



Gaussian Processes

a first introduction

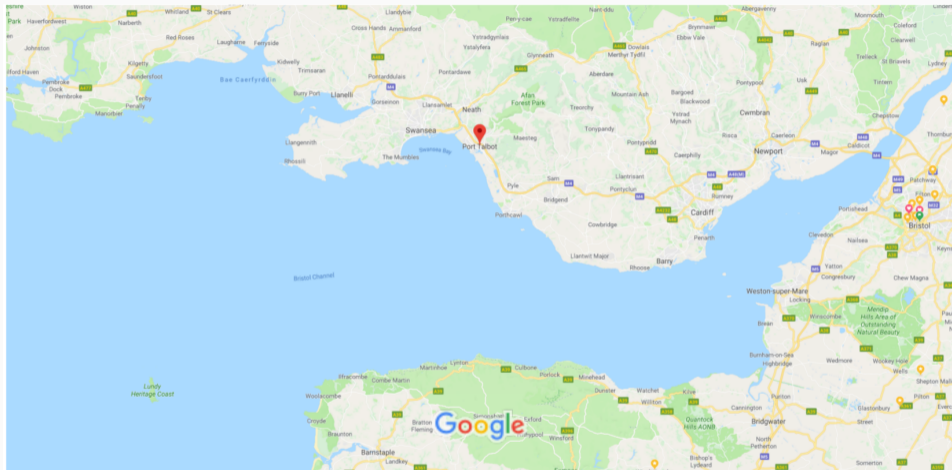
Carl Henrik Ek - che29@cam.ac.uk

September 9, 2024

<http://carlhenrik.com>



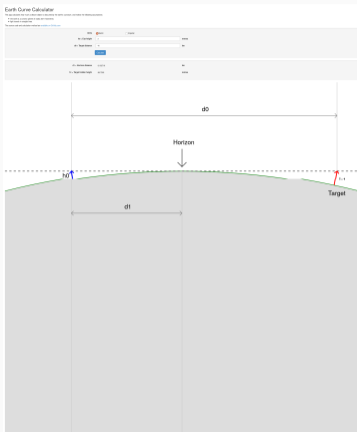












Distance to horizon $6.2km$

Hidden height $125.6m$





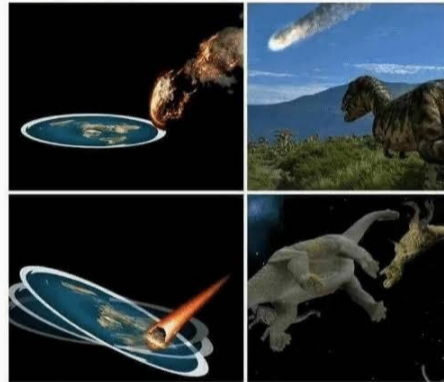
?

Flat Earth *The earth is flat and the water surface is flat, therefore I can see the building*

Flat Earth *The earth is flat and the water surface is flat, therefore I can see the building*

Spherical Earth *Due to the temperature gradient between the water and air, there is a dispersion of water molecules into the air proportional to the distance to the surface effectively creating a lens allowing us to see "around the bend" of the earths curvature*

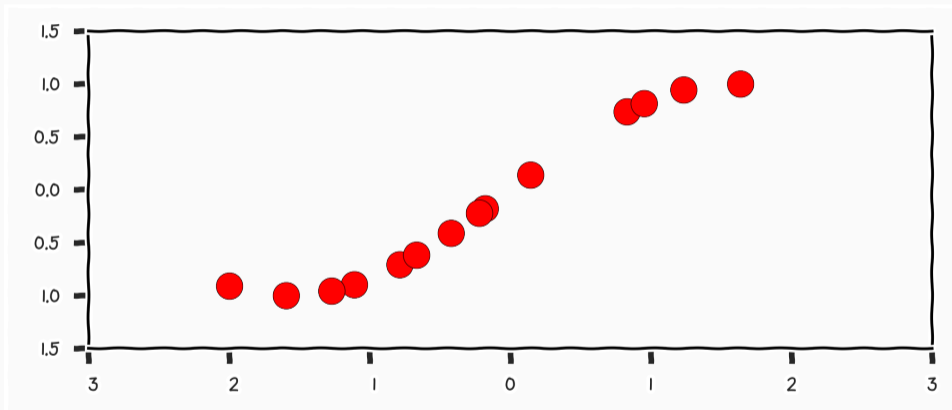
THE EXTINCTION OF THE DINOSAURS



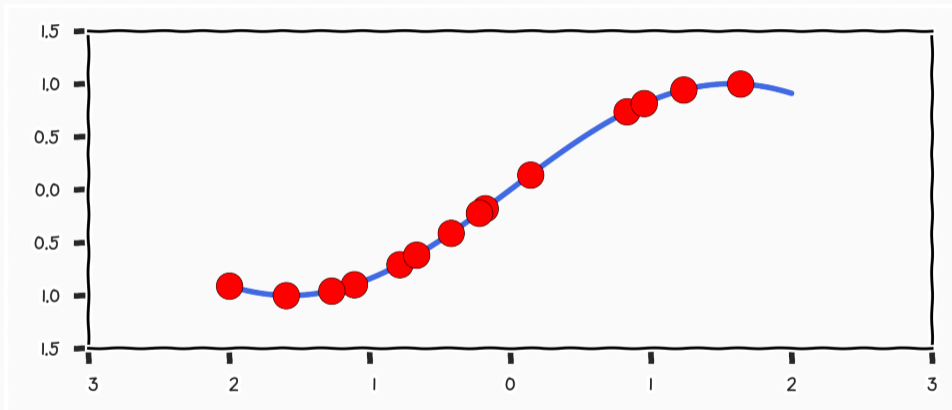
ACCORDING TO FLAT EARTHERS

ifunny.co

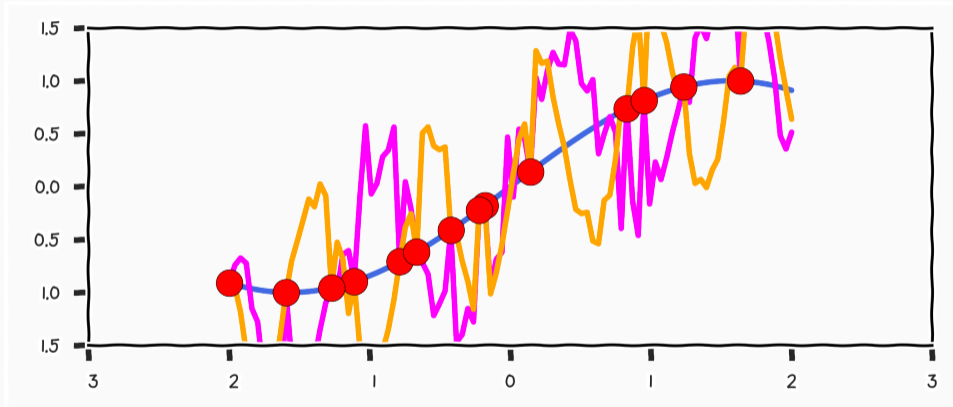
Curve Fitting



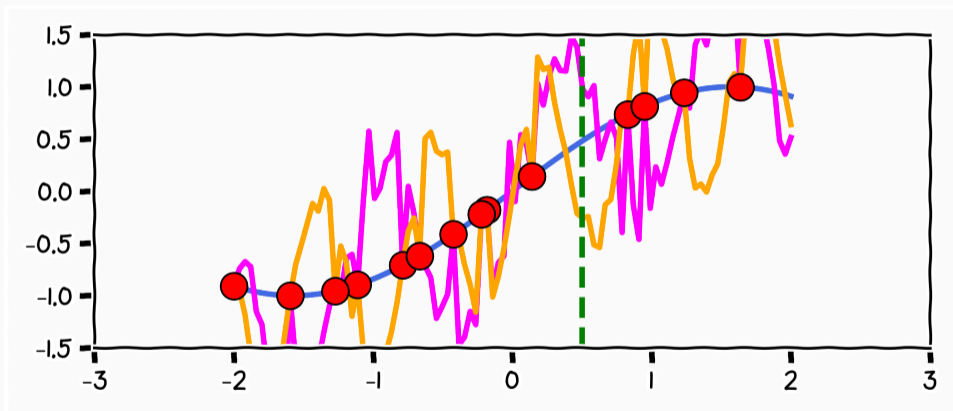
Curve Fitting

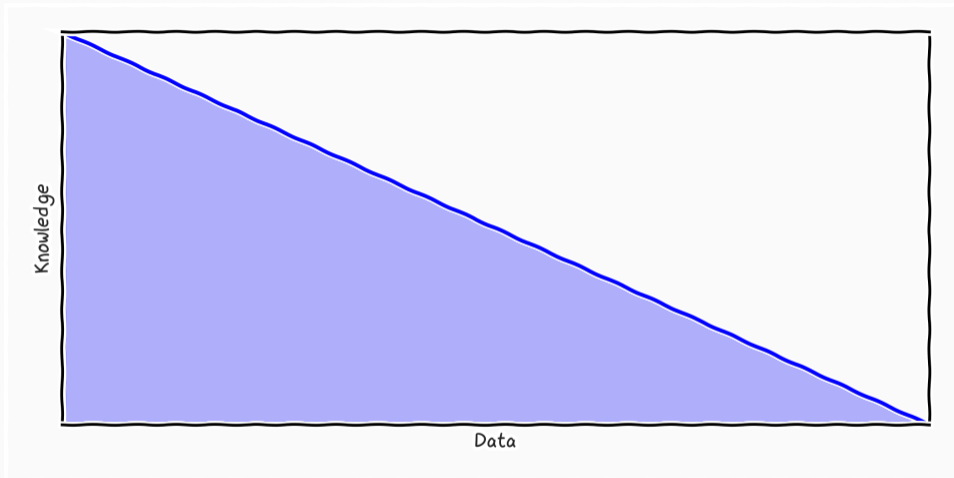


Curve Fitting



Curve Fitting





What is Machine Learning Machine Learning is the task of combining/integrating knowledge with observations to perform predictions using the subset of possible explanations that are consistent with both my knowledge and the observations

What is Machine Learning Machine Learning is the task of combining/integrating knowledge with observations to perform predictions using the subset of possible explanations that are consistent with both my knowledge and the observations

Isn't this Statistics? statistics cares about **parameters** of the knowledge while ML cares about the predictions we get from **using** the parameters we infer by combining knowledge and observations. (It is just a slight change of narrative)

Domain Set \mathcal{X} the set of measurements/objects that we want to label (input)

Domain Set \mathcal{X} the set of measurements/objects that we want to label (input)

Label Set \mathcal{Y} the set of outputs

Domain Set \mathcal{X} the set of measurements/objects that we want to label (input)

Label Set \mathcal{Y} the set of outputs

Training Data \mathcal{S} a finite sequence of pairs in $\mathcal{X} \times \mathcal{Y}$

Data Distribution \mathcal{D} probability distribution governing the measurements

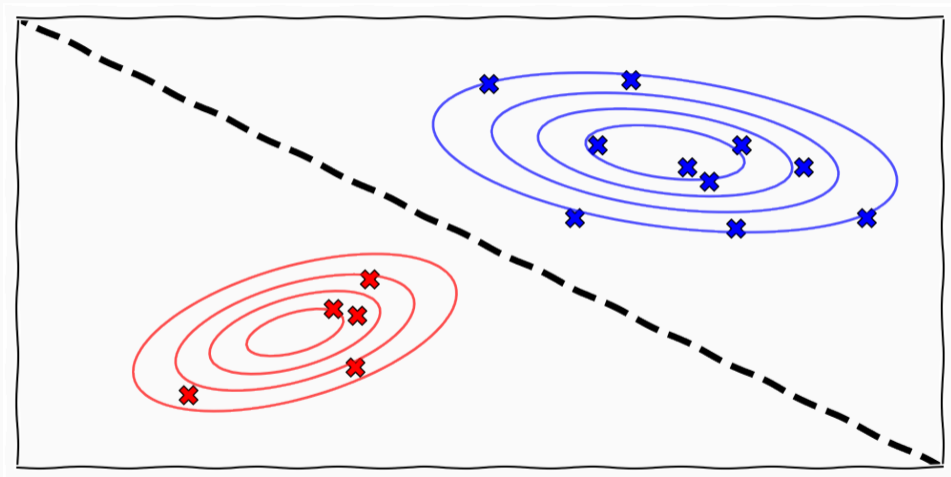
Data Distribution \mathcal{D} probability distribution governing the measurements

Data Generation $f : \mathcal{X} \rightarrow \mathcal{Y}$ the underlying generating process that we wish to recover

Data Distribution \mathcal{D} probability distribution governing the measurements

Data Generation $f : \mathcal{X} \rightarrow \mathcal{Y}$ the underlying generating process that we wish to recover

Prediction Rule $h : \mathcal{X} \rightarrow \mathcal{Y}$ what we wish to recover, the object that encodes the recovered knowledge



$$L_{\mathcal{D},f}(h) := \mathcal{D}(\{x : h(x) \neq f(x)\})$$

- measure of success as probability of misclassified points (true risk)

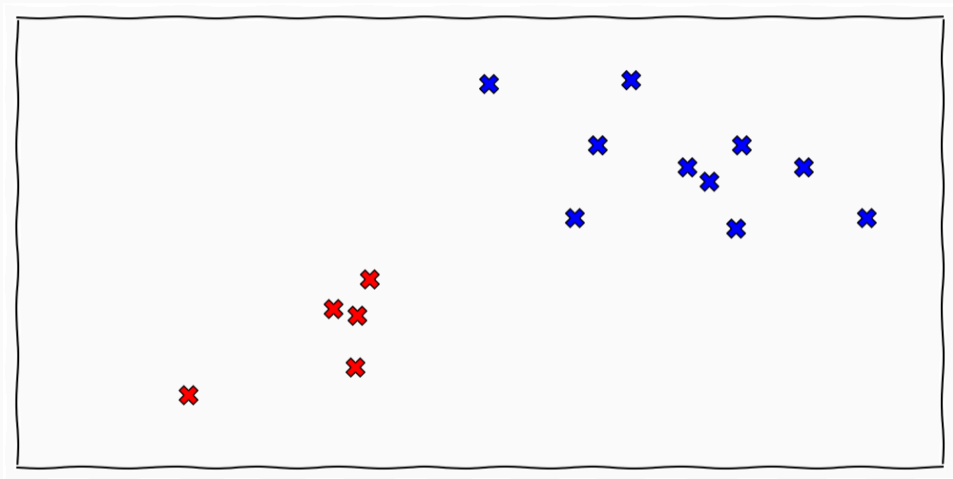
$$L_{\mathcal{D},f}(h) := \mathcal{D}(\{x : h(x) \neq f(x)\})$$

- measure of success as probability of misclassified points (true risk)
- we do not have access to \mathcal{D}

$$L_{\mathcal{D},f}(h) := \mathcal{D}(\{x : h(x) \neq f(x)\})$$

- measure of success as probability of misclassified points (true risk)
- we do not have access to \mathcal{D}
- we do not have access to f

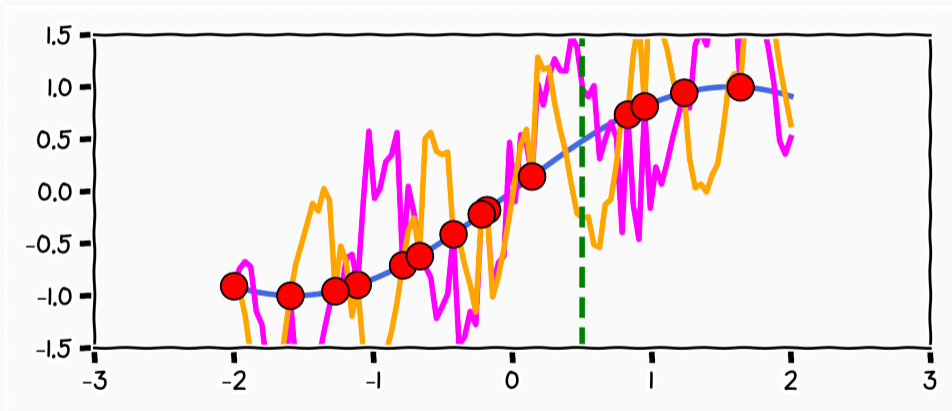
Classification



$$L_{\mathcal{S}}(h) := \frac{|\{i \in [m] : h(x_i) \neq y_i\}|}{m}$$

- We **assume** that $\mathcal{S} \sim \mathcal{D}$
- Empirical measure of risk

Curve Fitting

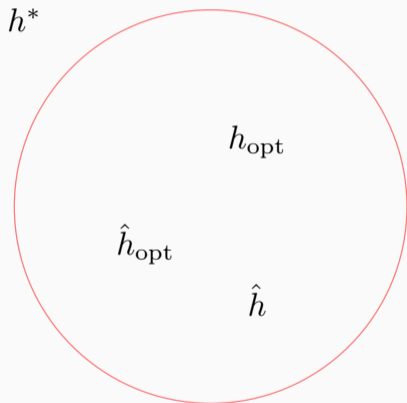


$$L_{\mathcal{S}}(A(\mathcal{S})) := \frac{|\{i \in [m] : h(x_i) \neq y_i\}|}{m}$$

- We use an algorithm $A : \mathcal{S} \rightarrow h$ to find a hypothesis

$$h_S \in \operatorname{argmin}_{h \in \mathcal{H}} L_S(h)$$

- We cannot parametrise **all** possible hypothesis



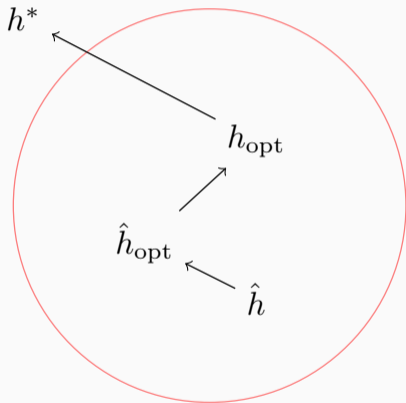
h^* the optimal predictor

h_{opt} the optimal hypothesis

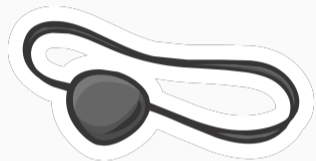
\hat{h}_{opt} the optimal hypothesis on training data

\hat{h} the hypothesis found by learning algorithm

Error Decomposition



$$\begin{aligned} & \epsilon(\hat{h}) - \epsilon(h^*) \\ &= \underbrace{\epsilon(h_{\text{opt}}) - \epsilon(h^*)}_{\text{Approximation}} \\ &+ \underbrace{\epsilon(\hat{h}_{\text{opt}}) - \epsilon(h_{\text{opt}})}_{\text{Estimation}} \\ &+ \underbrace{\epsilon(\hat{h}) - \epsilon(\hat{h}_{\text{opt}})}_{\text{Optimisation}} \end{aligned}$$

