UNIVERSITY OF CAMBRIDGE

Machine Learning and the Physical World

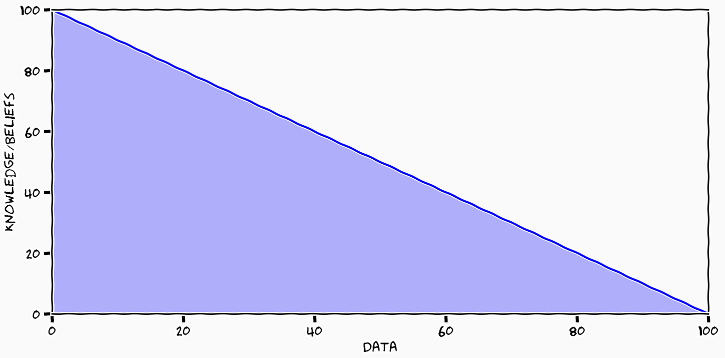
Lecture 3 : Gaussian Processes

Carl Henrik Ek - che29@cam.ac.uk

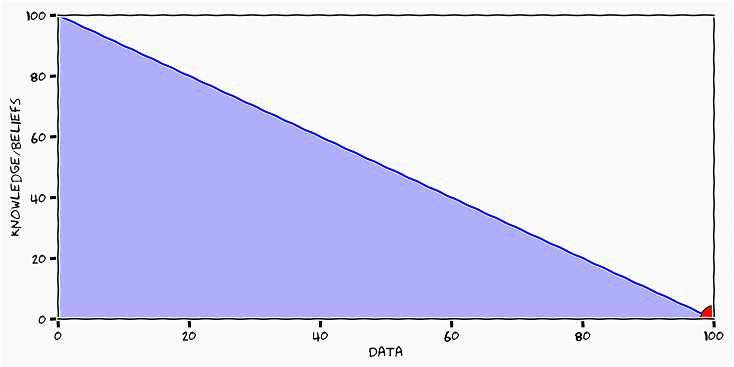
15th of October, 2020

http://carlhenrik.com

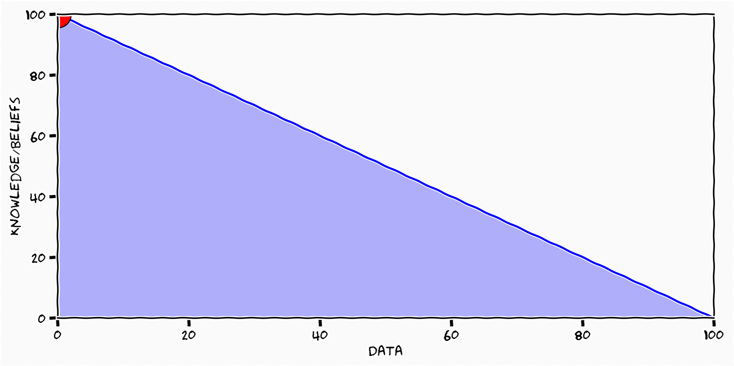
Data and Knowledge



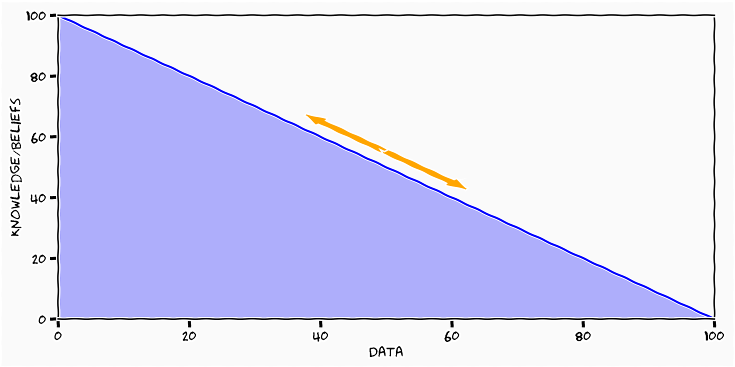
Data and Knowledge



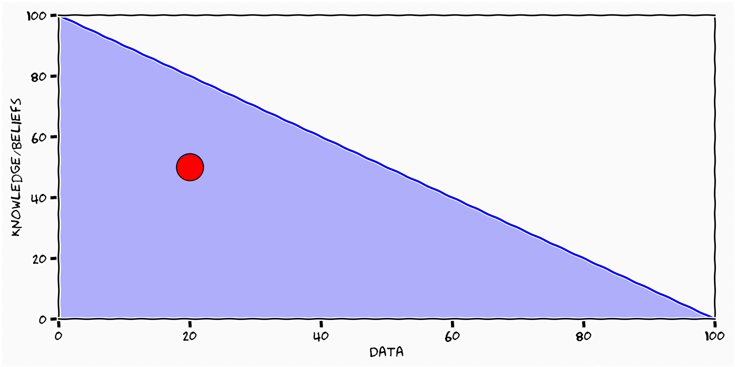
Data and Knowledge



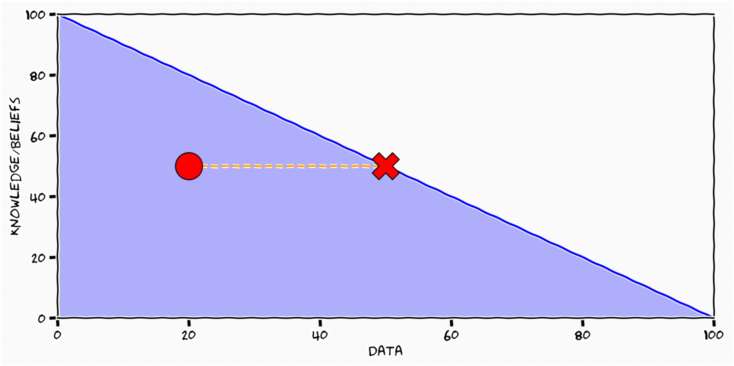
Data and Knowledge



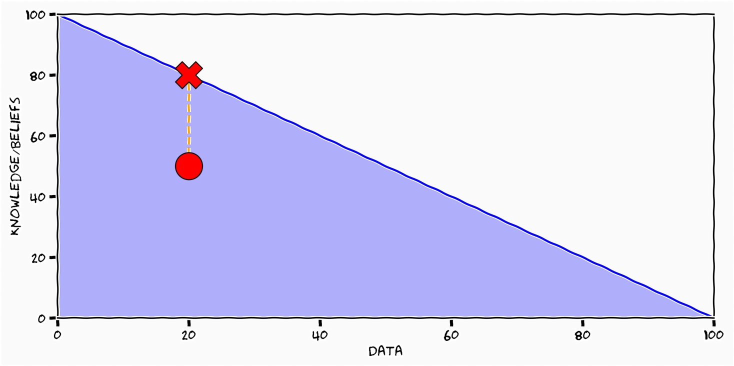
Data and Knowledge



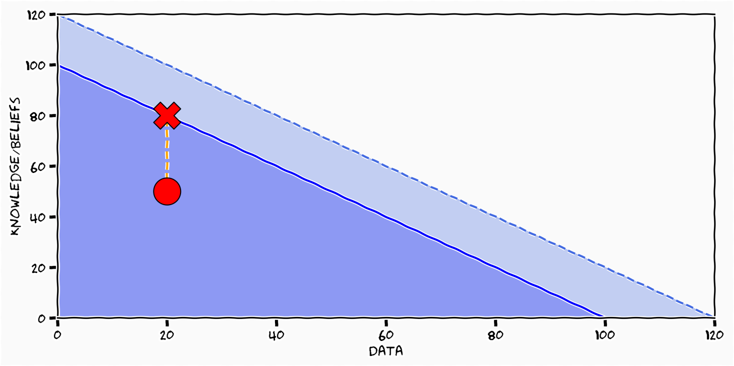
Data and Knowledge



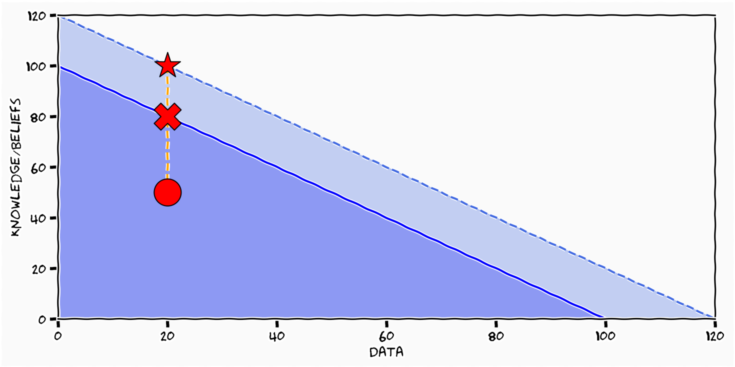
Data and Knowledge



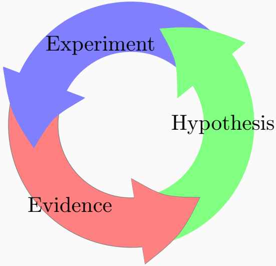
Data and Knowledge



Data and Knowledge

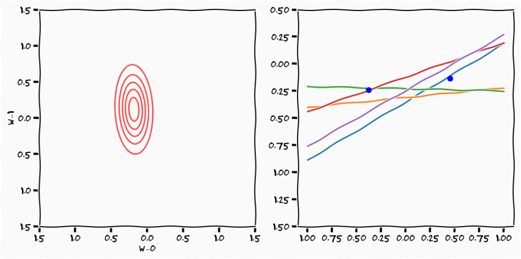


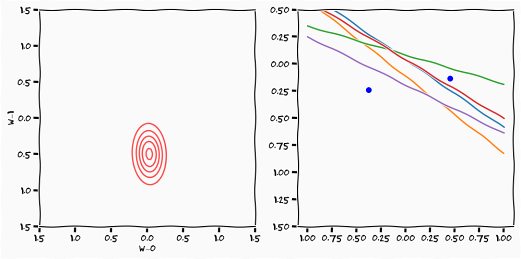
The Scientific Principle



$$p(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{\int p(y \mid \theta)p(\theta)\mathrm{d}\theta}$$

Knowledge is Relative





We must speak our minds openly, debate our disagreements honestly, but always pursue solidarity.

The leader of the free world

Machine Learning

- Machine Learning as a Framework

$$\text{Data} + \text{Model} \overbrace{\rightarrow}^{Compute} \text{Prediction}$$

Machine Learning

- Machine Learning as a Framework

$$\text{Data} + \text{Model} \overbrace{\rightarrow}^{Compute} \text{Prediction}$$

- Machine Learning as a Science
    - how to construct "handles" to allow us to input knowledge

Today

- How to specify beliefs over larger classes of hypothesis

Today

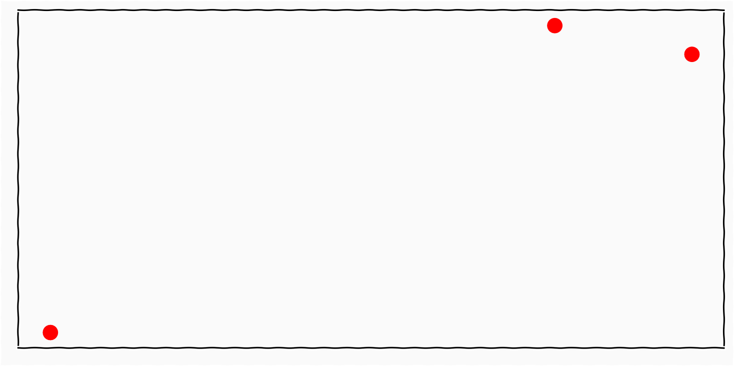
- How to specify beliefs over larger classes of hypothesis
- Non-parametrics

Today

- How to specify beliefs over larger classes of hypothesis
- Non-parametrics
- Gaussian processes

