Gaussian Processes

a second introduction

Carl Henrik Ek - che29@cam.ac.uk

September 11, 2023

http://carlhenrik.com

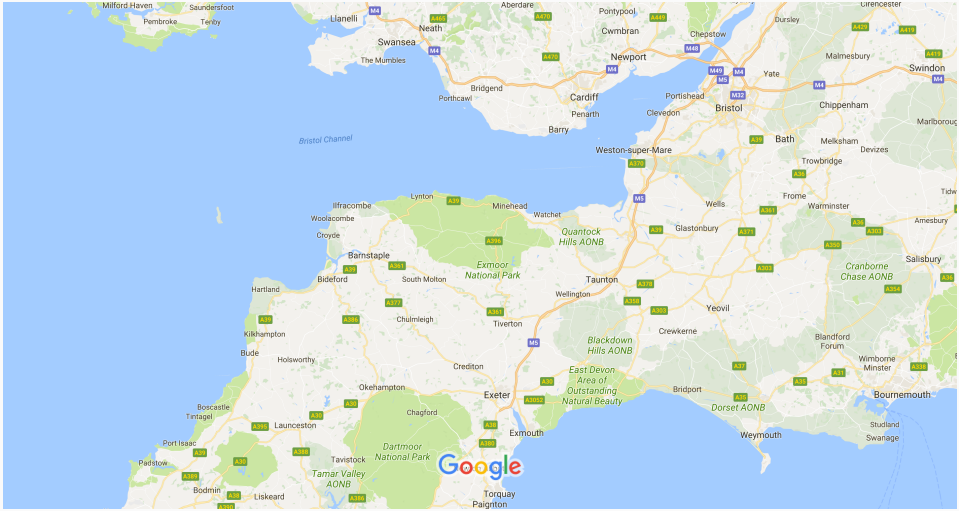

V
Velominati
KEEPERS OF THE COG

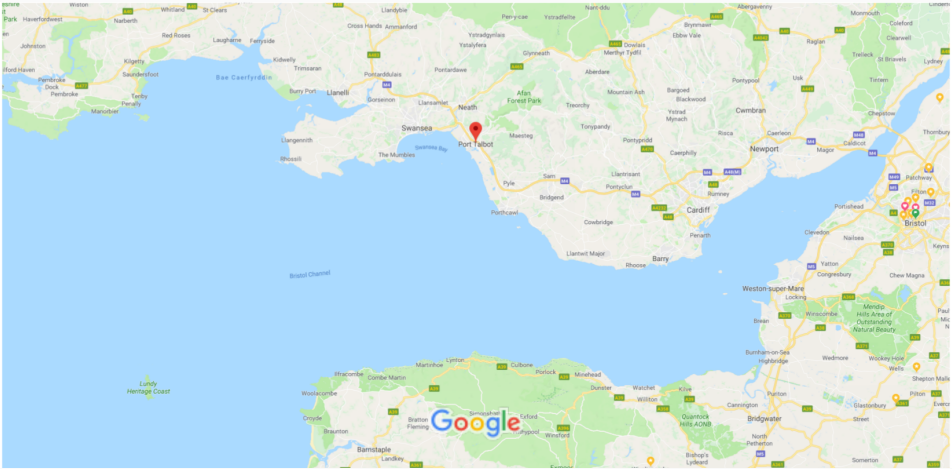
Rule #7 - Tan lines should be cultivated and kept razor sharp

"Under no circumstances should one be rolling up their sleeves or shorts in an effort to somehow diminish one's tan lines. Sleeveless jerseys are under no circumstances to be employed."

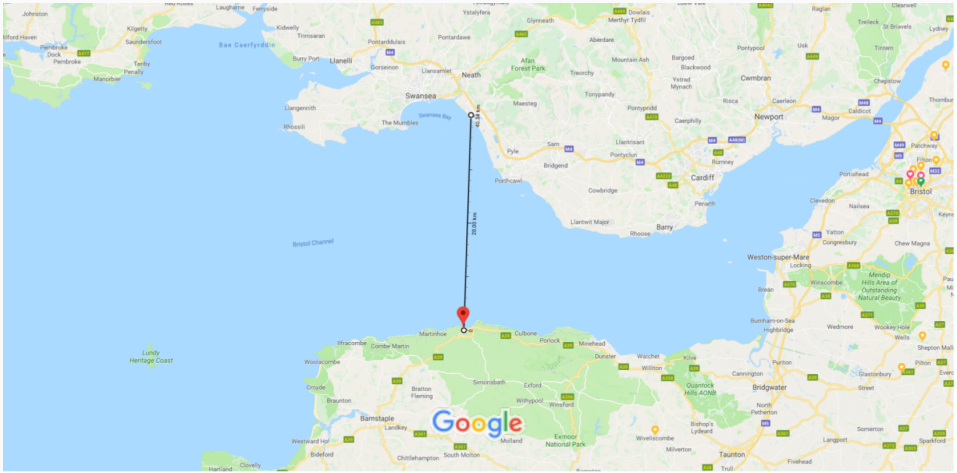
– https://www.velominati.com/

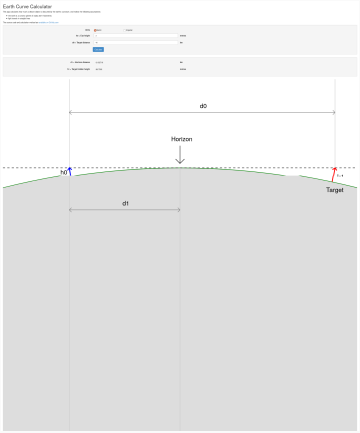

COLNAGO


Milford Haven
Saundersfoot
Pembroke
Tenby
Llanelli
Aberdare
Pontypool
Cirencester
Neath
Cwmbran
Dursley
Swansea
Chepstow
The Mumbles
Newport
Malmesbury
Swindon
Yate
Bridgend
Portishead
Porthcawl
Cardiff
Chippenham
Penarth
Clevedon
Bristol
Marlborough
Bristol Channel
Barry
Bath
Melksham
Weston-super-Mare
Devizes
Trowbridge
Ilfracombe
Lynton
Minehead
Frome
Wells
Warminster
Woolacombe
Watchet
Amesbury
Croyde
Quantock Hills AONB
Glastonbury
Barnstaple
Exmoor National Park
Salisbury
Bideford
South Molton
Taunton
Cranborne Chase AONB
Hartland
Wellington
Yeovil
Chulmleigh
Tiverton
Kilkhampton
Crewkerne
Blandford Forum
Blackdown Hills AONB
Bude
Holsworthy
Wimborne Minster
Crediton
East Devon Area of Outstanding Natural Beauty
Okehampton
Exeter
Bridport
Bournemouth
Boscastle
Tintagel
Chagford
Dorset AONB
Launceston
Exmouth
Studland
Swanage
Port Isaac
Dartmoor National Park
Weymouth
Padstow
Tavistock
Tamar Valley AONB
Bodmin
Liskeard
Torquay
Paignton
Google


