



Machine Learning and the Physical World

Lecture 4 : Gaussian Processes

Carl Henrik Ek - che29@cam.ac.uk

23rd of October, 2026ng

<http://carlhenrik.com>

Data + Model $\overset{\text{Compute}}{\rightarrow}$ Prediction

Data + Model $\xrightarrow{\text{Compute}}$ Prediction

- We do not have enough knowledge

Model $\xrightarrow[\text{?}]{\text{Compute}}$ Prediction

Model+Data $\rightarrow$ Predict

Data + Model $\xrightarrow{\text{Compute}}$ Prediction

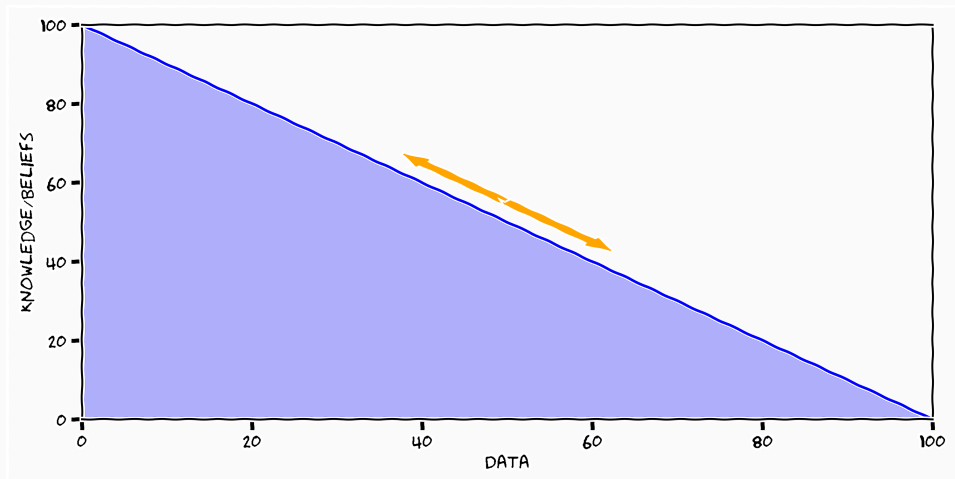
- We do not have enough knowledge

Model $\xrightarrow[\text{?}]{\text{Compute}}$ Prediction

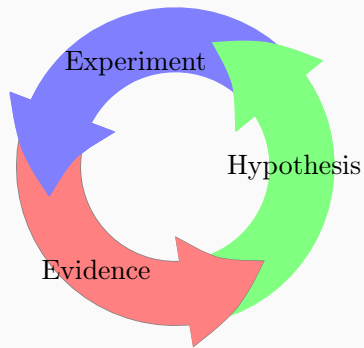
- We do not have enough data

Data $\xrightarrow[\text{?}]{\text{Compute}}$ Prediction

Data and Knowledge

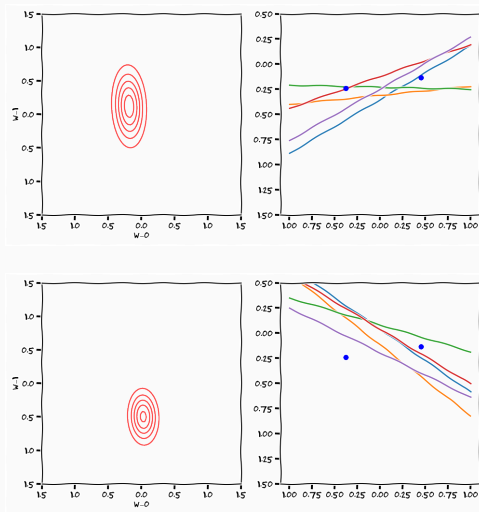


The Scientific Principle



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

Knowledge is Relative



$$\mathcal{A}_{\mathcal{H}}(\mathcal{S})$$

- $\mathcal{H}$ - hypothesis space
- $\mathcal{A}$ - "traverse" mechanism of the hypothesis space
- $\mathcal{S}$ - training data

- Machine Learning as a Framework

Data + Model $\xrightarrow{\text{Compute}}$ Prediction

- Machine Learning as a Framework

Data + Model $\xrightarrow{\text{Compute}}$ Prediction

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

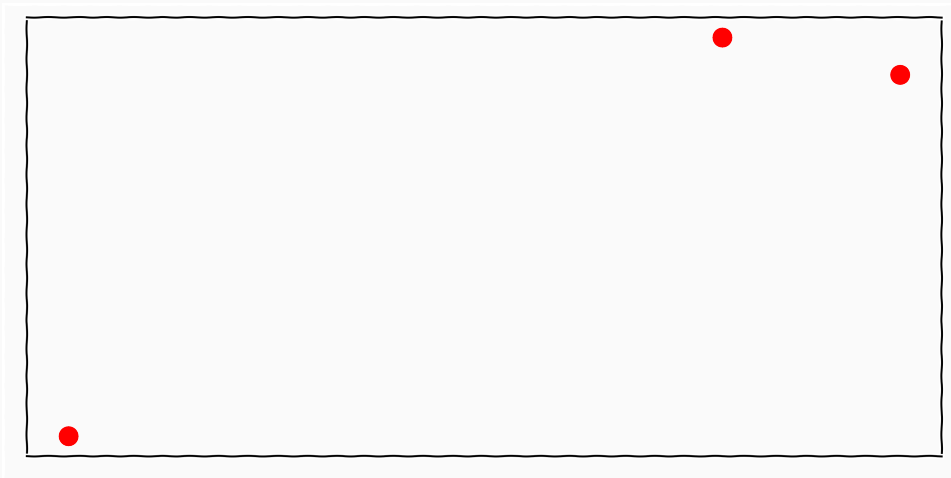
- How to specify beliefs over larger classes of hypothesis