Non-parametrics

Hypothesis Classes

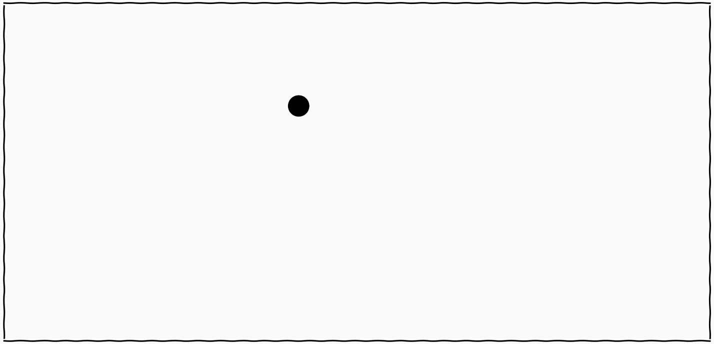


Non-parametrics

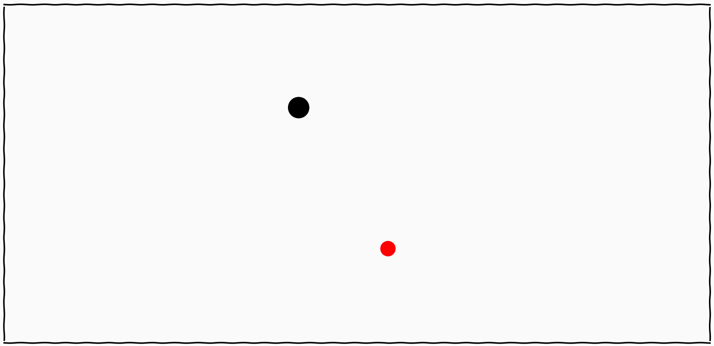
Non-parametrics

Non-parametrics

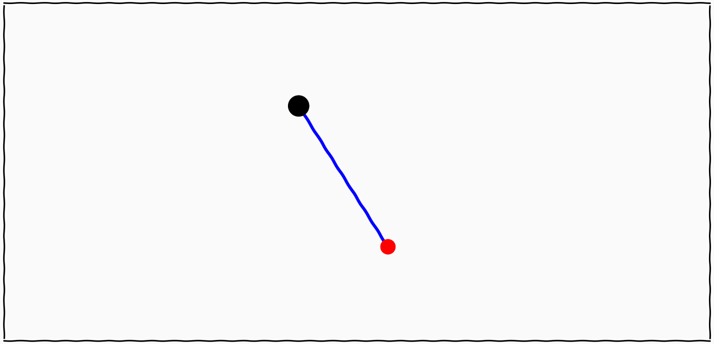
Example



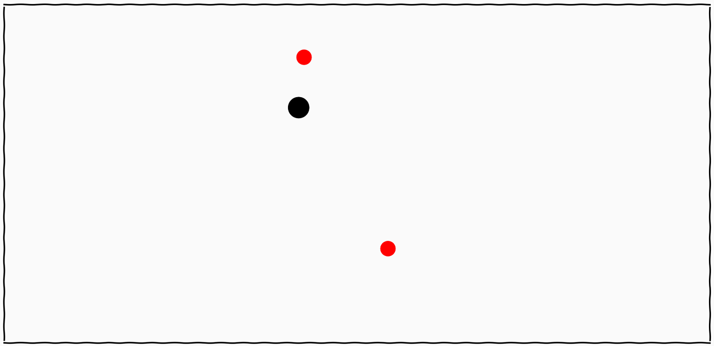
Example



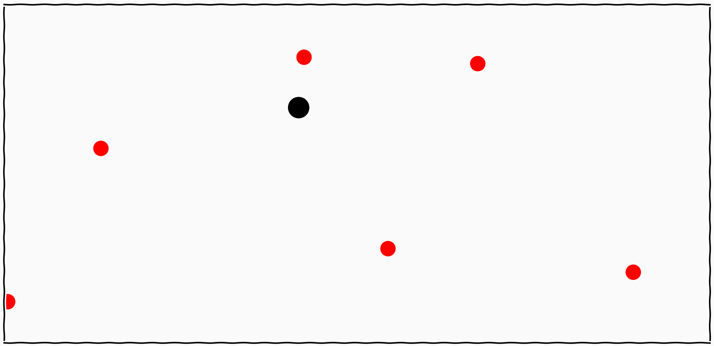
Example



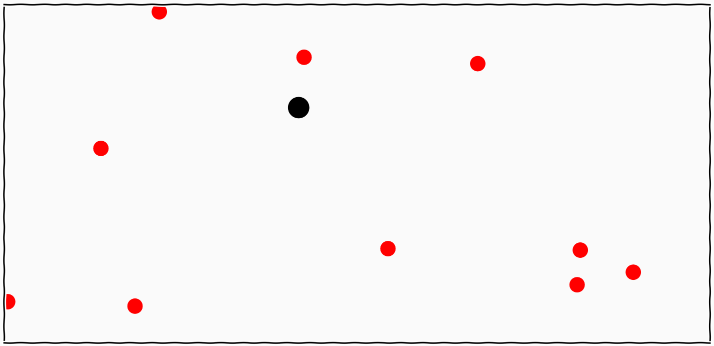
Example



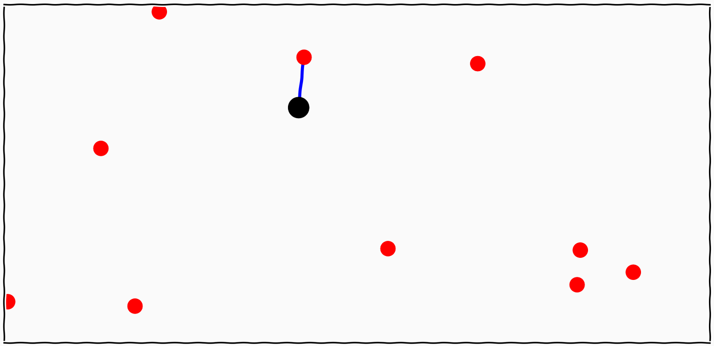
Example



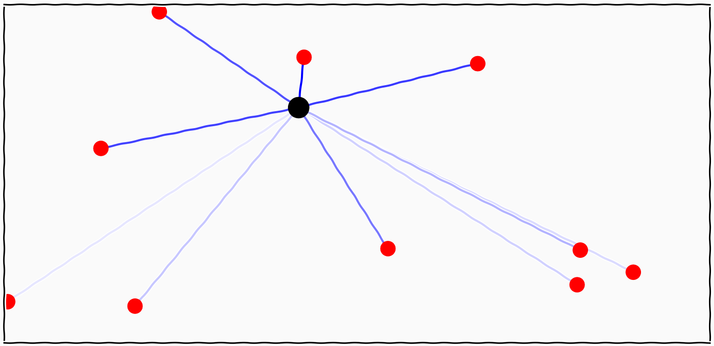
Example



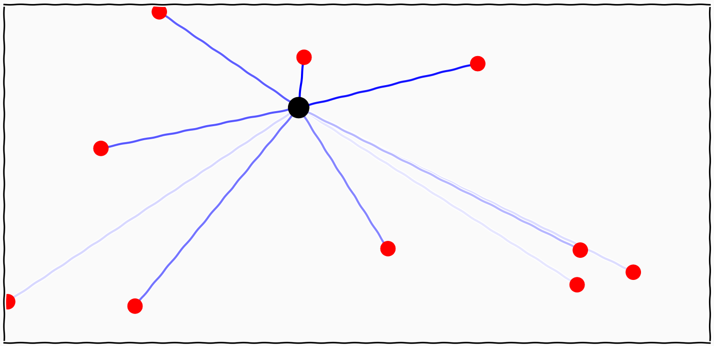
Example



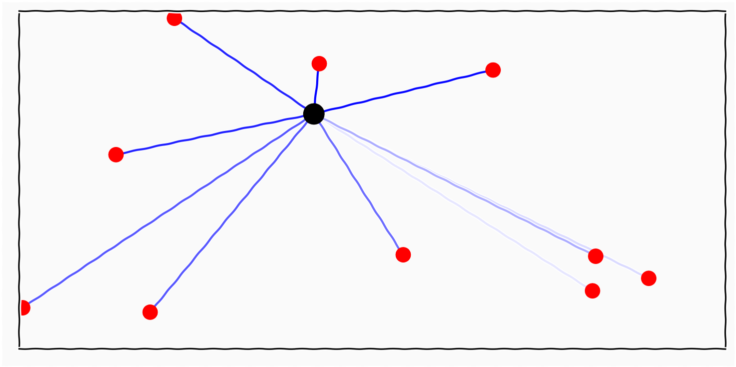
Example



Example

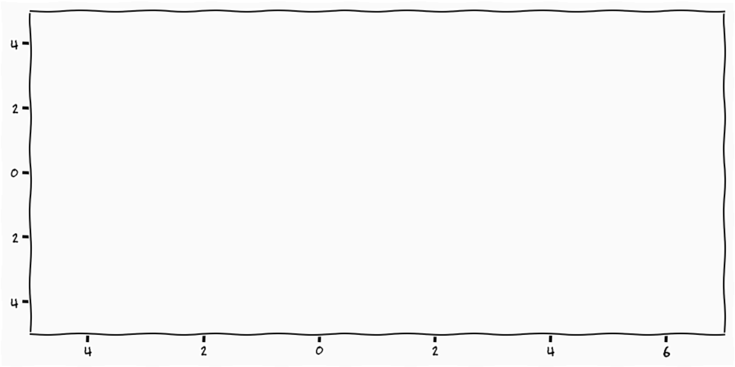


Example

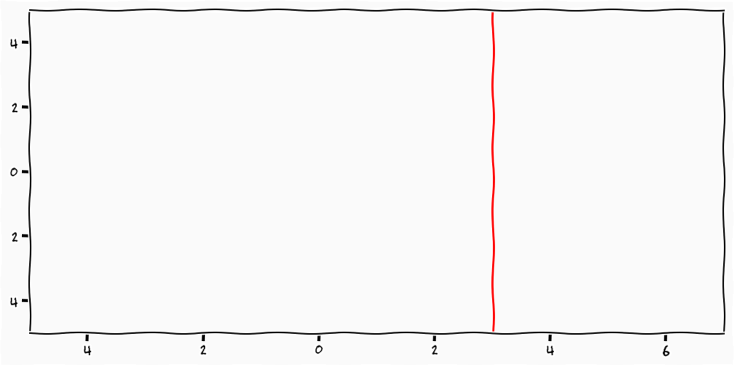


Non-parametric Functions

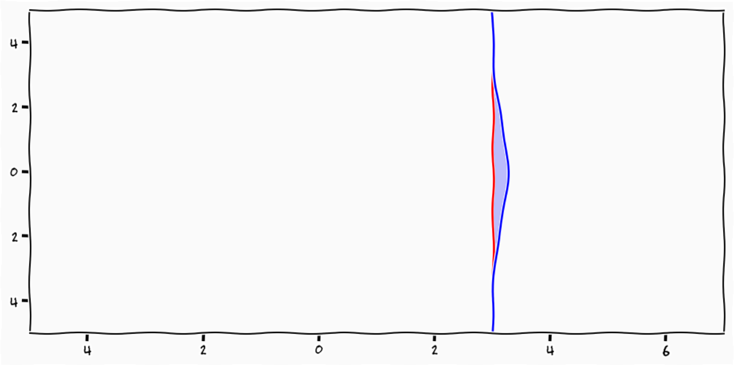
Lets talk about functions



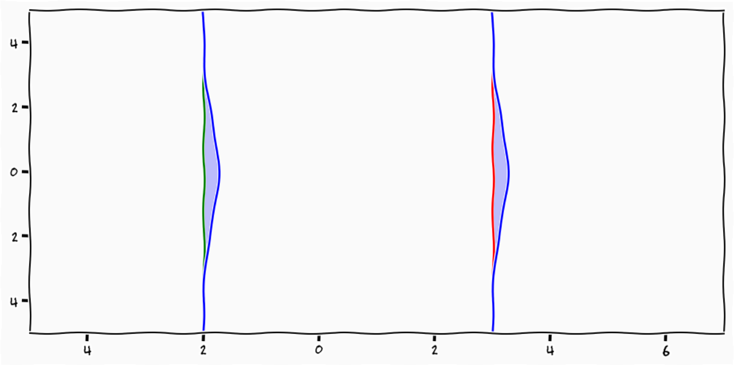
Lets talk about functions



Lets talk about functions



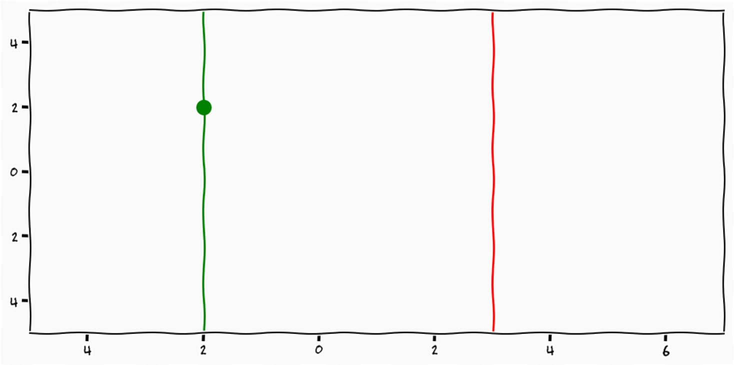
Lets talk about functions



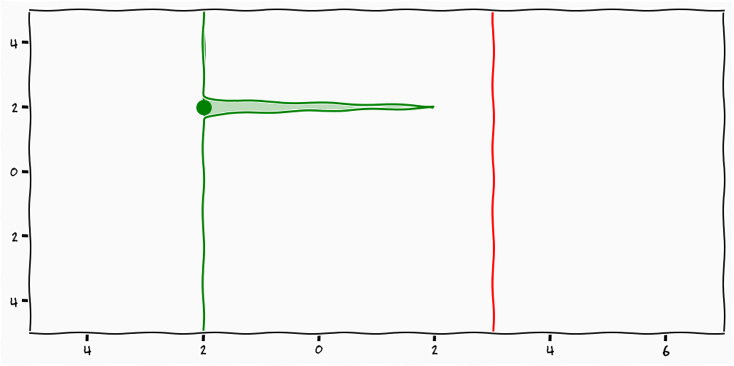
Jointly Gaussian function values

$$\begin{bmatrix} f_1 \\ f_2 \end{bmatrix} = \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{bmatrix} \right)$$