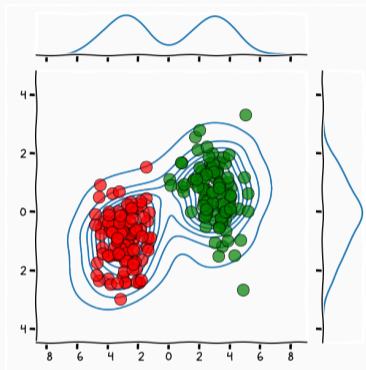


Statistical Learning



$$A_{\mathcal{H}}(S)$$

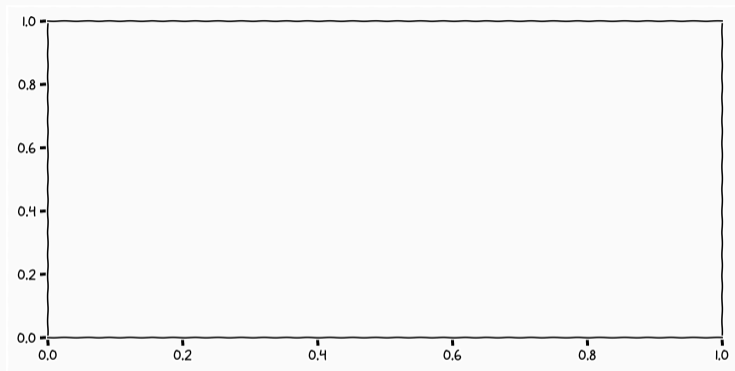




Statistical Learning

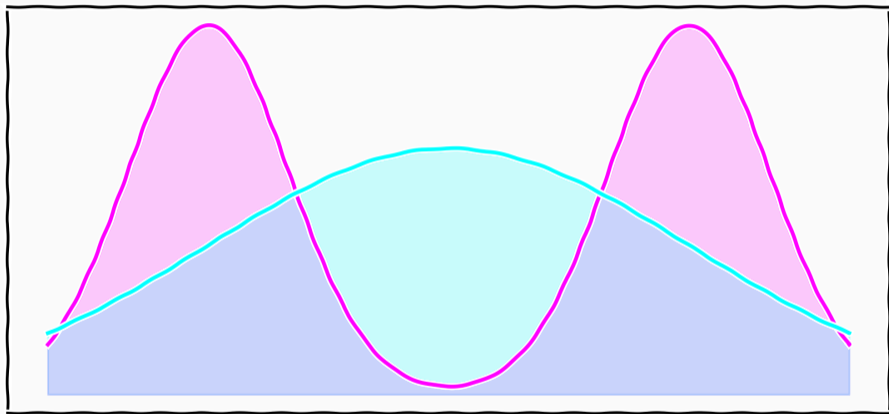
$$A_{\mathcal{H}}(\mathcal{S})$$

Assumptions: Hypothesis space



Statistical Learning

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



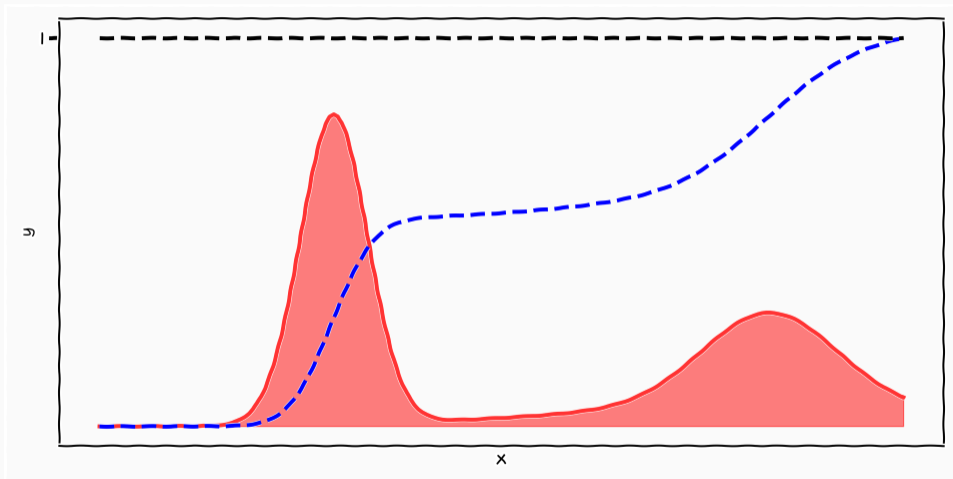
$$p(\theta | \mathcal{D}) = \frac{p(\mathcal{D} | \theta)p(\theta)}{p(\mathcal{D})}$$

$$p(\mathcal{D}) = \int p(\mathcal{D} | \theta)p(\theta)d\theta$$

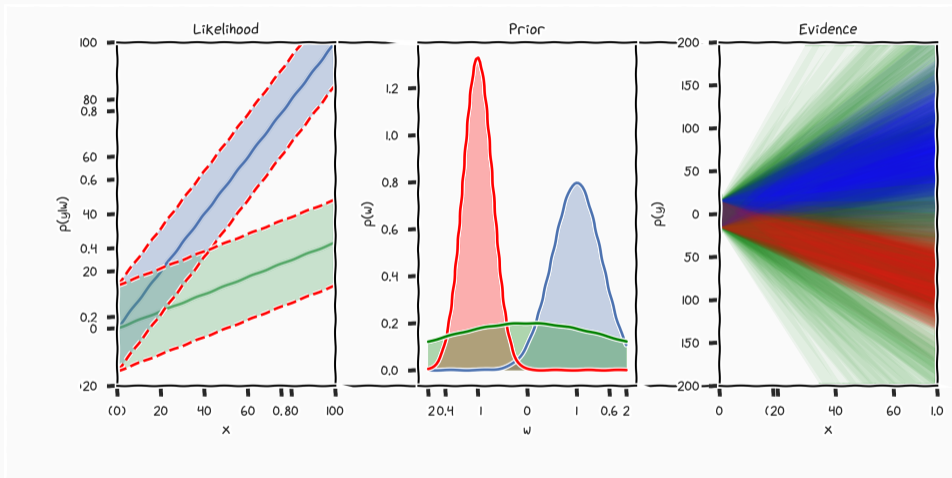
$$p(\mathcal{D}) = \int p(\mathcal{D} | \theta)p(\theta)d\theta$$

$$p(\mathcal{D}) = \int p(\mathcal{D} | \theta) p(\theta) d\theta$$

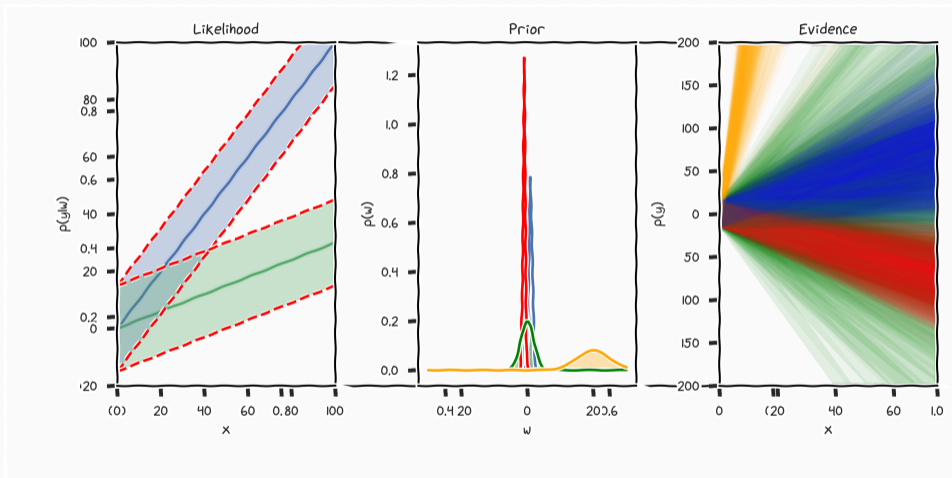
$$p(\mathcal{D}) = \int p(\mathcal{D} | \theta) \underbrace{p(\theta) d\theta}_{dt(\theta)}$$



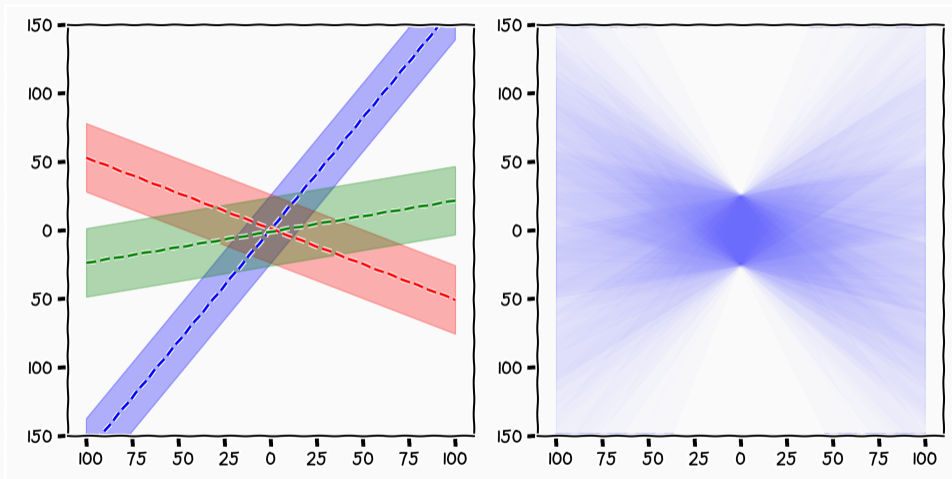
Marginalisation

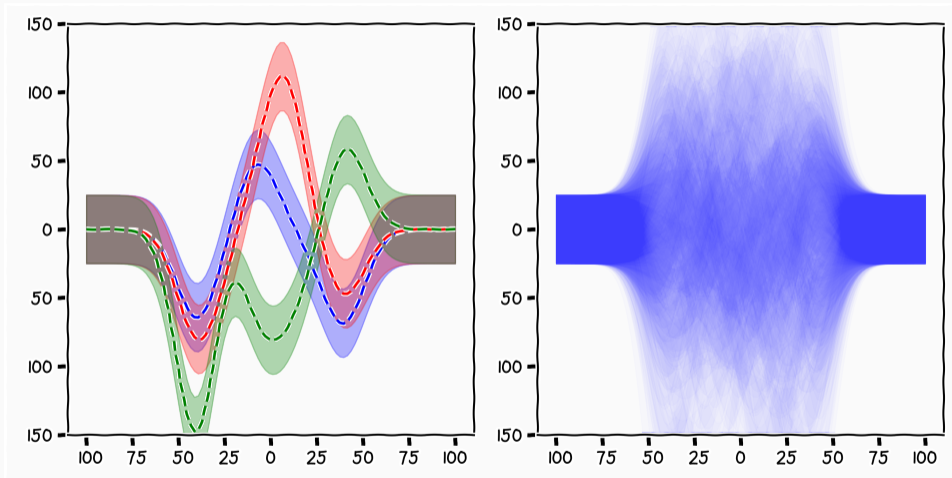


Marginalisation



Model Linear Linear





- The Bayesian argument implies that you try to re-parametrise the hypothesis space to reflect your beliefs

- The Bayesian argument implies that you try to re-parametrise the hypothesis space to reflect your beliefs
- A good analogy is to think about "space", the believable parameters gets a bigger space compared to the unlikely ones

- The Bayesian argument implies that you try to re-parametrise the hypothesis space to reflect your beliefs
- A good analogy is to think about "space", the believable parameters gets a bigger space compared to the unlikely ones
- Massive composite models can be thought of as directly altering the parameter space for the optimiser Roy et al., [2024](#)

Flexible such that we do not have to make trade-offs when including beliefs

Flexible such that we do not have to make trade-offs when including beliefs

Narrow such that we can reduce data-requirements

Flexible such that we do not have to make trade-offs when including beliefs

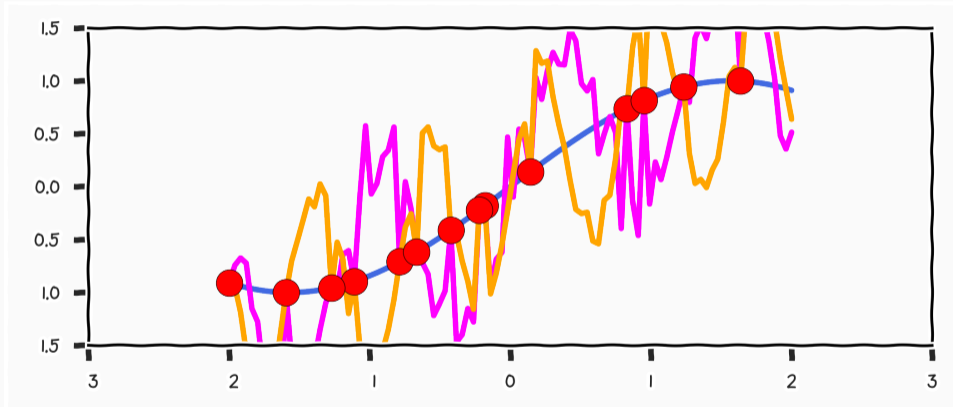
Narrow such that we can reduce data-requirements