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

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

Non-parametrics

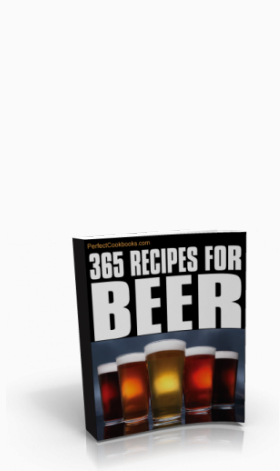
Hypothesis Classes



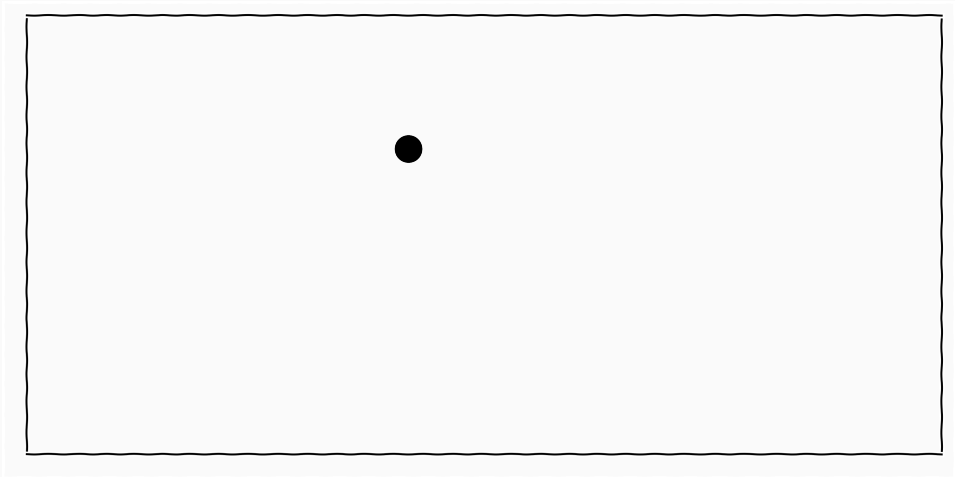




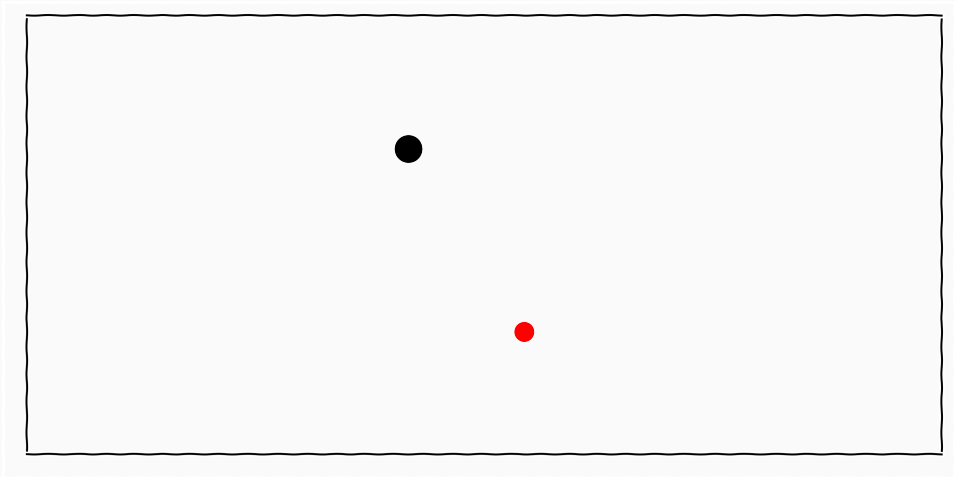
Non-parametrics



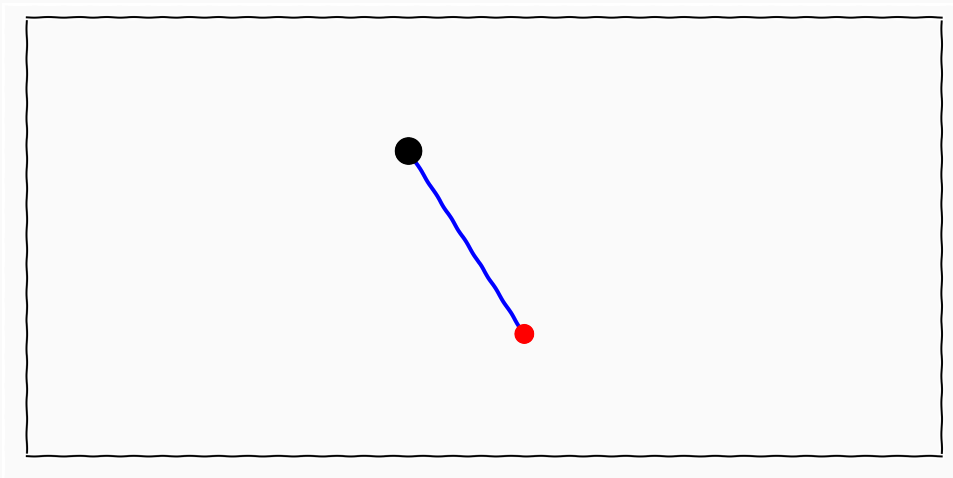
Example



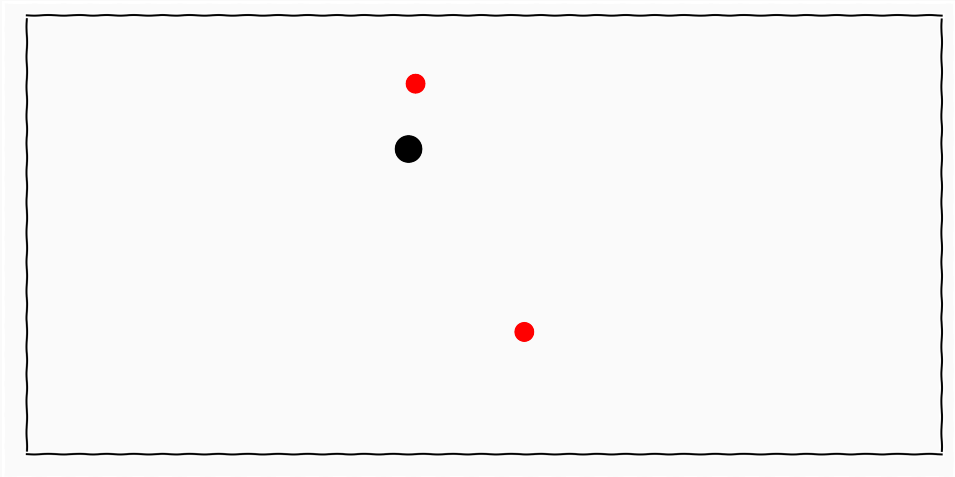
Example



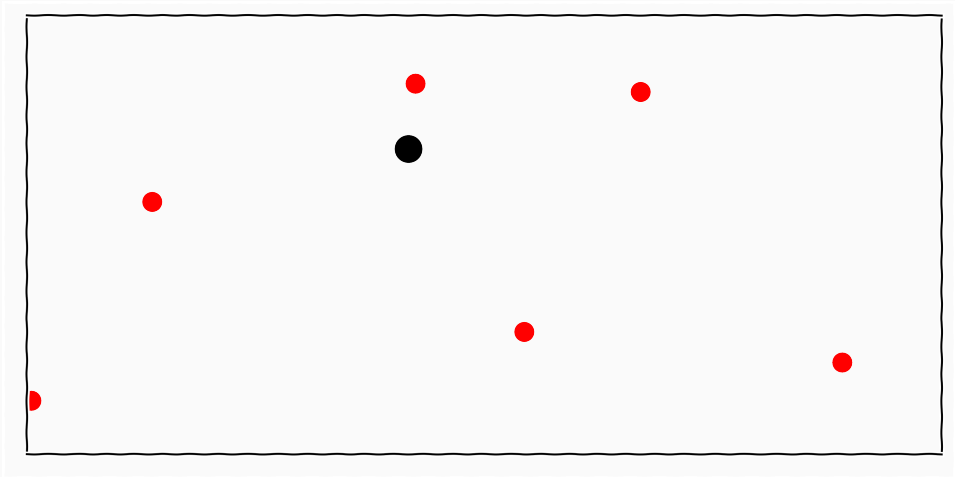
Example



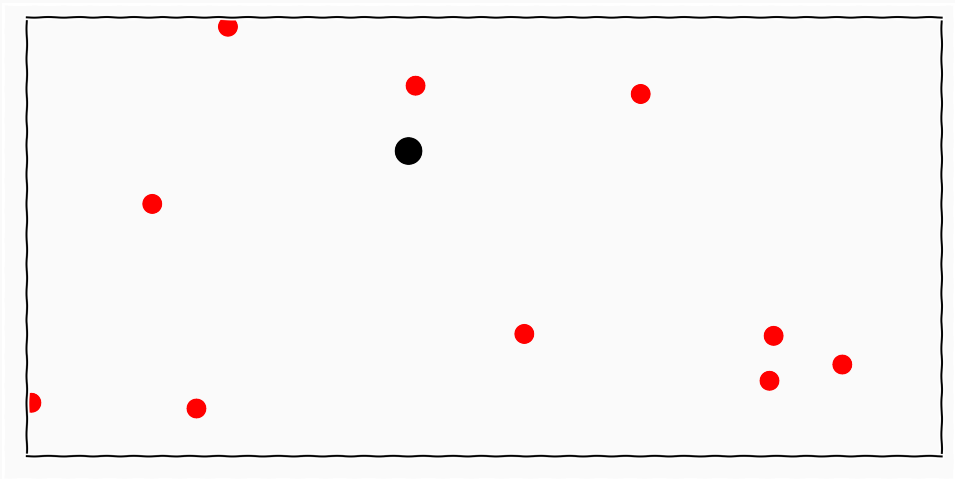
Example



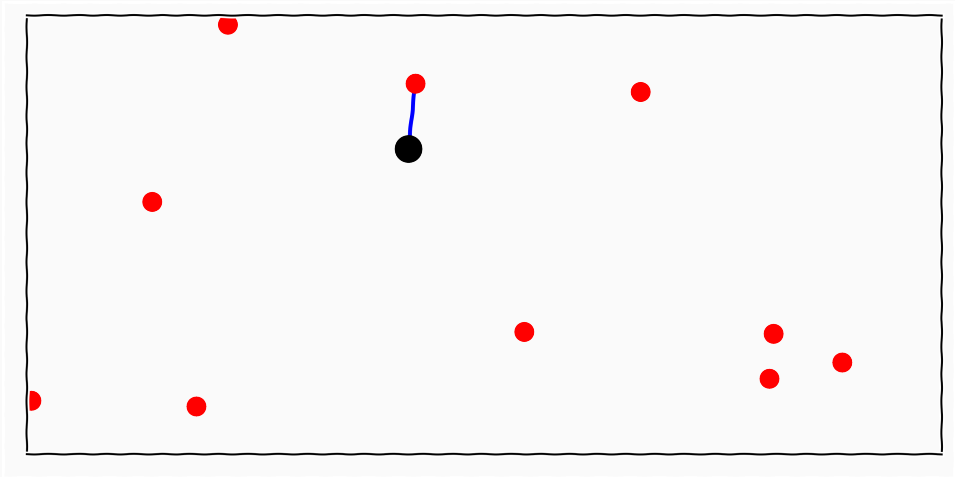
Example



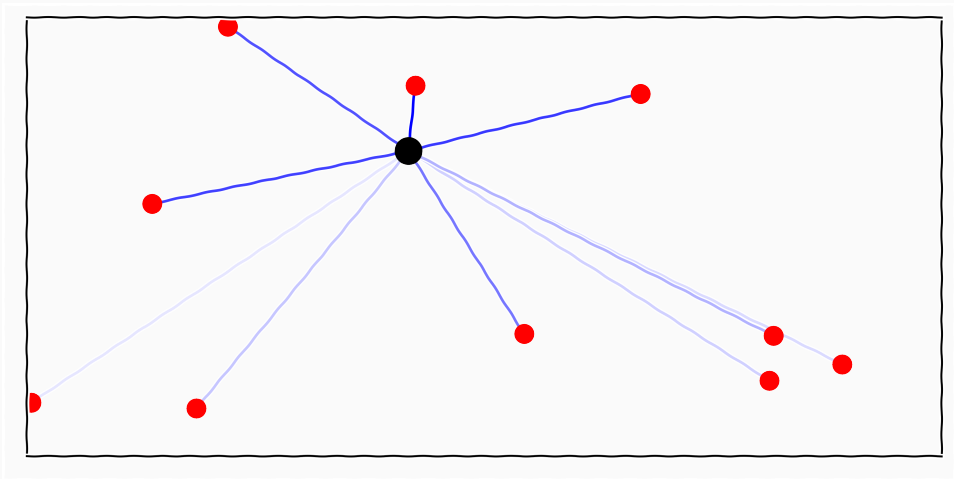
Example



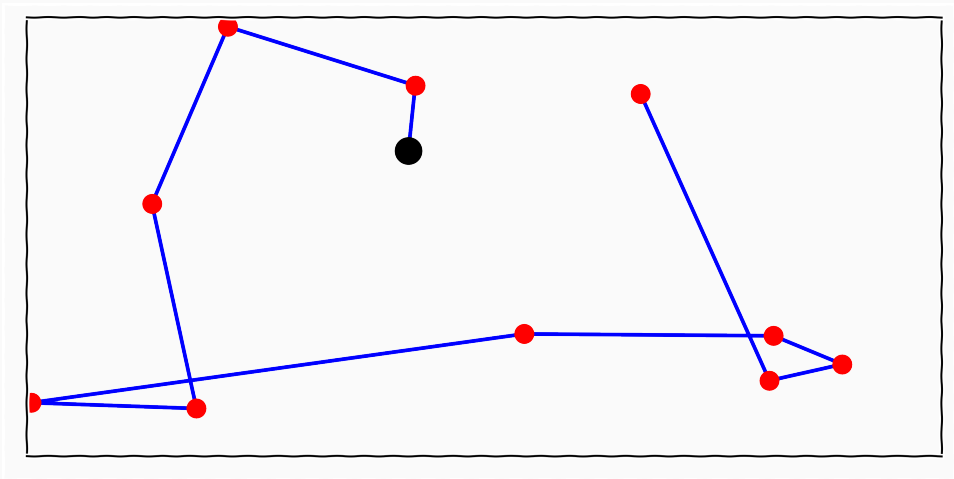
Example



Example

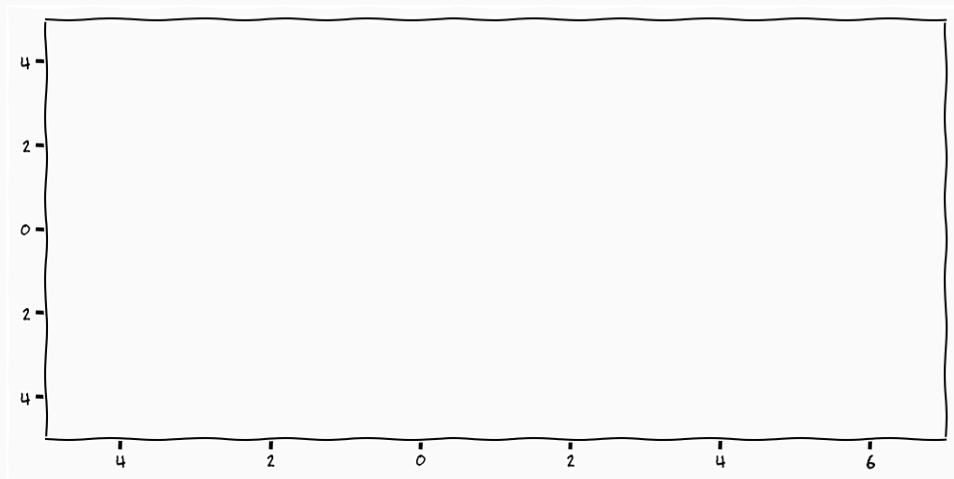


Example

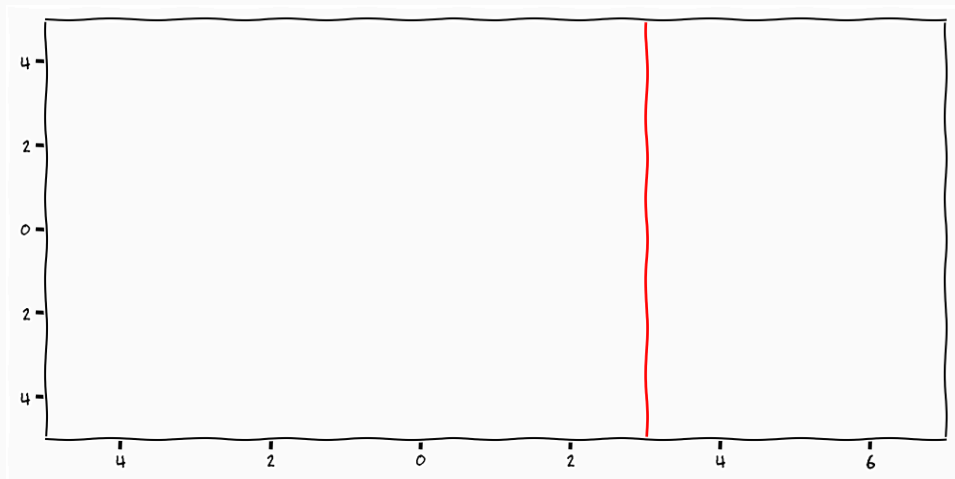


Non-parametric Functions

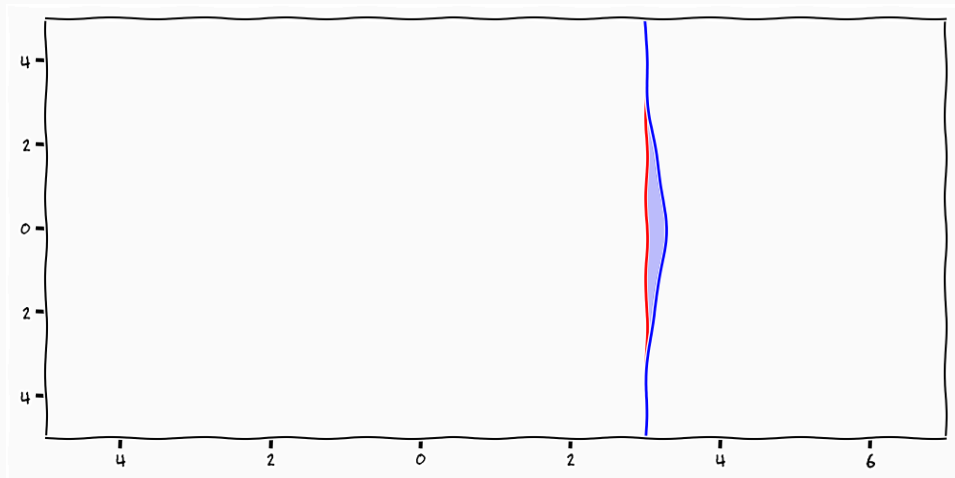
Lets talk about functions



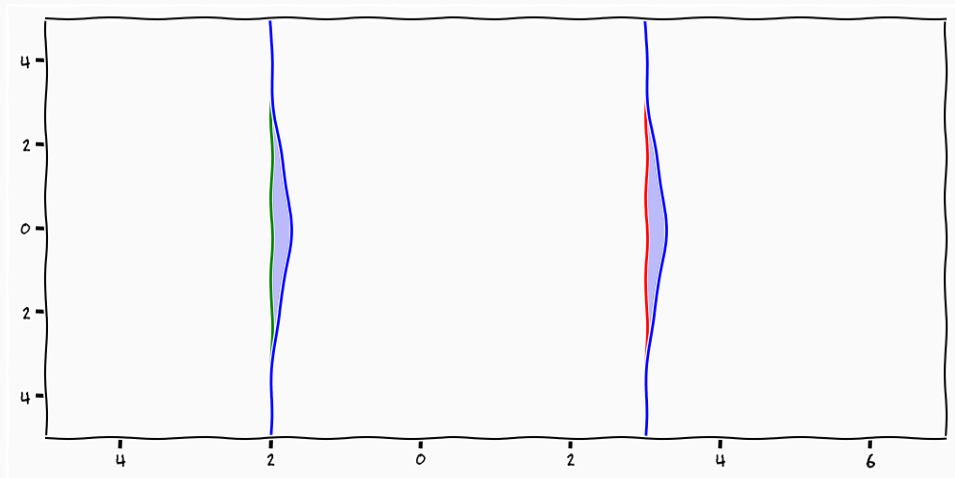
Lets talk about functions



Lets talk about functions



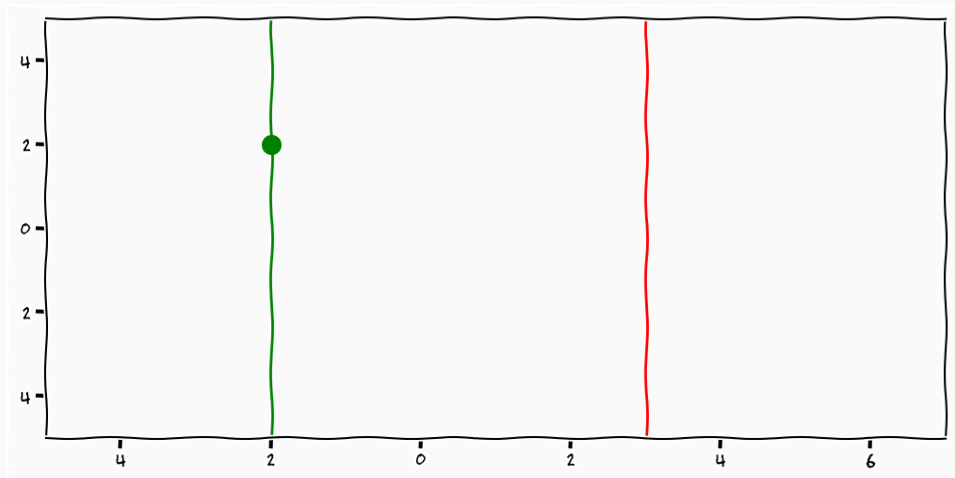
Lets talk about functions



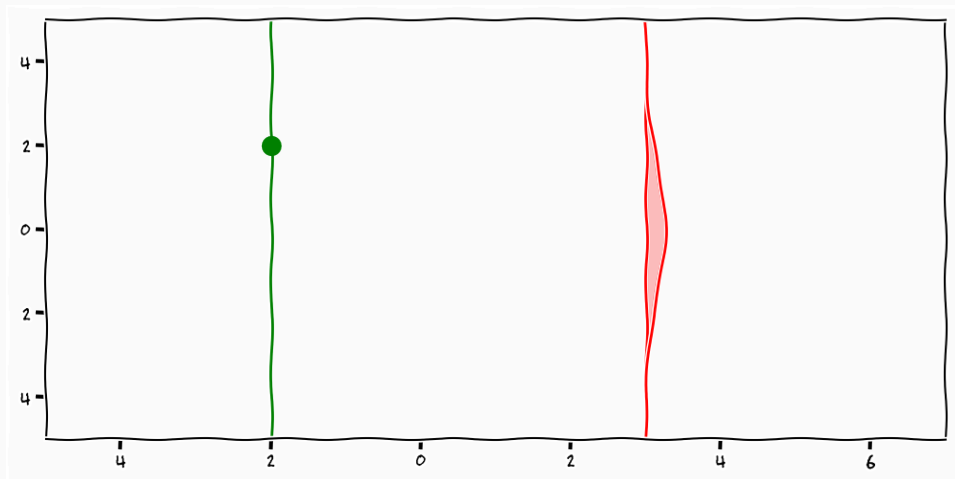
$$f_1 = \mathcal{N}(\mu_1, k_1)$$

$$f_2 = \mathcal{N}(\mu_2, k_2)$$

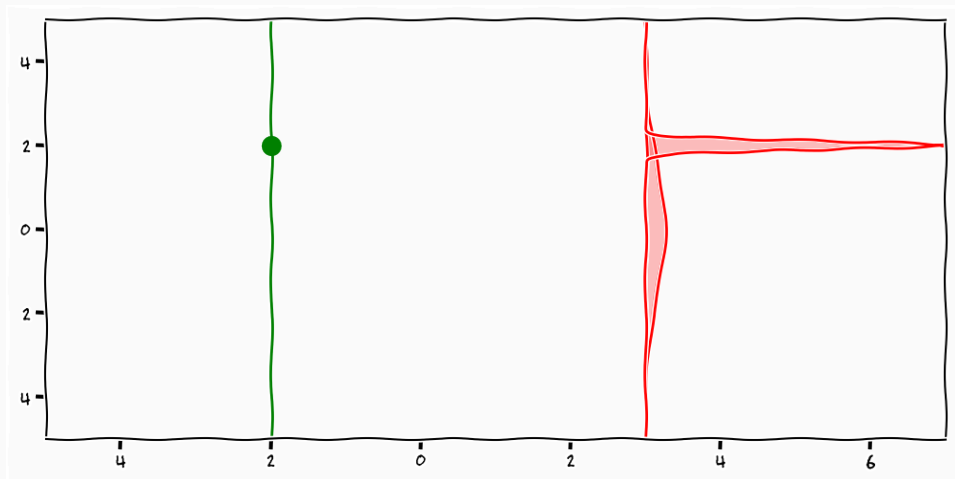
Non-parametric functions



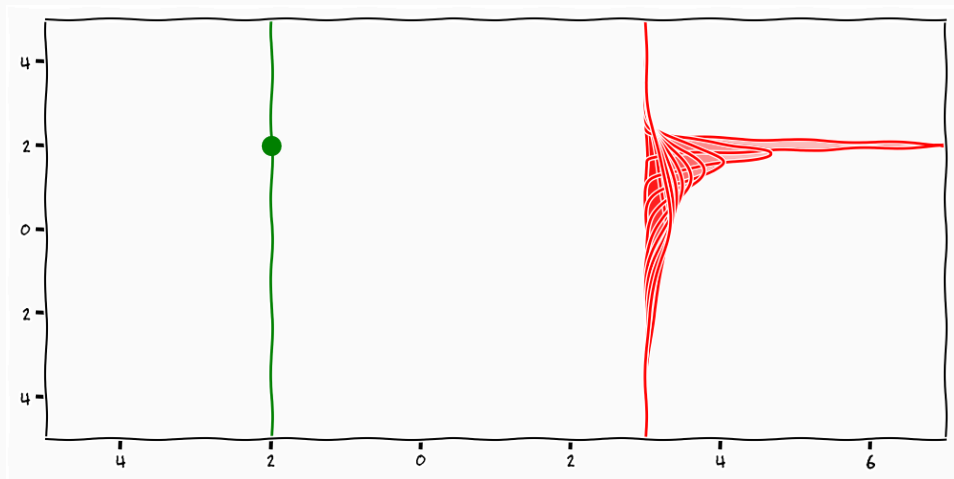
Non-parametric functions



Non-parametric functions



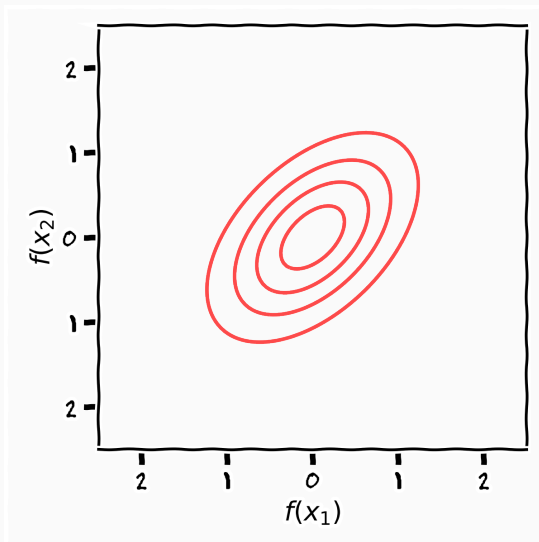
Non-parametric functions



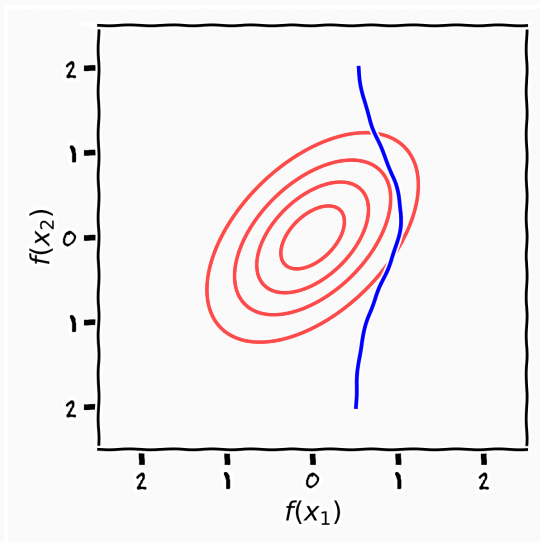
$$\begin{bmatrix} f_1 \\ f_2 \end{bmatrix} = \mathcal{N} \left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} k_{11} & ? \\ ? & k_{22} \end{bmatrix} \right)$$

$$\begin{bmatrix} f_1 \\ f_2 \end{bmatrix} = \mathcal{N} \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \right)$$

