

Introduction to Gaussian Processes

Pattern Recognition & Machine Learning Course, EPFL

December 2015

Carlos Becker, Emtiyaz Khan

Motivation

2

Goals of this lecture

- Understand what a Gaussian Process (GP) is.
- Learn how GPs can be used for regression.

More specific to GPs, you will learn:

- What a covariance matrix means from a GP point of view.
- How a GP defines a prior over functions, and its relationship to its covariance matrix and correlation terms.
- What “conditioning on the measurements” means, in a probabilistic sense as well as mathematically.

Note: GPs for classification are outside the scope of this lecture. But if you understand regression GPs it won't be too difficult to learn how classification GPs work. Please see Rasmussen and Williams's “Gaussian Processes for Machine Learning” book.

Motivation: why Gaussian Processes?

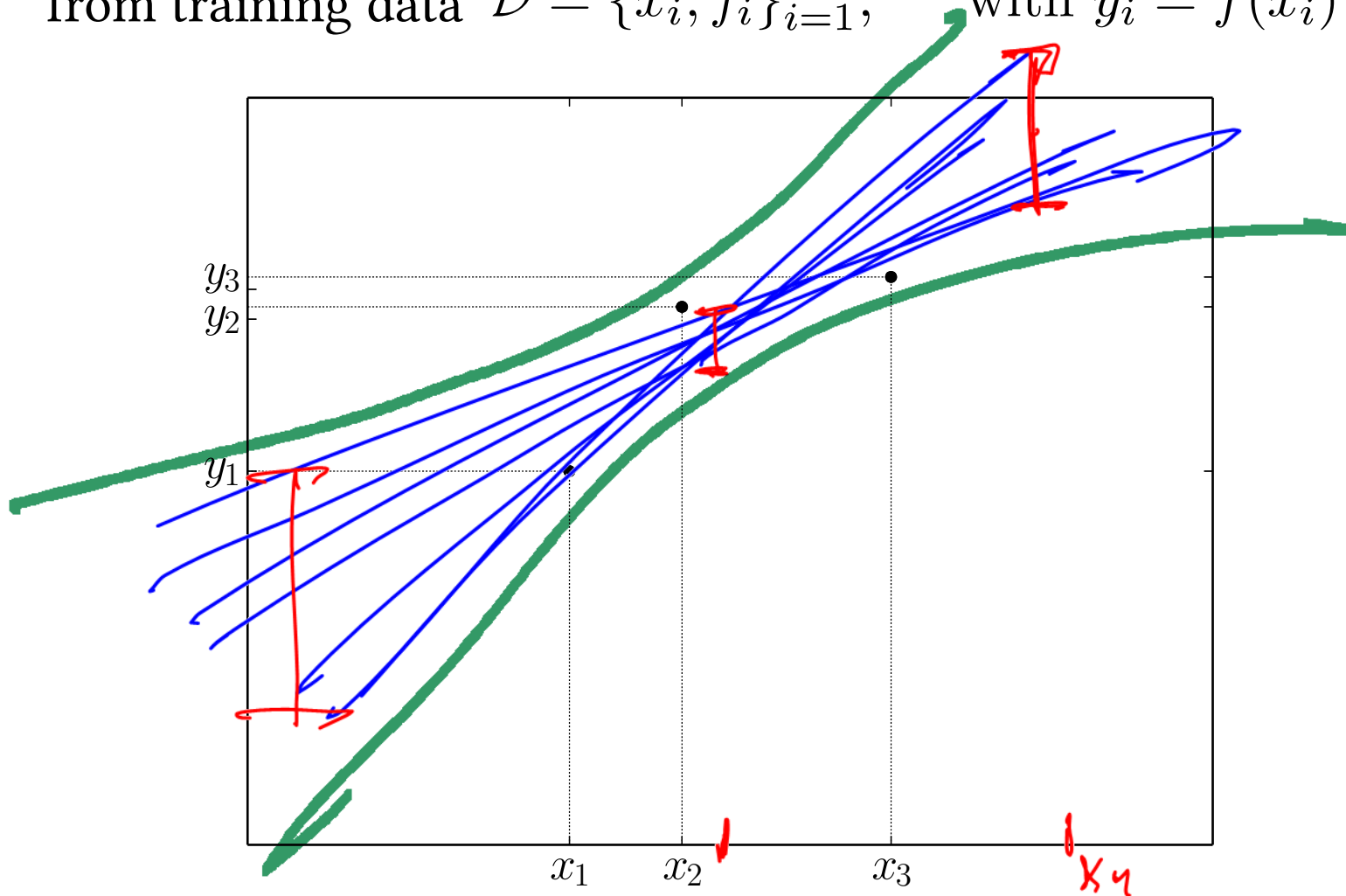


Motivation

4

Say we want to estimate a scalar function $f(x)$

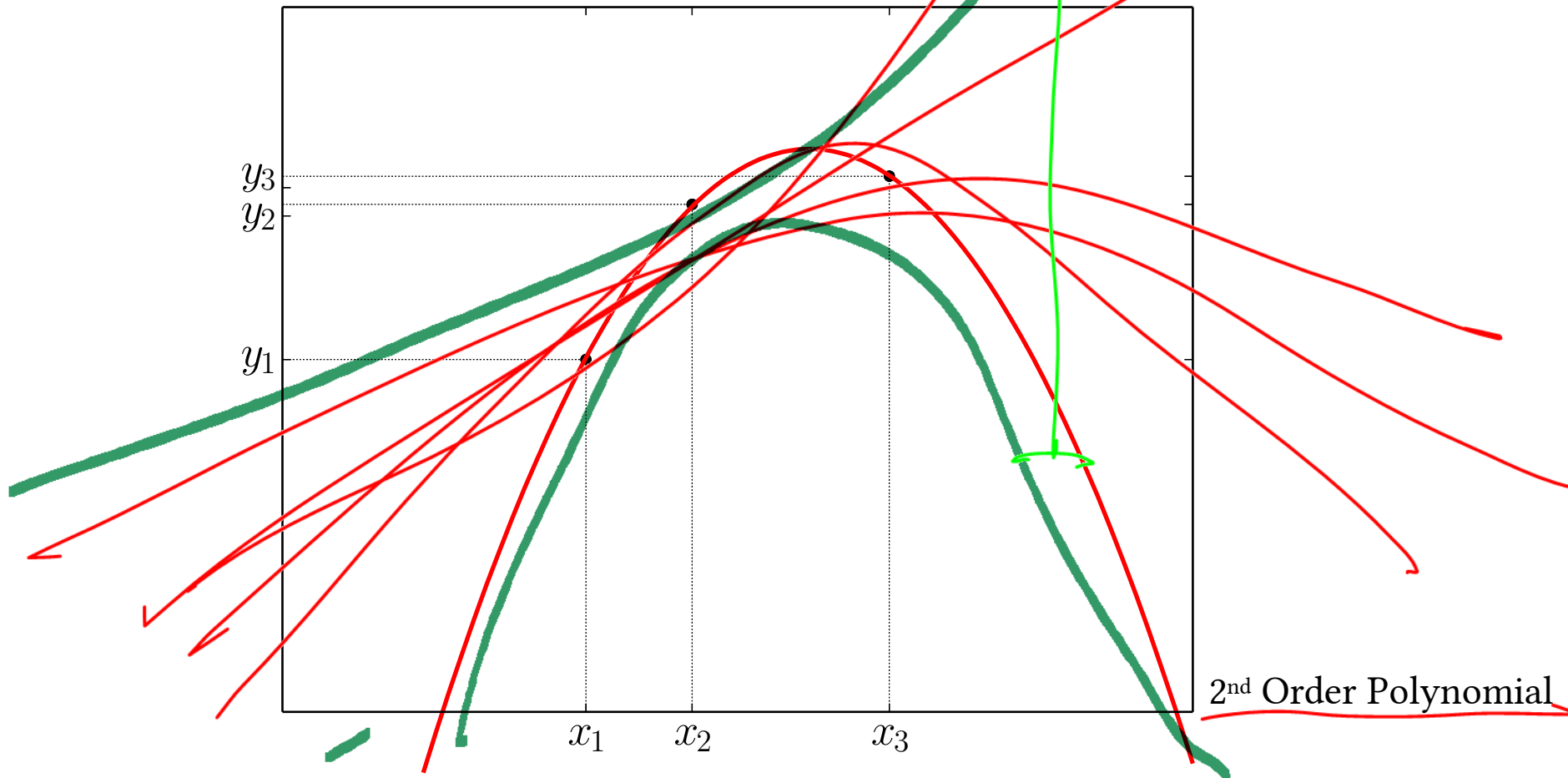
from training data $\mathcal{D} = \{x_i, f_i\}_{i=1}^N$, with $y_i = f(x_i) + \epsilon$



Motivation

5

Say we want to estimate a scalar function $f(x)$
from training data $\mathcal{D} = \{x_i, f_i\}_{i=1}^N$, with $y_i = f(x_i) + \epsilon$