Port Talbot
Swansea
Neath
Cardiff
Newport
Bristol
Barry
Weston-super-Mare
Bridgwater
Barnstaple
Bae Caerfyrddin
Bristol Channel
Lundy
Heritage Coast
Google

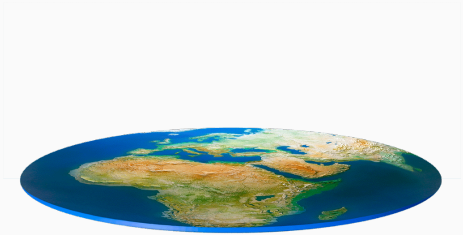

Swansea
Port Talbot
Llanelli
Neath
Cardiff
Newport
Barry
Bristol
Weston-super-Mare
Bridgwater
Taunton
Barnstaple
Bideford
Ilfracombe
Minehead
Bristol Channel
Lundy
Heritage Coast
Exmoor
National Park
Google


Swansea
Cardiff
Newport
Bristol
Barnstaple
Taunton
Bridgwater
Weston-super-Mare
Bristol Channel
Google



Distance to horizon $6.2km$

Hidden height $125.6m$



?

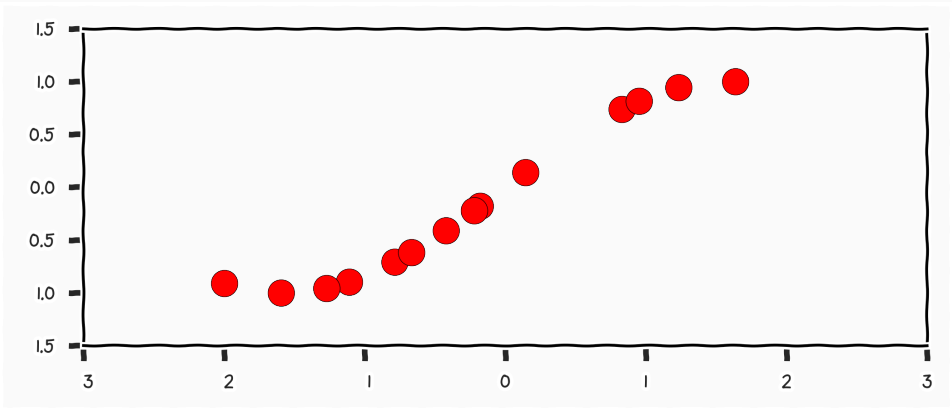
Flat Earth *The earth is flat and the water surface is flat, therefore I can see the building*

Occam's Razor

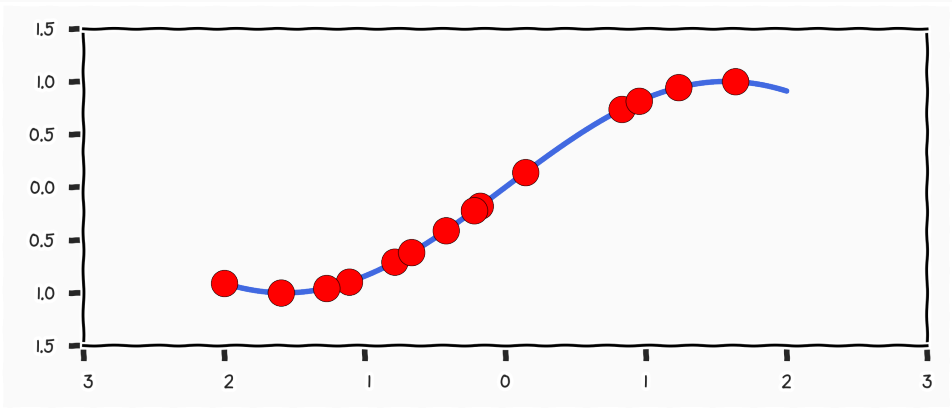
Flat Earth *The earth is flat and the water surface is flat, therefore I can see the building*

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*

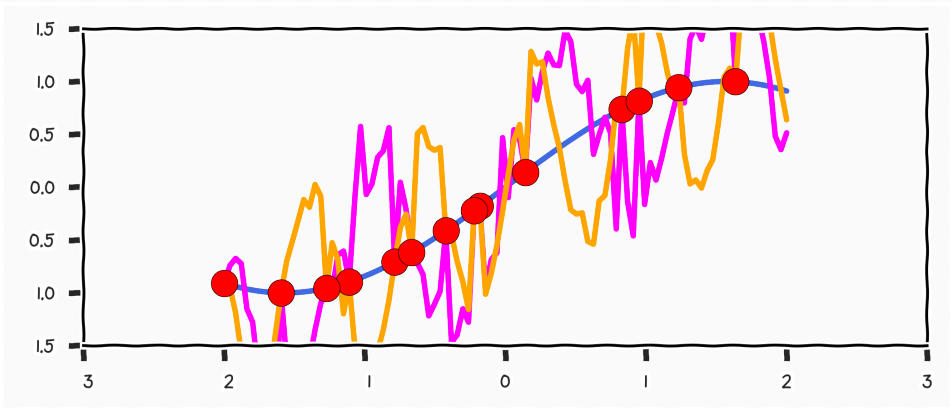
Curve Fitting

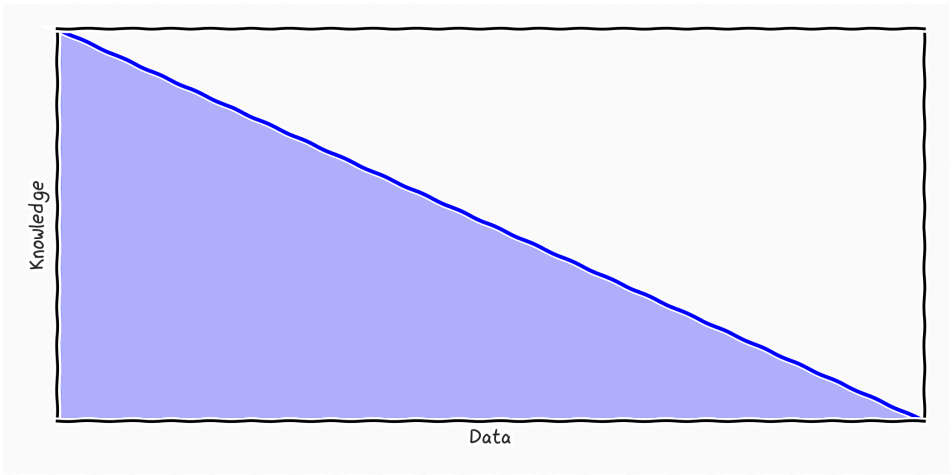


Curve Fitting



Curve Fitting




Knowledge
Data

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 change of narrative)

