Introduction to Gaussian Processes

Pattern Recognition & Machine Learning Course, EPFL

December 2015

Carlos Becker, Emtiyaz Khan

Motivation

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 Rasmusen and William's "Gaussian Processes for Machine Learning" book.

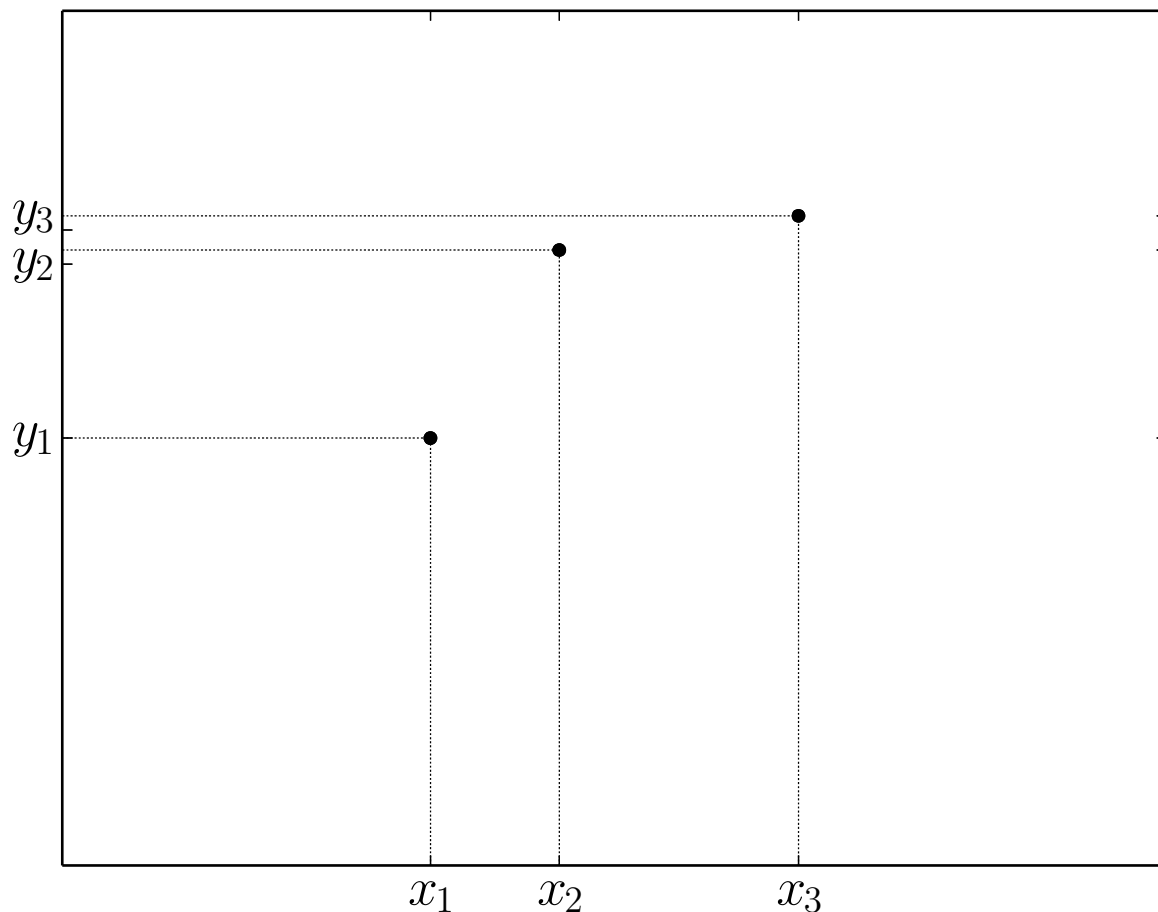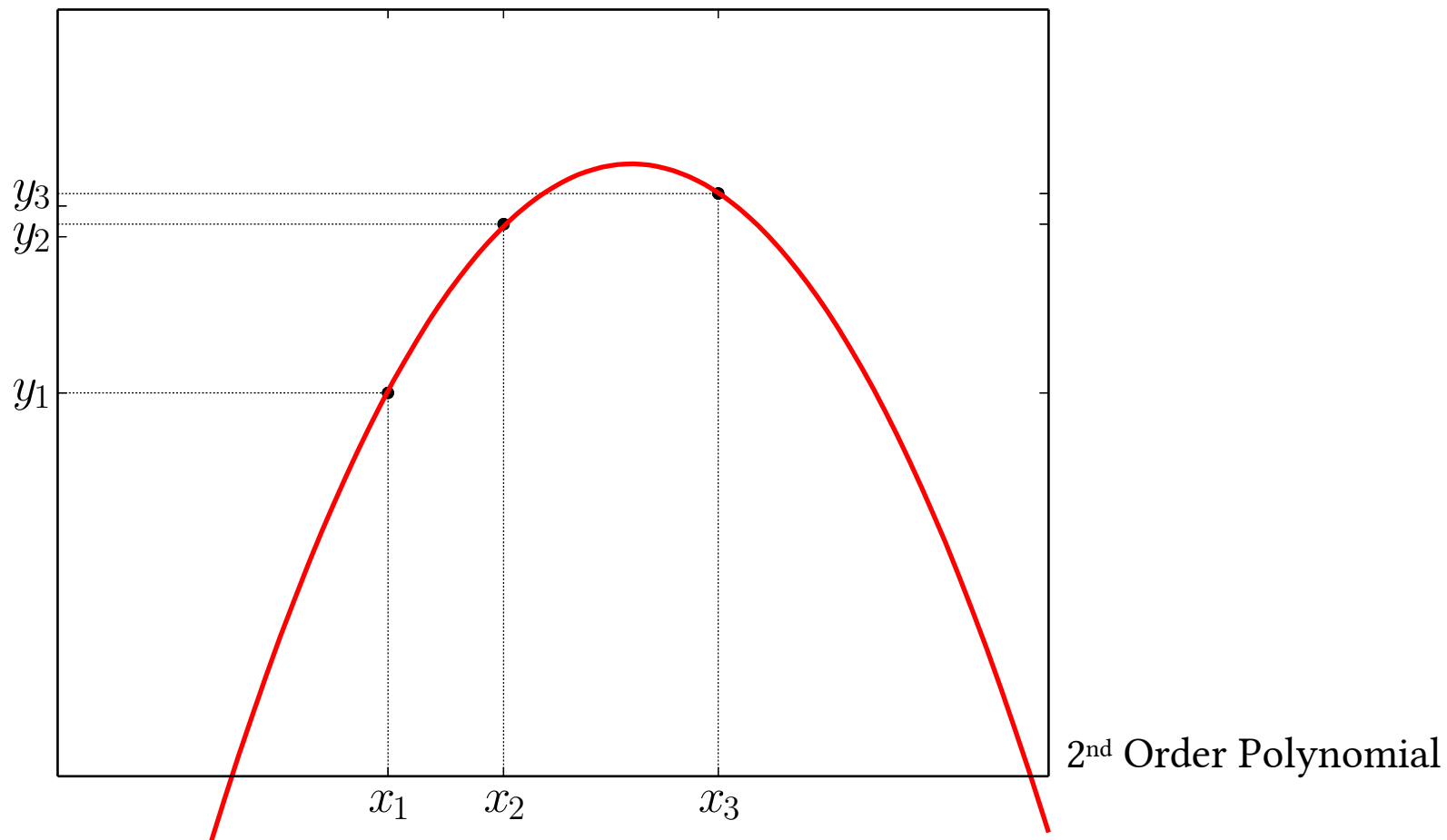# Motivation: why Gaussian Processes?