Interpretable so that we can translate our knowledge to the parametrisation



Non-parametrics

Curve Fitting

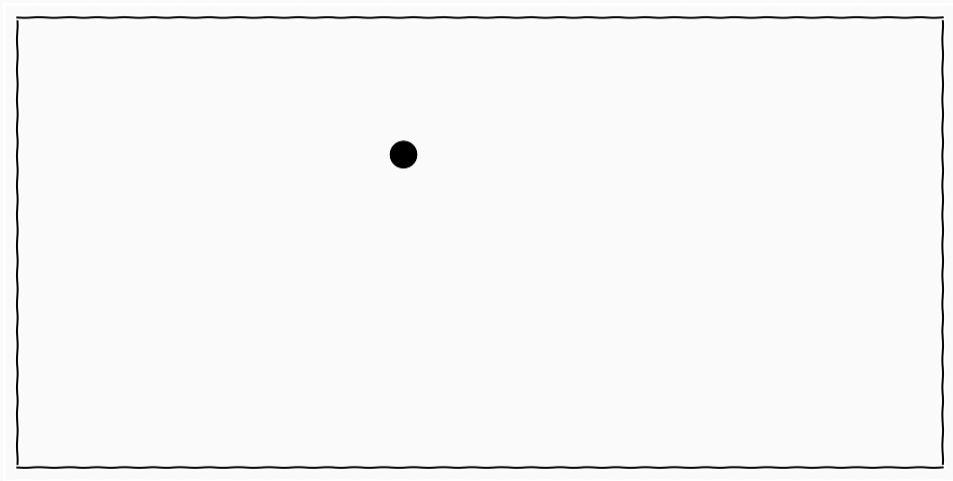




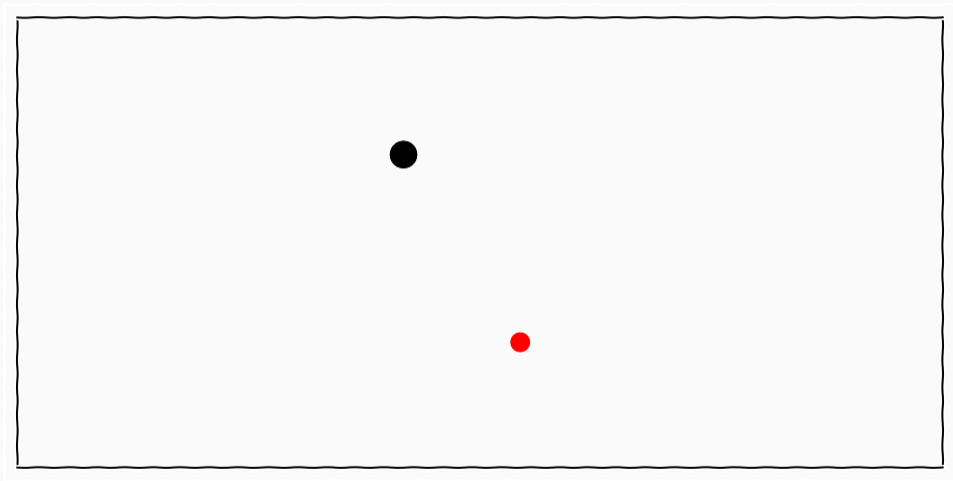




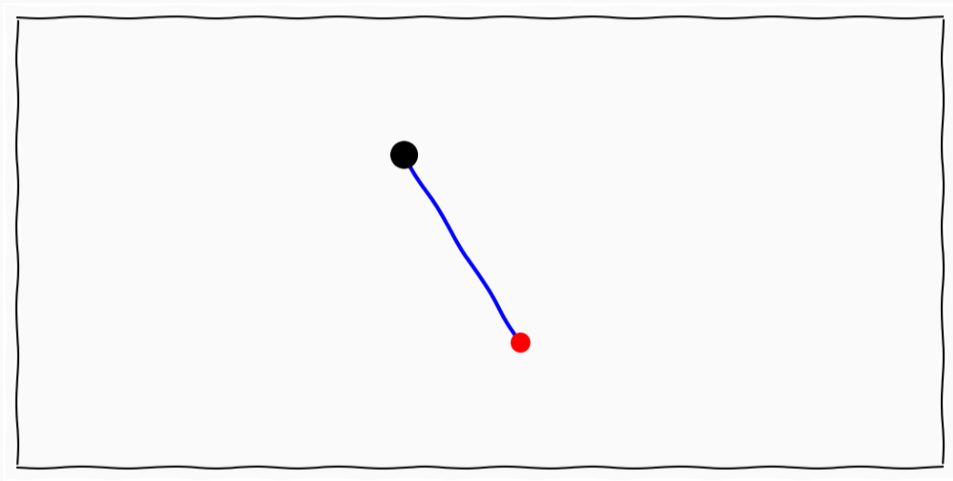
Example



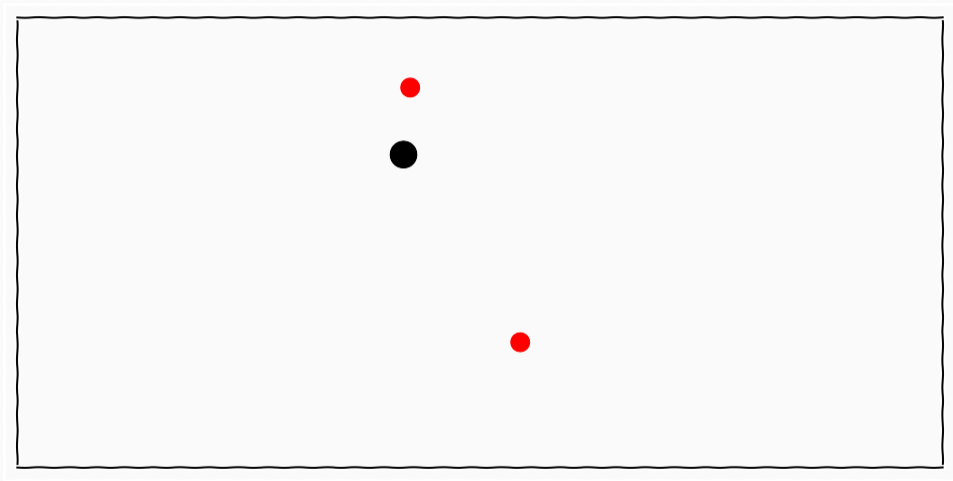
Example



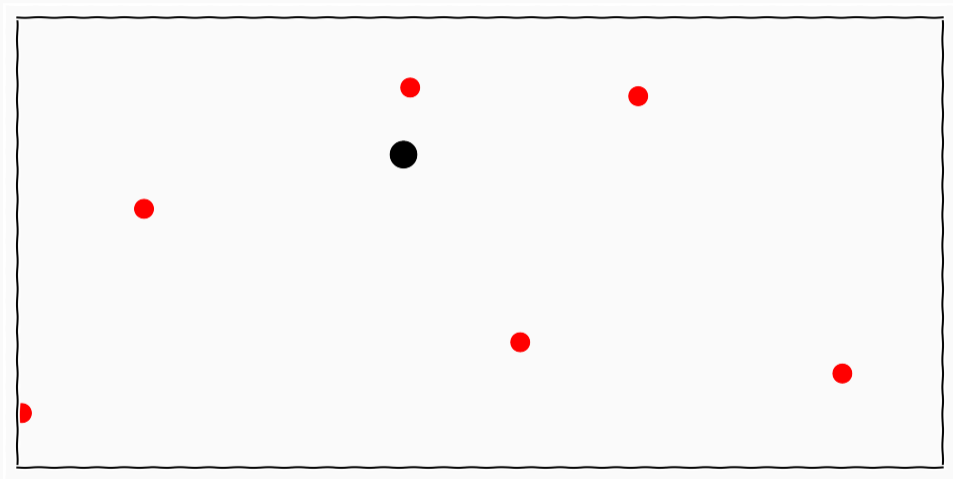
Example



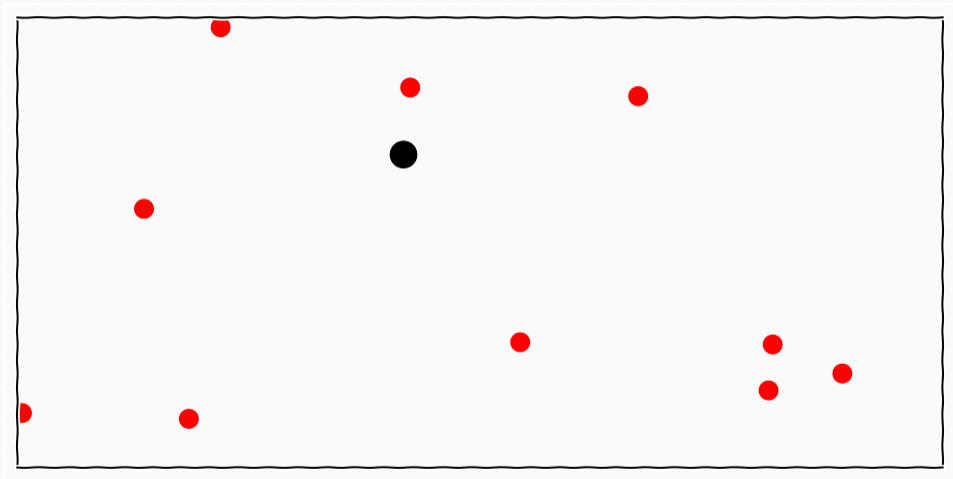
Example



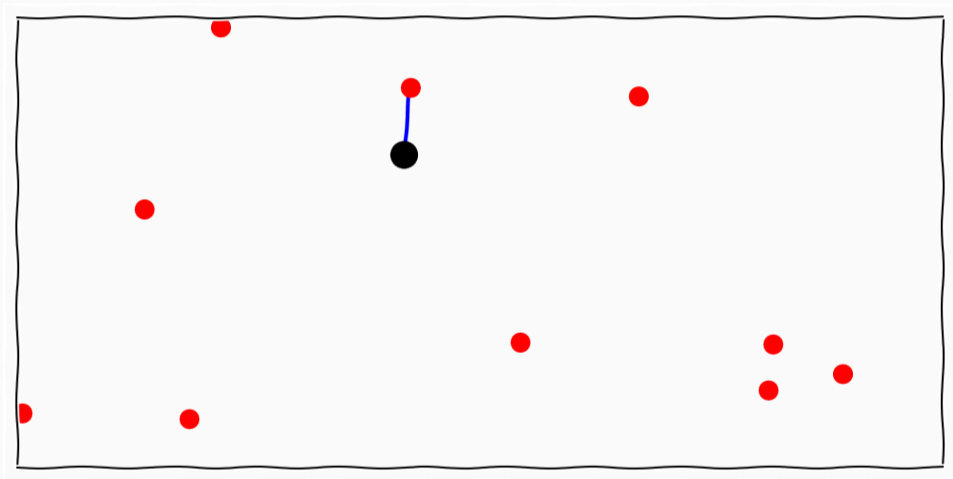
Example



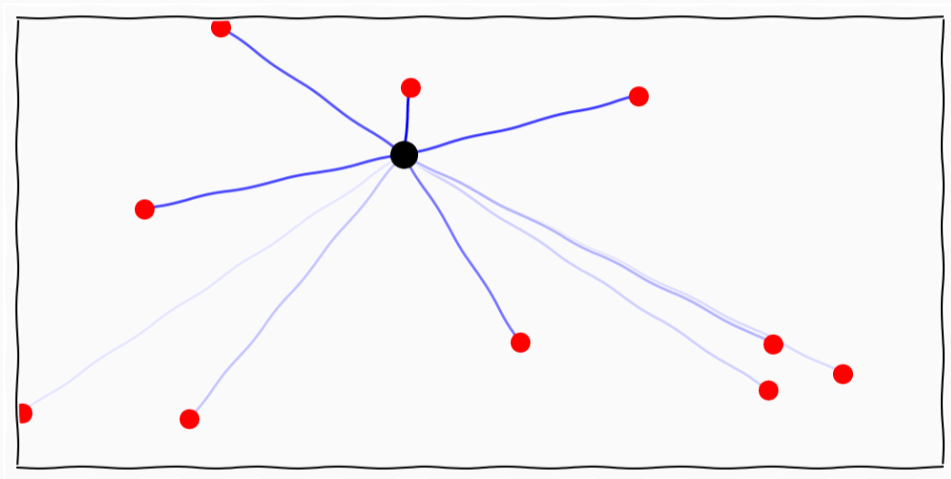
Example



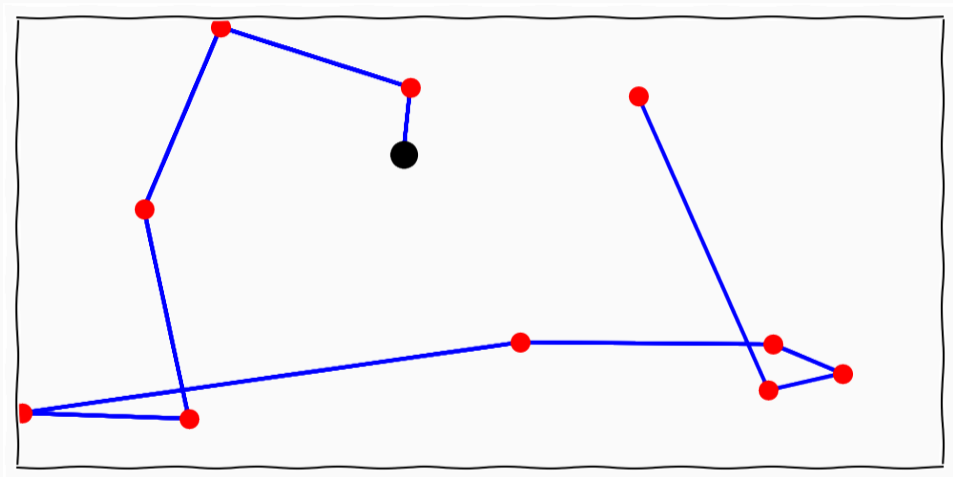
Example



Example

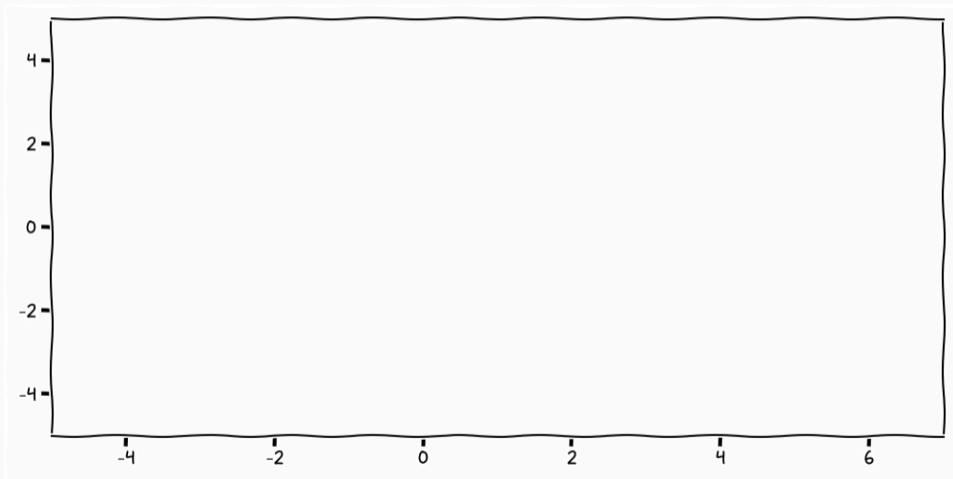


Example

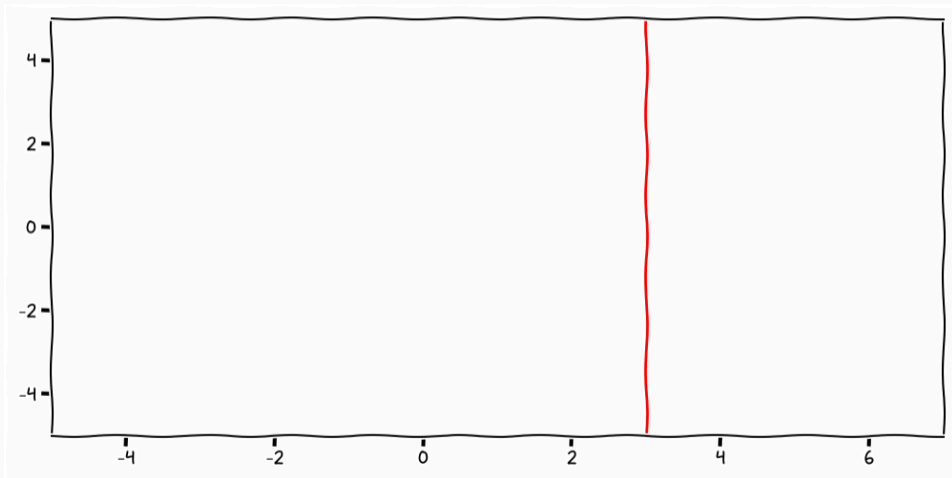


Non-parametric Models

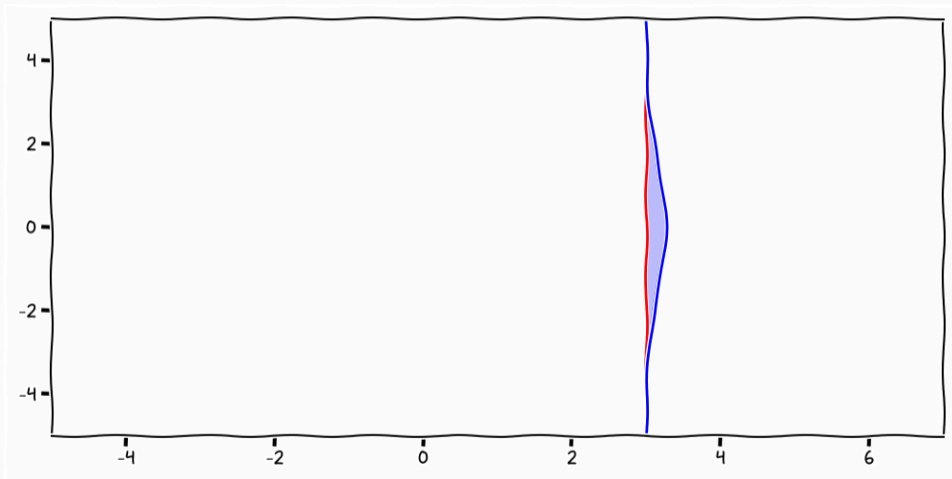
Lets talk about functions



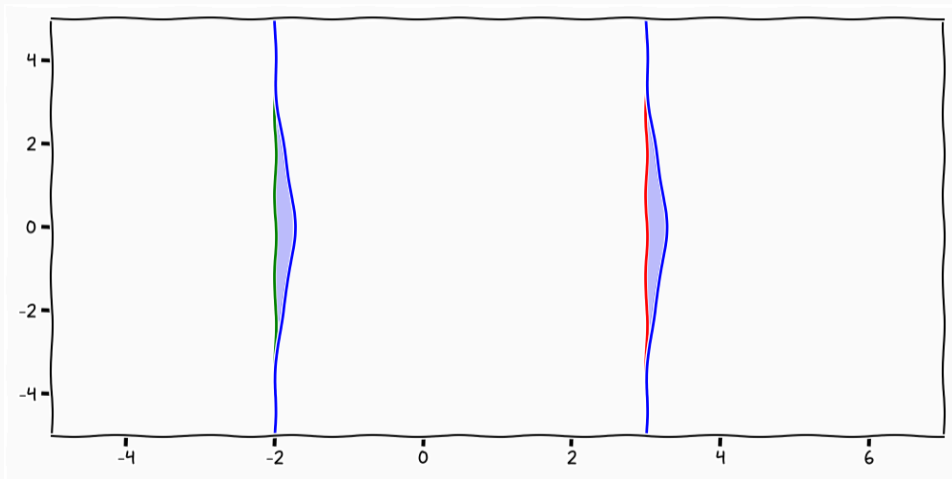
Lets talk about functions



Lets talk about functions



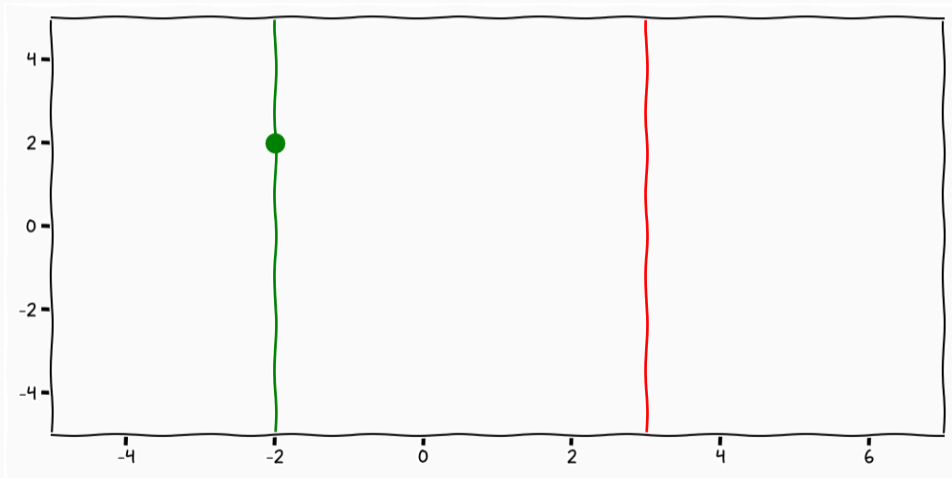
Lets talk about functions



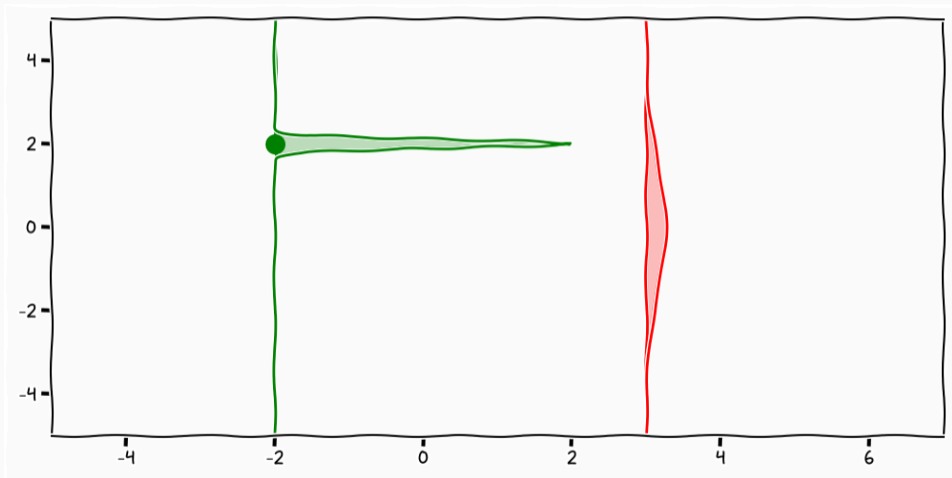
$$f_1 = \mathcal{N}(\mu_1, k_1)$$

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

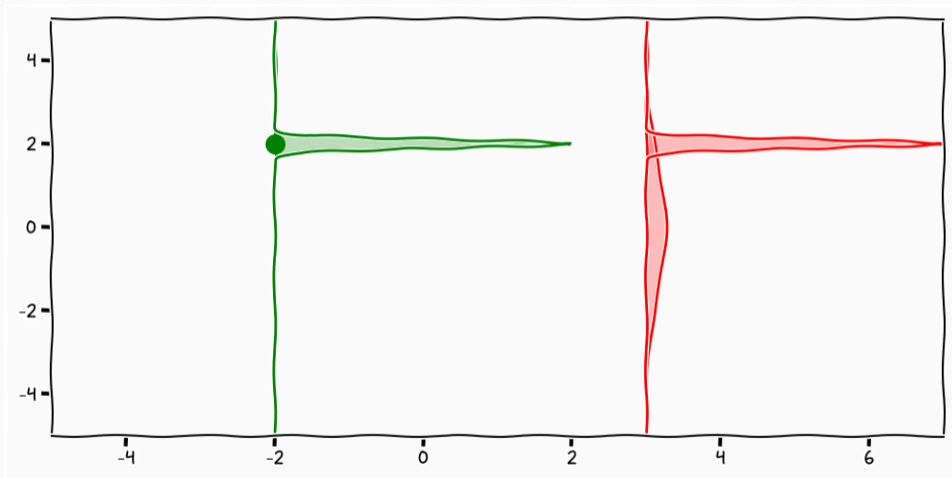
Non-parametric functions



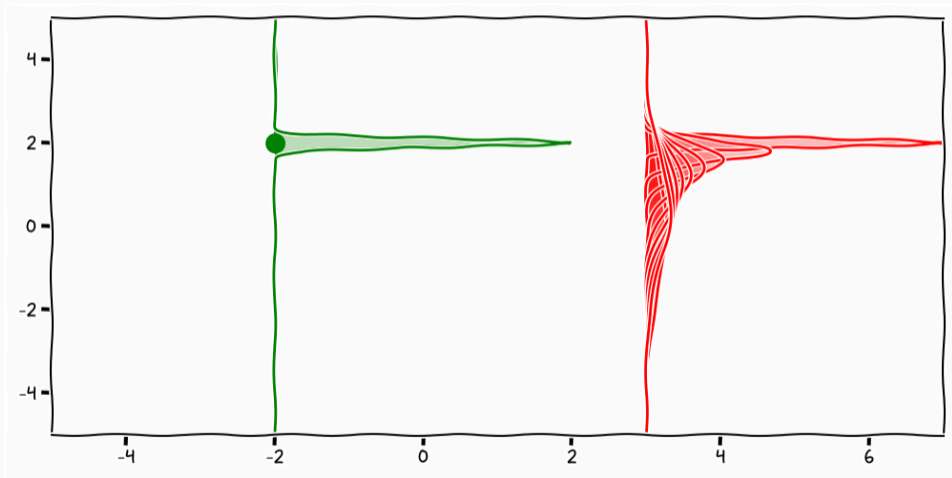
Non-parametric functions



Non-parametric functions

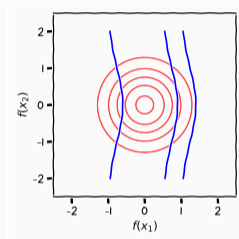
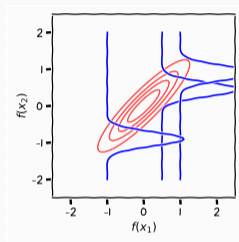
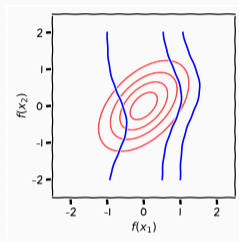


Non-parametric functions



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

Conditional Gaussians

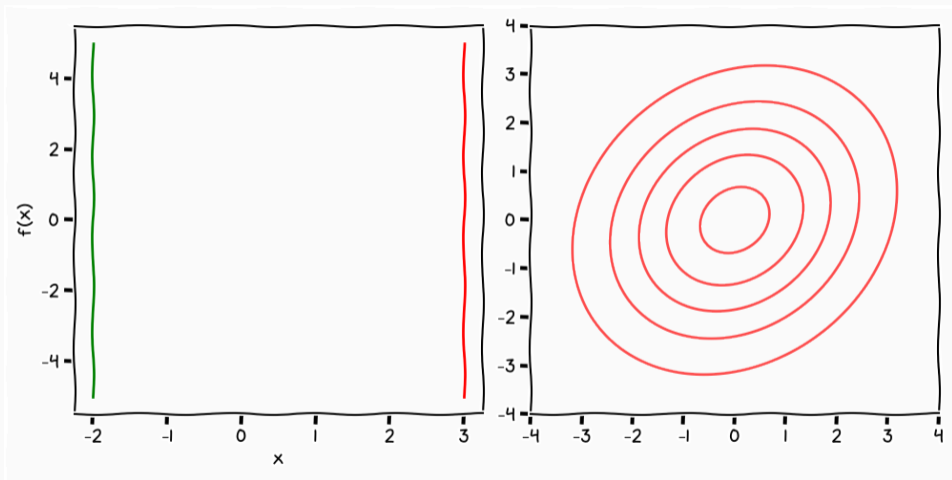


$$N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}\right)$$

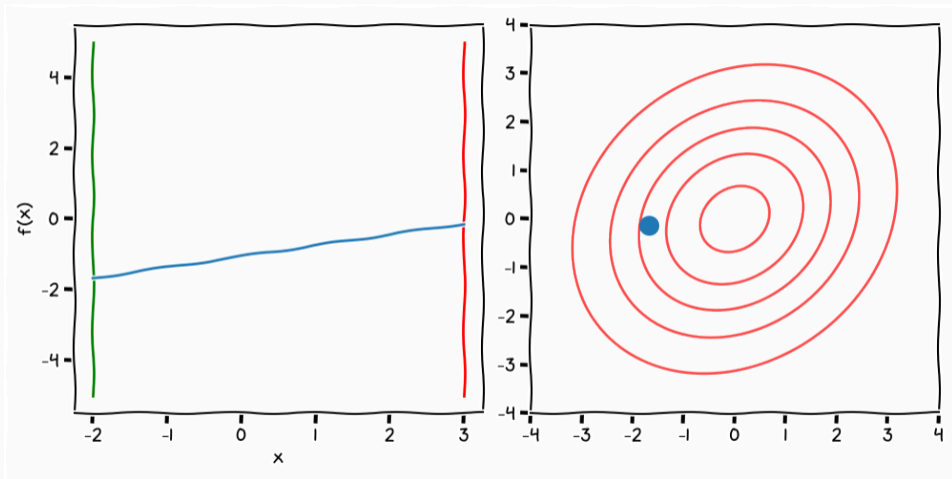
$$N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}\right)$$