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

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

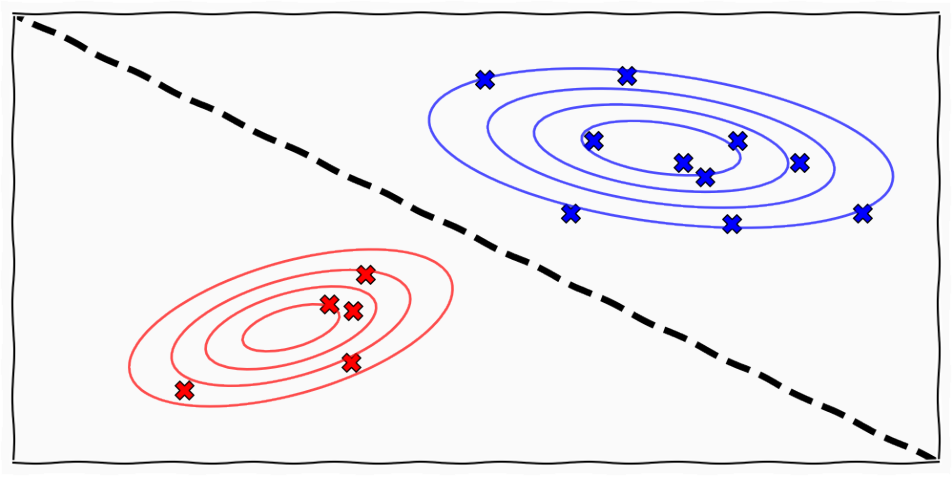
Formalism [Shalev-Shwartz et al., 2014]

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

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

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

Classification



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

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

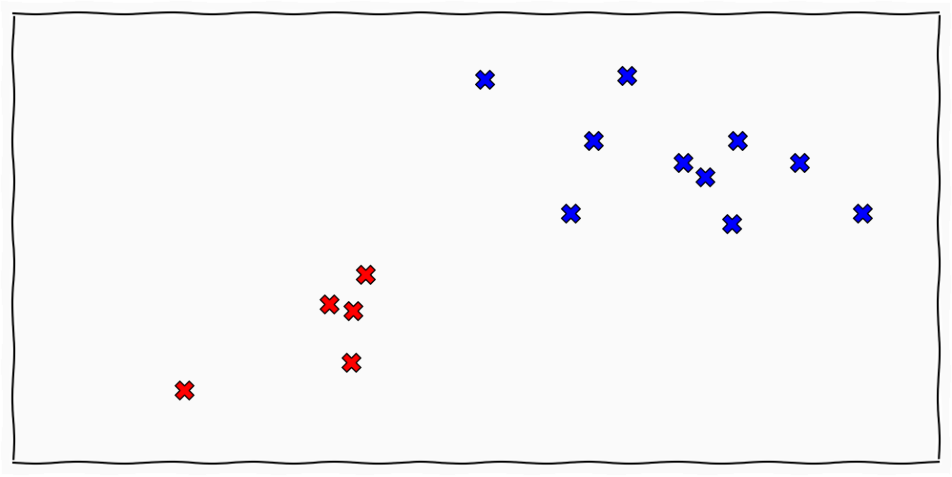
- measure of success as probability of misclassified points (true risk)
- we do not have access to $\mathcal{D}$

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

Algorithm

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

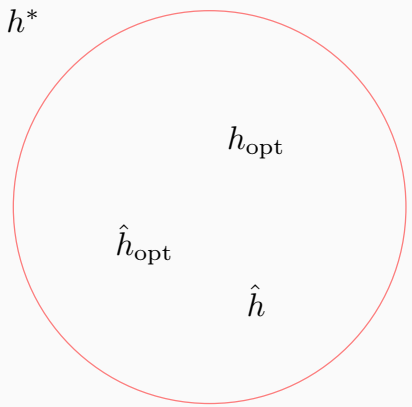
- We use an algorithm $A : \mathcal{S} \to h$ to find a hypothesis

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



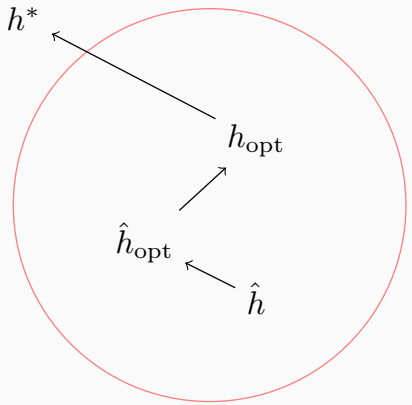
h^* the optimal predictor

$h_{\textbf{opt}}$ the optimal hypothesis

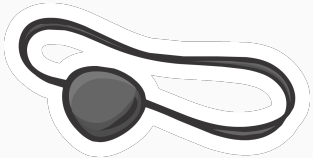
$\hat{h}_{\textbf{opt}}$ the optimal hypothesis on training data

$\hat{h}$ the hypothesis found by learning algorithm

Error Decomposition



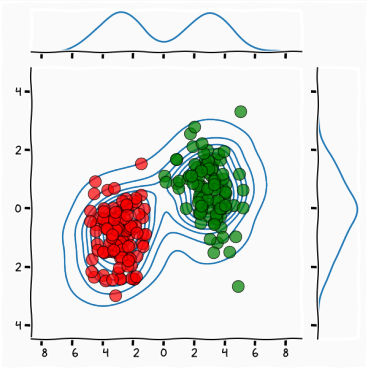
$$
\begin{aligned}
&\epsilon(\hat{h}) - \epsilon(h^*) \\
&= \underbrace{\epsilon(h_{\text{opt}}) - \epsilon(h^*)}_{\text{Approximation}} \\
&+ \underbrace{\epsilon(\hat{h}_{\text{opt}}) - \epsilon(h_{\text{opt}})}_{\text{Estimation}} \\
&+ \underbrace{\epsilon(\hat{h}) - \epsilon(\hat{h}_{\text{opt}})}_{\text{Optimisation}}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