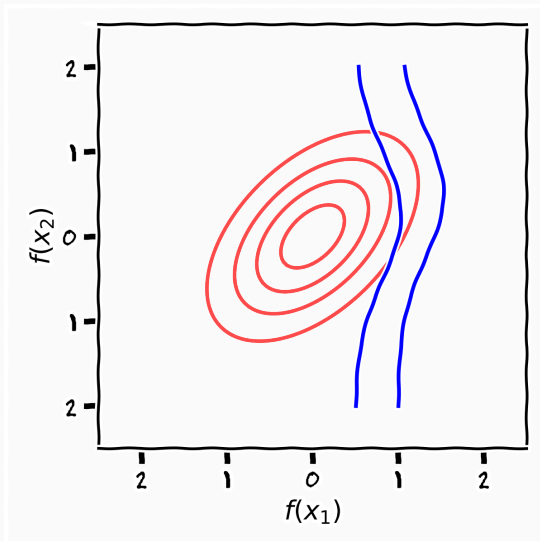
Conditional Gaussians



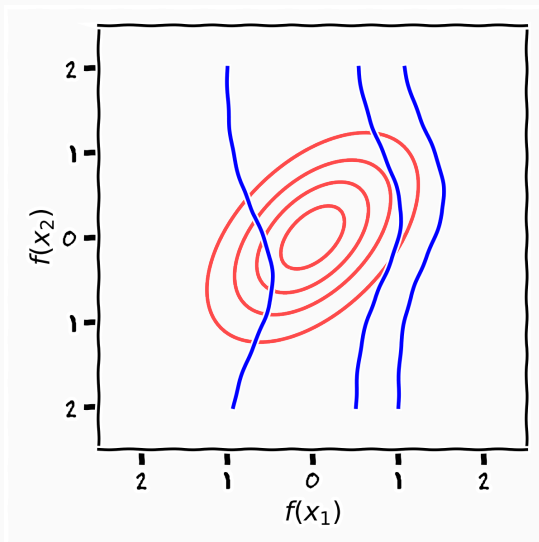
Conditional Gaussians



Conditional Gaussians

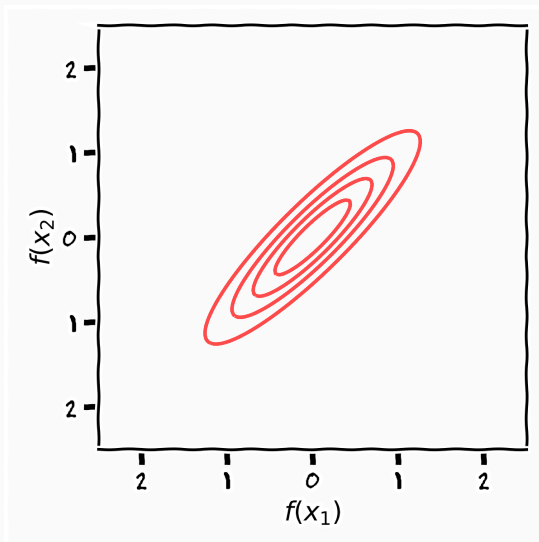


Conditional Gaussians

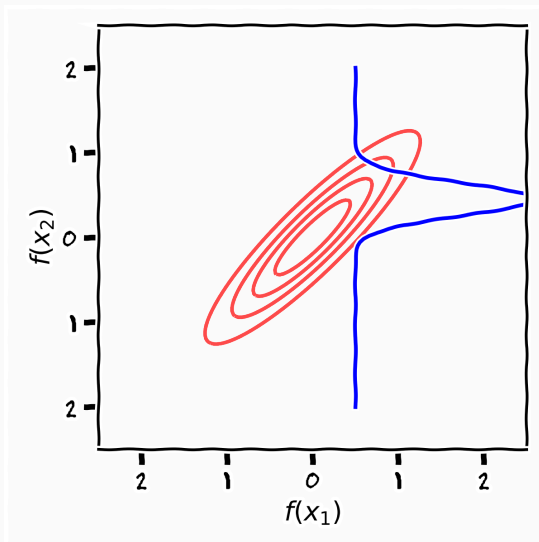


$$\begin{bmatrix} f_1 \\ f_2 \end{bmatrix} = \mathcal{N} \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix} \right)$$

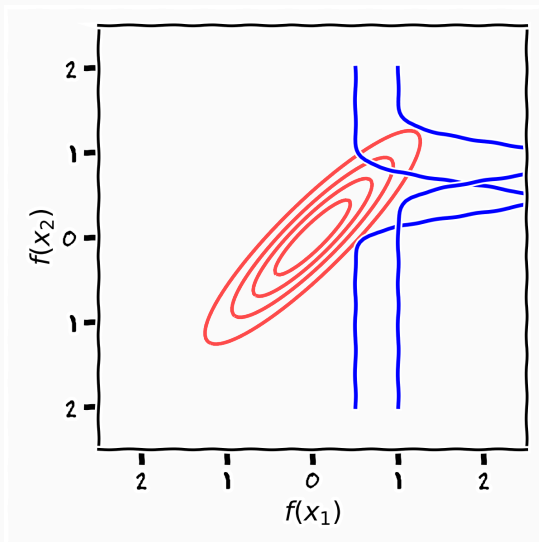
Conditional Gaussians



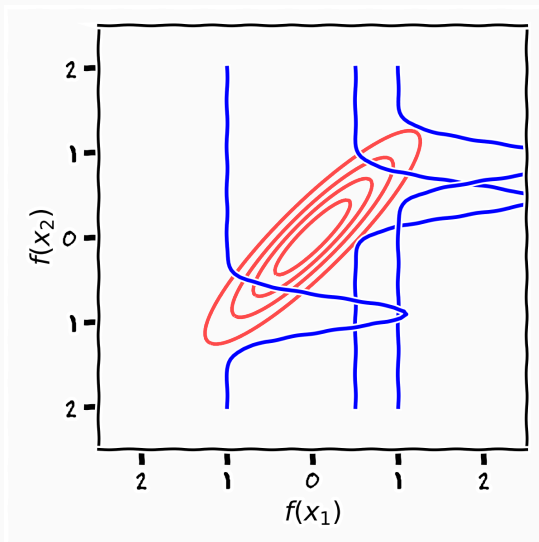
Conditional Gaussians



Conditional Gaussians

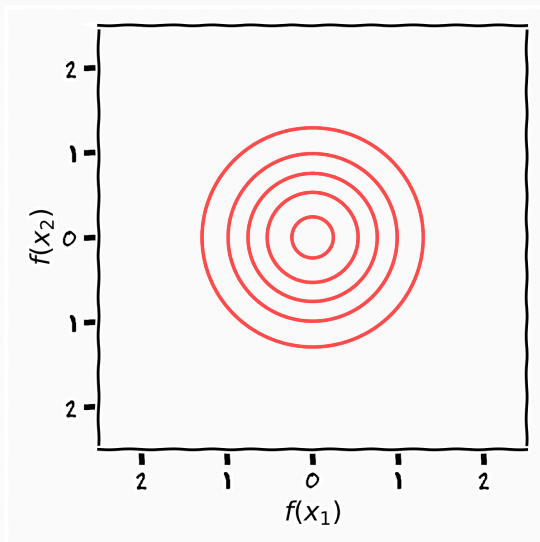


Conditional Gaussians

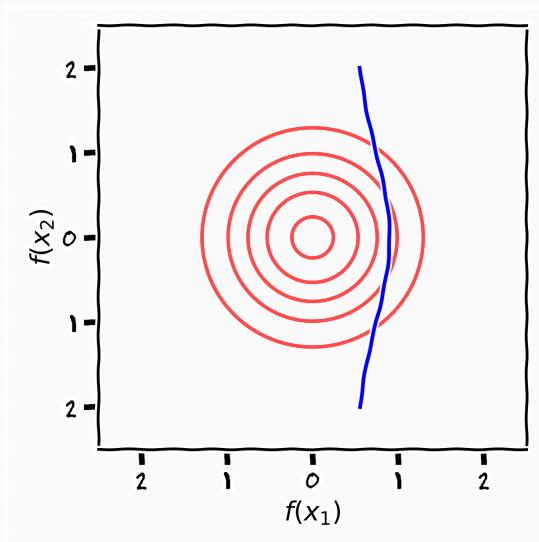


$$\begin{bmatrix} f_1 \\ f_2 \end{bmatrix} = \mathcal{N} \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right)$$

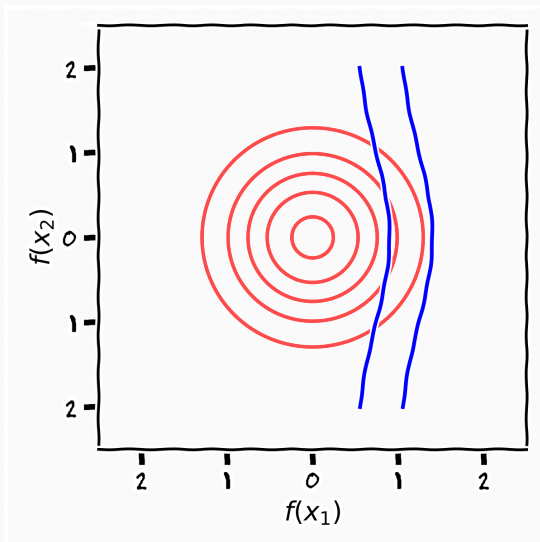
Conditional Gaussians



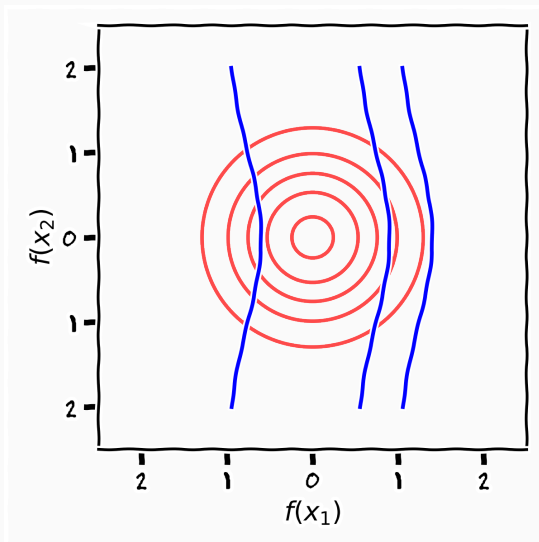
Conditional Gaussians



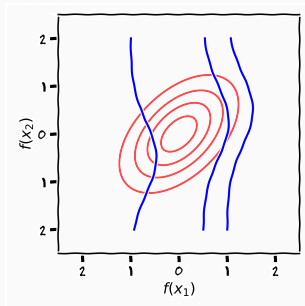
Conditional Gaussians



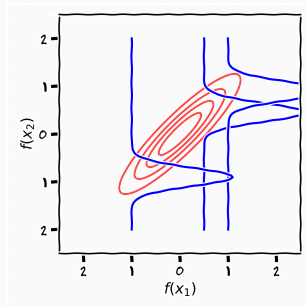
Conditional Gaussians



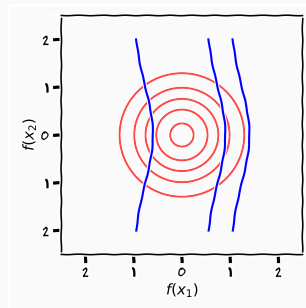
Conditional Gaussians



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

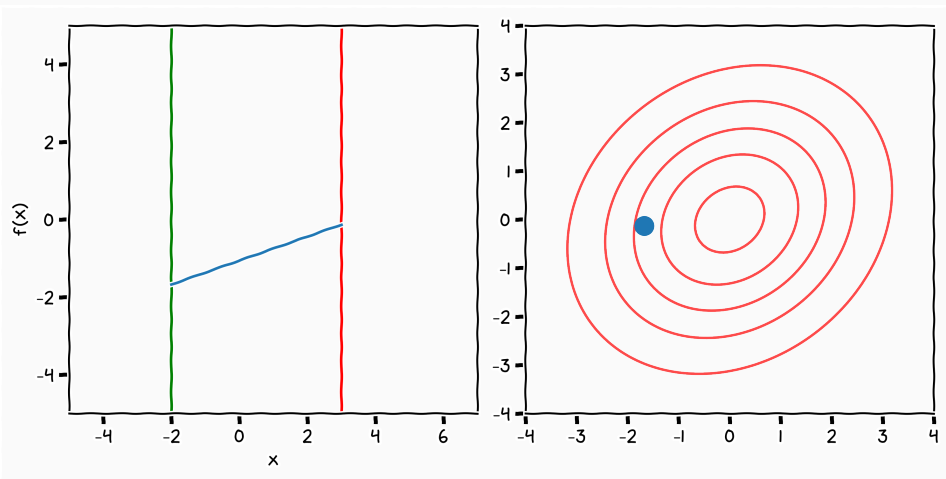


$$N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}\right)$$

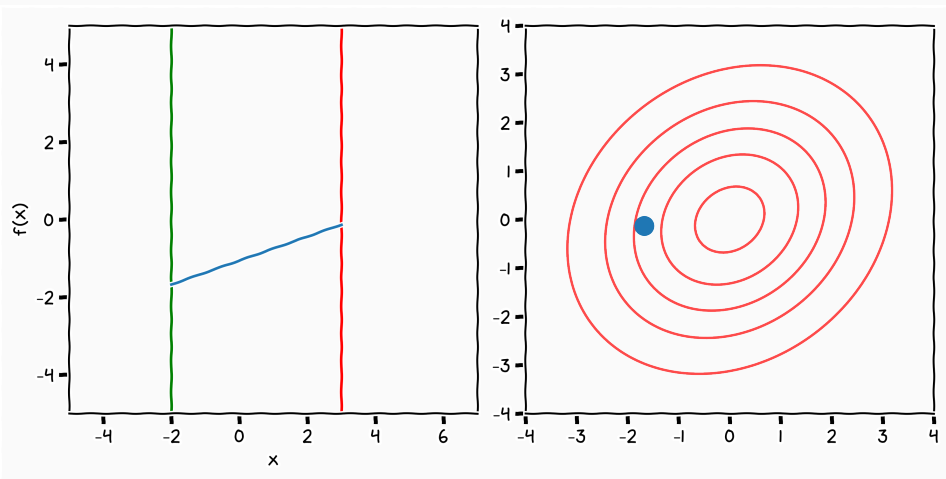


$$N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$$

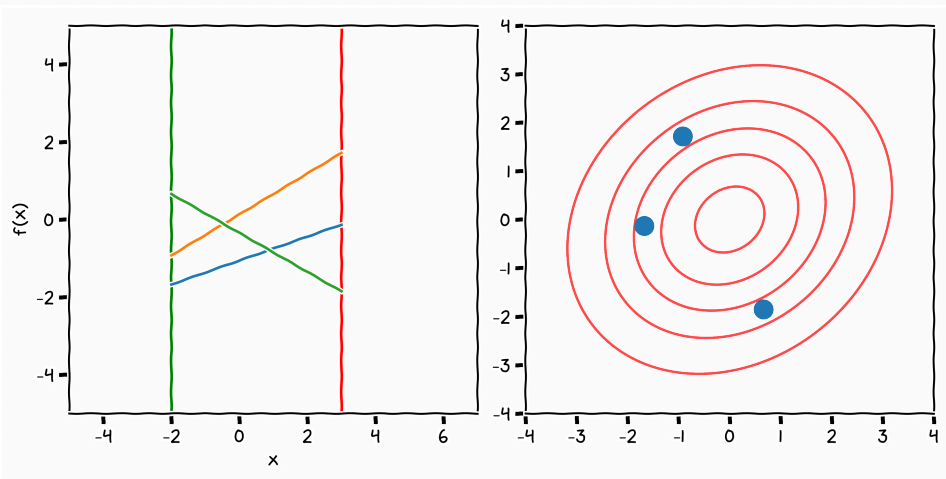
Gaussian Samples



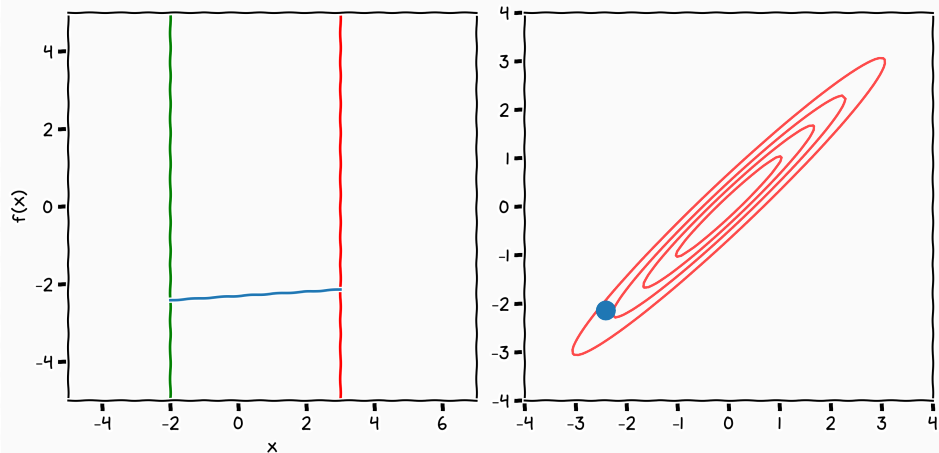
Gaussian Samples



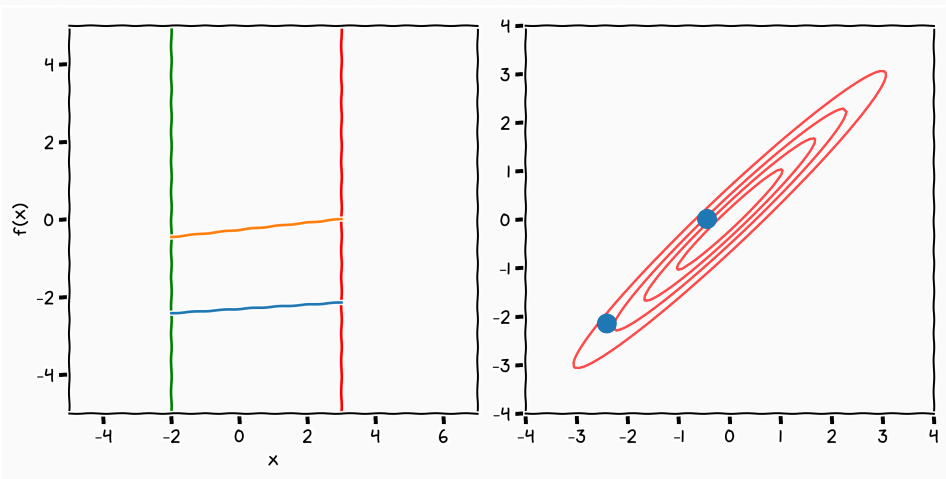
Gaussian Samples



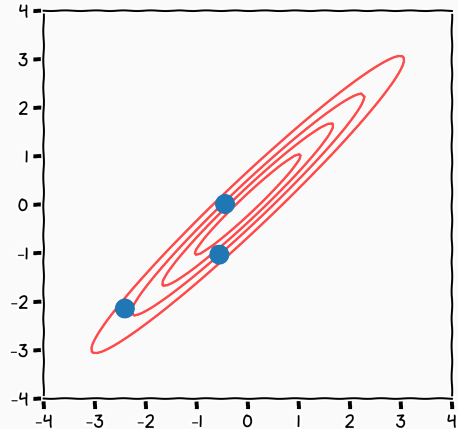
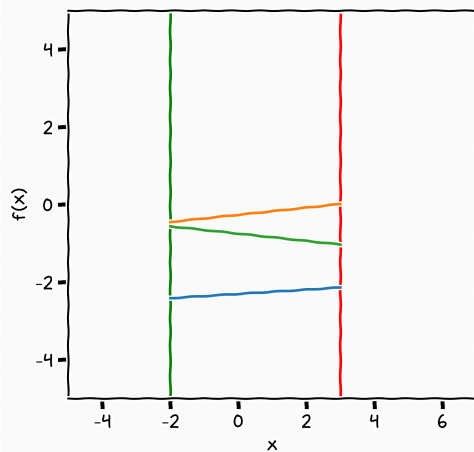
Gaussian Samples



