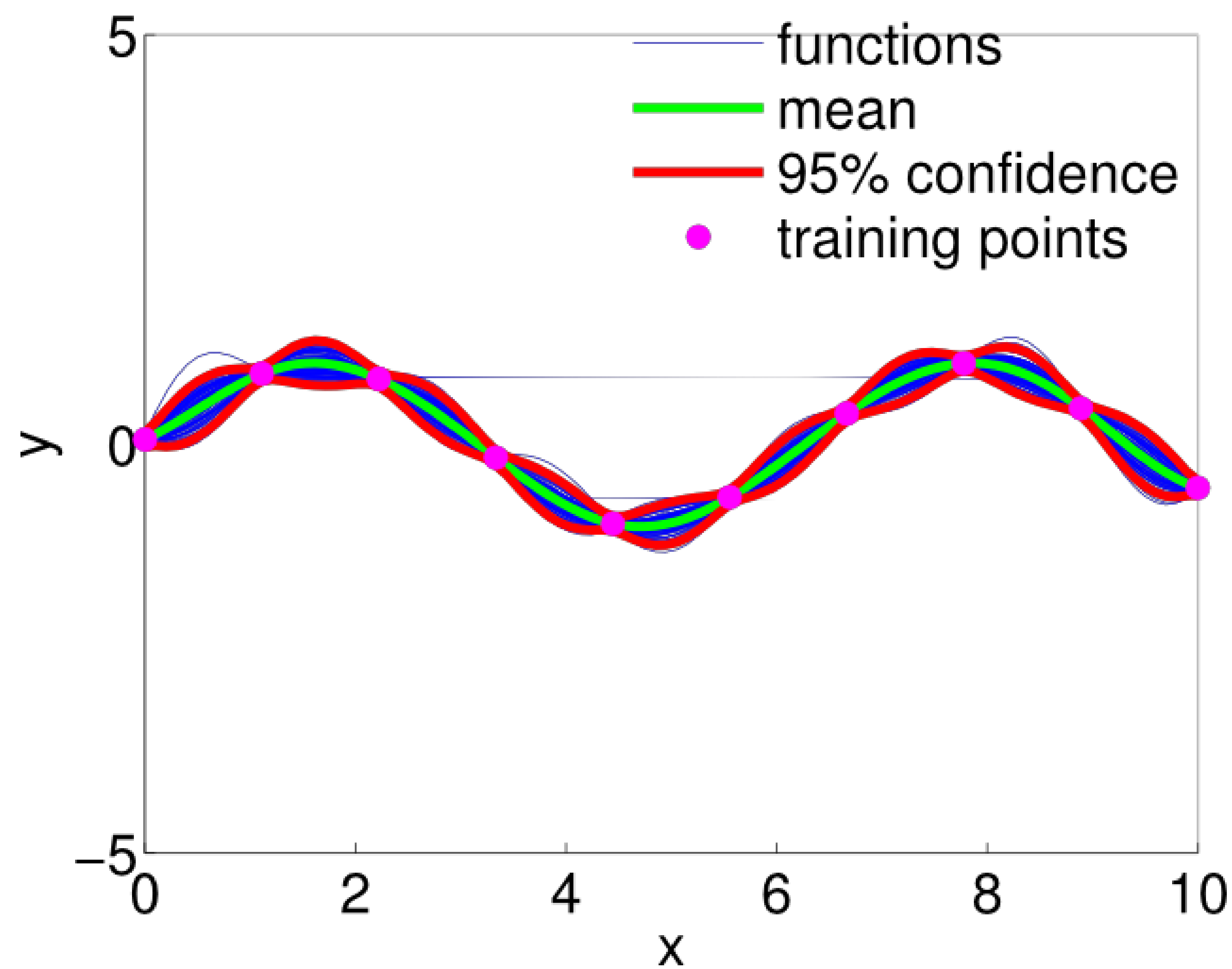
prior Gaussian process

Gaussian process: a collection of random variables, any finite number of which have a joint Gaussian distribution

Gaussian Process Regression



prior Gaussian process



posterior Gaussian process

Gaussian process: a collection of random variables, any finite number of which have a joint Gaussian distribution

Joint Prior Distribution

outputs
You have observed

$$\begin{bmatrix} y \\ f_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(X,X)+\sigma^2 I & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix}\right)$$

output you
have not observed (for a new query)

