

Machine Learning

Gaussian Processes

Carl Henrik Ek - carlhenrik.ek@bristol.ac.uk

October 21, 2019

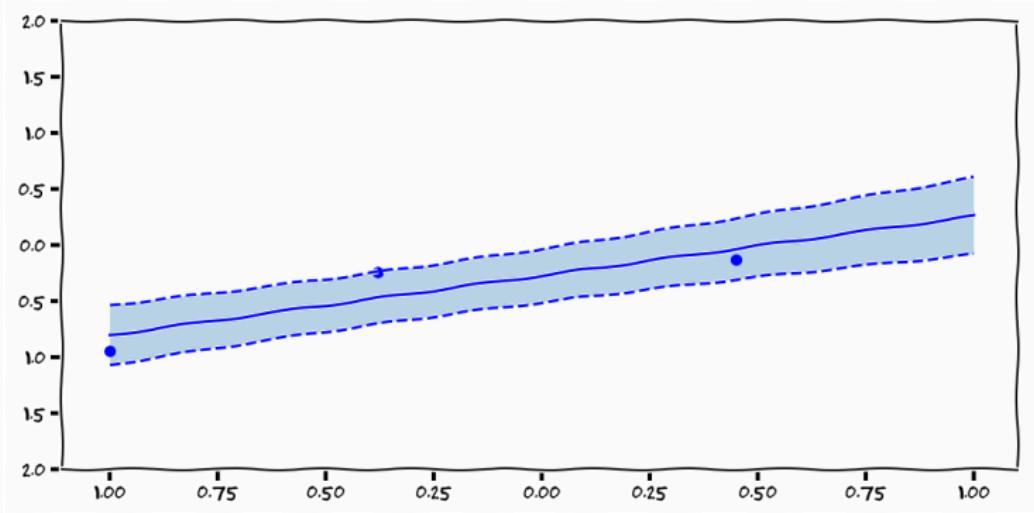
<http://carlhenrik.com>

Introduction

Linear Regression

$$\begin{aligned} p(\mathbf{t}|\mathbf{w}, \mathbf{x}) &= \prod_n^N p(t_n|\mathbf{w}, \mathbf{x}) = \prod_n^N \mathcal{N}(t_n|\mathbf{w}^T \mathbf{x}_n, \sigma^2 \mathbf{I}) \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{0}, \tau^2 \mathbf{I}) \\ p(\mathbf{w}|\mathbf{t}, \mathbf{x}) &= \frac{p(\mathbf{t}|\mathbf{w}, \mathbf{x})p(\mathbf{w})}{p(\mathbf{t})} \\ &\propto p(\mathbf{t}|\mathbf{w}, \mathbf{x})p(\mathbf{w}) \end{aligned}$$

Predictive Posterior



$$p(t_* | \mathbf{x}_*, \mathbf{X}, \mathbf{t}, \beta, \alpha) = \int p(t_* | \mathbf{x}_*, \mathbf{w}, \beta) p(\mathbf{w} | \mathbf{t}, \mathbf{X}, \alpha, \beta) d\mathbf{w}$$

Dual Linear Regression

$$\begin{aligned}\hat{\mathbf{w}} &= \operatorname{argmin}_{\mathbf{w}} -\log p(\mathbf{w}|\mathbf{t}, \mathbf{x}) \\ &= \frac{1}{2}(\mathbf{X}\mathbf{w} - \mathbf{t})^T(\mathbf{X}\mathbf{w} - \mathbf{t}) + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w} \\ \hat{\mathbf{w}} &= \mathbf{X}^T (\phi(\mathbf{X})^T \phi(\mathbf{X}) + \lambda \mathbf{I})^{-1} \mathbf{t}\end{aligned}$$

Linear Regression

Primal - explicit representation

$$\mathbf{y}(\mathbf{x}_*) = \mathbf{w}^T \phi(\mathbf{x}_*)$$

Dual - implicit representation

$$\mathbf{y}(\mathbf{x}_*) = \mathbf{w}^T \phi(\mathbf{x}_*) = k(\mathbf{x}_*, \mathbf{X})(\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{t}$$

$$k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^T \phi(\mathbf{x}')$$

Kernels

$$k(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$$

- kernel functions describe inner-products in an **induced** representation
- induced representation lives in what is called a **Hilbert Space**
- importantly the space is metric
- *we never need to know the mapping, only the inner-product*

What have we actually done

- Primal
 - See data
 - Encode relationship between variates using parameters w
 - Make predictions using w

What have we actually done

- Primal
 - See data
 - Encode relationship between variates using parameters w
 - Make predictions using w
- Dual
 - See Data
 - Encode relationship between variates using variates themselves

What have we actually done

- Primal
 - See data
 - Encode relationship between variates using parameters w
 - Make predictions using w
- Dual
 - See Data
 - Encode relationship between variates using variates themselves
 - *Model complexity depends on data*

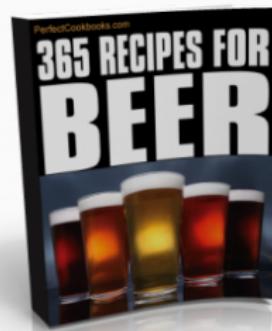
What have we actually done

- Primal
 - See data
 - Encode relationship between variates using parameters w
 - Make predictions using w
- Dual
 - See Data
 - Encode relationship between variates using variates themselves
 - *Model complexity depends on data*
 - Non-parametric model

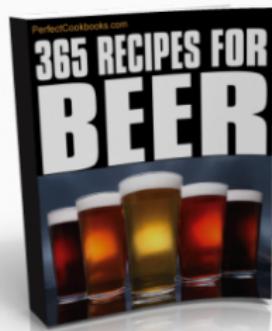
Non-parametrics



Non-parametrics



Non-parametrics



Kernels

Euclidean Distance

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 = \mathbf{x}_i^T \mathbf{x}_i - 2\mathbf{x}_i^T \mathbf{x}_j + \mathbf{x}_j^T \mathbf{x}_j$$

Kernelised Euclidean Distance

$$\|\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j)\|^2 = k(\mathbf{x}_i, \mathbf{x}_i) - 2k(\mathbf{x}_i, \mathbf{x}_j) + k(\mathbf{x}_j, \mathbf{x}_j)$$

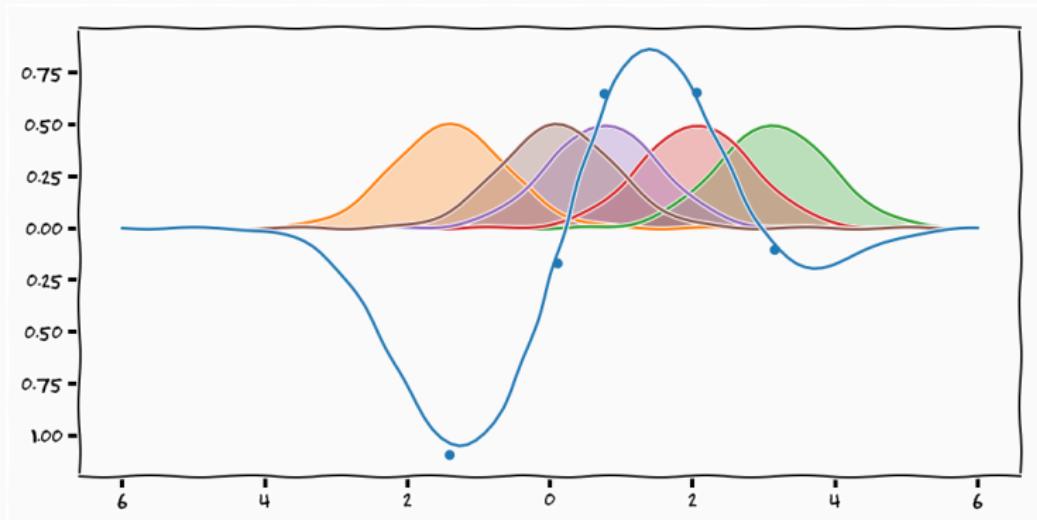
Kernel Example

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sigma^2 e^{-\frac{1}{2\ell^2} (\mathbf{x}_i - \mathbf{x}_j)^T (\mathbf{x}_i - \mathbf{x}_j)}$$

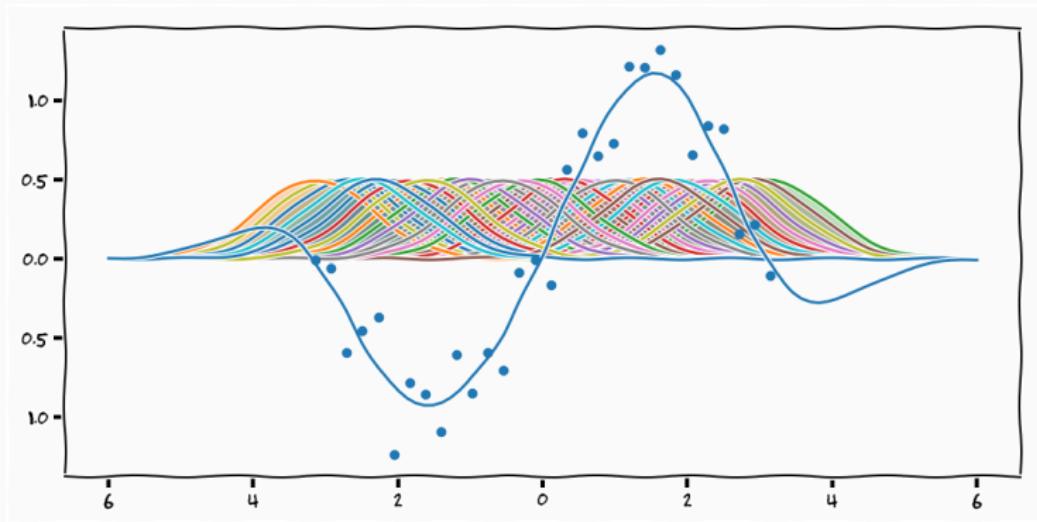
Exponented Quadratic

- How does the data vary along the dimensions spanned by the data
- RBF, Squared Exponential, Exponentiated Quadratic
- Co-variance smoothly decays with distance
- You can build new kernels out of other kernels [1] p. 296

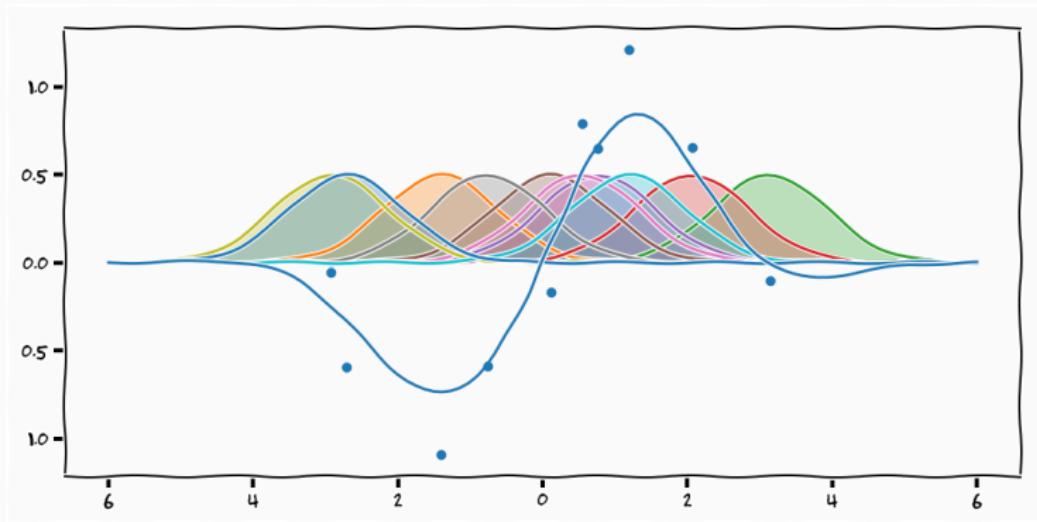
Kernel Regression



Kernel Regression



Kernel Regression



Uncertainty

- We have no uncertainty in our observed outputs

Uncertainty

- We have no uncertainty in our observed outputs
- We have no uncertainty in our mapping

Uncertainty

- We have no uncertainty in our observed outputs
- We have no uncertainty in our mapping
 - Linear, it is a line

Uncertainty

- We have no uncertainty in our observed outputs
- We have no uncertainty in our mapping
 - Linear, it is a line
 - Fixed basis functions hard to interpret

Uncertainty

- We have no uncertainty in our observed outputs
- We have no uncertainty in our mapping
 - Linear, it is a line
 - Fixed basis functions hard to interpret
 - Dual means no uncertainty in function