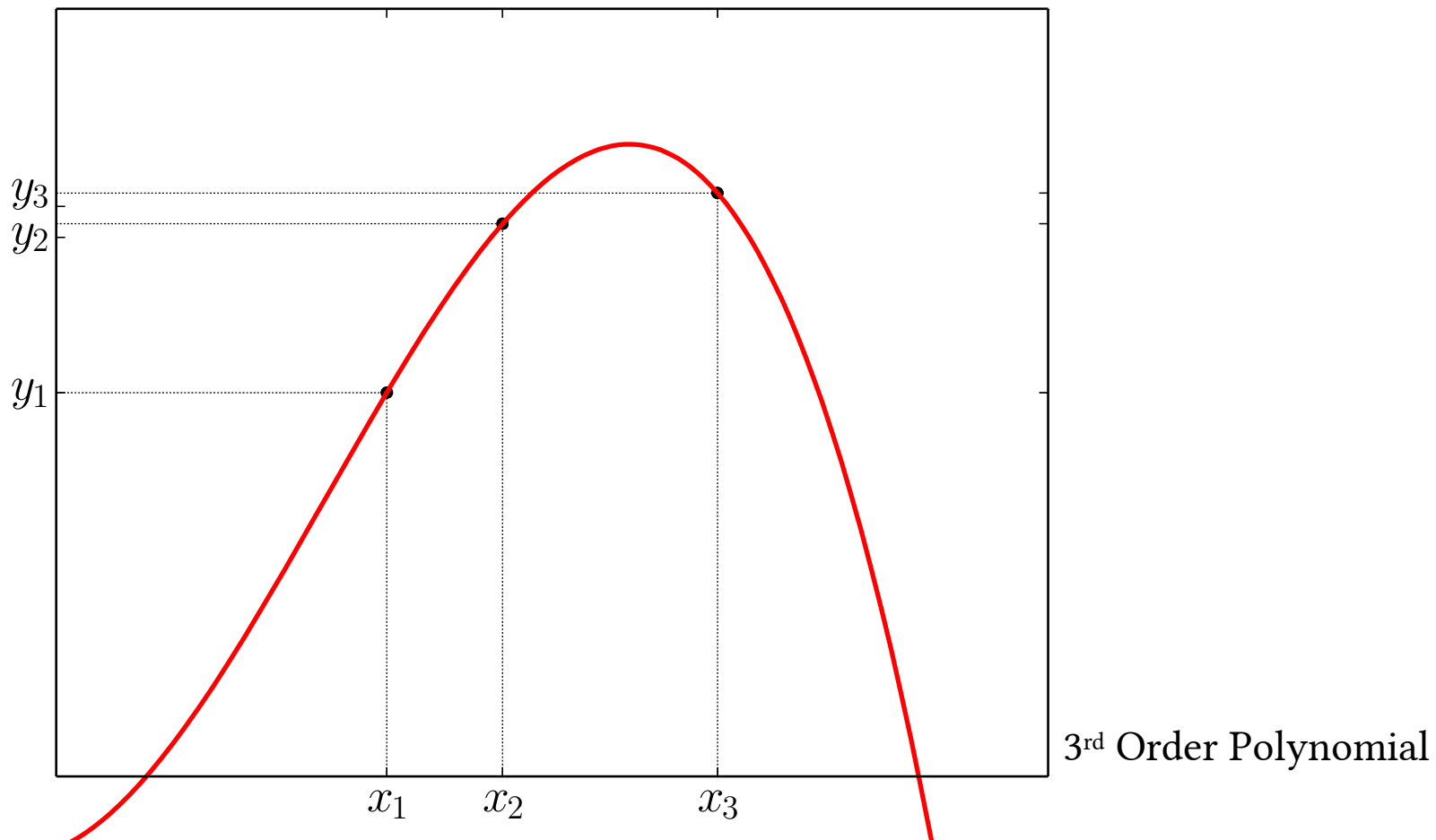


Motivation

6

Say we want to estimate a scalar function $f(x)$

from training data $\mathcal{D} = \{x_i, f_i\}_{i=1}^N$, with $y_i = f(x_i) + \epsilon$

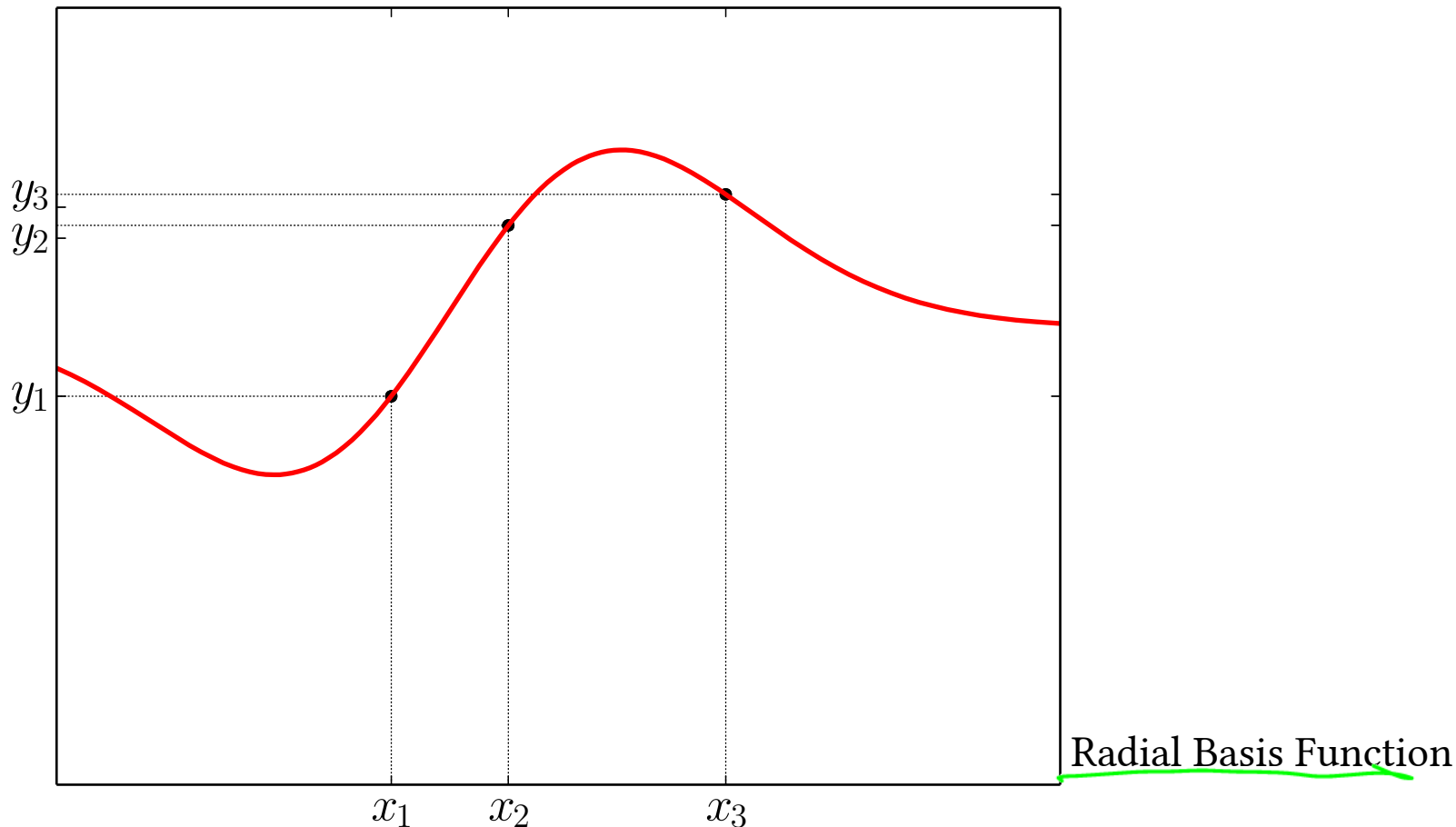


Motivation

7

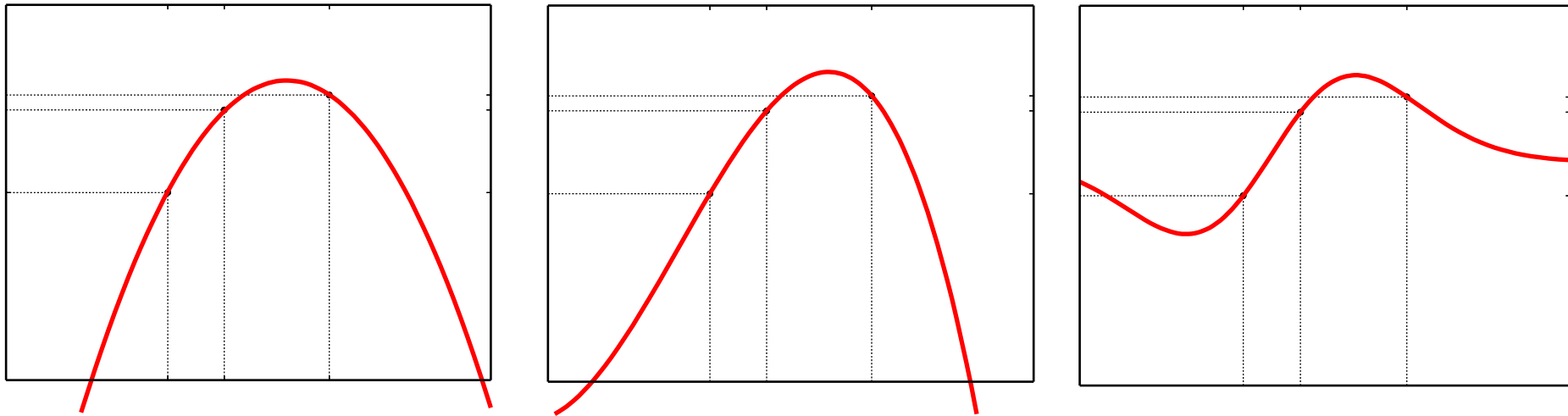
Say we want to estimate a scalar function $f(x)$

from training data $\mathcal{D} = \{x_i, f_i\}_{i=1}^N$, with $y_i = f(x_i) + \epsilon$



