Gaussian Processes

a first introduction

Carl Henrik Ek

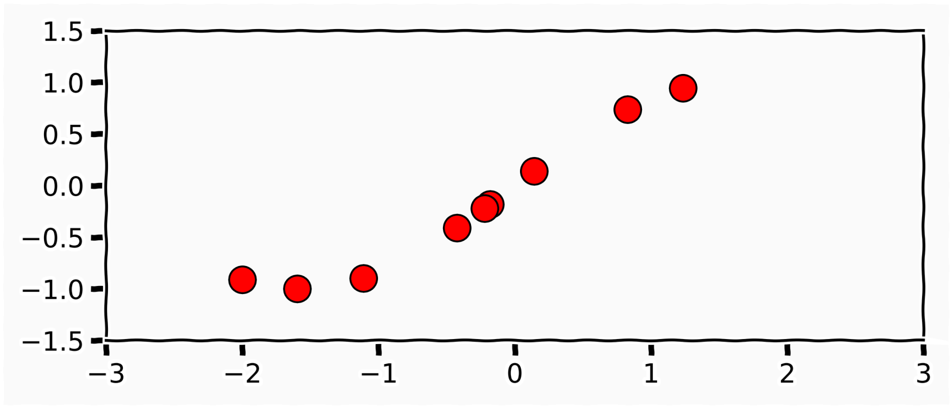
September 8, 2025

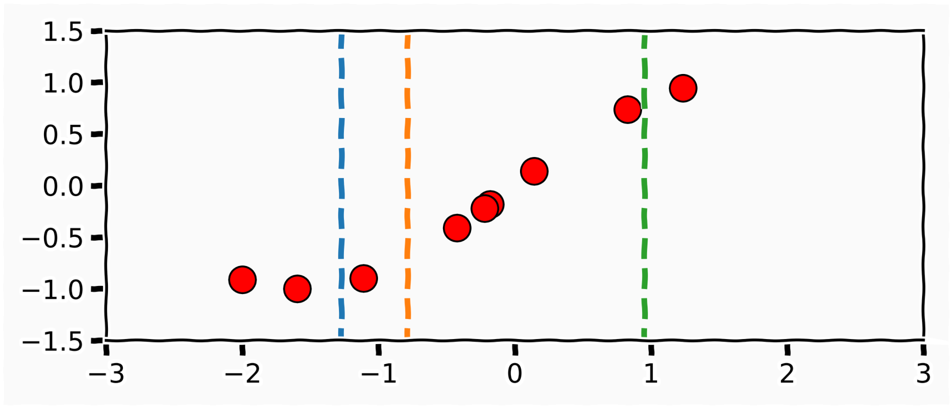
http://carlhenrik.com

Inductive Reasoning

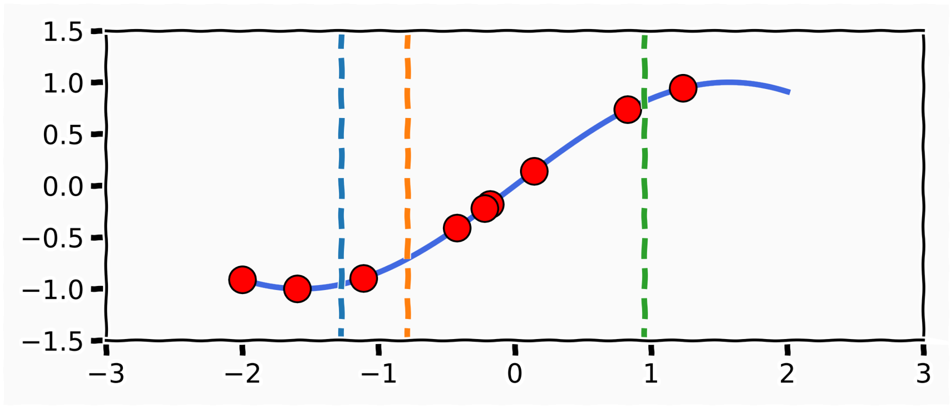
”In inductive inference, we go from the specific to the general. We make observations, discern a pattern, make a generalization, and infer an explanation or a theory”

– Wassertheil-Smoller

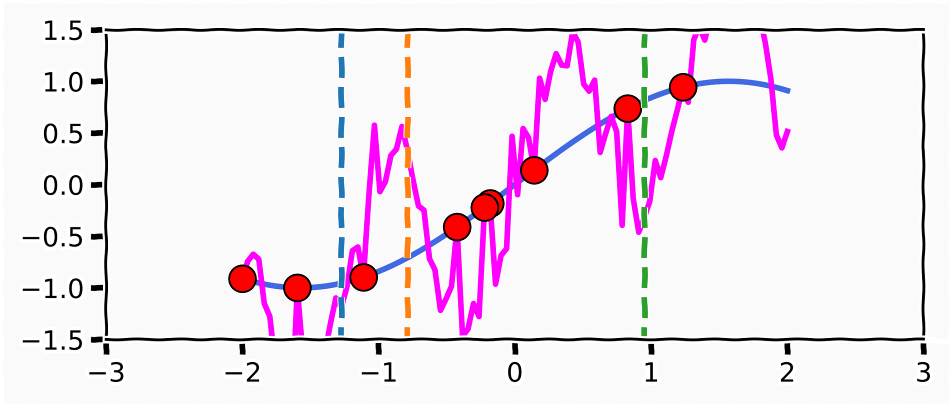

1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−3
−2
−1
0
1
2
3



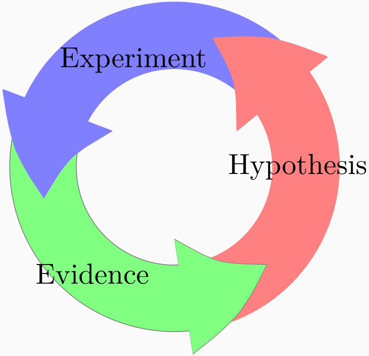
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−3
−2
−1
0
1
2
3


1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−3
−2
−1
0
1
2
3

Inductive Reasoning
Unlike deductive arguments, inductive reasoning allows for the possibility that the conclusion is false, even if all of the premises are true.


1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−3
−2
−1
0
1
2
3

The Scientific Principle

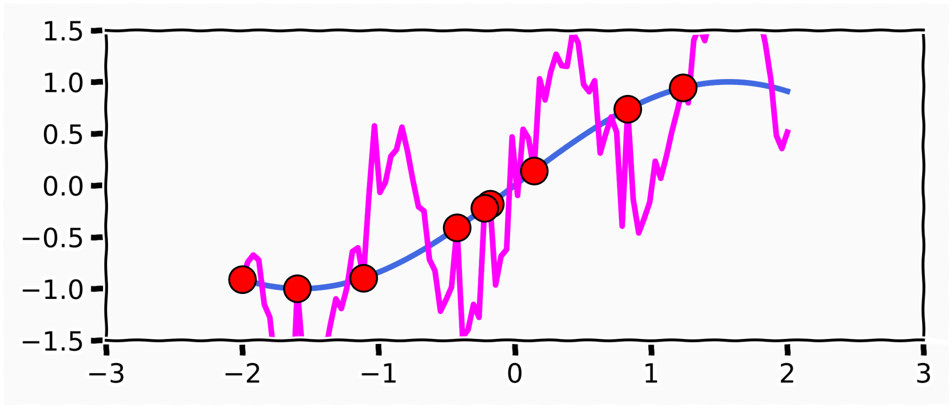


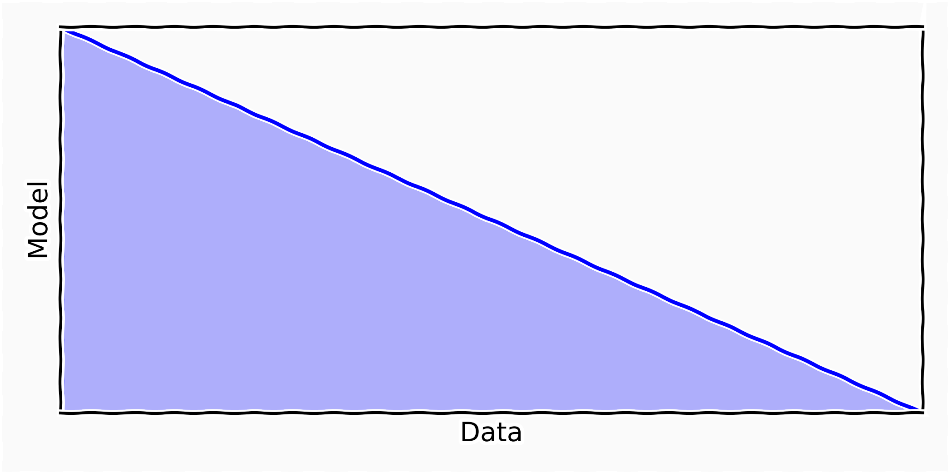
”The Machine Learning Principle”[1]

”There is a notion of success ... which I think is novel in the history of science. It interprets success as approximating unanalyzed data.”

– Prof. Noam Chomsky

[1] Chomsky et al., 1980


1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−3
−2
−1
0
1
2
3


Model
Data

What is machine learning?

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?

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 but important change of narrative)

Formalism [Shalev-Shwartz et al., 2014]

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

Formalism [Shalev-Shwartz et al., 2014]

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

Label Set $\mathcal{Y}$ the set of outputs

Formalism [Shalev-Shwartz et al., 2014]

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 $\mathcal{X} \times \mathcal{Y}$

Formalism [Shalev-Shwartz et al., 2014]

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

Formalism [Shalev-Shwartz et al., 2014]

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

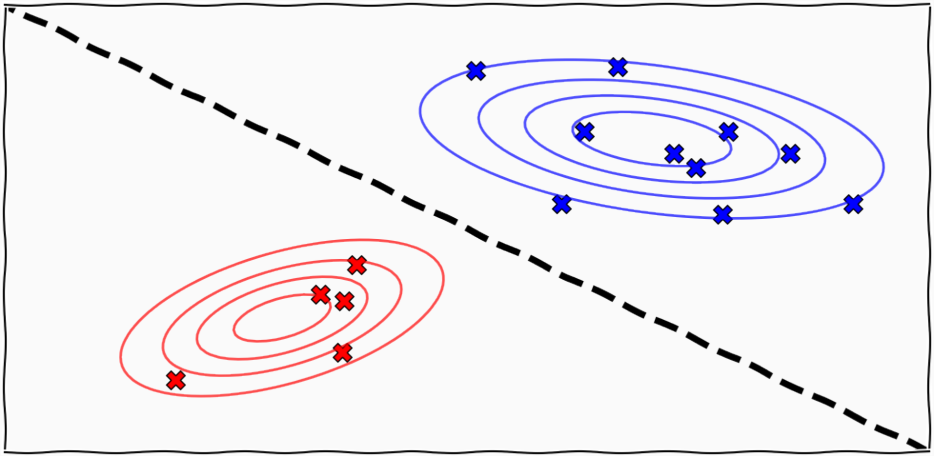
Data Generation $f : \mathcal{X} \to \mathcal{Y}$ the underlying generating process that we wish to recover

Formalism [Shalev-Shwartz et al., 2014]

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



Measure of Success

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

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

Measure of Success

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

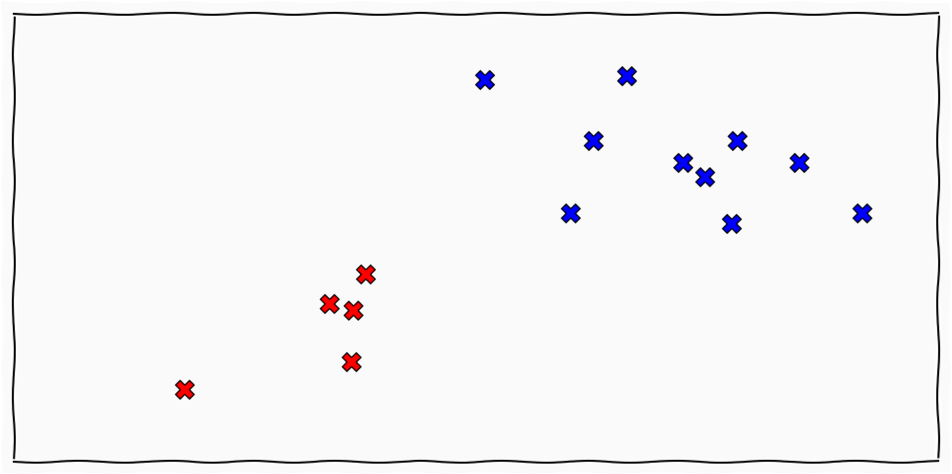
- measure of success as probability of misclassified points (true risk)
- we do not have access to $\mathcal{D}$

Measure of Success

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

- 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

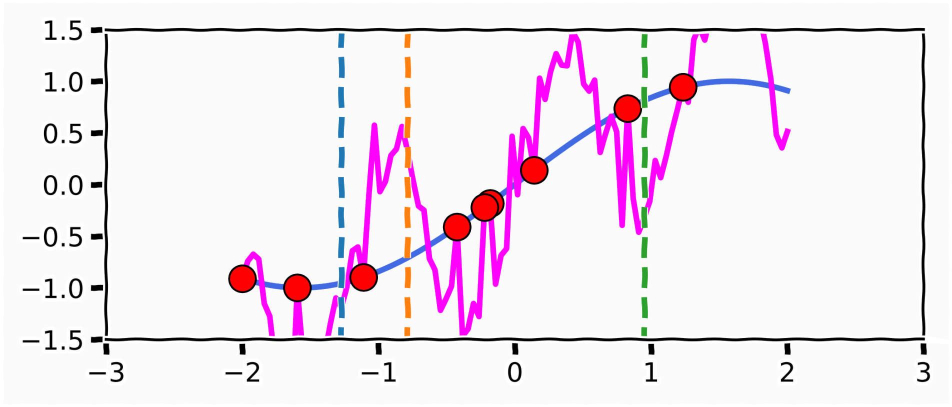
Classification



Empirical Risk Minimisation

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

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


1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−3
−2
−1
0
1
2
3

Algorithm

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

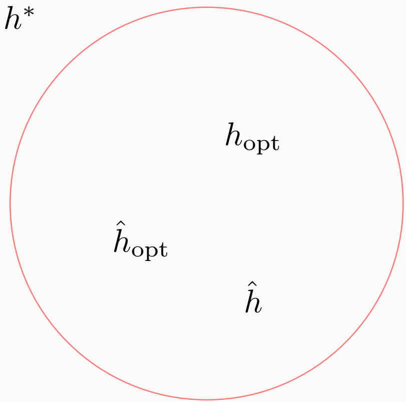
- We use an algorithm $A : \mathcal{S} \to h$ to find a hypothesis

Finite Hypothesis Classes

$$h_{\mathcal{S}} \in \operatorname*{argmin}_{h \in \mathcal{H}} L_{\mathcal{S}}(h)$$

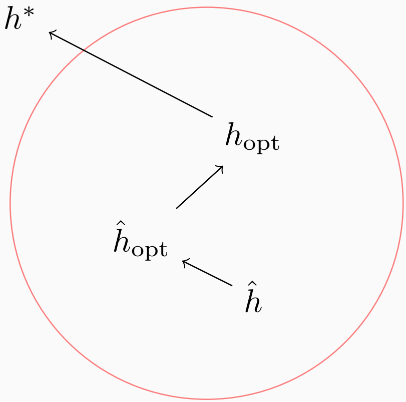
- We cannot parametrise all possible hypothesis

Error Decomposition



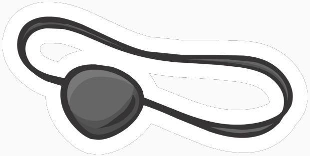
- h^* the optimal predictor
- h_{opt} the optimal hypothesis
- $\hat{h}_{\text{opt}}$ the optimal hypothesis on training data
- $\hat{h}$ the hypothesis found by learning algorithm

Error Decomposition



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

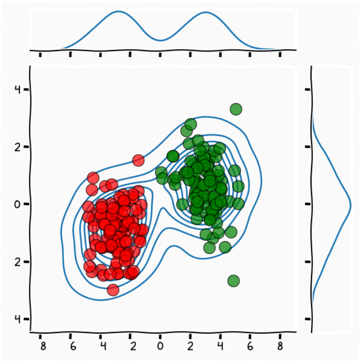
Assumptions: Algorithms



Statistical Learning

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

Assumptions: Biased Sample



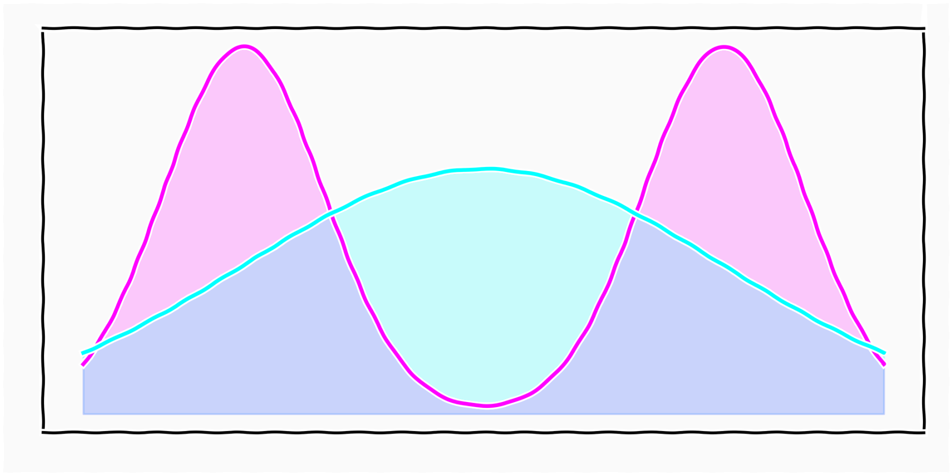
Statistical Learning

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

Assumptions: Hypothesis space

Statistical Learning

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



Bayes' Rule

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

Marginalisation

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

Marginalisation

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

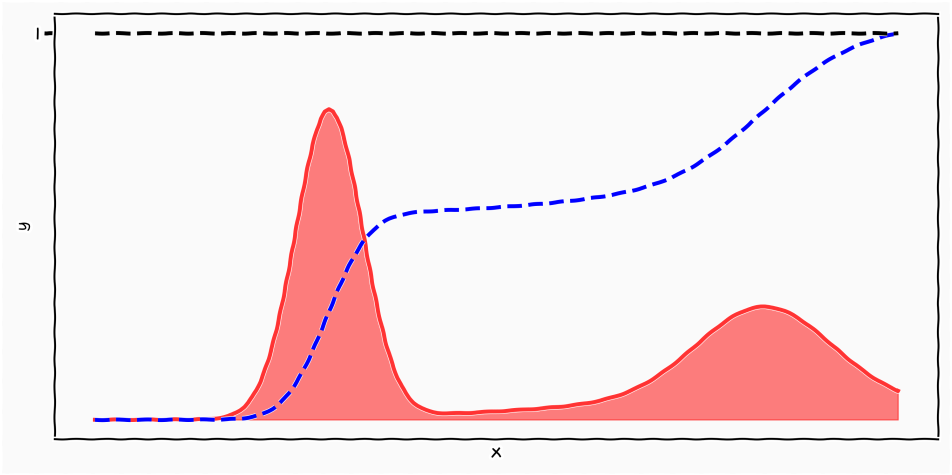
Marginalisation

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

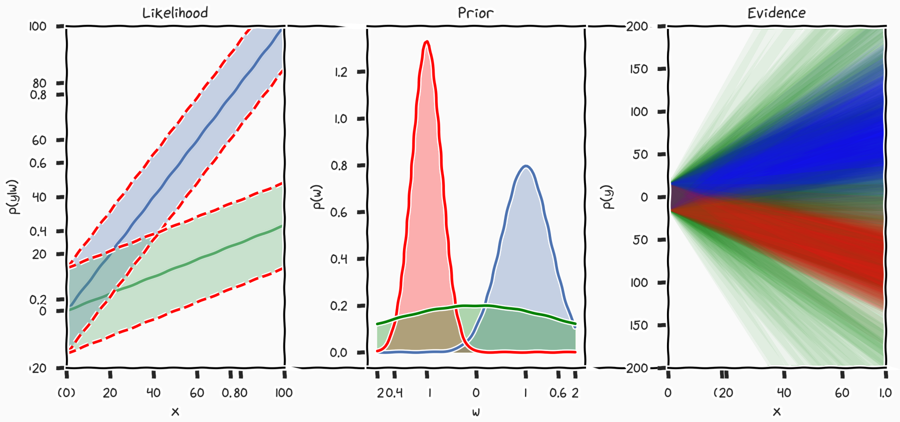
Marginalisation

$$p(\mathcal{D}) = \int p(\mathcal{D} \mid \theta) \underbrace{p(\theta)\mathrm{d}\theta}_{\mathrm{d}t(\theta)}$$