$K(X,X) \in \mathbb{R}^{n\times n}$ covariance of the training outputs with the training outputs

$K(X,X_*) \in \mathbb{R}^{n\times n_*}$ covariance of the training outputs with the test outputs

$K(X_*,X) \in \mathbb{R}^{n_*\times n}$ covariance of the test outputs with the training outputs

$K(X_*,X_*) \in \mathbb{R}^{n_*\times n_*}$ covariance of the test outputs with the test outputs

Gaussian process: a collection of random variables, any finite number of which have a joint Gaussian distribution

Predictive Distribution

$$\begin{bmatrix} y \\ f_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(X,X)+\sigma^2 I & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix}\right)$$

condition the joint prior distribution on the training data
(this operation is a property of Gaussians)

$$\mathbf{f}_*|X,y,X_* \sim \mathcal{N}(\bar{\mathbf{f}}_*, \mathrm{cov}(\mathbf{f}_*))$$

predictive distribution

$$\bar{\mathbf{f}}_* = K(X_*,X)\left[K(X,X)+\sigma^2 I\right]^{-1}\mathbf{y}$$

$$\mathrm{cov}(\bar{\mathbf{f}}_*) =$$

$$K(X_*,X_*) - K(X_*,X)[K(X,X)+\sigma^2 I]^{-1}K(X,X_*)$$

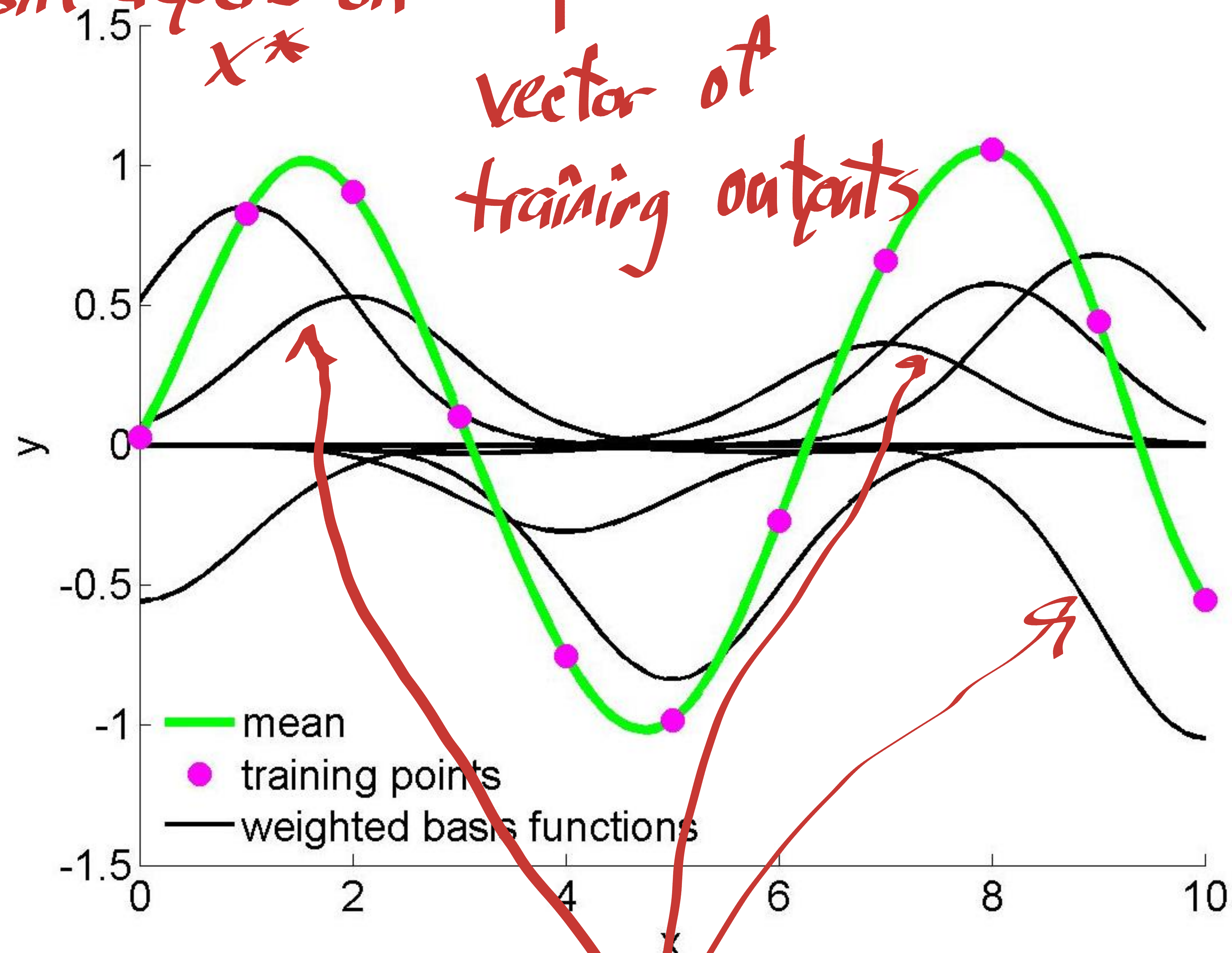
Mean of Predictive Distribution

$$\bar{f}_*(\mathbf{x}_*) = \overbrace{\mathbf{k}(\mathbf{x}_*, X)}^{\text{how close is } x^* \text{ to training input?}} \underbrace{\left[K(X,X) + \sigma^2 I\right]^{-1}}_{\text{doesn't depend on } x^*} \mathbf{y}$$

vector of training outputs

$$\bar{f}_*(\mathbf{x}_*) = \sum_{i=1}^{n} \alpha_i k(\mathbf{x}_*, \mathbf{x_i})$$

linear combination of basis functions each centered at a training point



basis functions

weights

$$\alpha_i = (K(X,X) + \sigma^2 I)_i^{-1} y_i$$