Motivation

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

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

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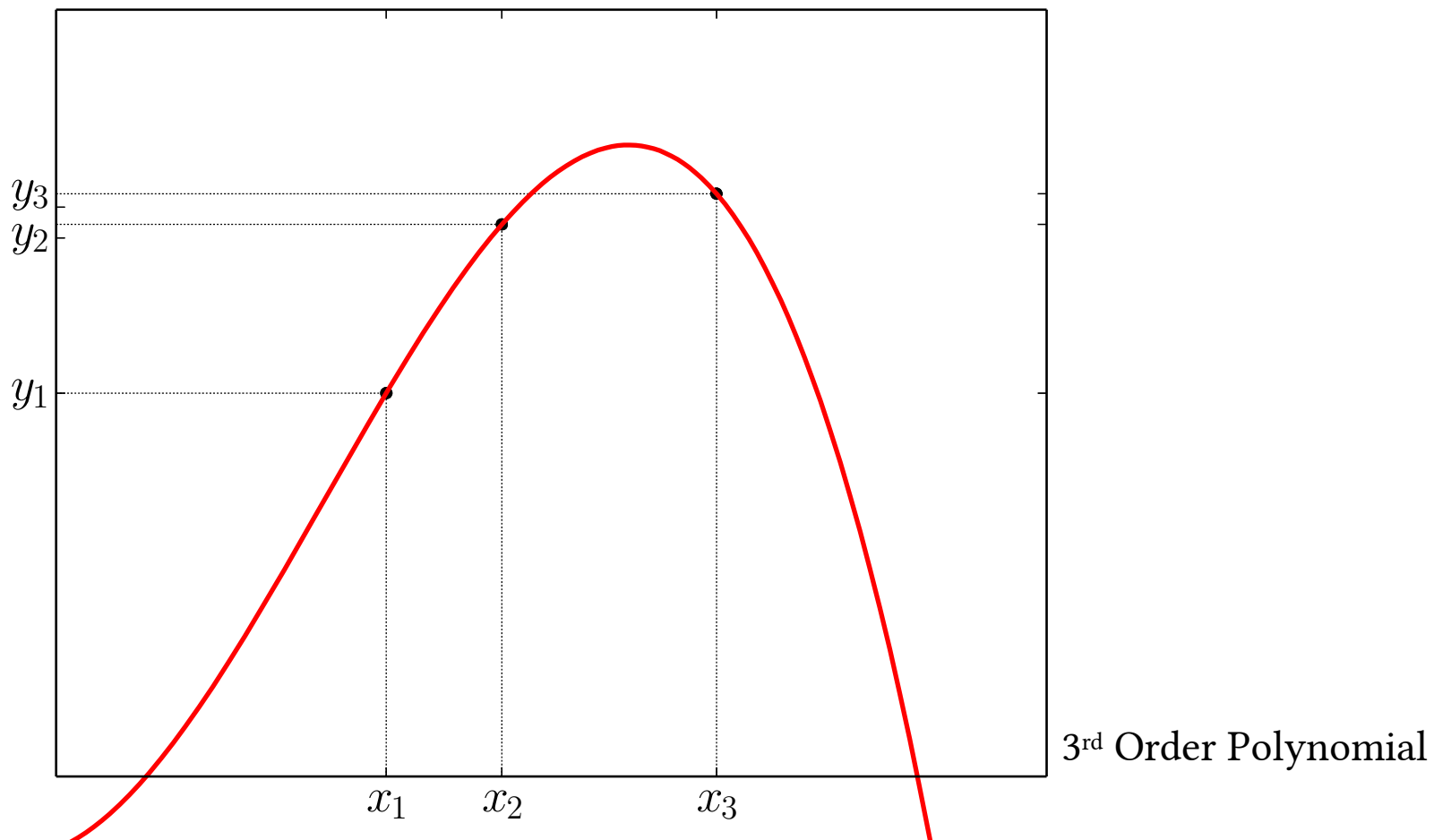