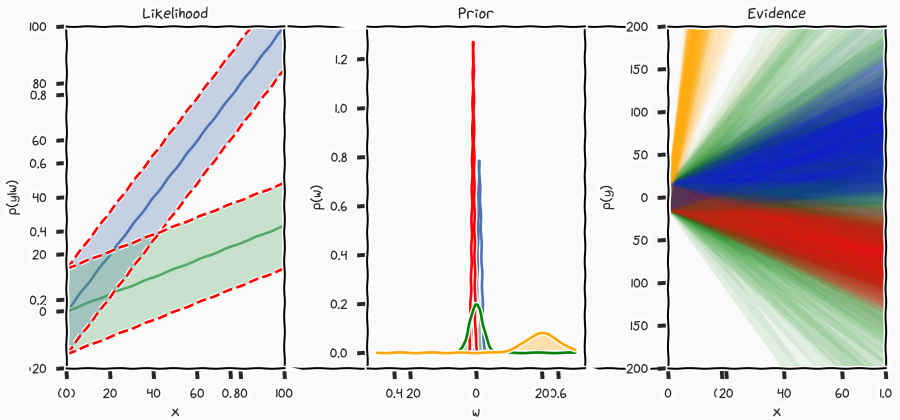
Marginalisation



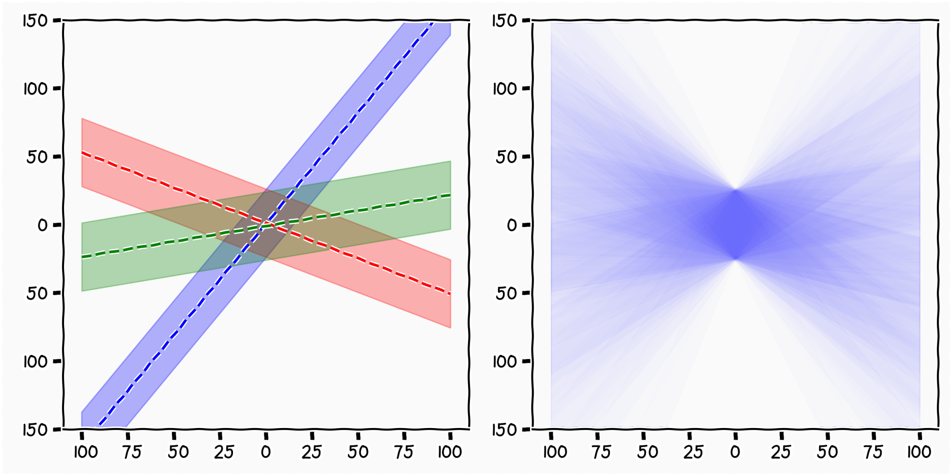
Marginalisation



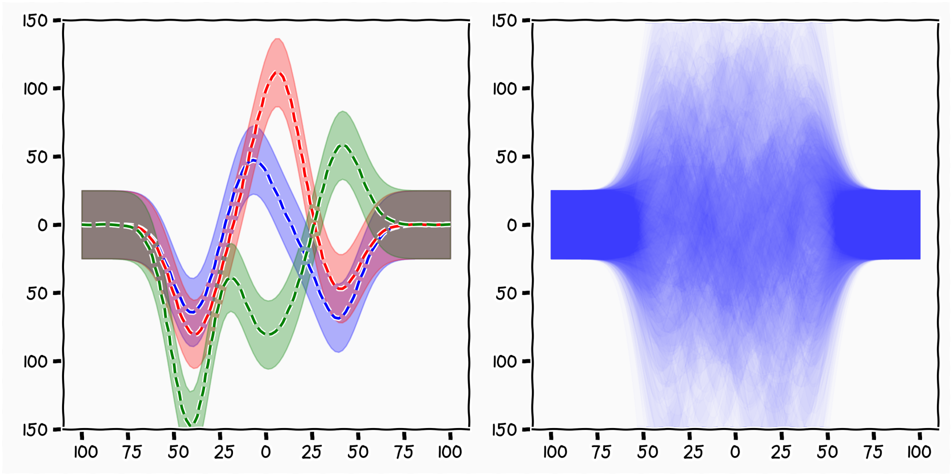
Marginalisation



Model Linear Linear



Model Linear



Bayesian

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

Bayesian

- 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

Bayesian

- 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

”Good” parametrisation

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

”Good” parametrisation

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

Narrow such that we can reduce data-requirements

”Good” parametrisation

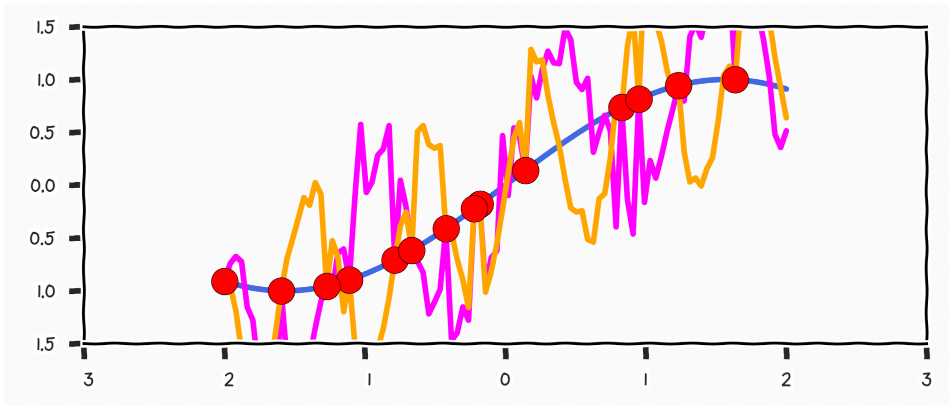
Flexible such that we do not have to make trade-offs when including beliefs

Narrow such that we can reduce data-requirements

Interpretable so that we can translate our knowledge to the parametrisation

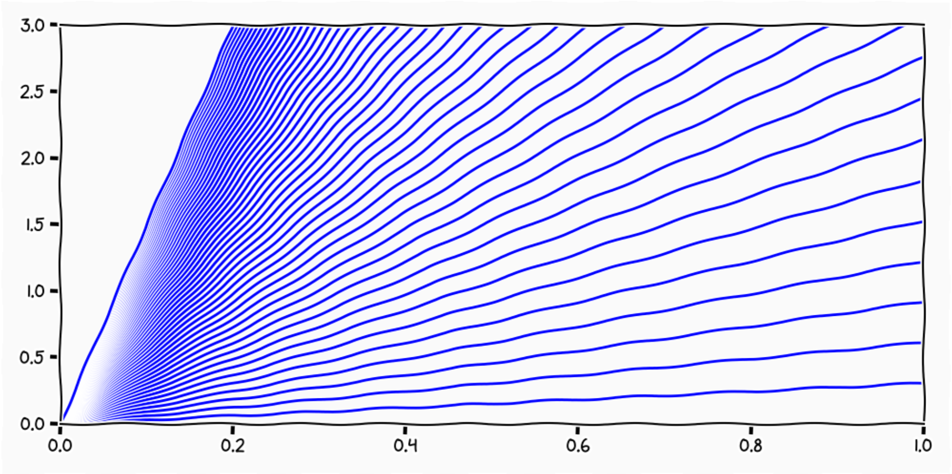
Non-parametrics

Curve Fitting

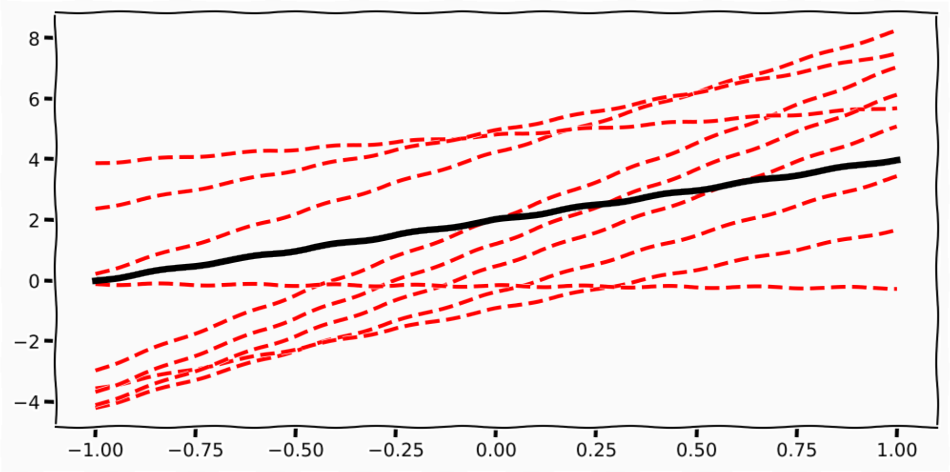


Parametrisations

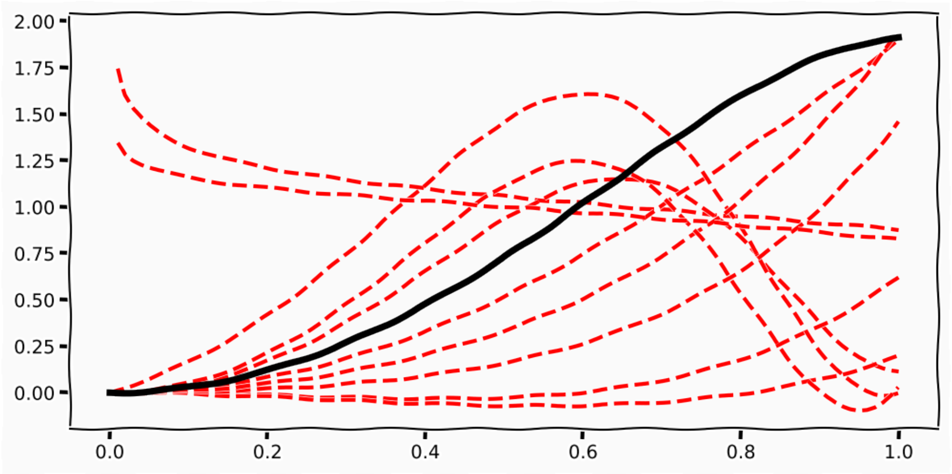
$$f(x) = \beta_1 + \beta_2 \cdot x + \beta_3 \cdot x^2 + \ldots + \beta_k \sin(x) + \ldots$$



$$f(x) = \beta \cdot x$$



$$f(x) = \beta_1 + \beta_2 \cdot x$$



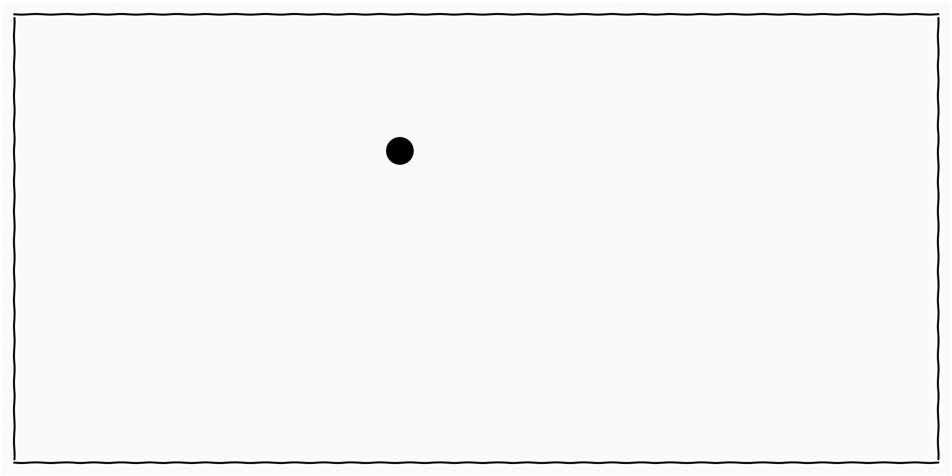
$$x^{\beta_0} + \sin(\beta_1 \cdot x^2)$$