Motivation

8



Gaussian Processes let us place a **prior** on the 'shape' of $f(x)$

And this **prior** is formulated probabilistically

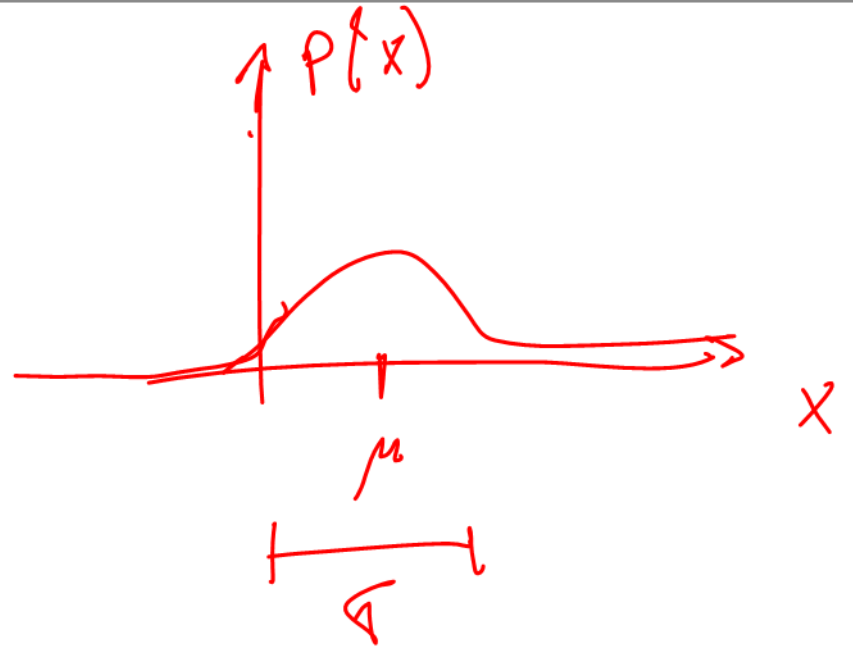
Let's get started!



Just before...

10

$$X \sim N(\mu, \sigma^2)$$

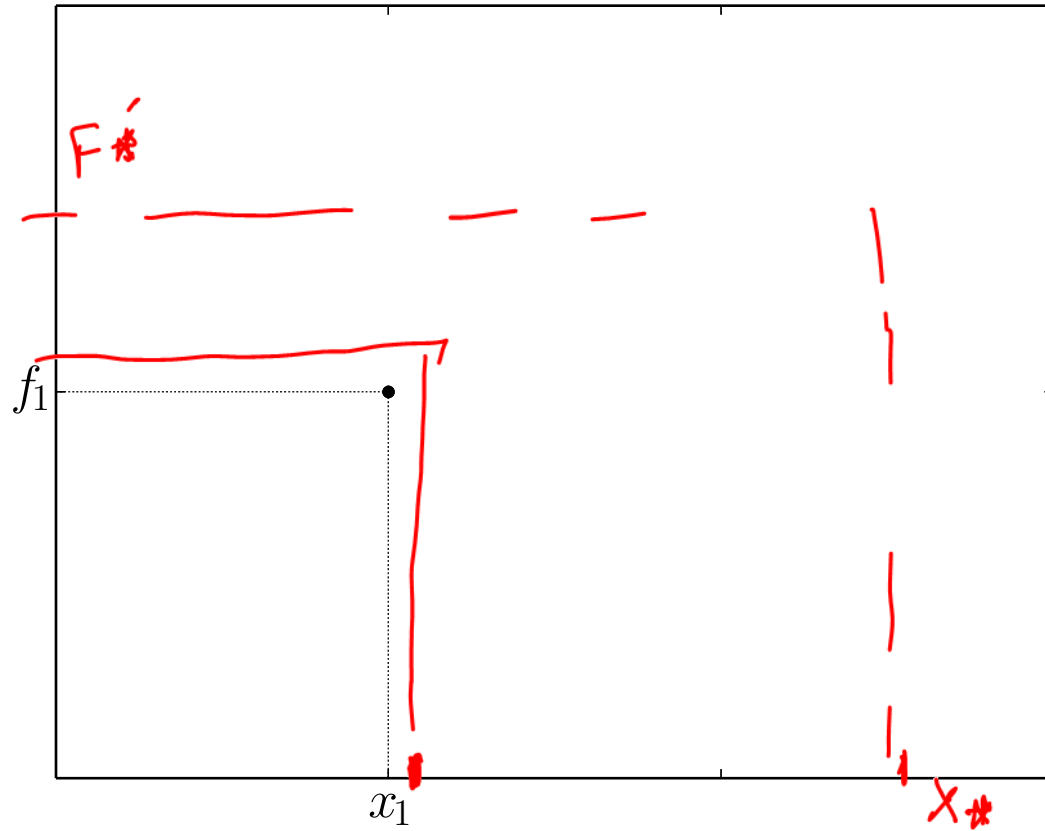


Gaussian Processes

11

Let's get back to finding $f(x)$, but now

- From a single observation $\{x_1, f_1\}$
- We want to predict f_* = $f(x_*)$



Intuition?

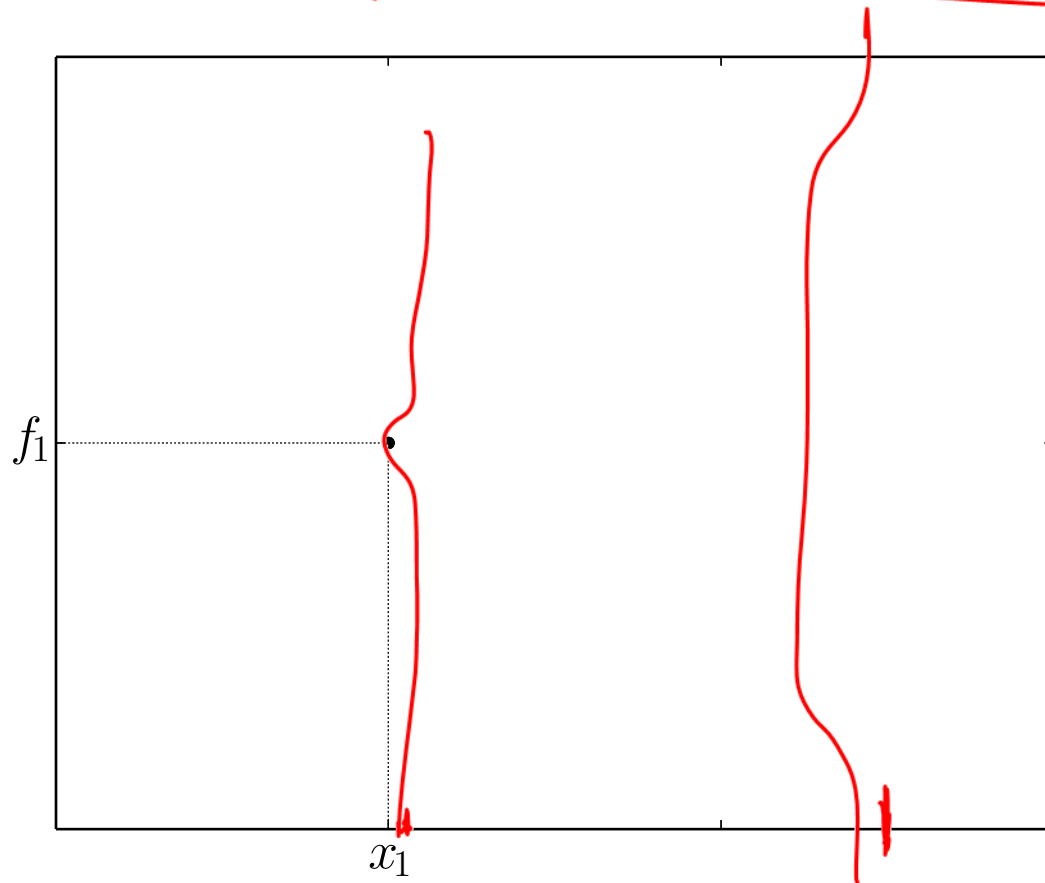
$$X_* \sim X_1 \Rightarrow f_* \sim f_1$$
$$|X_* - X_1| \gg 0 \Rightarrow f_* ?$$

Gaussian Processes

12

Now, in a probabilistic manner...

The estimated f_* is now a **Random Variable**, with a corresponding **PDF**.