2nd Order Polynomial
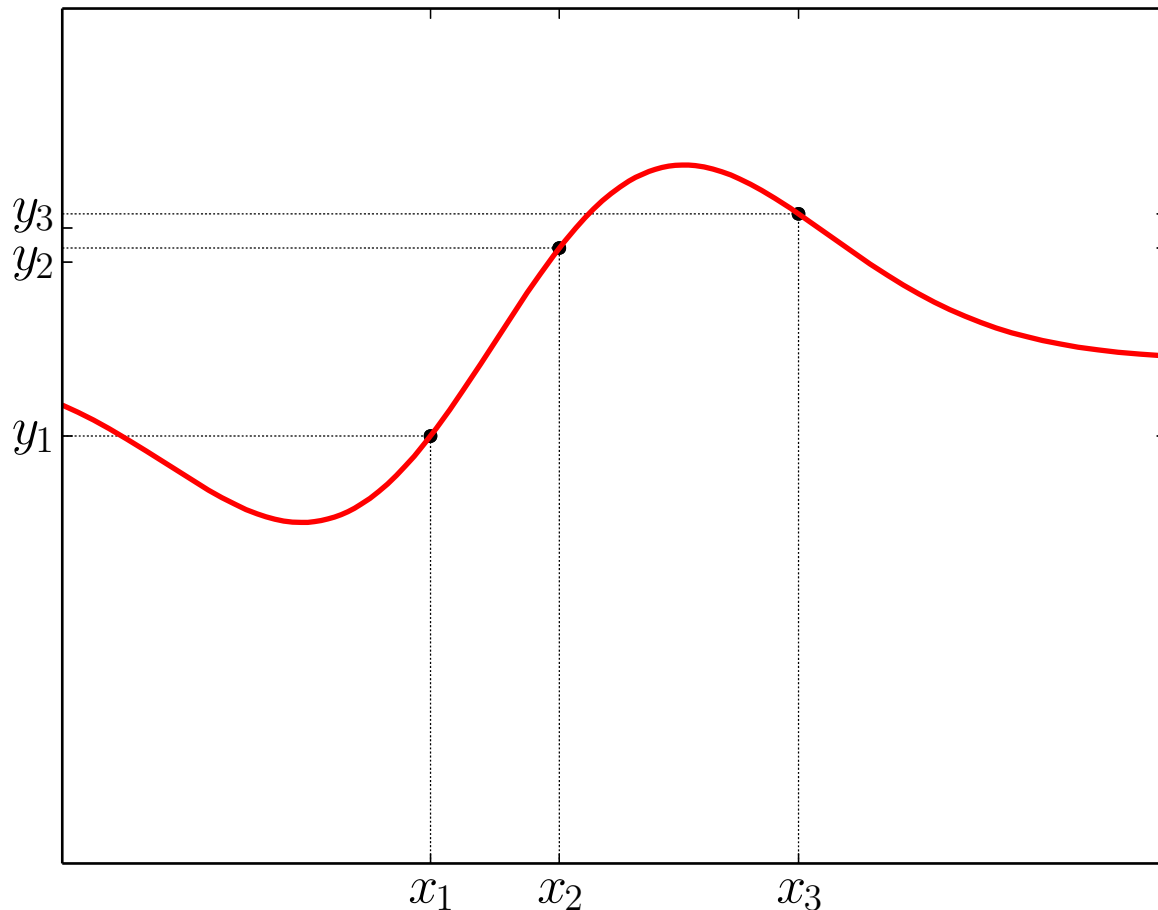
Motivation

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

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

3rd Order Polynomial

Motivation

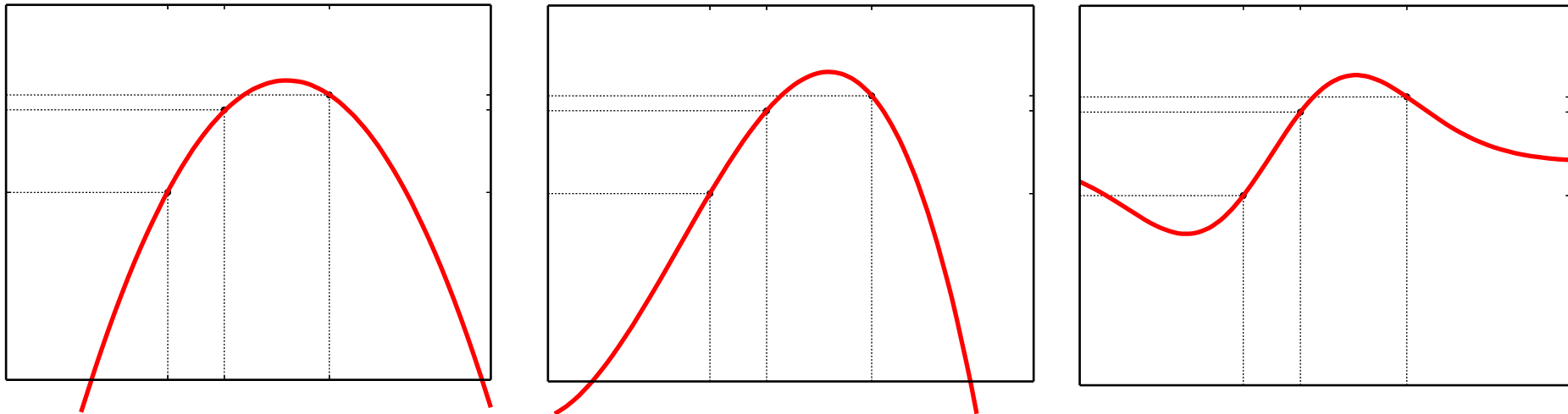Say we want to estimate a scalar function $f(x)$

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

Radial Basis Function

Motivation

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

And this **prior** is formulated probabilistically
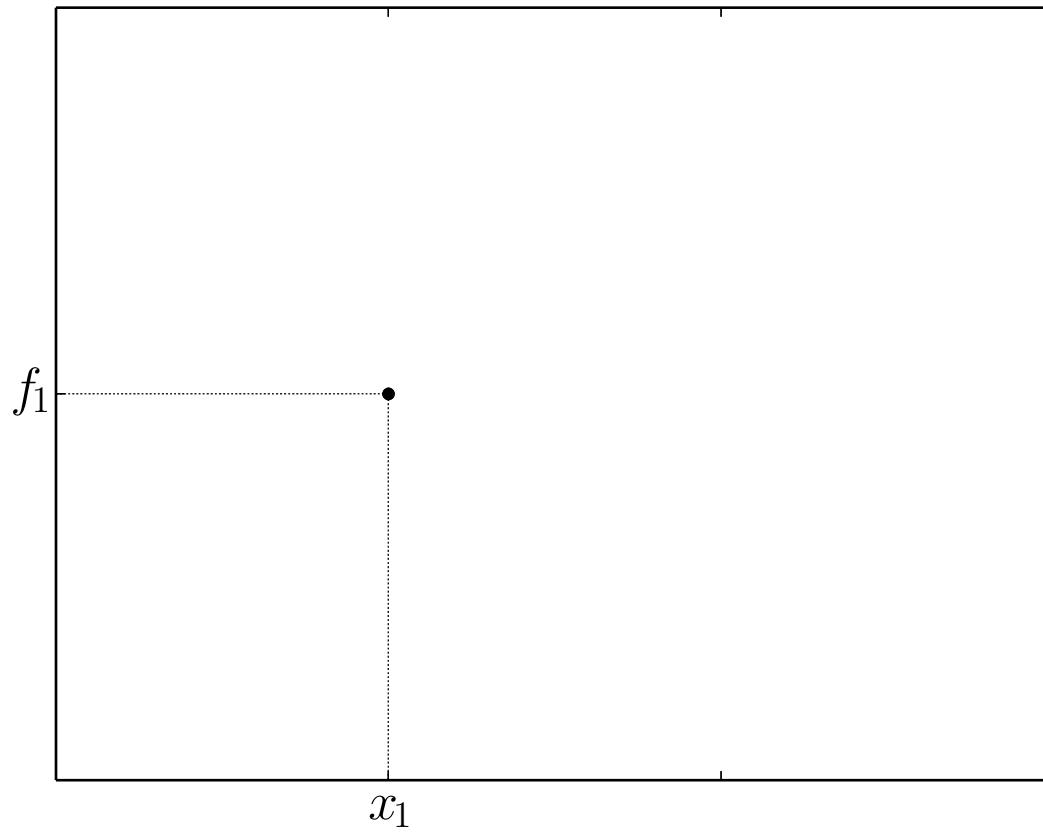
Let's get started!