Statistical Learning

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

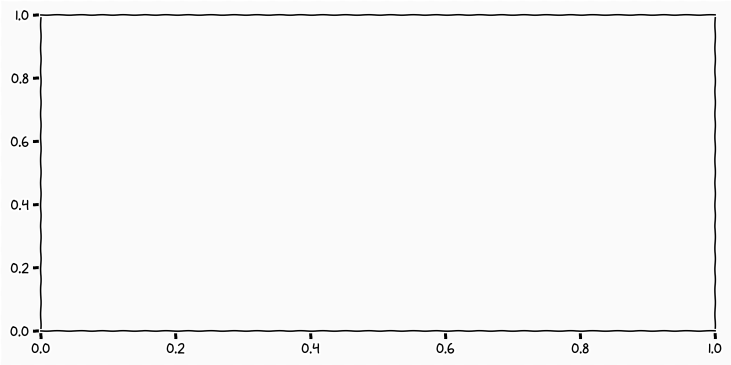
Assumptions: Biased Sample



Statistical Learning

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

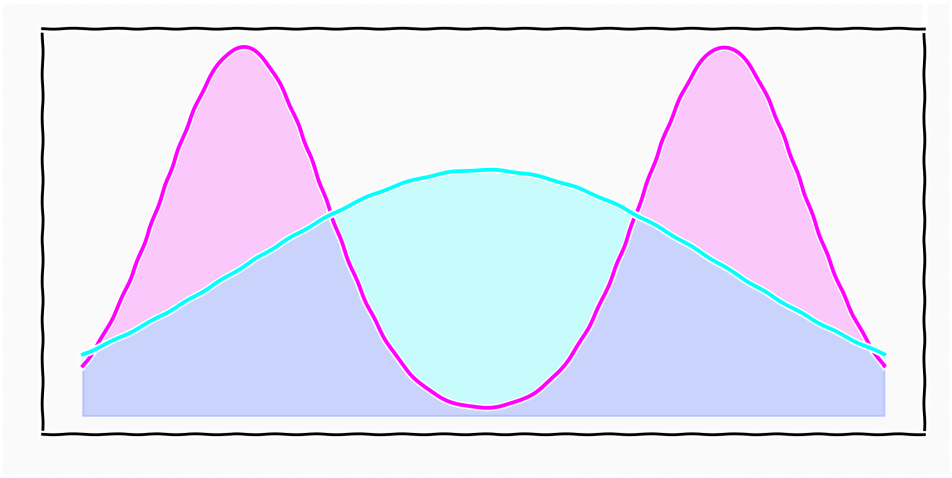
Assumptions: Hypothesis space



Statistical Learning

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

Quantifying Knowledge



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

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

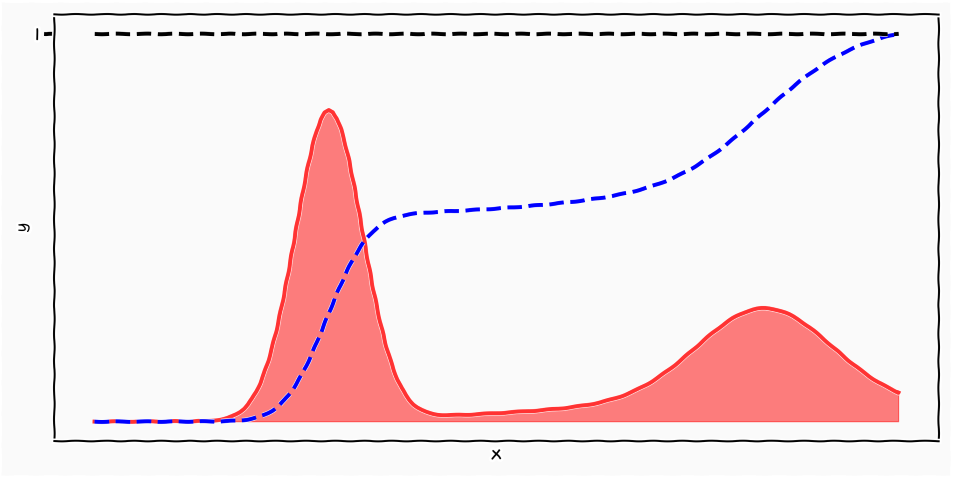
Marginalisation

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

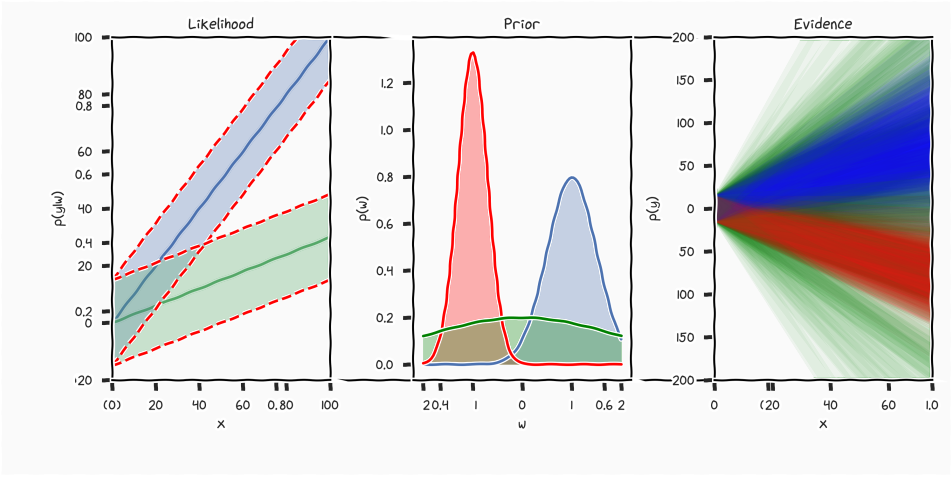
Marginalisation

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

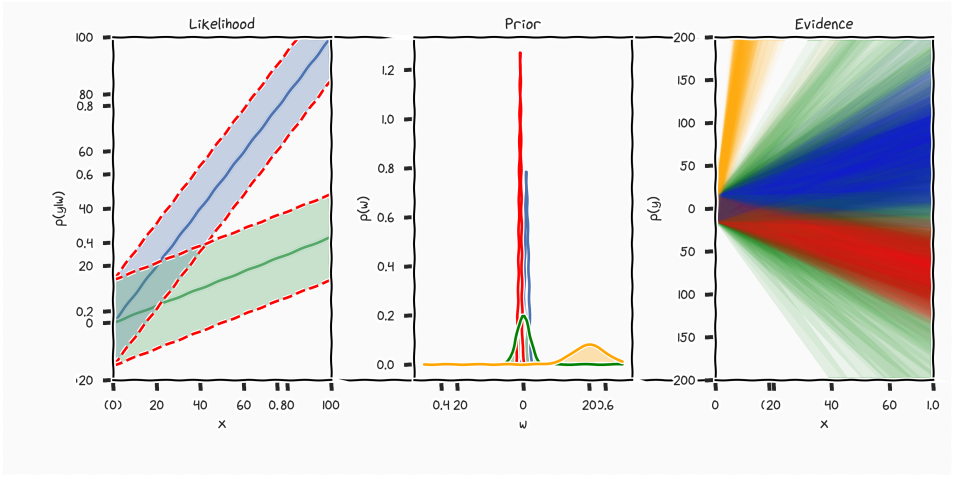
Marginalisation



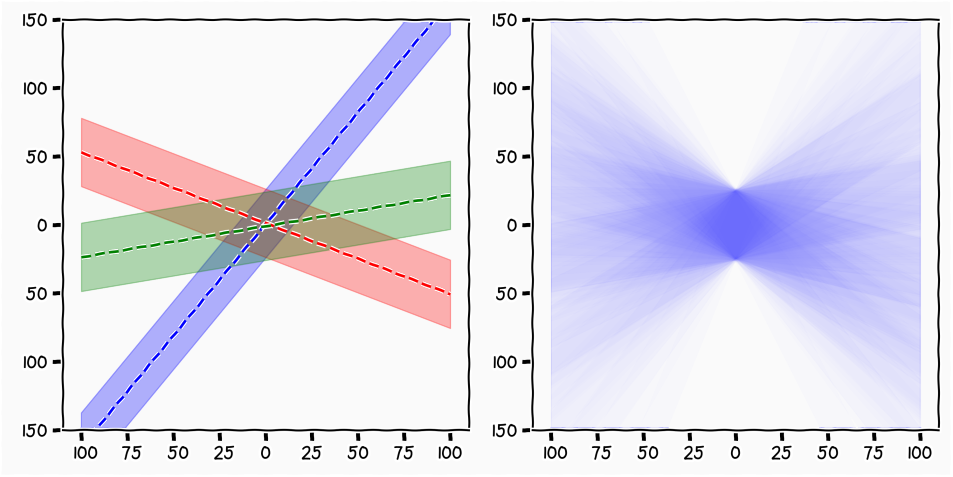
Marginalisation



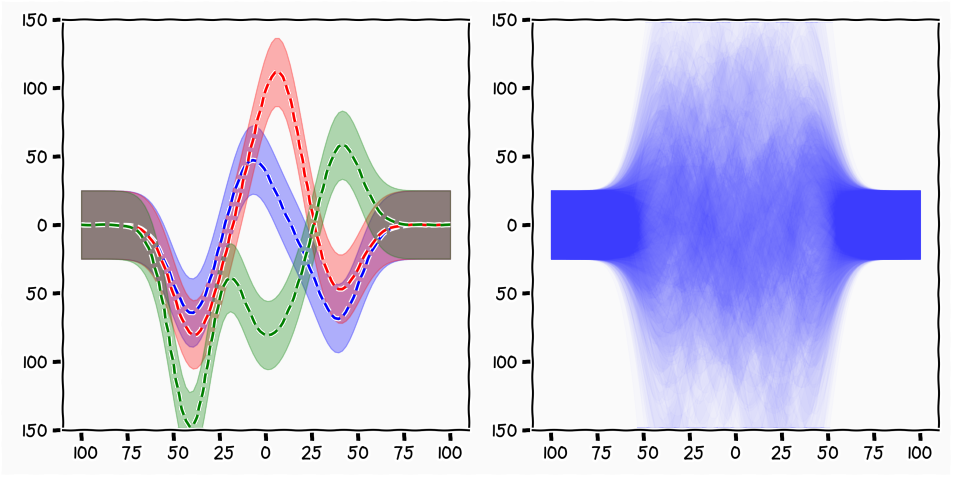
Marginalisation



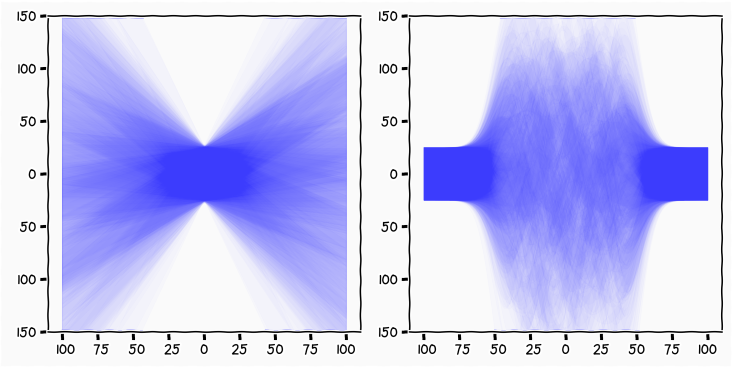
Model Linear Linear



Model Linear



Model Selection