Uncertainty

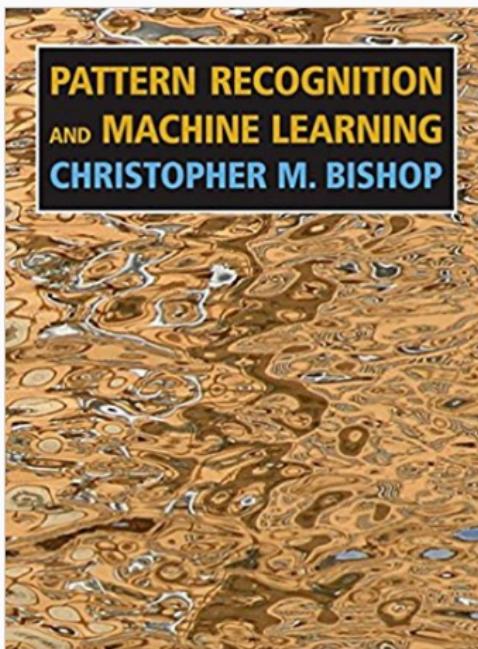
- We have no uncertainty in our observed outputs
- We have no uncertainty in our mapping
 - Linear, it is a line
 - Fixed basis functions hard to interpret
 - Dual means no uncertainty in function
- *need to make assumption over the space of functions*

Uncertainty

- We have no uncertainty in our observed outputs
- We have no uncertainty in our mapping
 - Linear, it is a line
 - Fixed basis functions hard to interpret
 - Dual means no uncertainty in function
- *need to make assumption over the space of functions*
- A distribution over the space of functions