$$k(\mathbf{x}_*, \mathbf{x_i}) = p_1 e^{-\frac{(\mathbf{x}_* - \mathbf{x_i})^\top (\mathbf{x}_* - \mathbf{x_i})}{2p_2^2}}$$

Covariance of Predictive Distribution

$$\mathrm{cov}(\bar{\mathbf{f}}_*) =$$

$$K(X_*, X_*) - K(X_*, X)[K(X, X) + \sigma^2 I]^{-1} K(X, X_*)$$

prior test point covariance

additional information gained from the training inputs

How to Pick the Hyperparameters

$$k(\mathbf{x}, \mathbf{x}') = p_1 e^{-\frac{(\mathbf{x}-\mathbf{x}')^\top (\mathbf{x}-\mathbf{x}')}{2p_2^2}} \qquad \theta = [p_1 \; p_2]$$

marginal likelihood or evidence of the data

$$p(\mathbf{y}|X,\theta) = \int \underbrace{p(\mathbf{y}|\mathbf{f}, X, \theta)}_{\text{probability of observing the training outputs given, the inputs and the model}} \underbrace{p(\mathbf{f}|X,\theta)}_{\text{probability of the model}} d\mathbf{f}$$

$$\log p(\mathbf{y}|X,\theta) =$$

$$-\frac{1}{2}\mathbf{y}^\top (K + \sigma^2 I)^{-1}\mathbf{y} - \frac{1}{2}\log\left|K + \sigma^2 I\right| - \frac{n}{2}\log 2\pi$$

data fit term

model complexity penalty

normalization term

$$\underset{\theta}{\text{maximize}} \quad \log p(\mathbf{y}|X,\theta)$$

Incorporate Fixed Basis Functions

$$g(\mathbf{x}) = f(\mathbf{x}) + \mathbf{x}^\top \mathbf{w}$$

$K_x = K(x_*, x_*)$

$K = K(X, X)$

$$\mathbf{w} \sim \mathcal{N}(\mathbf{b}, B)$$

prior parameter distribution

$$f(\mathbf{x}) \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$$

zero-mean GP

$$g(\mathbf{x}) \sim \mathcal{GP}(\mathbf{x}^\top \mathbf{b}, k(\mathbf{x}, \mathbf{x}') + \mathbf{x}^\top B \mathbf{x})$$

combined GP

$$\bar{\mathbf{g}}(X_*) = X_*^\top \bar{\mathbf{w}} + \overbrace{K_*^\top K^{-1}(\mathbf{y} - X^\top \bar{\mathbf{w}})}$$