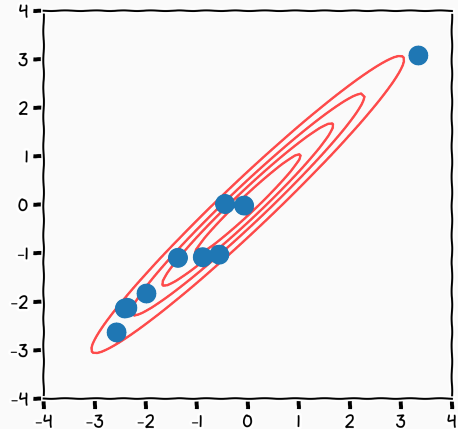
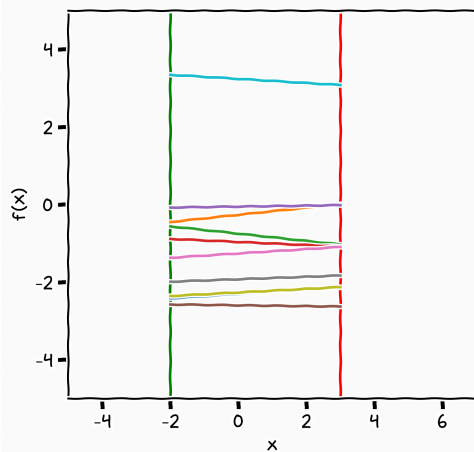
Gaussian Samples



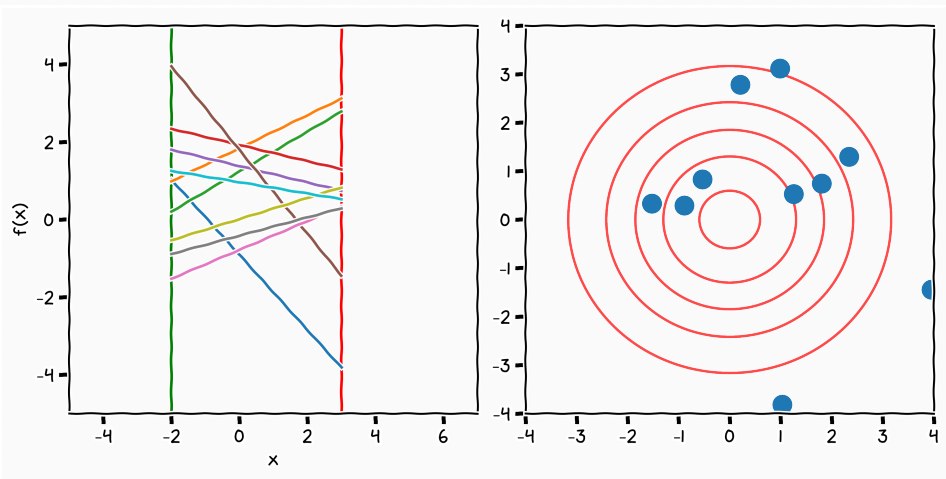
Gaussian Samples



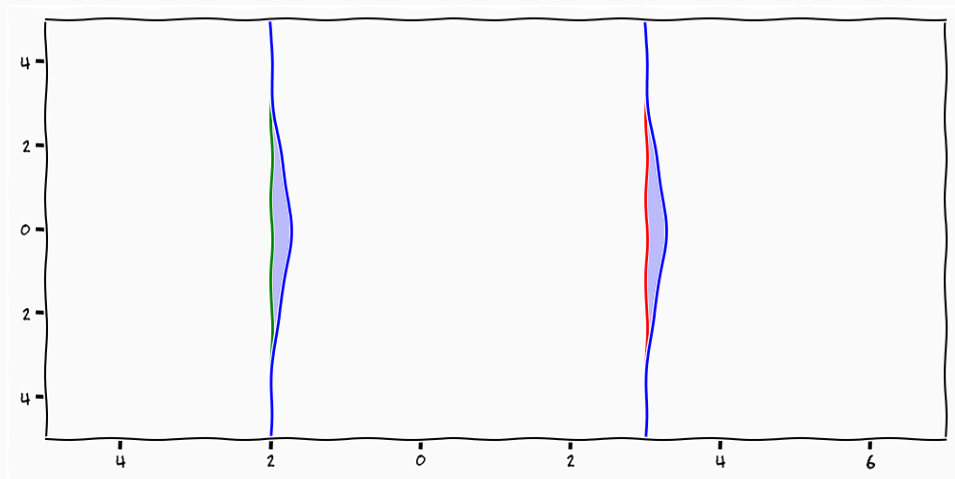
Gaussian Samples



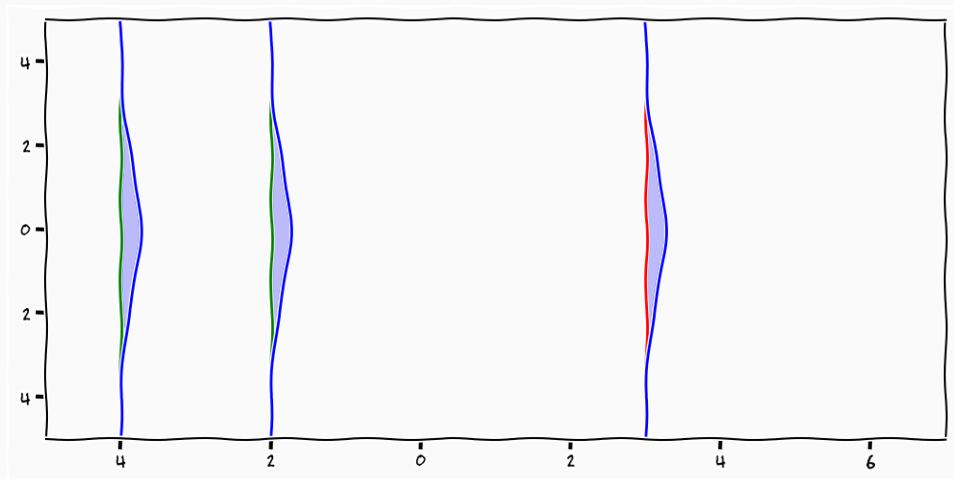
Gaussian Samples



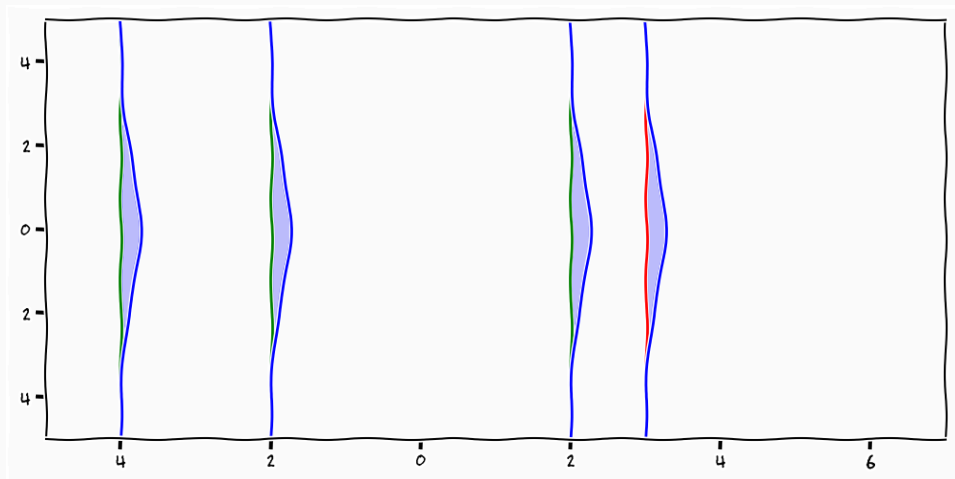
Lets talk about functions



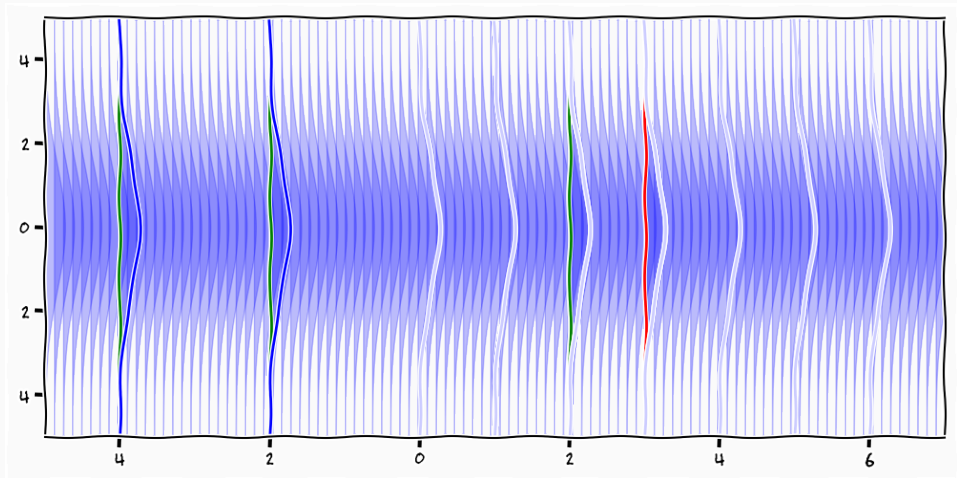
Non-parametric functions



Non-parametric functions



Non-parametric functions



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

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

$$p(\mathbf{x}_1, x_2) = \mathcal{N} \left(\begin{array}{c} \mathbf{x}_1 \\ x_2 \end{array} \middle| \begin{array}{cc} \mu_1 & k_{11} \\ \mu_2 & k_{21} \end{array}, \begin{array}{cc} k_{12} & k_{22} \end{array} \right)$$
$$\Rightarrow p(\mathbf{x}_1) = \int_{x_2} p(\mathbf{x}_1, x_2) = \underline{\mathcal{N}(\mathbf{x}_1 \mid \mu_1, k_{11})}$$

$$p(\mathbf{x}_1, x_2) = \mathcal{N} \left(\begin{array}{c|cc} \mathbf{x}_1 & \mu_1 & k_{11} \\ x_2 & \mu_2 & k_{21} \end{array} \begin{array}{c} k_{12} \\ k_{22} \end{array} \right)$$

$$\Rightarrow p(\mathbf{x}_1) = \int_{x_2} p(\mathbf{x}_1, x_2) = \underline{\mathcal{N}(\mathbf{x}_1 \mid \mu_1, k_{11})}$$

$$p(\mathbf{x}_1, x_2, \dots, x_N) = \mathcal{N} \left(\begin{array}{c|cccc} \mathbf{x}_1 & \mu_1 & k_{11} & k_{12} & \cdots & k_{1N} \\ x_2 & \mu_2 & k_{21} & k_{22} & \cdots & k_{2N} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_N & \mu_N & k_{N1} & k_{N2} & \cdots & k_{NN} \end{array} \right)$$

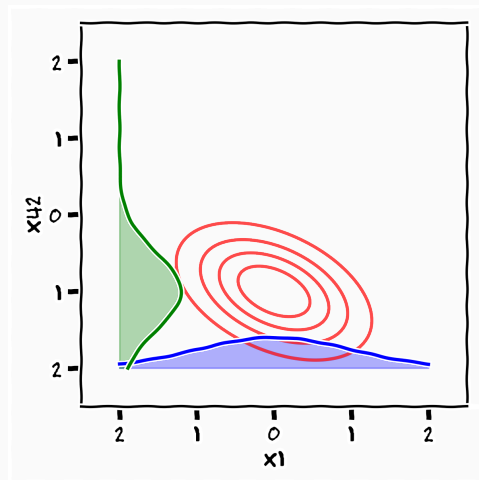
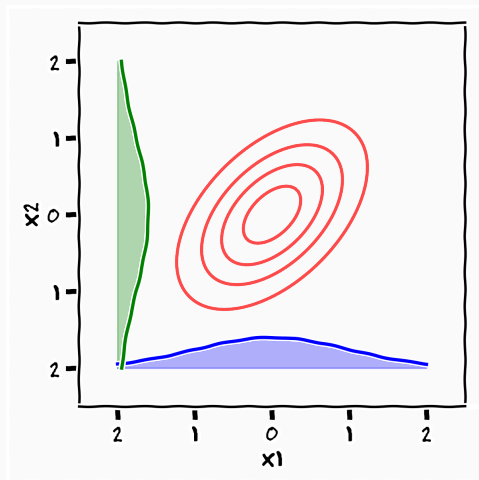
$$p(\mathbf{x}_1, x_2) = \mathcal{N} \left(\begin{array}{c|cc} \mathbf{x}_1 & \mu_1 & k_{11} & k_{12} \\ x_2 & \mu_2 & k_{21} & k_{22} \end{array} \right)$$

$$\Rightarrow p(\mathbf{x}_1) = \int_{x_2} p(\mathbf{x}_1, x_2) = \underline{\mathcal{N}(\mathbf{x}_1 \mid \mu_1, k_{11})}$$

$$p(\mathbf{x}_1, x_2, \dots, x_N) = \mathcal{N} \left(\begin{array}{c|cccc} \mathbf{x}_1 & \mu_1 & k_{11} & k_{12} & \cdots & k_{1N} \\ x_2 & \mu_2 & k_{21} & k_{22} & \cdots & k_{2N} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_N & \mu_N & k_{N1} & k_{N2} & \cdots & k_{NN} \end{array} \right)$$

$$\Rightarrow p(\mathbf{x}_1) = \int_{x_2, \dots, x_N} p(\mathbf{x}_1, x_2, \dots, x_N) = \underline{\mathcal{N}(\mathbf{x}_1 \mid \mu_1, k_{11})}$$

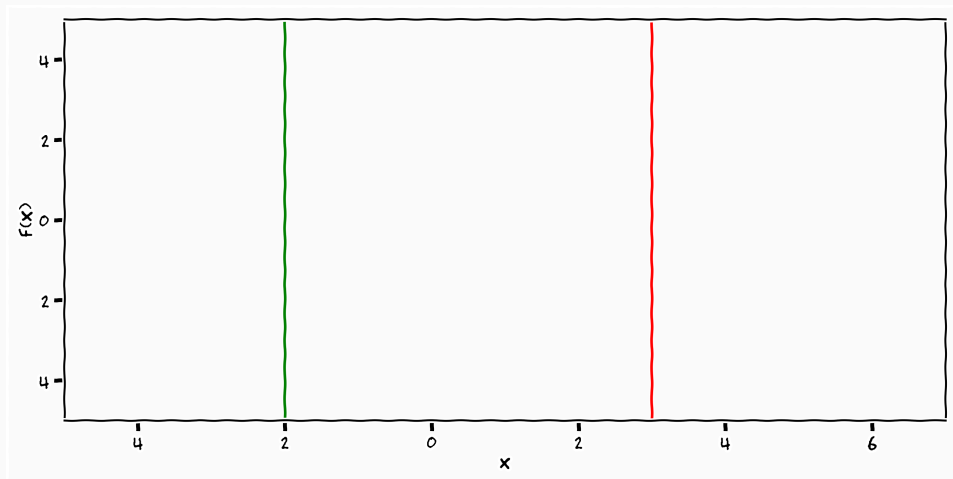
Gaussian Distribution - Marginal



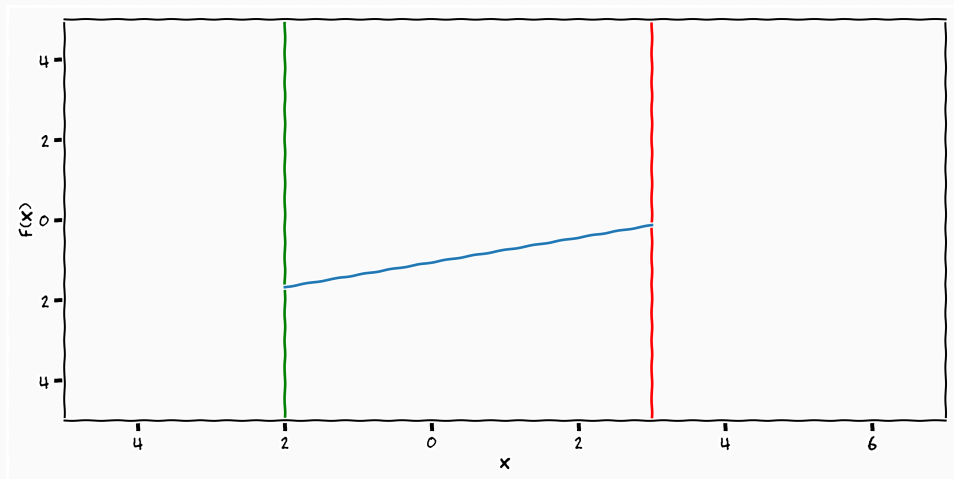
For all measurable sets $F_i \subseteq \mathbb{R}^n$ and probability measure $\mathcal{N}$