Gaussian Processes

Book

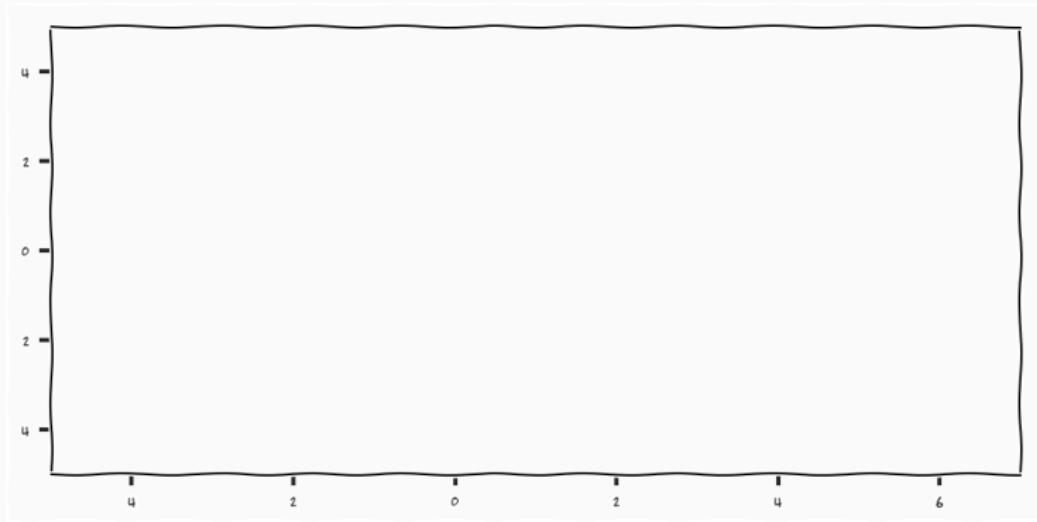




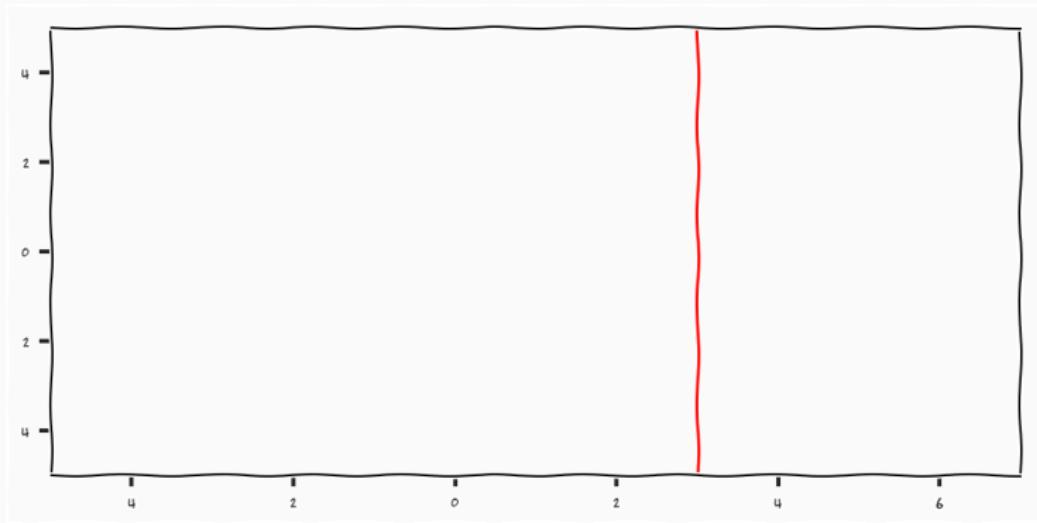
IUDICIUM POSTERIUM DISCIPULUS EST PRIORIS

¹<http://gpss.cc>

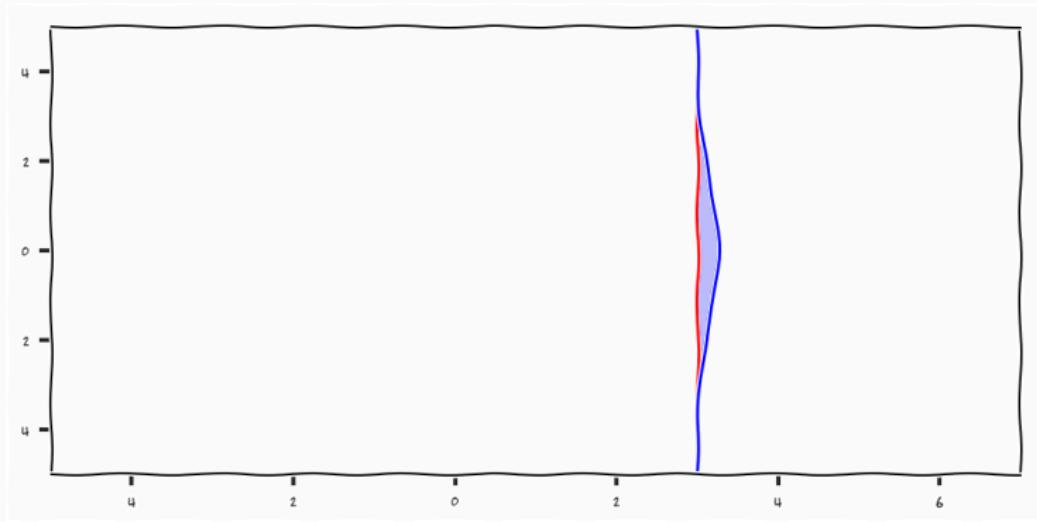
Gaussian Processes



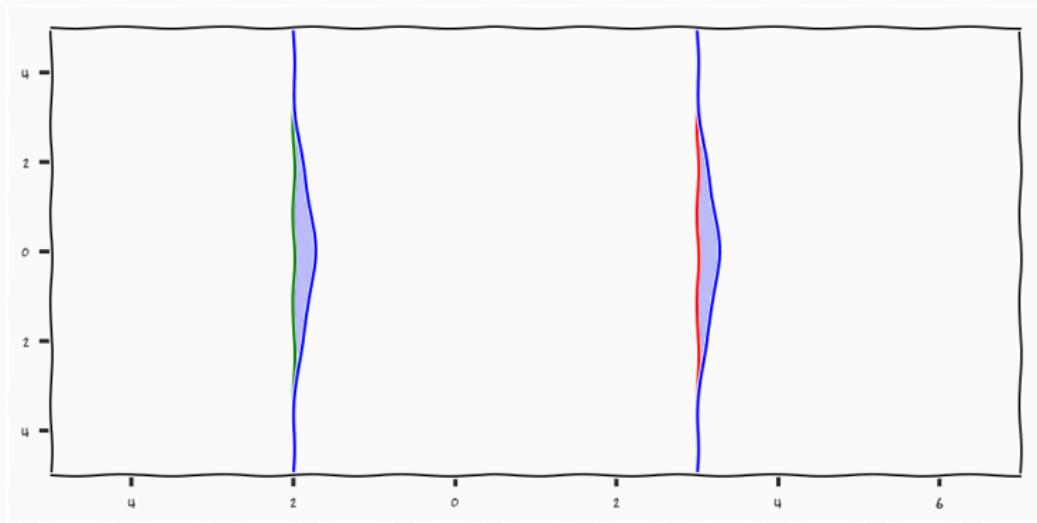
Gaussian Processes



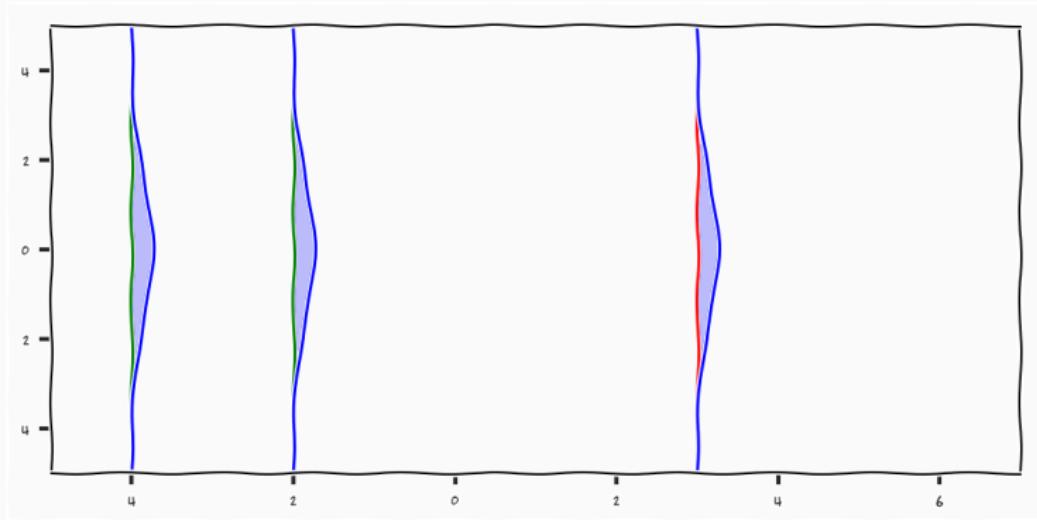
Gaussian Processes



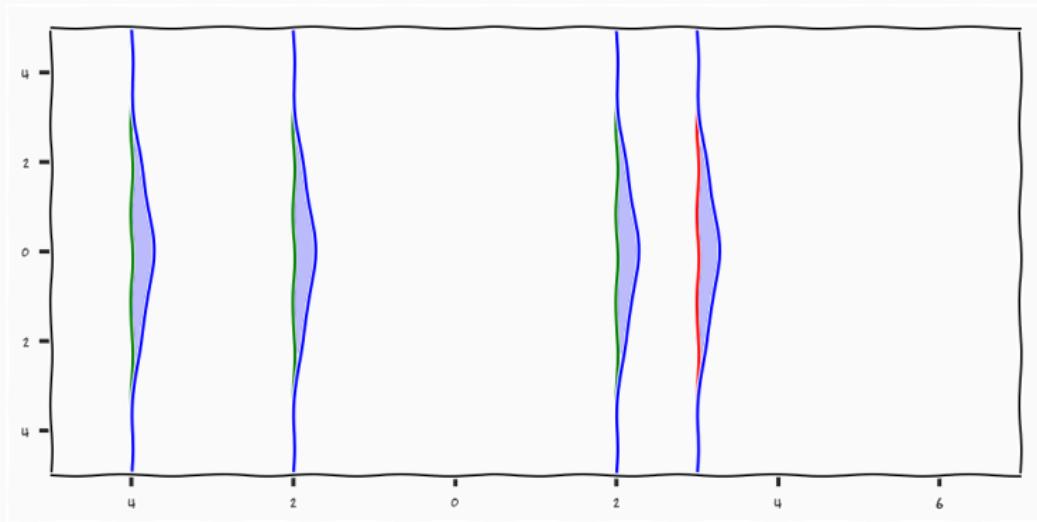
Gaussian Processes



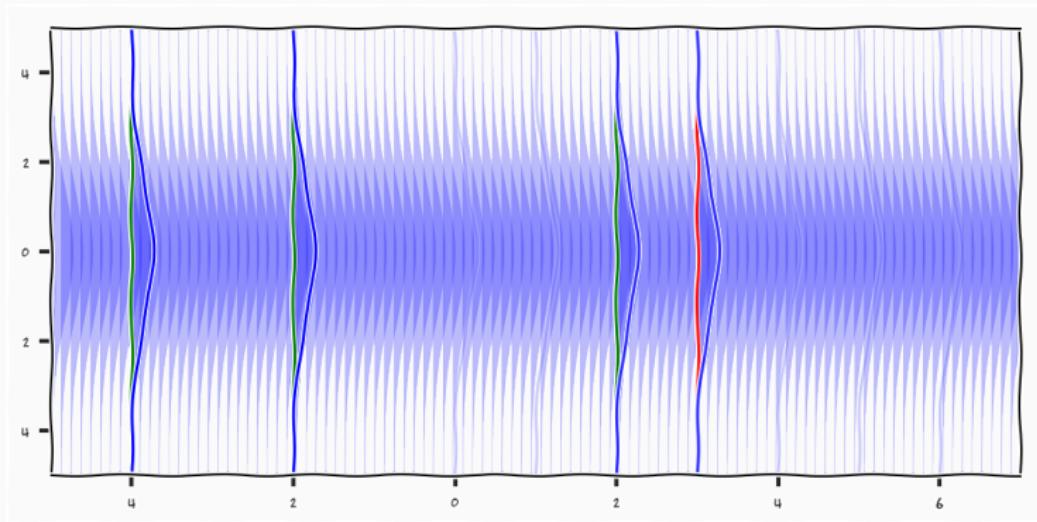
Gaussian Processes



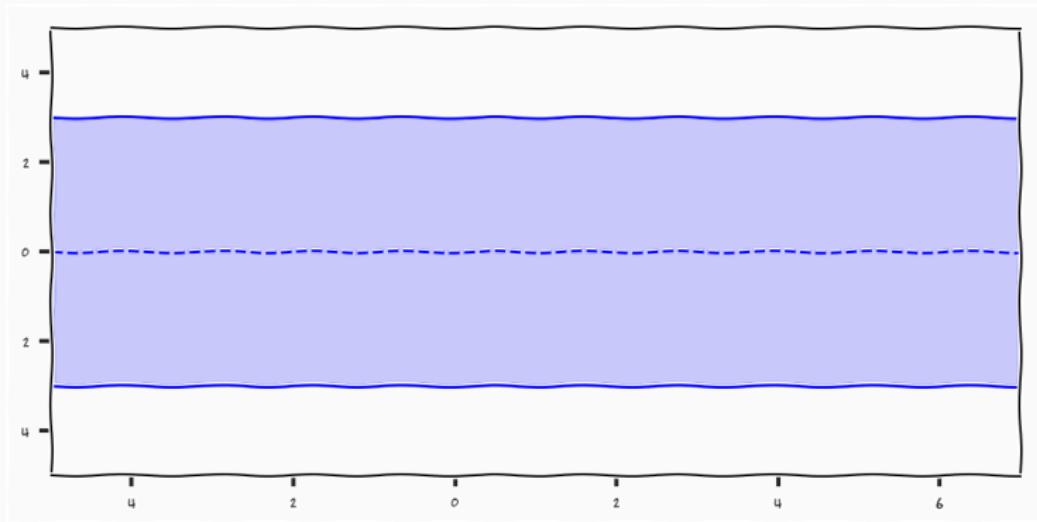
Gaussian Processes



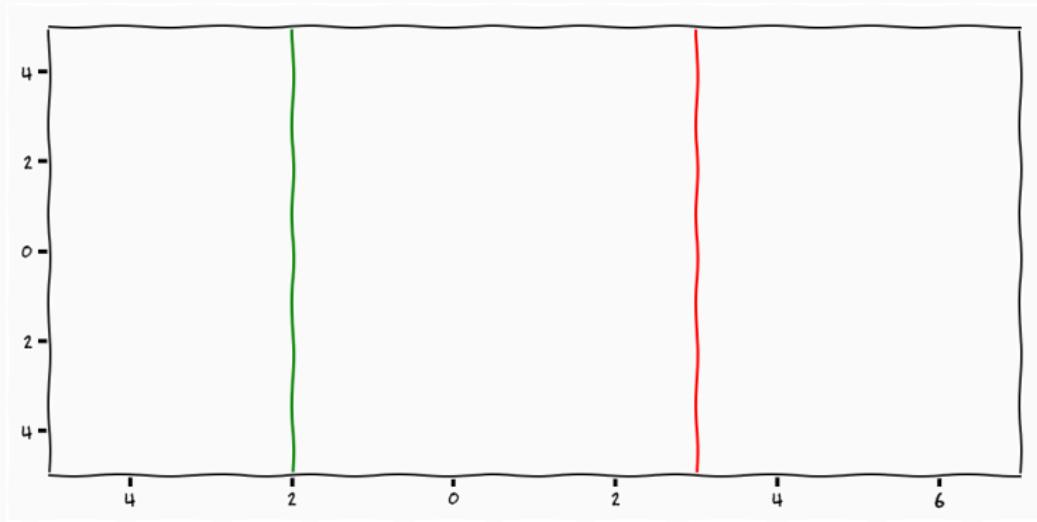
Gaussian Processes



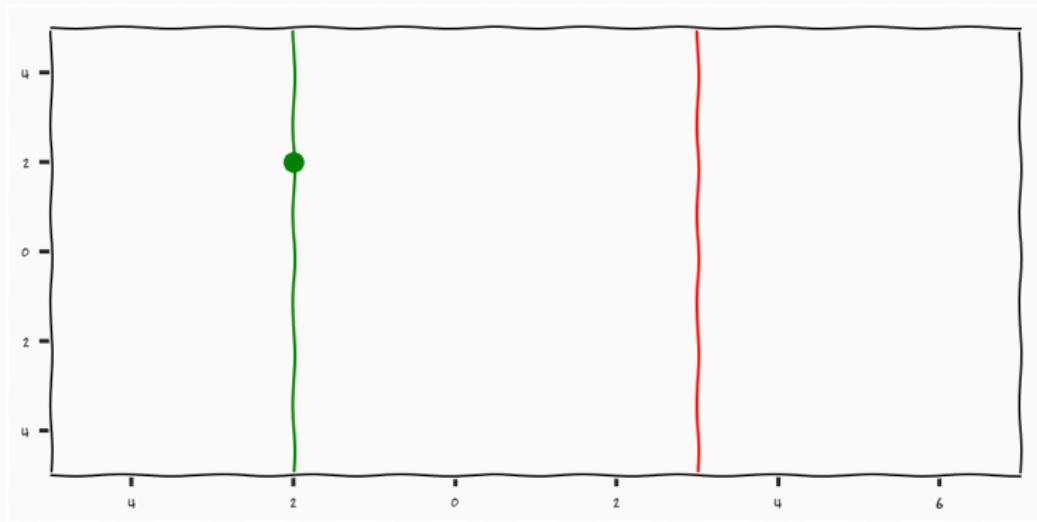
Gaussian Processes



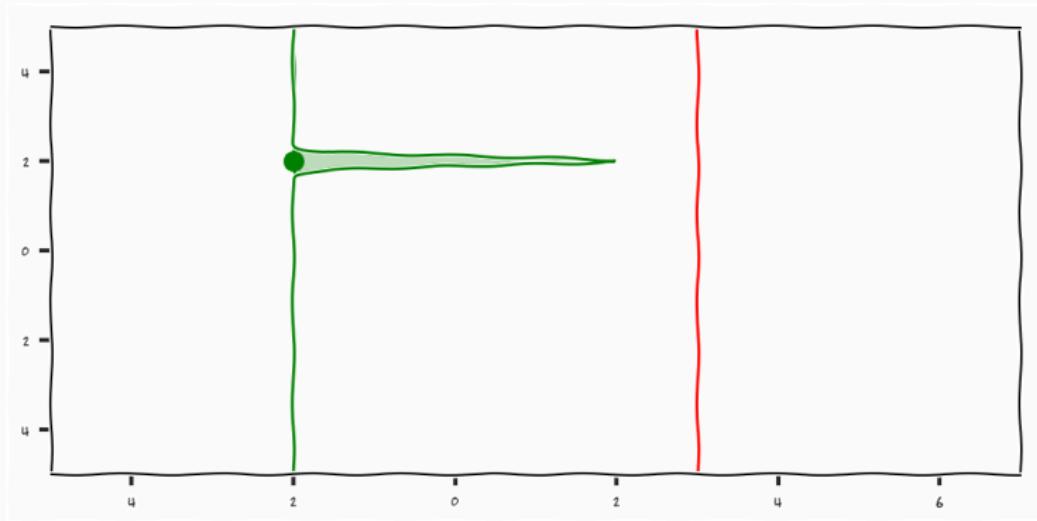