Non-parametrics

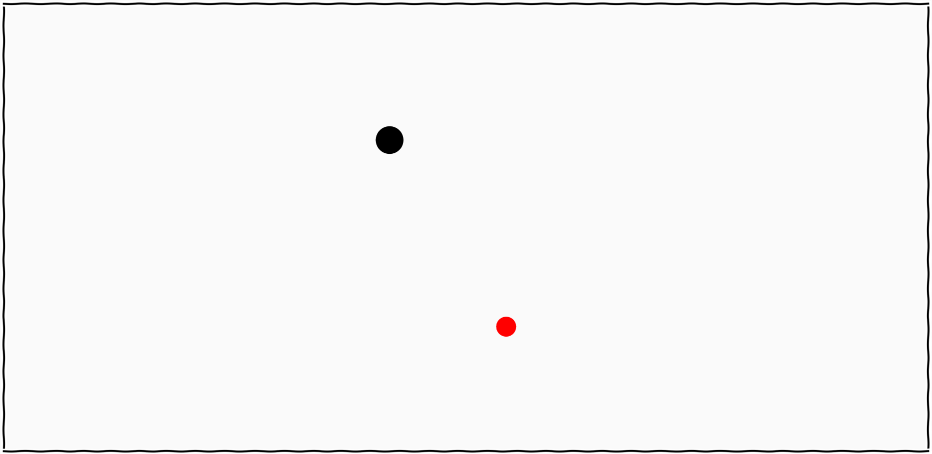
Non-parametrics

Non-parametrics

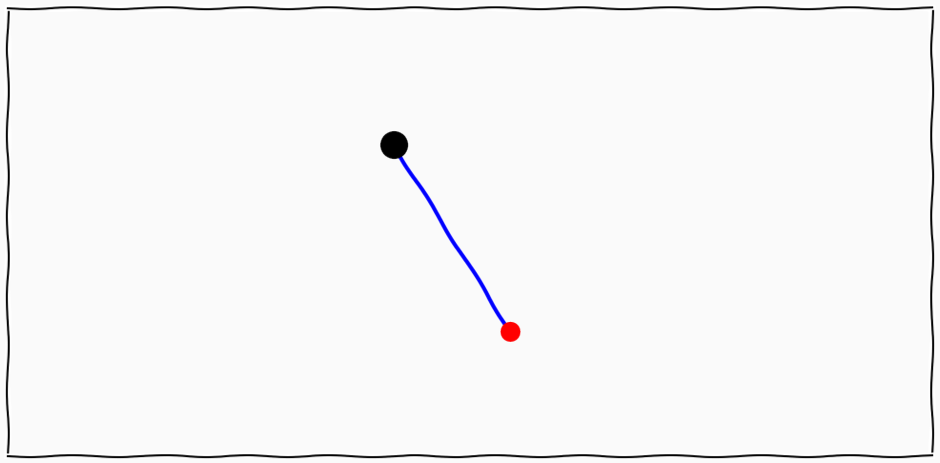
Example



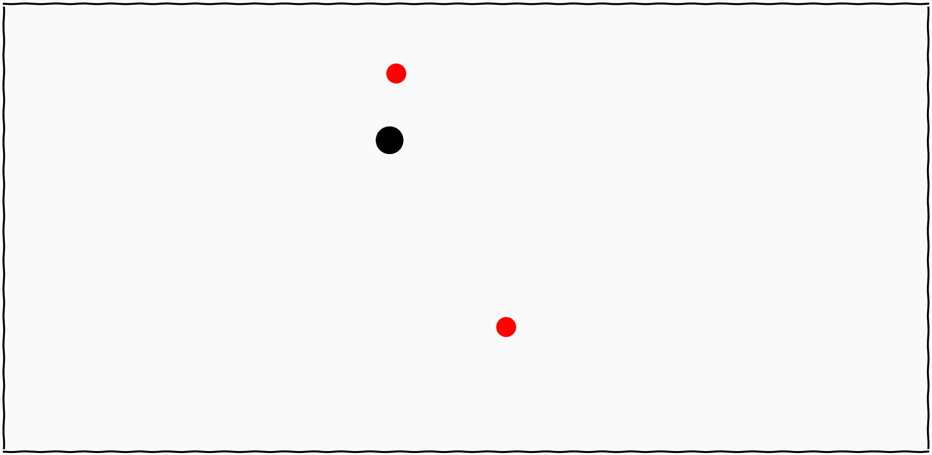
Example



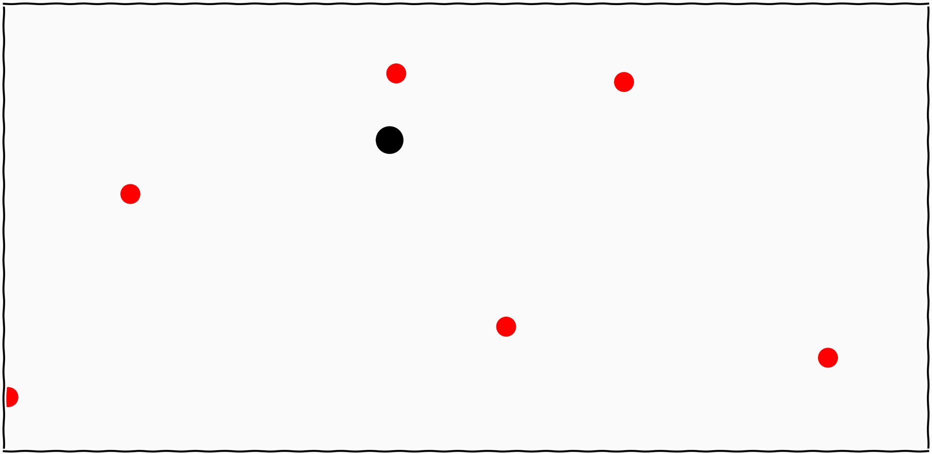
Example



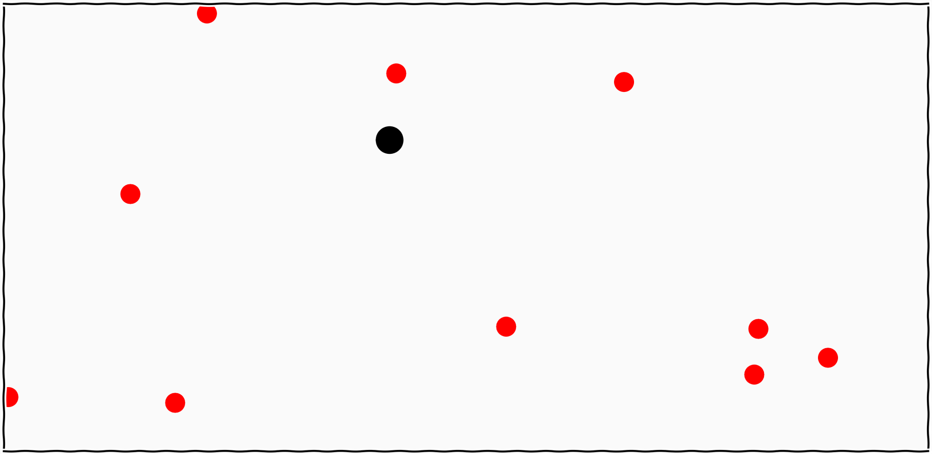
Example



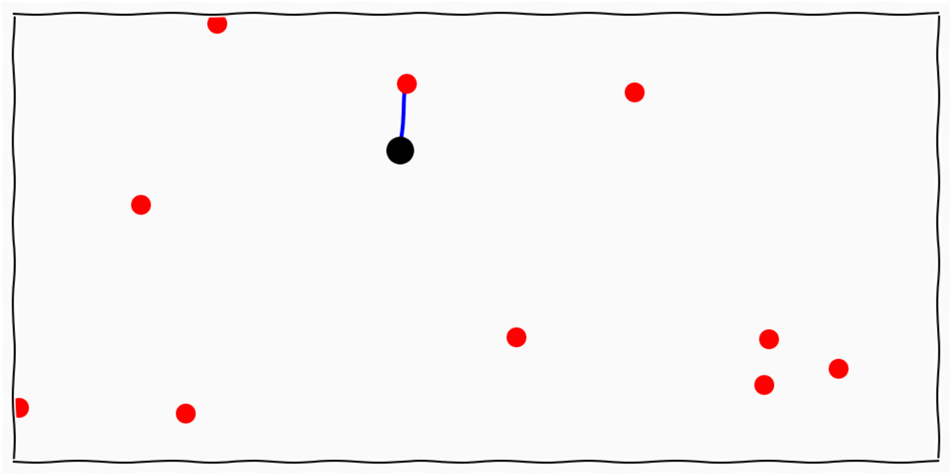
Example



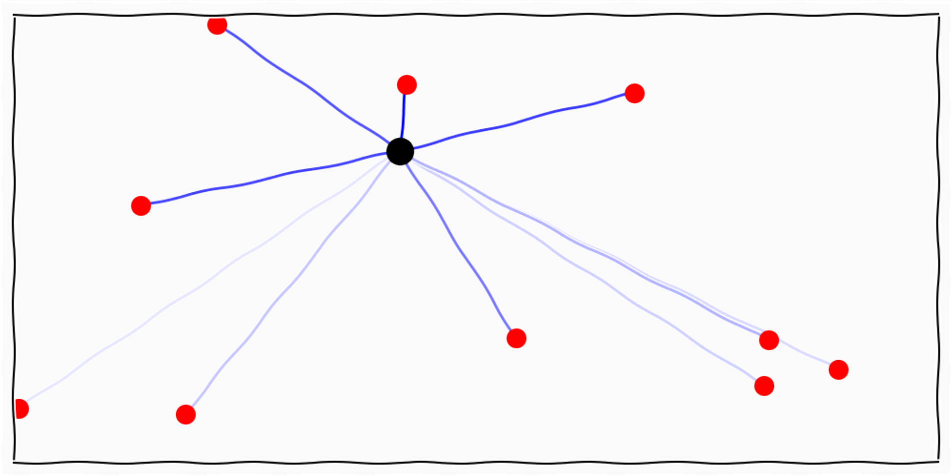
Example



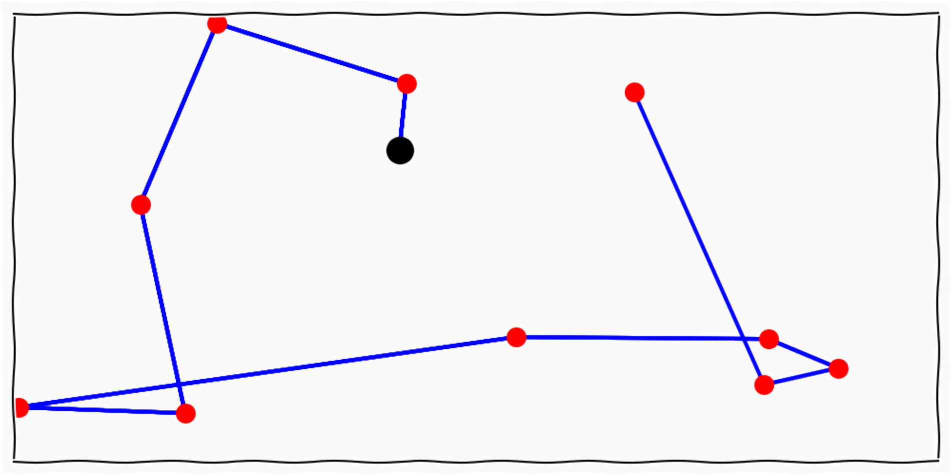
Example



Example

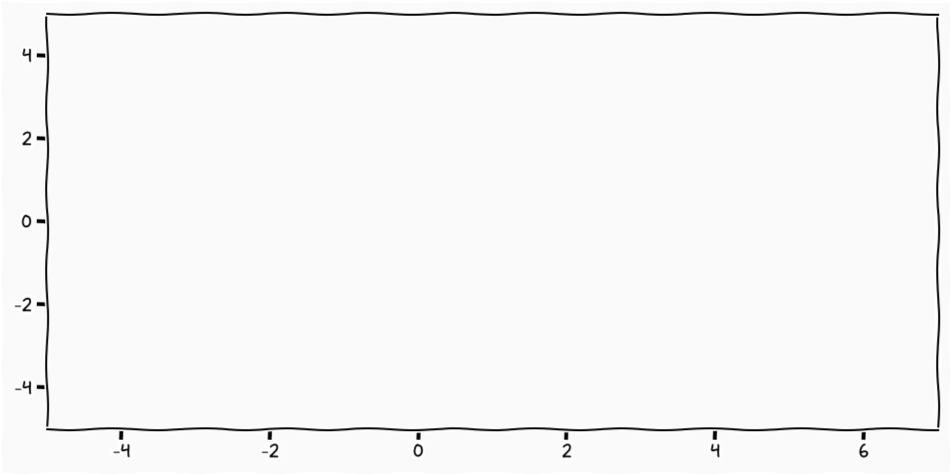


Example

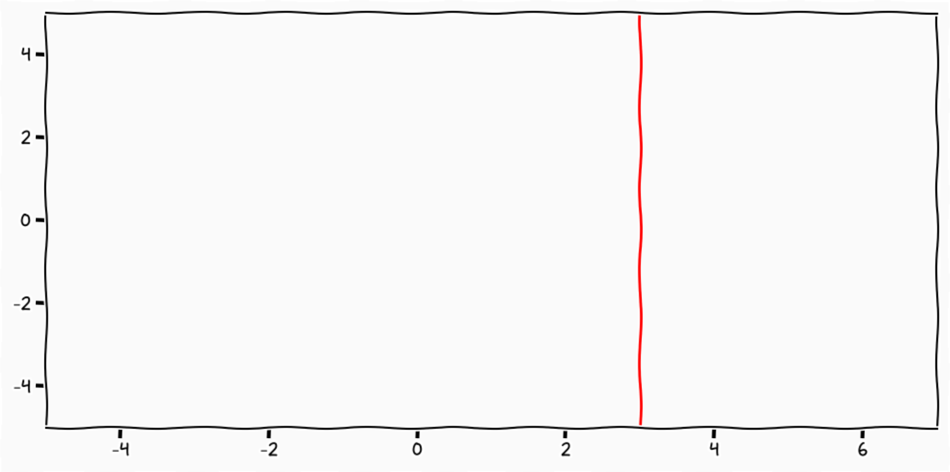


Non-parametric Models

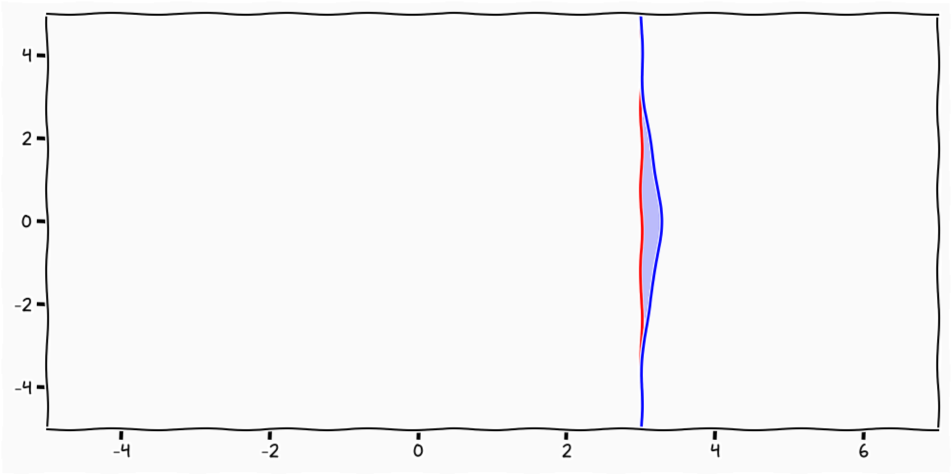
Lets talk about functions



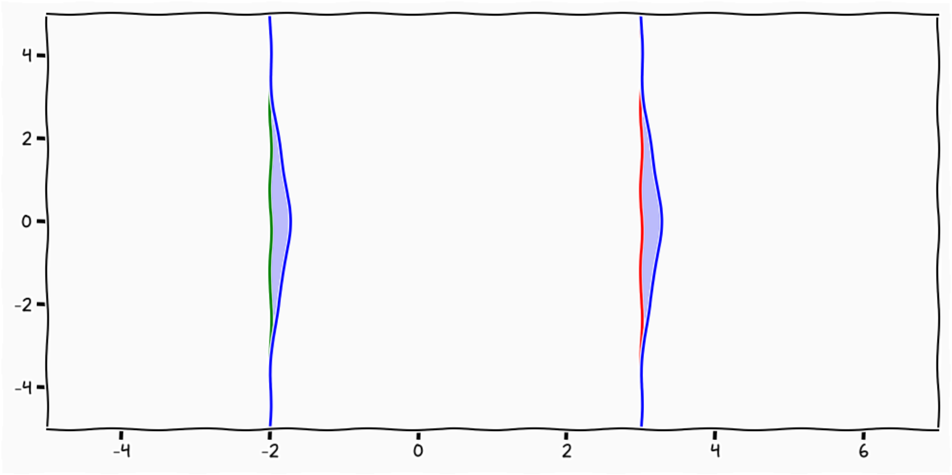
Lets talk about functions



Lets talk about functions



Lets talk about functions

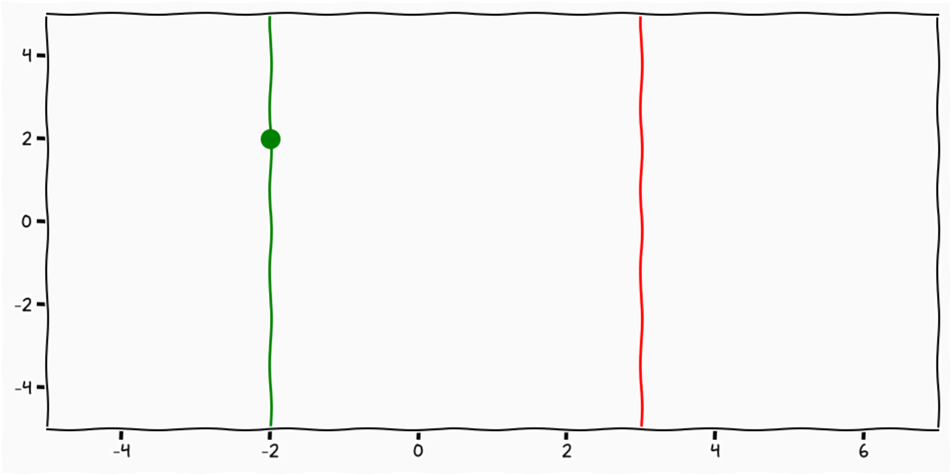


Gaussian function values

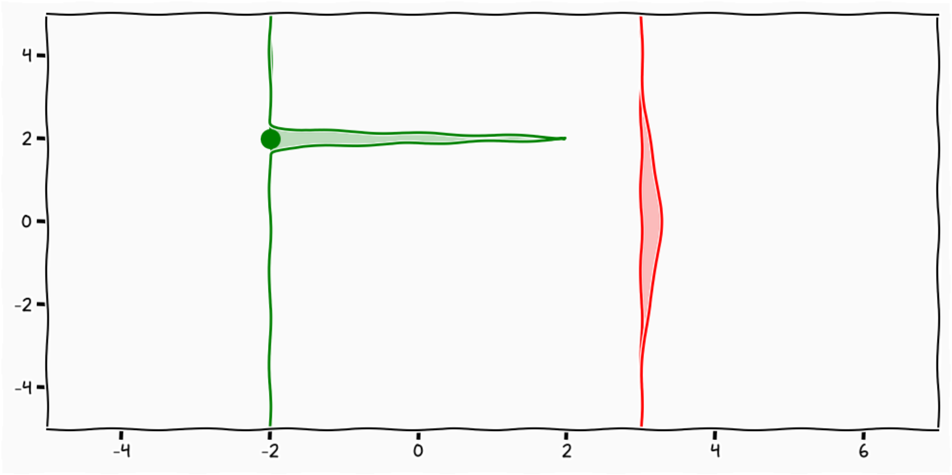
$$f_1 = \mathcal{N}(\mu_1, k_1)$$

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

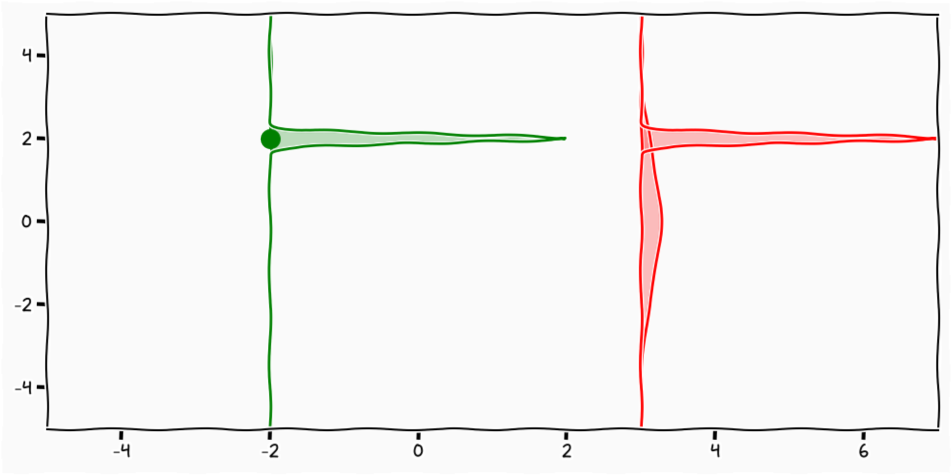
Non-parametric functions



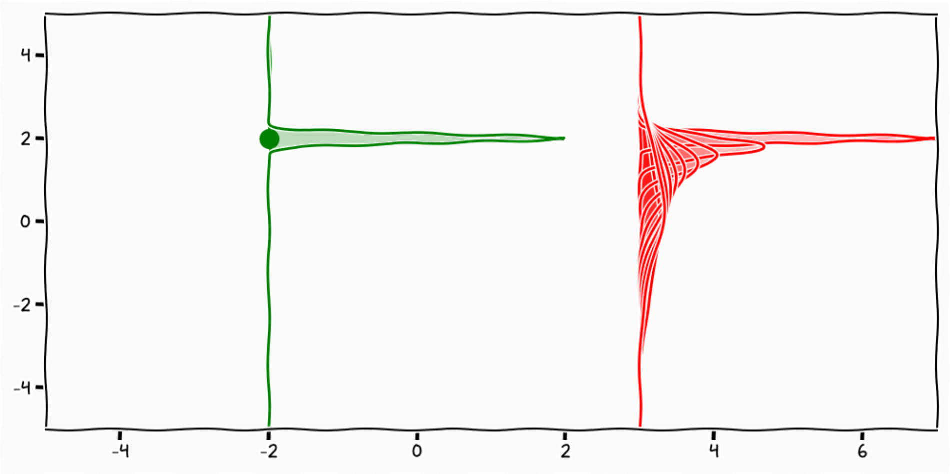
Non-parametric functions



Non-parametric functions



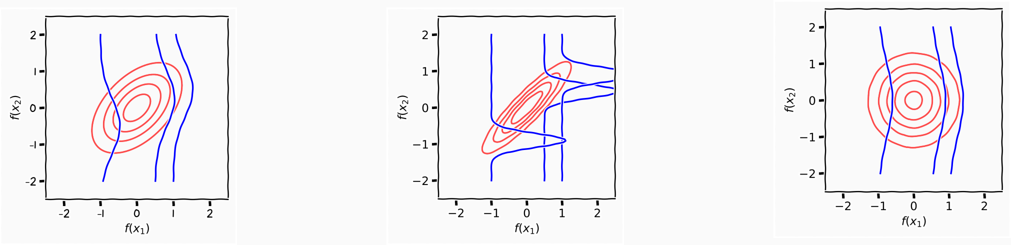
Non-parametric functions



Jointly Gaussian function values

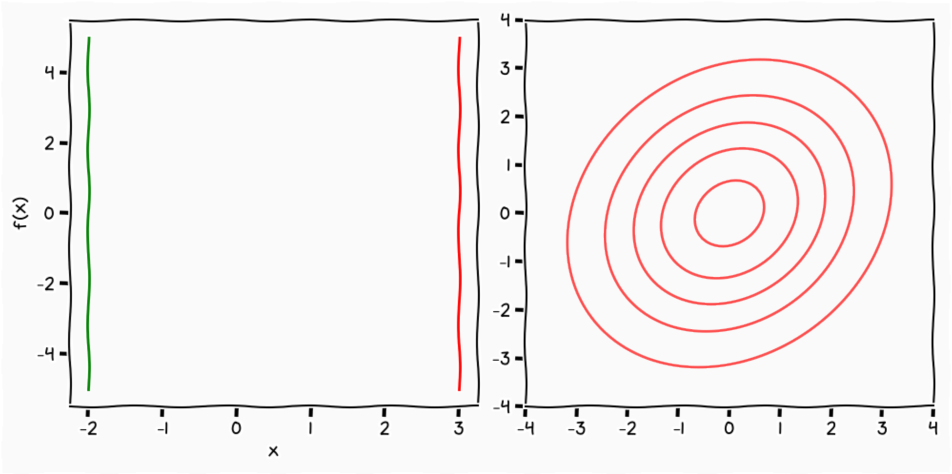
$$\begin{bmatrix} f_1 \\ f_2 \end{bmatrix} = \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} k_{11} & ? \\ ? & k_{22} \end{bmatrix} \right)$$

Conditional Gaussians

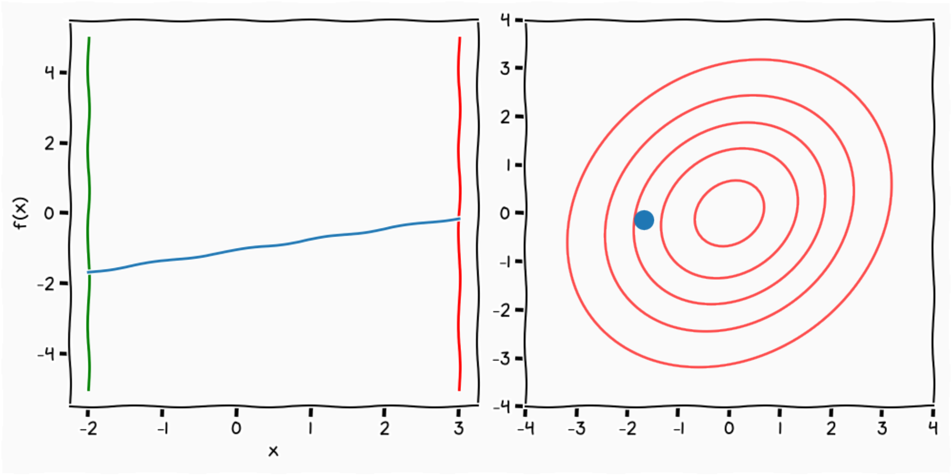


$$N\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}1 & 0.5\\0.5 & 1\end{bmatrix}\right)\qquad N\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}1 & 0.9\\0.9 & 1\end{bmatrix}\right)\qquad N\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right)$$