Just before…

Gaussian Processes

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

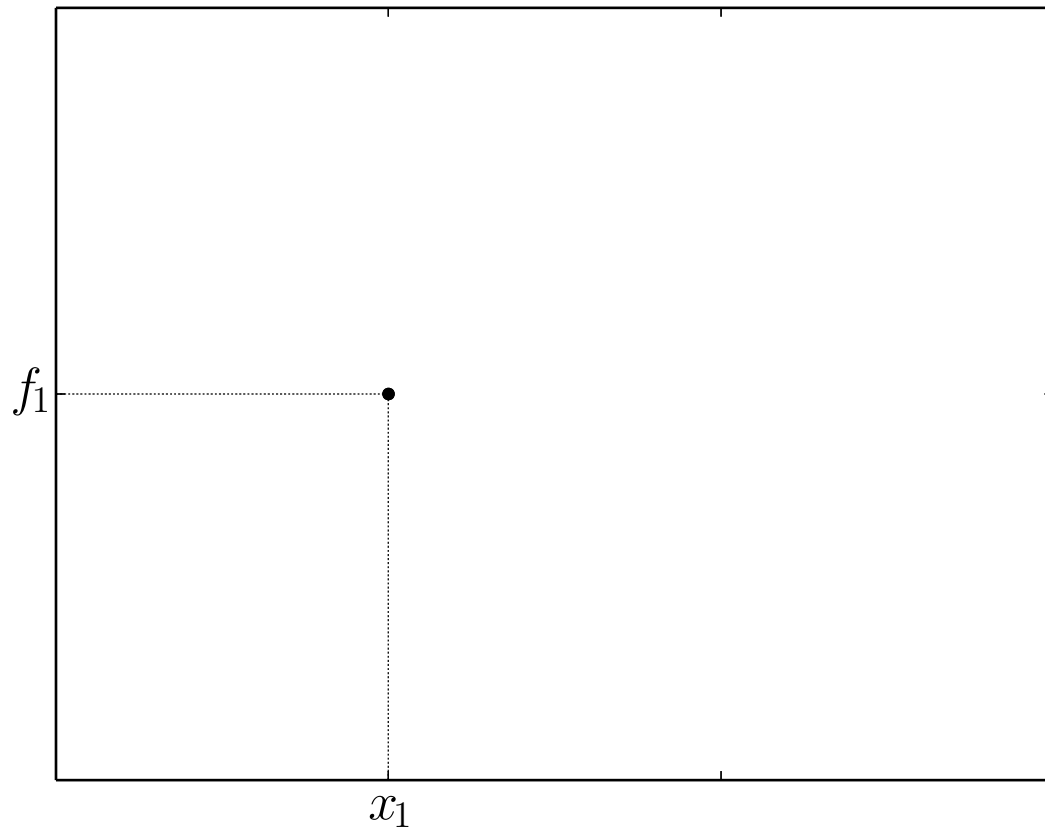- From a single observation $\{x_1, f_1\}$
- We want to predict $f_* = f(x_*)$

Intuition?

Gaussian Processes

Now, in a probabilistic manner...

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

Intuition?

f_1 •

x_1

Gaussian Processes

Getting there...

The estimated f_* is now a Gaussian RV: $\boxed{f_* \sim \mathcal{N}\left(\mu_*, \sigma_*^2\right)}$
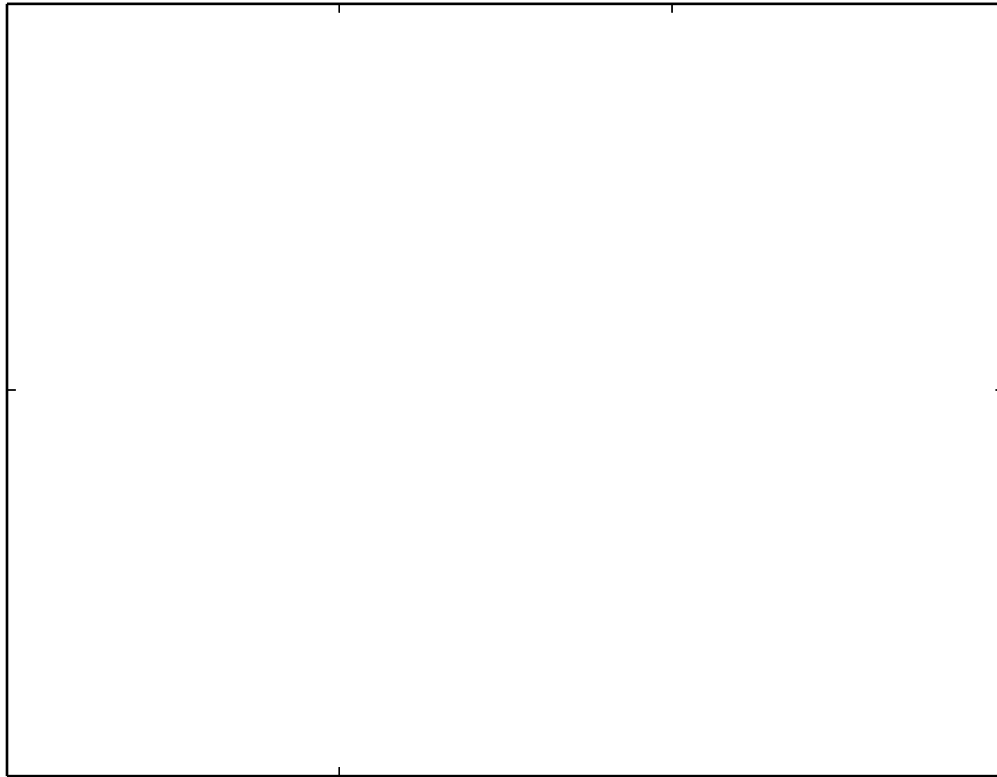
Intuition?
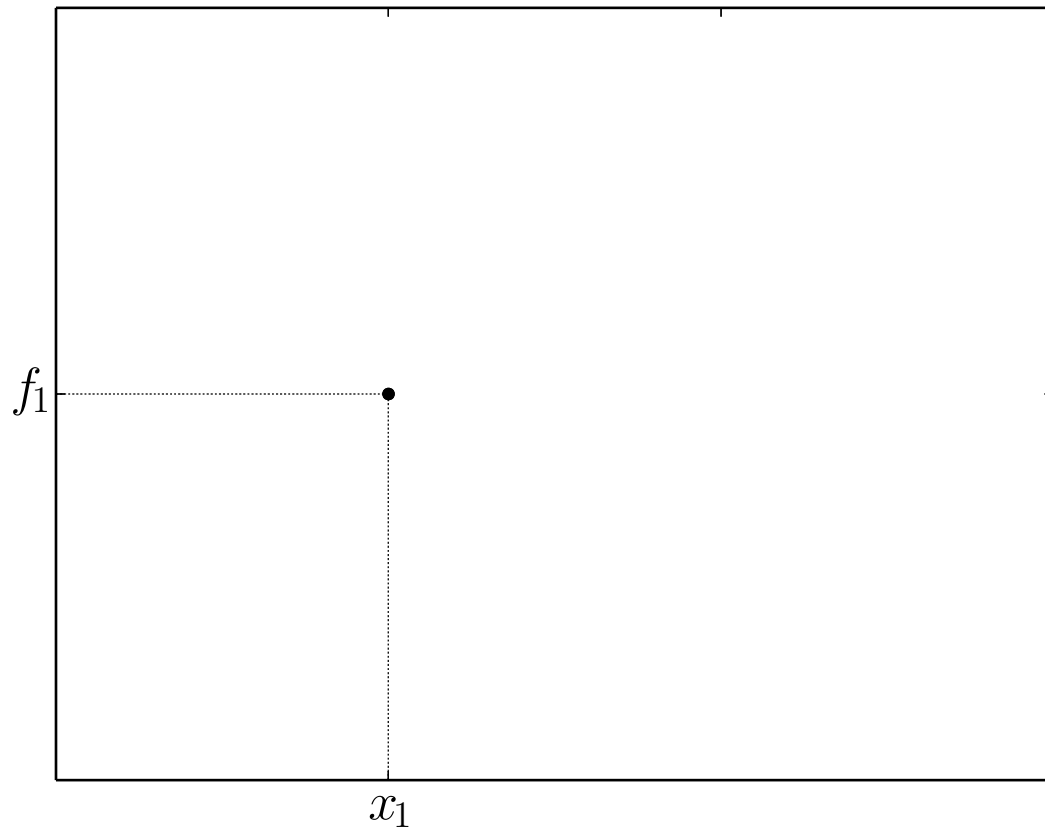
Gaussian Processes