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

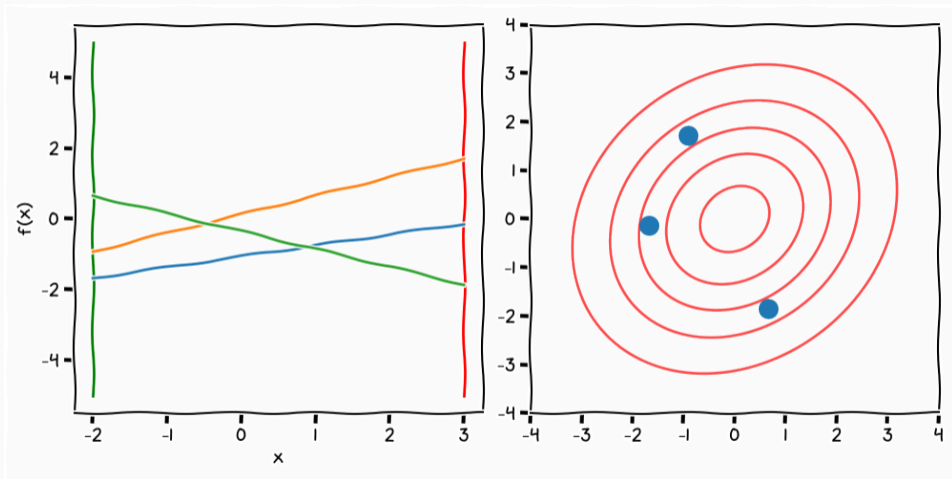
Gaussian Samples



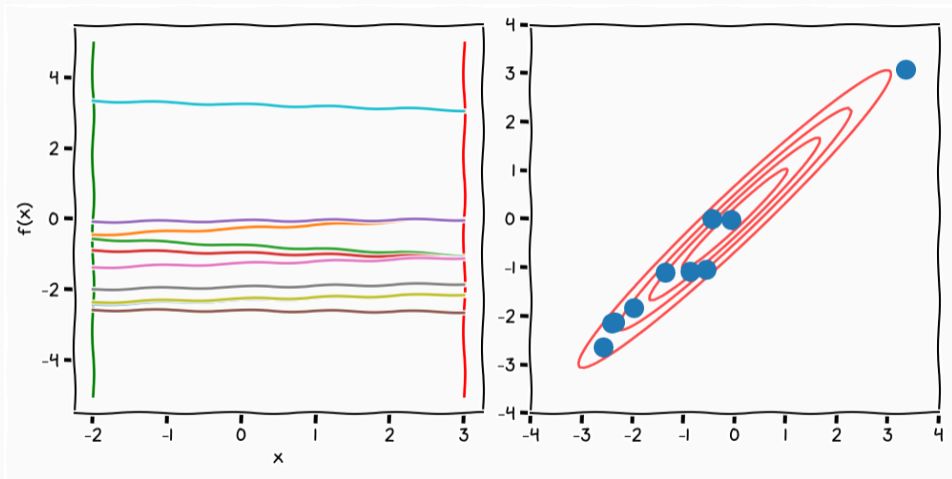
Gaussian Samples



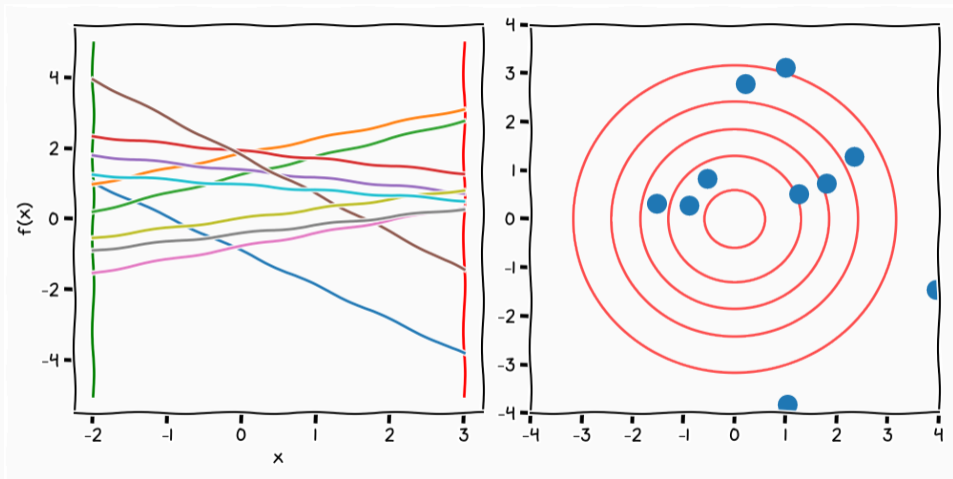
Gaussian Samples



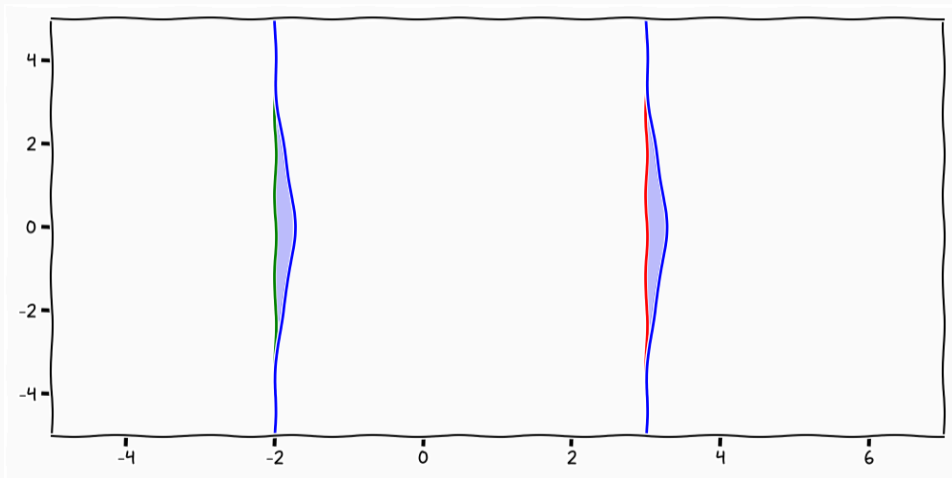
Gaussian Samples



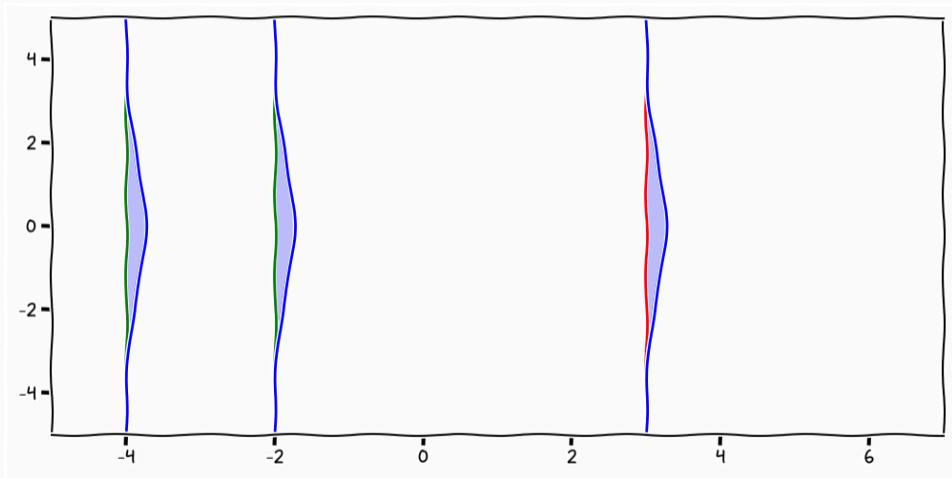
Gaussian Samples



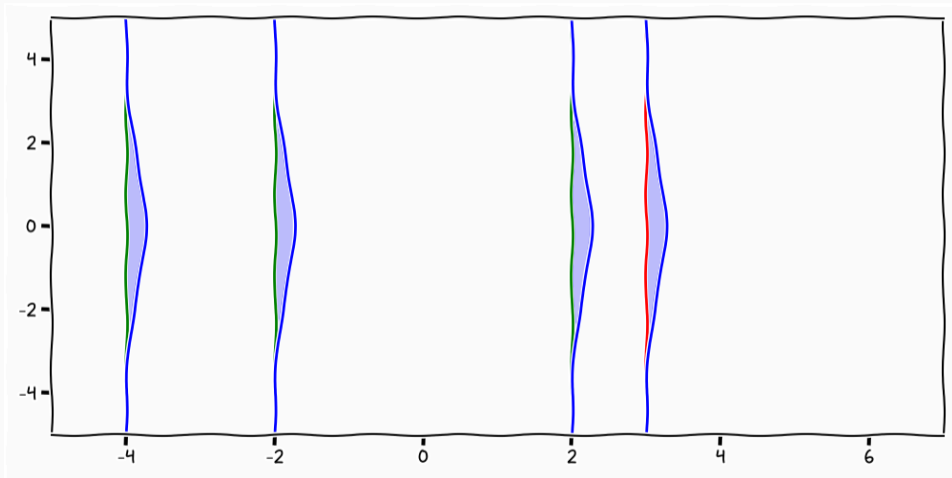
Lets talk about functions



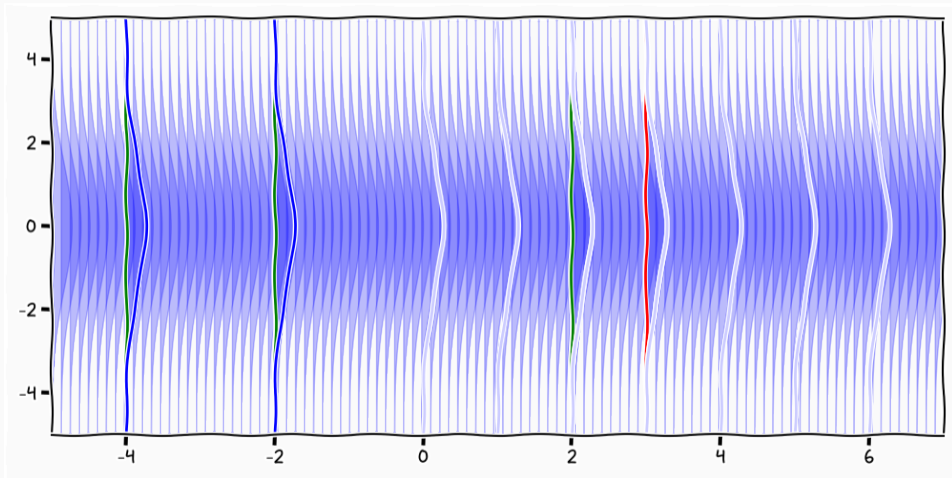
Non-parametric functions



Non-parametric functions



Non-parametric functions



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

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

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

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

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

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

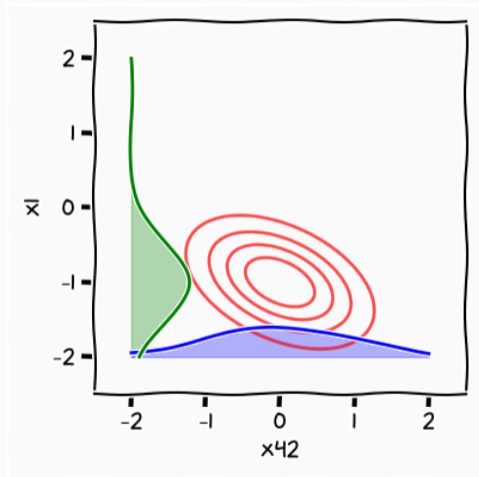
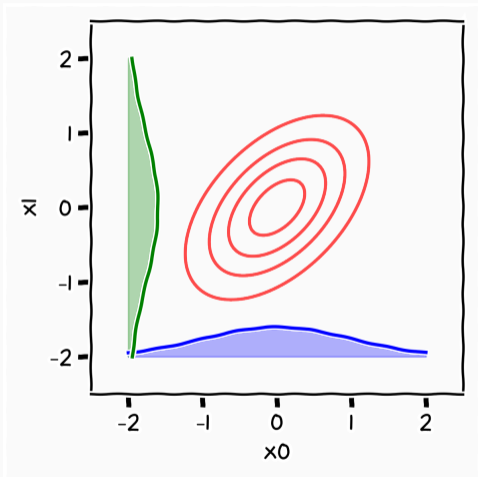
$$p(\mathbf{x}_1, x_2) = \mathcal{N} \left(\begin{array}{c|cc} \mathbf{x}_1 & \mu_1 & k_{11} \\ x_2 & \mu_2 & k_{21} \end{array} \left| \begin{array}{cc} k_{12} & \\ & k_{22} \end{array} \right. \right)$$

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

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

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

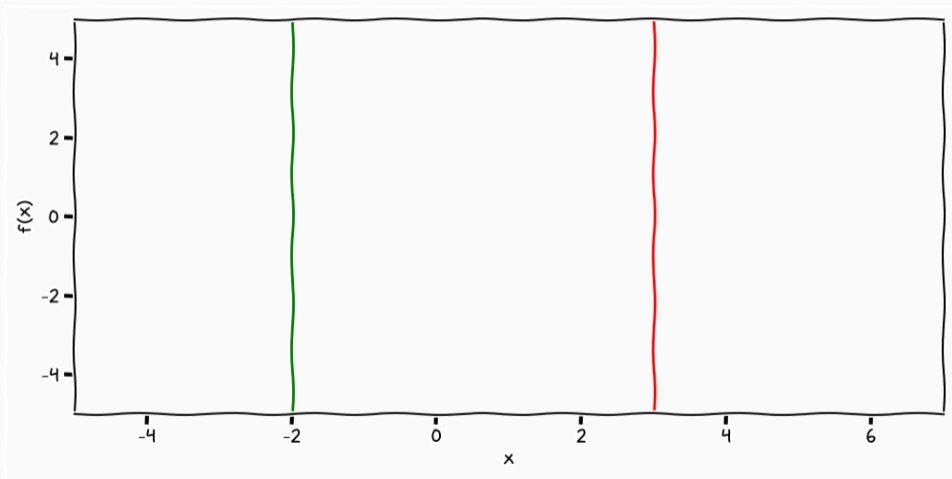
Gaussian Distribution - Marginal



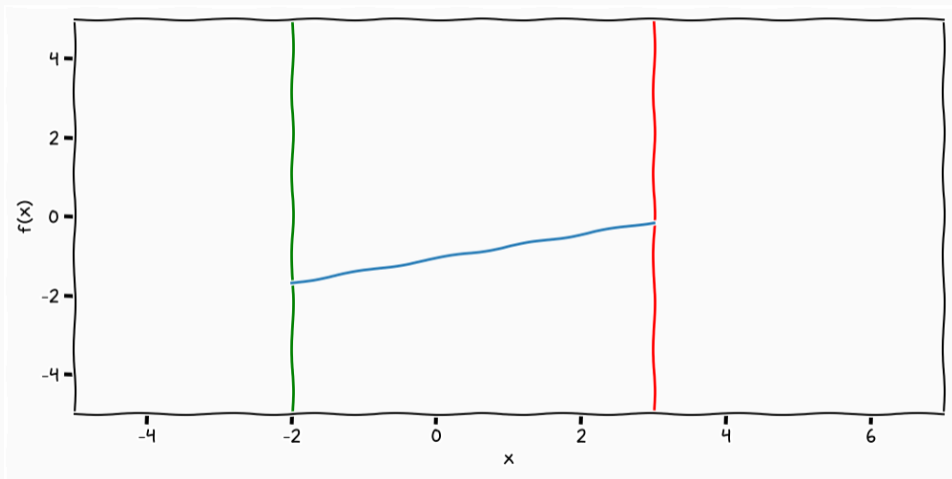
For all measurable sets $F_i \subseteq \mathbb{R}^n$ and probability measure \mathcal{N}