Gaussian Processes



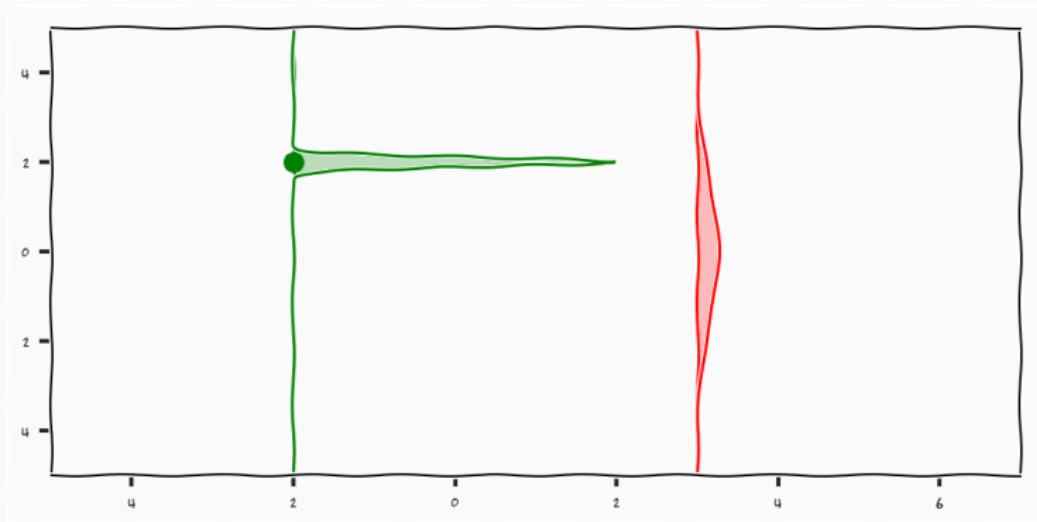
Gaussian Processes



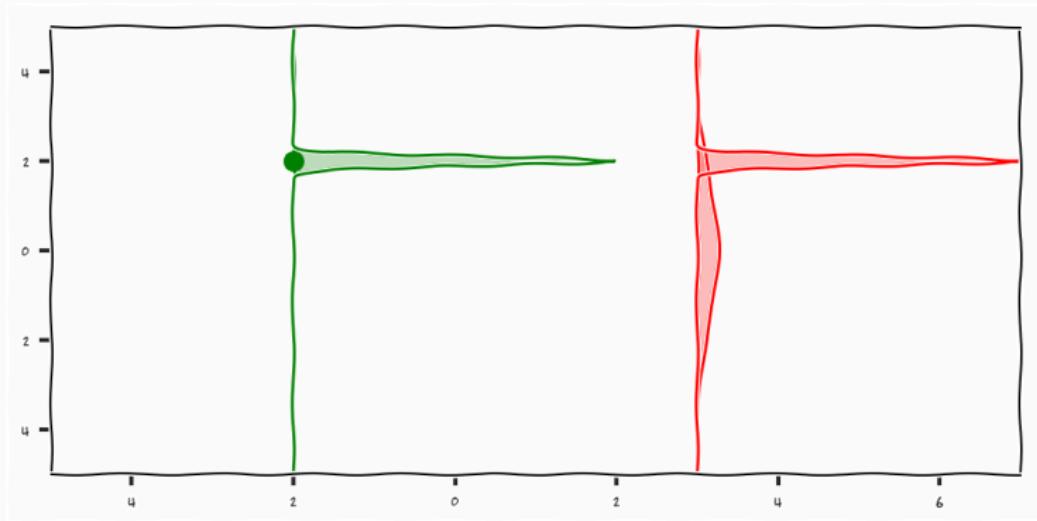
Gaussian Processes



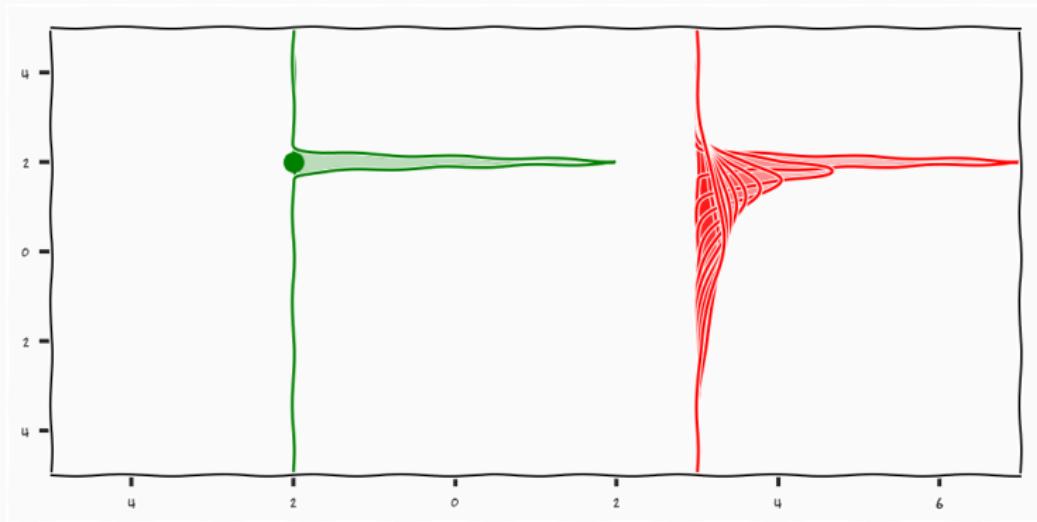
Gaussian Processes



Gaussian Processes



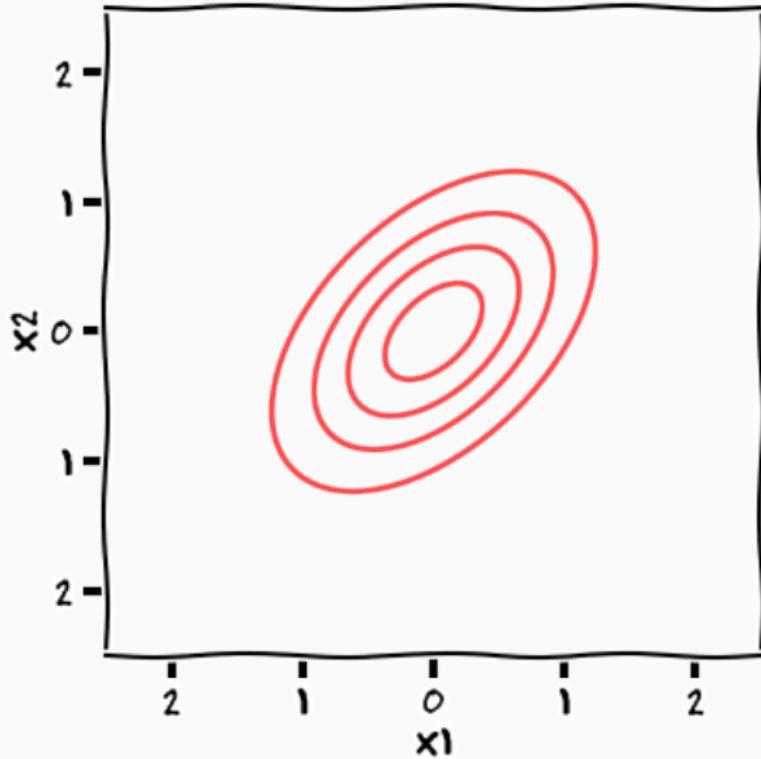
Gaussian Processes



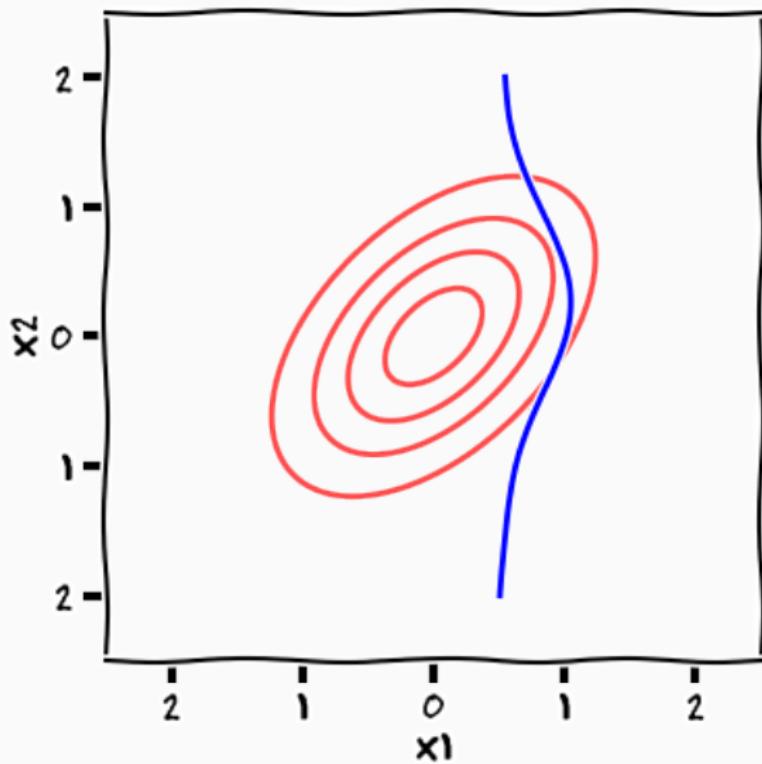
Conditional Gaussians

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

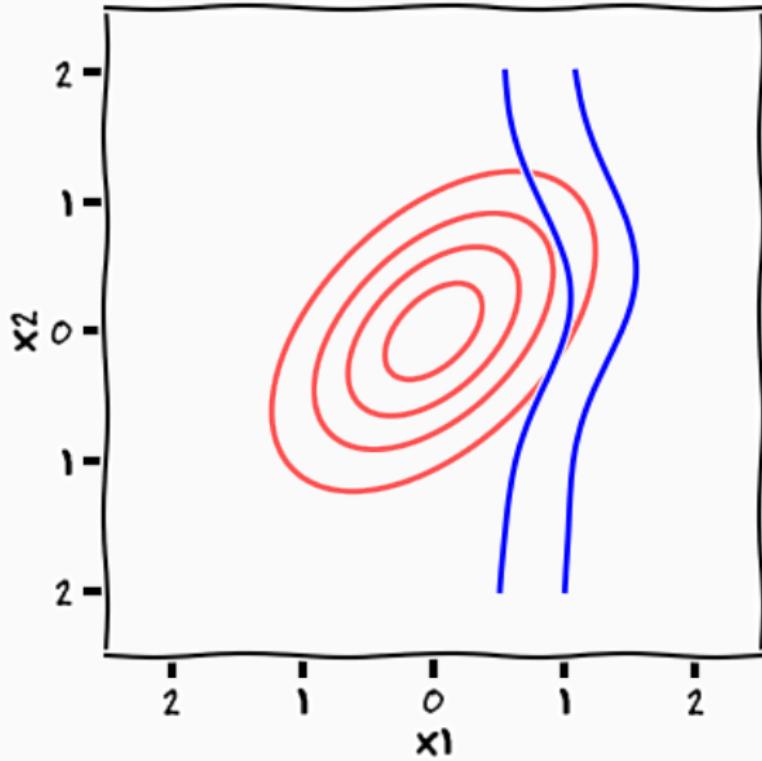
Conditional Gaussians



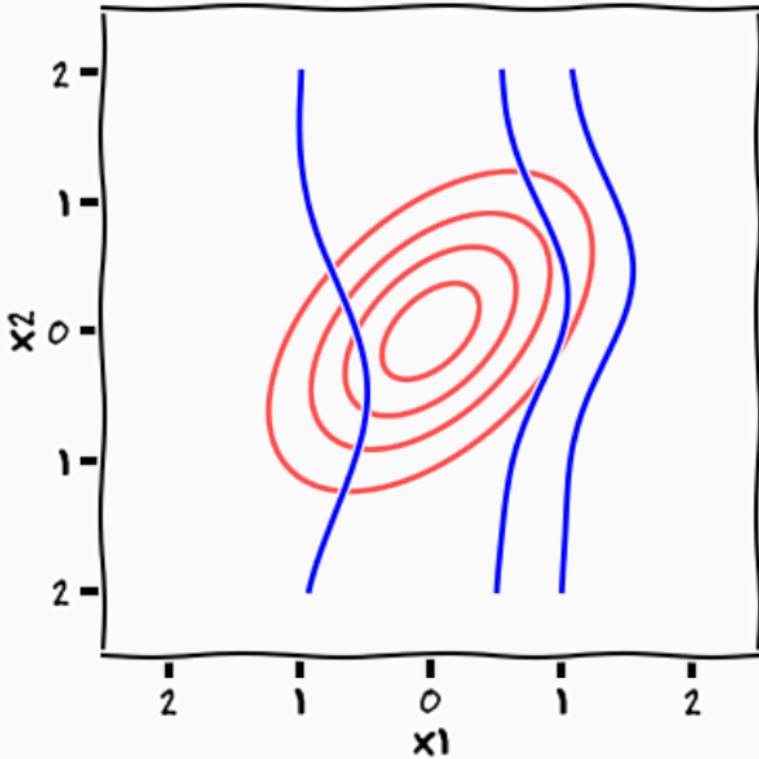
Conditional Gaussians



Conditional Gaussians



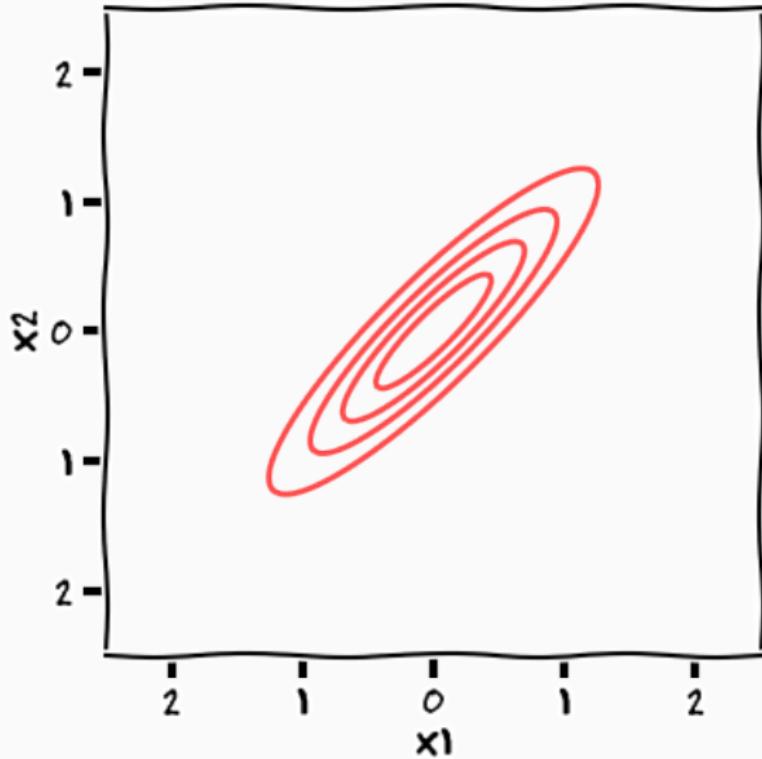
Conditional Gaussians



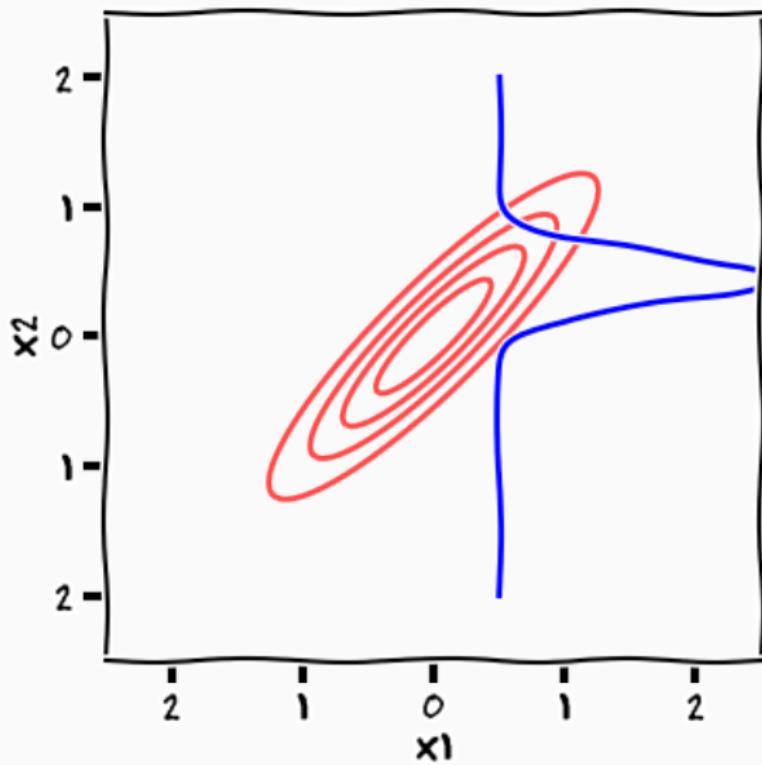