What about multiple training points?

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

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

Gaussian Processes

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

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

More precisely,

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

Gaussian Processes

Incorporating the measurement:

Gaussian Processes

Incorporating the measurement:

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

Gaussian Processes

Incorporating the measurement:

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

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

Gaussian Processes

Incorporating the measurement:

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

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

And we want to estimate

Gaussian Processes

Incorporating the measurement:

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

Our model is $\quad p(f_1, f_* \mid x_1, x_*)$

And we want to estimate

What we want is $\quad p(f_* \mid f_1, x_1, x_*)$

Gaussian Processes

Incorporating the measurement:

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

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

And we want to estimate

What we want is $p(f_* \mid f_1, x_1, x_*)$
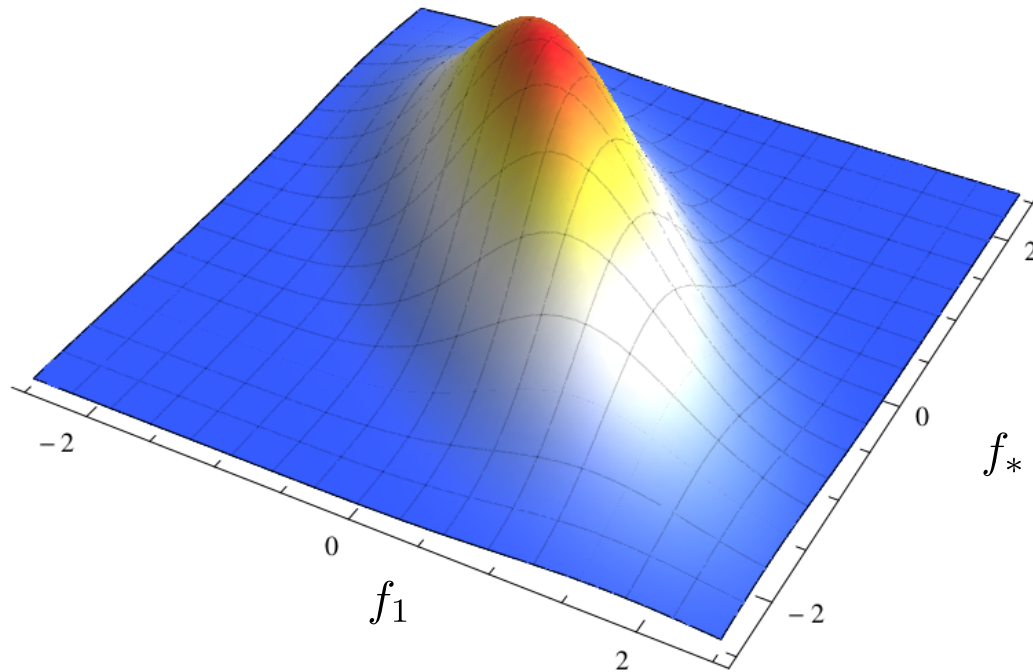
Gaussian Processes

Summarizing our 2-point Gaussian Process:

· Our model or **prior** is

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

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

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

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

We get a probability distribution as the output.

Exercise

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

· For our model prior

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

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

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

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

4) What if $x_* = 10$?

Time to get to the Real GP

Gaussian Processes

A gaussian process defines a **prior over functions f**,

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

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

1) A prior over functions

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

Example 1: RBF kernel $k(\mathbf{x}_n, \mathbf{x}_m) = e^{-||\mathbf{x}_n - \mathbf{x}_m||^2/L^2}$
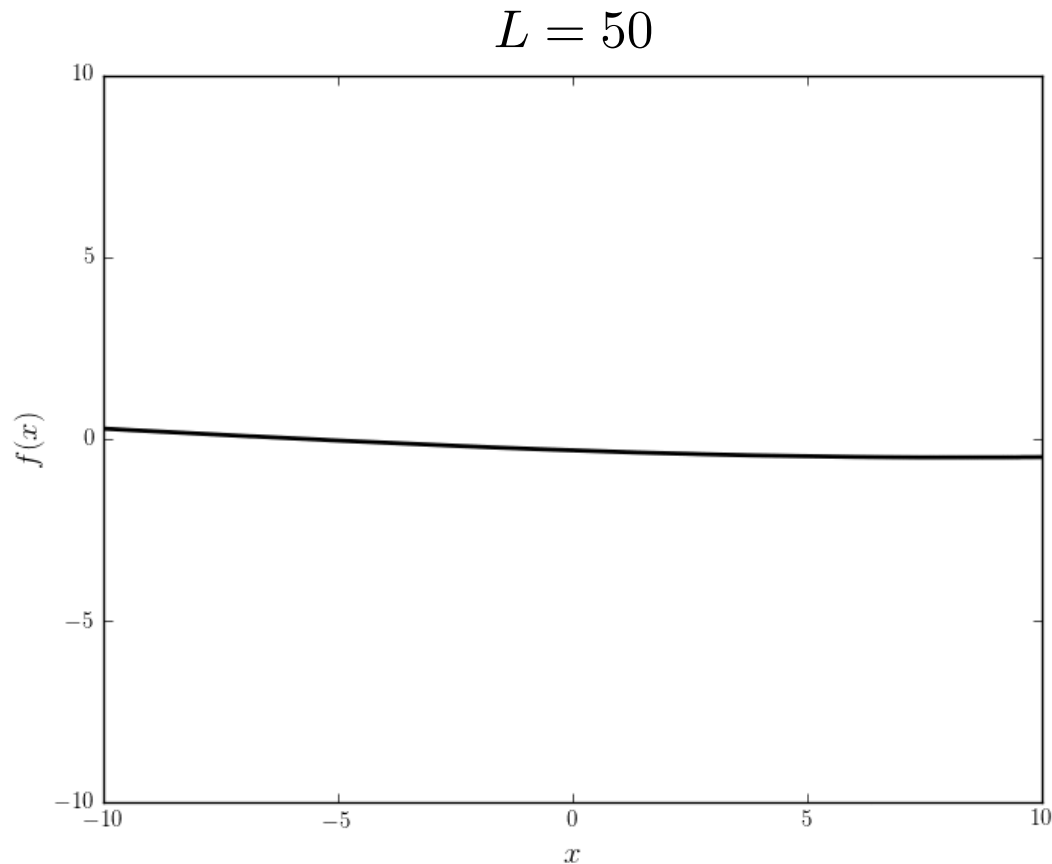
$$L = 50$$