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

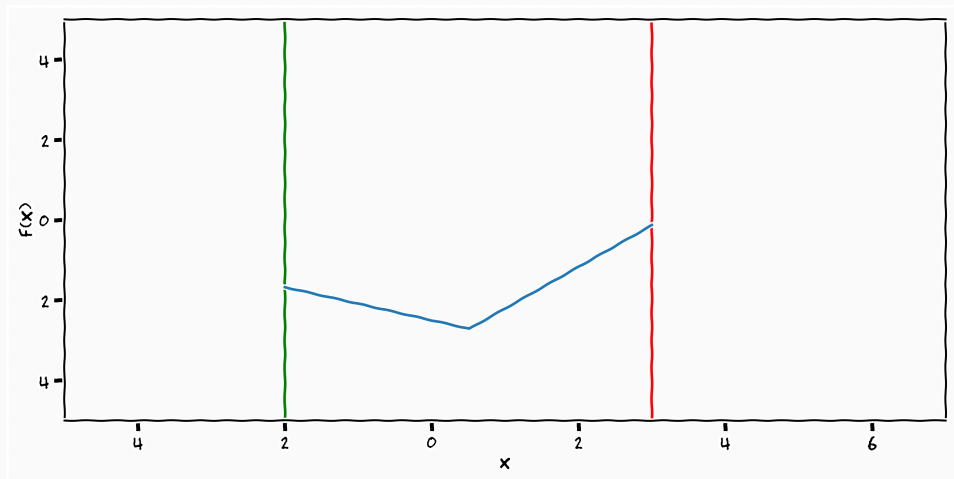
Gaussian Samples



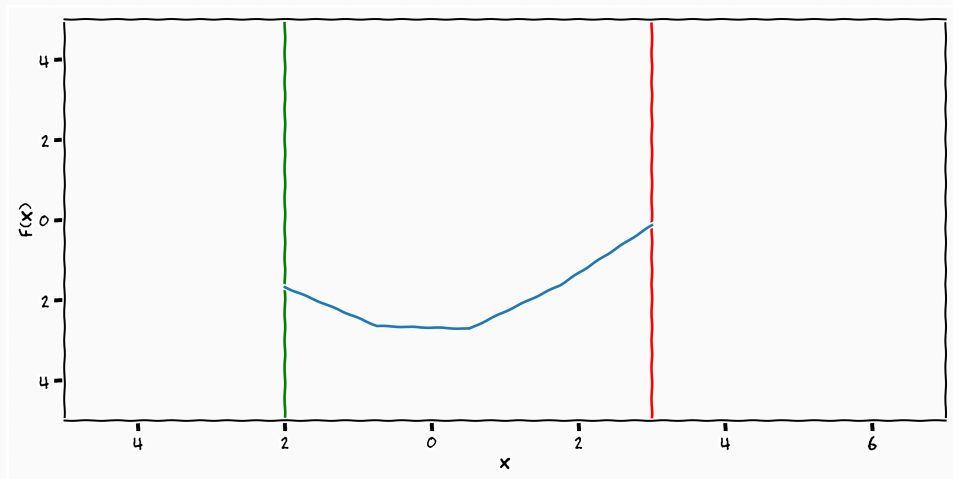
Gaussian Samples



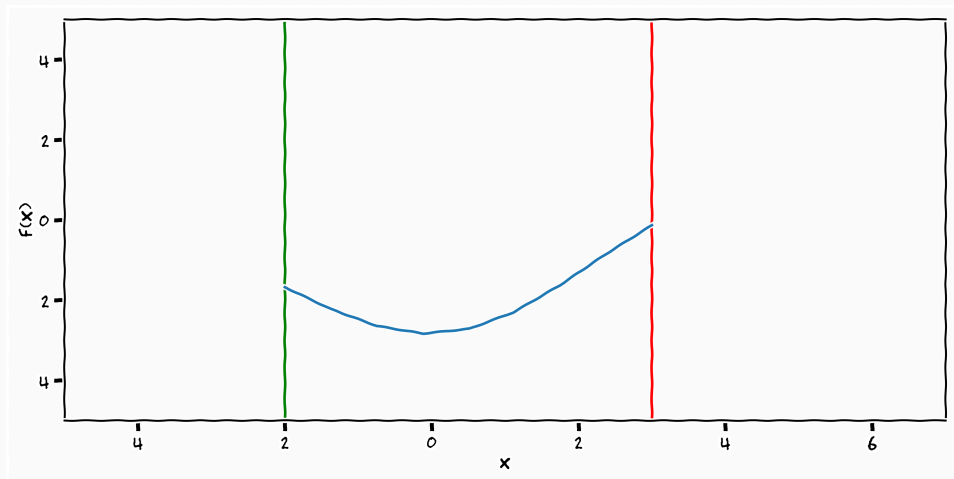
Gaussian Samples



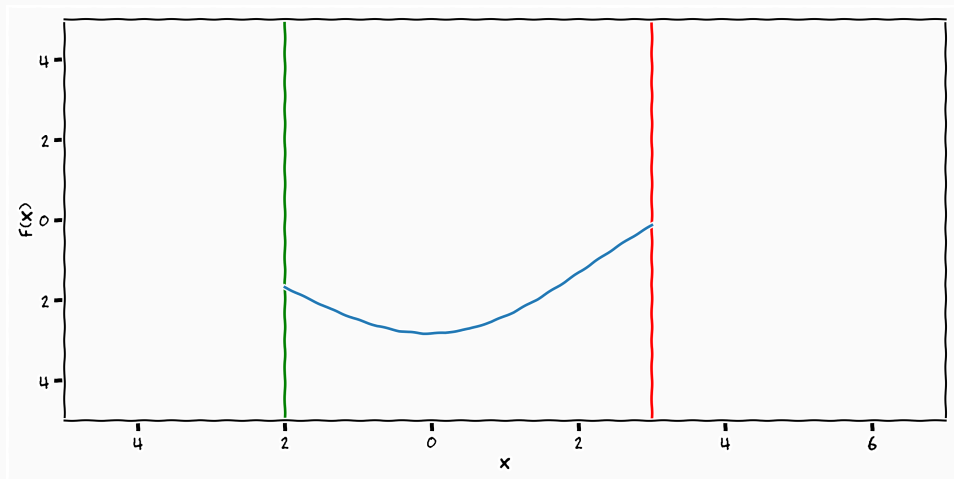
Gaussian Samples



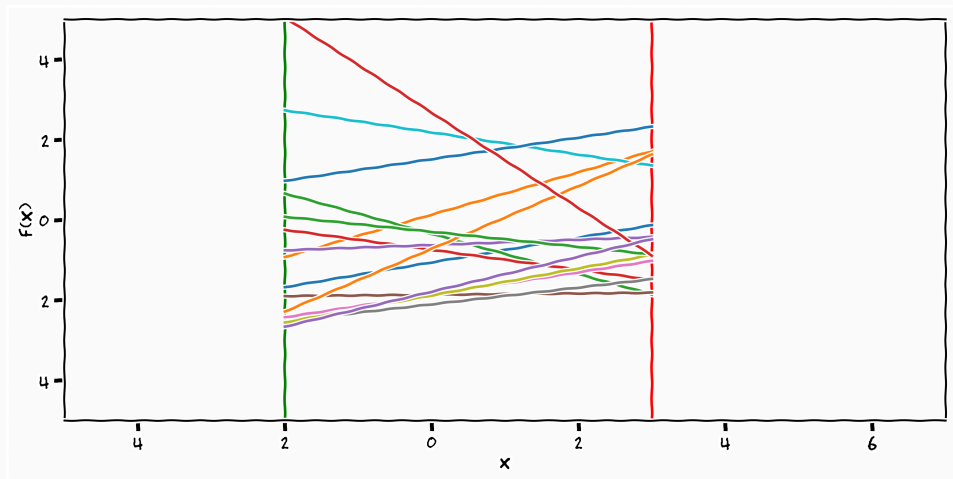
Gaussian Samples



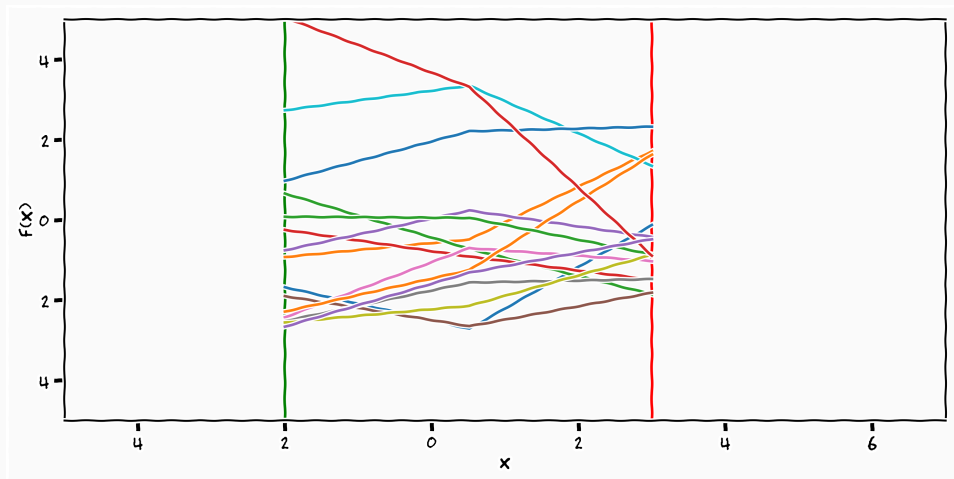
Gaussian Samples



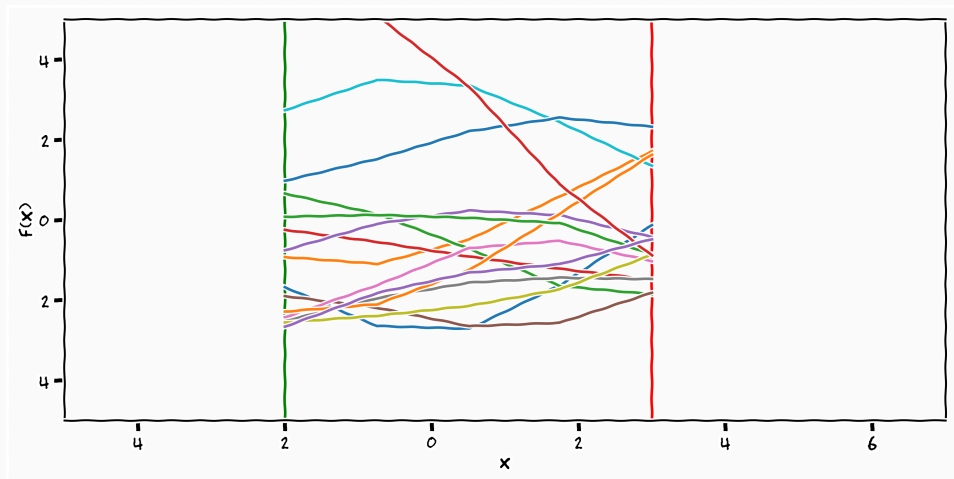
Gaussian Samples



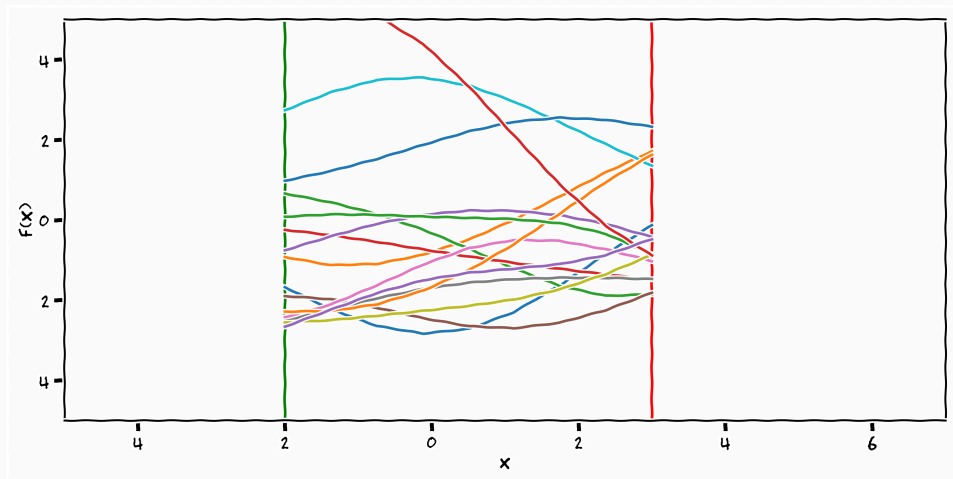
Gaussian Samples



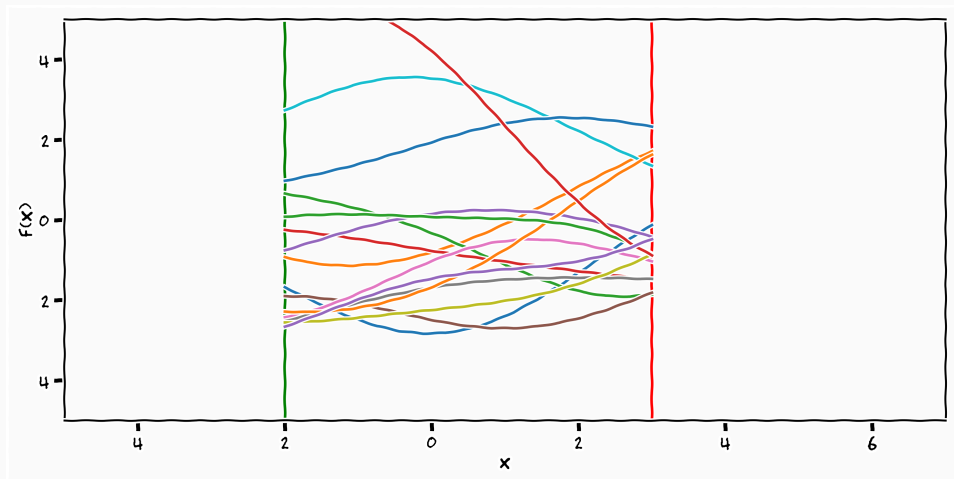
Gaussian Samples



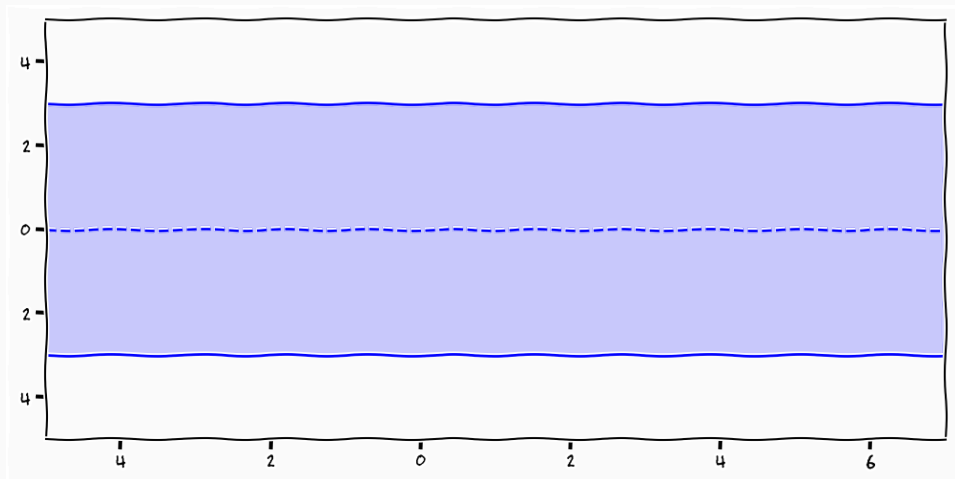
Gaussian Samples



Gaussian Samples



Gaussian Processes



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

$$\begin{array}{ccc} \mathcal{GP}(\cdot, \cdot) & & \mathcal{N}(\cdot, \cdot) \\ & M \in \mathbb{R}^{\infty \times N} & \\ & \rightarrow & \\ \infty & & N \end{array}$$

The Gaussian distribution is the projection of the infinite Gaussian process

- The "function" is completely specified by
 - a mean vector $\boldsymbol{\mu} \in \mathbb{R}^\infty$
 - a co-variance matrix $\mathbf{K}^{\infty \times \infty}$

- The "function" is completely specified by
 - a mean vector $\boldsymbol{\mu} \in \mathbb{R}^\infty$
 - a co-variance matrix $\mathbf{K}^{\infty \times \infty}$
- They are not without restrictions as they need to parametrise a valid Gaussian
 - $\mathbf{K}_{ii} \geq 0$
 - $\mathbf{K} \succcurlyeq 0$

$$k_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$$

- We parameterise the covariance as a function of the input

$$k_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$$

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

$$k_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$$

- 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

$$k_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$$

- 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*

$$k_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$$

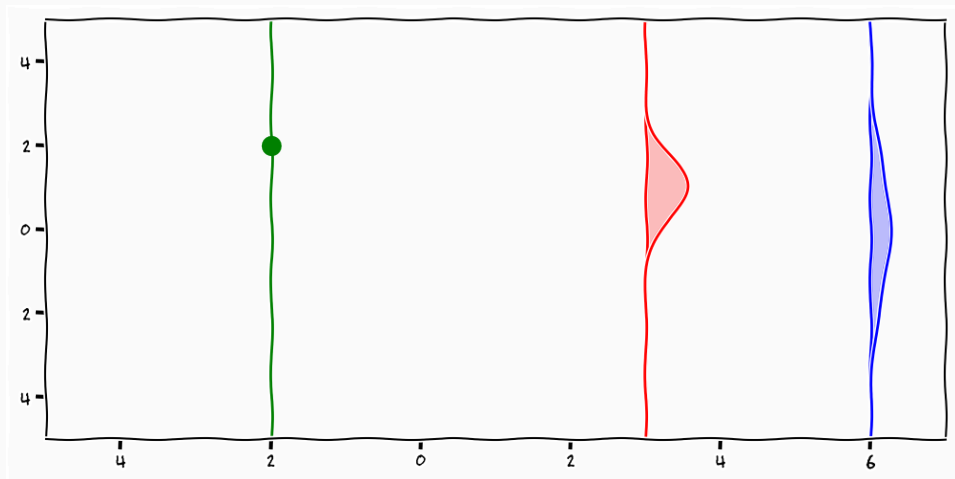
- We parameterise the covariance as a function of the input
- the index set of the measure is the uncountable infinity
- Your "handle" to input your knowledge into a GP is the covariance function
 - *you specify the degree of covariance between data-points*
- *This is your "handle" to input knowledge*