$$\hat{p} = \max\{p_{\text{lin-lin}}, p_{\text{lin}}\}$$

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 predictions does the opposite and alters the search algorithm but keeps the space the same.

"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

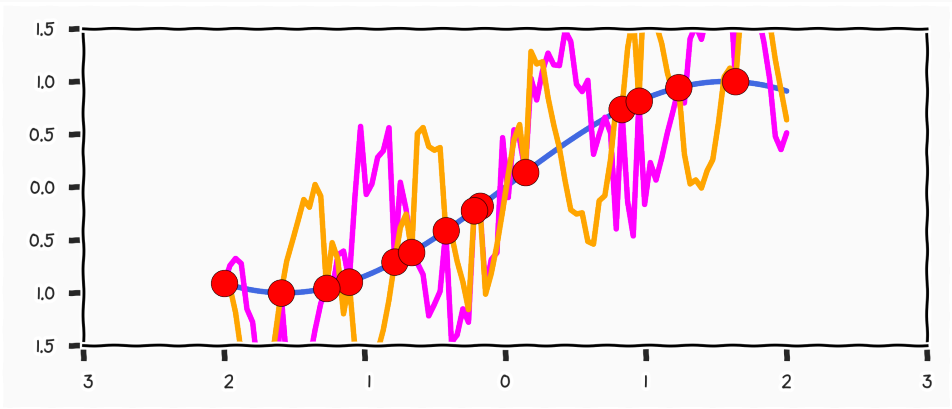
Flexible such that we do not have to make trade-offs when including beliefs

Narrow such that we can reduce data-requirements

Interpretable so that we can translate our knowledge to the parametrisation

Non-parametrics

Curve Fitting



Rule #47 - Drink Trippels, don't ride trippels

"Cycling and beer are so intertwined we may never understand the full relationship. Beer is a recovery drink, an elixir for post-ride trash talking and a just plain excellent thing to pour down the neck. We train to drink so don't fool around. Drink quality beer from real breweries. If it is brewed with rice instead of malted barley or requires a lime, you are off the path. Know your bittering units like you know your gear length. Life is short, don't waste it on piss beer."