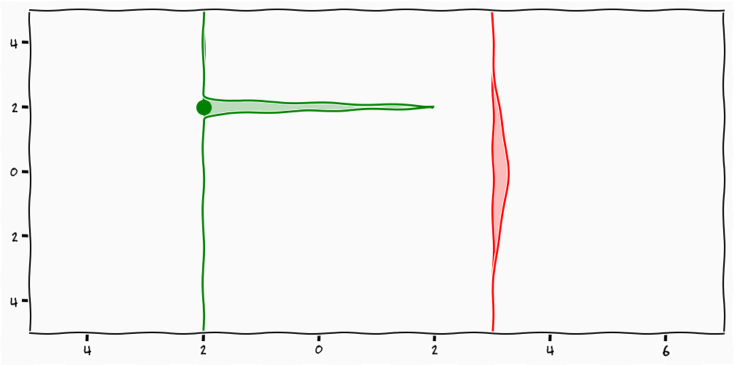
Non-parametric functions



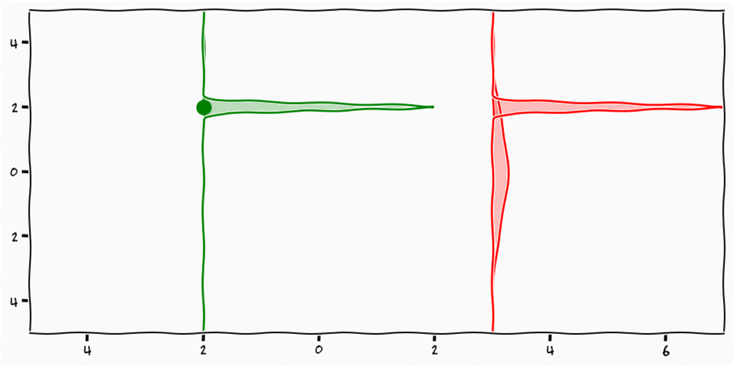
Non-parametric functions



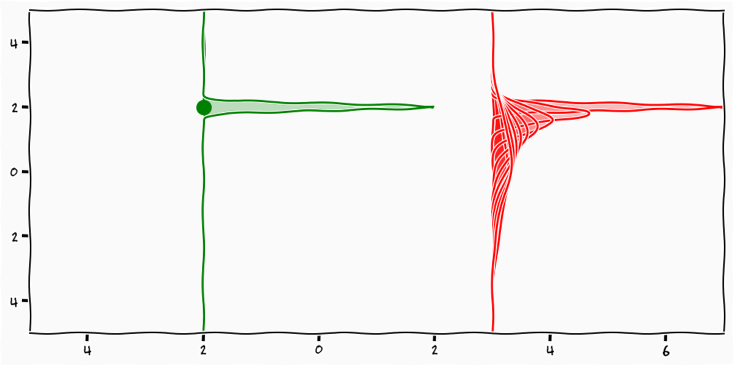
Non-parametric functions



Non-parametric functions



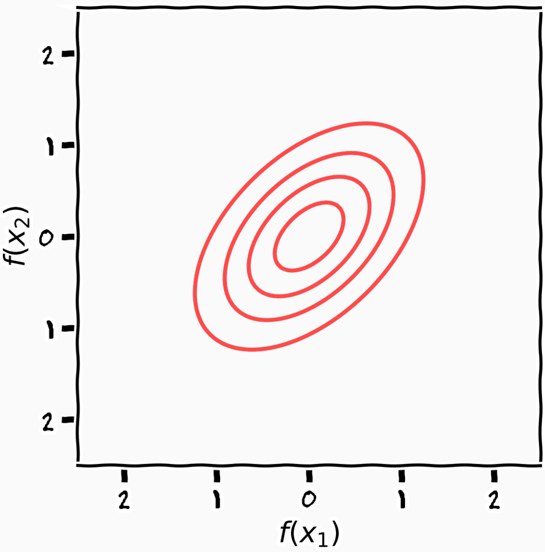
Non-parametric functions



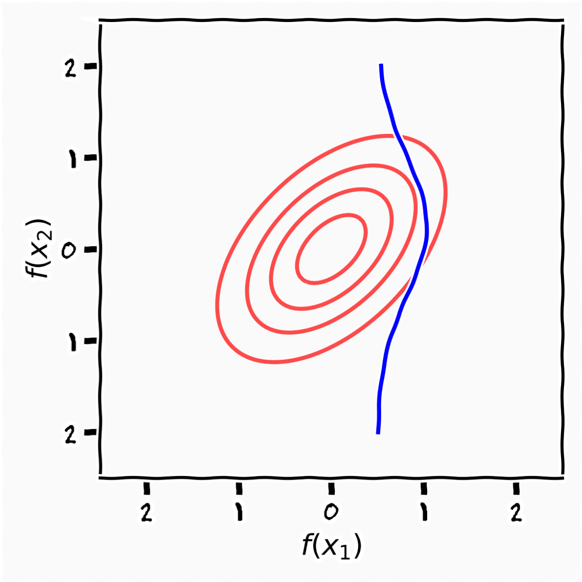
Conditional Gaussians

$$\mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}\right)$$

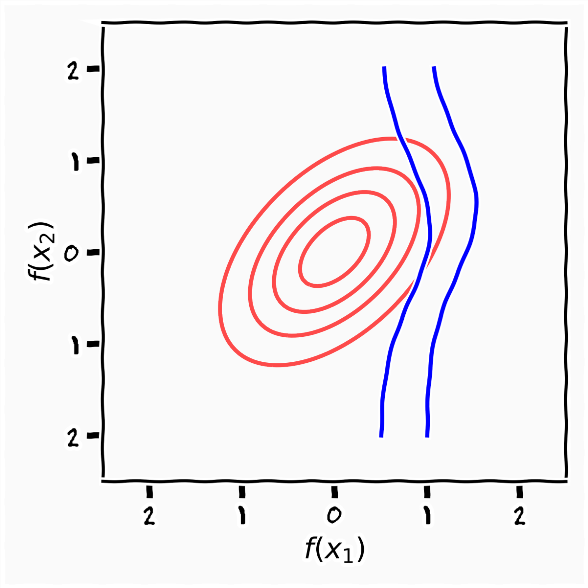
Conditional Gaussians



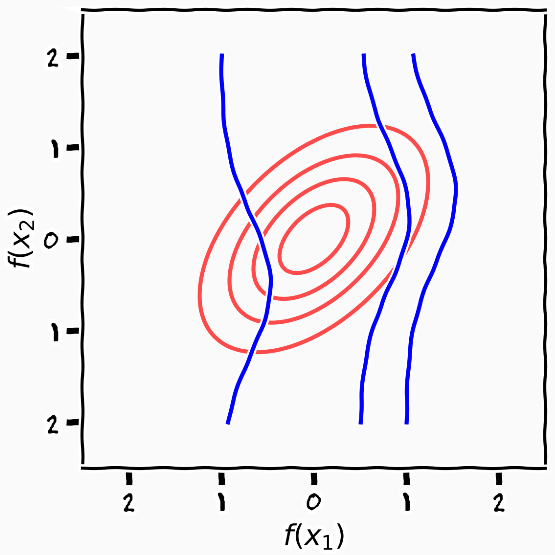
Conditional Gaussians



Conditional Gaussians



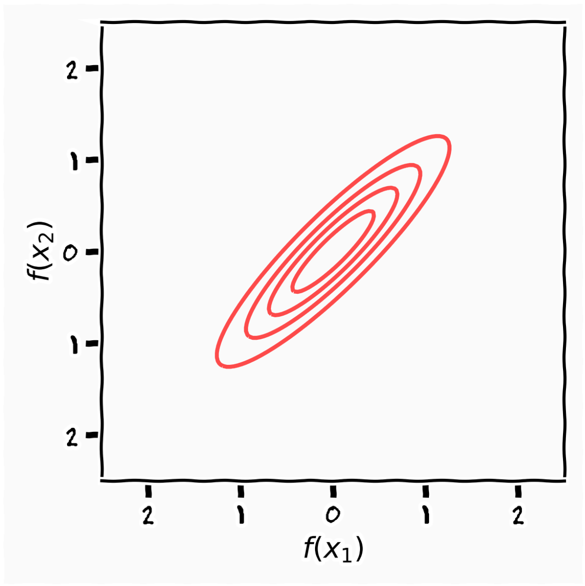
Conditional Gaussians

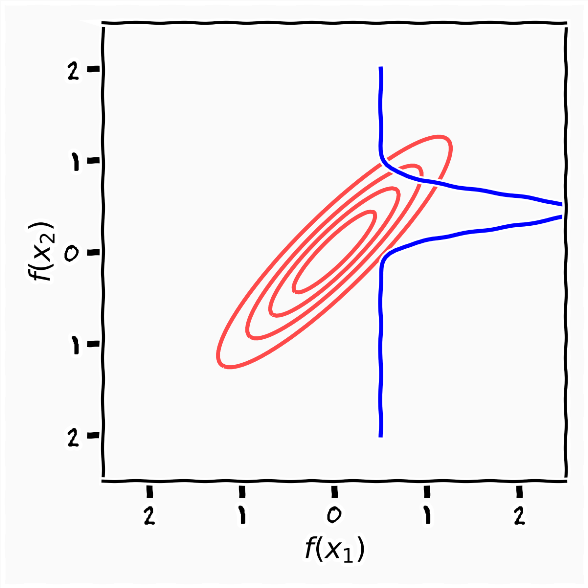


Conditional Gaussians

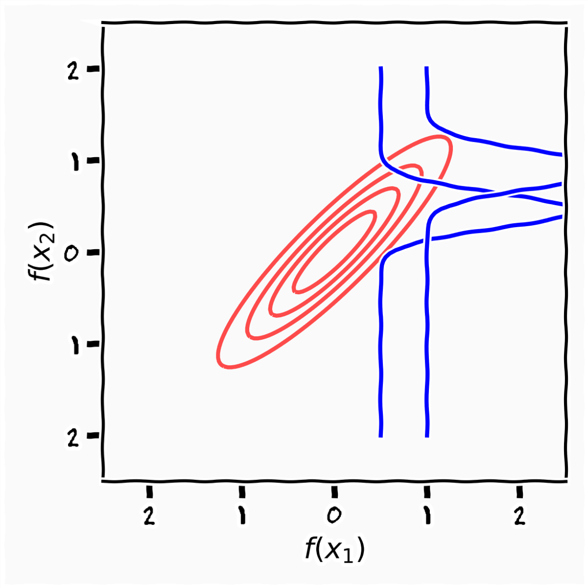
$$\mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}\right)$$