Gaussian Processes

1) A prior over functions

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

Example 1: RBF kernel $k(\mathbf{x}_n, \mathbf{x}_m) = e^{-||\mathbf{x}_n - \mathbf{x}_m||^2/L^2}$
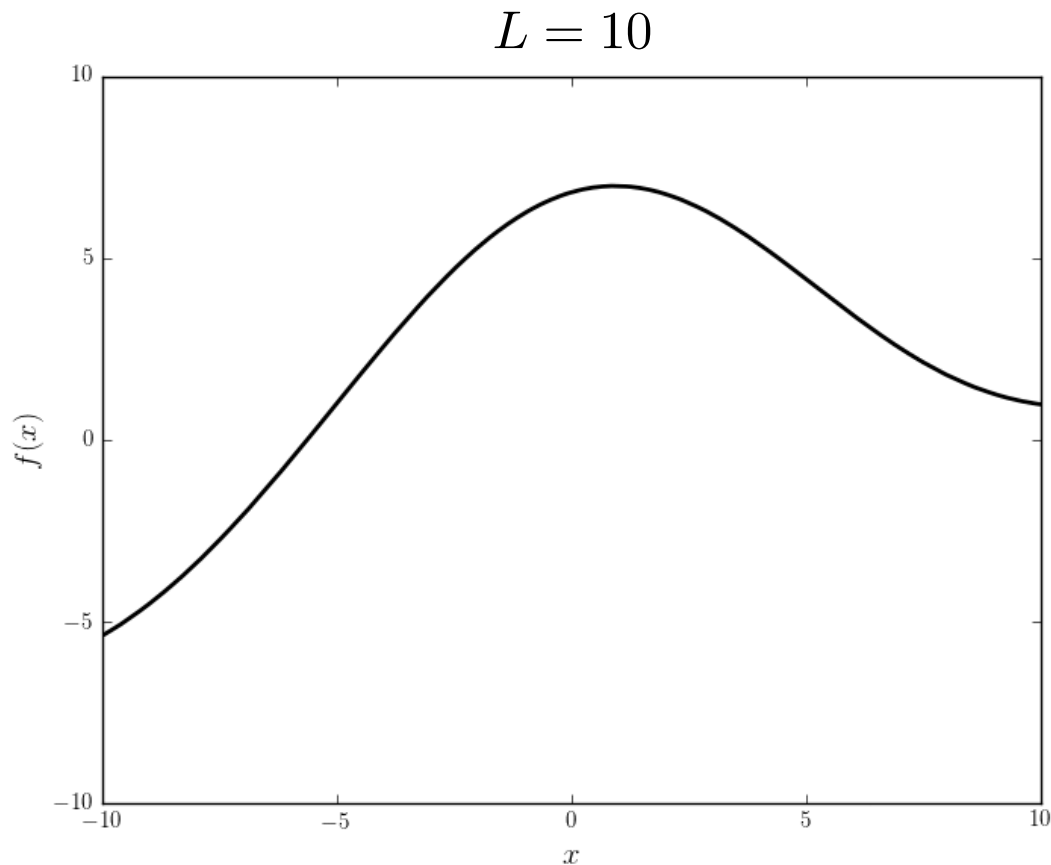
$$L = 10$$