Conditional Gaussians

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

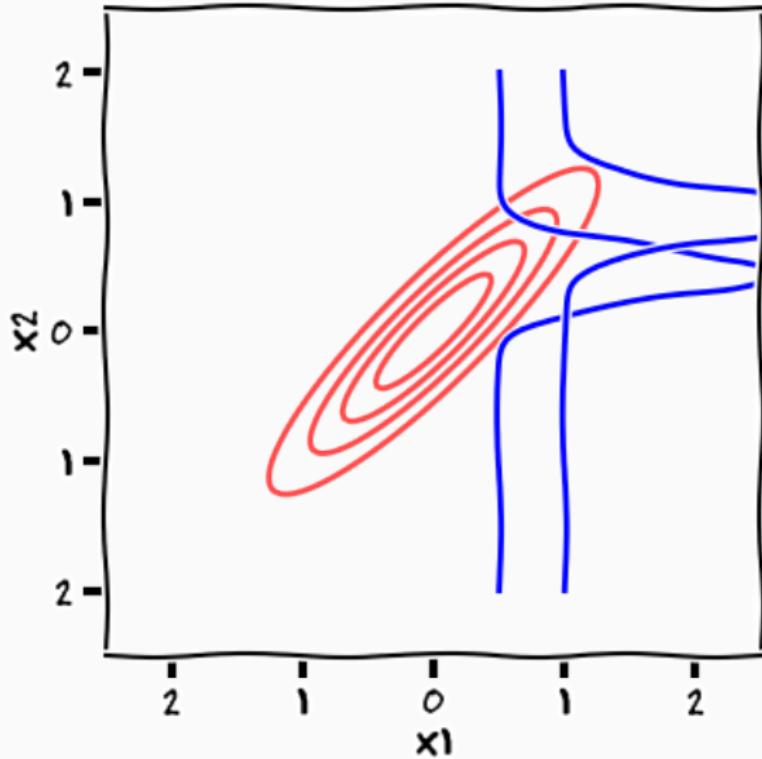
Conditional Gaussians



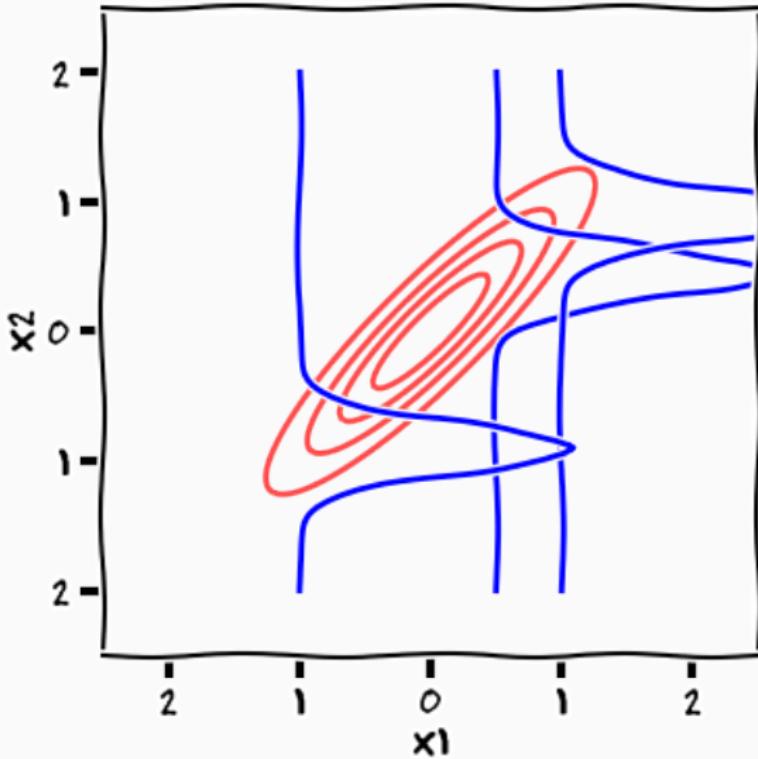
Conditional Gaussians



Conditional Gaussians



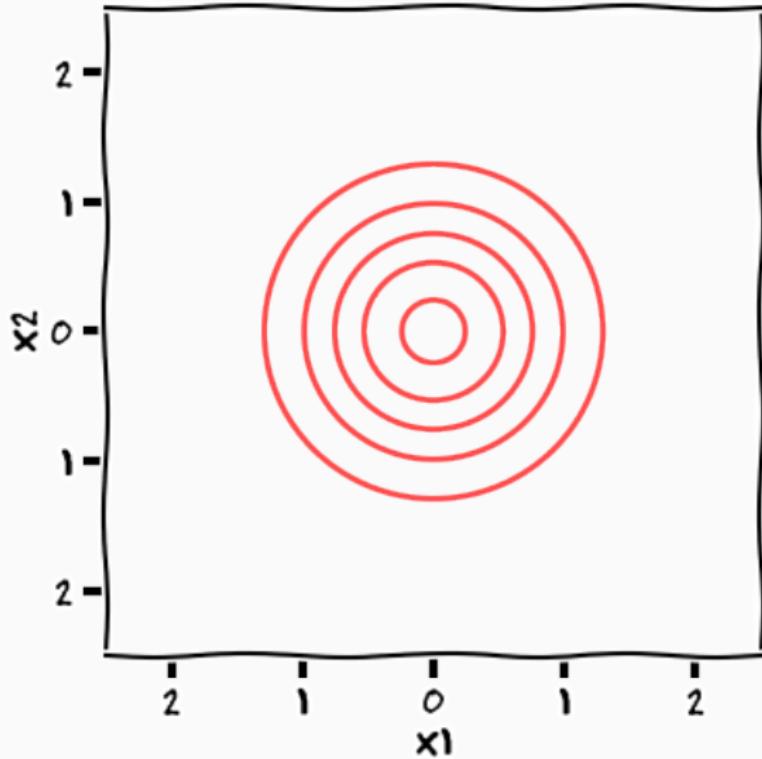
Conditional Gaussians



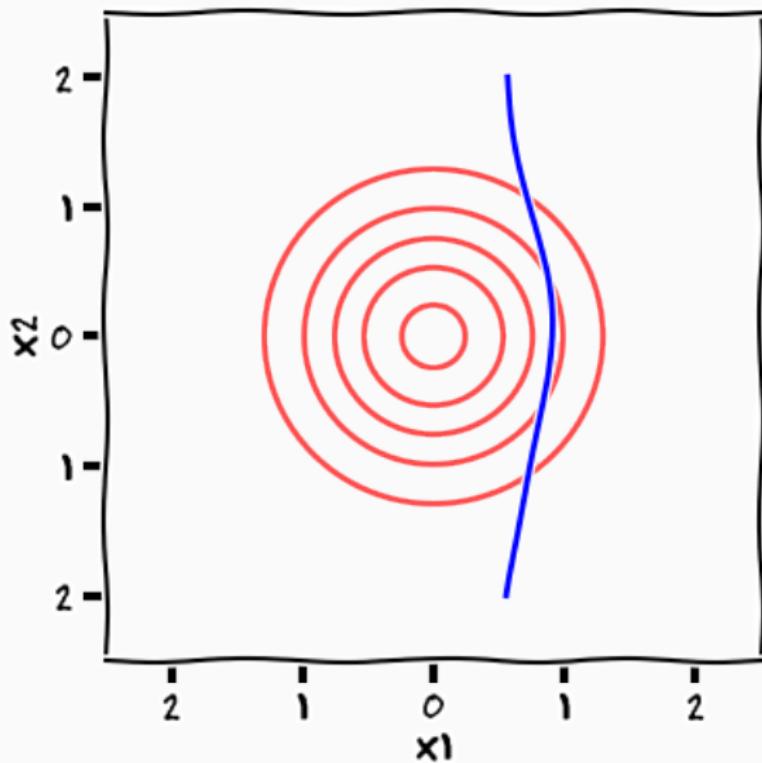
Conditional Gaussians

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

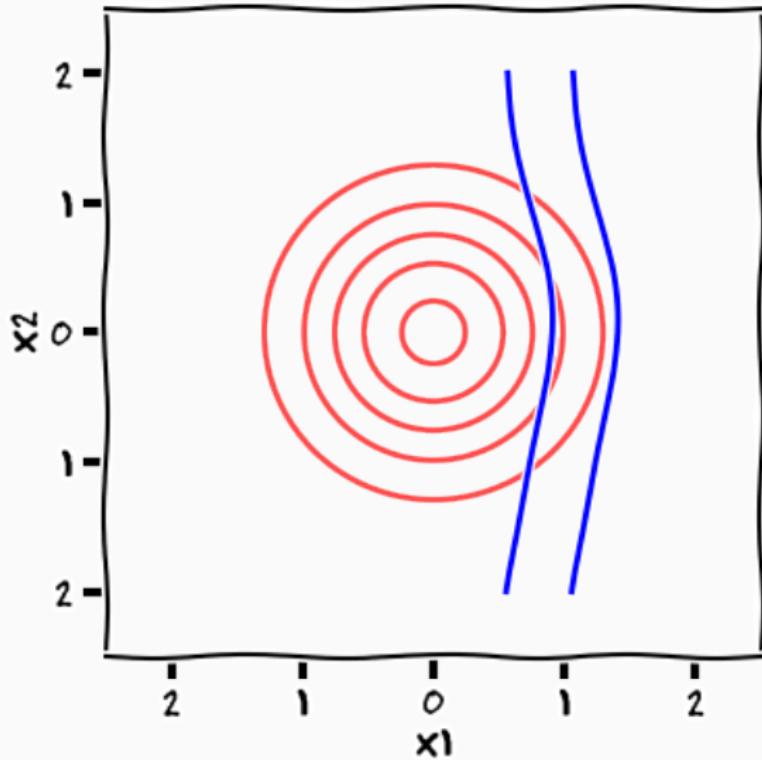
Conditional Gaussians



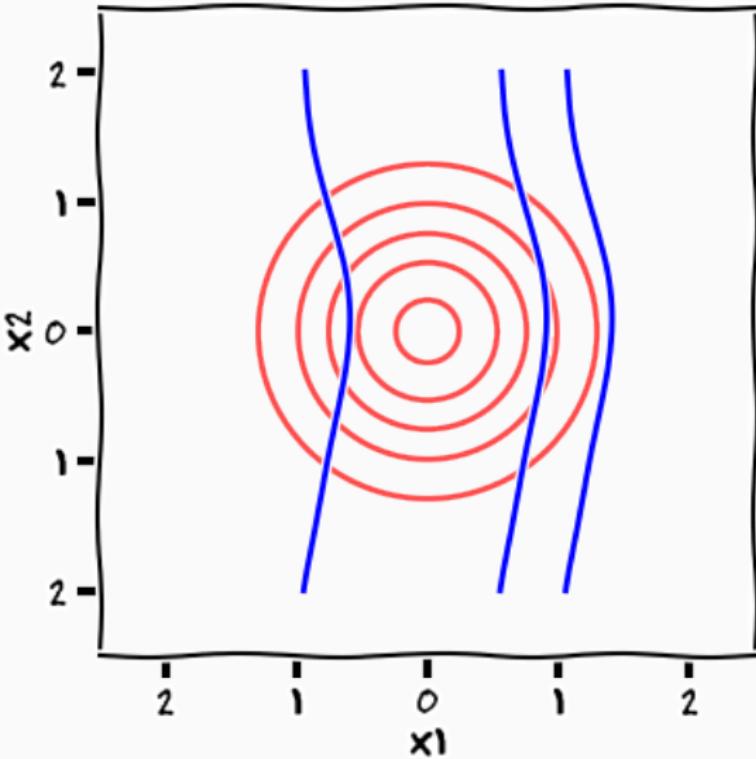
Conditional Gaussians



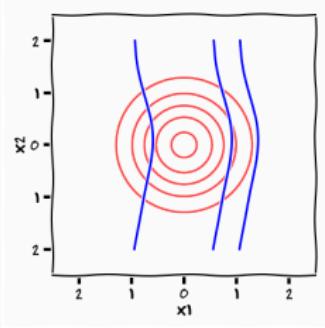
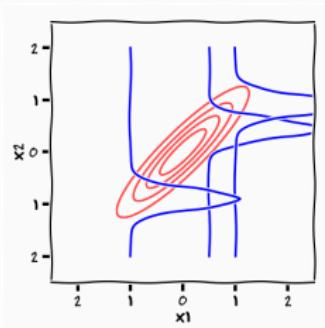
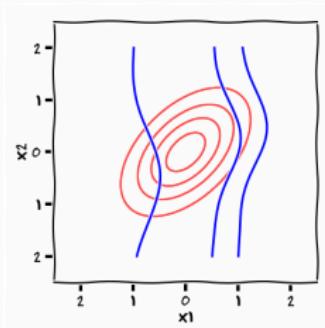
Conditional Gaussians