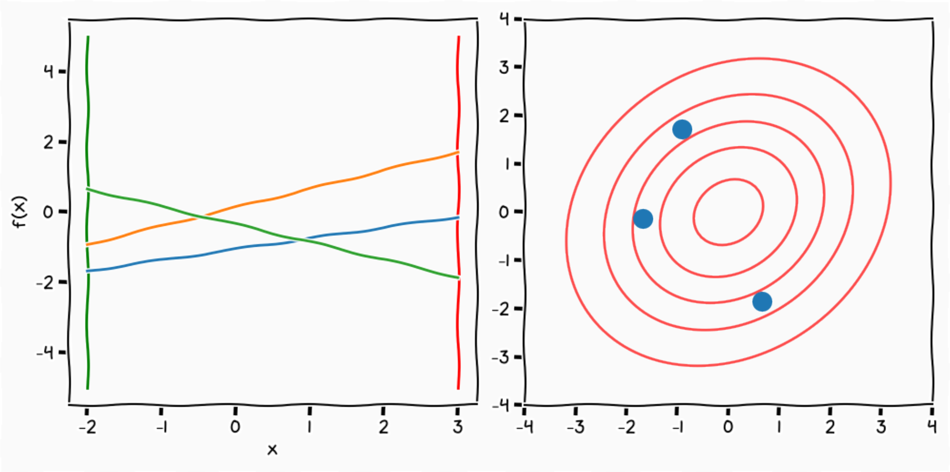
Gaussian Samples



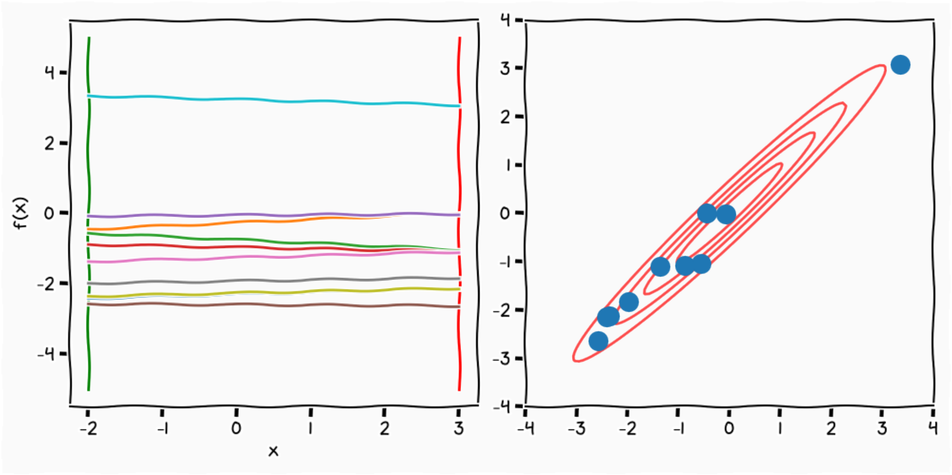
Gaussian Samples



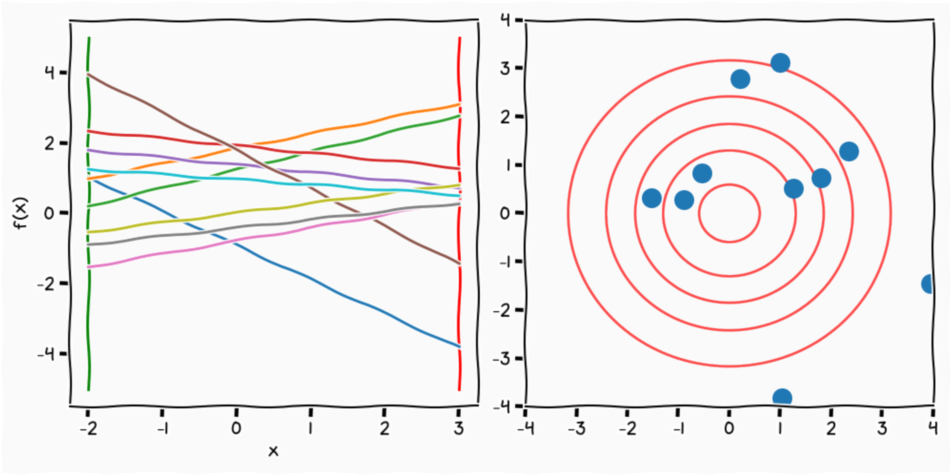
Gaussian Samples



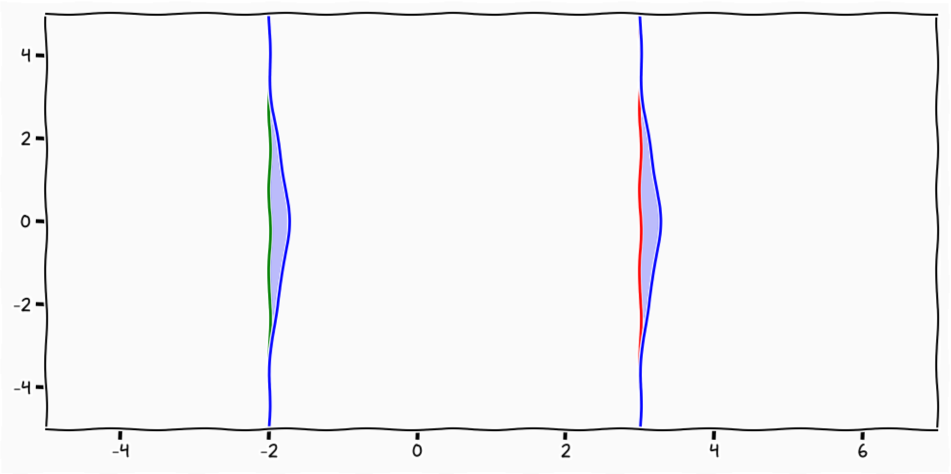
Gaussian Samples

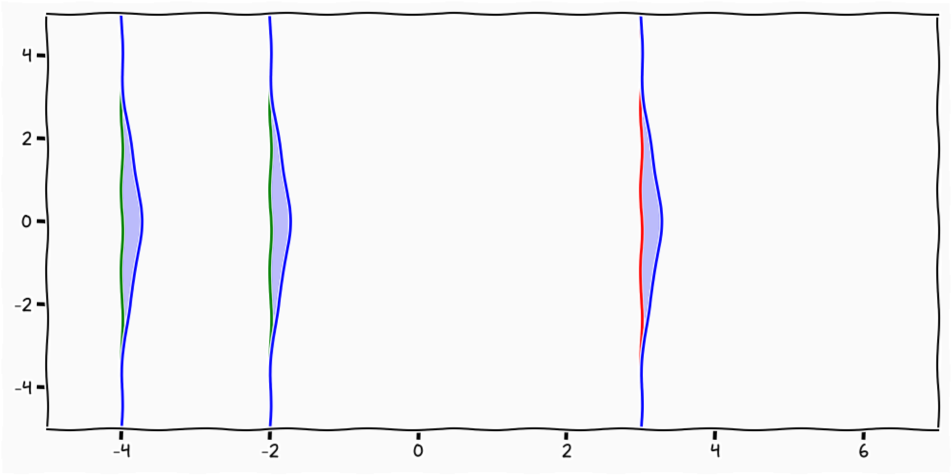


Gaussian Samples



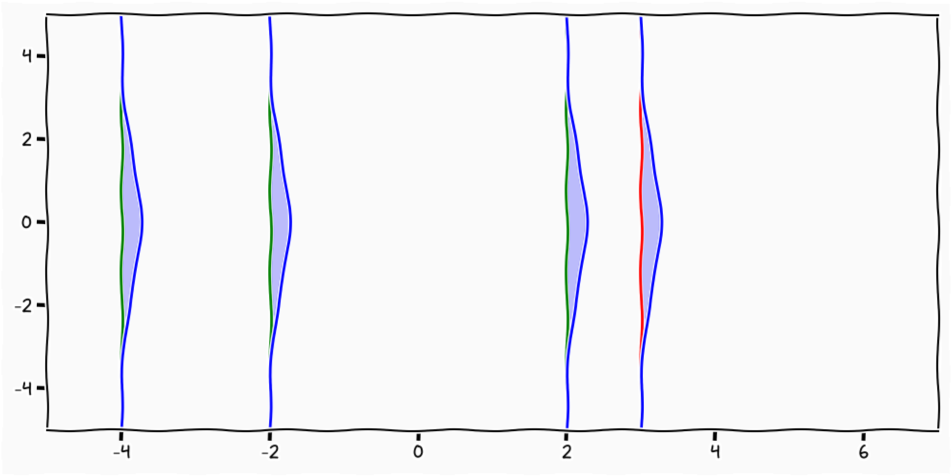
Lets talk about functions



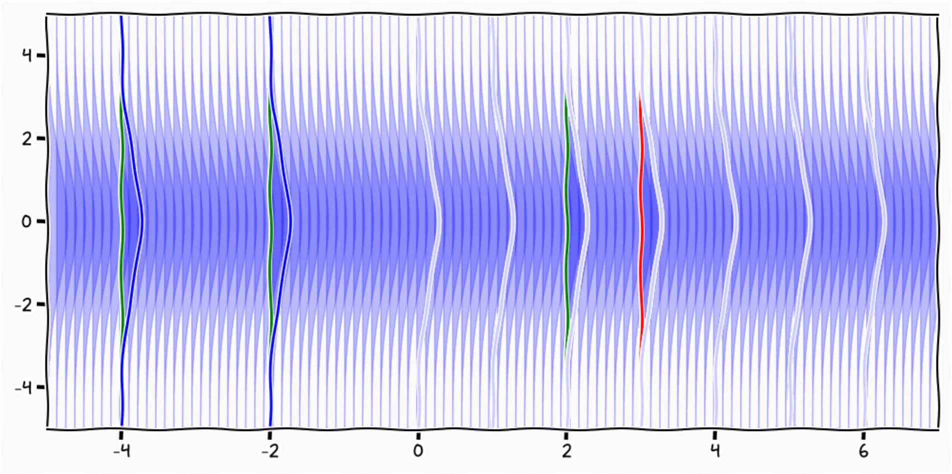


4
2
0
-2
-4
-4
-2
0
2
4
6

Non-parametric functions



Non-parametric functions



Jointly Gaussian functions II

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

Gaussian Distribution - Marginal

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

Gaussian Distribution - Marginal

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

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

Gaussian Distribution - Marginal

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

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

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

Gaussian Distribution - Marginal

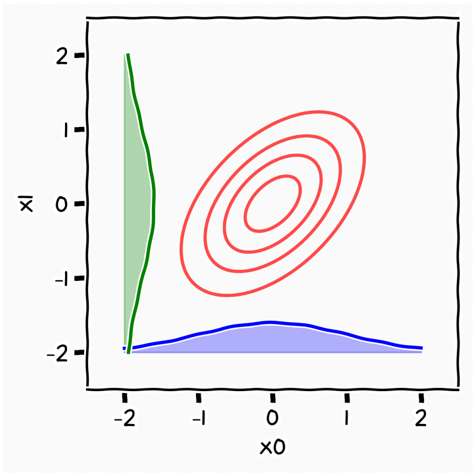
$$p(x_1, x_2) = \mathcal{N}\left(\begin{array}{c|c} x_1 & \mu_1 \\ x_2 & \mu_2 \end{array}, \begin{array}{cc} k_{11} & k_{12} \\ k_{21} & k_{22} \end{array}\right)$$

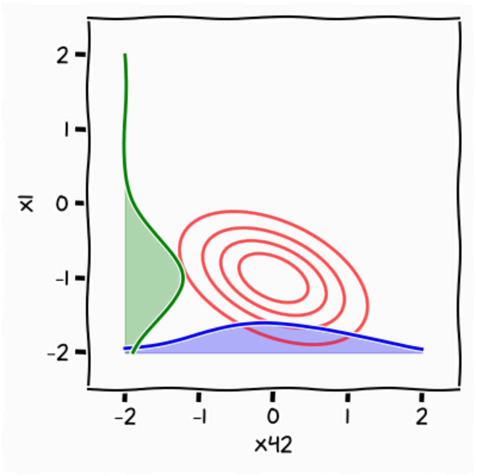
$$\Rightarrow p(x_1) = \int_{x_2} p(x_1, x_2) = \underline{\mathcal{N}\left(x_1 \mid \mu_1, k_{11}\right)}$$

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

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

Gaussian Distribution - Marginal



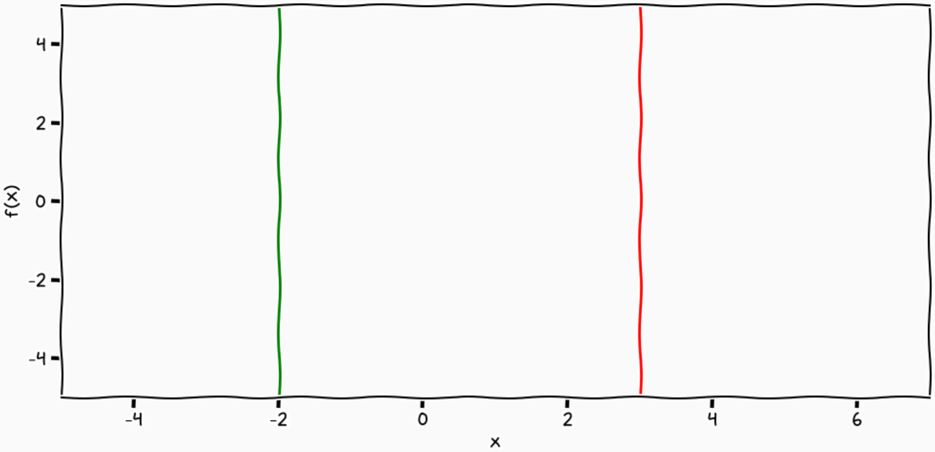


Marginal Property (Consistency)