Intuition?

$$X_* \sim X_r$$

uncertainty \rightarrow PDF

and depends on

$$X_1, F_1, X_*$$

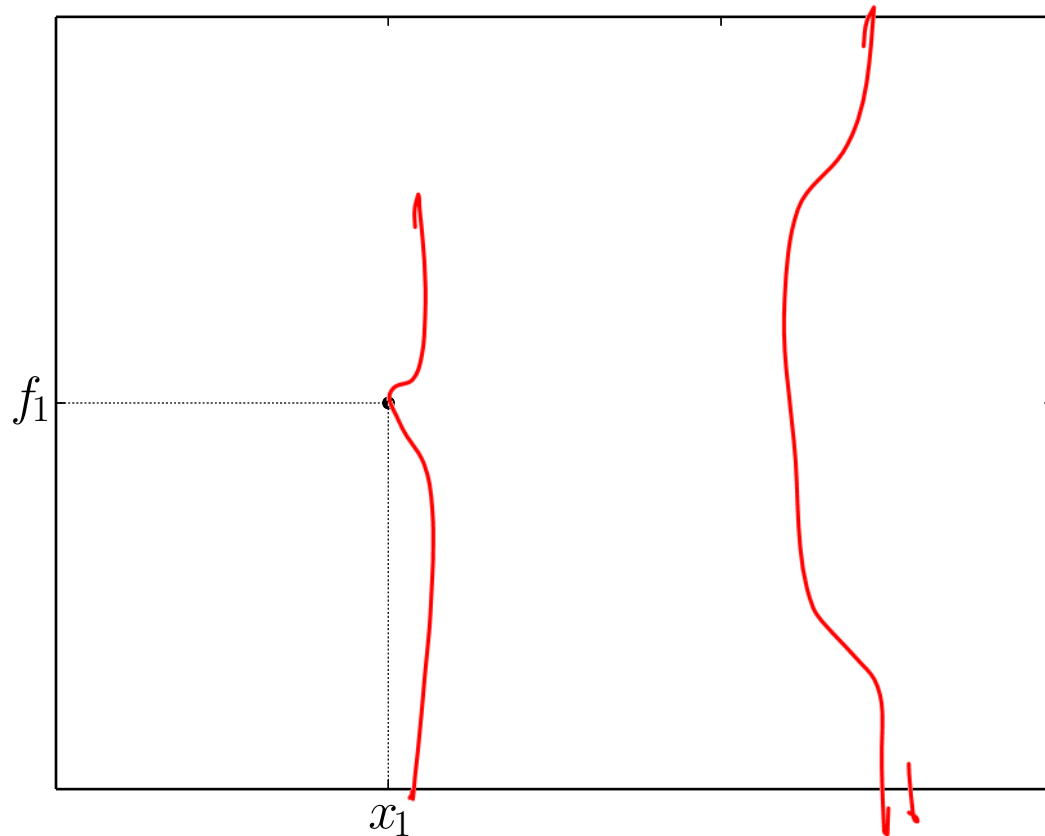
Gaussian Processes

Getting there...

GRV

The estimated f_* is now a Gaussian RV:

$$f_* \sim \mathcal{N}(\mu_*, \sigma_*^2)$$



Intuition?

$$X_* \sim X_1 \Rightarrow \mu_* \sim F_1$$

$\sigma_*^2 \downarrow$

$$|X_* - X_1| \gg 0 \Rightarrow \mu_* ?$$

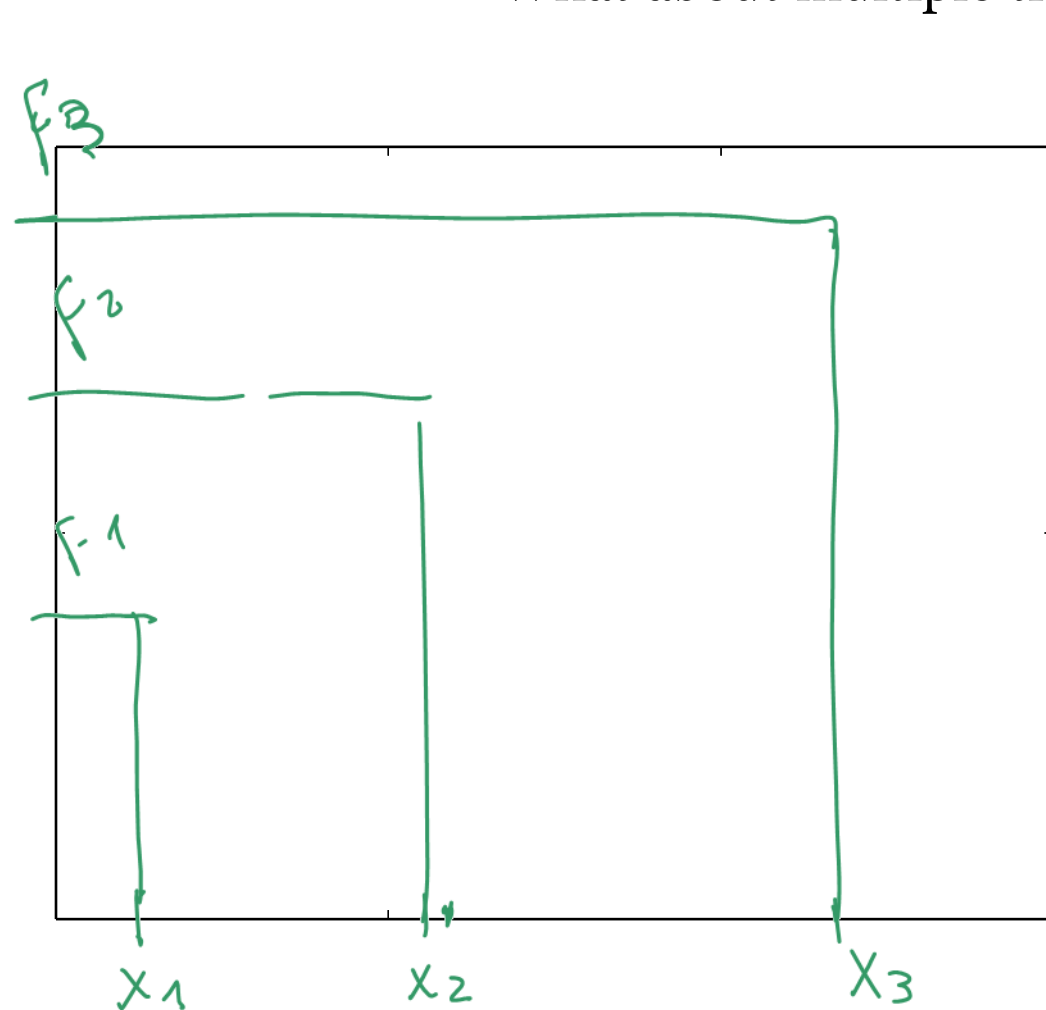
$\sigma_*^2 \uparrow$

$$\mu_* (x_1, f_1, x_*)$$
$$\sigma_*^2 \begin{pmatrix} n & n & n \end{pmatrix}$$

Gaussian Processes

14

What about multiple training points?



$$F_* \sim N(\mu_*, \Sigma_*^2)$$

$$\mu_* (x_1, f_1, x_2, f_2, x_3, f_3, x_*)$$
$$\Sigma_*^2 \left(\begin{array}{c} \uparrow \\ \uparrow \end{array} \right)$$

CONCL: μ_*, Σ_*^2 depend
both on tr. data
and pred point x_*

Gaussian Processes

So far we expressed f_* as a function of the training data

But Gaussian Processes work in a slightly different way...

Gaussian Processes

We take f_1 and f_* to be RVs, with a joint Gaussian pdf

$$p(\underline{f_1}, \underline{f_*} \mid x_1, x_*)$$

Gaussian Processes

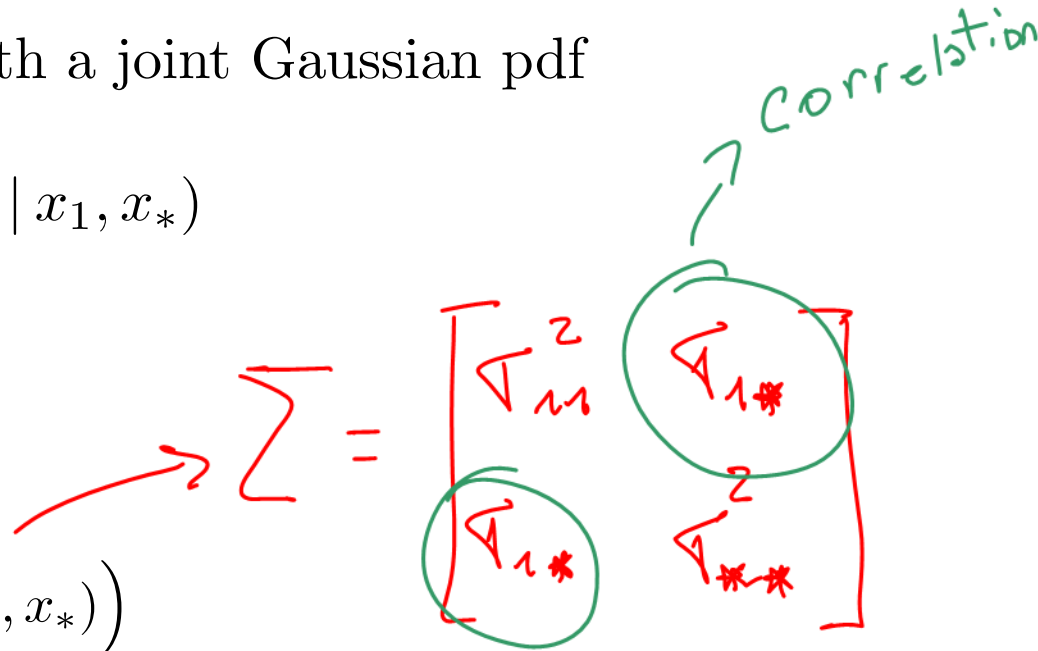
17

We take f_1 and f_* to be RVs, with a joint Gaussian pdf

$$p(f_1, f_* | x_1, x_*)$$

More precisely,

$$\begin{bmatrix} f_1 \\ f_* \end{bmatrix} \sim \mathcal{N}(\underline{\mathbf{0}}, K(x_1, x_*))$$



$$x_1 \sim x_* \Rightarrow \sigma_{1*} \uparrow$$

$$|x_1 - x_*| \gg 0 \Rightarrow \sigma_{1*} \rightarrow 0$$

Gaussian Processes

18

Incorporating the measurement:

Gaussian Processes

19

Incorporating the measurement:

Our model is $p(f_1, f_* | x_1, x_*)$

prior
(before any
measurements)

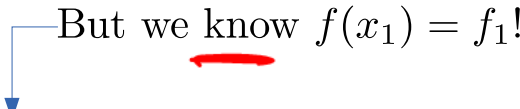
Gaussian Processes

20

Incorporating the measurement:

Our model is $p(f_1, f_* | x_1, x_*)$

But we know $f(x_1) = f_1!$



Gaussian Processes

21

Incorporating the measurement:

Our model is $p(f_1, f_* | x_1, x_*)$

But we know $f(x_1) = f_1!$

And we want to estimate

Gaussian Processes

22

Incorporating the measurement:

Our model is $p(f_1, f_* | x_1, x_*)$

But we know $f(x_1) = f_1!$

And we want to estimate

What we want is $p(f_* | f_1, x_1, x_*)$

Gaussian Processes

23

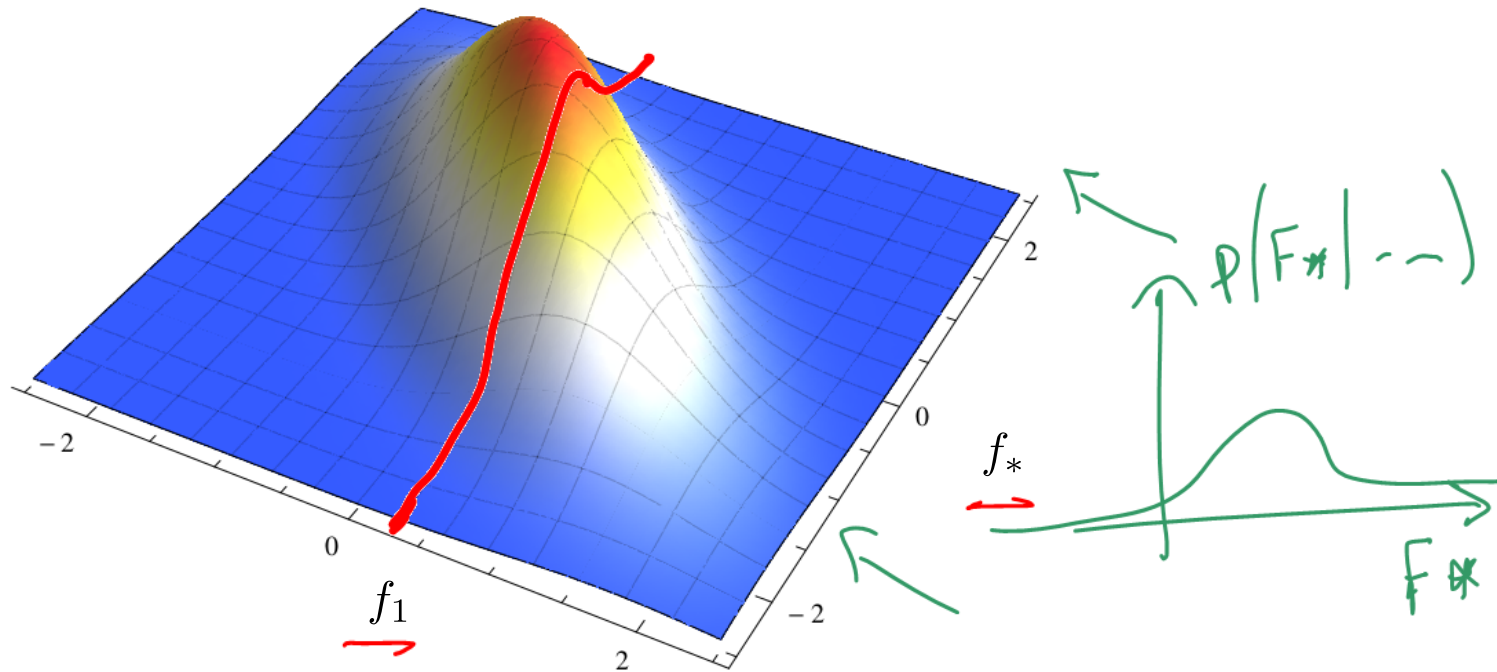
Incorporating the measurement:

Our model is $p(f_1, f_* | x_1, x_*)$

But we know $f(x_1) = f_1!$

And we want to estimate

What we want is $p(f_* | f_1, x_1, x_*)$



Gaussian Processes

Summarizing our 2-point Gaussian Process:

- Our model or prior is

$$p(f_1, f_* | x_1, x_*) = \mathcal{N}\left(\mathbf{0}, K(x_1, x_*)\right)$$

$$K = [K_{mn}] = [k(x_m, x_n)]$$

- If we have a measurement $f(x_1) = f_1$, we can condition on it to estimate f_* :

$$p(f_* | f_1, x_1, x_*) = \mathcal{N}\left(\mu_*, \sigma_*^2\right)$$

We get a probability distribution as the output.

Exercise

25

Let's use the RBF kernel $k(x_n, x_m) = e^{-(x_n - x_m)^2 / 2}$

• For our model prior

$$p(f_1, f_* | x_1, x_*) = \mathcal{N}(\mathbf{0}, K)$$

$$K = \begin{bmatrix} k(x_1, x_1) & k(x_1, x_*) \\ k(x_1, x_*) & k(x_*, x_*) \end{bmatrix}$$

$$K = [K_{mn}] = [k(x_m, x_n)]$$

$$K = \begin{bmatrix} 1 & 0.6 \\ 0.6 & 1 \end{bmatrix}$$

- 1) How many rows and columns K has? 2×2
- 2) Compute K for $x_1 = 1$, $x_* = 2$.
- 3) Are f_1 and f_* strongly correlated?

$$K = \begin{bmatrix} 1 & \sim 0 \\ \sim 0 & 1 \end{bmatrix}$$

(Hint: $e^{-1/2} \approx 0.6$)

- 4) What if $x_* = 10$?

Time to get to the Real GP



Gaussian Processes

27

A gaussian process defines a **prior over functions \mathbf{f}** ,

$$p(\mathbf{f}|X) = \mathcal{N}(\mathbf{f}|\mathbf{0}, K(X)) \leftarrow$$

$$\mathbf{F} = \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ \vdots \end{bmatrix}$$