Conditional Gaussians



Conditional Gaussians

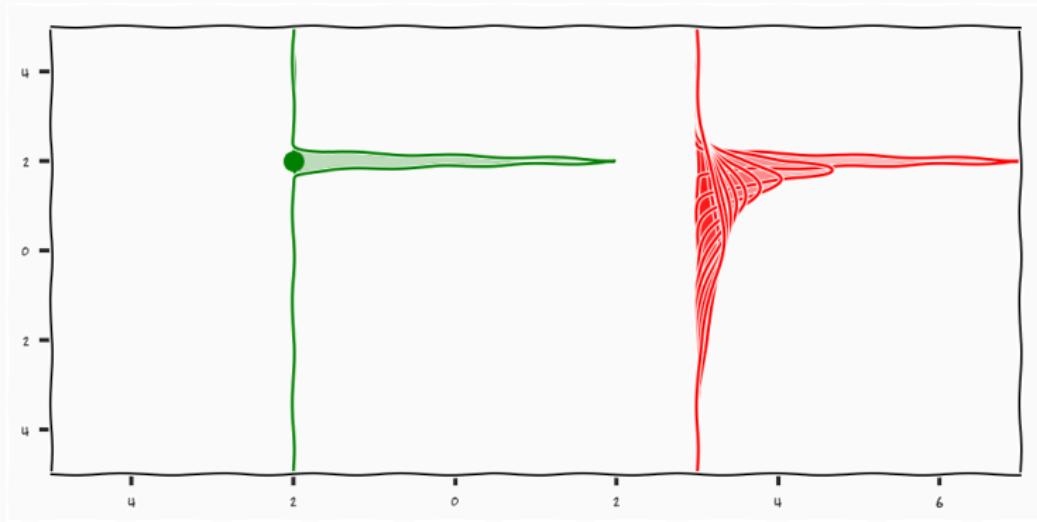


$$N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}\right)$$

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

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

Gaussian Processes



$$p(\mathbf{f}|\mathbf{x}, \theta) \sim \mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}))$$

Definition (Gaussian Process)

A Gaussian Process is an infinite collection of random variables who **any** subset is jointly gaussian. The process is specified by a mean function $\mu(\cdot)$ and a co-variance function $k(\cdot, \cdot)$

Gaussian Marginal

$$p(f_1, f_2, \dots, f_N, \dots | \mathbf{x}, \boldsymbol{\theta}) = \mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}))$$

$$= \mathcal{N} \left(\begin{bmatrix} \mu(x_1) \\ \mu(x_2) \\ \vdots \\ \mu(x_N) \\ \vdots \end{bmatrix}, \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & \cdots & k(x_1, x_N) & \cdots \\ k(x_2, x_1) & k(x_2, x_2) & \cdots & k(x_2, x_N) & \cdots \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ k(x_N, x_1) & k(x_N, x_2) & \cdots & k(x_N, x_N) & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \right)$$

Gaussian Processes

$$p(\mathbf{f}|\mathbf{x}, \theta) \sim \mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}))$$

$$\mathbf{y}_i = f_i + \epsilon$$

$$\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$$

$$p(\mathbf{y}|\mathbf{x}, \theta) = \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{x}, \theta)d\mathbf{f}$$

\mathcal{GP} is infinite, but we only observe finite amount of data. This means conditioning on a subset of the data, the \mathcal{GP} is just a Gaussian distribution

Uncertainty over functions

- Regression model,

$$\begin{aligned}\mathbf{y}_i &= f(\mathbf{x}_i) + \epsilon \\ \epsilon &\sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})\end{aligned}$$

- Introduce f_i as *instantiation* of function at location x_i

$$f_i = f(\mathbf{x}_i),$$

- as a new random variable.
- now we have a "handle" to specify our assumptions over

Uncertainty over functions

Joint

$$p(\mathbf{y}, \mathbf{f}|\mathbf{x}, \theta) = p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{x}, \theta)$$

Prior

$$p(\mathbf{f}|\mathbf{x}, \theta),$$

Likelihood

$$p(\mathbf{y}|\mathbf{f}) = \prod_i^N p(y_i|f_i)$$

The Mean Function

- Function of only the input location
- What do I expect the function value to be **only** accounting for the input location

The Covariance Function

- Function of **two** input locations
- How should the information from other locations with **known** function value observations effect my estimate

The Mean Function

- Function of only the input location
- What do I expect the function value to be **only** accounting for the input location
- We will assume this to be constant

The Covariance Function

- Function of **two** input locations
- How should the information from other locations with **known** function value observations effect my estimate

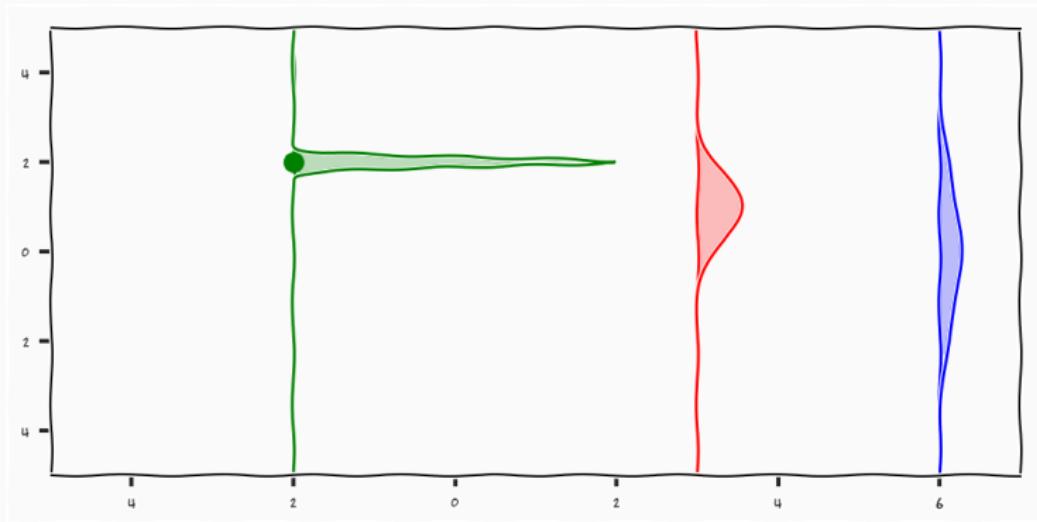
The Mean Function

- Function of only the input location
- What do I expect the function value to be **only** accounting for the input location
- We will assume this to be constant

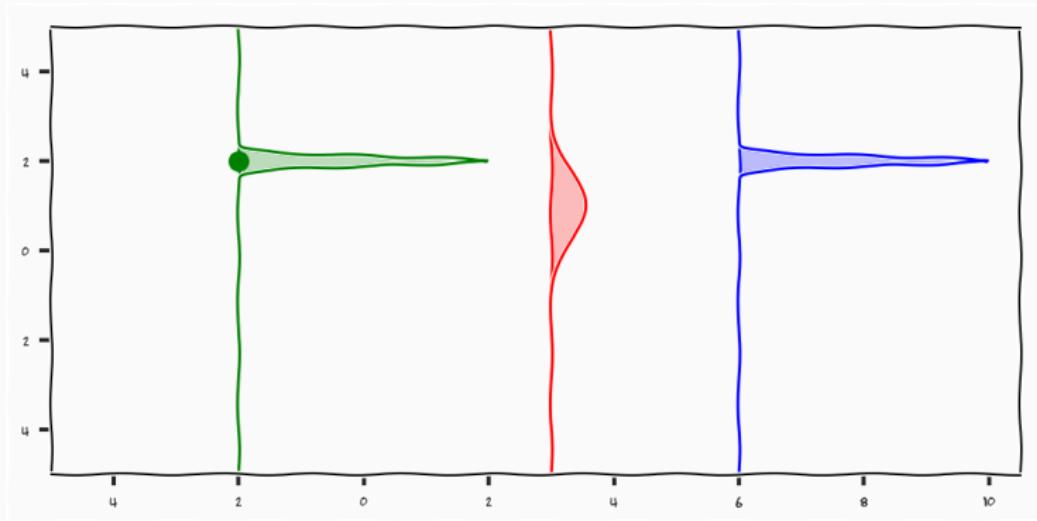
The Covariance Function

- Function of **two** input locations
- How should the information from other locations with **known** function value observations effect my estimate
- Encodes the behavior of the function

Gaussian Processes



Gaussian Processes

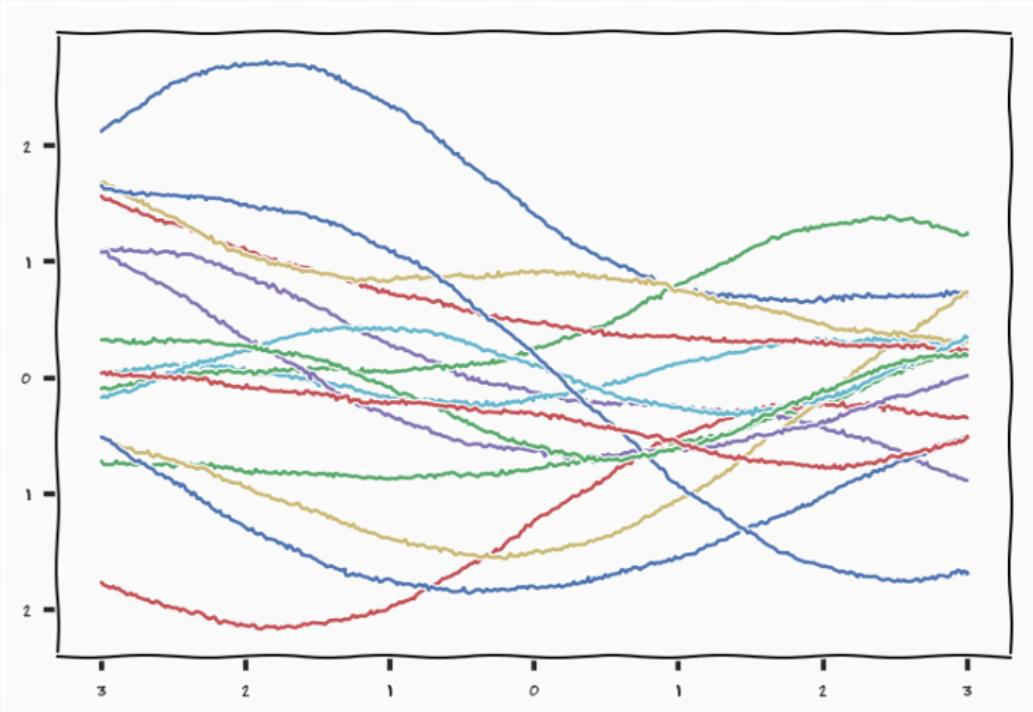


Gaussian Process: Samples

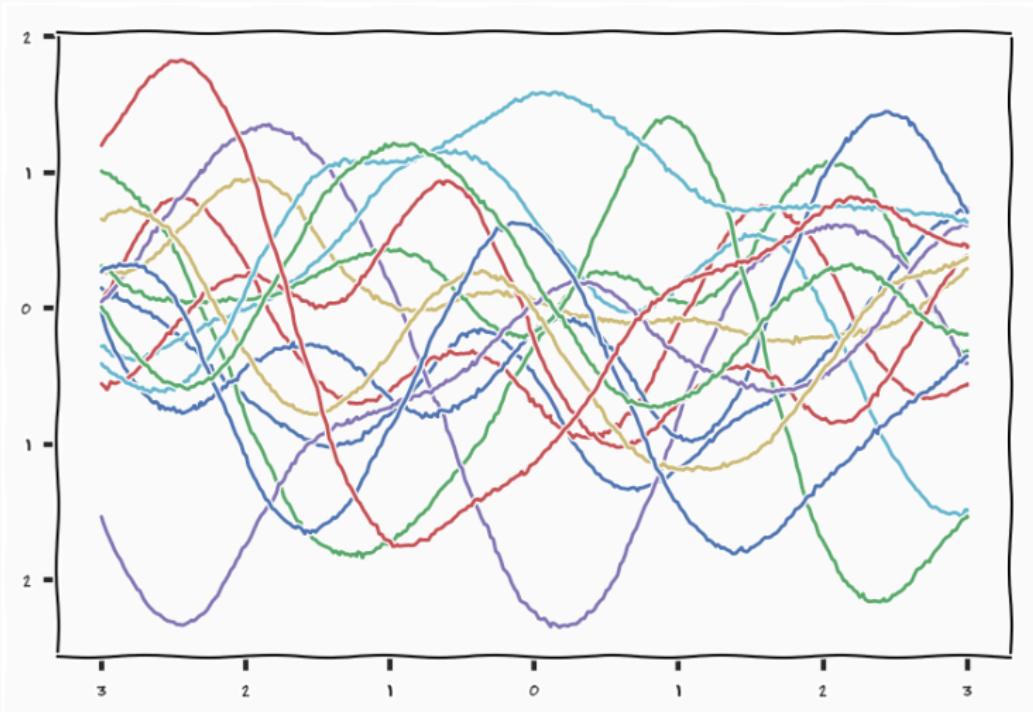
$$p(f_1, f_2, \dots, f_N, \dots | \mathbf{x}, \boldsymbol{\theta}) = \mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}))$$