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

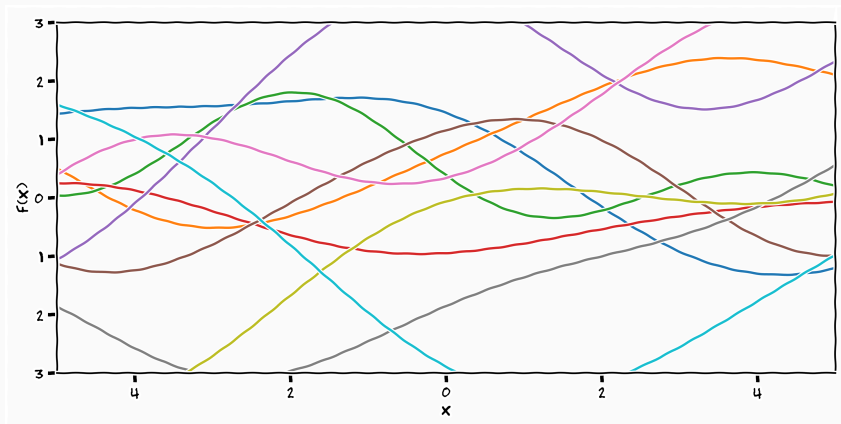
Definition (Gaussian Process)

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

Gaussian Processes

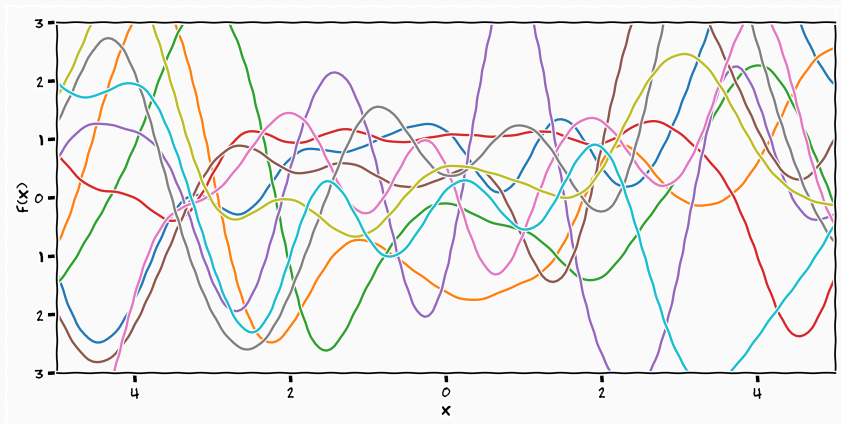


Gaussian Processes Samples



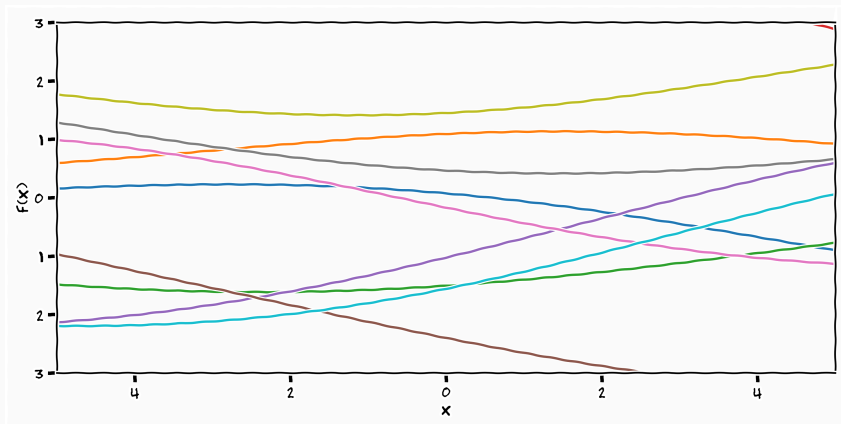
$$k(x_i, x_j) = 3 \cdot e^{-\frac{(x_i - x_j)^2}{15}}$$

Gaussian Processes Samples



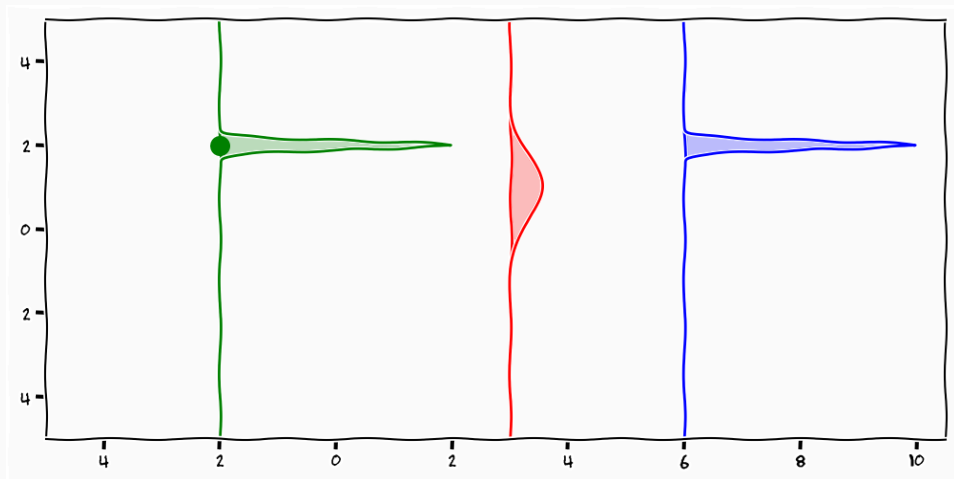
$$k(x_i, x_j) = 3 \cdot e^{-\frac{(x_i - x_j)^2}{1}}$$

Gaussian Processes Samples



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

Gaussian Processes



Bayesian Inference

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

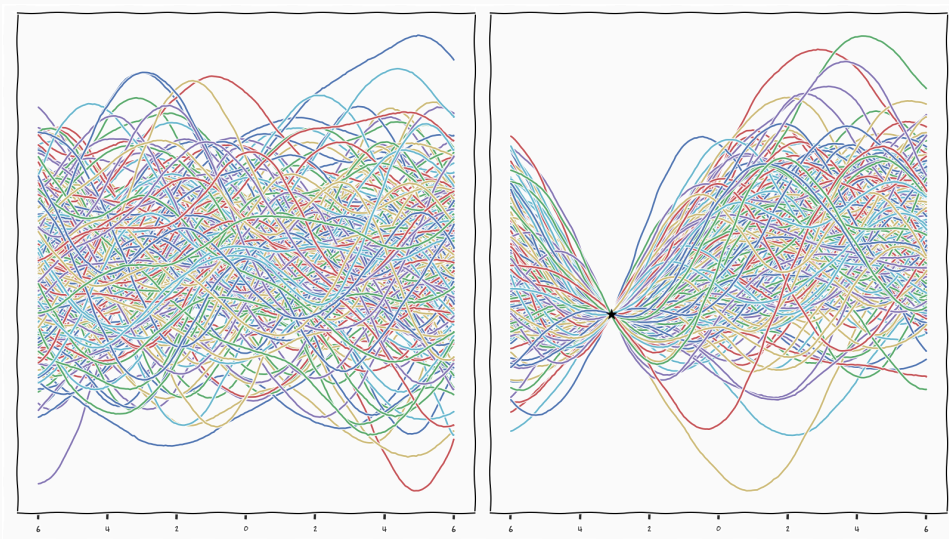
$$\int p(\mathbf{f}, \mathbf{f}_*) d\mathbf{f}_* = \int p(\mathbf{f} | \mathbf{f}_*) p(\mathbf{f}_*) d\mathbf{f}_*$$

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

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

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

Gaussian Processes: Posterior Samples



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

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

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

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

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

- 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 $p(\mathbf{f}_* | \mathbf{f})$ is Gaussian

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

- 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 $p(\mathbf{f}_* | \mathbf{f})$ is Gaussian
- $\Rightarrow$ We can just solve for $p(\mathbf{f}_* | \mathbf{f})$

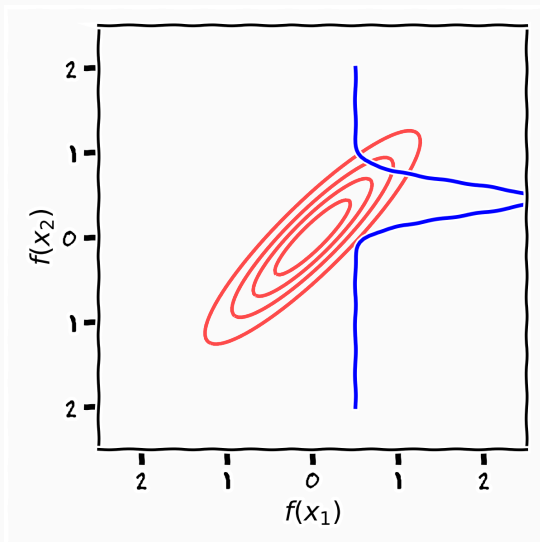
- All instantiations are jointly Gaussian

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

- Conditional Gaussian

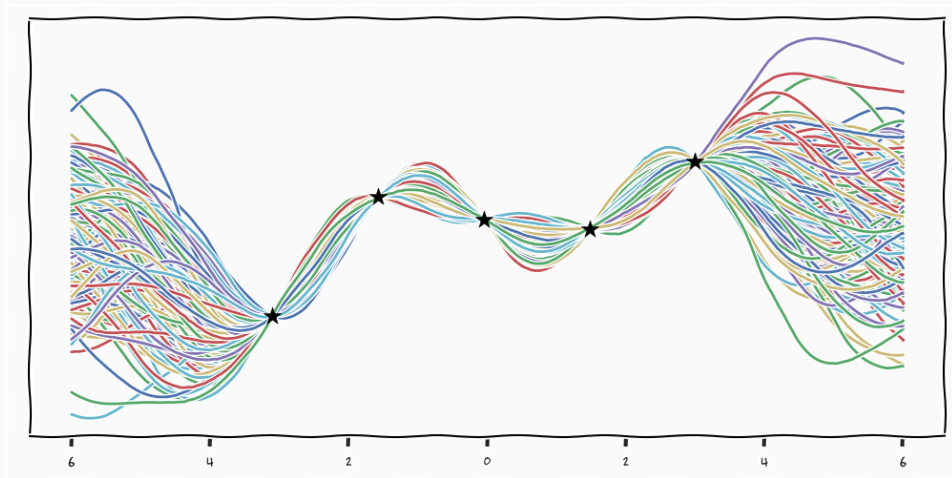
$$p(f_* | \mathbf{x}_*, \mathbf{x}, \mathbf{f}) = \mathcal{N}(k(\mathbf{x}_*, \mathbf{x})^\top k(\mathbf{x}, \mathbf{x})^{-1} \mathbf{f}, \\ k(\mathbf{x}_*, \mathbf{x}_*) - k(\mathbf{x}_*, \mathbf{x})^\top k(\mathbf{x}, \mathbf{x})^{-1} k(\mathbf{x}, \mathbf{x}_*))$$

Conditional Gaussians

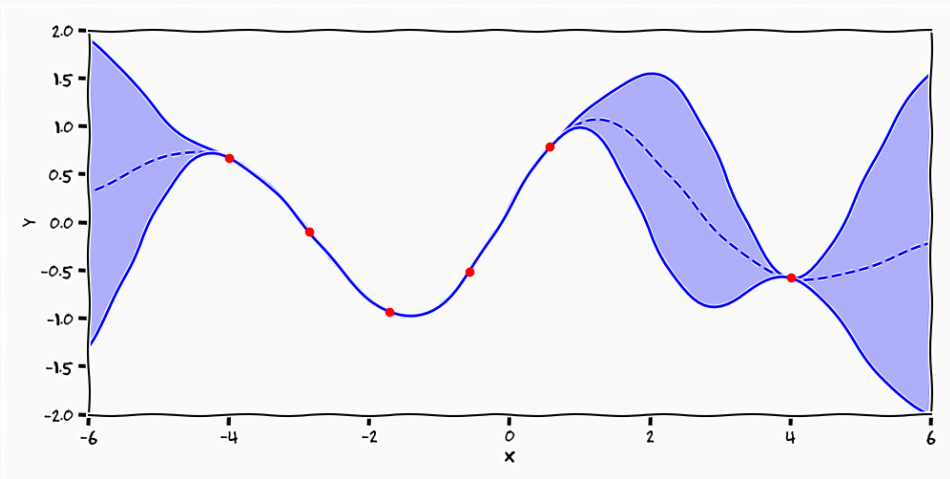


Gaussian Processes: "Predictive Posterior Samples"

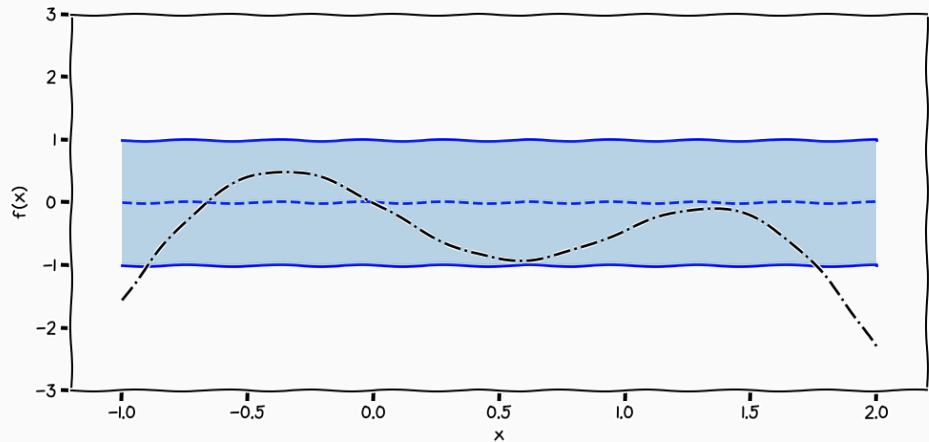
l



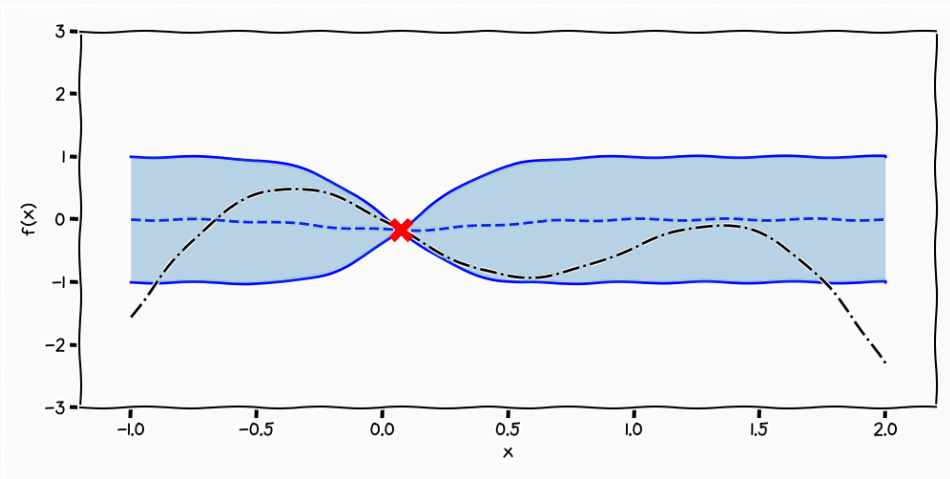
Gaussian Processes: "Predictive Posterior Process"