Conditional Gaussians

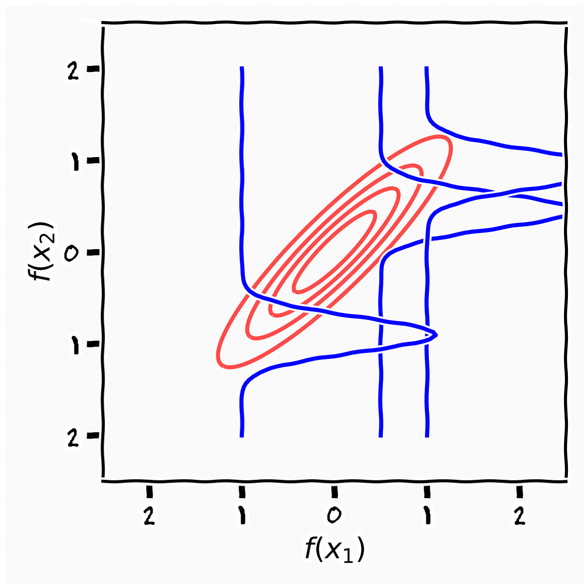



2
1
0
1
2
$f(x_2)$
2
1
0
1
2
$f(x_1)$

Conditional Gaussians



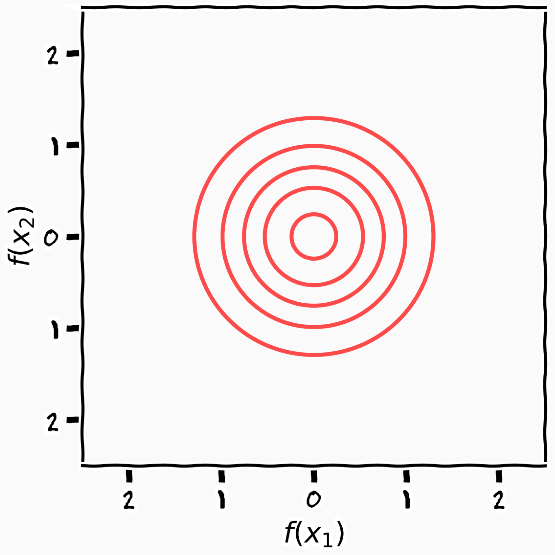
Conditional Gaussians



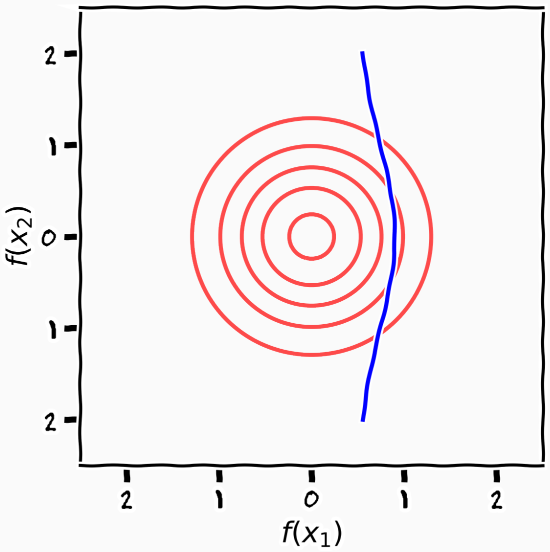
Conditional Gaussians

$$\mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$$

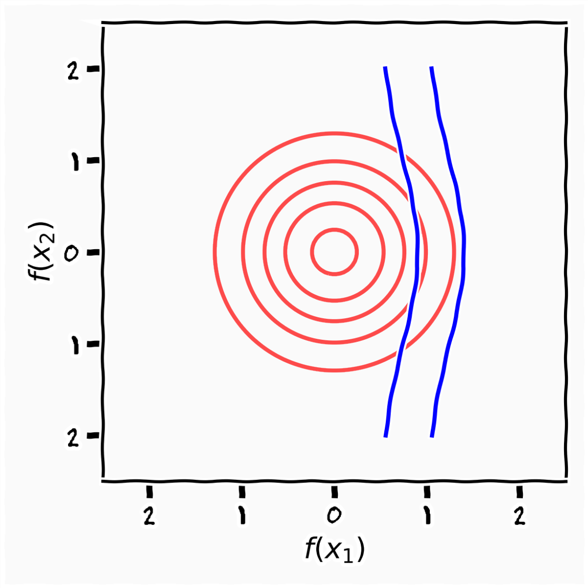
Conditional Gaussians



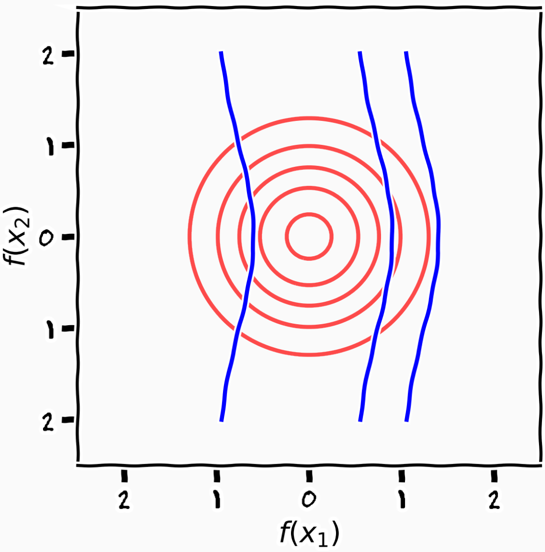
Conditional Gaussians



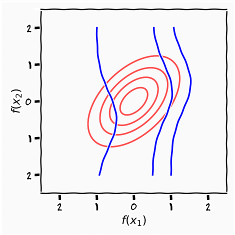
Conditional Gaussians



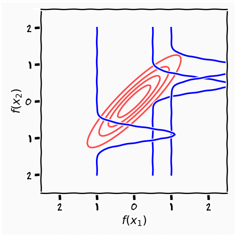
Conditional Gaussians



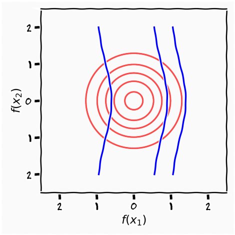
Conditional Gaussians



$$\mathrm{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}\right)$$

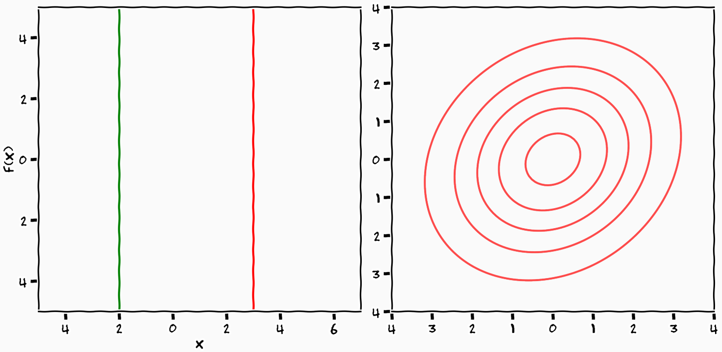


$$\mathrm{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}\right)$$

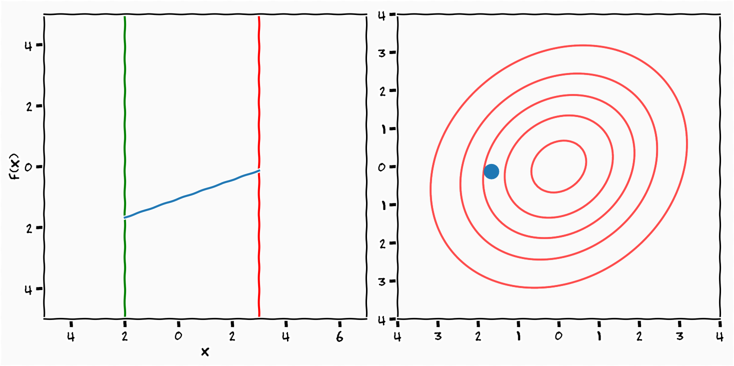


$$\mathrm{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$$

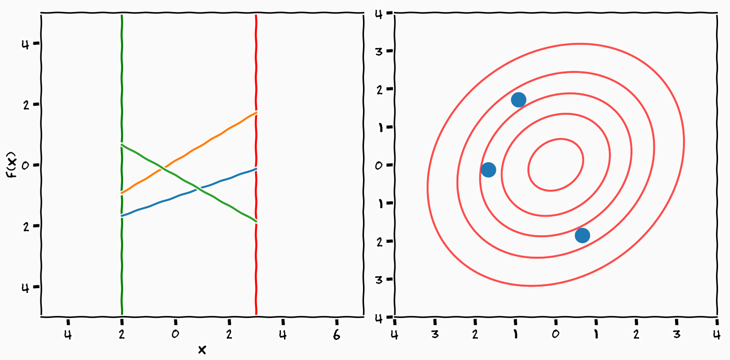
Gaussian Samples



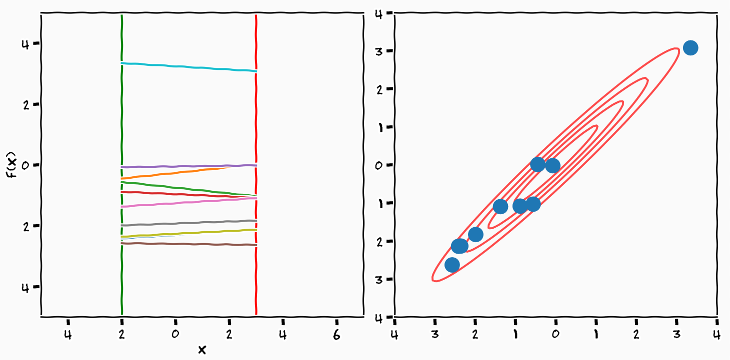
Gaussian Samples



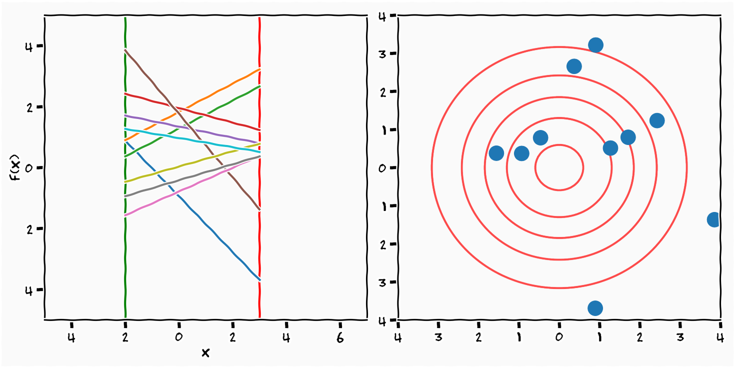
Gaussian Samples



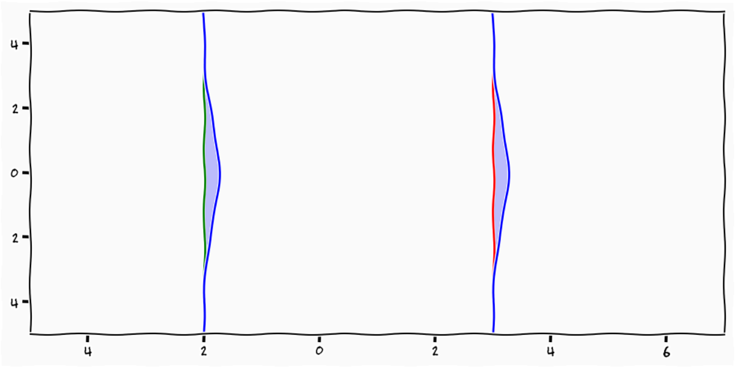
Gaussian Samples



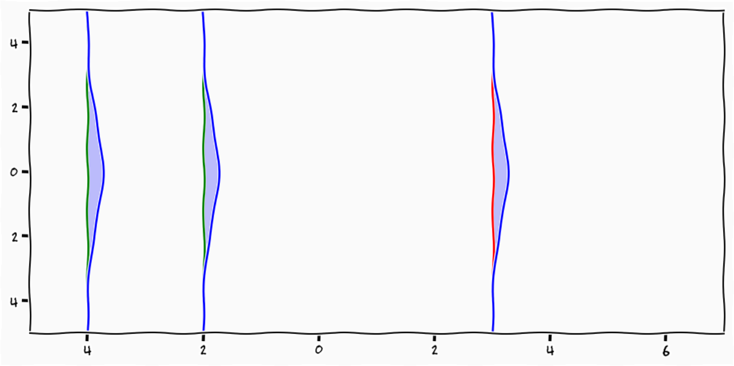
Gaussian Samples



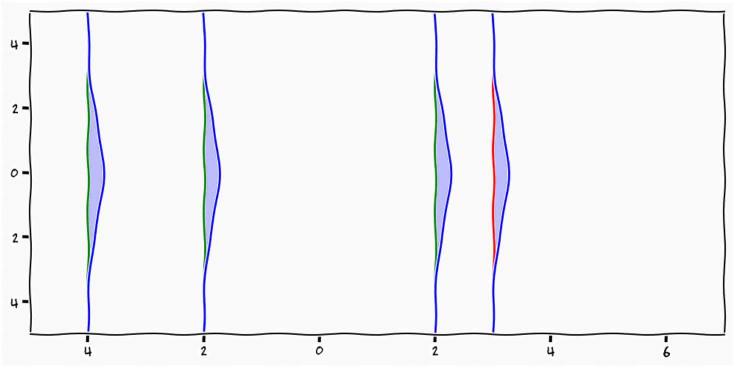
Lets talk about functions



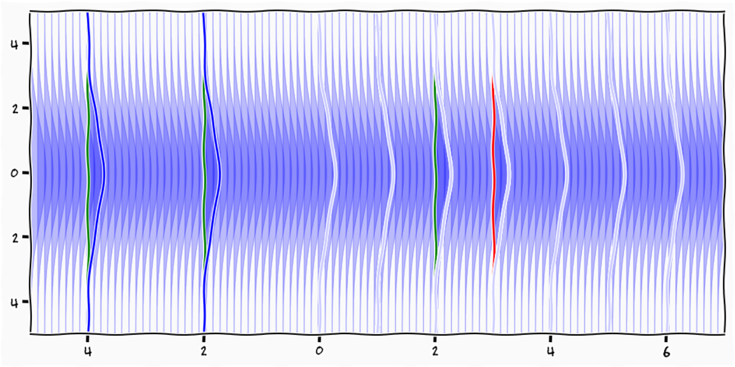
Non-parametric functions



Non-parametric functions



Non-parametric functions



Jointly Gaussian functions II

$$p(\mathbf{f}) = \mathcal{N}\left(\begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_N \end{bmatrix} \middle| \begin{bmatrix} \mu(x_1) \\ \mu(x_2) \\ \vdots \\ \mu(x_N) \end{bmatrix}, \begin{bmatrix} k_{11} & k_{12} & \dots & k_{1N} \\ k_{21} & k_{22} & \dots & k_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ k_{N1} & k_{N2} & \dots & k_{NN} \end{bmatrix} \right)$$

Gaussian Distribution - Marginal

$$p(x_1, x_2) = \mathcal{N}\left(\begin{array}{c|c} x_1 & \mu_1 \\ x_2 & \mu_2 \end{array} , \begin{array}{cc} k_{11} & k_{12} \\ k_{21} & k_{22} \end{array} \right)$$

Gaussian Distribution - Marginal

$$p(x_1, x_2) = \mathcal{N}\left(\begin{array}{c} x_1 \\ x_2 \end{array} \middle| \begin{array}{c} \mu_1 \\ \mu_2 \end{array}, \begin{array}{cc} k_{11} & k_{12} \\ k_{21} & k_{22} \end{array} \right)$$

$$\Rightarrow p(x_1) = \int_{x_2} p(x_1, x_2) = \underline{\mathcal{N}\left(x_1 \mid \mu_1, k_{11}\right)}$$

Gaussian Distribution - Marginal

$$p(x_1, x_2) = \mathcal{N}\left(\begin{array}{c|c} x_1 & \mu_1 \\ x_2 & \mu_2 \end{array}, \begin{array}{cc} k_{11} & k_{12} \\ k_{21} & k_{22} \end{array}\right)$$

$$\Rightarrow p(x_1) = \int_{x_2} p(x_1, x_2) = \underline{\mathcal{N}\left(x_1 \mid \mu_1, k_{11}\right)}$$

$$p(x_1, x_2, \ldots, x_N) = \mathcal{N}\left(\begin{array}{c|c} x_1 & \mu_1 \\ x_2 & \mu_2 \\ \vdots & \vdots \\ x_N & \mu_N \end{array}, \begin{array}{cccc} k_{11} & k_{12} & \cdots & k_{1N} \\ k_{21} & k_{22} & \cdots & k_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ k_{N1} & k_{N2} & \cdots & k_{NN} \end{array}\right)$$

Gaussian Distribution - Marginal

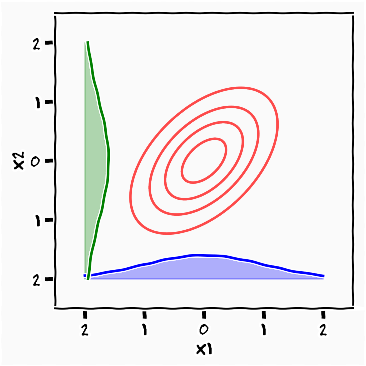
$$p(x_1, x_2) = \mathcal{N}\left(\begin{array}{c|c} x_1 & \mu_1 \\ x_2 & \mu_2 \end{array}, \begin{array}{cc} k_{11} & k_{12} \\ k_{21} & k_{22} \end{array} \right)$$

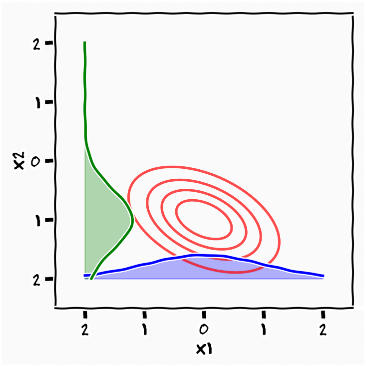
$$\Rightarrow p(x_1) = \int_{x_2} p(x_1, x_2) = \underline{\mathcal{N}\left(x_1 \mid \mu_1, k_{11}\right)}$$

$$p(x_1, x_2, \ldots, x_N) = \mathcal{N}\left(\begin{array}{c|c} x_1 & \mu_1 \\ x_2 & \mu_2 \\ \vdots & \vdots \\ x_N & \mu_N \end{array}, \begin{array}{cccc} k_{11} & k_{12} & \cdots & k_{1N} \\ k_{21} & k_{22} & \cdots & k_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ k_{N1} & k_{N2} & \cdots & k_{NN} \end{array} \right)$$

$$\Rightarrow \quad p(x_1) = \int_{x_2, \ldots, x_N} p(x_1, x_2, \ldots, x_N) = \underline{\mathcal{N}\left(x_1 \mid \mu_1, k_{11}\right)}$$

Gaussian Distribution - Marginal



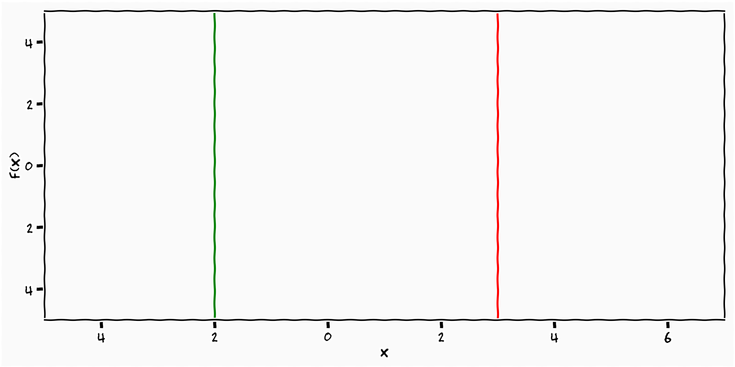


Marginal Property (Consistency)

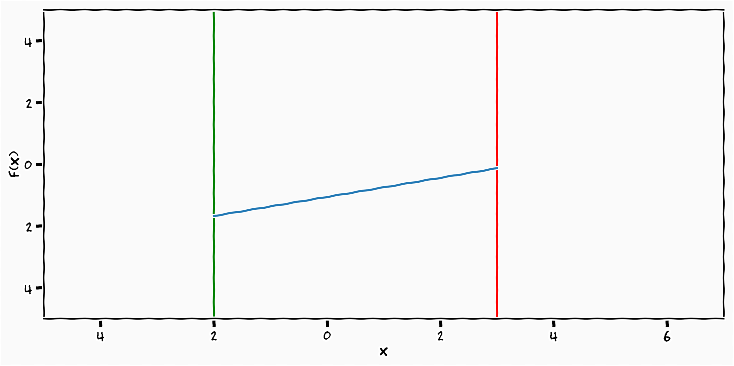
For all measurable sets $F_i \subseteq \mathbb{R}^n$ and probability measure $\mathcal{N}$

$$\mathcal{N}_{t_1 \cdot t_k}\left(F_1 \times \cdot \times F_k\right) = \mathcal{N}_{t_1 \cdots t_k, t_{k+1} \cdot t_{k+m}}\left(F_1 \times \cdot \times F_k \times \mathbb{R}^n \times \cdot \times \mathbb{R}^n\right)$$