Gaussian Processes

1) A prior over functions

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

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

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 $L \to 0$?**

Gaussian Processes

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 $L \to 0$?**

What will happen to the correlation between different points?
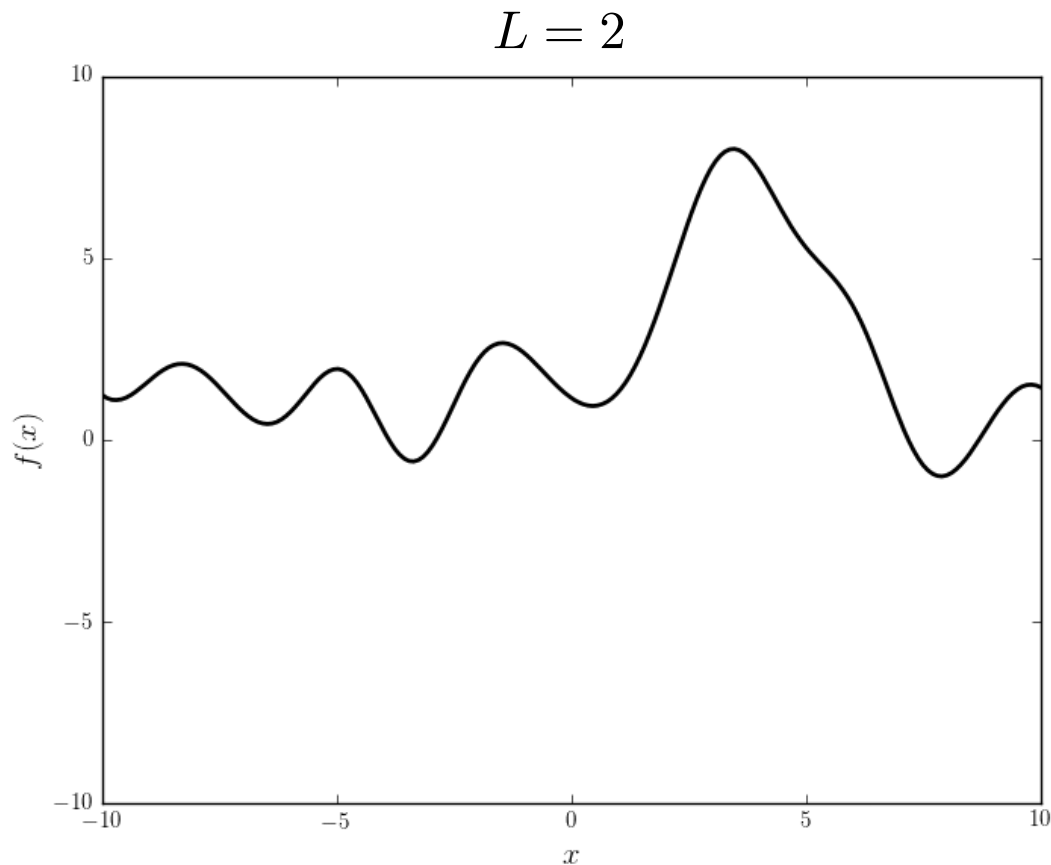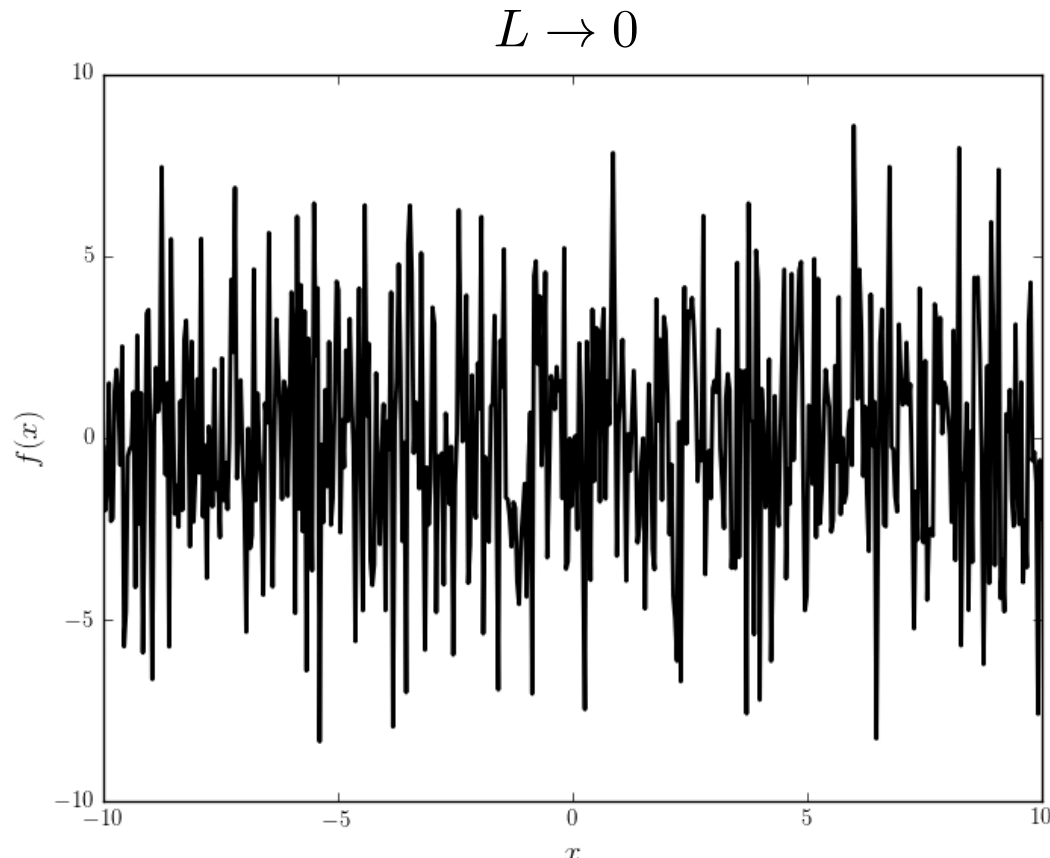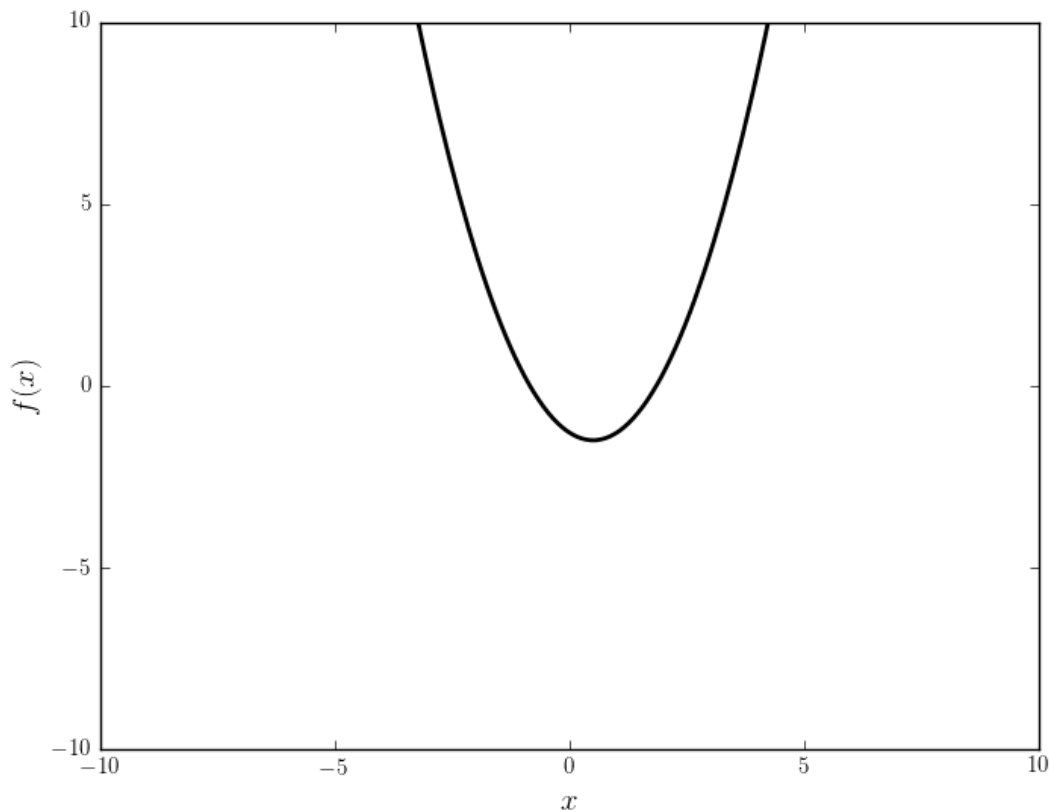
$$L \to 0$$