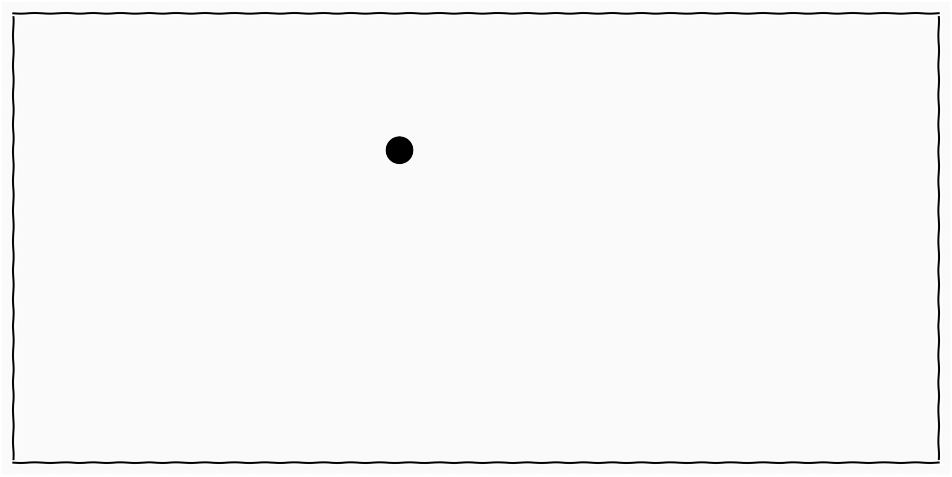
– https://www.velominati.com/

Non-parametrics

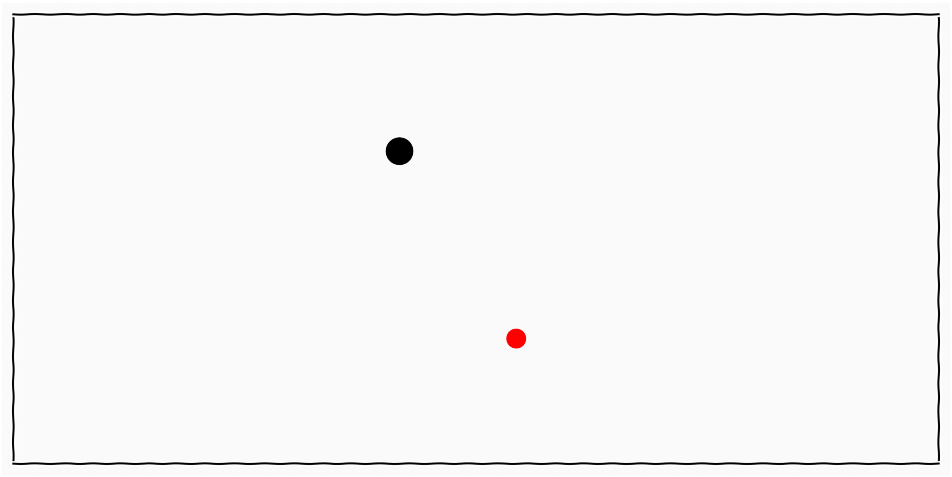
Non-parametrics

Non-parametrics

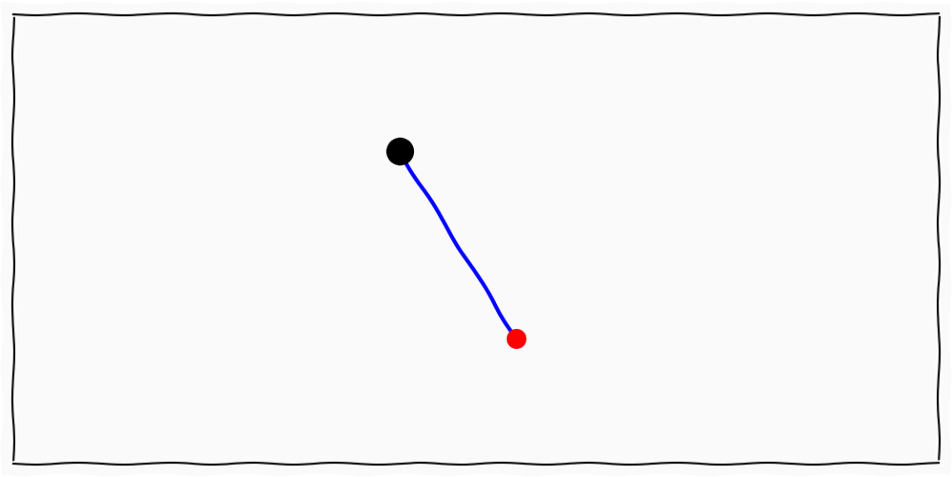
Example



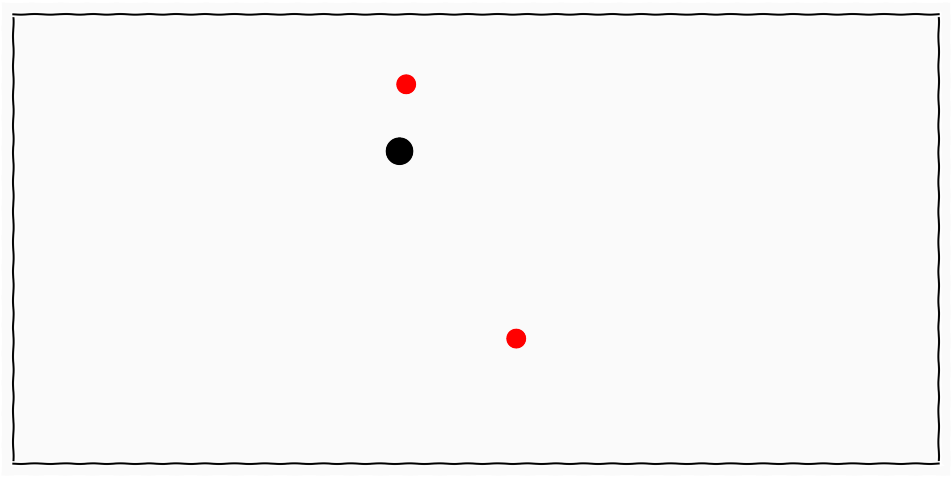
Example