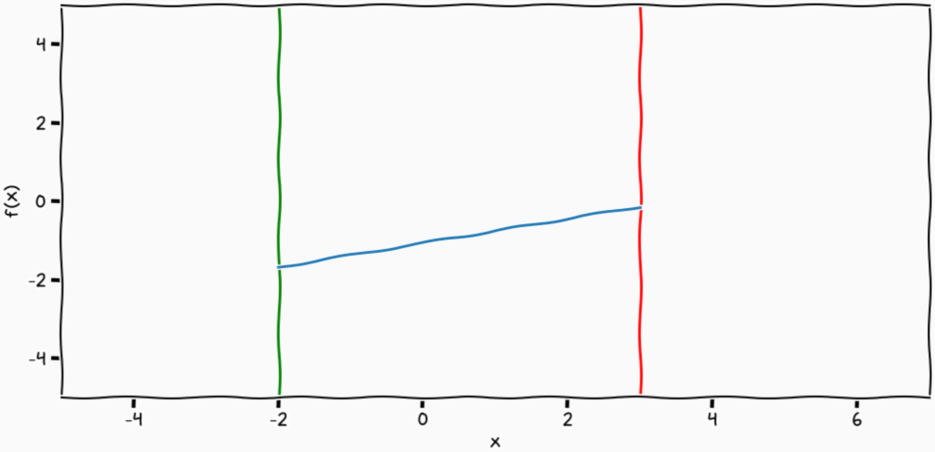
For all measurable sets $F_i \subseteq \mathbb{R}^n$ and probability measure $\mathcal{N}$

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

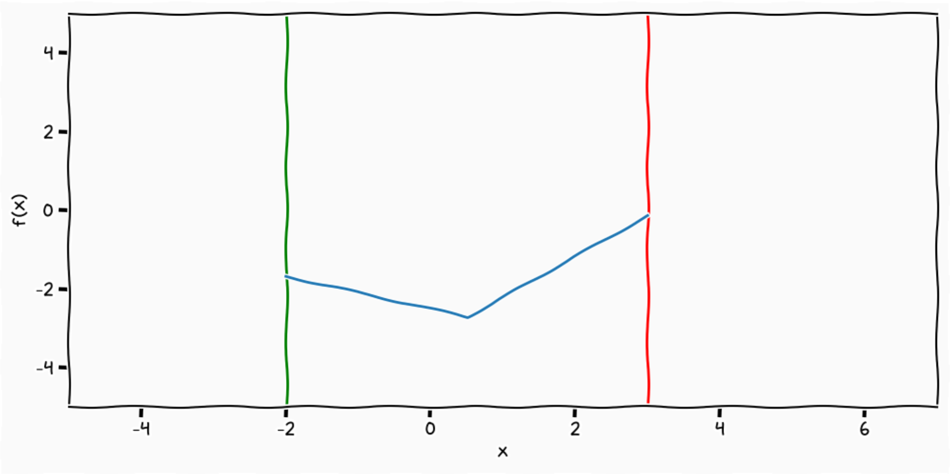
Gaussian Samples



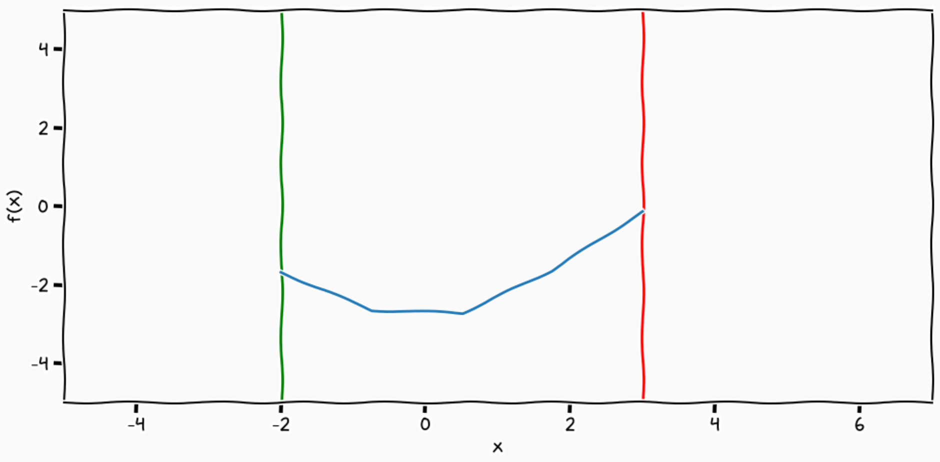
Gaussian Samples



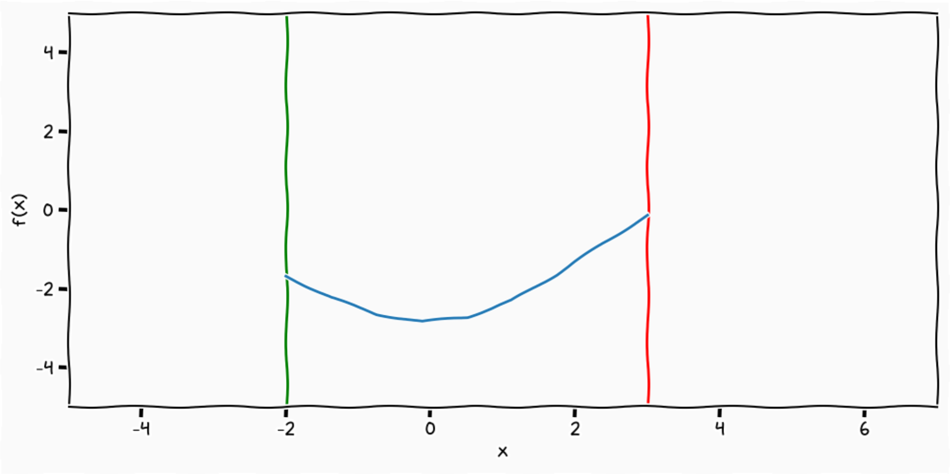
Gaussian Samples



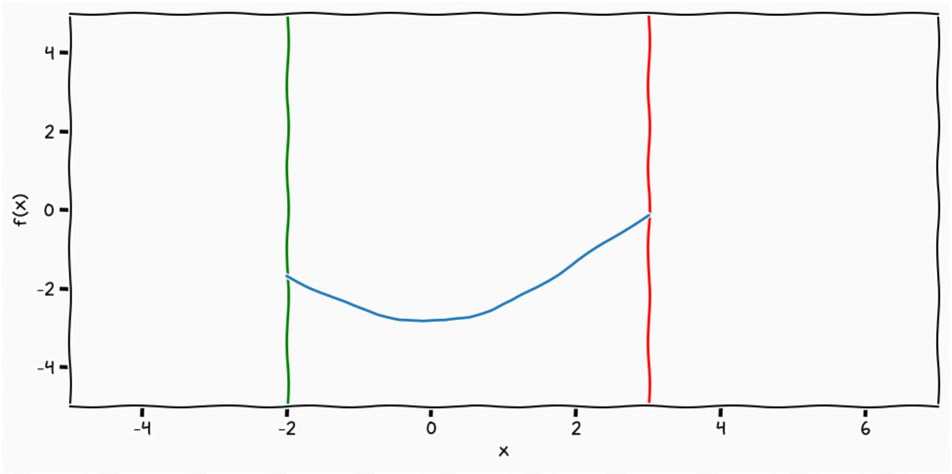
Gaussian Samples



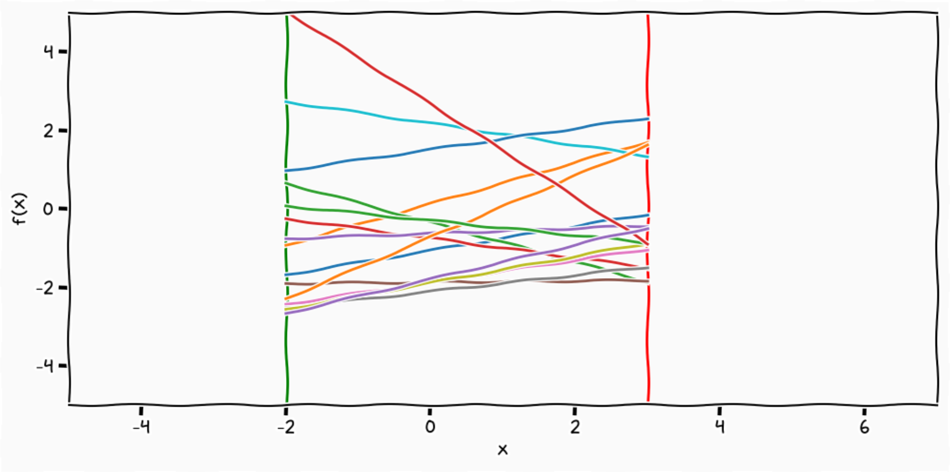
Gaussian Samples



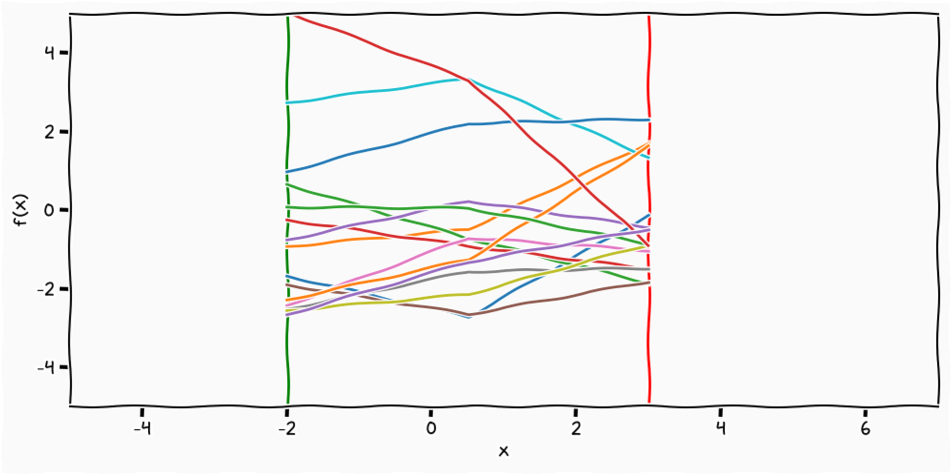
Gaussian Samples



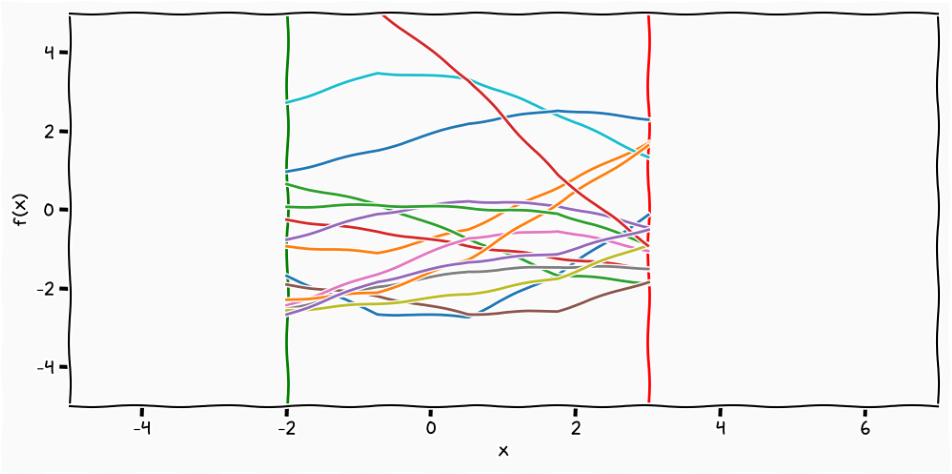
Gaussian Samples



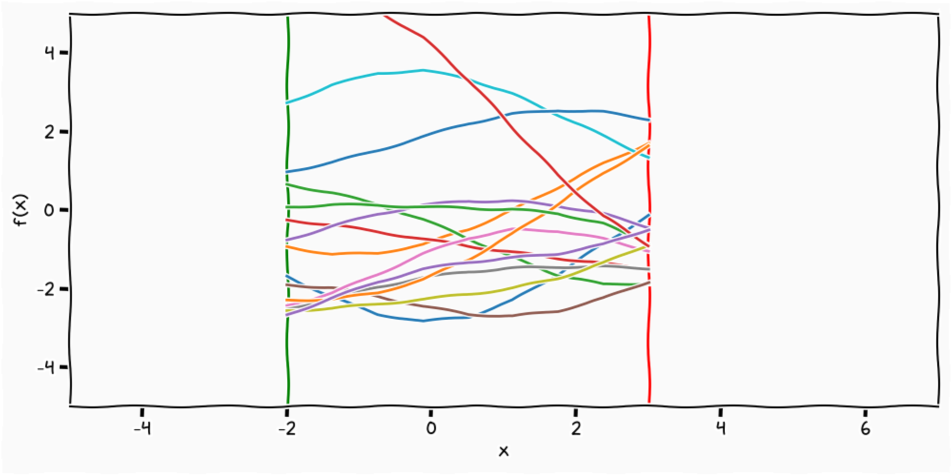
Gaussian Samples



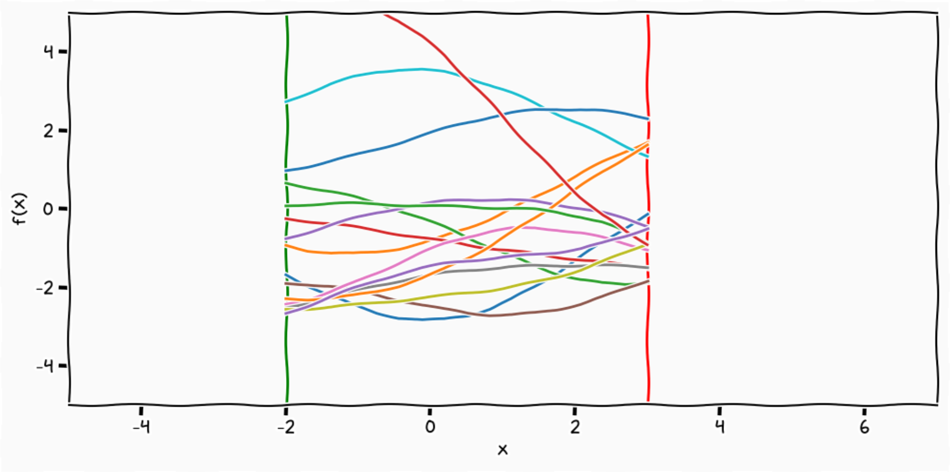
Gaussian Samples



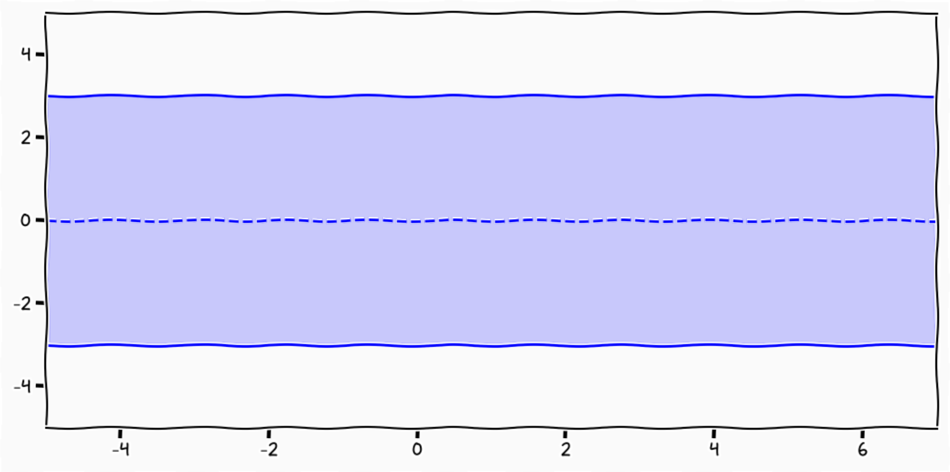
Gaussian Samples



Gaussian Samples



Gaussian Processes



Gaussian Processes: Formalism

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

Gaussian processes

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

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

The Gaussian distribution is the projection of the infinite Gaussian process

Definition (Gaussian Process)

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

Gaussian Processes: Formalism II

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

”Parametrisation”

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

- We parameterise the covariance as a function of the input

”Parametrisation”

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

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

”Parametrisation”

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

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

”Parametrisation”

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

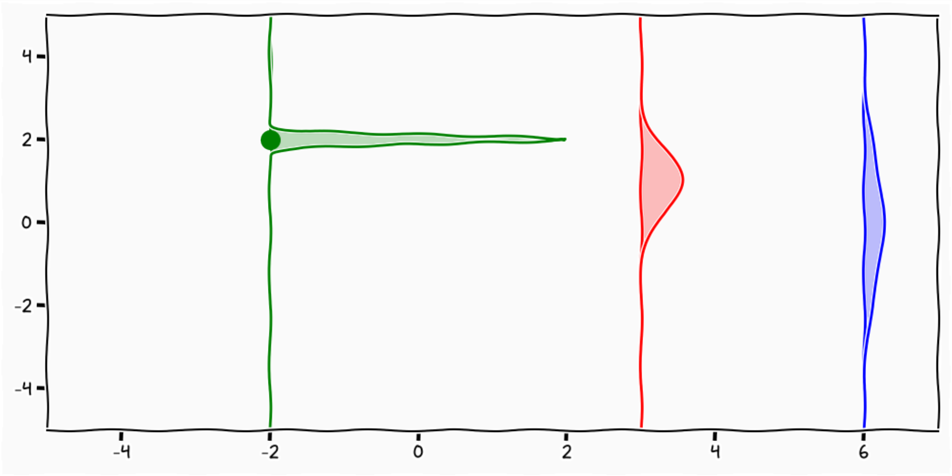
- We parameterise the covariance as a function of the input
- the index set of the measure is the uncountable infinity
- Your ”handle” to input your knowledge into a GP is the covariance function
  - *you specify the degree of covariance between data-points*

”Parametrisation”