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

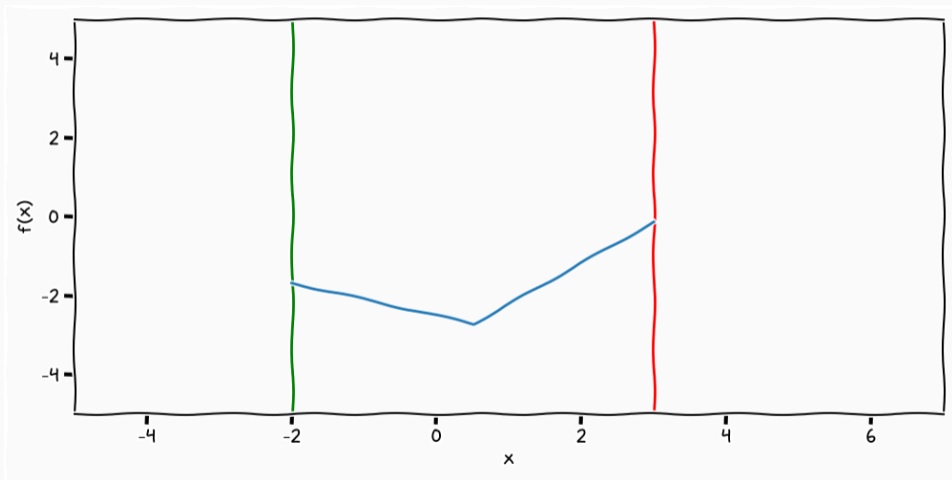
Gaussian Samples



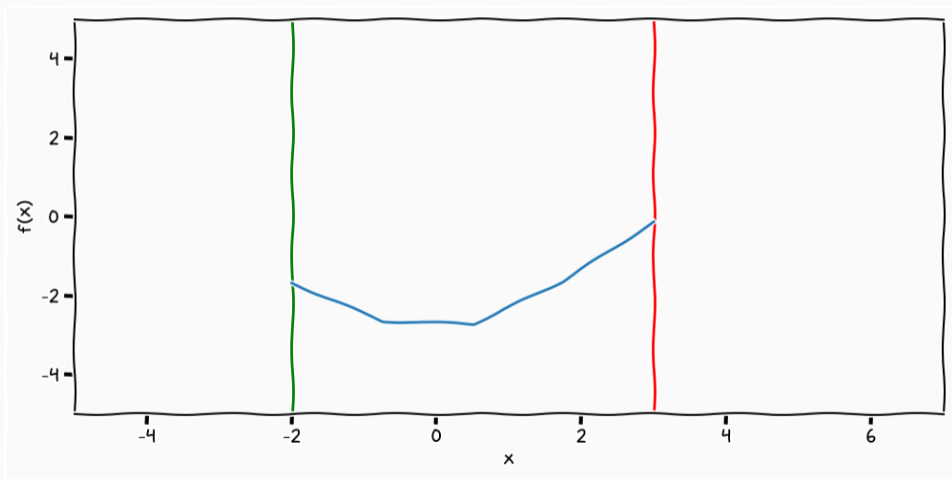
Gaussian Samples



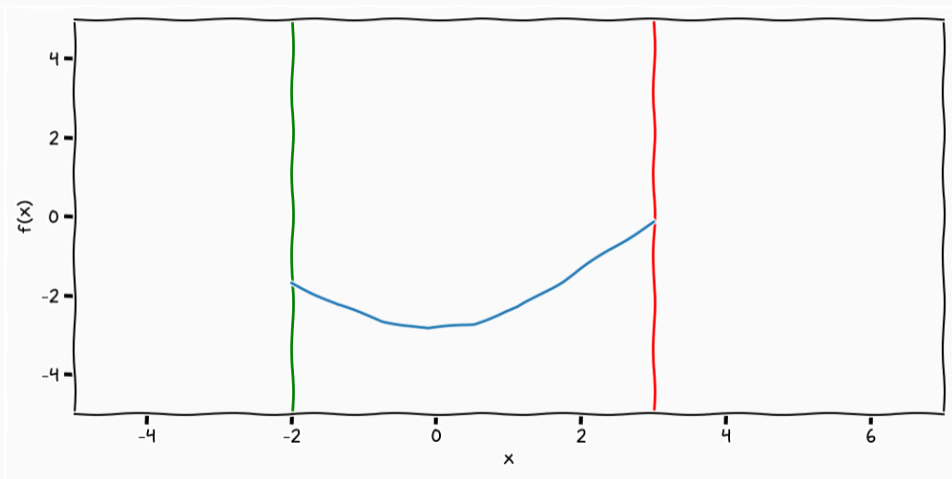
Gaussian Samples



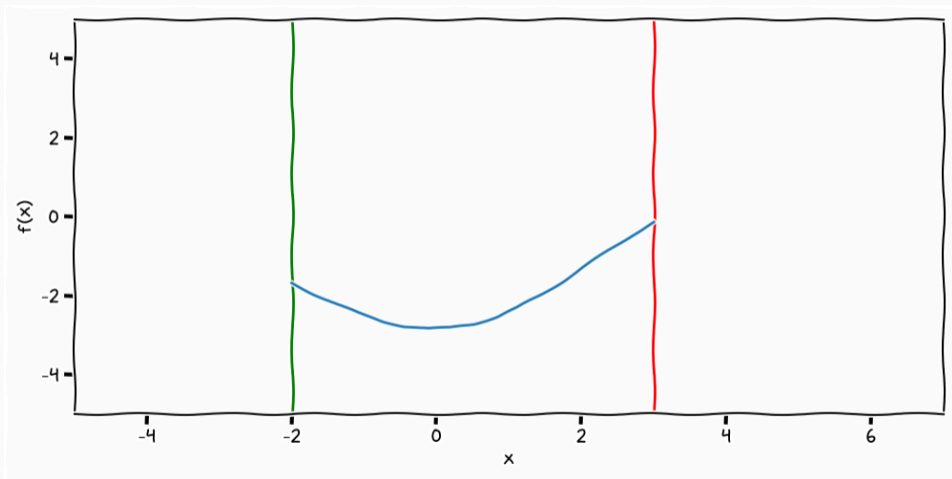
Gaussian Samples



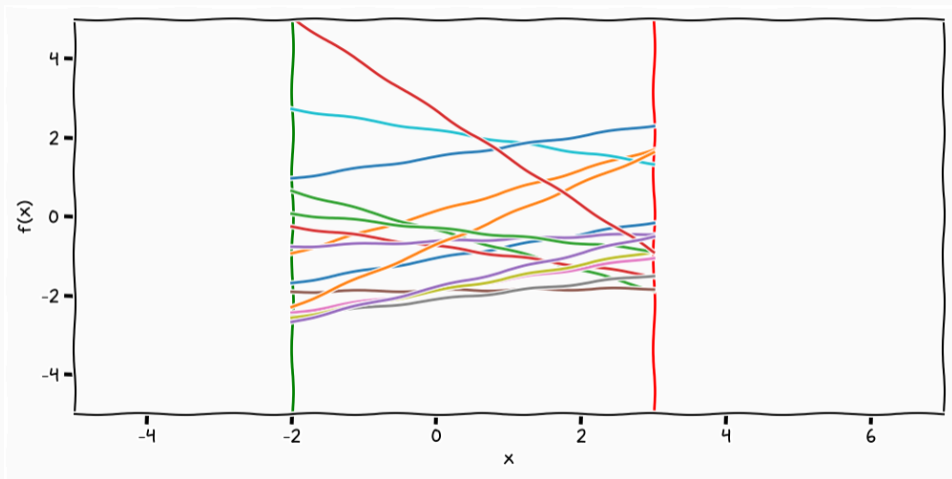
Gaussian Samples



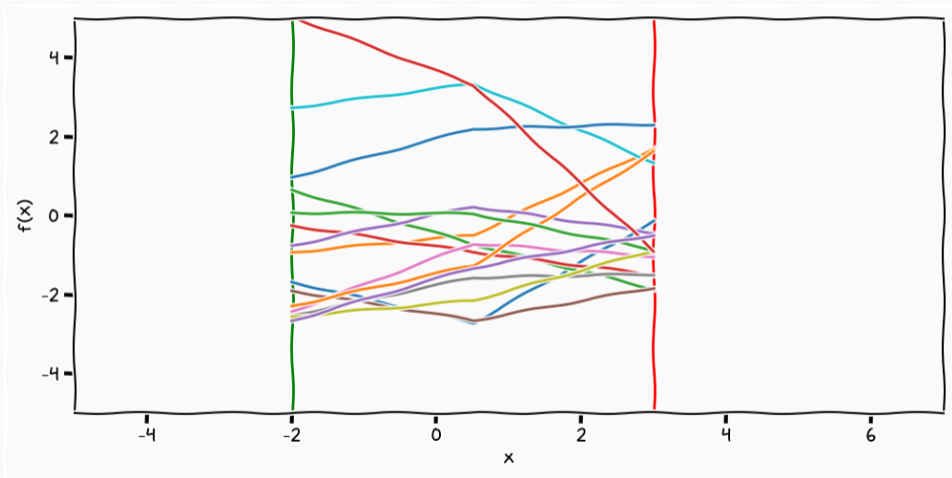
Gaussian Samples



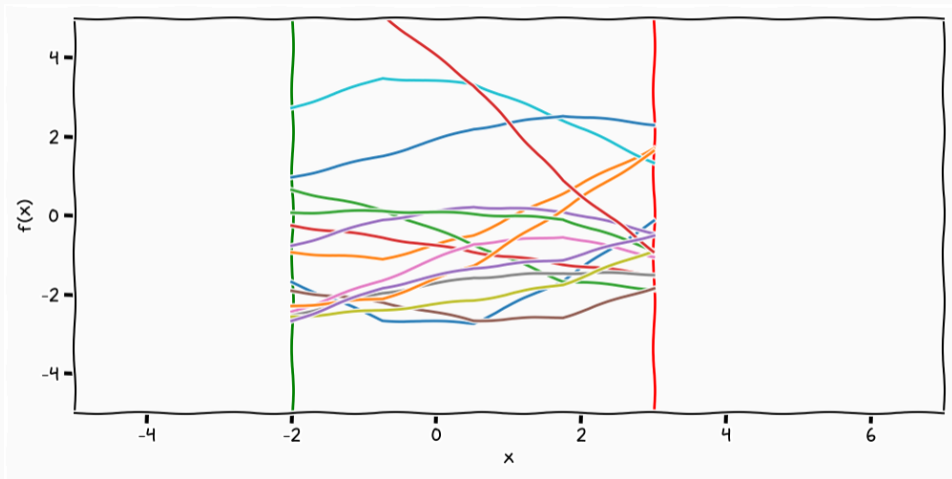
Gaussian Samples



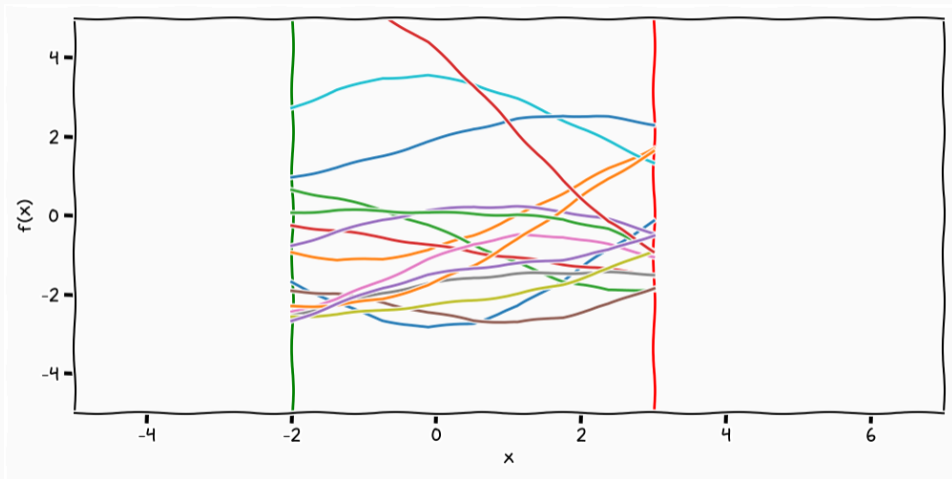
Gaussian Samples



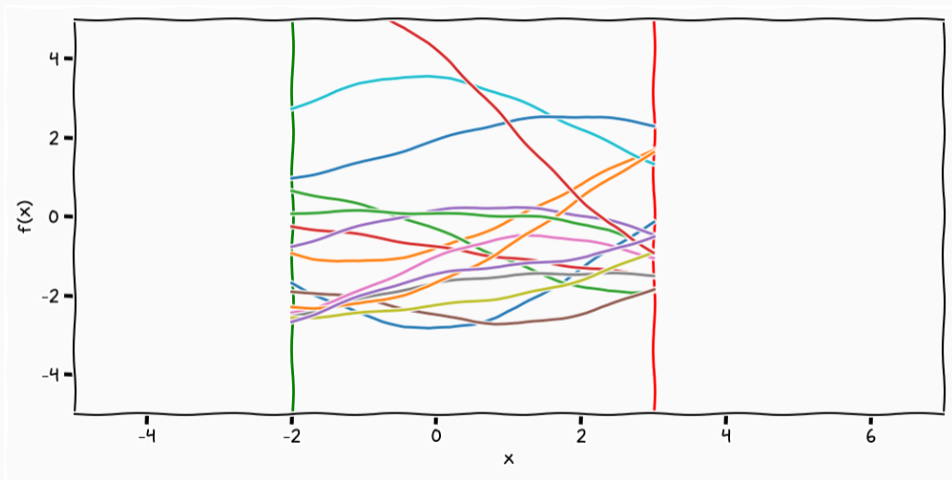
Gaussian Samples



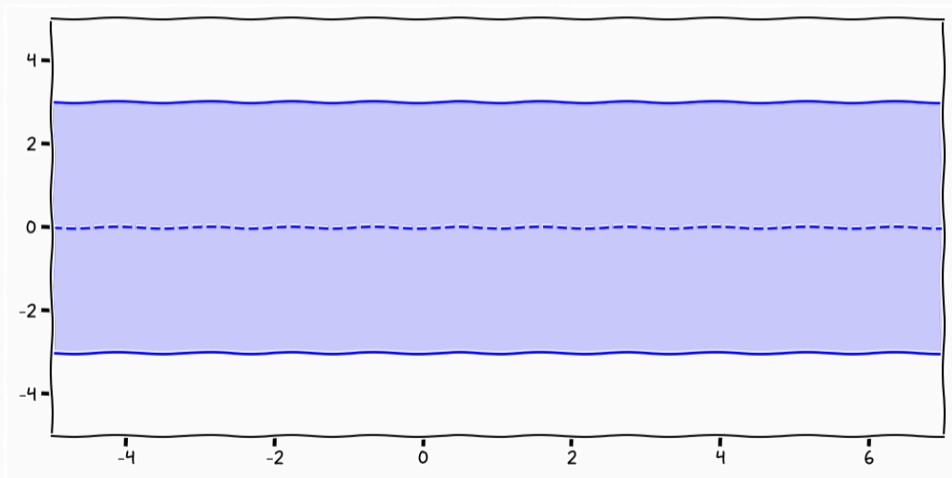
Gaussian Samples



Gaussian Samples



Gaussian Processes



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

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

The Gaussian distribution is the projection of the infinite Gaussian process

Definition (Gaussian Process)

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

$$p(\mathbf{f}) = \mathcal{N} \left(\begin{array}{c} \left[\begin{array}{c} f_1 \\ f_2 \\ \vdots \\ f_N \\ \vdots \end{array} \right] \parallel \left[\begin{array}{c} \mu(x_1) \\ \mu(x_2) \\ \vdots \\ \mu(x_N) \\ \vdots \end{array} \right], \left[\begin{array}{cccccc} k(x_1, x_1) & k(x_1, x_2) & \dots & k(x_1, x_N) & \dots \\ k(x_2, x_1) & k(x_2, x_2) & \dots & k(x_2, x_N) & \dots \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ k(x_N, x_1) & k(x_N, x_2) & \dots & k(x_N, x_N) & \dots \\ \vdots & \vdots & \dots & \vdots & \ddots \end{array} \right] \end{array} \right)$$

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

- We parameterise the covariance as a function of the input

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

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

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

- 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

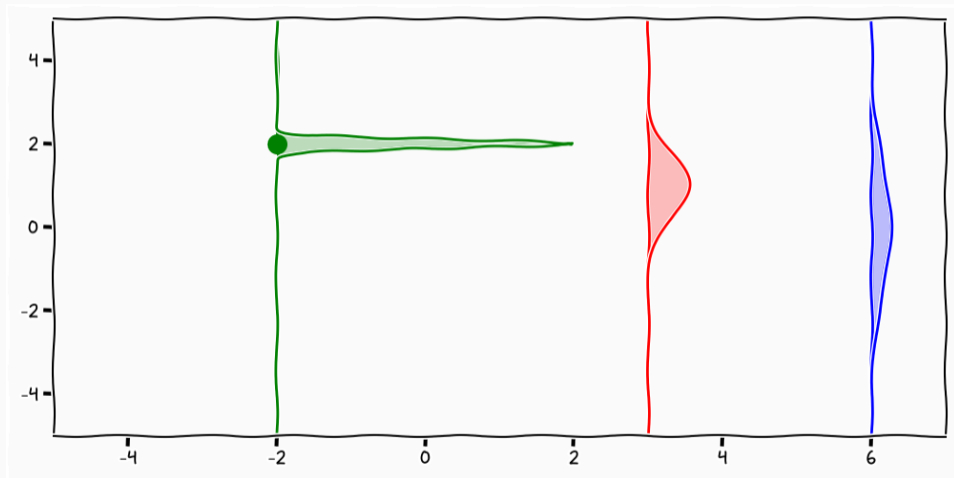
$$k_{ij} = k(x_i, x_j)$$

- 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*

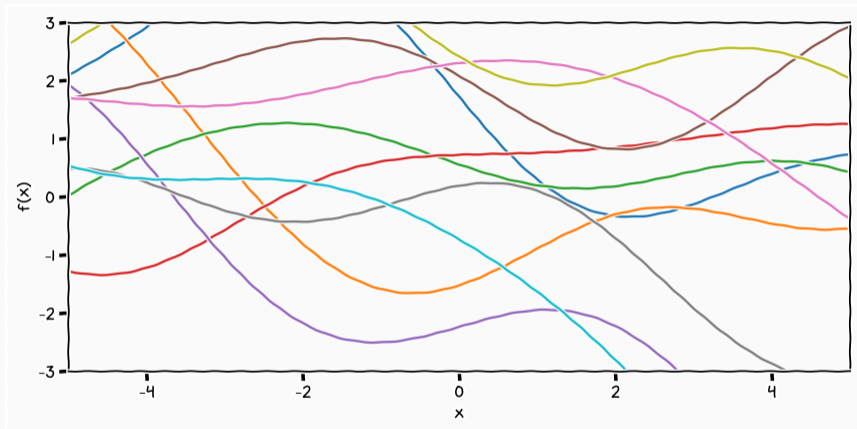
$$k_{ij} = k(x_i, x_j)$$

- 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

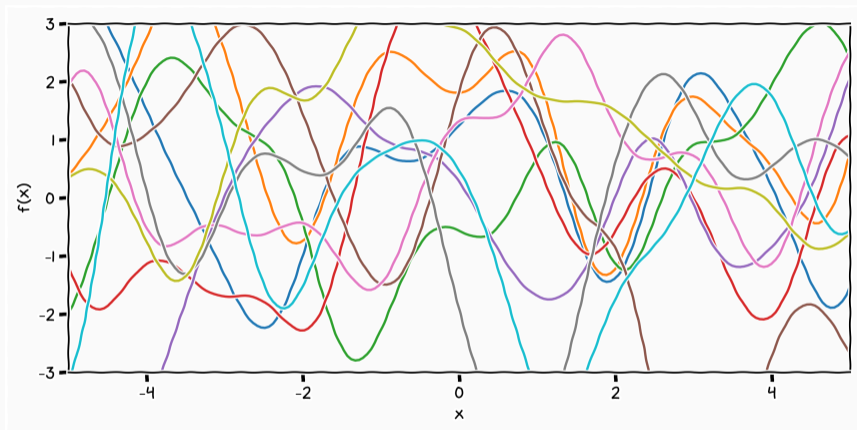


Gaussian Processes Samples



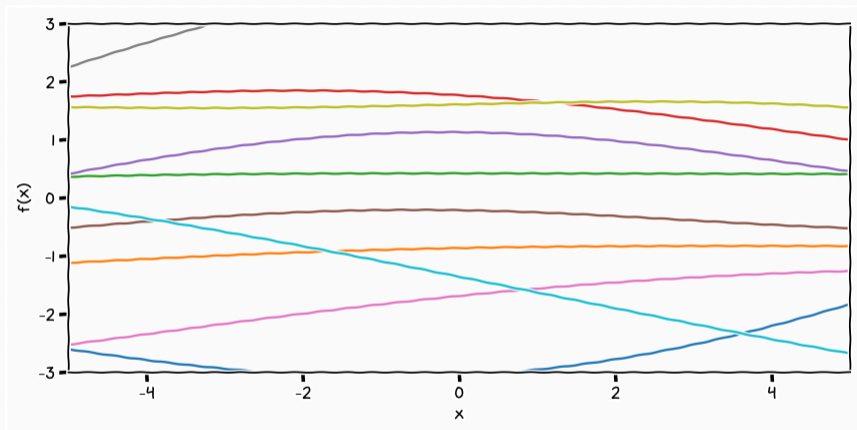
$$k(x_i, x_j) = 3 \cdot e^{-\frac{(x_i - x_j)^2}{15}}$$

Gaussian Processes Samples



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

Gaussian Processes Samples



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

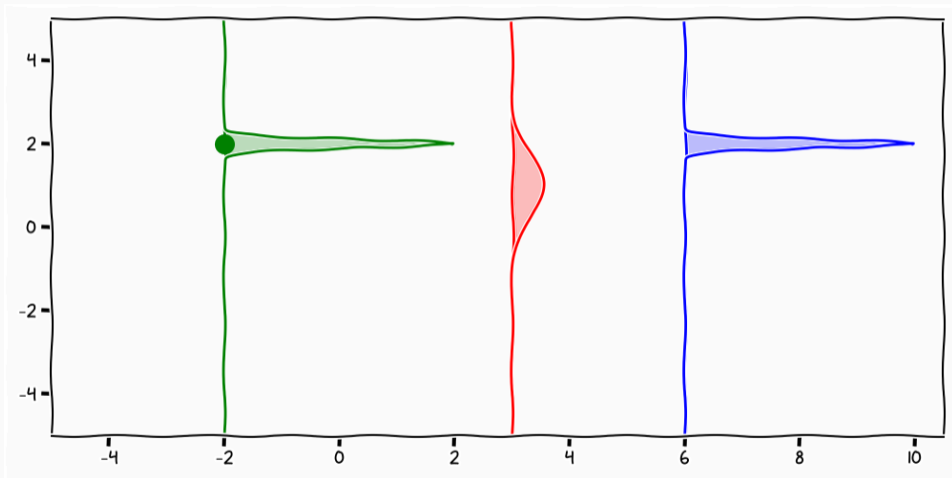
Code

```
x = np.linspace(-5,5,200)
x = x.reshape((-1,1))

Sigma = 3.0*np.exp(-np.power(cdist(x,x),2)/lengthScale)
mu = np.zeros(x.shape)

y = np.random.multivariate_normal(mu.flatten(),Sigma,10)
ax.plot(x,y.T)
```

Gaussian Processes



$$k(x, x') = ck_1(x, x')$$

$$k(x, x') = f(x)k_1(x, x')f(x')$$

$$k(x, x') = q(k_1(x, x'))$$

$$k(x, x') = \exp(k_1(x, x'))$$

$$k(x, x') = k_1(x, x') + k_2(x, x')$$

$$k(x, x') = k_1(x, x')k_2(x, x')$$

$$k(x, x') = k_3(\phi(x), \phi(x'))$$

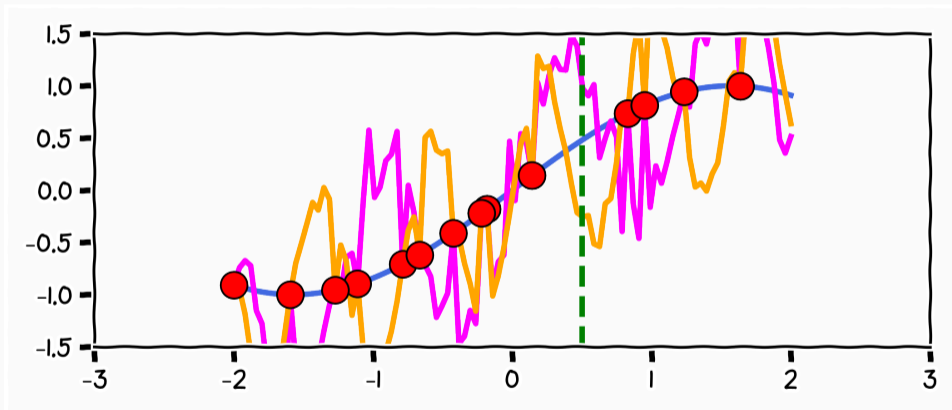
$$k(x, x') = x^T \mathbf{A}x'$$

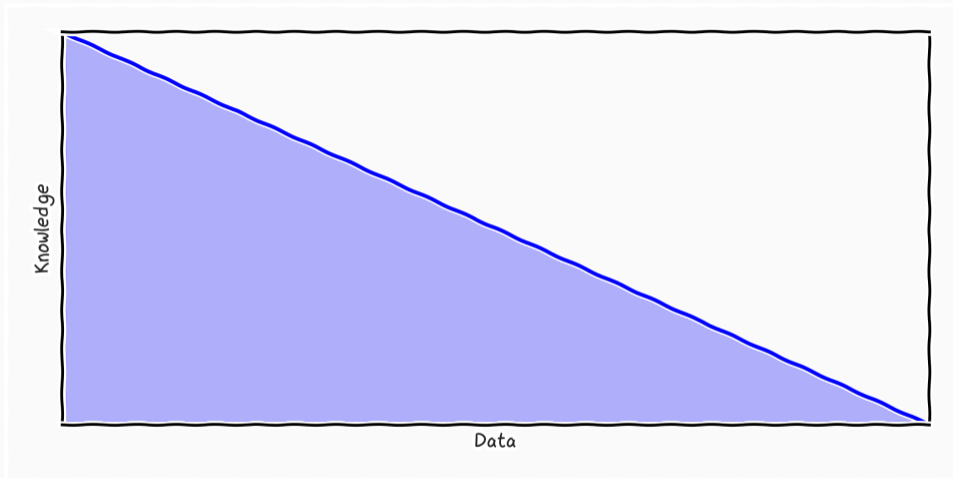
$$k(x, x') = k_a(x_a, x'_a) + k_b(x_b, x'_b)$$

$$k(x, x') = k_a(x_a, x'_a)k_b(x_b, x'_b)$$

¹Bishop, 2006.