$$= \mathcal{N} \left(\begin{bmatrix} \mu(x_1) \\ \mu(x_2) \\ \vdots \\ \mu(x_N) \\ t \end{bmatrix}, \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & \cdots & k(x_1, x_N) & \cdots \\ k(x_2, x_1) & k(x_2, x_2) & \cdots & k(x_2, x_N) & \cdots \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ k(x_N, x_1) & k(x_N, x_2) & \cdots & k(x_N, x_N) & \cdots \end{bmatrix} \right)$$

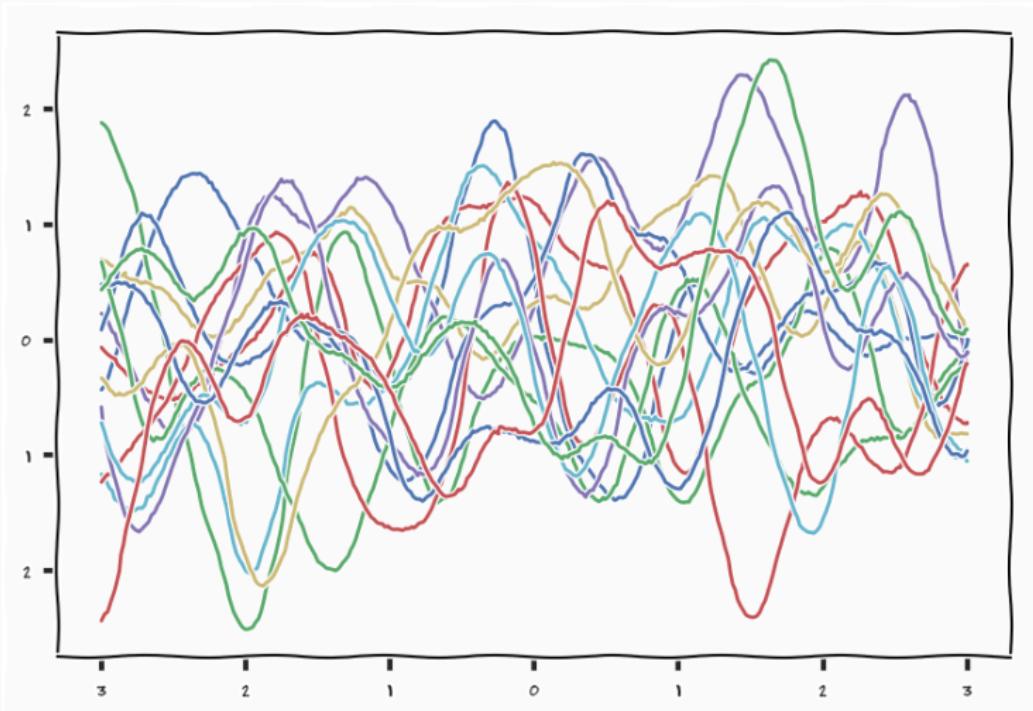
Gaussian Processes: Samples



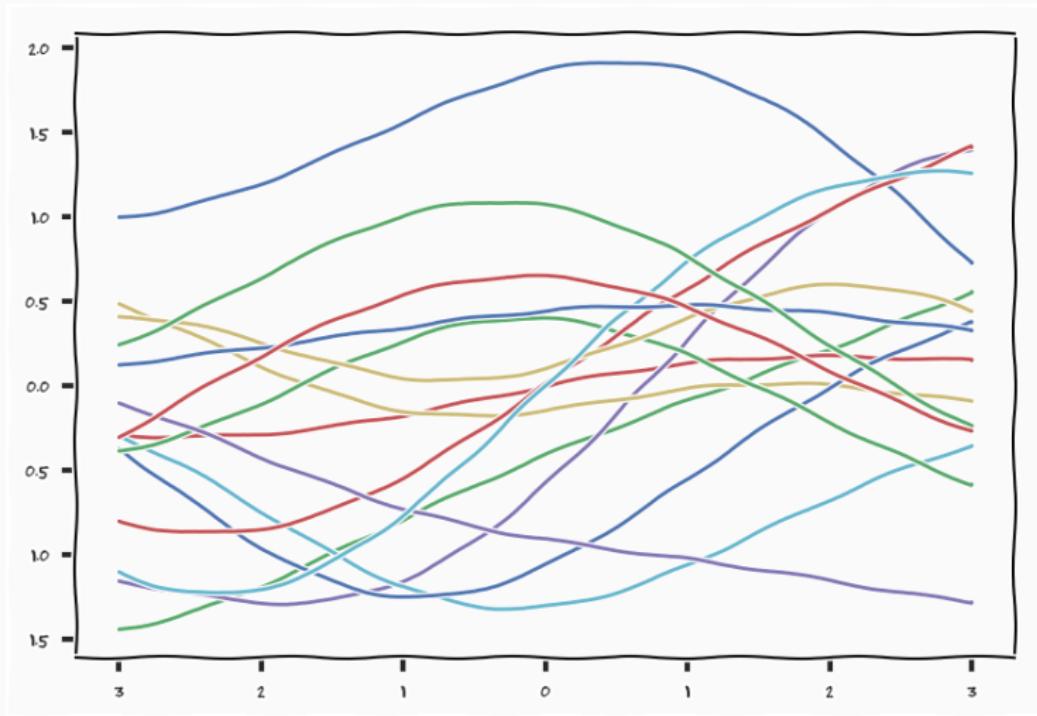
Gaussian Processes: Samples



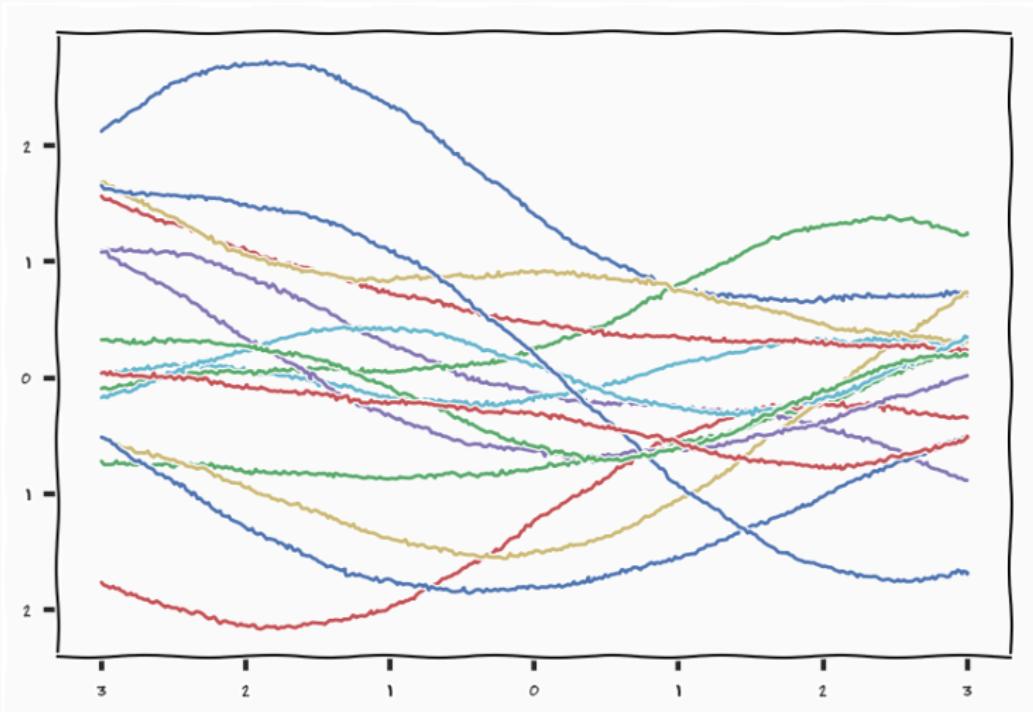
Gaussian Processes: Samples



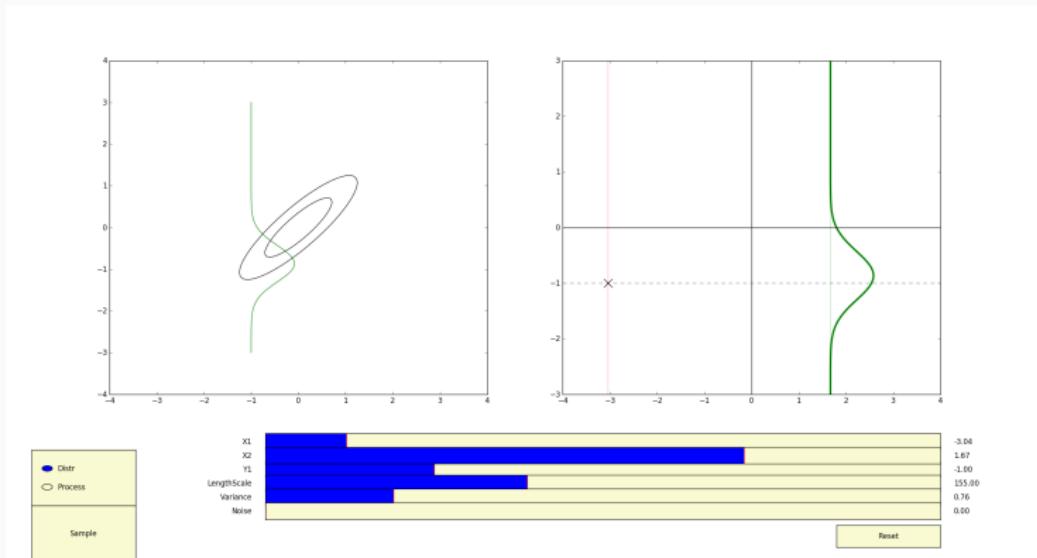
Gaussian Processes: Samples



Gaussian Processes: Samples



Demo



Gaussian Process: Posterior

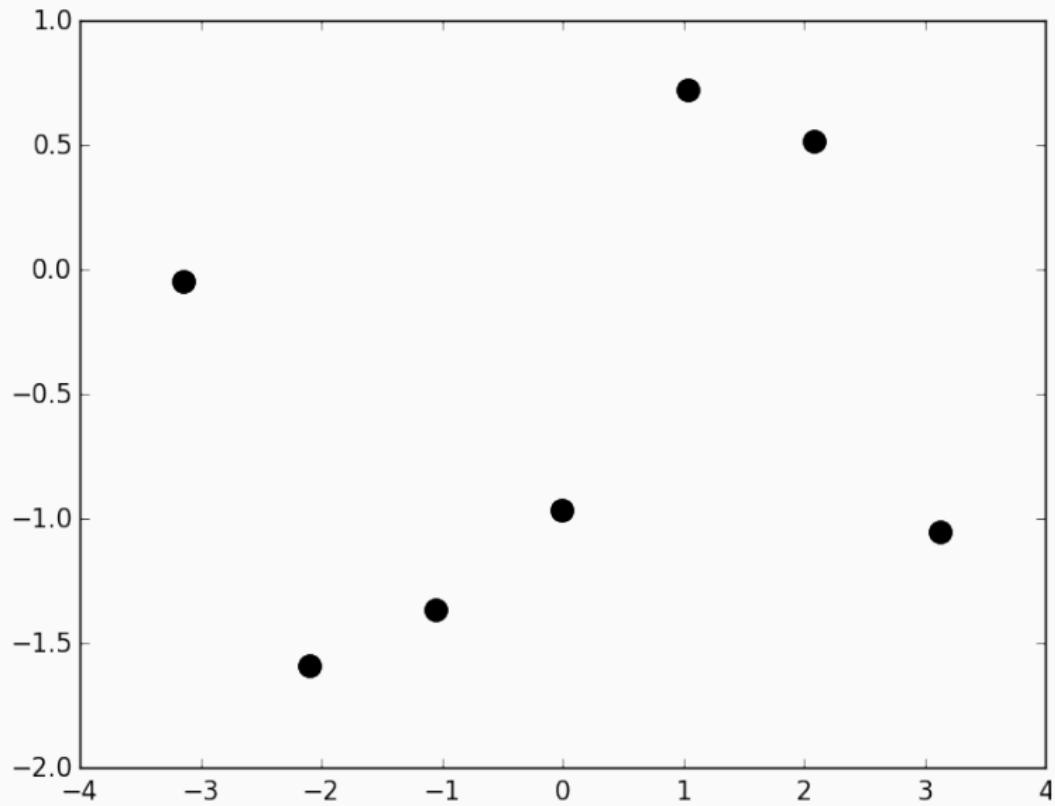
- All instantiations are jointly Gaussian