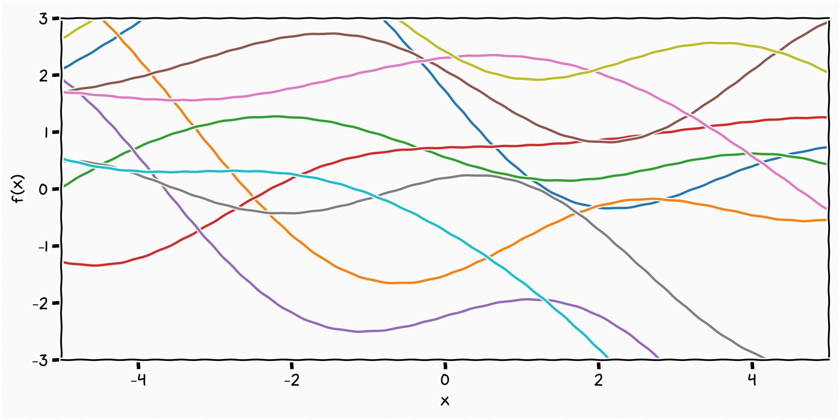
$$k_{ij} = k(x_i, x_j)$$

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

Gaussian Processes

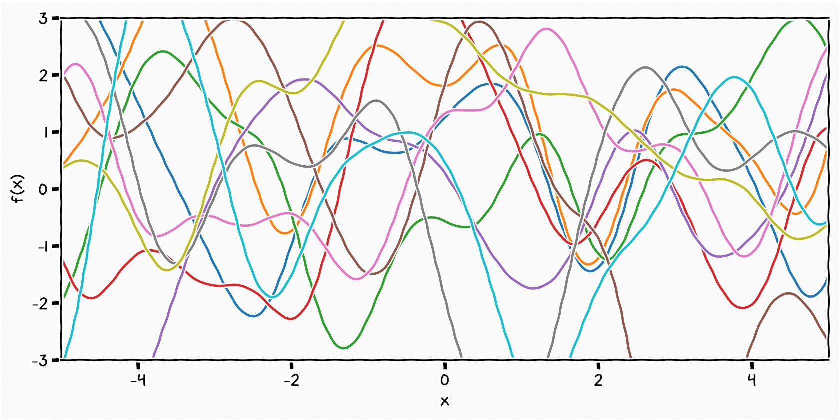


Gaussian Processes Samples



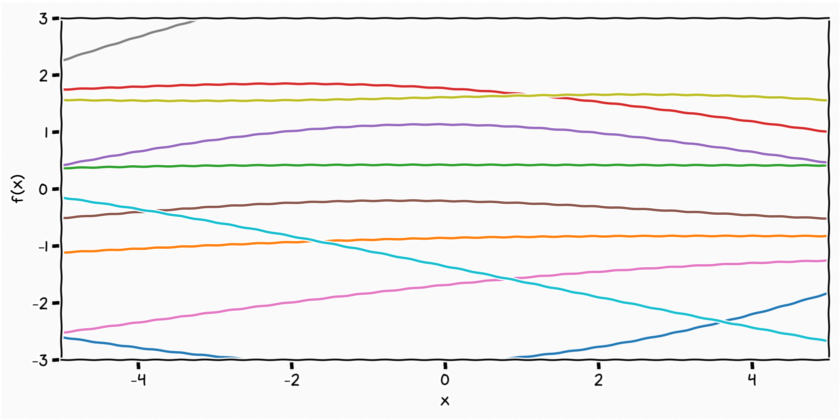
$$k(x_i, x_j) = 3 \cdot e^{-\frac{(x_i - x_j)^2}{15}}$$

Gaussian Processes Samples



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

Gaussian Processes Samples



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

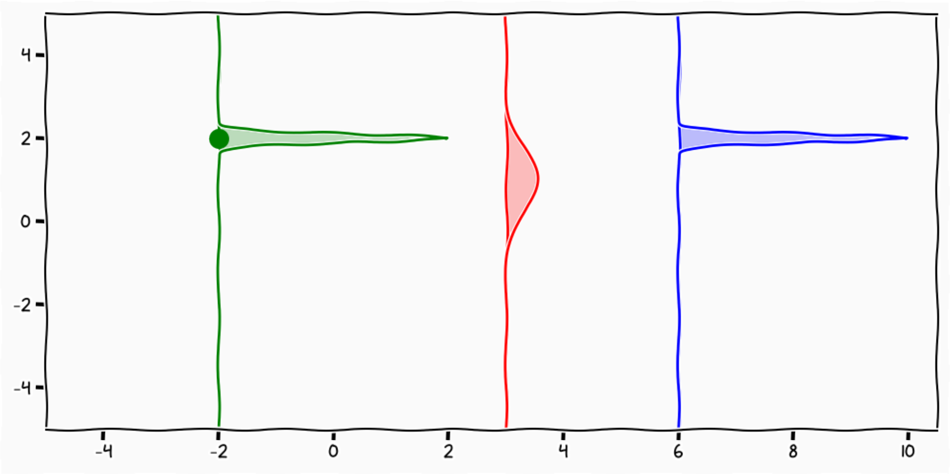
Code

```
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)
```

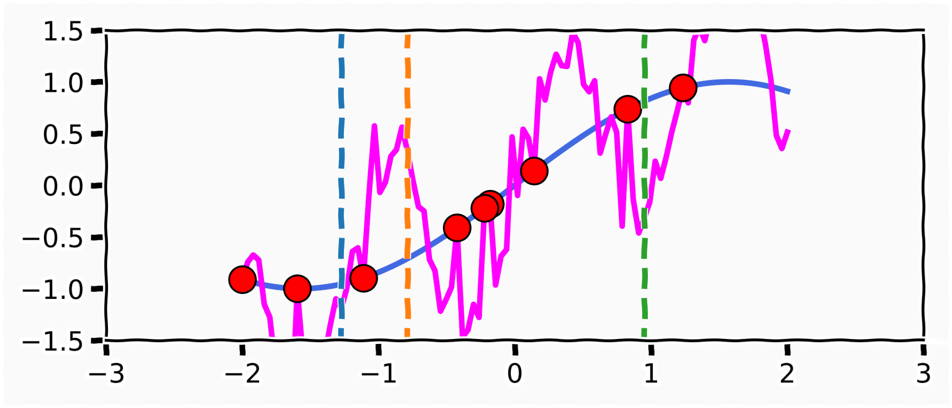
Gaussian Processes



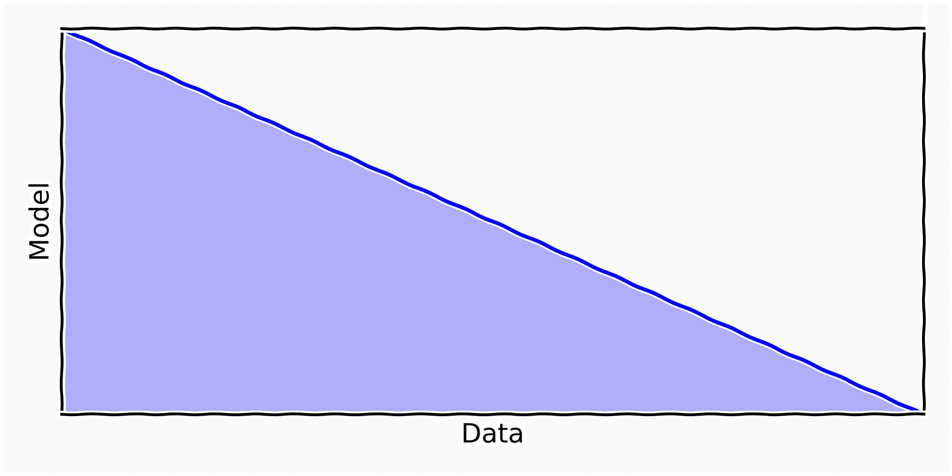
Choosing Covariances[2]

$$k(x, x') = ck_1(x, x')$$
$$k(x, x') = f(x)k_1(x, x')f(x')$$
$$k(x, x') = q(k_1(x, x')$$
$$k(x, x') = \exp(k_1(x, x'))$$
$$k(x, x') = k_1(x, x') + k_2(x, x')$$
$$k(x, x') = k_1(x, x')k_2(x, x')$$
$$k(x, x') = k_3(\phi(x), \phi(x'))$$
$$k(x, x') = x^{\mathrm{T}}\mathbf{A}x'$$
$$k(x, x') = k_a(x_a, x'_a) + k_b(x_b, x'_b)$$
$$k(x, x') = k_a(x_a, x'_a)k_b(x_b, x'_b)$$

[2] Bishop, 2006.



1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−3
−2
−1
0
1
2
3


Model
Data

Inference

Bayes' Rule

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

Marginal Likelihood

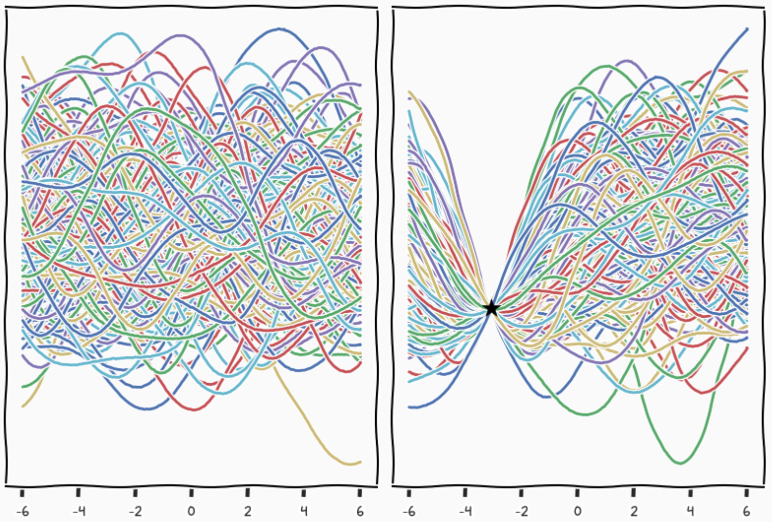
$$\int p(\mathbf{f}, \mathbf{f}_*) \mathrm{d}\mathbf{f}_* = \int p(\mathbf{f} \mid \mathbf{f}_*) p(\mathbf{f}_*) \mathrm{d}\mathbf{f}_*$$

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

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

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

Gaussian Processes: Posterior Samples



Gaussian Process: ”Predictive Posterior”

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

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

Gaussian Process: ”Predictive Posterior”

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

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

Gaussian Process: ”Predictive Posterior”

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

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

Gaussian Process: ”Predictive Posterior”

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

- We have defined $p(\mathbf{f}, \mathbf{f}_*)$ 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 $p(\mathbf{f}_* \mid \mathbf{f})$ is Gaussian process
- $\Rightarrow$ *We can just solve for* $p(\mathbf{f}_* \mid \mathbf{f})$

Gaussian Process: ”Predictive Posterior”

- All instantiations are jointly Gaussian

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

Gaussian Process: ”Predictive Posterior”

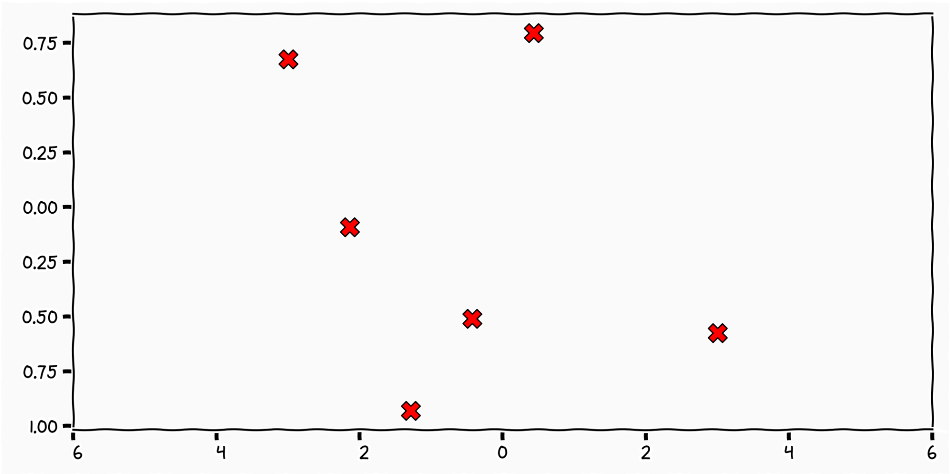
- All instantiations are jointly Gaussian

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

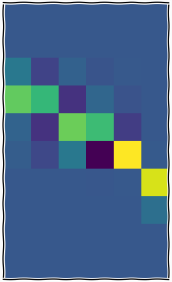
- Conditional Gaussian

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

Intuition

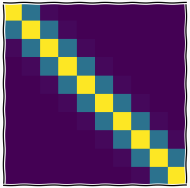


Does it make sense: Mean



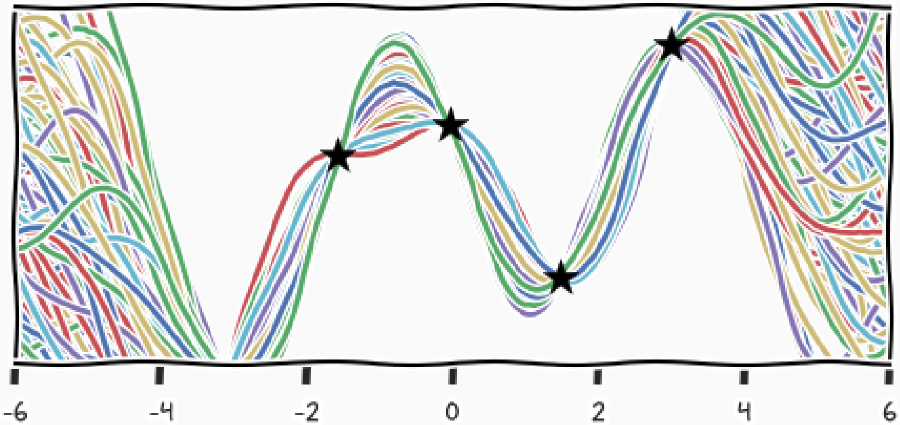
$$k(\mathbf{x}_*, \mathbf{X})^{\mathrm{T}} k(\mathbf{X}, \mathbf{X})^{-1} \mathbf{f}$$

Does it make sense: Covariance

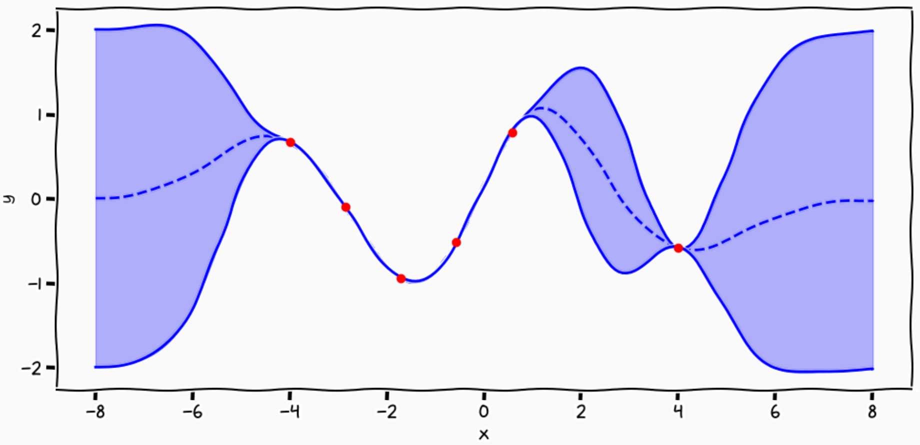


$$k(\mathbf{x}_*, \mathbf{x}_*) - k(\mathbf{x}_*, \mathbf{x})^{\mathrm{T}} k(\mathbf{x}, \mathbf{x})^{-1} k(\mathbf{x}, \mathbf{x}_*)$$

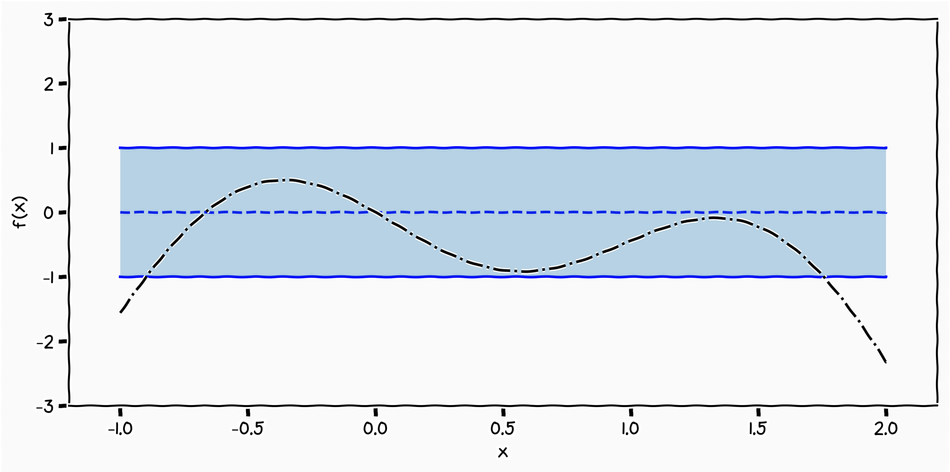
Gaussian Processes: "Predictive Posterior Samples"



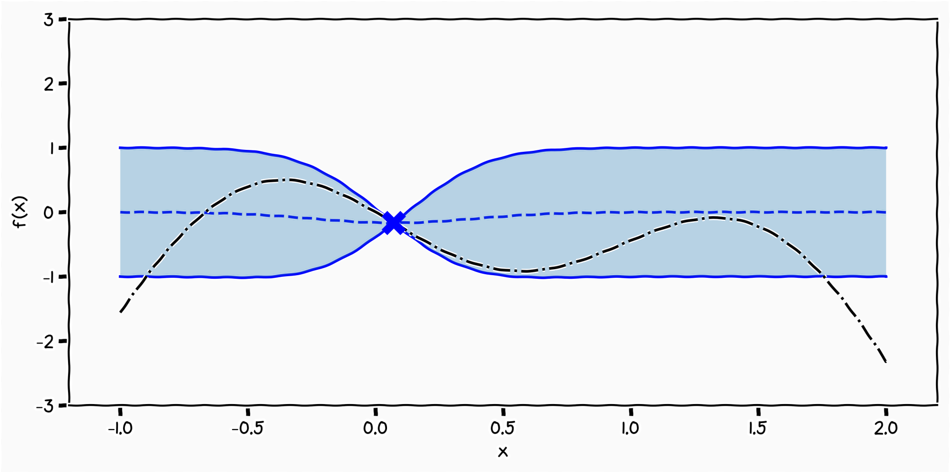
Gaussian Processes: "Predictive Posterior Process"