Gaussian Processes

1) A prior over functions

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

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

Gaussian Processes

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

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

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

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

Gaussian Processes

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

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

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

with
$$\mu' = K(X_*, X) \left[K(X, X) + \sigma_n^2 I\right]^{-1} \mathbf{y}$$
$$\Sigma' = K(X_*, X_*) \,-\, K(X_*, X) \left[K(X, X) + \sigma_n^2 I\right] K(X, X_*)$$
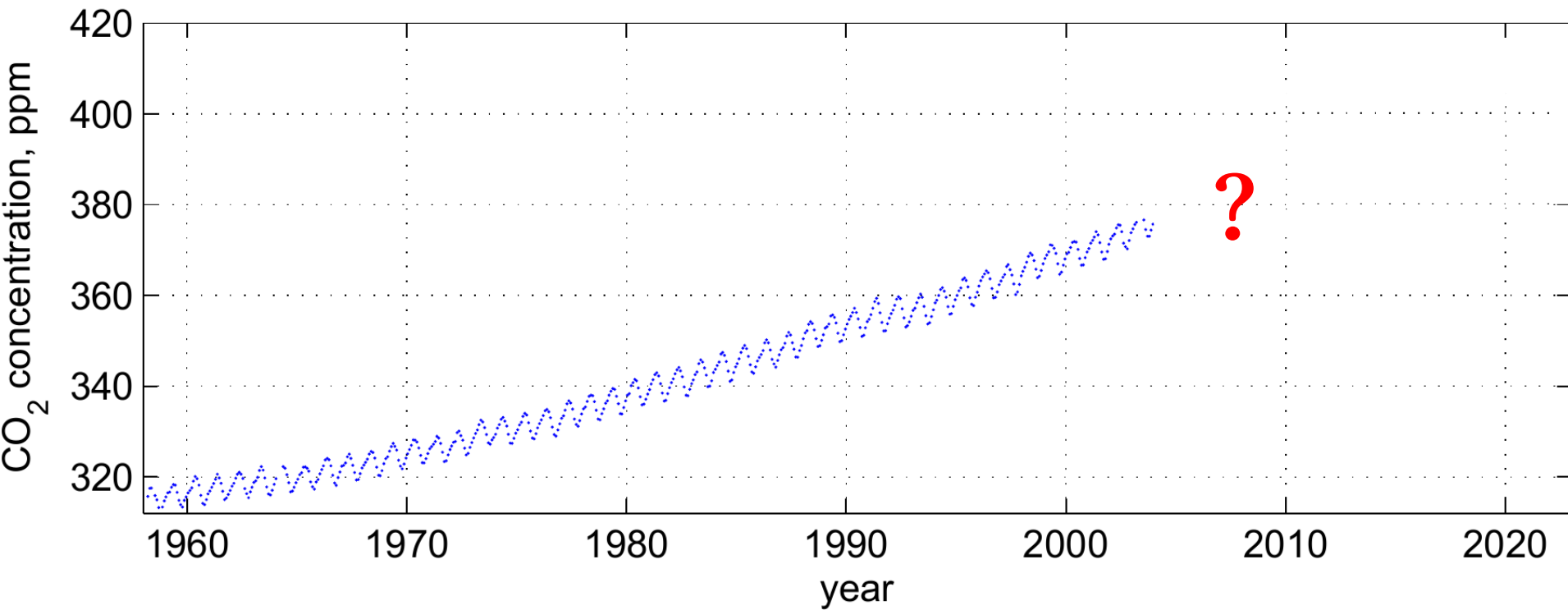
Gaussian Processes

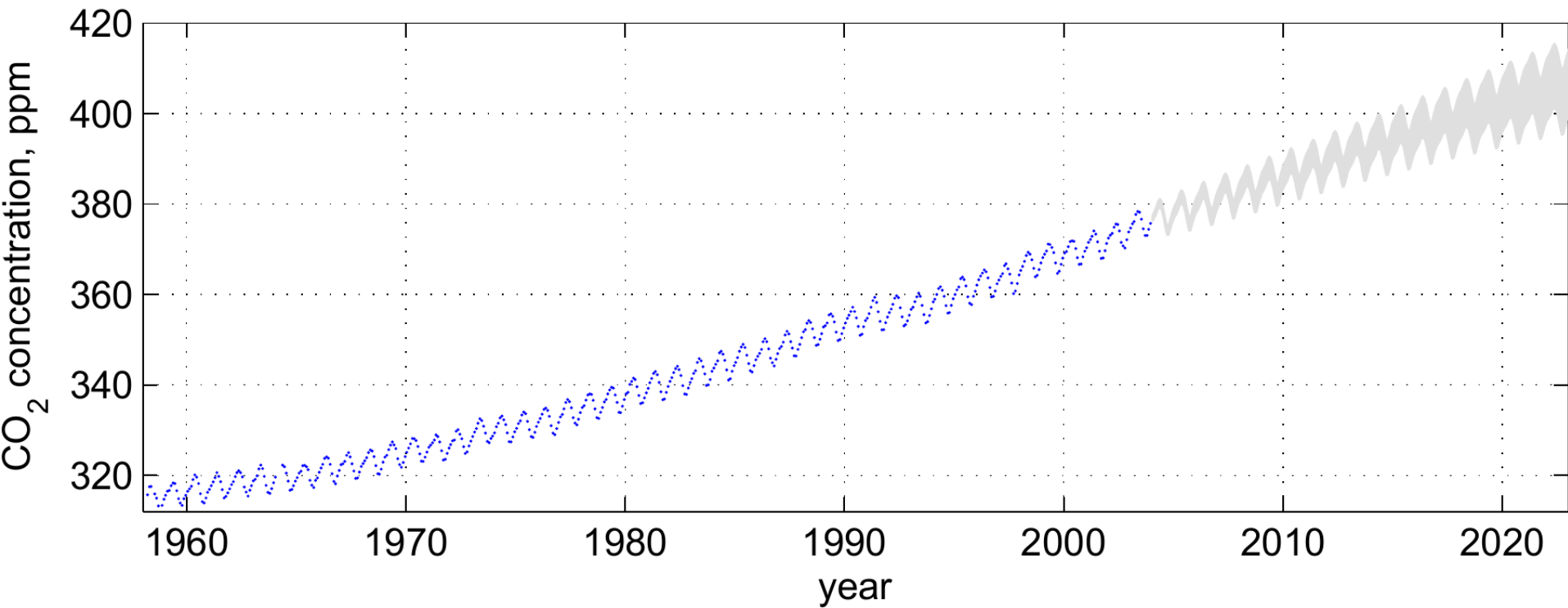**Demo Time**

Gaussian Processes

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

Gaussian Processes

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

Gaussian Processes

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

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

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

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

What about Classification?

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

what would you suggest doing?

Gaussian Processes

Summary

- GPs place a prior over functions through $p(\mathbf{f}|X) = \mathcal{N}\Big(\mathbf{f}|\mathbf{0}, K(X)\Big)$

    $\rightarrow K(X)$ 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...)