Defining the kernel function $k(\mathbf{x}_n, \mathbf{x}_m)$ defines the prior

1. We can sample *functions* from this prior
2. We can use the prior + measurements to generate predictions

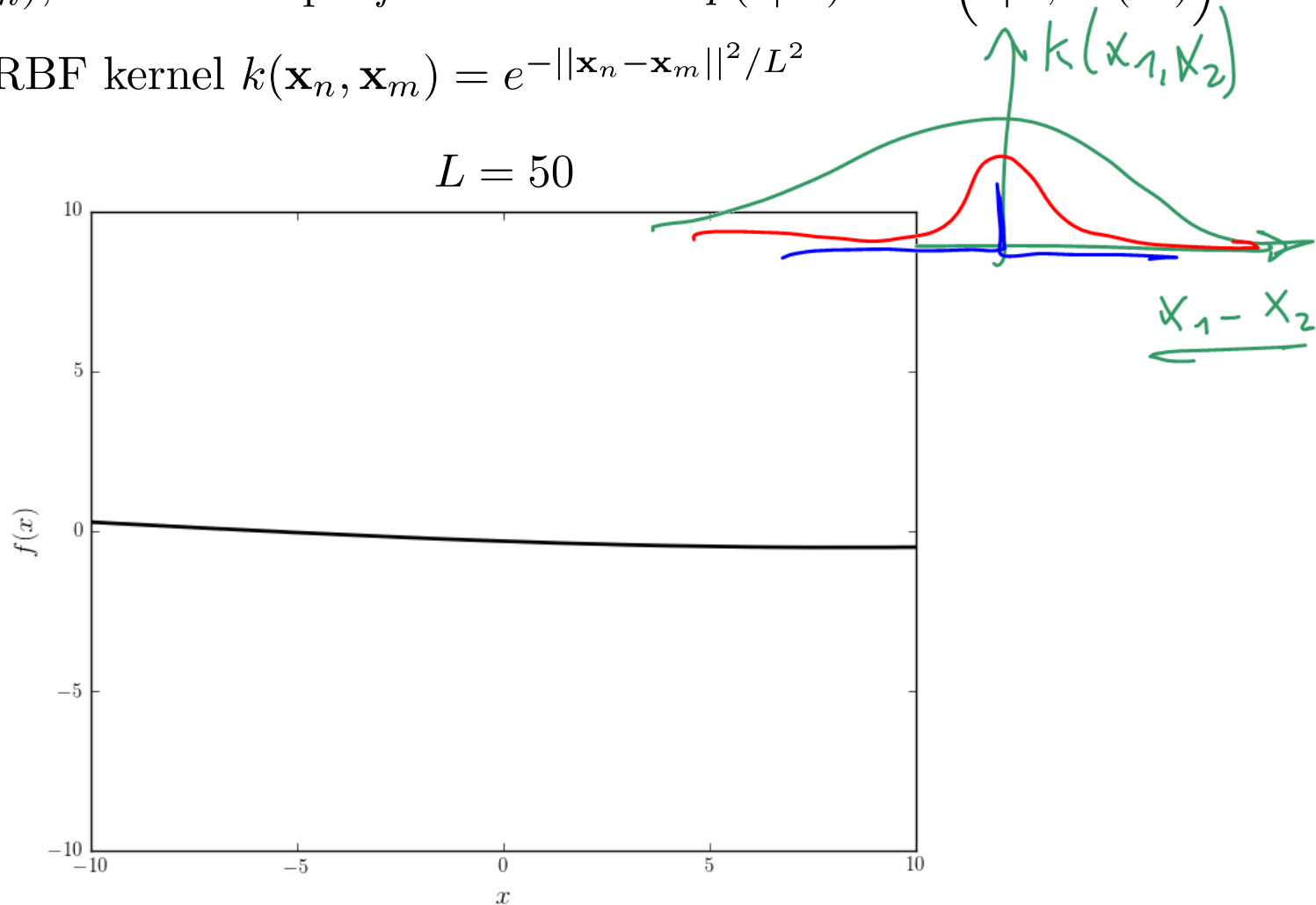
Gaussian Processes

28

1) A prior over functions

Given $k(x_n, x_m)$, we can sample *functions* from $p(\mathbf{f}|X) = \mathcal{N}(\mathbf{f}|\mathbf{0}, K(X))$

Example 1: RBF kernel $k(\mathbf{x}_n, \mathbf{x}_m) = e^{-\|\mathbf{x}_n - \mathbf{x}_m\|^2/L^2}$



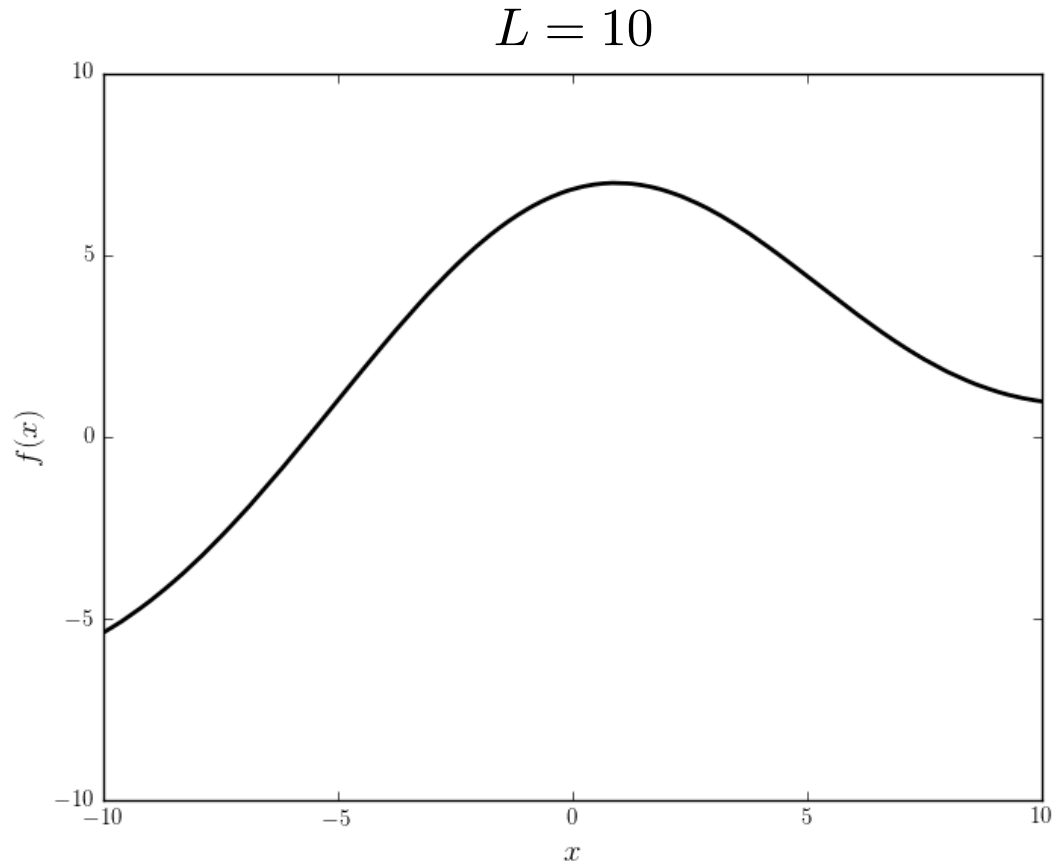
Gaussian Processes

29

1) A prior over functions

Given $k(x_n, x_m)$, we can sample *functions* from $p(\mathbf{f}|X) = \mathcal{N}(\mathbf{f}|\mathbf{0}, K(X))$

Example 1: RBF kernel $k(\mathbf{x}_n, \mathbf{x}_m) = e^{-\|\mathbf{x}_n - \mathbf{x}_m\|^2/L^2}$



Gaussian Processes

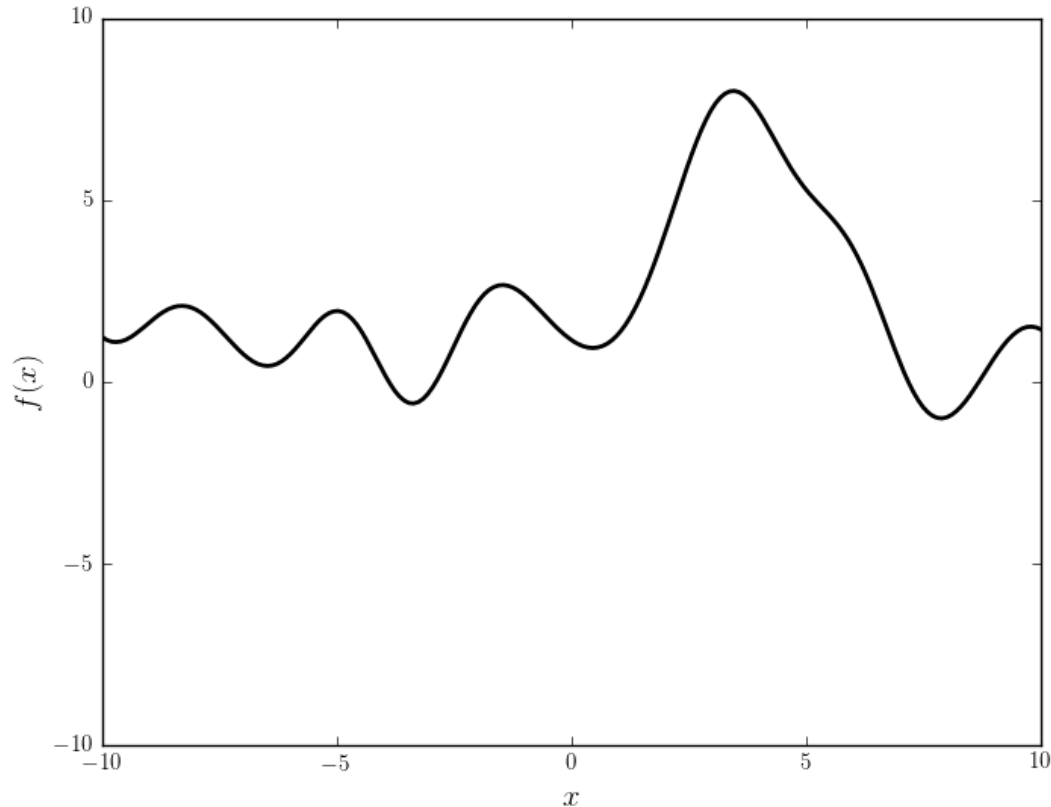
30

1) A prior over functions

Given $k(x_n, x_m)$, we can sample *functions* from $p(\mathbf{f}|X) = \mathcal{N}(\mathbf{f}|\mathbf{0}, K(X))$