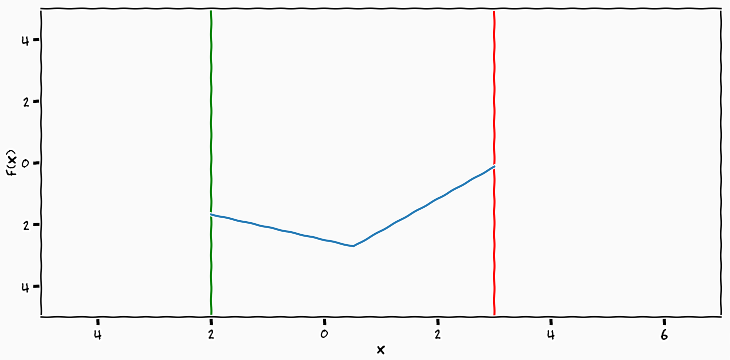
Gaussian Samples



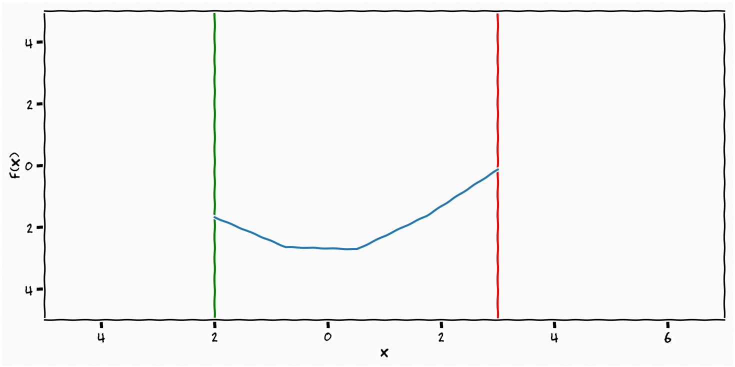
Gaussian Samples



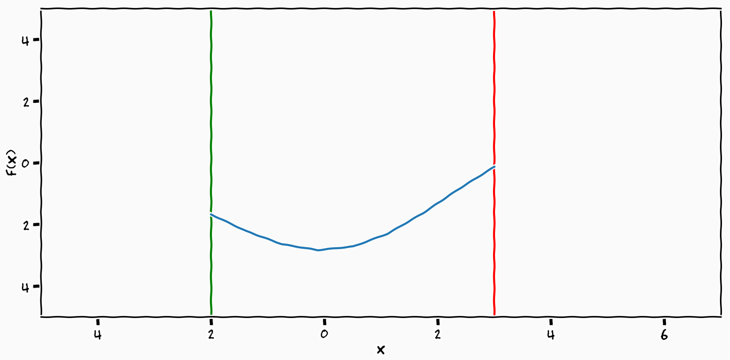
Gaussian Samples



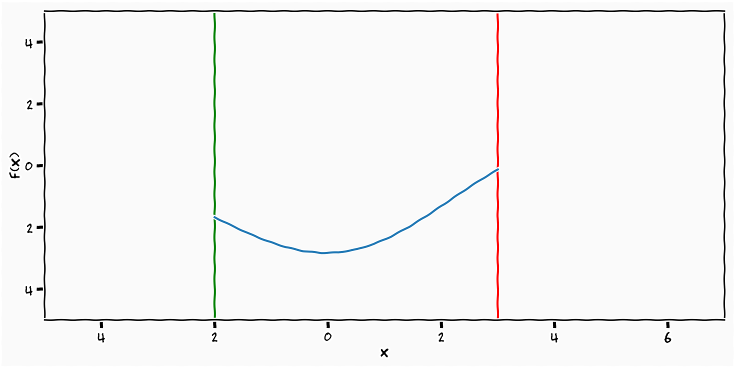
Gaussian Samples



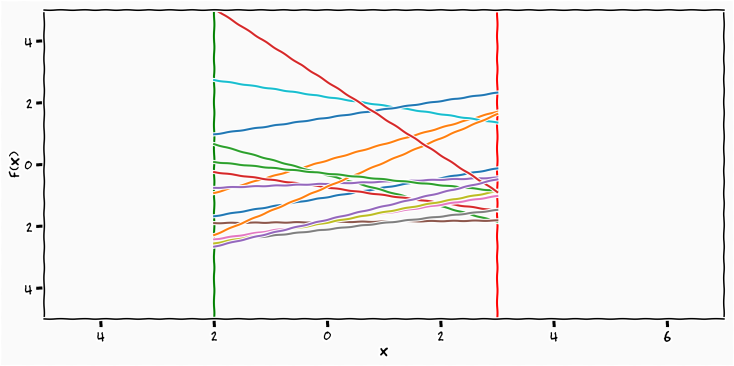
Gaussian Samples



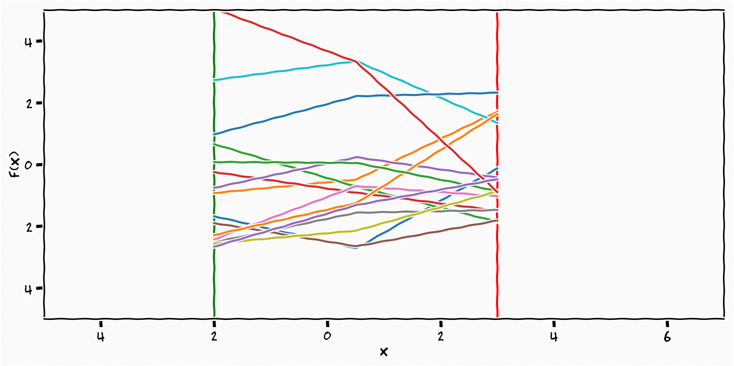
Gaussian Samples



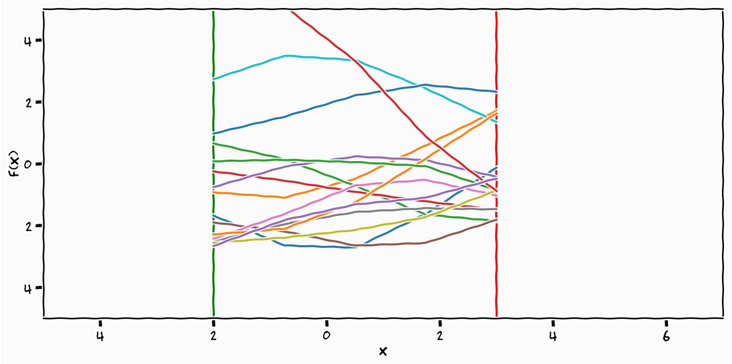
Gaussian Samples



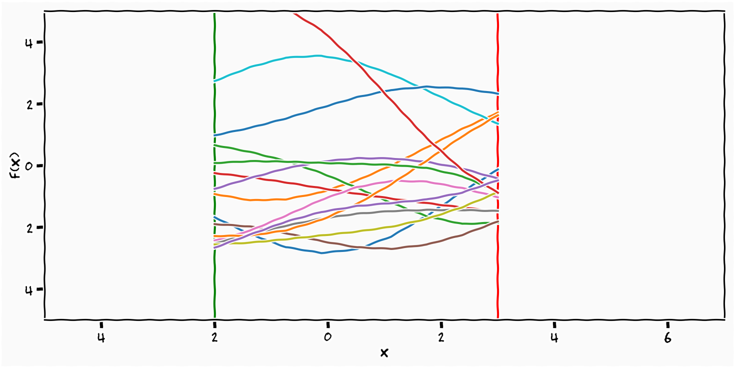
Gaussian Samples



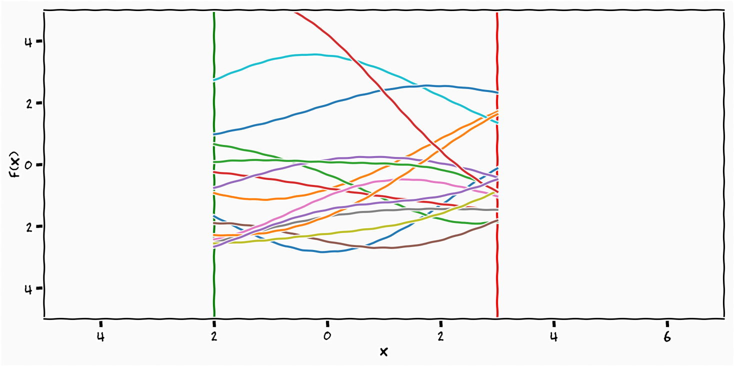
Gaussian Samples

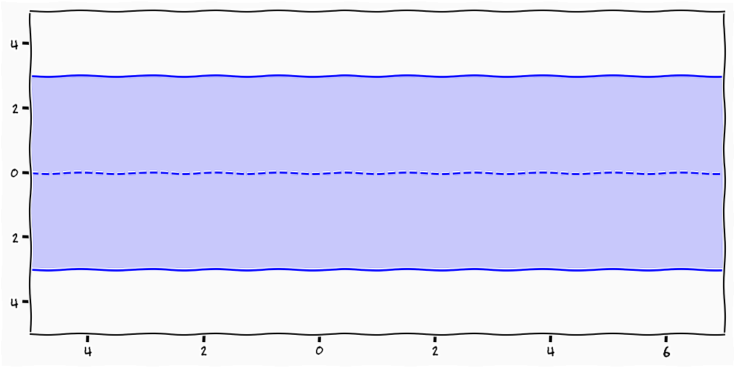


Gaussian Samples



Gaussian Samples




4
2
0
2
4
4
2
0
2
4
6

Gaussian Processes: Formalism

$$p(\mathbf{f}) = \mathcal{N}\left(\begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_N \\ \vdots \end{bmatrix} \middle| \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_N \\ \vdots \end{bmatrix}, \begin{bmatrix} k_{11} & k_{12} & \dots & k_{1N} & \dots \\ k_{21} & k_{22} & \dots & k_{2N} & \dots \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ k_{N1} & k_{N2} & \dots & k_{NN} & \dots \\ \vdots & \vdots & \dots & \vdots & \ddots \end{bmatrix} \right)$$

$$\mathcal{GP}(\cdot,\cdot) \qquad M \in \mathbb{R}^{\infty \times N} \qquad \mathcal{N}(\cdot,\cdot)$$

$$\infty \qquad \rightarrow \qquad N$$

The Gaussian distribution is the projection of the infinite Gaussian process

Gaussian Process

Definition (Gaussian Process)
A Gaussian process is a collection of random variables who are jointly Gaussian distributed index by a infinite index set

Gaussian Processes: Formalism II

$$p(\mathbf{f}) = \mathcal{N}\left(\begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_N \\ \vdots \end{bmatrix} \middle| \begin{bmatrix} \mu(x_1) \\ \mu(x_2) \\ \vdots \\ \mu(x_N) \\ \vdots \end{bmatrix}, \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & \dots & k(x_1, x_N) & \dots \\ k(x_2, x_1) & k(x_2, x_2) & \dots & k(x_2, x_N) & \dots \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ k(x_N, x_1) & k(x_N, x_2) & \dots & k(x_N, x_N) & \dots \\ \vdots & \vdots & \dots & \vdots & \ddots \end{bmatrix}\right)$$

"Parametrisation"

$$k_{ij} = k(x_i, x_j)$$

- We parameterise the covariance as a function of the input

"Parametrisation"

$$k_{ij} = k(x_i, x_j)$$

- We parameterise the covariance as a function of the input
- the index set of the measure is the uncountable infinity

"Parametrisation"

$$k_{ij} = k(x_i, x_j)$$

- We parameterise the covariance as a function of the input
- the index set of the measure is the uncountable infinity
- Your "handle" to input your knowledge into a GP is the covariance function

"Parametrisation"

$$k_{ij} = k(x_i, x_j)$$

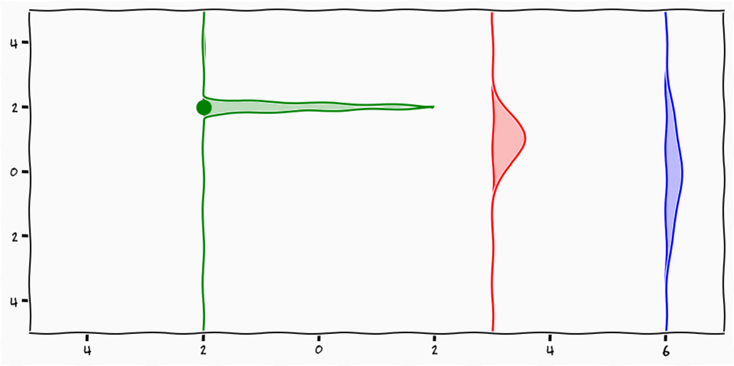
- We parameterise the covariance as a function of the input
- the index set of the measure is the uncountable infinity
- Your "handle" to input your knowledge into a GP is the covariance function
    - *you specify the degree of covariance between data-points*

"Parametrisation"

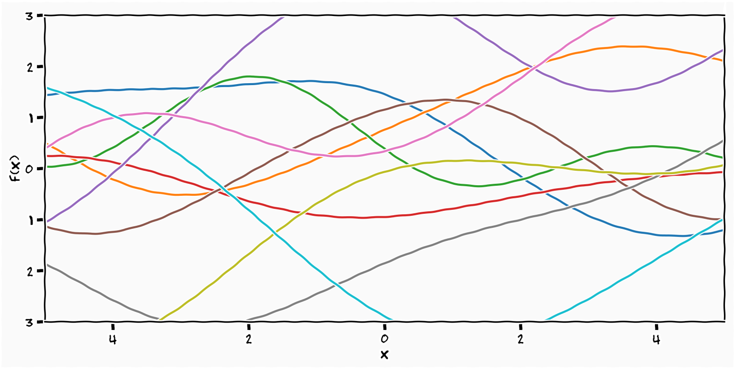
$$k_{ij} = k(x_i, x_j)$$

- We parameterise the covariance as a function of the input
- the index set of the measure is the uncountable infinity
- Your "handle" to input your knowledge into a GP is the covariance function
  - *you specify the degree of covariance between data-points*
- If this "parametrisation" aligns well with your knowledge a GP is the way forward!