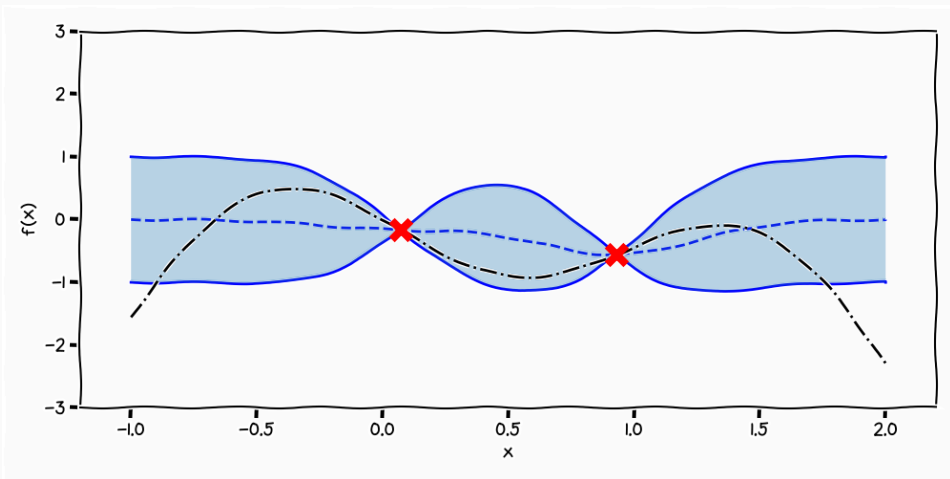
Posterior Processes



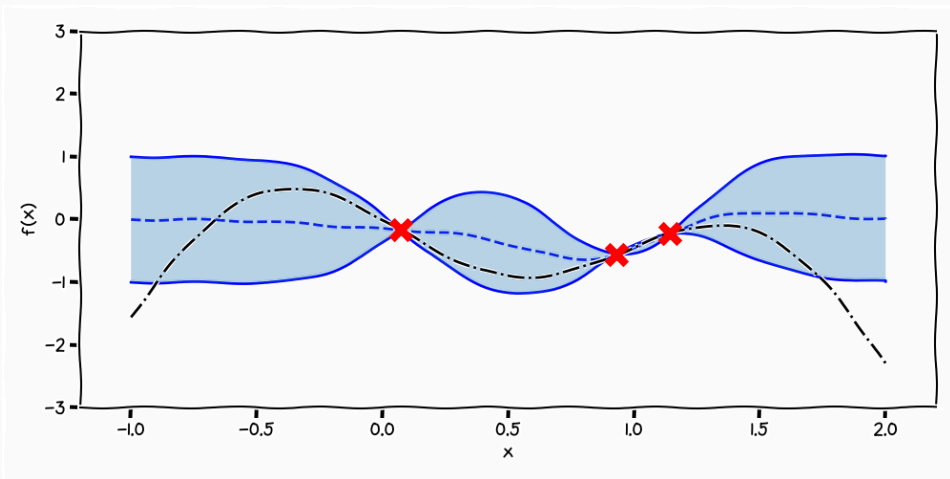
Posterior Processes



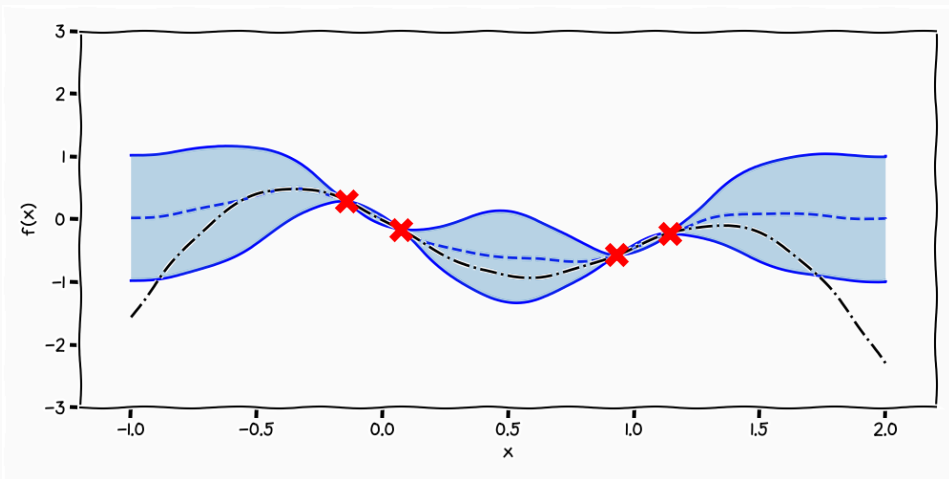
Posterior Processes



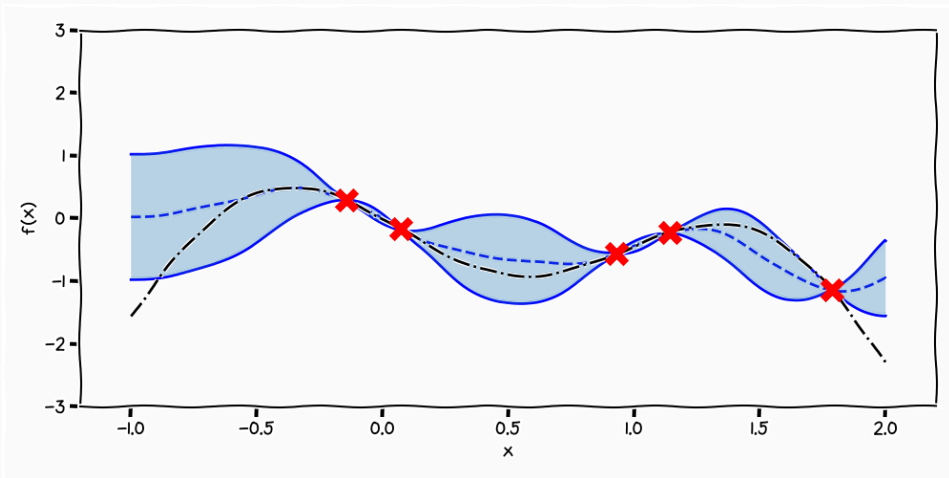
Posterior Processes



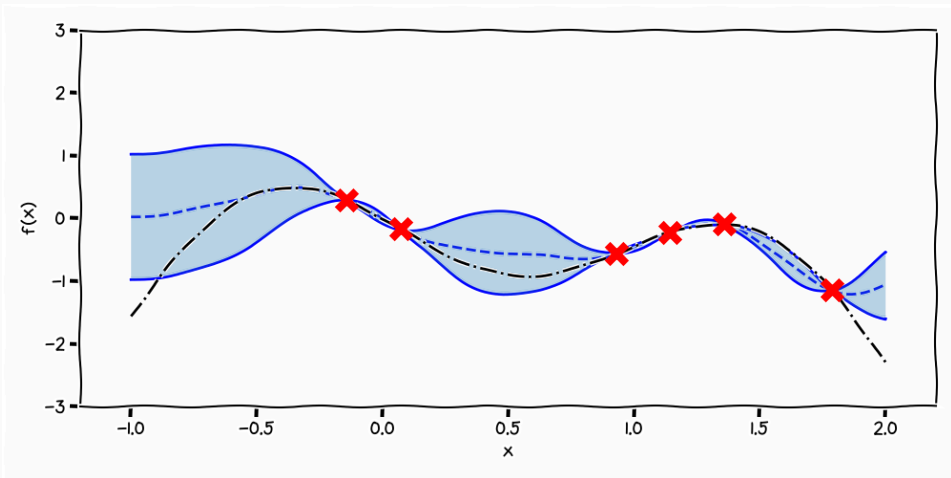
Posterior Processes



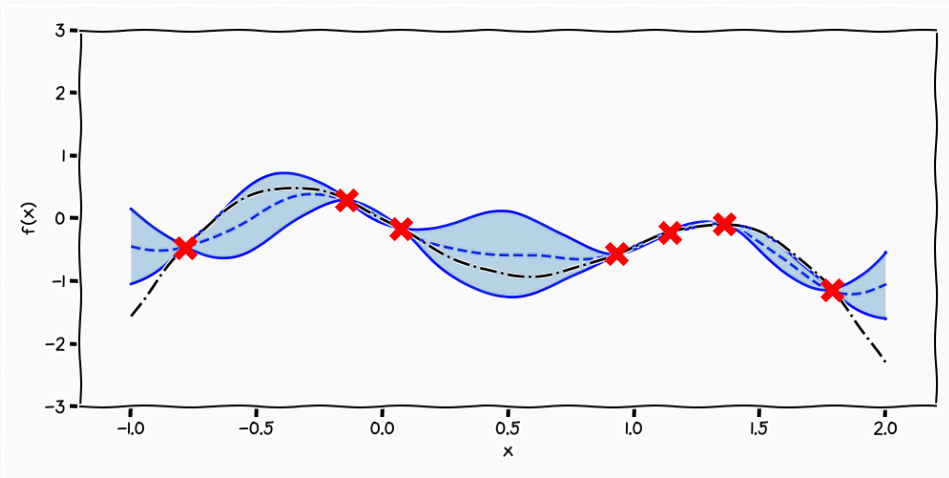
Posterior Processes



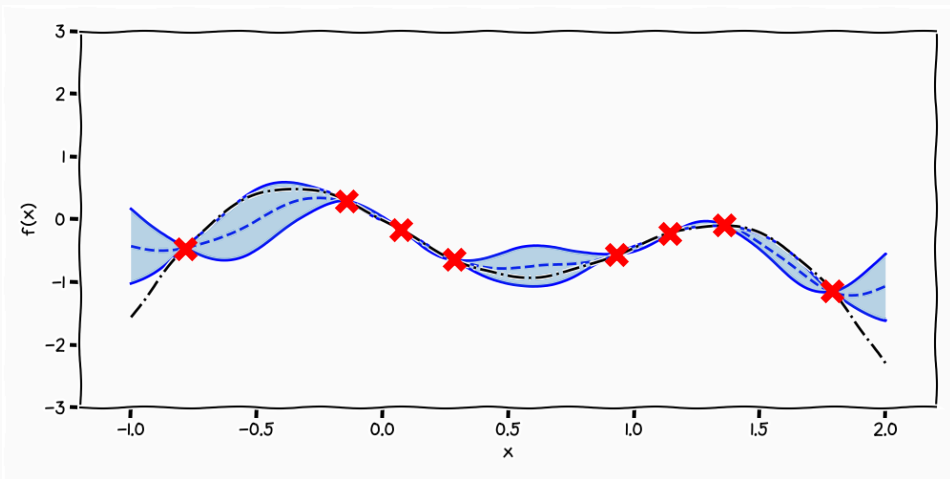
Posterior Processes



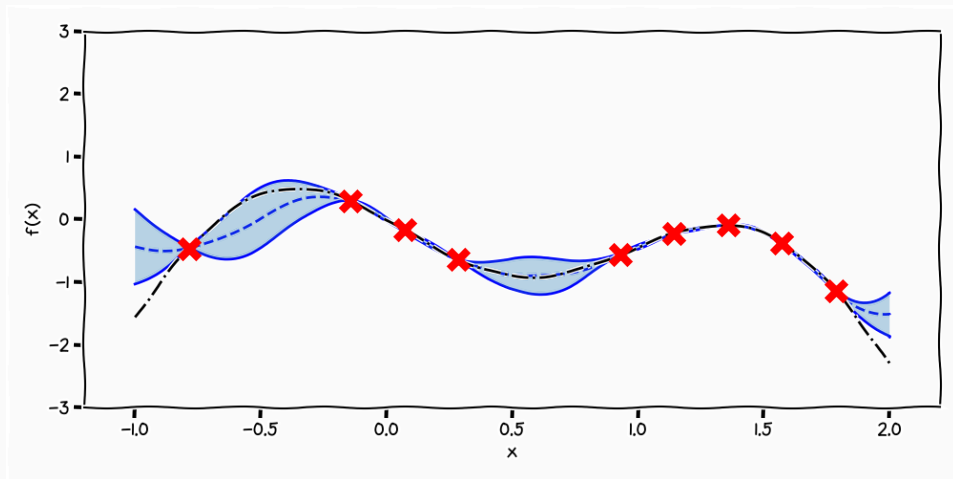
Posterior Processes



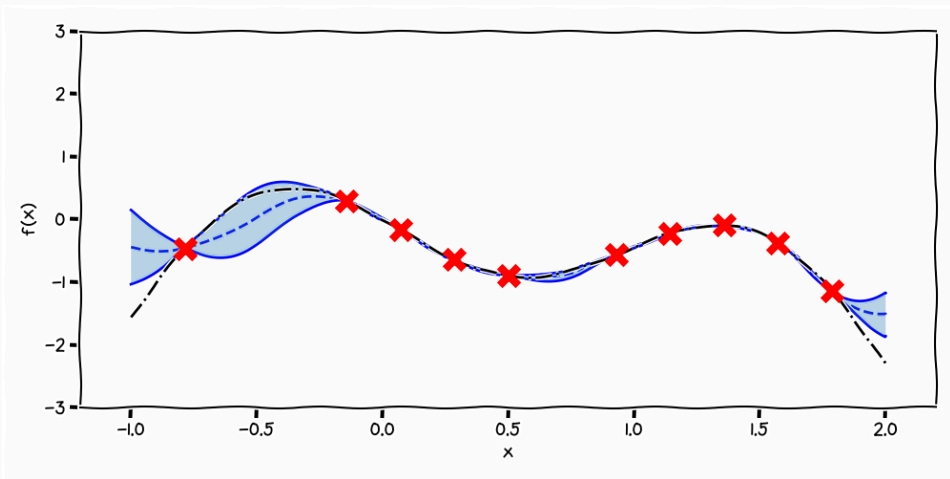
Posterior Processes



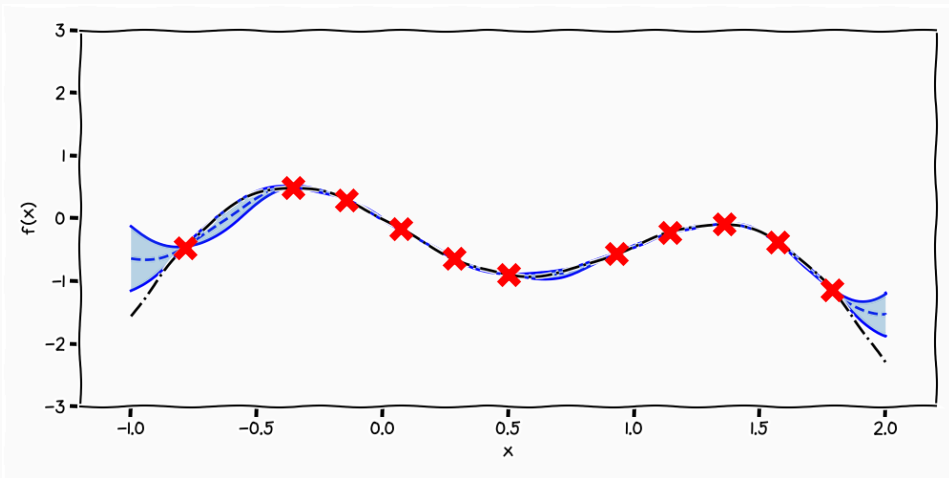
Posterior Processes



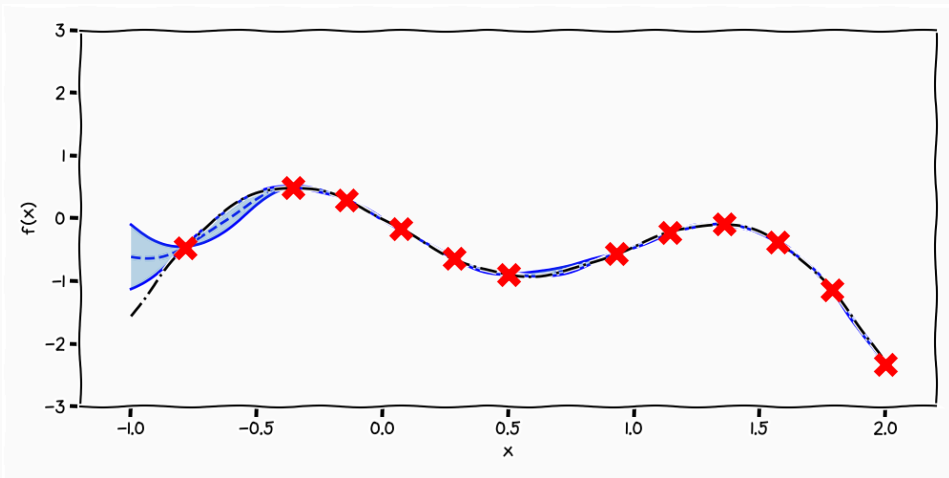
Posterior Processes



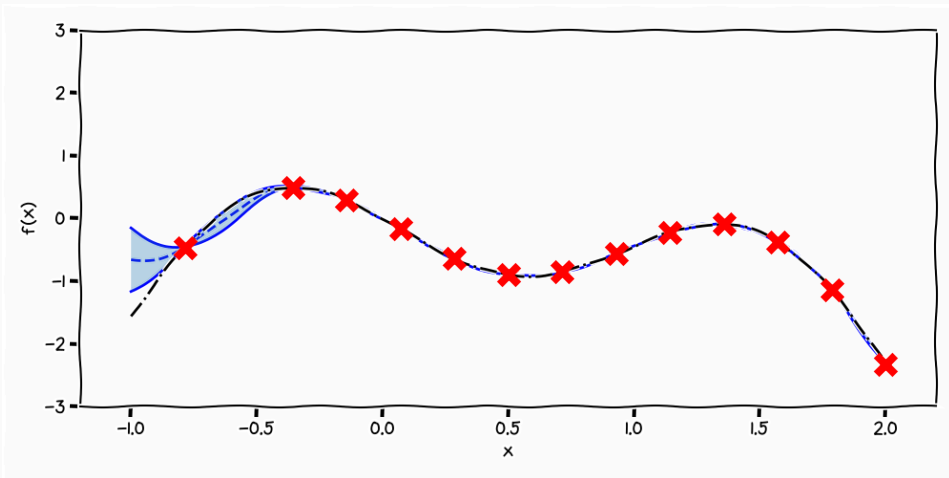
Posterior Processes



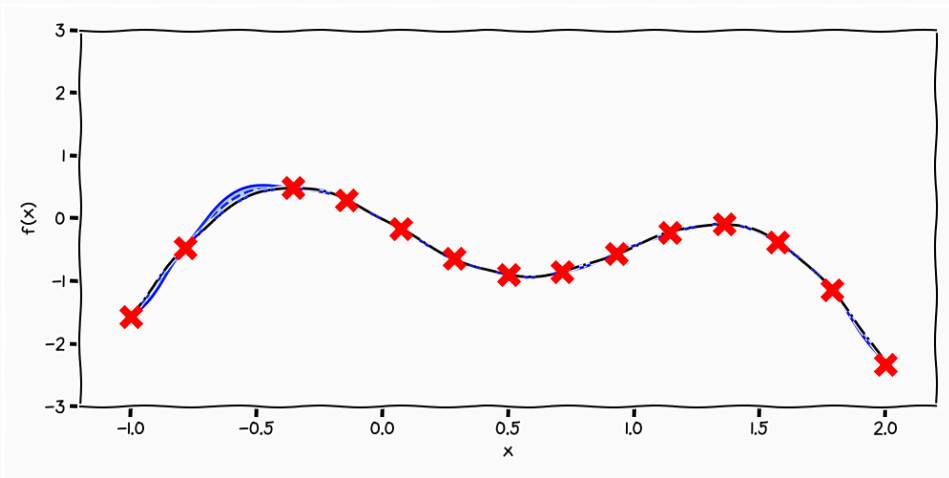
Posterior Processes



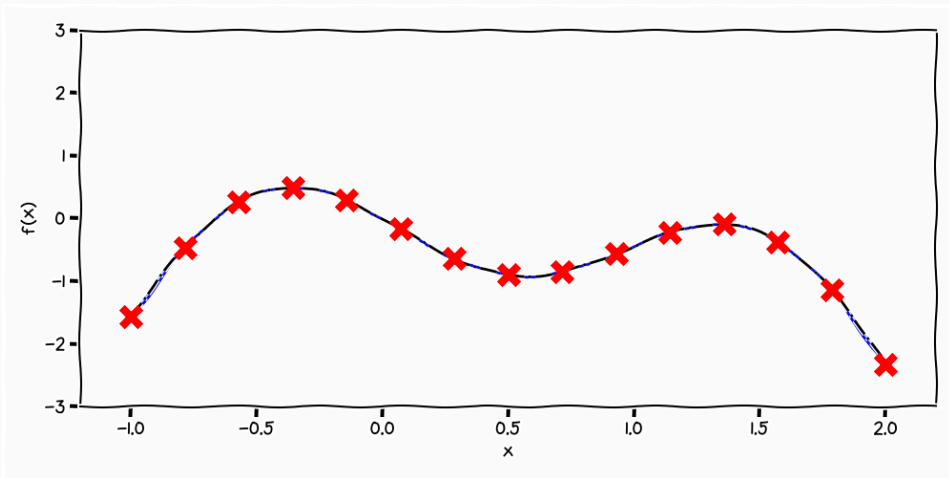
Posterior Processes



Posterior Processes



Posterior Processes



- 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

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

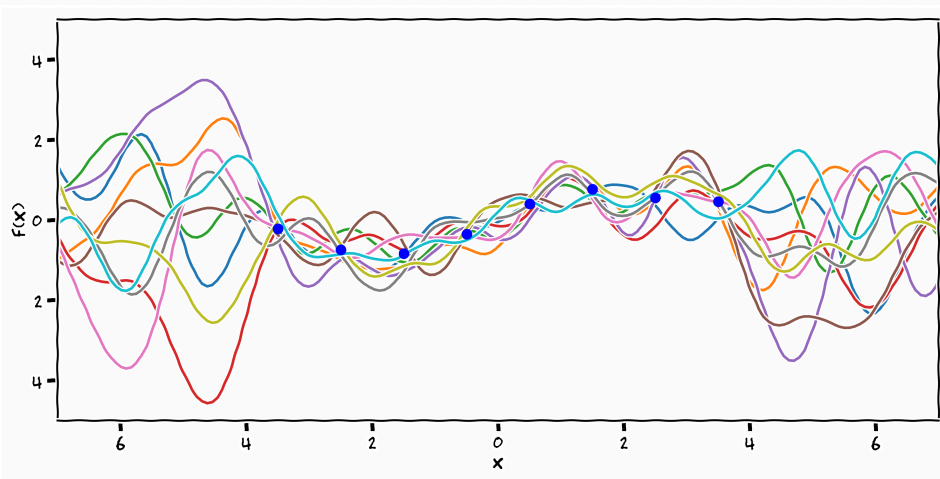
$$p(\mathbf{y} \mid \mathbf{f}) = \mathcal{N}(\mathbf{y} \mid \mathbf{f}, \sigma^2 \mathbf{I})$$

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

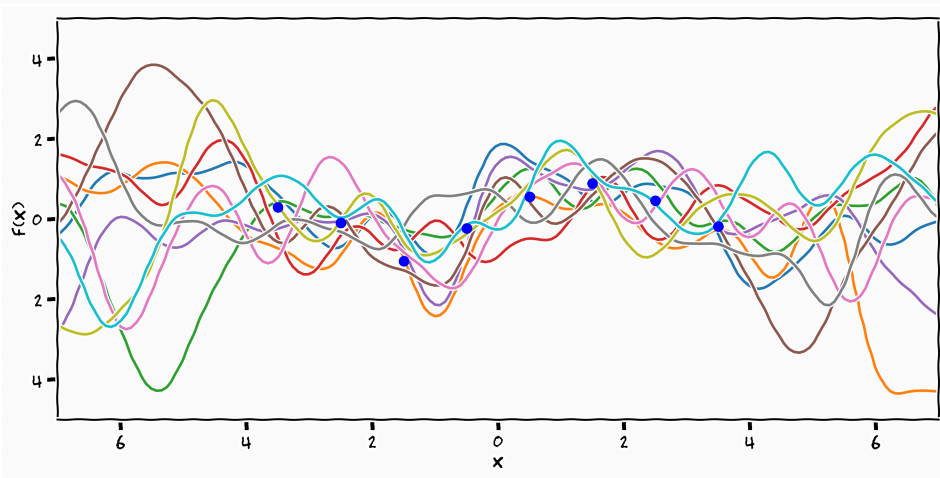
$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} k(\mathbf{x}, \mathbf{x}) + \sigma^2 \mathbf{I} & k(\mathbf{x}, \mathbf{x}_*) \\ k(\mathbf{x}_*, \mathbf{x}) & k(\mathbf{x}_*, \mathbf{x}_*) \end{bmatrix}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

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

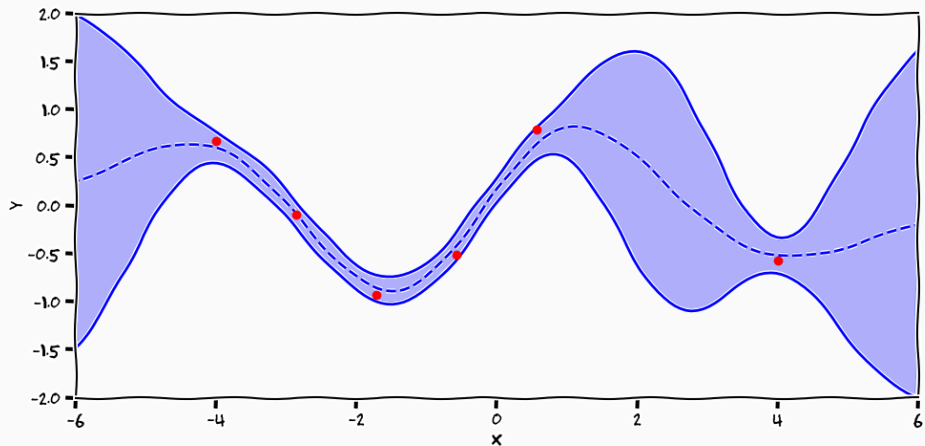
Posterior Samples



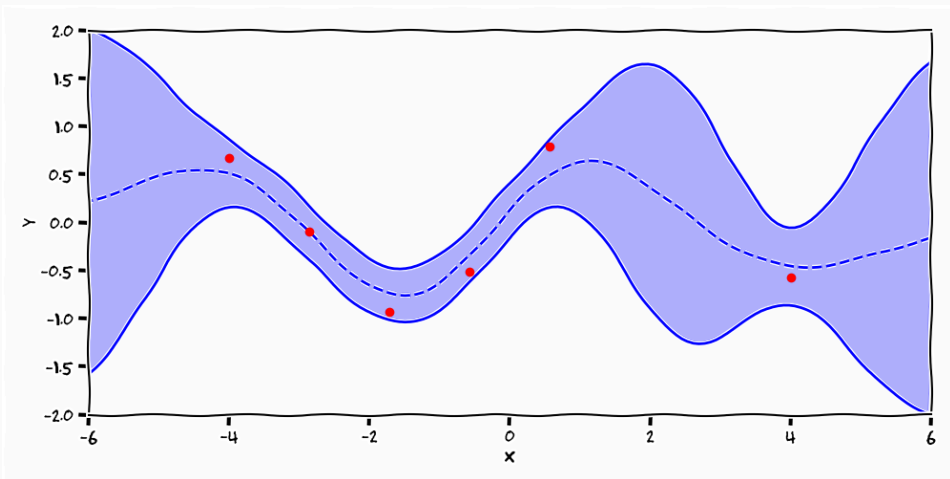
Posterior Samples



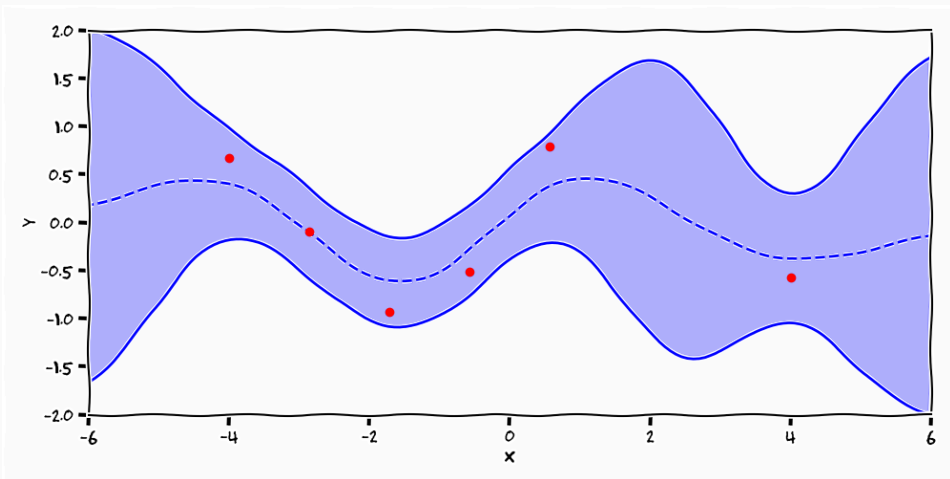
Gaussian Processes: Predictive Posterior Process



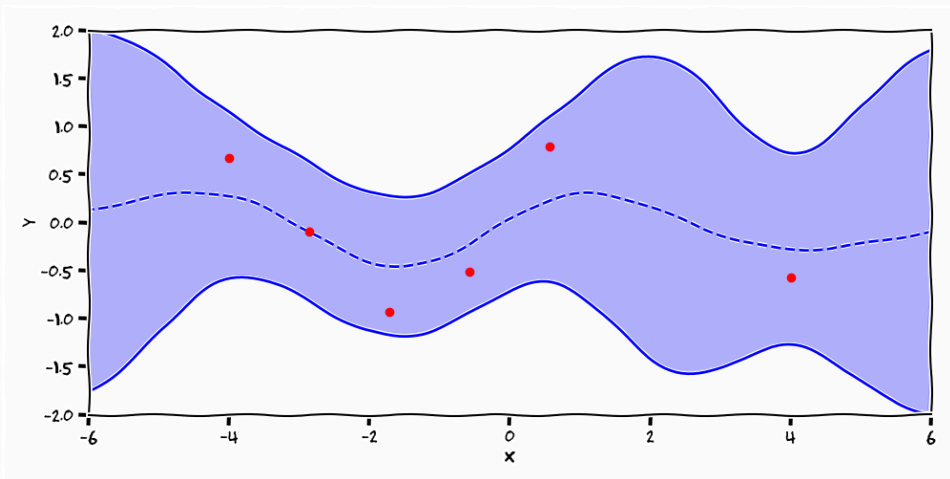
Gaussian Processes: Predictive Posterior Process



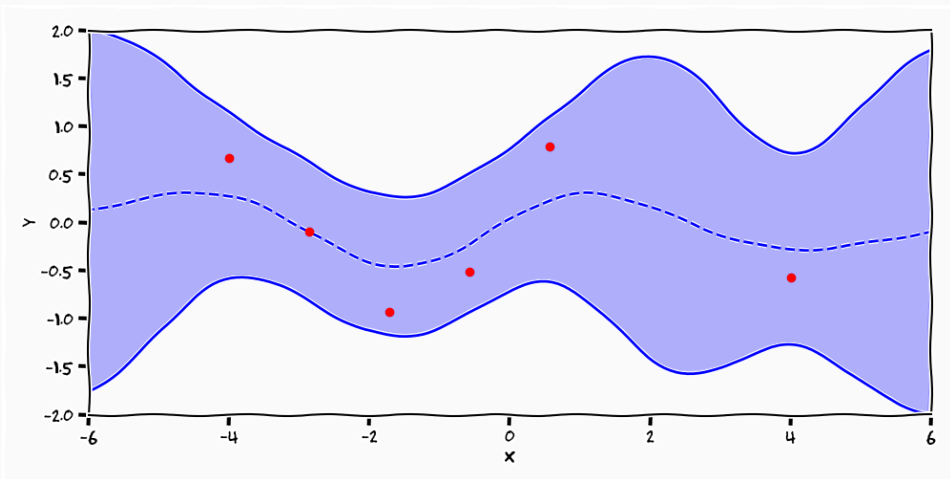
Gaussian Processes: Predictive Posterior Process



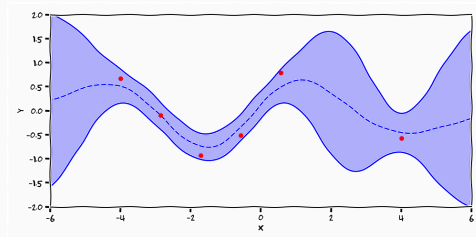
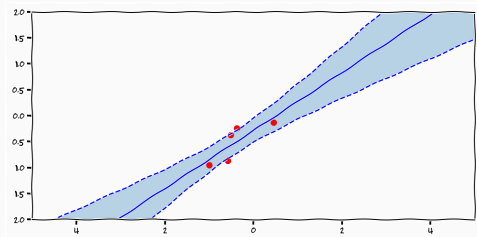
Gaussian Processes: Predictive Posterior Process



Gaussian Processes: Predictive Posterior Process



Two views



$$\mu_*^{\text{GP}} = k(\mathbf{x}_*, \mathbf{x}) (\sigma^2 \mathbf{I} + k(\mathbf{x}, \mathbf{x}))^{-1} \mathbf{y}$$

$$\mu_*^{\text{Lin}} = \mathbf{x}_*^{\text{T}} \mathbf{x} (\beta(\alpha \mathbf{I} + \beta \mathbf{x}^{\text{T}} \mathbf{x}))^{-1} \mathbf{y}$$

- both means are linear with respect to the output data

¹or use a kernel $k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^{\text{T}} \mathbf{x}$

$$\mu_*^{\text{GP}} = k(\mathbf{x}_*, \mathbf{x}) (\sigma^2 \mathbf{I} + k(\mathbf{x}, \mathbf{x}))^{-1} \mathbf{y}$$

$$\mu_*^{\text{Lin}} = \mathbf{x}_*^T \mathbf{x} (\beta(\alpha \mathbf{I} + \beta \mathbf{x}^T \mathbf{x}))^{-1} \mathbf{y}$$

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

¹or use a kernel $k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T \mathbf{x}$

$$\mu_*^{\text{GP}} = k(\mathbf{x}_*, \mathbf{x}) (\sigma^2 \mathbf{I} + k(\mathbf{x}, \mathbf{x}))^{-1} \mathbf{y}$$