mean of predictive distribution

parameterized model term

GP model term

Remember This

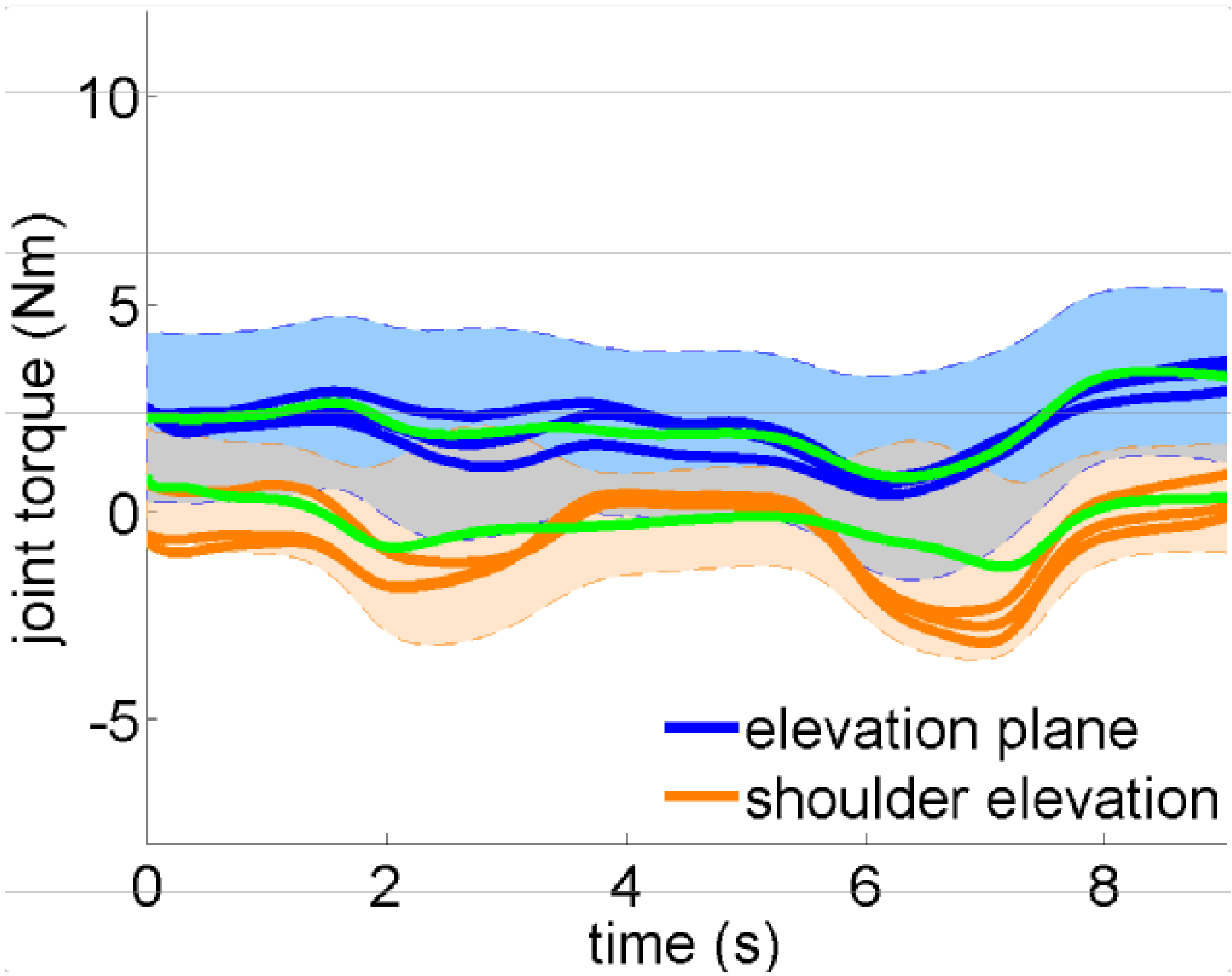
1. You get to write down an analytical expression for the distribution of your predicted output
2. You didn’t have to solve a nonlinear optimization problem.