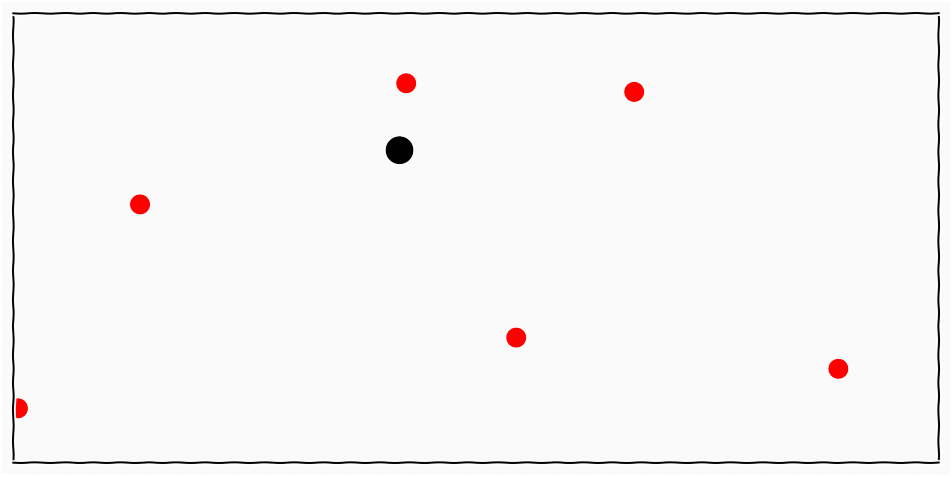
Example



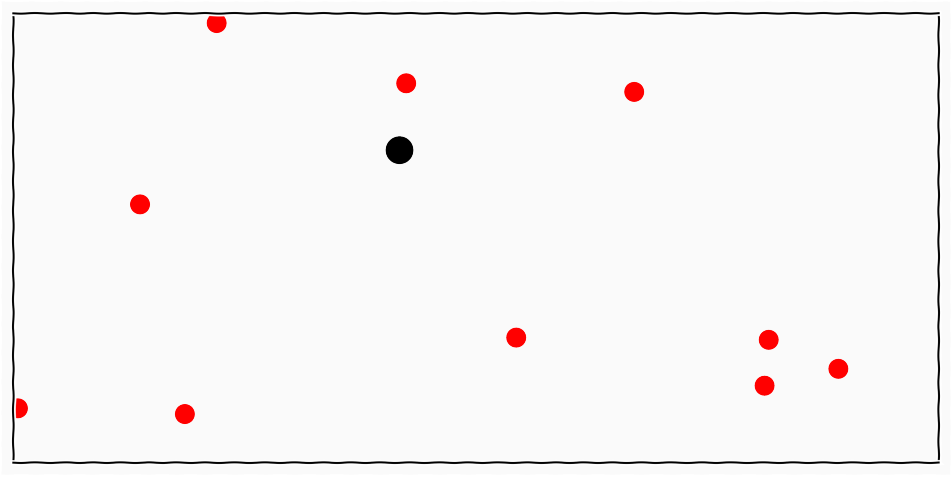
Example



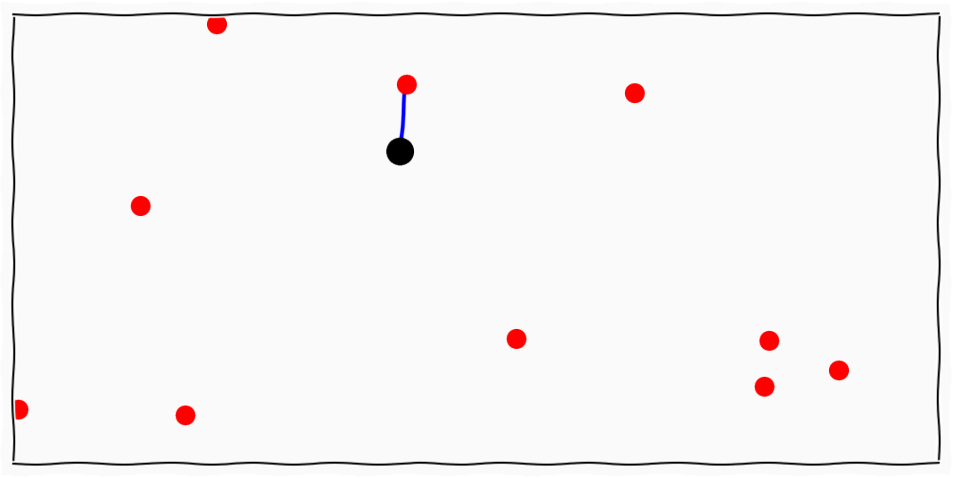
Example



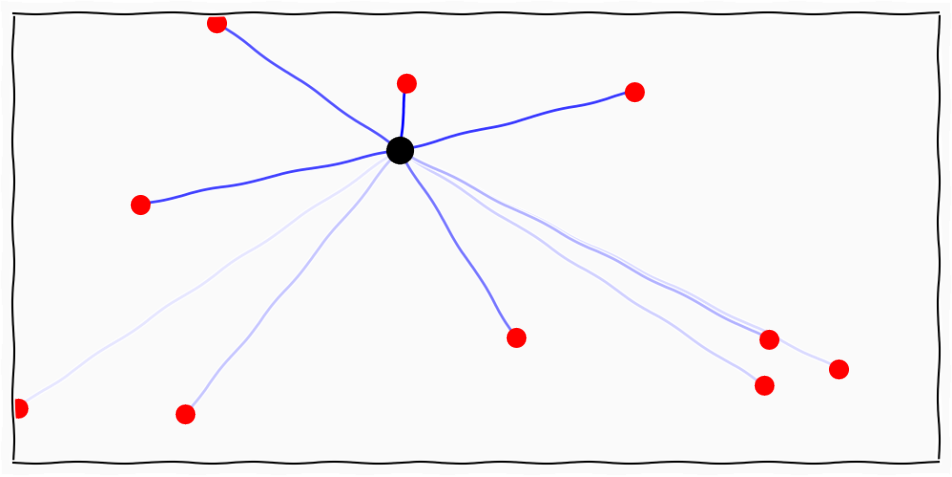
Example



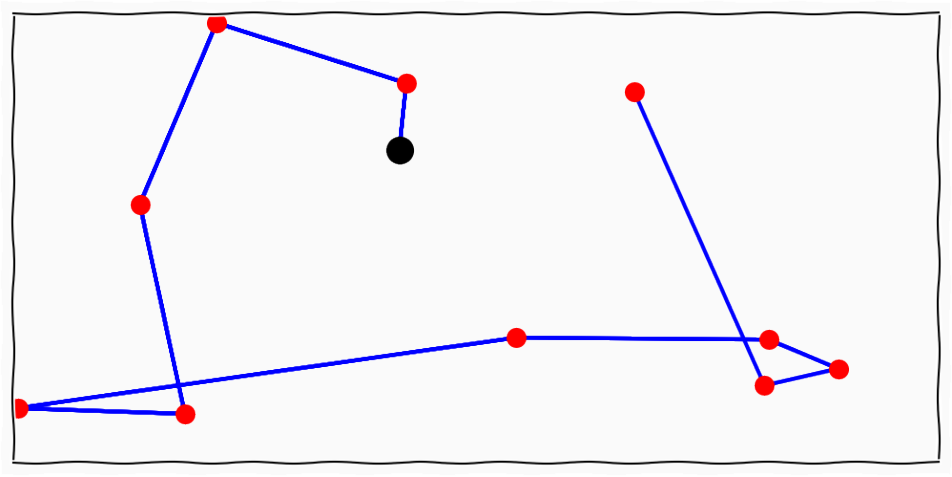
Example



Example

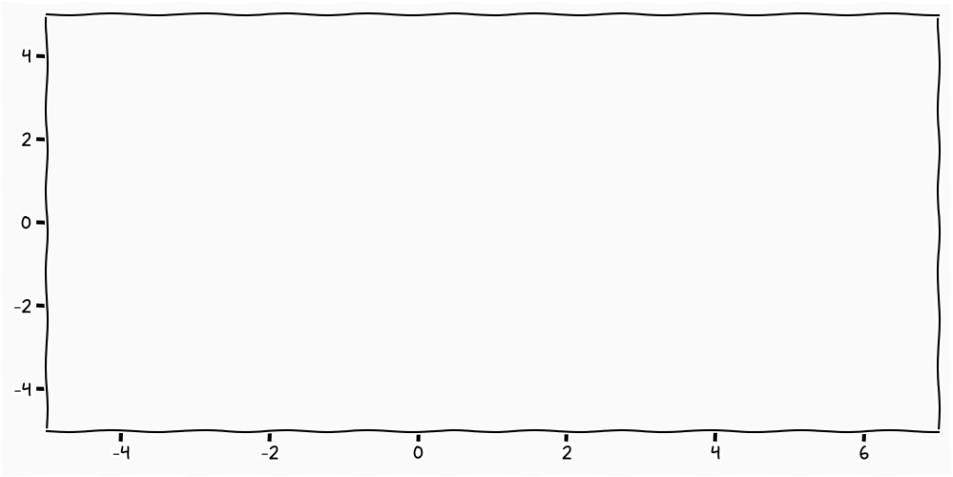