Example 1: RBF kernel $k(\mathbf{x}_n, \mathbf{x}_m) = e^{-\|\mathbf{x}_n - \mathbf{x}_m\|^2/L^2}$

$$L = 2$$



Gaussian Processes

31

1) A prior over functions

$$k(\mathbf{x}_n, \mathbf{x}_m) = e^{-\|\mathbf{x}_n - \mathbf{x}_m\|^2 / L^2}$$

Example 1: RBF kernel, **What will happen if $L \rightarrow 0$?**

Gaussian Processes

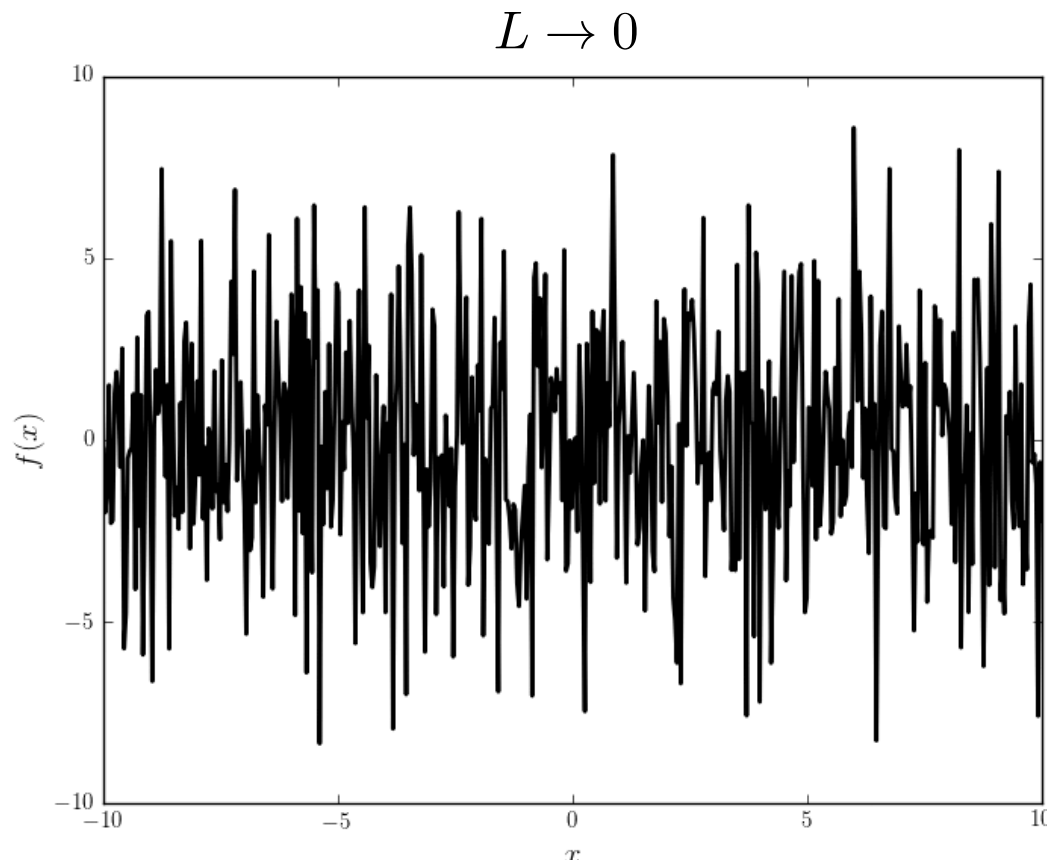
32

1) A prior over functions

$$k(\mathbf{x}_n, \mathbf{x}_m) = e^{-\|\mathbf{x}_n - \mathbf{x}_m\|^2 / L^2}$$

Example 1: RBF kernel, **What will happen if $L \rightarrow 0$?**

What will happen to the correlation between different points?



$$k = I$$

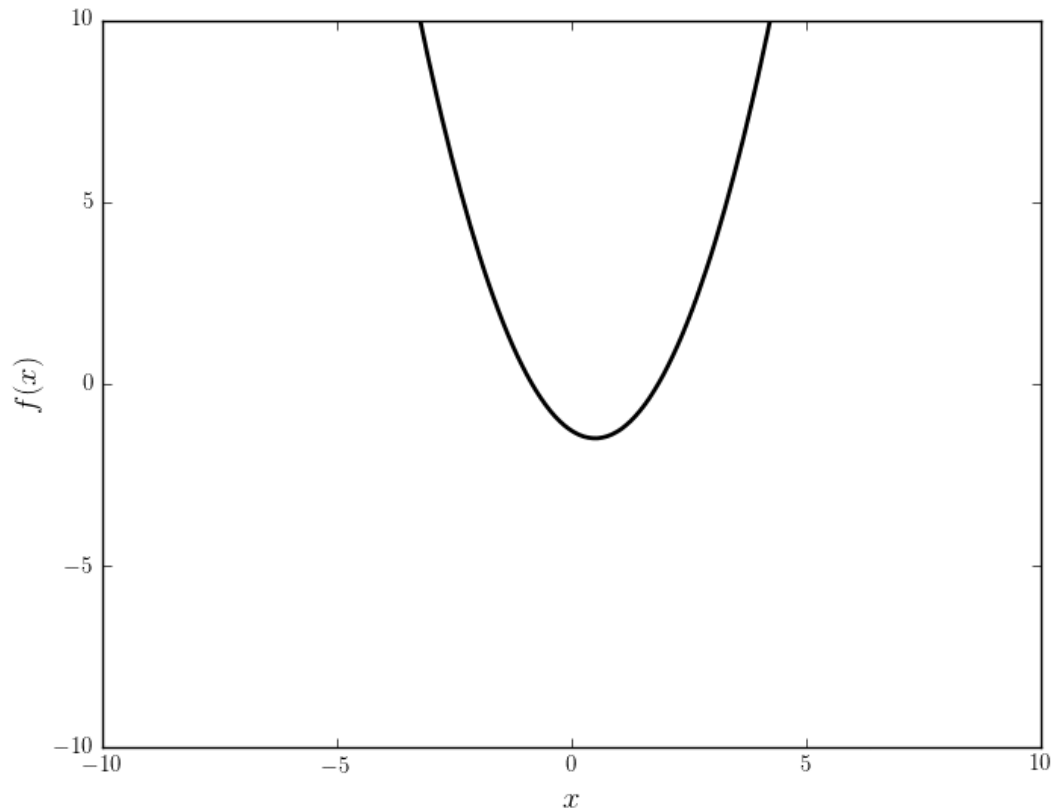
Gaussian Processes

33

1) A prior over functions

Given $k(x_n, x_m)$, we can sample *functions* from $p(\mathbf{f}|X) = \mathcal{N}(\mathbf{f}|\mathbf{0}, K(X))$

Example 2: Quadratic kernel $k(\mathbf{x}_n, \mathbf{x}_m) = (1 + \mathbf{x}_n^T \mathbf{x}_m)^2$



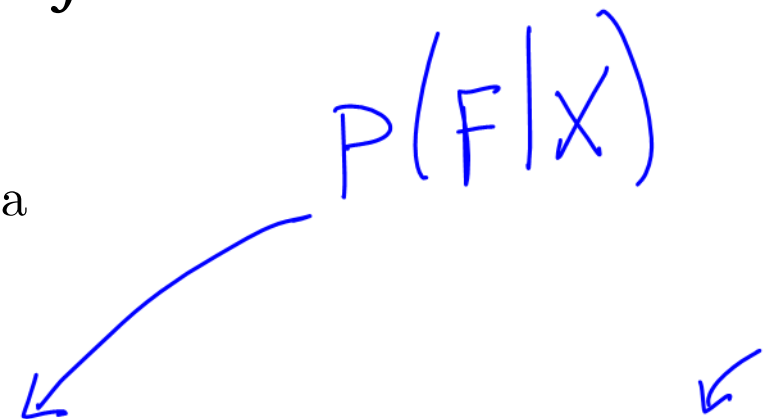
Gaussian Processes

34

2) Incorporating *noise-free* measurements

Notation:

- \mathbf{f} , X : training data
- \mathbf{f}_* , X_* : prediction

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(X, X) & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) \end{bmatrix}\right)$$


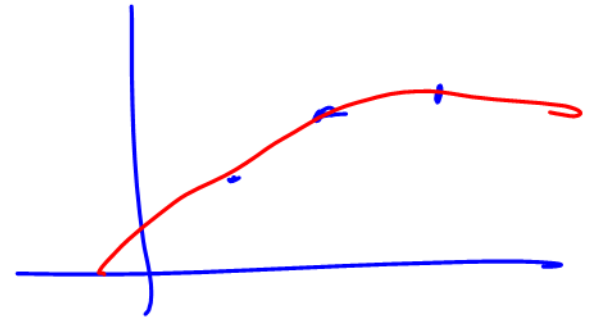
Gaussian Processes

35

2) Incorporating noise-free measurements

Notation:

- \mathbf{f} , X : training data
- \mathbf{f}_* , X_* : prediction



$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(X, X) & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) \end{bmatrix}\right)$$

Conditioning on \mathbf{f} (training data) we get

$$\mathbf{f}_* \mid \mathbf{f}, X_*, X \sim \mathcal{N}(\mu, \Sigma)$$

$$\begin{aligned} \text{with } \mu &= K(X_*, X) K(X, X)^{-1} \mathbf{f} \\ \Sigma &= K(X_*, X_*) - K(X_*, X) K(X, X)^{-1} K(X, X_*) \end{aligned}$$

Gaussian Processes

36

3) Incorporating *noisy* measurements (as in real life)

Assume measurements y are noisy such that

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

and ϵ is i.i.d. with $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$

Therefore, $\text{cov}(\mathbf{y}) = K(X, X) + \sigma_n^2 I$, and we can write

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

Gaussian Processes

37

3) Incorporating *noisy* measurements (as in real life)