Gaussian Processes

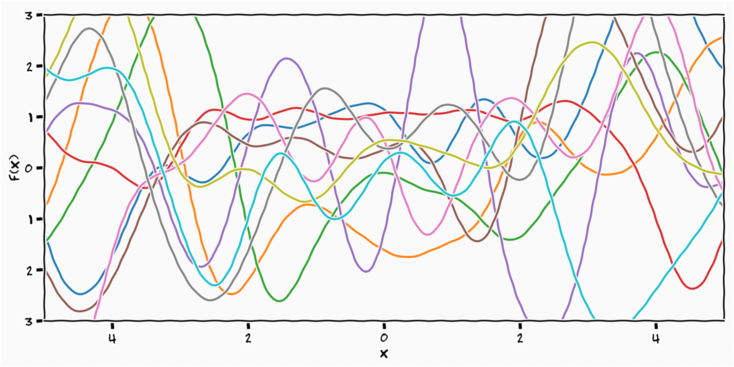


Gaussian Processes Samples



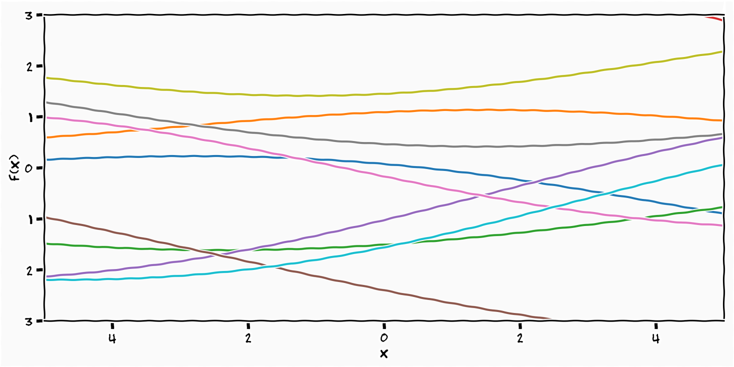
$$k(x_i, x_j) = 3 \cdot e^{-\frac{(x_i - x_j)^2}{15}}$$

Gaussian Processes Samples

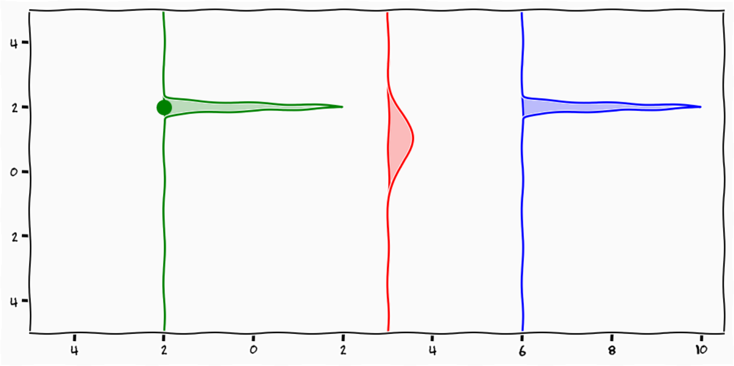


$$k(x_i, x_j) = 3 \cdot e^{-\frac{(x_i - x_j)^2}{1}}$$

Gaussian Processes Samples



$$k(x_i, x_j) = 3 \cdot e^{-\frac{(x_i - x_j)^2}{150}}$$


4
2
0
2
4
4
2
0
2
4
6
8
10

Bayesian Inference

Bayes' Rule

$$p(\mathbf{f}_* \mid \mathbf{f}) = \frac{p(\mathbf{f}, \mathbf{f}_*)}{p(\mathbf{f})} = \frac{p(\mathbf{f}, \mathbf{f}_*)}{\int p(\mathbf{f}, \mathbf{f}_*) \mathrm{d}\mathbf{f}_*}$$

Marginal Likelihood

$$\int p(\mathbf{f}, \mathbf{f}_*) \mathrm{d}\mathbf{f}_* = \int p(\mathbf{f} \mid \mathbf{f}_*) p(\mathbf{f}_*) \mathrm{d}\mathbf{f}_*$$

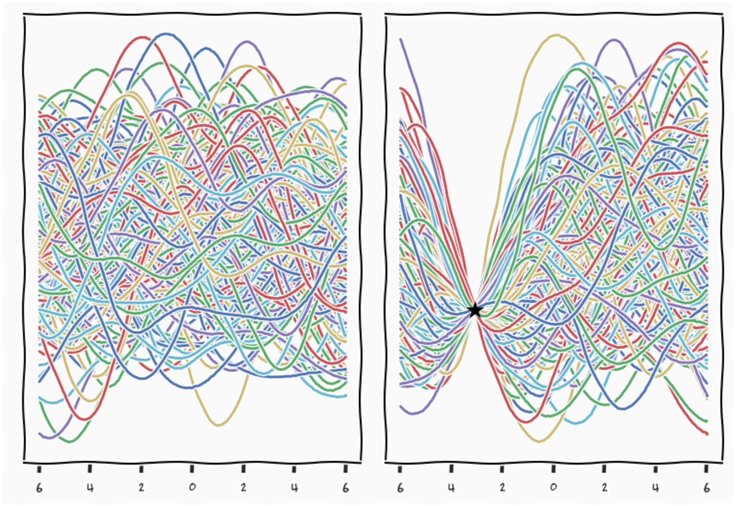
- Take every possible function value/marginal $\mathbf{f}_*$ at location $\mathbf{x}_*$ according to their probability

Marginal Likelihood

$$\int p(\mathbf{f}, \mathbf{f}_*) \mathrm{d}\mathbf{f}_* = \int p(\mathbf{f} \mid \mathbf{f}_*) p(\mathbf{f}_*) \mathrm{d}\mathbf{f}_*$$

- Take every possible function value/marginal $\mathbf{f}_*$ at location $\mathbf{x}_*$ according to their probability
- Check if these marginals are consistent with the marginals we observe $\mathbf{f}$ at location $\mathbf{x}$

Gaussian Processes: Posterior Samples



Gaussian Process: "Predictive Posterior"

$$p(\mathbf{f}, \mathbf{f}_*) = p(\mathbf{f}_* \mid \mathbf{f})p(\mathbf{f})$$

- We have defined $p(\mathbf{f}, \mathbf{f}_*)$

Gaussian Process: "Predictive Posterior"

$$p(\mathbf{f}, \mathbf{f}_*) = p(\mathbf{f}_* \mid \mathbf{f})p(\mathbf{f})$$

- We have defined $p(\mathbf{f}, \mathbf{f}_*)$
- We know through the marginal property of the Gaussian that $p(\mathbf{f})$ is consistent

Gaussian Process: "Predictive Posterior"

$$p(\mathbf{f}, \mathbf{f}_*) = p(\mathbf{f}_* \mid \mathbf{f})p(\mathbf{f})$$

- We have defined $p(\mathbf{f}, \mathbf{f}_*)$
- We know through the marginal property of the Gaussian that $p(\mathbf{f})$ is consistent
- We know that $p(\mathbf{f}_* \mid \mathbf{f})$ is Gaussian

Gaussian Process: "Predictive Posterior"

$$p(\mathbf{f}, \mathbf{f}_*) = p(\mathbf{f}_* \mid \mathbf{f})p(\mathbf{f})$$

- We have defined $p(\mathbf{f}, \mathbf{f}_*)$
- We know through the marginal property of the Gaussian that $p(\mathbf{f})$ is consistent
- We know that $p(\mathbf{f}_* \mid \mathbf{f})$ is Gaussian
- $\Rightarrow$ *We can just solve for* $p(\mathbf{f}_* \mid \mathbf{f})$

Gaussian Process: "Predictive Posterior"

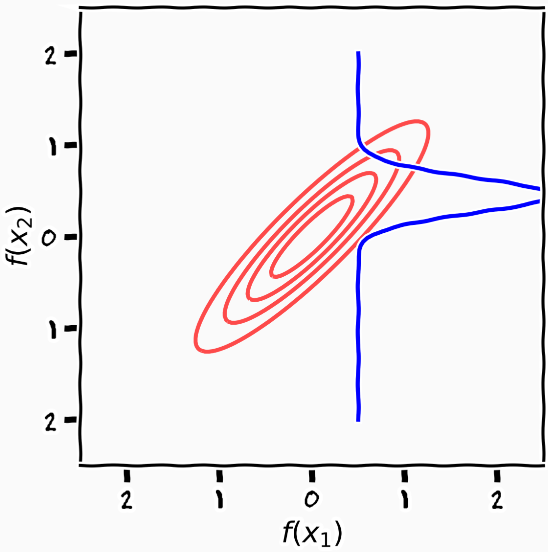
- All instantiations are jointly Gaussian

$$\left[\begin{array}{c}\mathbf{f}\\ \mathbf{f}_*\end{array}\right] \sim \mathcal{N}\left(\left[\begin{array}{c}\mathbf{0}\\ \mathbf{0}\end{array}\right], \left[\begin{array}{cc}k(\mathbf{x},\mathbf{x}) & k(\mathbf{x},\mathbf{x}_*)\\ k(\mathbf{x}_*,\mathbf{x}) & k(\mathbf{x}_*,\mathbf{x}_*)\end{array}\right]\right)$$

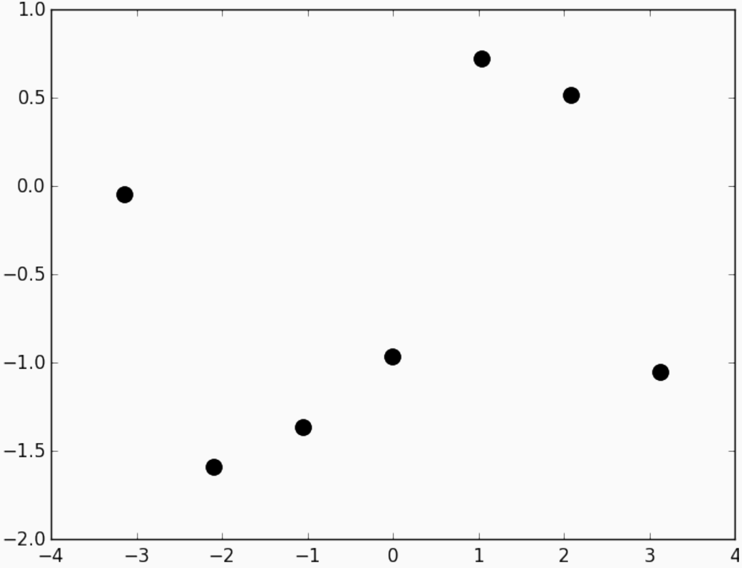
- Conditional Gaussian

$$\begin{aligned} p(f_*|\mathbf{x}_*,\mathbf{x},\mathbf{f}) = \mathcal{N}(&k(\mathbf{x}_*,\mathbf{x})^{\mathrm{T}}k(\mathbf{x},\mathbf{x})^{-1}\mathbf{f},\\ &k(\mathbf{x}_*,\mathbf{x}_*) - k(\mathbf{x}_*,\mathbf{x})^{\mathrm{T}}k(\mathbf{x},\mathbf{x})^{-1}k(\mathbf{x},\mathbf{x}_*)) \end{aligned}$$

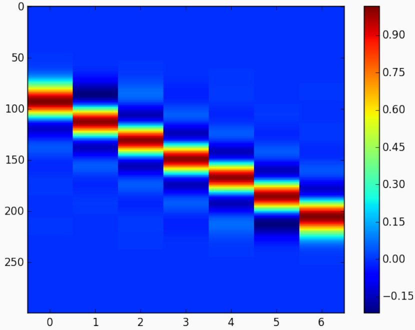
Conditional Gaussians



Gaussian Process: "Predictive Posterior"

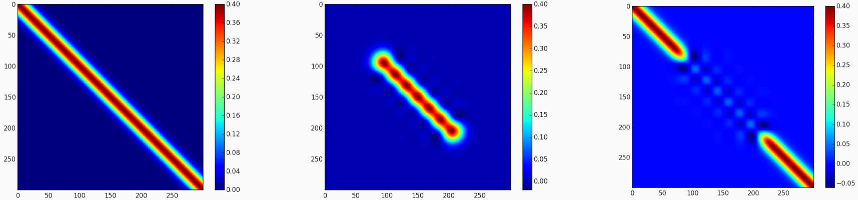


Does it make sense: Mean



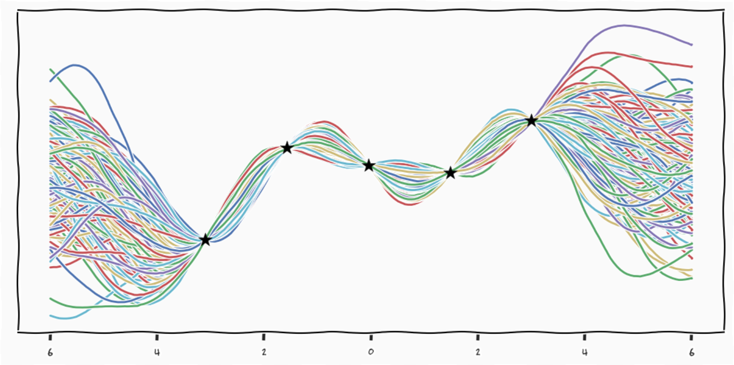
$$k(\mathbf{x}_*, \mathbf{X})^{\mathrm{T}} k(\mathbf{X}, \mathbf{X})^{-1} \mathbf{f}$$

Does it make sense: Covariance

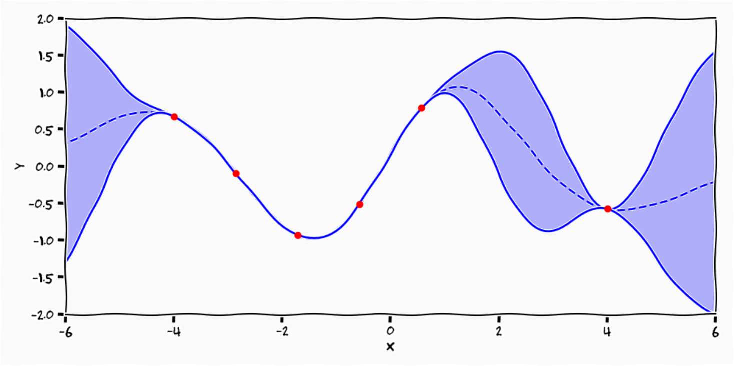


$$k(\mathbf{x}_*, \mathbf{x}_*) - k(\mathbf{x}_*, \mathbf{x})^{\mathrm{T}} k(\mathbf{x}, \mathbf{x})^{-1} k(\mathbf{x}, \mathbf{x}_*)$$

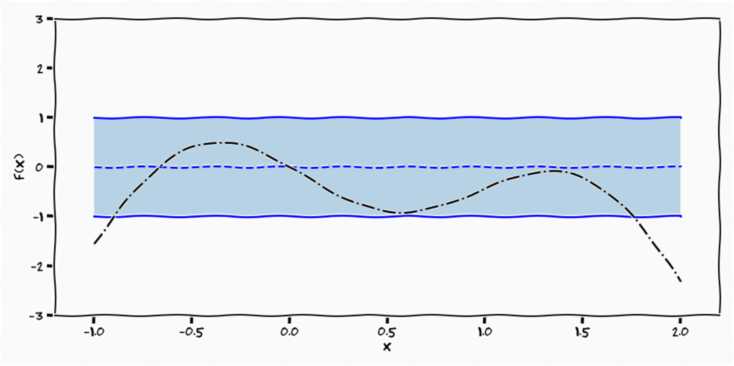
Gaussian Processes: "Predictive Posterior Samples"



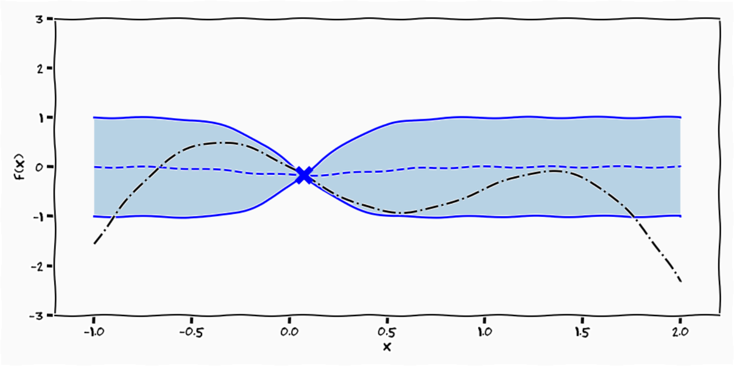
Gaussian Processes: "Predictive Posterior Process"