$$\mu_*^{\text{Lin}} = \mathbf{x}_*^T \mathbf{x} (\beta(\alpha \mathbf{I} + \beta \mathbf{x}^T \mathbf{x}))^{-1} \mathbf{y}$$

- both means are linear with respect to the output data
- how about if we take a basis function approach such that $\mathbf{x}_*^T \mathbf{x} = \Phi(\mathbf{x}_*)^T \Phi(\mathbf{x})$
- if we think of $\mathbf{x}$ as the center of each basis function¹

¹or use a kernel $k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T \mathbf{x}$

$$\mu_*^{\text{GP}} = k(\mathbf{x}_*, \mathbf{x}) (\sigma^2 \mathbf{I} + k(\mathbf{x}, \mathbf{x}))^{-1} \mathbf{y}$$

$$\mu_*^{\text{Lin}} = \mathbf{x}_*^T \mathbf{x} (\beta(\alpha \mathbf{I} + \beta \mathbf{x}^T \mathbf{x}))^{-1} \mathbf{y}$$

- both means are linear with respect to the output data
- how about if we take a basis function approach such that $\mathbf{x}_*^T \mathbf{x} = \Phi(\mathbf{x}_*)^T \Phi(\mathbf{x})$
- if we think of $\mathbf{x}$ as the center of each basis function¹
 - a basis function per data point

¹or use a kernel $k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T \mathbf{x}$

$$\mu_*^{\text{GP}} = k(\mathbf{x}_*, \mathbf{x}) (\sigma^2 \mathbf{I} + k(\mathbf{x}, \mathbf{x}))^{-1} \mathbf{y}$$

$$\mu_*^{\text{Lin}} = \mathbf{x}_*^T \mathbf{x} (\beta(\alpha \mathbf{I} + \beta \mathbf{x}^T \mathbf{x}))^{-1} \mathbf{y}$$

- both means are linear with respect to the output data
- how about if we take a basis function approach such that $\mathbf{x}_*^T \mathbf{x} = \Phi(\mathbf{x}_*)^T \Phi(\mathbf{x})$
- if we think of $\mathbf{x}$ as the center of each basis function¹
 - a basis function per data point
- if we could parametrise $\Phi(\mathbf{x}_*)^T \Phi(\mathbf{x}) = k(\mathbf{x}_*, \mathbf{x})$ as a function

¹or use a kernel $k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T \mathbf{x}$

$$\mu_*^{\text{GP}} = k(\mathbf{x}_*, \mathbf{x}) (\sigma^2 \mathbf{I} + k(\mathbf{x}, \mathbf{x}))^{-1} \mathbf{y}$$

$$\mu_*^{\text{Lin}} = \mathbf{x}_*^T \mathbf{x} (\beta(\alpha \mathbf{I} + \beta \mathbf{x}^T \mathbf{x}))^{-1} \mathbf{y}$$

- both means are linear with respect to the output data
- how about if we take a basis function approach such that $\mathbf{x}_*^T \mathbf{x} = \Phi(\mathbf{x}_*)^T \Phi(\mathbf{x})$
- if we think of $\mathbf{x}$ as the center of each basis function¹
 - a basis function per data point
- if we could parametrise $\Phi(\mathbf{x}_*)^T \Phi(\mathbf{x}) = k(\mathbf{x}_*, \mathbf{x})$ as a function
- this leads to the interpretation of GPs as infinite basis functions

¹or use a kernel $k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T \mathbf{x}$

Summary

- Non-parametrics
 - parametrise relationship between data

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

Kolmogorov's Extension Theorem

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

1. Exchangeable

$$\nu_{t_{\pi(1)} \dots t_{\pi(k)}} (F_{\pi(1)} \times \dots \times F_{\pi(k)}) = \nu_{t_1 \dots t_k} (F_1 \times \dots \times F_k)$$

2. Marginal

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

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

eof