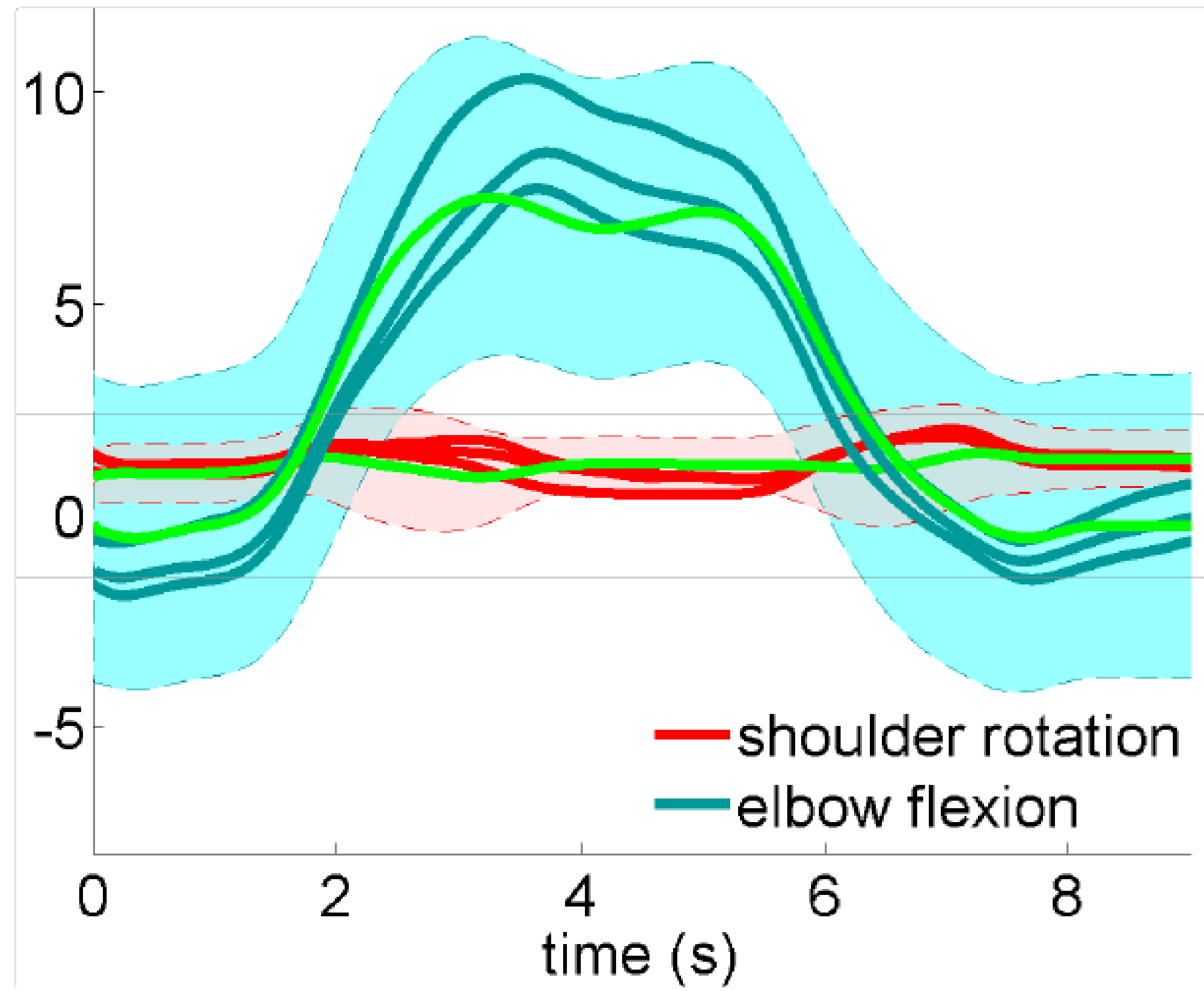
Inverse Dynamics of the Human Arm

Given: shoulder and elbow joint angles, velocities, and accelerations

Predict: shoulder and elbow joint torques to drive arm along a trajectory

Mean Predictions with Confidence Intervals of Joint Torques





Overview

$w \sim N(m, \sigma^2)$

linear regression $f = w^T x$

GPR $f \sim GP$

1. What is regression?
2. Why are some regression problems hard?
3. What is a Gaussian process?
4. How does Gaussian process regression work?
5. How does Gaussian process regression compare to other nonlinear regression methods?

Gaussian Process Regression and Other Nonlinear Regression Methods

Method	Number of Parameters/ function complexity	Nonlinear optimization required with new data?	Model Selection	Analytical Predictive Distribution	Speed of computation
Gaussian process regression	low	not really	minimize ML	yes	slow for big data sets
Artificial neural networks	high	yes	ad hoc	no	fast
Radial basis functions	high	no	K-means clustering	yes	fast
Locally weighted regression	low	no	ad hoc	yes	fast