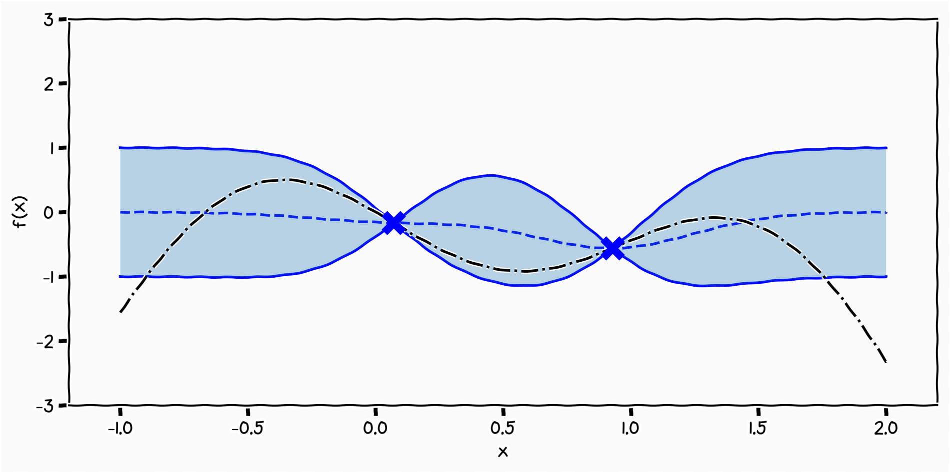
Posterior Processes



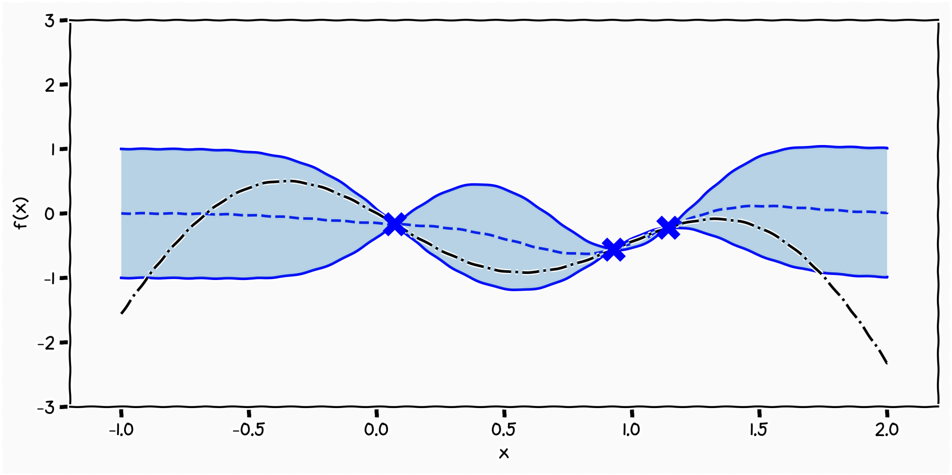
Posterior Processes



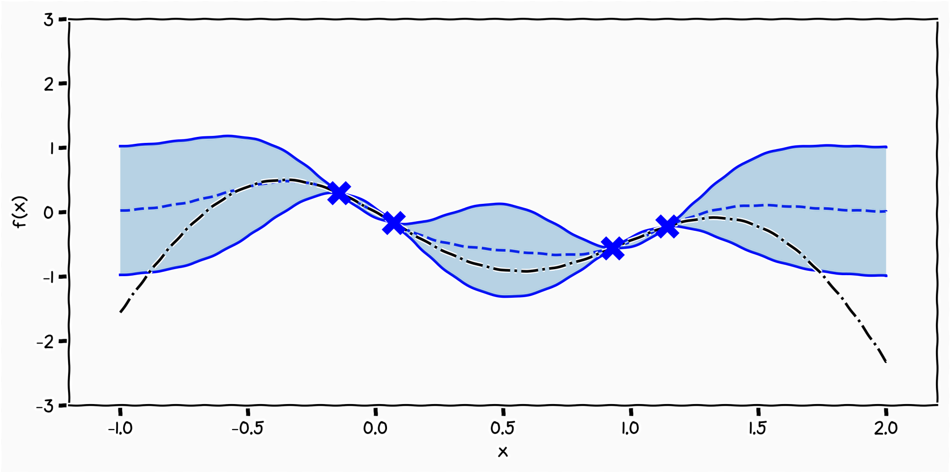
Posterior Processes

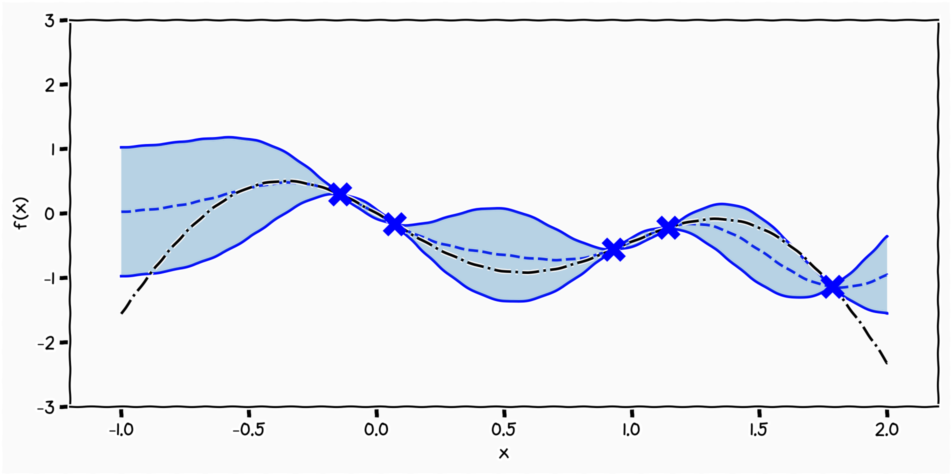


Posterior Processes

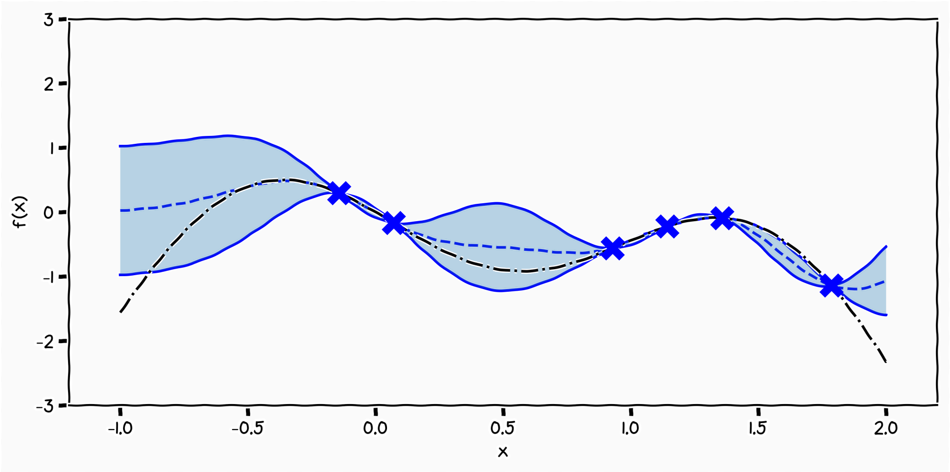


Posterior Processes

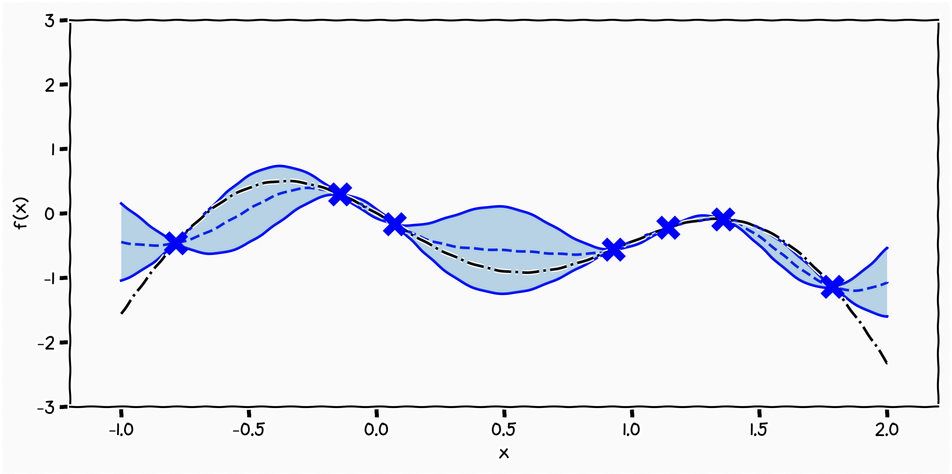



3
2
1
0
-1
-2
-3
f(x)
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
x

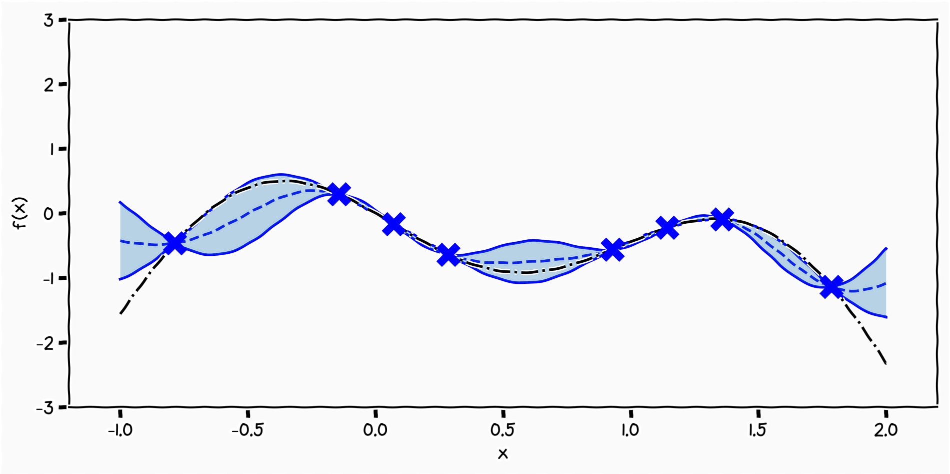
Posterior Processes



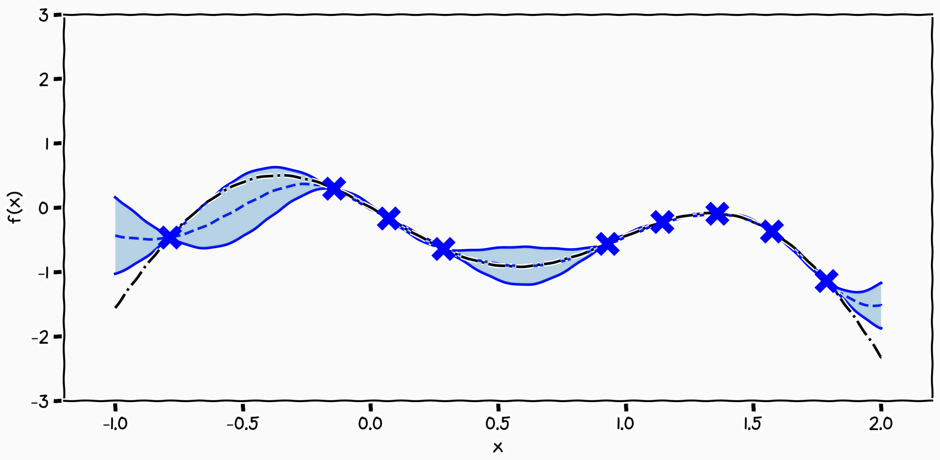
Posterior Processes

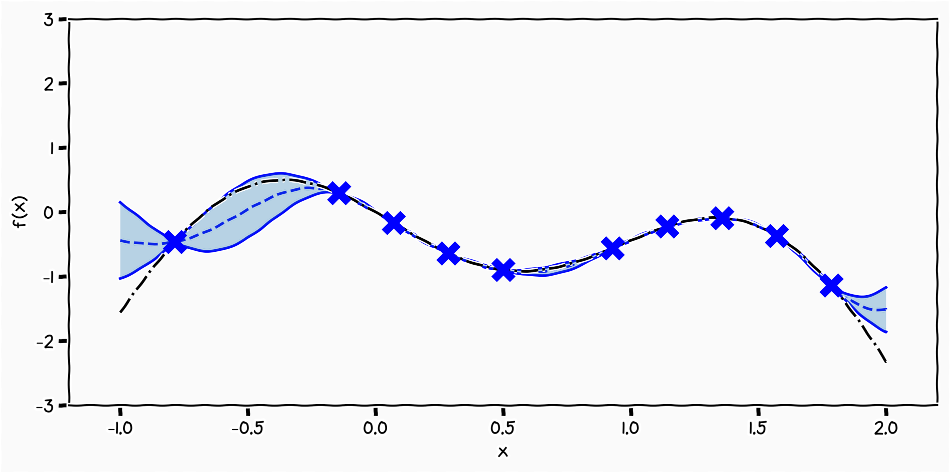


Posterior Processes

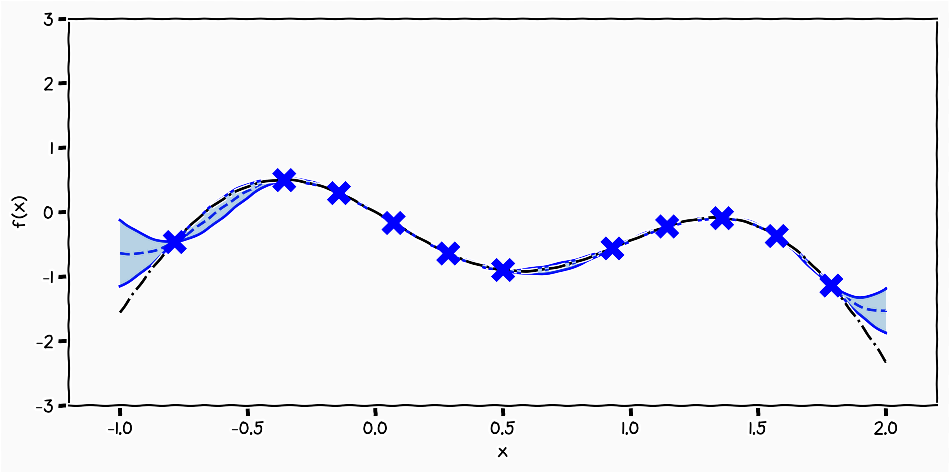


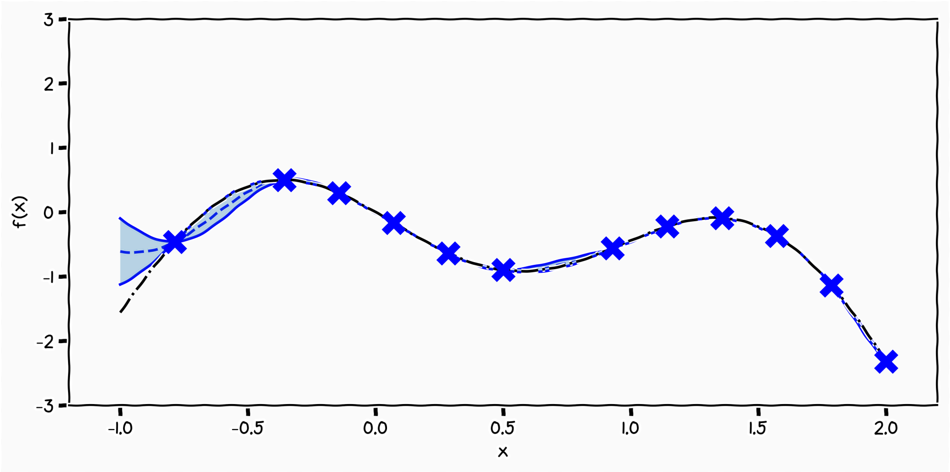
Posterior Processes




3
2
1
0
-1
-2
-3
f(x)
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
x

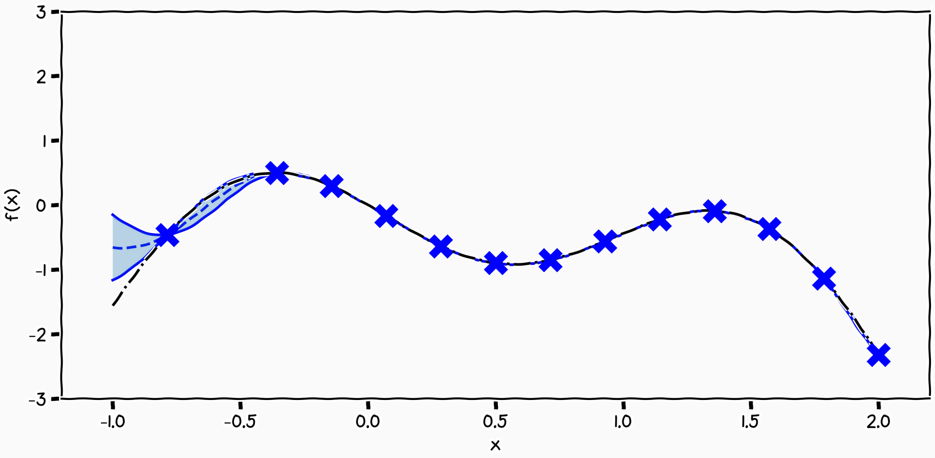
Posterior Processes



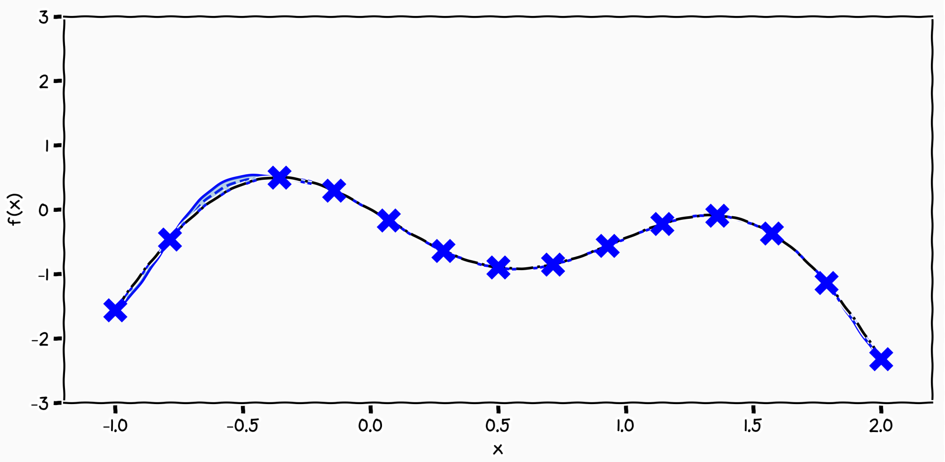


3
2
1
0
-1
-2
-3
f(x)
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
x

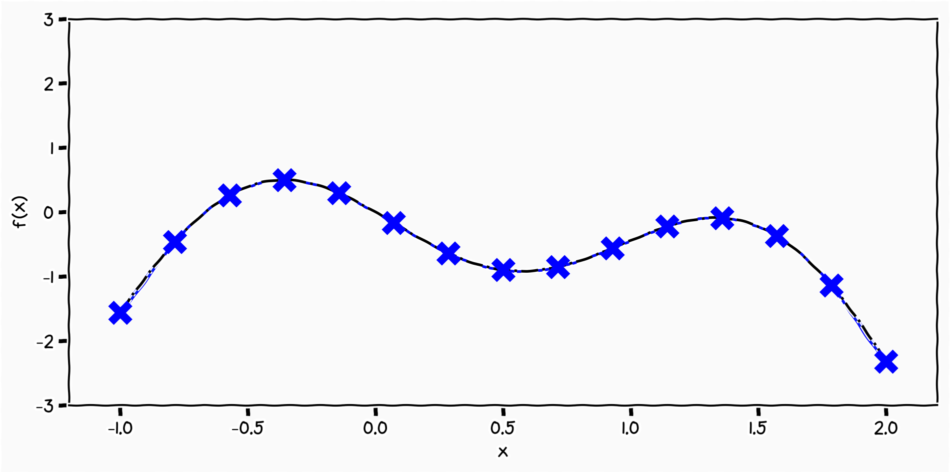
Posterior Processes



Posterior Processes



Posterior Processes



Taking Stock

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

- we have defined a measure over functions

Taking Stock

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

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

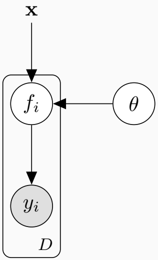
Taking Stock

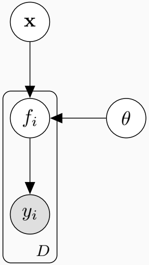
$$p(\mathbf{f}) \sim \mathcal{N}(\mathbf{f} \mid \mu(\cdot), k(\cdot,\cdot)), p(\mathbf{f}_*|\mathbf{f}) = \mathcal{N}(\mathbf{f}_*(\mathbf{x}_*, \mathbf{x})^{\mathrm{T}} k(\mathbf{x}, \mathbf{x})^{-1}\mathbf{f},$$
$$k(\mathbf{x}_*, \mathbf{x}_*) - k(\mathbf{x}_*, \mathbf{x})^{\mathrm{T}} k(\mathbf{x},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

- 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

Models

Learning

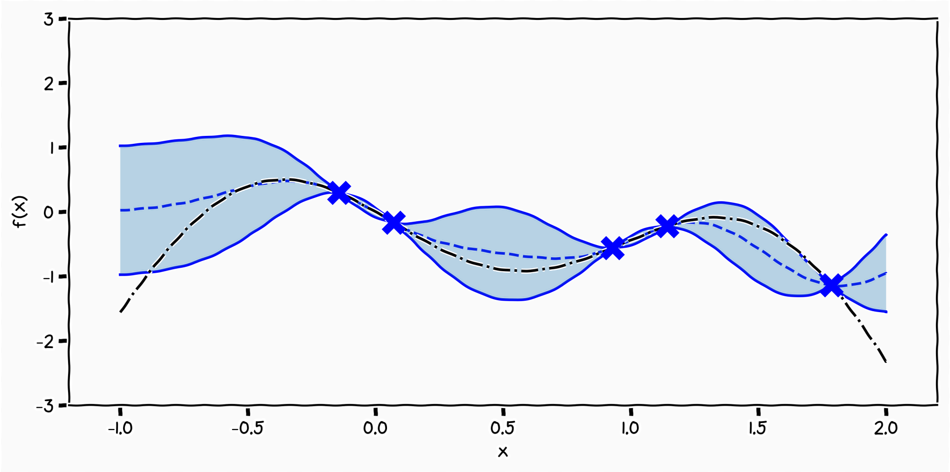




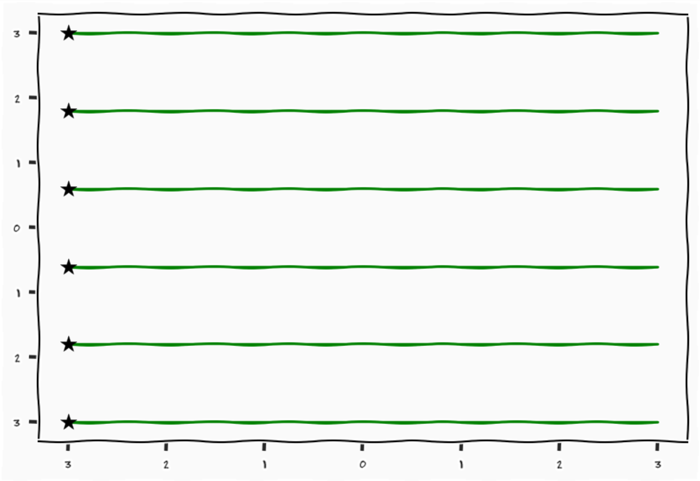
$$p(y|x) = \int p(y \mid f)p(f \mid x)\mathrm{d}f$$

$$p(y) = \int p(y \mid f)p(f \mid x)p(x)\mathrm{d}f\mathrm{d}x$$

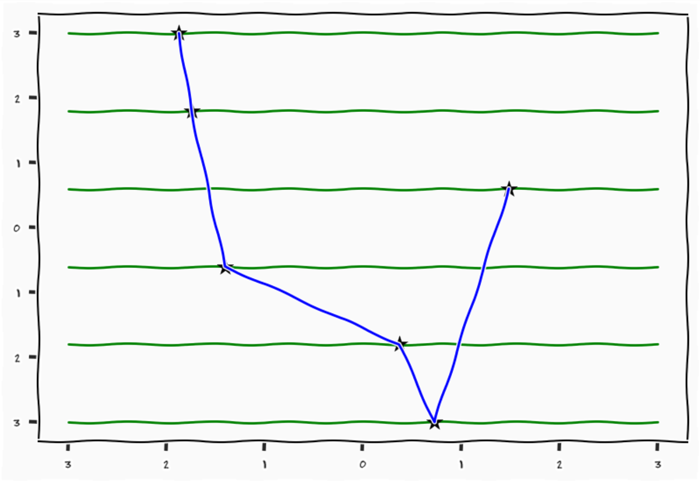
Posterior Processes

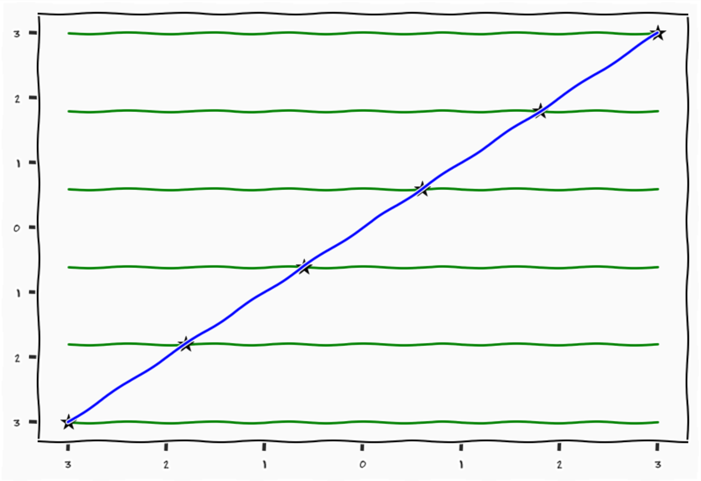


Unsupervised Learning

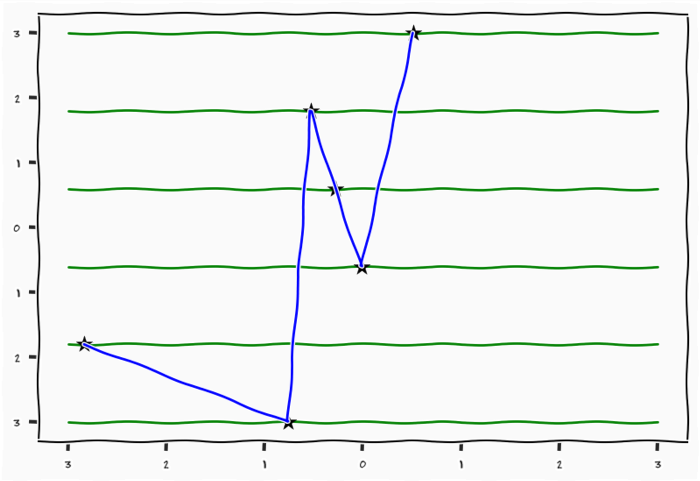


Unsupervised Learning

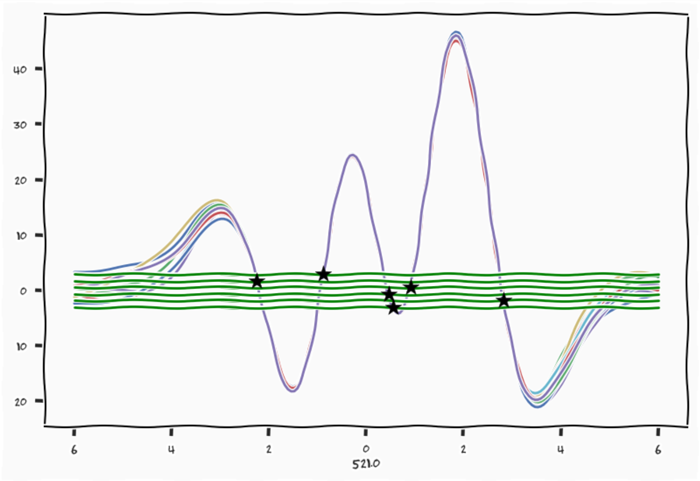



3
2
1
0
1
2
3
3
2
1
0
1
2
3

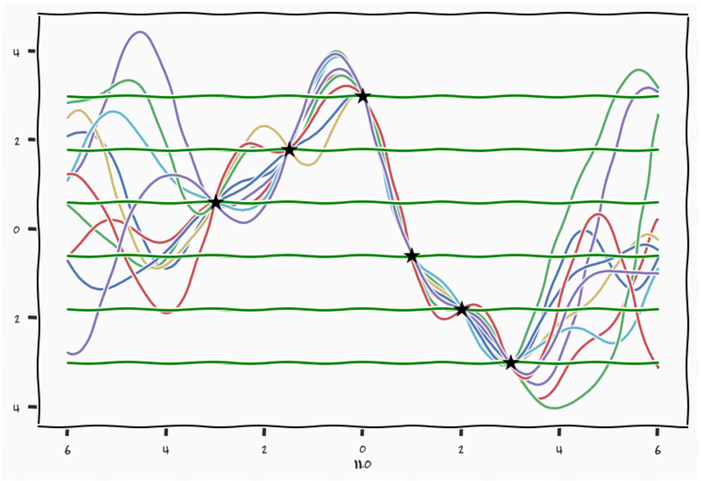
Unsupervised Learning



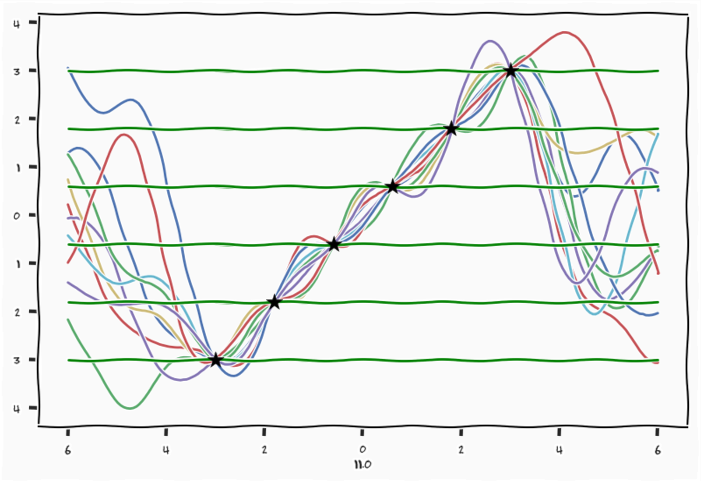
Unsupervised Learning

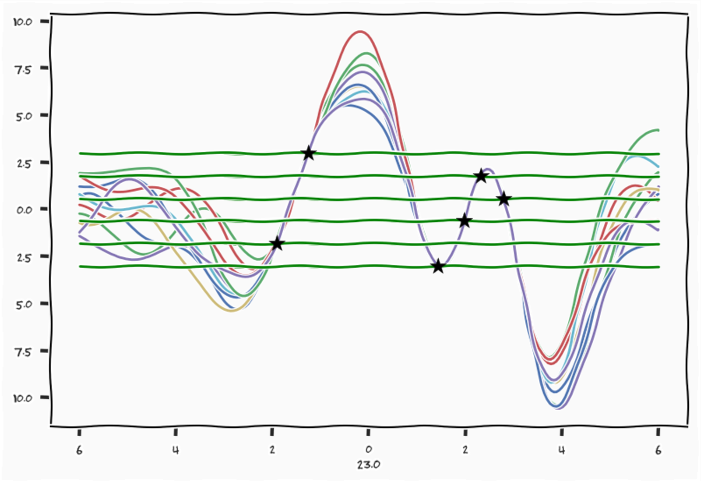


Unsupervised Learning



Unsupervised Learning




10.0
7.5
5.0
2.5
0.0
2.5
5.0
7.5
10.0
6
4
2
0
2
4
6
23.0

Being Bayesian

$$p(y) = \int p(y \mid f_2)p(f_2 \mid f_1)\mathrm{d}f_2\mathrm{d}f_1$$