Curve Fitting





Inference

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

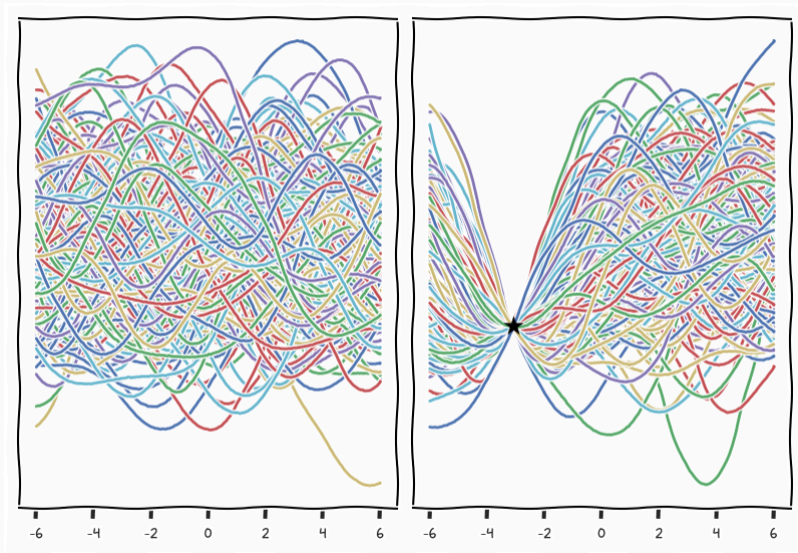
$$\int p(\mathbf{f}, \mathbf{f}_*) d\mathbf{f}_* = \int p(\mathbf{f} | \mathbf{f}_*) p(\mathbf{f}_*) d\mathbf{f}_*$$

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

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

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

Gaussian Processes: Posterior Samples



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

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

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

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

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

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

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

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

- All instantiations are jointly Gaussian

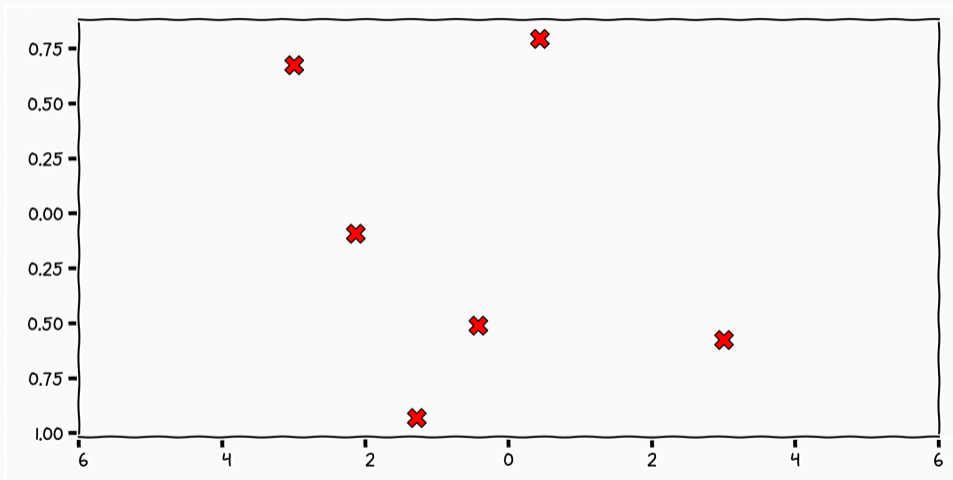
$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} k(\mathbf{x}, \mathbf{x}) & k(\mathbf{x}, \mathbf{x}_*) \\ k(\mathbf{x}_*, \mathbf{x}) & k(\mathbf{x}_*, \mathbf{x}_*) \end{bmatrix} \right)$$

- All instantiations are jointly Gaussian

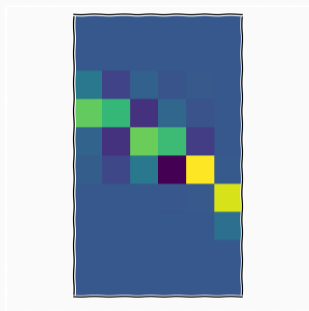
$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} k(\mathbf{x}, \mathbf{x}) & k(\mathbf{x}, \mathbf{x}_*) \\ k(\mathbf{x}_*, \mathbf{x}) & k(\mathbf{x}_*, \mathbf{x}_*) \end{bmatrix} \right)$$

- Conditional Gaussian

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

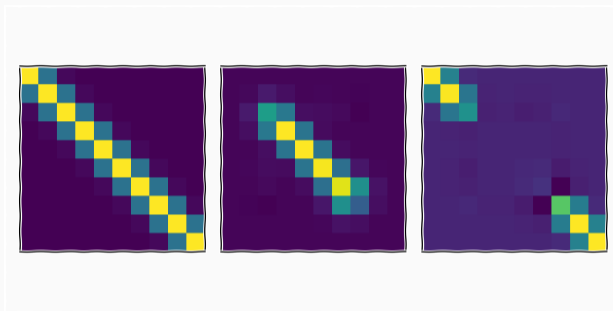


Does it make sense: Mean



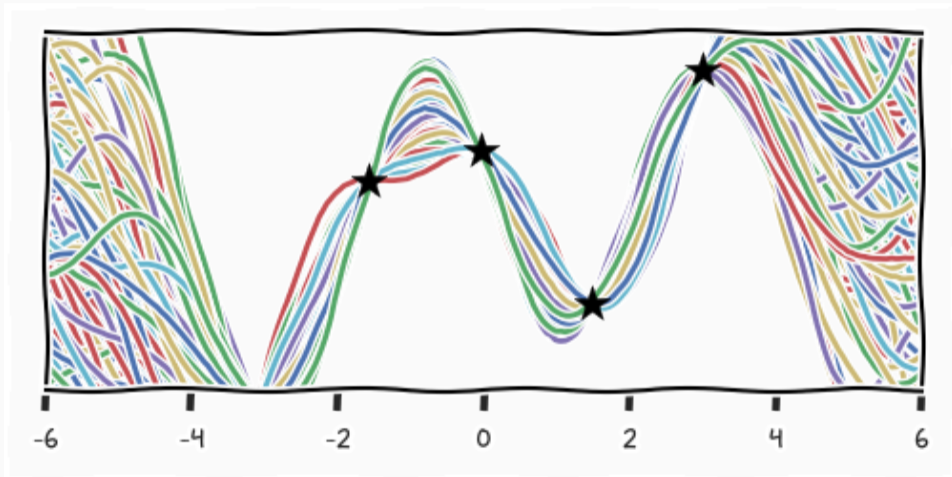
$$k(\mathbf{x}_*, \mathbf{X})^T k(\mathbf{X}, \mathbf{X})^{-1} \mathbf{f}$$

Does it make sense: Covariance

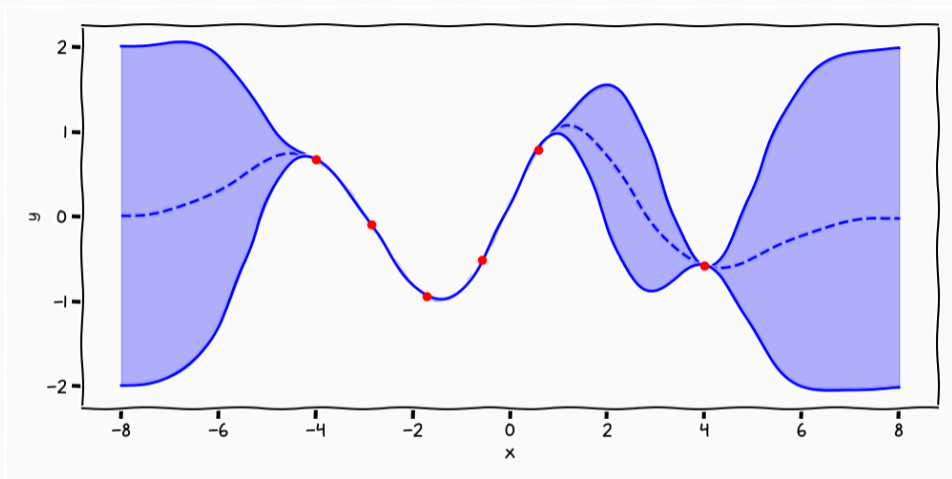


$$k(\mathbf{x}_*, \mathbf{x}_*) - k(\mathbf{x}_*, \mathbf{x})^T k(\mathbf{x}, \mathbf{x})^{-1} k(\mathbf{x}, \mathbf{x}_*)$$

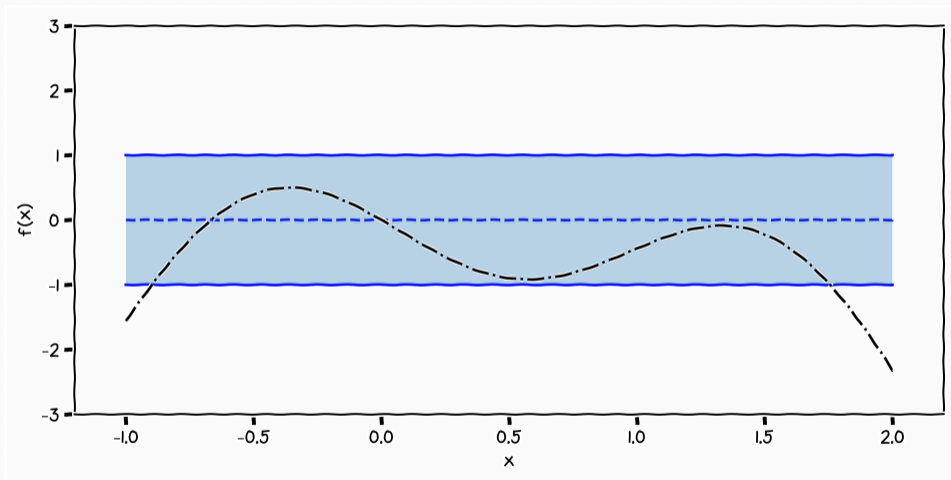
Gaussian Processes: "Predictive Posterior Samples"



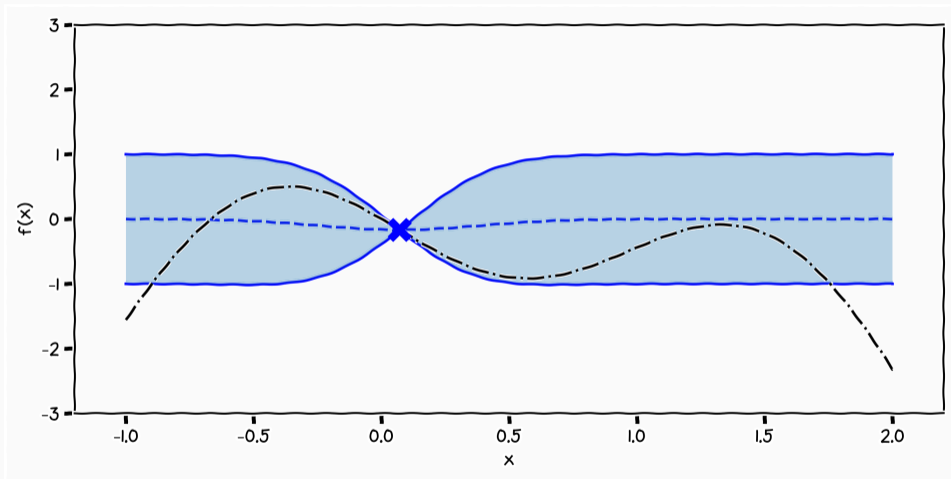
Gaussian Processes: "Predictive Posterior Process"