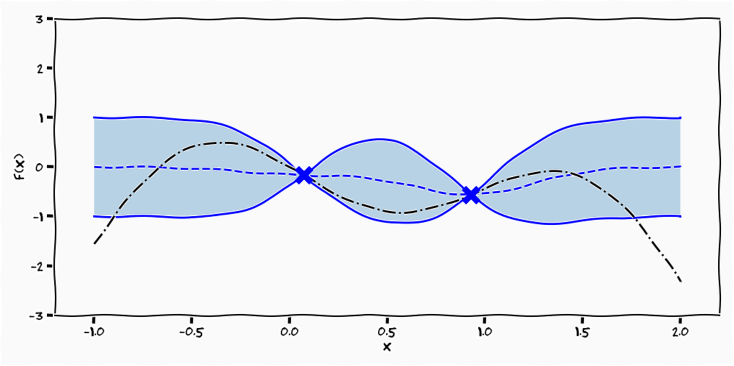
Posterior Processes



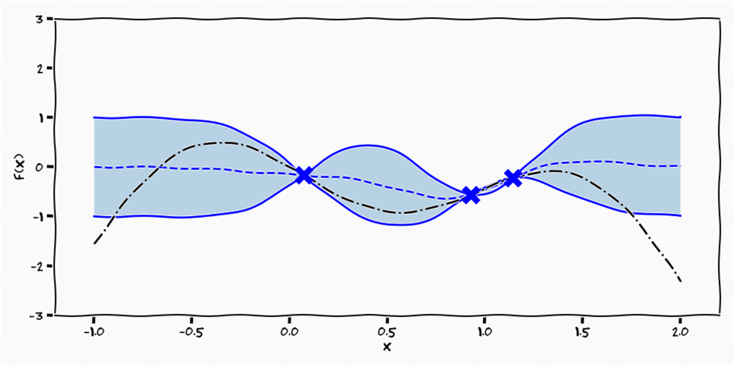
Posterior Processes



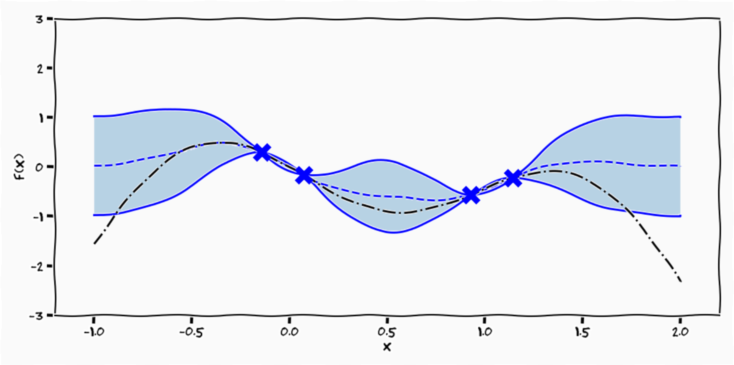
Posterior Processes



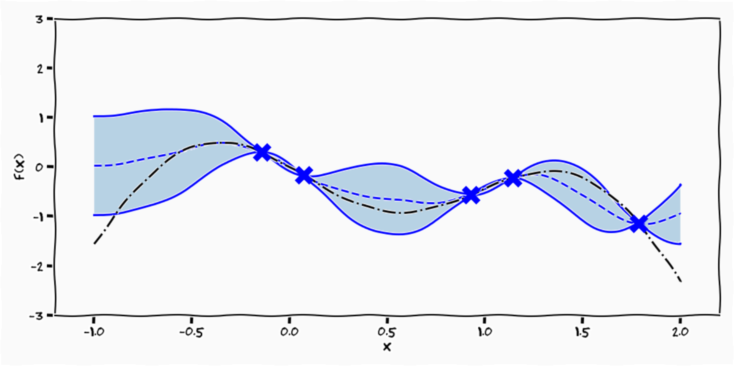
Posterior Processes



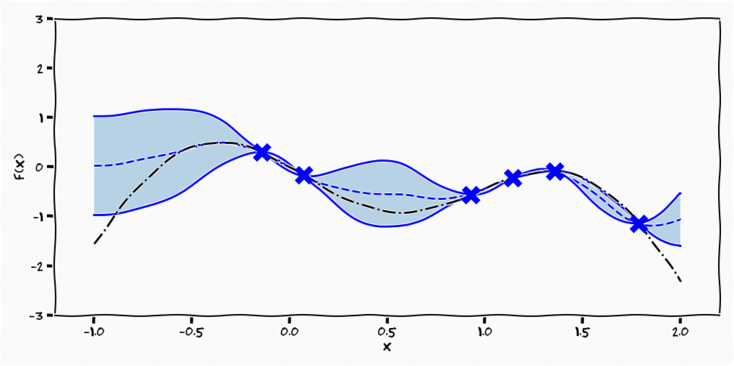
Posterior Processes



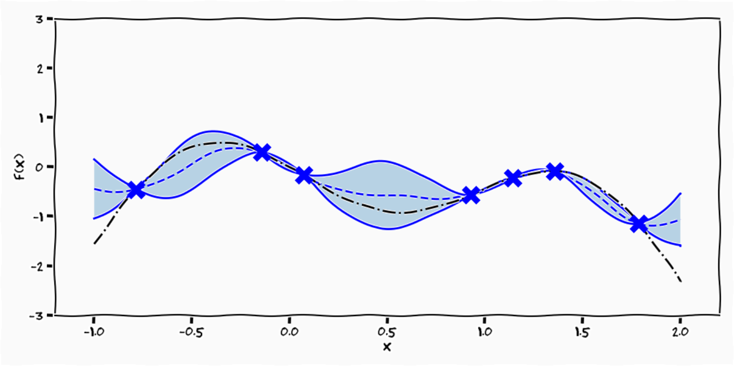
Posterior Processes



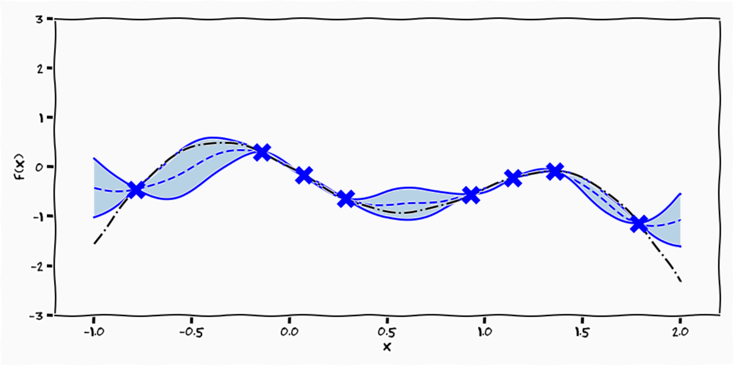
Posterior Processes



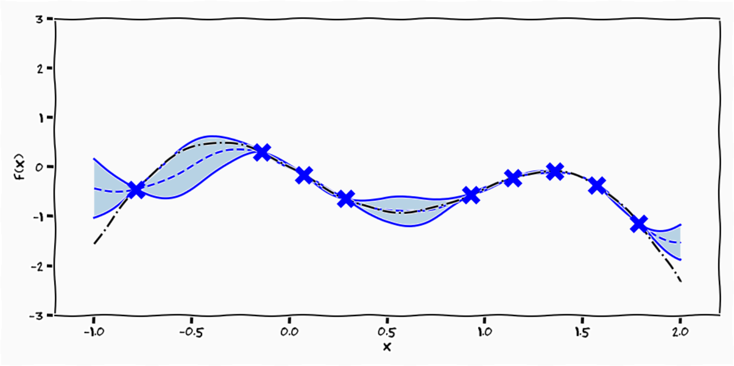
Posterior Processes



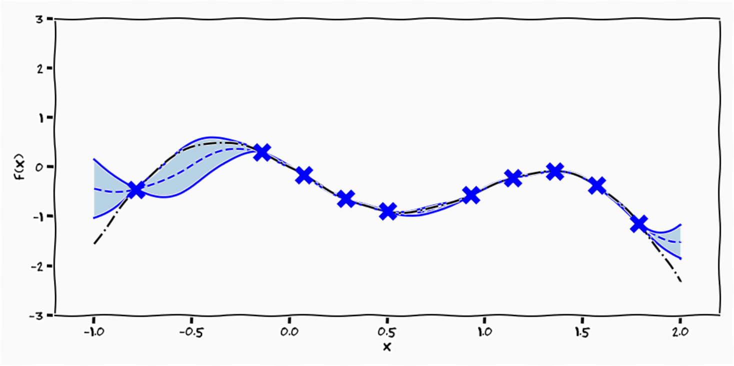
Posterior Processes



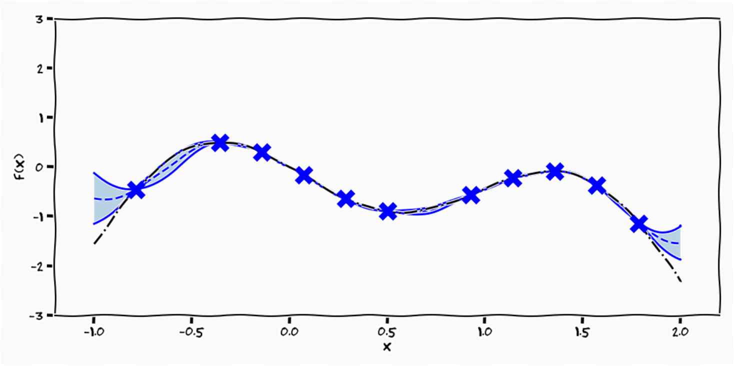
Posterior Processes



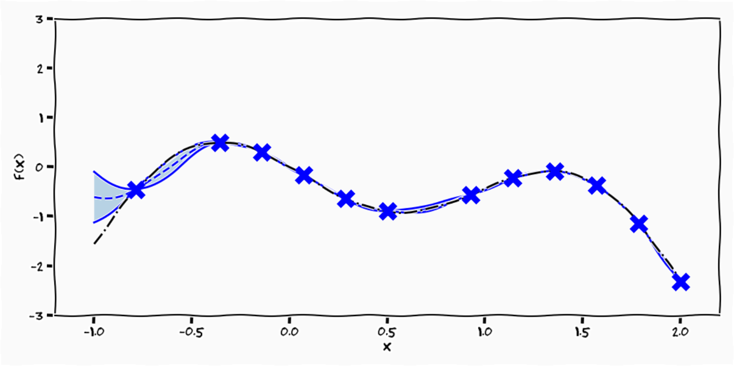
Posterior Processes



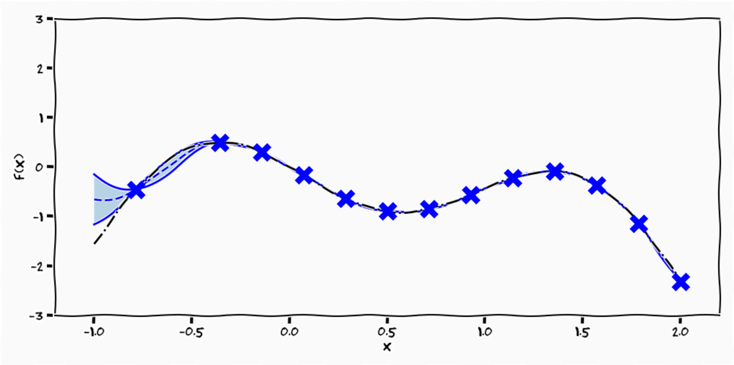
Posterior Processes



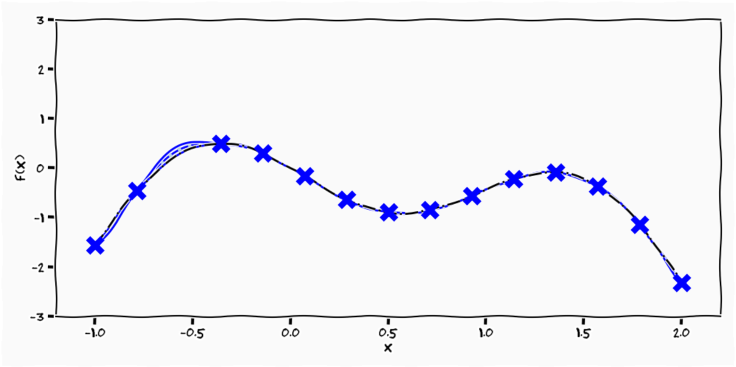
Posterior Processes



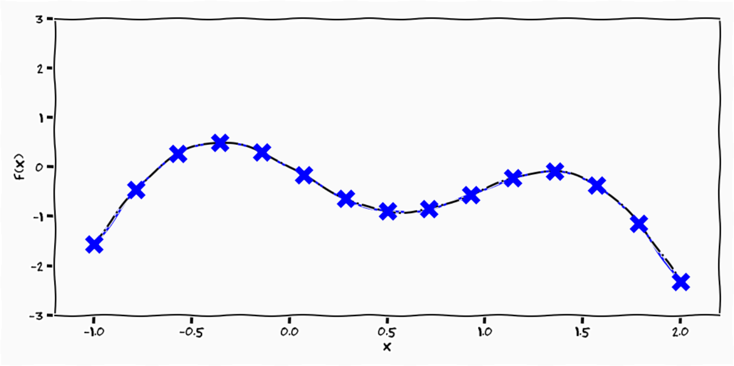
Posterior Processes



Posterior Processes



Posterior Processes



Model

- So far we have only looked at the prior
  - the same as when we sampled from $p(\mathbf{w})$ in the previous lecture
  - the "predictive posterior" has really been the "conditional prior"
- Now lets introduce a likelihood

Model

$$p(\mathbf{y},\mathbf{f}) = p(\mathbf{y} \mid \mathbf{f})p(\mathbf{f}) = p(\mathbf{f}) \prod_{i=1} \mathcal{N}(p(y_i \mid f_i))$$

$$p(\mathbf{y} \mid \mathbf{f}) = \mathcal{N}(\mathbf{y} \mid \mathbf{f}, \sigma^2 \mathbf{I})$$

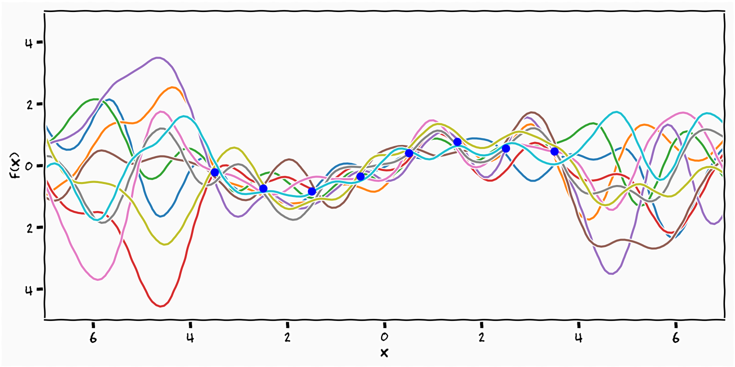
- Same motivation as with linear regression
- we want to explain away our ignorance from the data using σ