Assume measurements y are noisy such that $y = f(\mathbf{x}) + \epsilon$

and ϵ is i.i.d. with $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$

Therefore, $\text{cov}(\mathbf{y}) = K(X, X) + \sigma_n^2 I$, and we can write

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

Conditioning on \mathbf{y} (training data) we get

$$\mathbf{f}_* \mid \mathbf{y}, X_*, X \sim \mathcal{N}(\mu', \Sigma')$$

with

$$\begin{aligned} \mu' &= K(X_*, X) [K(X, X) + \sigma_n^2 I]^{-1} \mathbf{y} \\ \Sigma' &= K(X_*, X_*) - K(X_*, X) [K(X, X) + \sigma_n^2 I]^{-1} K(X, X_*) \end{aligned}$$

Gaussian Processes

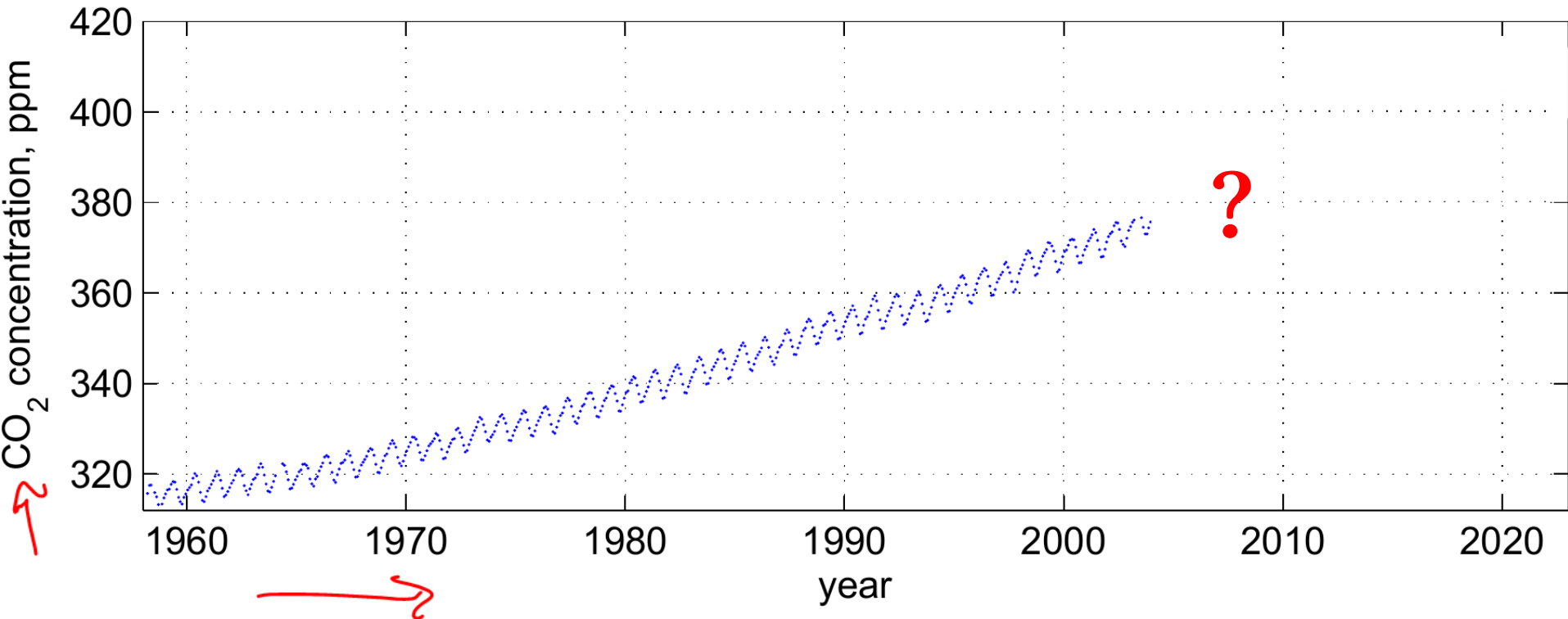
38

Demo Time

Gaussian Processes

39

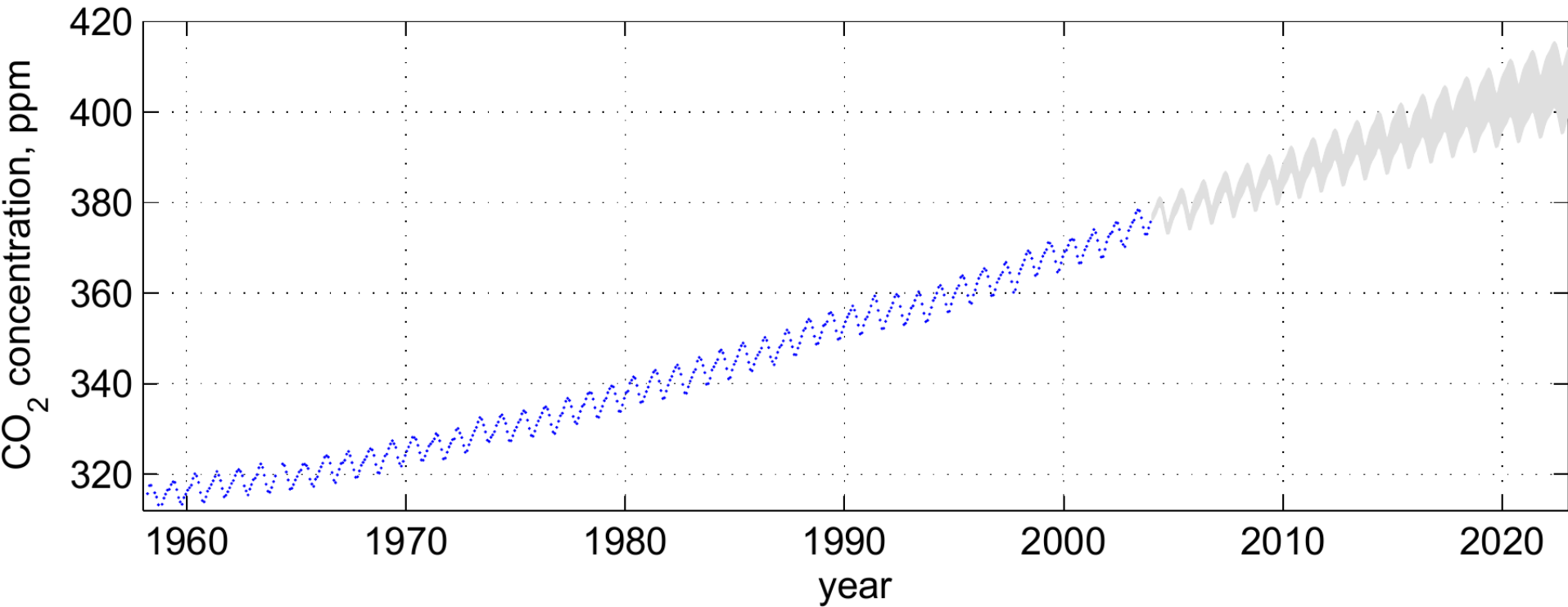
A Real Example [Rasmussen, Williams, Gaussian Processes for Machine Learning]



Gaussian Processes

40

A Real Example [Rasmussen, Williams, Gaussian Processes for Machine Learning]



Gaussian Processes

41

A Real Example [Rasmussen, Williams, Gaussian Processes for Machine Learning]

Kernel Design: $k(x, x') = \underbrace{k_1(x, x')} + \underbrace{k_2(x, x')} + \underbrace{k_3(x, x')} + \underbrace{k_4(x, x')}$

Long-term smoothness

$$k_1(x, x') = \theta_1^2 \exp\left(-\frac{(x - x')^2}{\theta_2^2}\right) \text{ RBF}$$

→ decay

Seasonal trend (periodicity)

$$k_2(x, x') = \theta_3^2 \exp\left[-\frac{-2 \sin^2(\pi(x - x'))}{\theta_5^2}\right] \exp\left(-\frac{1}{2} \frac{(x - x')^2}{\theta_4^2}\right)$$

Short- and medium-term anomaly

$$k_3(x, x') = \theta_6^2 \left(1 + \frac{(x - x')^2}{2\theta_8\theta_7^2}\right)^{-\theta_8}$$

11 parameters

Noise

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

Gaussian Processes

42

What about Classification?

So far $f(\mathbf{x})$ is a real function, not optimal for classification

what would you suggest doing?

$$F(x) \sim N(\mu(x), k(x))$$

$$P(F^* | F_{--}) \sim N(-)$$

Regression

$$P(y=1 | F) = \sigma(F)$$
$$F \sim N(\dots)$$

↓

$$\sigma(F) \text{ is NOT}$$

Classification

GPML matlabs toolbox

Gaussian Processes

Summary

- GPs place a prior over functions through $p(\mathbf{f}|X) = \mathcal{N}(\mathbf{f}|\mathbf{0}, K(X))$
→ $K(X)$ defines 'shape' and prior knowledge about our problem
- Prediction = Prior | Measurements
(Tightly linked to Bayesian Estimation)
- GPs can be applied to classification
(and many other applications, eg. Dimensionality Reduction, Latent Variable Models...)