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

Gaussian Process Latent Variable Model

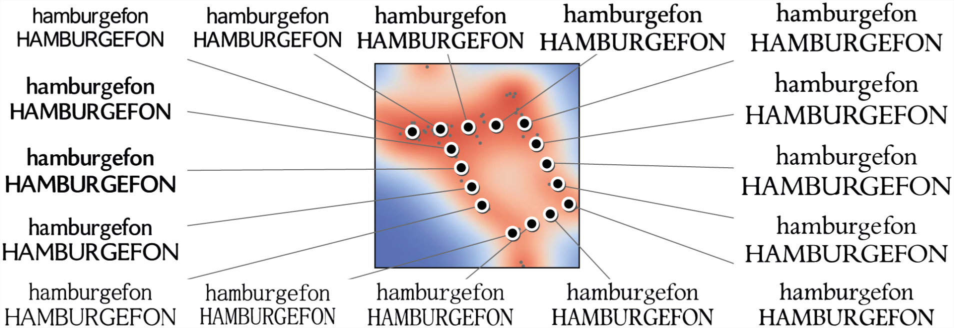
Regression there are infinite number of possible functions that connects the data equally well. A GP provides a measure over these solutions that makes the problem ”well-posed”.

Gaussian Process Latent Variable Model

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”

Fonts “Learning a manifold of fonts”



URL

Multi-Fidelity Models

Summary

Summary

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

Summary

- 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

Summary

- 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 religously about being Bayesian or not boils down to do if you agree with the process of marginalisation

Summary

- 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 religously 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

Non-parametrics

- infinite capacity by parametrising the model through relationship between data

Non-parametrics

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

Non-parametrics

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

Non-parametrics

- 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

Non-parametrics

- 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

Kolmogrovs Extension Theorem

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

1. Exchangeable

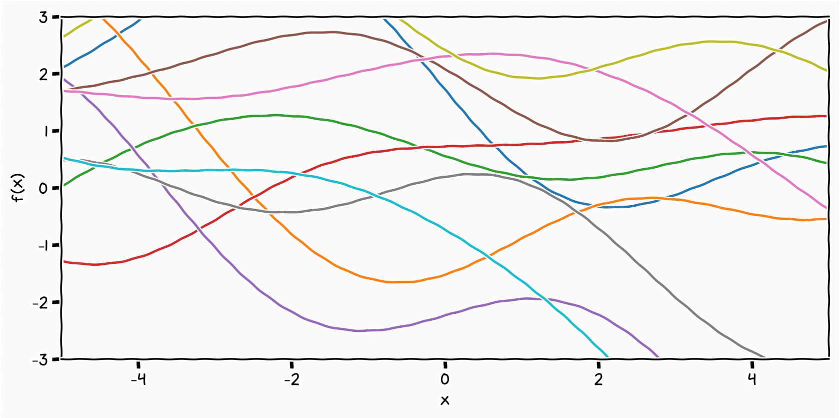
$$\nu_{t_{\pi(1)} \cdots t_{\pi(k)}} \left(F_{\pi(1)} \times \cdot \times F_{\pi(k)}\right) = \nu_{t_1 \cdots t_k} \left(F_1 \times \cdots \times F_k\right)$$

2. Marginal

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

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

Are Gaussian Processes good parametrisations?



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 eof

References

- Bishop, Christopher M. (2006). ***Pattern Recognition and Machine Learning (Information Science and Statistics).*** Secaucus, NJ, USA: Springer-Verlag New York, Inc.
- Campbell, NDF and J Kautz (July 2014). **“Learning a manifold of fonts”.** In: *ACM Transactions on Graphics (TOG)* 33.4, p. 91.
- Chomsky, Noam A and Jerry A Fodor (1980). **“The inductivist fallacy”.** In: *Language and Learning: The Debate between Jean Piaget and Noam Chomsky.*

Roy, Hrittik, Marco Miani, Carl Henrik Ek, Philipp Hennig, Marvin Pförtner, Lukas Tatzel, and Søren Hauberg (2024). **"Reparameterization Invariance in Approximate Bayesian Inference".** In: *Advances in Neural Information Processing Systems (NeurIPS)*.

Shalev-Shwartz, Shai and Shai Ben-David (2014). ***Understanding Machine Learning: From Theory to Algorithms.*** New York, NY, USA: Cambridge University Press.