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

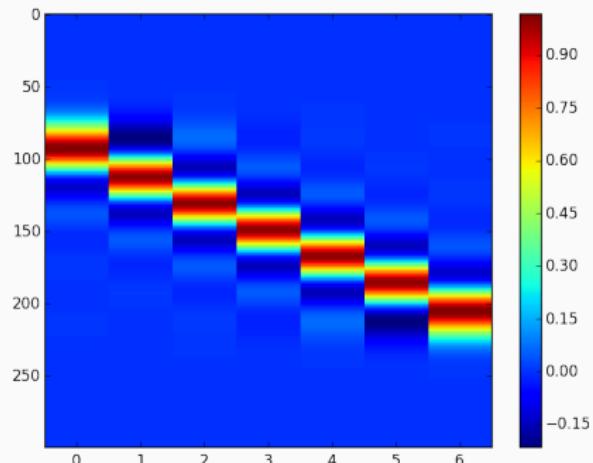
- Conditional Gaussian (same as always)

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

Gaussian Process: Posterior

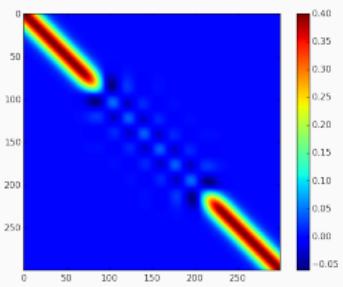
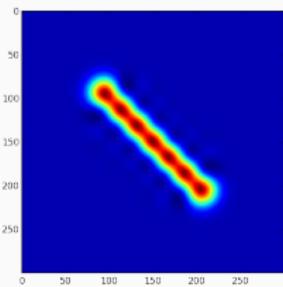
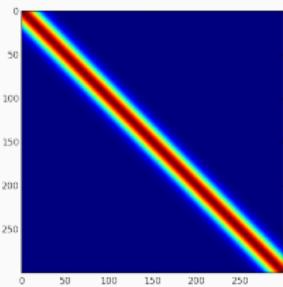


Does it make sense



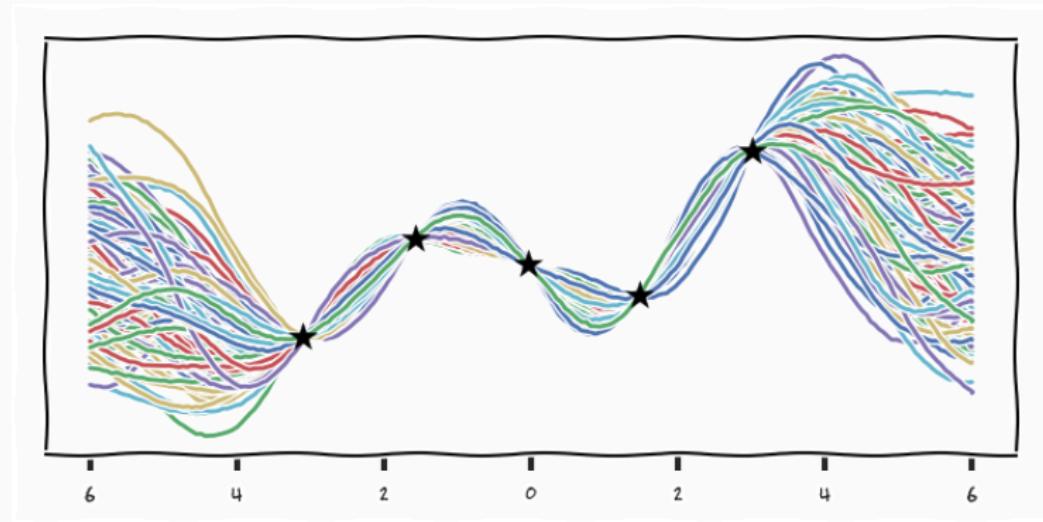
$$k(\mathbf{x}_*, \mathbf{X})^T k(\mathbf{X}, \mathbf{X})^{-1} \mathbf{f}$$

Does it make sense

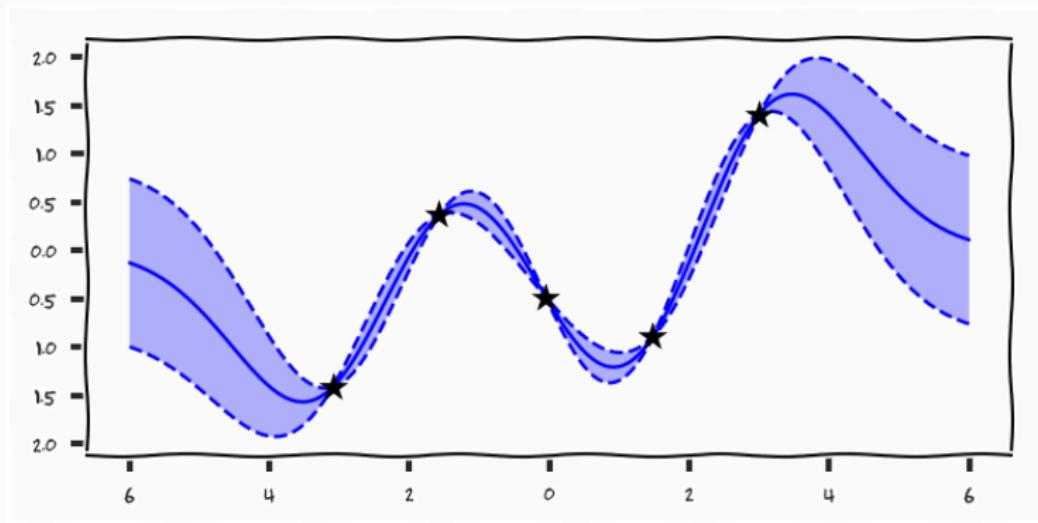


$$k(\mathbf{x}_*, \mathbf{x}_*) - k(\mathbf{x}_*, \mathbf{X})^T k(\mathbf{X}, \mathbf{X})^{-1} k(\mathbf{X}, \mathbf{x}_*)$$

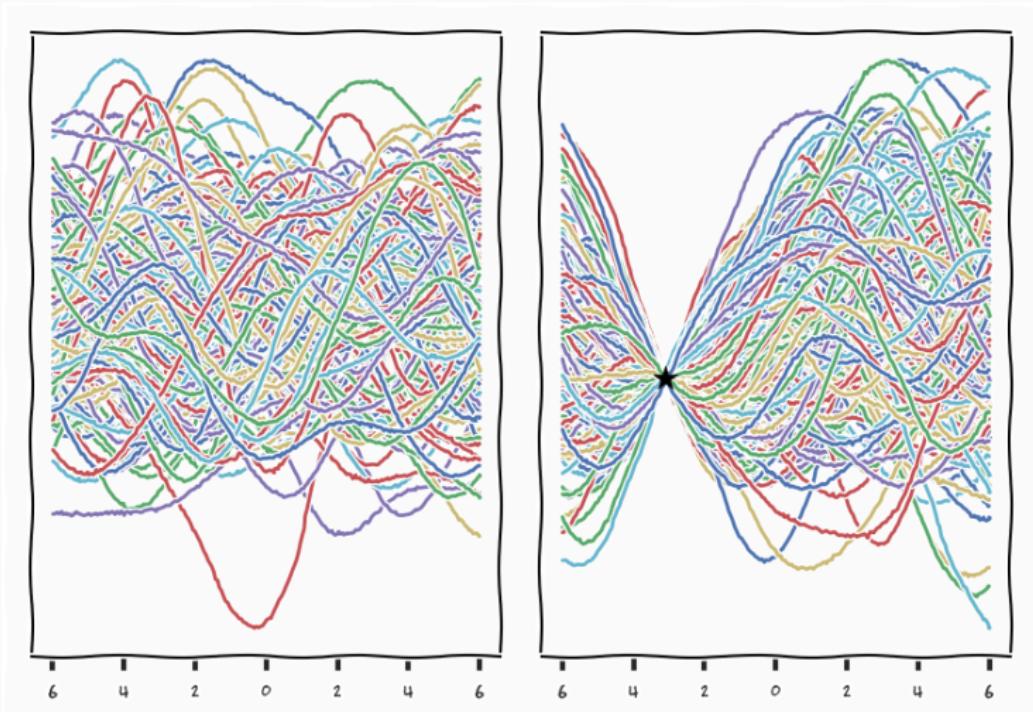
Gaussian Processes Posterior



Gaussian Processes Posterior



Gaussian Processes: Posterior Samples



Gaussian Processes: Noisy observations

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

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

- Add noise to observations

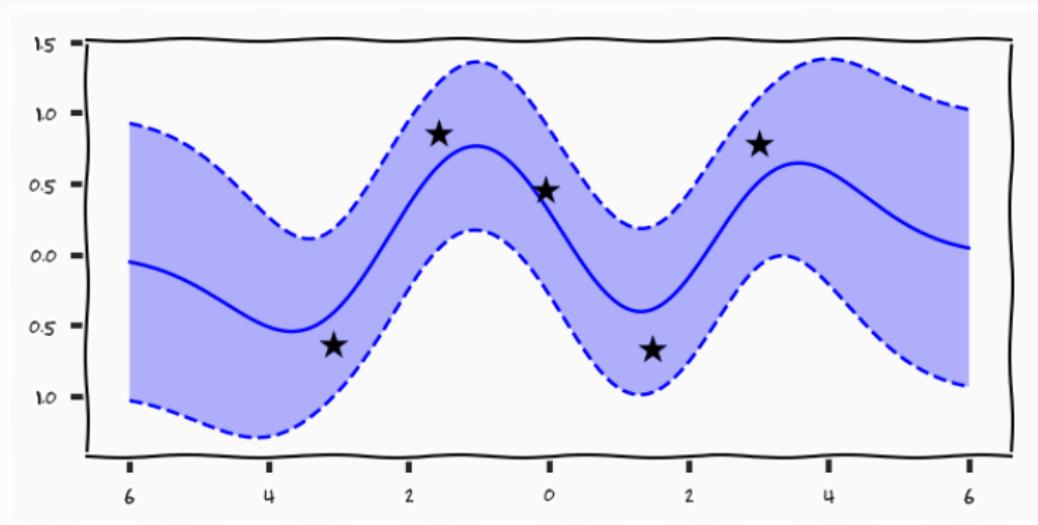
Gaussian Processes: Noisy observations

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

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

- Add noise to observations
- *Do you recognise the mean?*

Gaussian Processes



Summary

Summary

- Repeat of the machine learning procedure
 - assumption + data + compute → updated assumption
 - don't worry it will become clear eventually
- Gaussian processes
 - infinite generalisation of Gaussian distribution
 - prior over the space of functions
 - contains **all** functions

Useful?



eof

References

 Christopher M. Bishop.

*Pattern Recognition and Machine Learning (Information
Science and Statistics).*

Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.