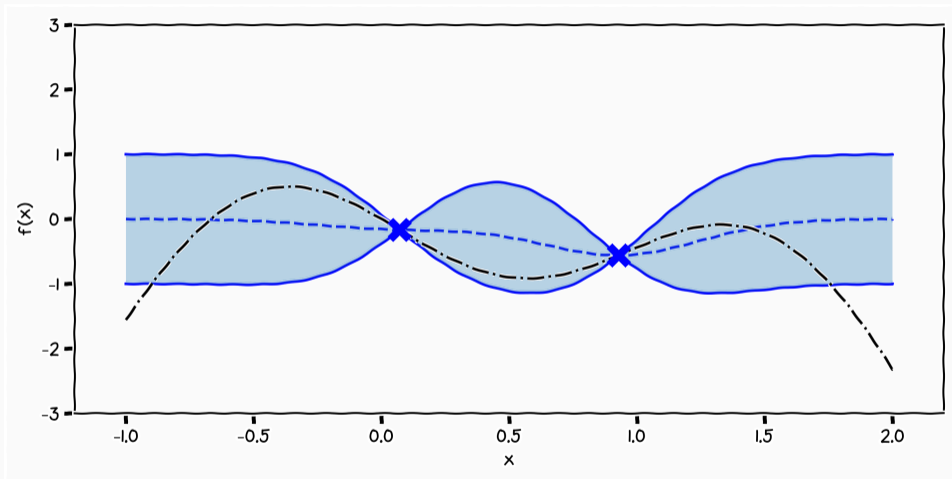
Posterior Processes



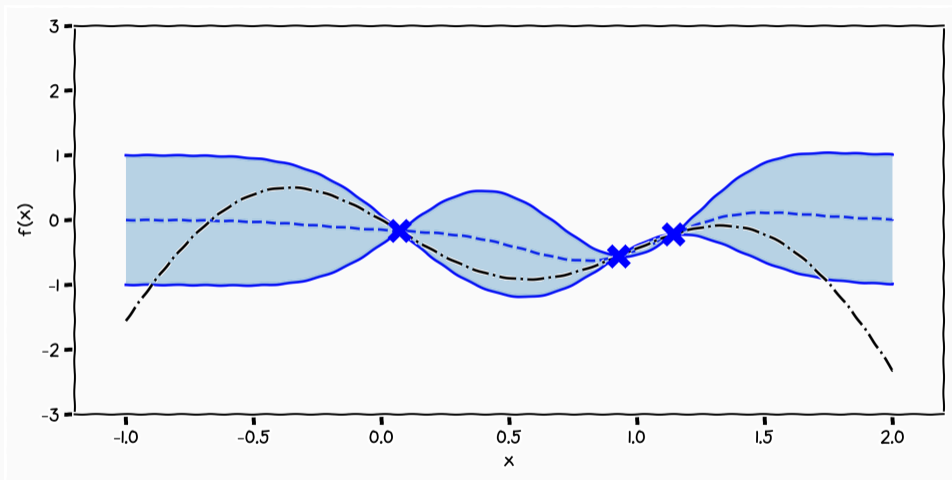
Posterior Processes



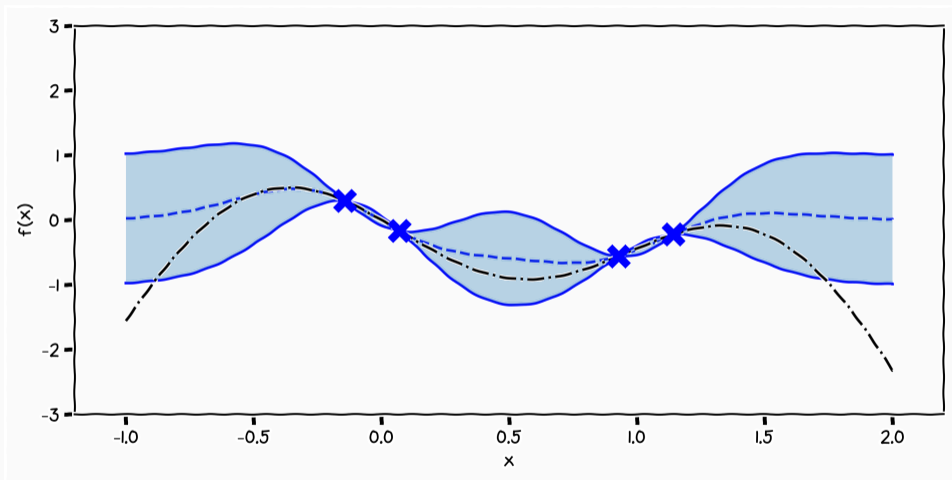
Posterior Processes



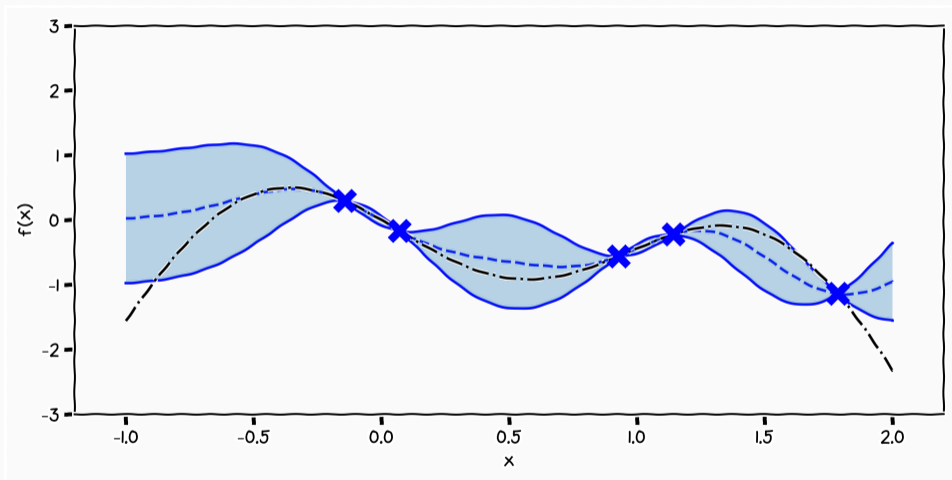
Posterior Processes



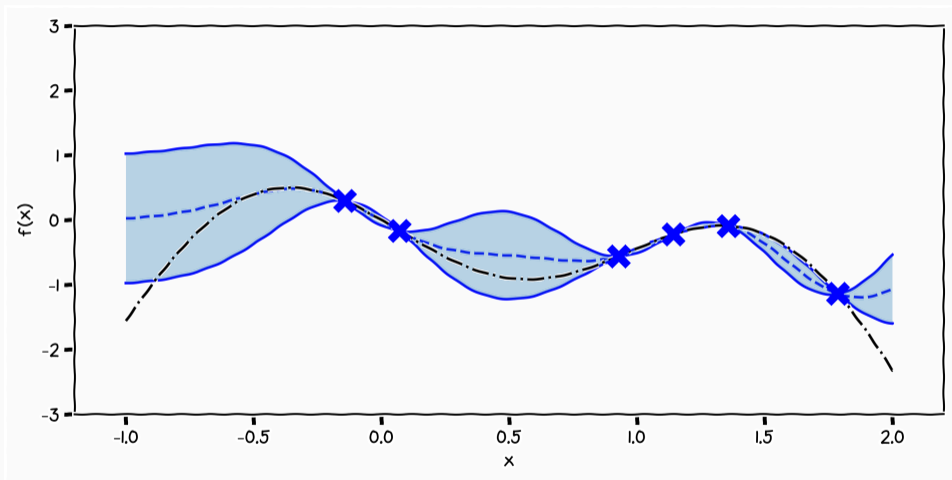
Posterior Processes



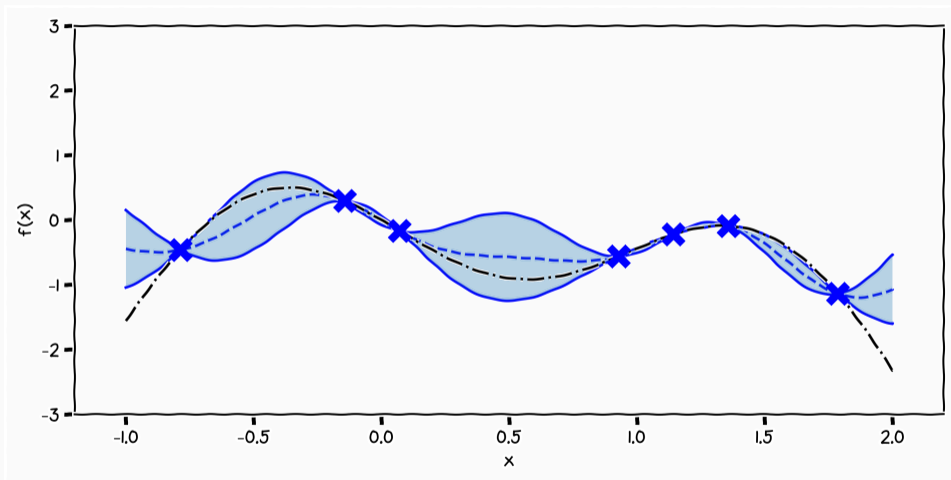
Posterior Processes



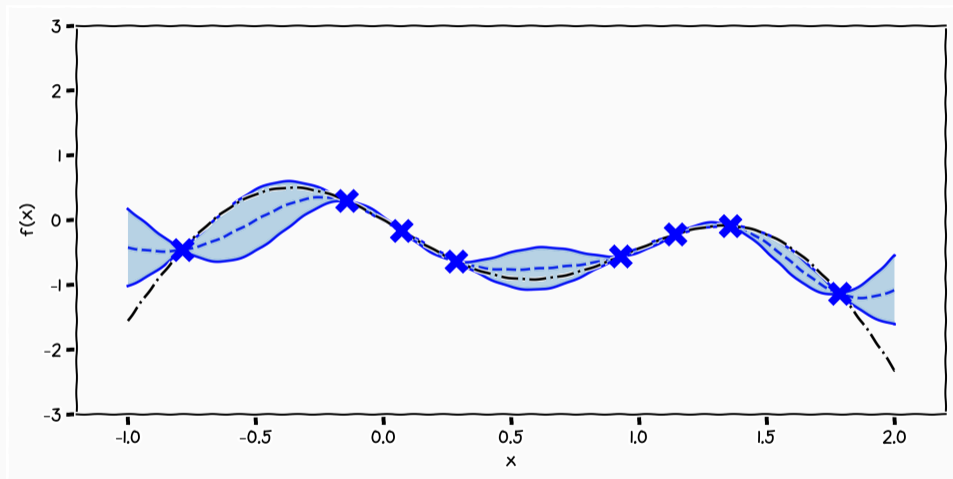
Posterior Processes



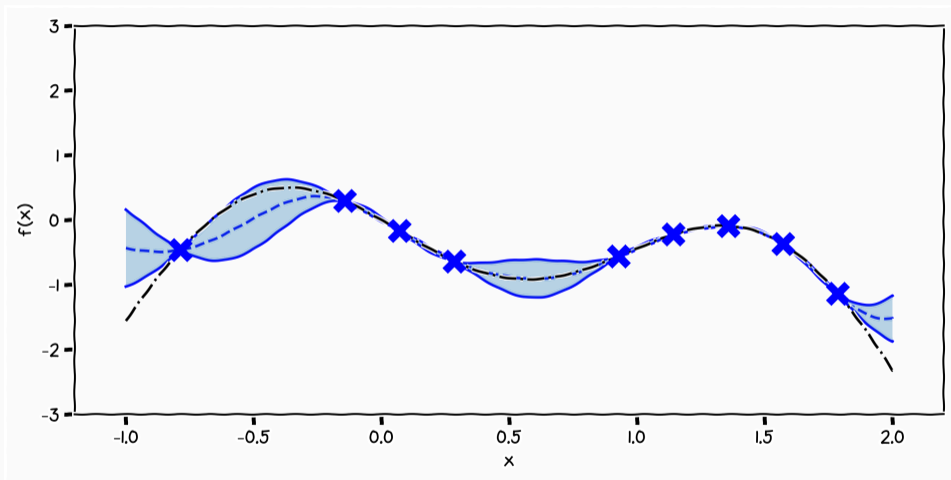
Posterior Processes



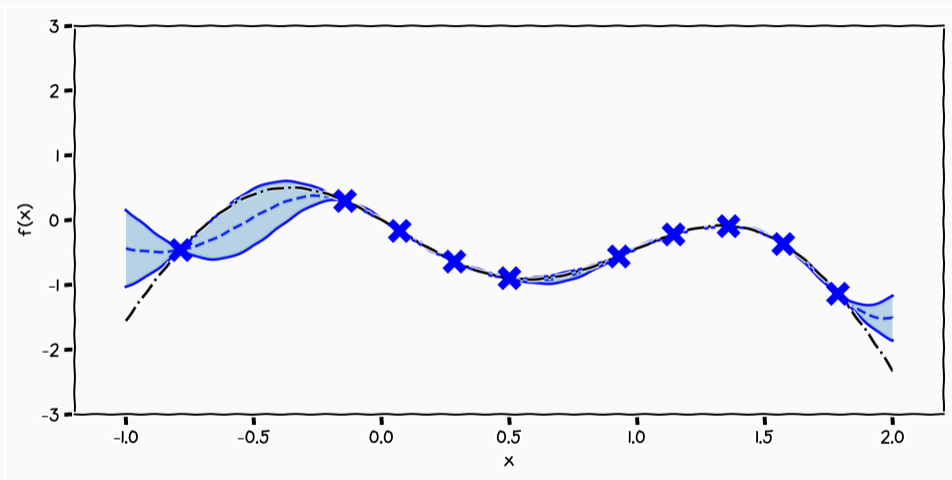
Posterior Processes



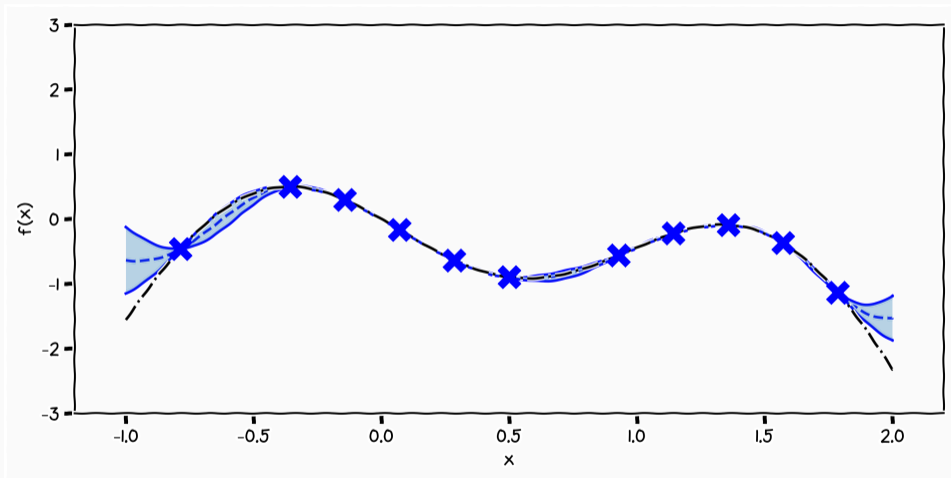
Posterior Processes



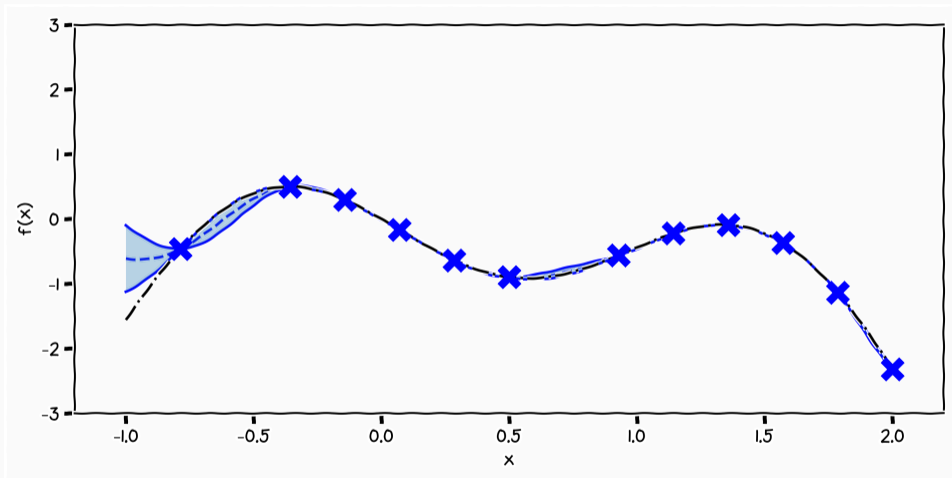
Posterior Processes



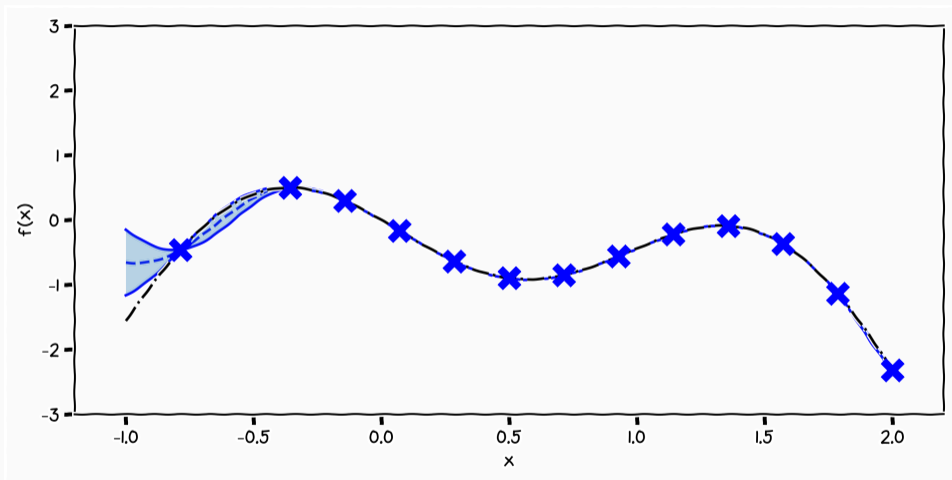
Posterior Processes



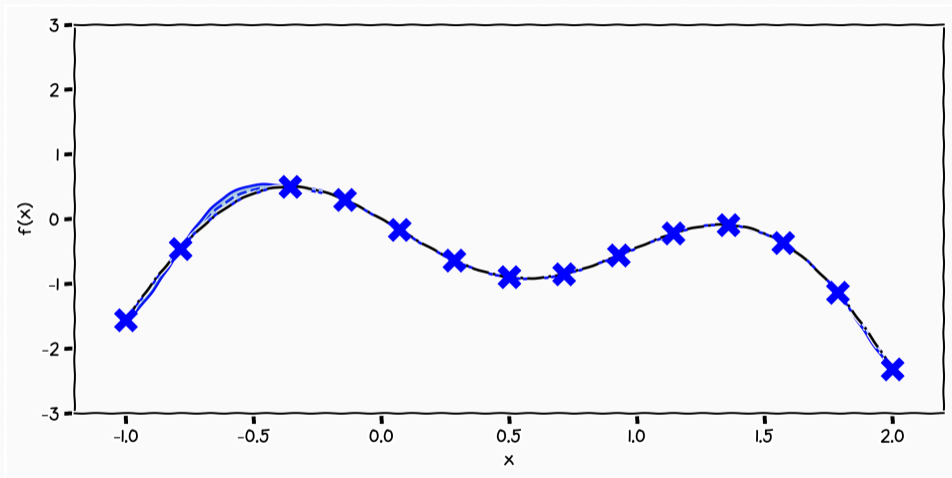
Posterior Processes



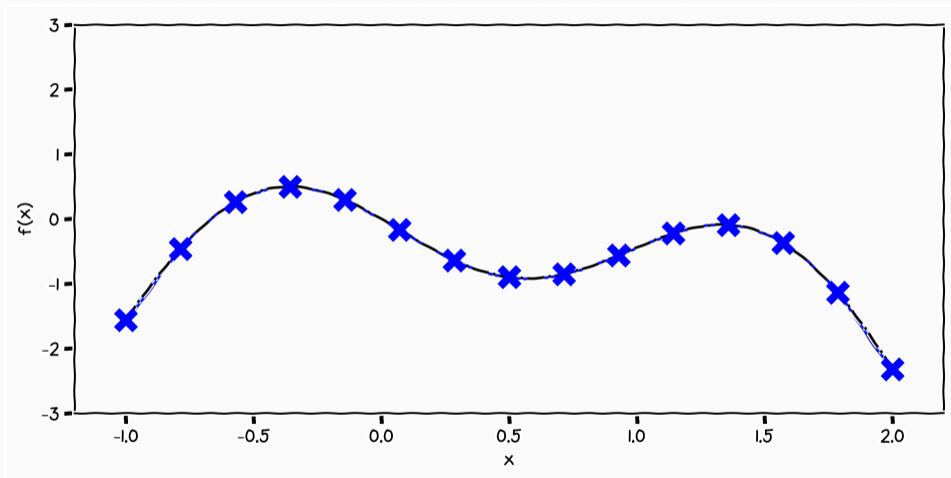
Posterior Processes



Posterior Processes



Posterior Processes



$$p(\mathbf{f}) \sim \mathcal{N}(\mathbf{f} \mid \mu(\cdot), k(\cdot, \cdot)), p(\mathbf{f}_* \mid \mathbf{f}) = \mathcal{N}(\mathbf{f}_*(\mathbf{x}_*, \mathbf{x})^\top k(\mathbf{x}, \mathbf{x})^{-1} \mathbf{f}, \\ k(\mathbf{x}_*, \mathbf{x}_*) - k(\mathbf{x}_*, \mathbf{x})^\top k(\mathbf{x}, \mathbf{x})^{-1} k(\mathbf{x}, \mathbf{x}_*))$$

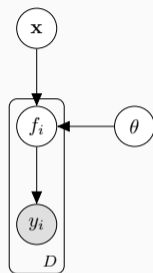
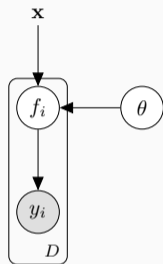
- we have defined a measure over functions

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

- we have defined a measure over functions
- we can parametrise this measure to reflect our knowledge

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

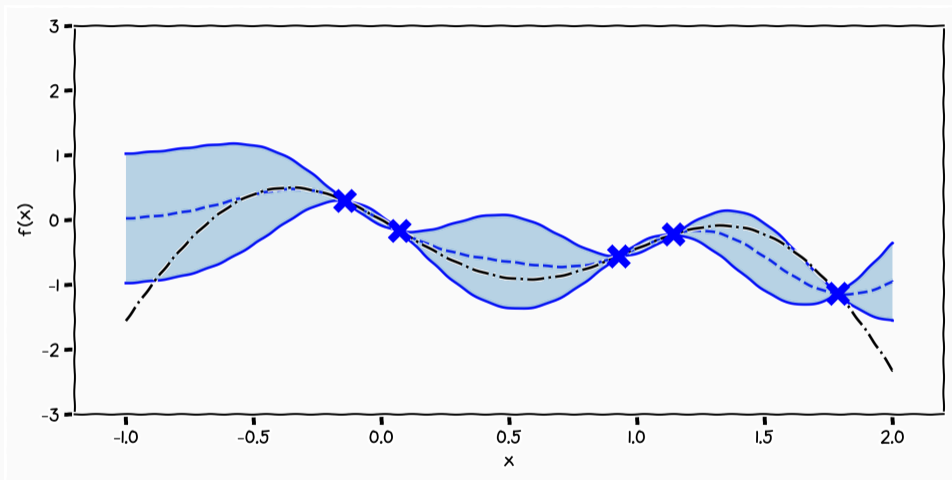
- we have defined a measure over functions
- we can parametrise this measure to reflect our knowledge
- we can get an updated measure that combines our knowledge with data