Predictive Posterior Process

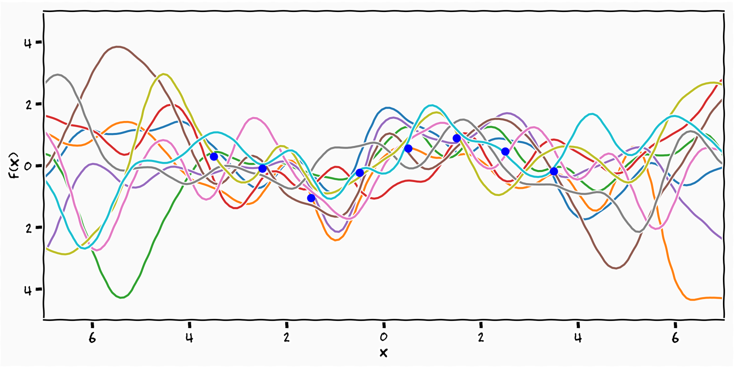
$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} k(\mathbf{x}, \mathbf{x}) + \sigma^2 \mathbf{I} & k(\mathbf{x}, \mathbf{x}_*) \\ k(\mathbf{x}_*, \mathbf{x}) & k(\mathbf{x}_*, \mathbf{x}_*) \end{bmatrix} \right)$$

$$\begin{aligned} p(f_* | \mathbf{x}_*, \mathbf{x}, \mathbf{y}, \boldsymbol{\theta}) = \mathcal{N}(& k(\mathbf{x}_*, \mathbf{x})^{\mathrm{T}} (K(\mathbf{x}, \mathbf{x}) + \sigma^2 \mathbf{I}))^{-1} \mathbf{y}, \\ & k(\mathbf{x}_*, \mathbf{x}_*) - k(\mathbf{x}_*, \mathbf{x})^{\mathrm{T}} (K(\mathbf{x}, \mathbf{x}) + \sigma^2 \mathbf{I})^{-1} K(\mathbf{x}, \mathbf{x}_*)) \end{aligned}$$

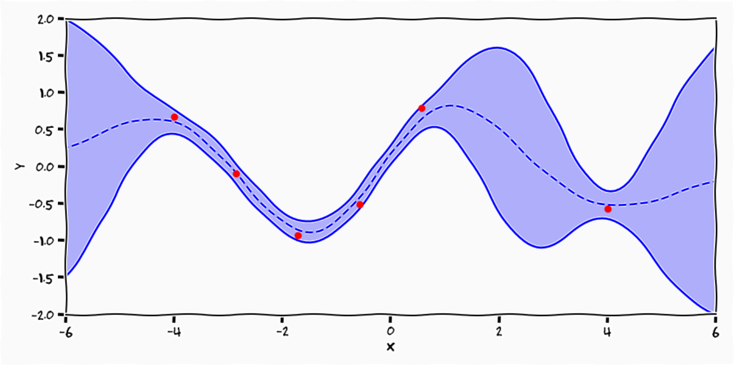
Posterior Samples



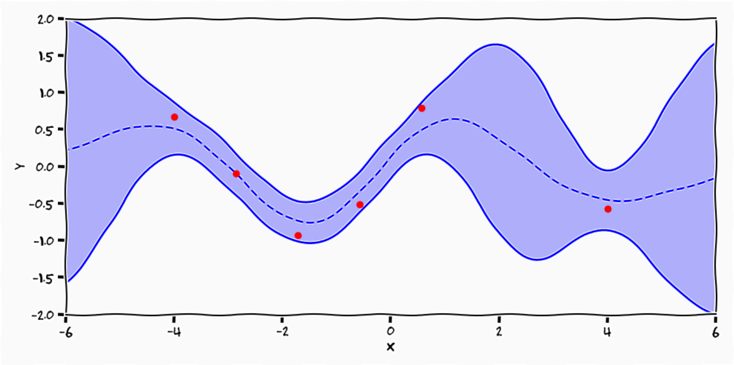
Posterior Samples



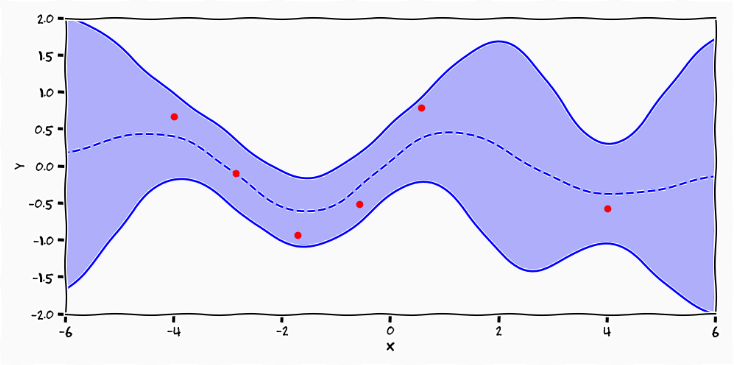
Gaussian Processes: Predictive Posterior Process



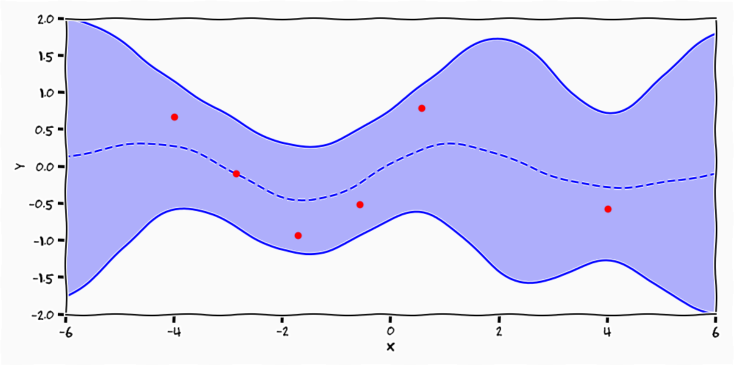
Gaussian Processes: Predictive Posterior Process



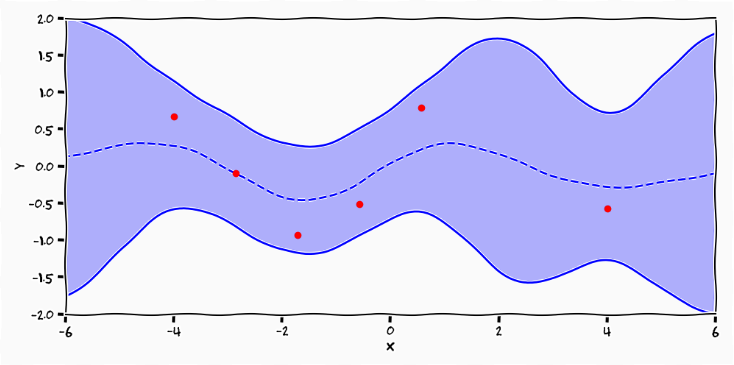
Gaussian Processes: Predictive Posterior Process



Gaussian Processes: Predictive Posterior Process



Gaussian Processes: Predictive Posterior Process



Predictive Posteriors

- Gaussian Process - *Non-parametric formulation*

$$
\begin{aligned}
p(y_* \mid \mathbf{y}, \mathbf{x}_*, \mathbf{X}, \theta) = \mathcal{N}(y_* \mid \mu_*, K_*) \\
\mu_* &= k(x_*, x)k(x, x)^{-1}\mathbf{y} \\
K_* &= k(x_*, x_*) - k(x_*, x)k(x, x)^{-1}k(x, x_*)
\end{aligned}
$$

Predictive Posteriors

- Gaussian Process - *Non-parametric formulation*

$$
\begin{aligned}
p(y_* \mid \mathbf{y}, \mathbf{x}_*, \mathbf{X}, \theta) &= \mathcal{N}(y_* \mid \mu_*, K_*) \\
\mu_* &= k(x_*, x)k(x, x)^{-1}\mathbf{y} \\
K_* &= k(x_*, x_*) - k(x_*, x)k(x, x)^{-1}k(x, x_*)
\end{aligned}
$$

- Linear Regression - *Parametric formulation*

$$
\begin{aligned}
p(y_*|\mathbf{y}, \mathbf{x}_*, \mathbf{X}, \alpha, \beta) &= \int p(y_*|\mathbf{x}_*, \mathbf{w}, \beta)p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \alpha, \beta)\mathrm{d}\mathbf{w} \\
&= \mathcal{N}\left((y_* \mid \mu_*(x_*), \Sigma_*(x_*)\right) \\
\mu_* &= \left(\beta(\alpha\mathbf{I} + \beta\mathbf{x}^{\mathrm{T}}\mathbf{x})^{-1}\mathbf{x}\mathbf{y}\right)^{\mathrm{T}}\mathbf{x}_* \\
&= \mathbf{x}_*^{\mathrm{T}}\mathbf{x}\left(\beta(\alpha\mathbf{I} + \beta\mathbf{x}^{\mathrm{T}}\mathbf{x})\right)^{-1}\mathbf{y} \\
\Sigma_* &= \frac{1}{\beta} + \mathbf{x}_*^{\mathrm{T}}(\alpha\mathbf{I} + \beta\mathbf{x}^{\mathrm{T}}\mathbf{x})^{-1}\mathbf{x}
\end{aligned}
$$

Regression Models: Mean

$$k(x_*, x)k(x, x)^{-1}\mathbf{y} \qquad \mathbf{x}_*^{\mathrm{T}}\mathbf{x}\left(\beta(\alpha\mathbf{I} + \beta\mathbf{x}^{\mathrm{T}}\mathbf{x})\right)^{-1}\mathbf{y}$$

- both means are linear with respect to the output data

Regression Models: Mean

$$k(x_*, x)k(x, x)^{-1}\mathbf{y} \qquad\qquad \mathbf{x}_*^{\mathrm{T}}\mathbf{x}\left(\beta(\alpha\mathbf{I} + \beta\mathbf{x}^{\mathrm{T}}\mathbf{x})\right)^{-1}\mathbf{y}$$

- both means are linear with respect to the output data
- how about if we take a basis function approach such that $\mathbf{x}_*^{\mathrm{T}}\mathbf{x} = \Phi(\mathbf{x}_*)^{\mathrm{T}}\Phi(\mathbf{x})$

Regression Models: Mean

$$k(x_*, x)k(x, x)^{-1}\mathbf{y} \qquad \mathbf{x}_*^{\mathrm{T}}\mathbf{x}\left(\beta(\alpha\mathbf{I} + \beta\mathbf{x}^{\mathrm{T}}\mathbf{x})\right)^{-1}\mathbf{y}$$

- both means are linear with respect to the output data
- how about if we take a basis function approach such that $\mathbf{x}_*^{\mathrm{T}}\mathbf{x} = \Phi(\mathbf{x}_*)^{\mathrm{T}}\Phi(\mathbf{x})$
- if we think of $\mathbf{x}$ as the center of each basis function

Regression Models: Mean

$$k(x_*, x)k(x, x)^{-1}\mathbf{y} \qquad\qquad \mathbf{x}_*^{\mathrm{T}}\mathbf{x}\left(\beta(\alpha\mathbf{I} + \beta\mathbf{x}^{\mathrm{T}}\mathbf{x})\right)^{-1}\mathbf{y}$$

- both means are linear with respect to the output data
- how about if we take a basis function approach such that $\mathbf{x}_*^{\mathrm{T}}\mathbf{x} = \Phi(\mathbf{x}_*)^{\mathrm{T}}\Phi(\mathbf{x})$
- if we think of $\mathbf{x}$ as the center of each basis function
    - a basis function per data point

Regression Models: Mean

$$k(x_*, x)k(x, x)^{-1}\mathbf{y} \qquad\qquad \mathbf{x}_*^{\mathrm{T}}\mathbf{x}\left(\beta(\alpha\mathbf{I} + \beta\mathbf{x}^{\mathrm{T}}\mathbf{x})\right)^{-1}\mathbf{y}$$

- both means are linear with respect to the output data
- how about if we take a basis function approach such that $\mathbf{x}_*^{\mathrm{T}}\mathbf{x} = \Phi(\mathbf{x}_*)^{\mathrm{T}}\Phi(\mathbf{x})$
- if we think of $\mathbf{x}$ as the center of each basis function
  - a basis function per data point
- if we could parametrise $\Phi(\mathbf{x}_*)^{\mathrm{T}}\Phi(\mathbf{x}) = k(\mathbf{x}_*, \mathbf{x})$ as a function

$$k(x_*, x)k(x, x)^{-1}\mathbf{y} \qquad \mathbf{x}_*^{\mathrm{T}}\mathbf{x}\left(\beta(\alpha\mathbf{I} + \beta\mathbf{x}^{\mathrm{T}}\mathbf{x})\right)^{-1}\mathbf{y}$$

- both means are linear with respect to the output data
- how about if we take a basis function approach such that $\mathbf{x}_*^{\mathrm{T}}\mathbf{x} = \Phi(\mathbf{x}_*)^{\mathrm{T}}\Phi(\mathbf{x})$
- if we think of $\mathbf{x}$ as the center of each basis function
    - a basis function per data point
- if we could parametrise $\Phi(\mathbf{x}_*)^{\mathrm{T}}\Phi(\mathbf{x}) = k(\mathbf{x}_*, \mathbf{x})$ as a function
- this leads to the interpretation of GPs as infinite basis functions

Summary

Summary

- Non-parametrics
  - parametrise relationship between data

Summary

- Non-parametrics
  - parametrise relationship between data
- Gaussian processes
  - Implementation: "just a big big Gaussian"
  - Theory: projection of an infinite stochastic process

Kolmogrovs Extension Theorem

For all permutations π, measurable sets $F_i \subseteq \mathbb{R}^n$ and probability measure ν

1. Exchangeable

$$\nu_{t_{\pi(1)}\cdots t_{\pi(k)}}\left(F_{\pi(1)} \times \cdot \times F_{\pi(k)}\right) = \nu_{t_1 \cdots t_k}\left(F_1 \times \cdots \times F_k\right)$$

2. Marginal

$$\nu_{t_1 \cdot t_k}\left(F_1 \times \cdot \times F_k\right) = \nu_{t_1 \cdots t_k, t_{k+1} \cdot t_{k+m}}\left(F_1 \times \cdot \times F_k \times \mathbb{R}^n \times \cdot \times \mathbb{R}^n\right)$$

In this case the finite dimensional probability measure is a realisation of an underlying stochastic process eof