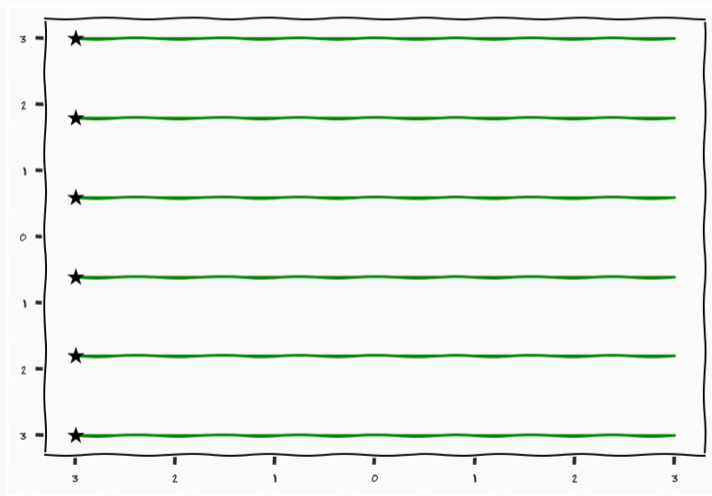
$$p(y|x) = \int p(y | f)p(f | x)df$$

$$p(y) = \int p(y | f)p(f | x)p(x)dfdx$$

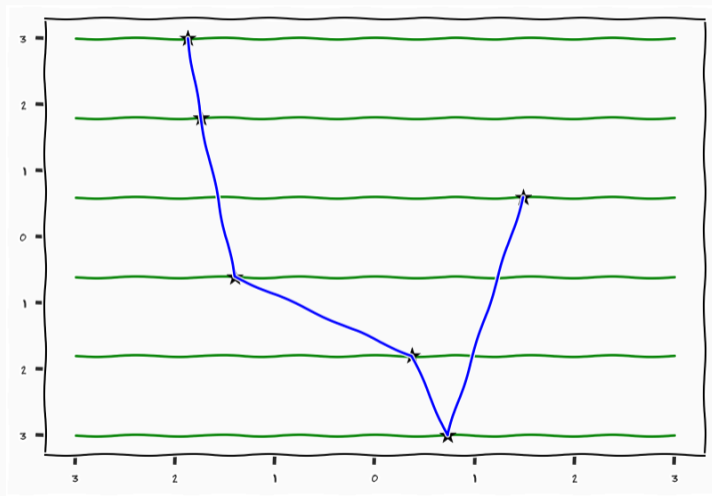
Posterior Processes



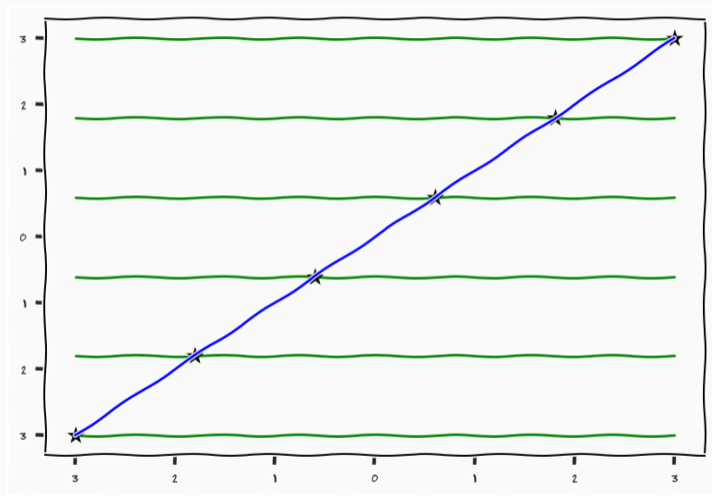
Unsupervised Learning



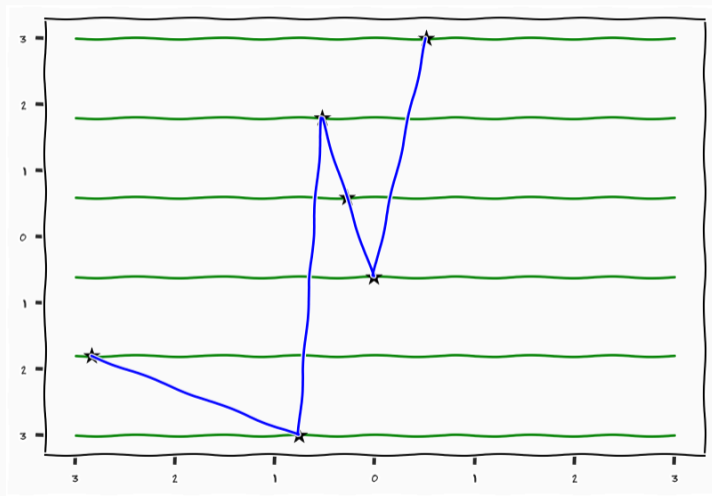
Unsupervised Learning



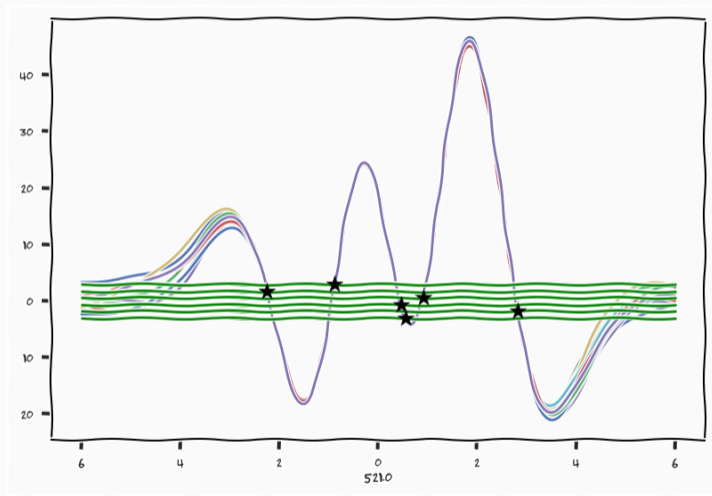
Unsupervised Learning



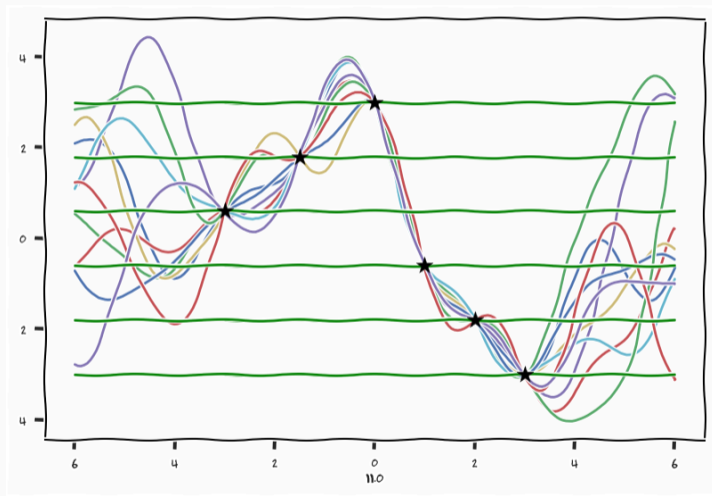
Unsupervised Learning



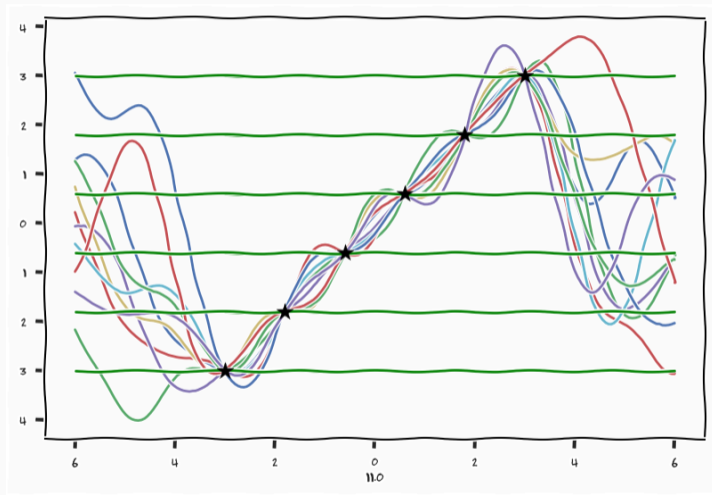
Unsupervised Learning



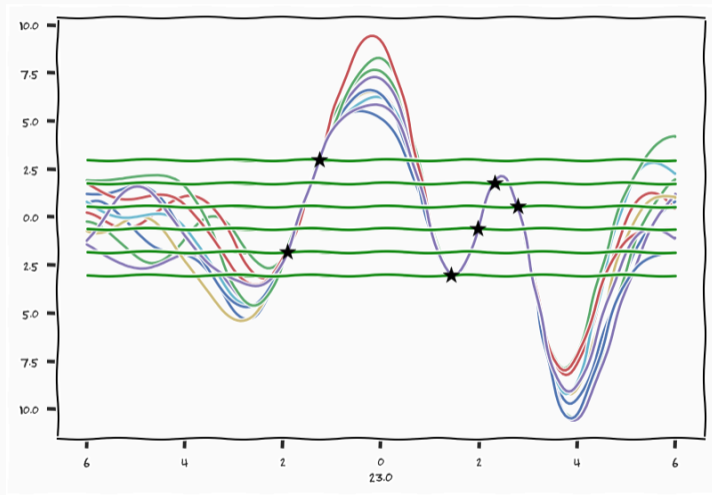
Unsupervised Learning



Unsupervised Learning



Unsupervised Learning



$$p(y) = \int p(y | f_2)p(f_2 | f_1)df_2df_1$$

- The process of Marginalisation allows me to convert one measure to another measure

Regression there are infinite number of possible functions that connects the data equally well. A GP provides a measure over these solutions that makes the problem "well-posed".

Regression there are infinite number of possible functions that connects the data equally well. A GP provides a measure over these solutions that makes the problem "well-posed".

Unsupervised Learning there are infinite number of possible combinations of input locations and functions that generate the data equally well. A GP and a latent space prior jointly provides a measure over these solutions to make the problem "well-posed"

Approximate Inference

$$\hat{\theta} = \operatorname{argmax}_{\theta} p(y | \theta) = \int p(y | f)p(f | x, \theta)p(x)dfdx$$

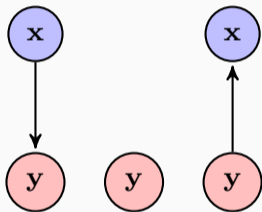
$$\hat{\theta} = \operatorname{argmax}_{\theta} p(y | \theta) = \int p(y | f)p(f | x, \theta)p(x)dfdx$$

- each evaluation of $p(y | \theta)$ is $\mathcal{O}(n^3)$

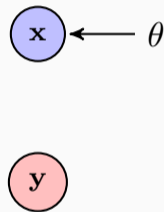
$$\hat{\theta} = \operatorname{argmax}_{\theta} p(y | \theta) = \int p(y | f)p(f | x, \theta)p(x)dfdx$$

- each evaluation of $p(y | \theta)$ is $\mathcal{O}(n^3)$
- integrating over $p(x)$ is generally analytically intractable

$$p(y) = \int p(y | x)p(x)dx$$



$$p(y) = \int_x p(y|x)p(x) = \frac{p(y|x)p(x)}{p(x|y)}$$

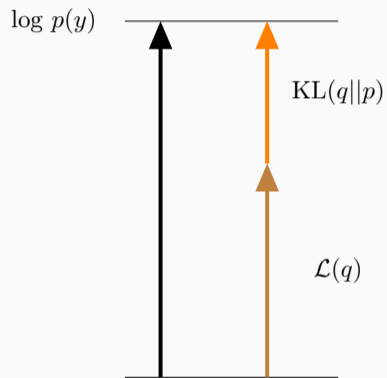


$$q_{\theta}(x) \approx p(x|y)$$

$$\begin{aligned}\log p(y) &= \int q(x) \log \frac{1}{q(x)} dx + \int q(x) \log p(x, y) dx + \int q(x) \log \frac{q(x)}{p(x|y)} dx \\ &\geq - \int q(x) \log q(x) dx + \int q(x) \log p(x, y) dx\end{aligned}$$

- The Evidence Lower BOnd
- Tight if $q(x) = p(x|y)$

Deterministic Approximation



- All Bayesian models are generally computational expensive and intractable

- All Bayesian models are generally computational expensive and intractable
- 90% of your work is on coming up with approximations

- All Bayesian models are generally computational expensive and intractable
- 90% of your work is on coming up with approximations
- As scientists, worry about formulating the best possible model to start with, then worry about inference.

- All Bayesian models are generally computational expensive and intractable
- 90% of your work is on coming up with approximations
- As scientists, worry about formulating the best possible model to start with, then worry about inference.
- Understand the Bayesian modelling principles, understand Gaussian processes, first

Summary

- There is no such thing as a free lunch, anything that learns something does so by being biased

- There is no such thing as a free lunch, anything that learns something does so by being biased
- Any explanation of a result can only ever be interpreted relative to the bias that has been included

- There is no such thing as a free lunch, anything that learns something does so by being biased
- Any explanation of a result can only ever be interpreted relative to the bias that has been included
- Arguing religiously about being Bayesian or not boils down to do if you agree with the process of marginalisation

- There is no such thing as a free lunch, anything that learns something does so by being biased
- Any explanation of a result can only ever be interpreted relative to the bias that has been included
- Arguing religiously about being Bayesian or not boils down to do if you agree with the process of marginalisation
 - I believe you can be pragmatically non-bayesian, but it is very hard to motivate philosophically

- infinite capacity by parametrising the model through relationship between data

- infinite capacity by parametrising the model through relationship between data
- model of non-parametric parametrisation leads to stochastic processes

- infinite capacity by parametrising the model through relationship between data
- model of non-parametric parametrisation leads to stochastic processes
- Gaussian processes

- infinite capacity by parametrising the model through relationship between data
- model of non-parametric parametrisation leads to stochastic processes
- Gaussian processes
practical use simple manipulation with multi-variate normals

- infinite capacity by parametrising the model through relationship between data
- model of non-parametric parametrisation leads to stochastic processes
- Gaussian processes
 - practical use** simple manipulation with multi-variate normals
 - theoretically** beautiful semantic in terms of stochastic processes

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

1. Exchangeable

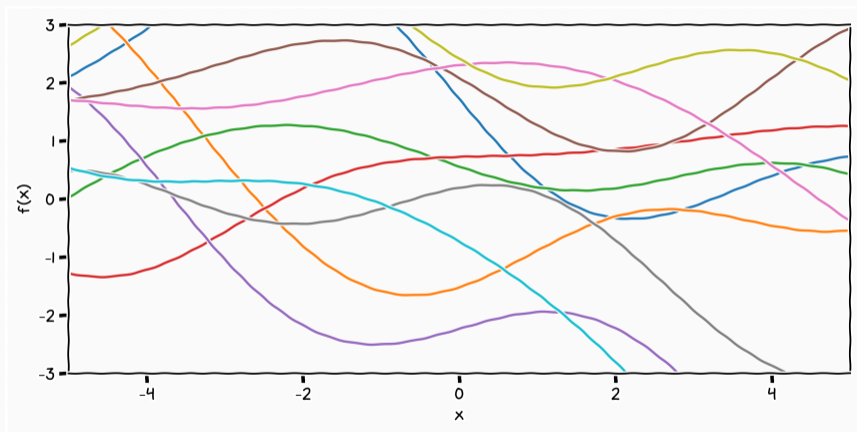
$$\nu_{t_{\pi(1)} \dots t_{\pi(k)}} (F_{\pi(1)} \times \dots \times F_{\pi(k)}) = \nu_{t_1 \dots t_k} (F_1 \times \dots \times F_k)$$

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

Are Gaussian Processes good parametrisations?



Yes being non-parametric it is only our lack of knowledge of appropriate measures of correlation that forces us to compromise

Are Gaussian Processes good parametrisations?

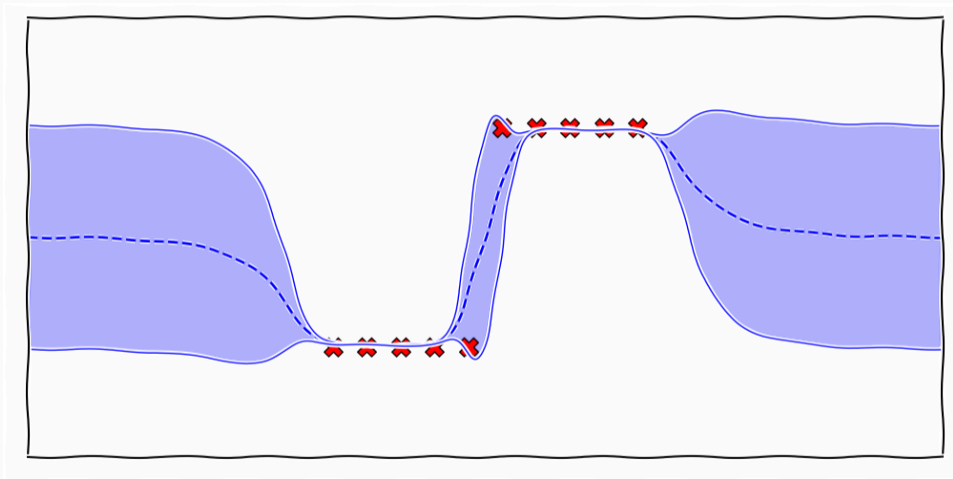
- Yes** being non-parametric it is only our lack of knowledge of appropriate measures of correlation that forces us to compromise
- Yes** their parametrisation is very well aligned to the knowledge we have of many problems, most complex knowledge (like beer) is relative

Are Gaussian Processes good parametrisations?

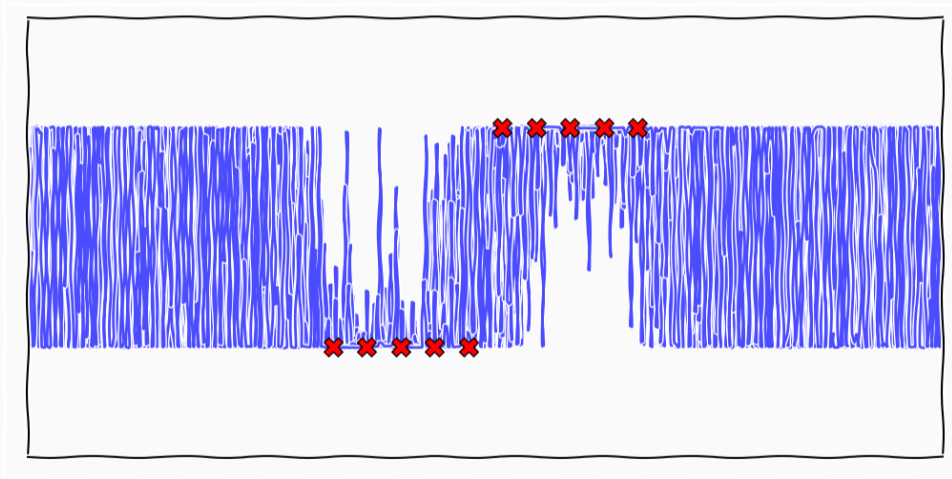
- Yes** being non-parametric it is only our lack of knowledge of appropriate measures of correlation that forces us to compromise
- Yes** their parametrisation is very well aligned to the knowledge we have of many problems, most complex knowledge (like beer) is relative
- Yes** they are incredibly "narrow" but have infinite coverage

pink-elephant.png

$$f(x) = f_L \circ f_{L-1} \circ \cdots \circ f_0(x)$$



Composite GP Step



Composite GPs potentially interesting but inference is a huge issue and

Composite GPs potentially interesting but inference is a huge issue and

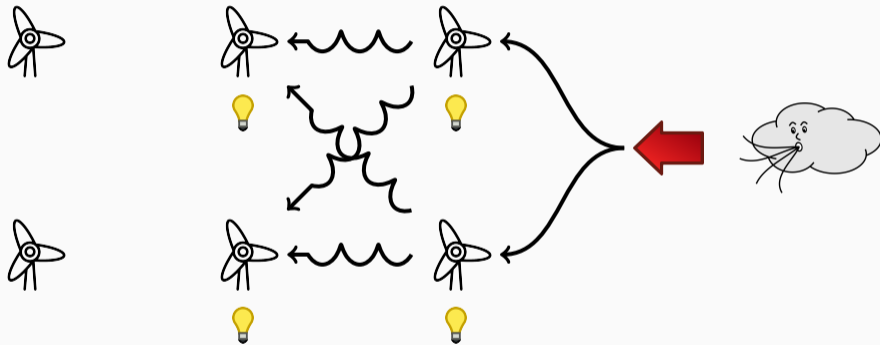
BNN worst of both worlds, a prior we do not understand, in a structure we do not get, means that we are effectively spending a huge computational overload to implement a regulariser

Composite GPs potentially interesting but inference is a huge issue and

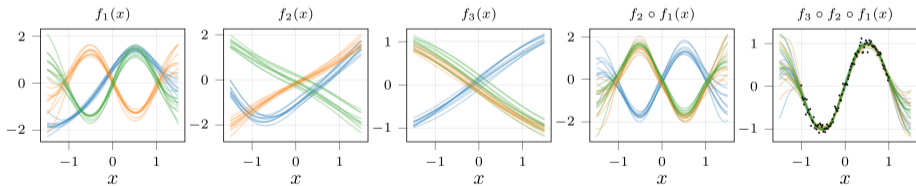
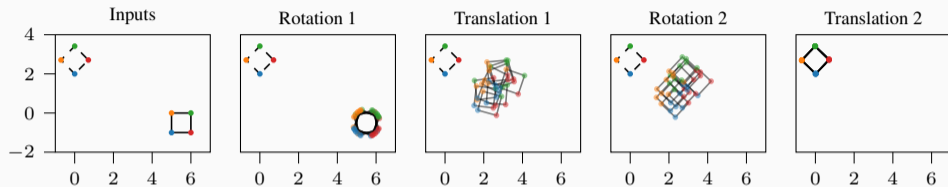
BNN worst of both worlds, a prior we do not understand, in a structure we do not get, means that we are effectively spending a huge computational overload to implement a regulariser

When should we use composite models when our knowledge is composite

$$p(\mathbf{s}_{t+1}) = \int p(\mathbf{s}_{t+1} \mid \mathbf{f}, \mathbf{s}_t, \mathbf{a}_t) p(\mathbf{a}_t \mid \pi, \mathbf{s}_t) p(\mathbf{s}_t) p(\mathbf{f}) p(\pi) d\mathbf{a}_t d\mathbf{s}_t d\mathbf{f} d\pi,$$





Composite Model²



²Ustyuzhaninov et al., 2020

eof

References

-  Bishop, Christopher M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc.
-  Kaiser, M., C. Otte, T. Runkler, and C. H. Ek (2018). “Bayesian Alignments of Warped Multi-Output Gaussian Processes.” In: *Advances in Neural Information Processing Systems 32, [NIPS Conference, Montreal, Quebec, Canada, December 3 - December 8, 2018]*.

-  Roy, Hrittik et al. (2024). “Reparameterization Invariance in Approximate Bayesian Inference.” In: *CoRR*.
-  Shalev-Shwartz, Shai and Shai Ben-David (2014). *Understanding Machine Learning: From Theory to Algorithms*. New York, NY, USA: Cambridge University Press.
-  Ustyuzhaninov, Ivan et al. (2020). “Compositional uncertainty in deep Gaussian processes.” In: *Proceedings of the Thirty-Sixth Conference on Uncertainty in Artificial Intelligence, UAI 2020, virtual online, August 3-6, 2020*. Ed. by Ryan P. Adams and Vibhav Gogate. Vol. 124. Proceedings of Machine Learning Research. AUAI Press, pp. 480–489.