Example

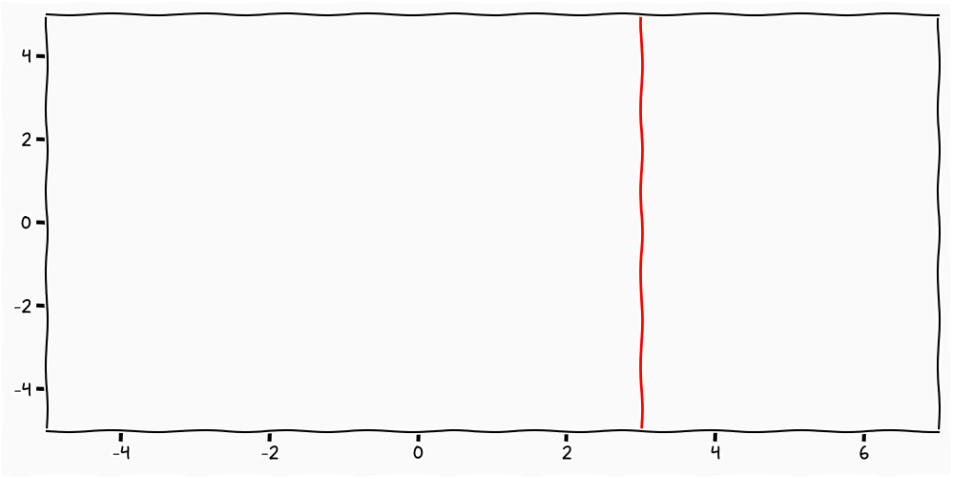


Non-parametric Models

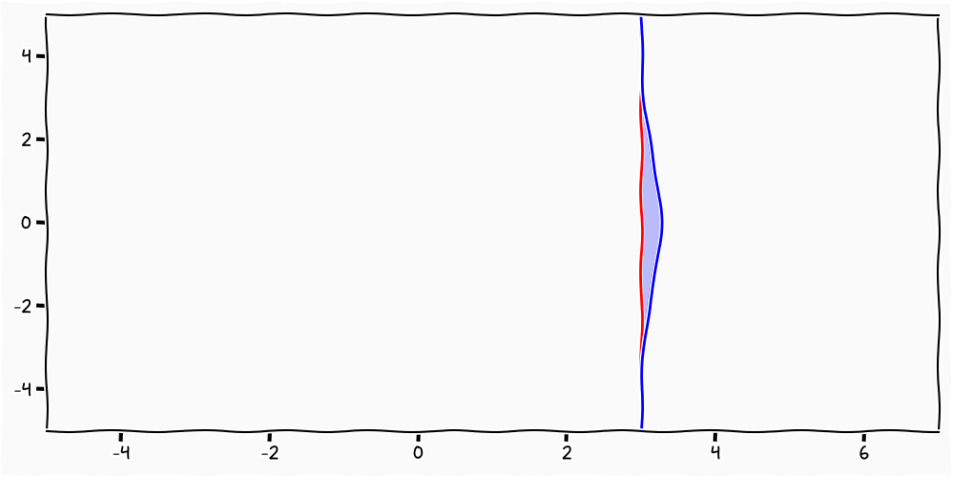
Lets talk about functions



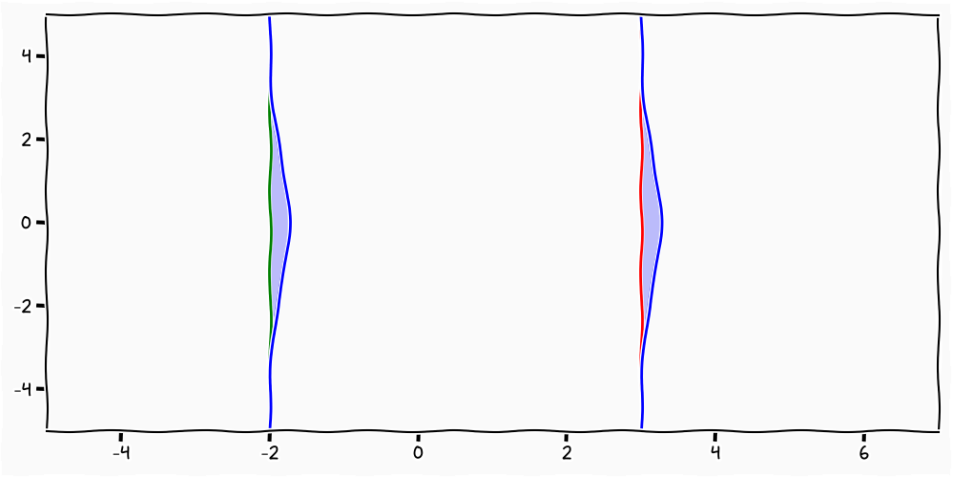
Lets talk about functions



Lets talk about functions



Lets talk about functions

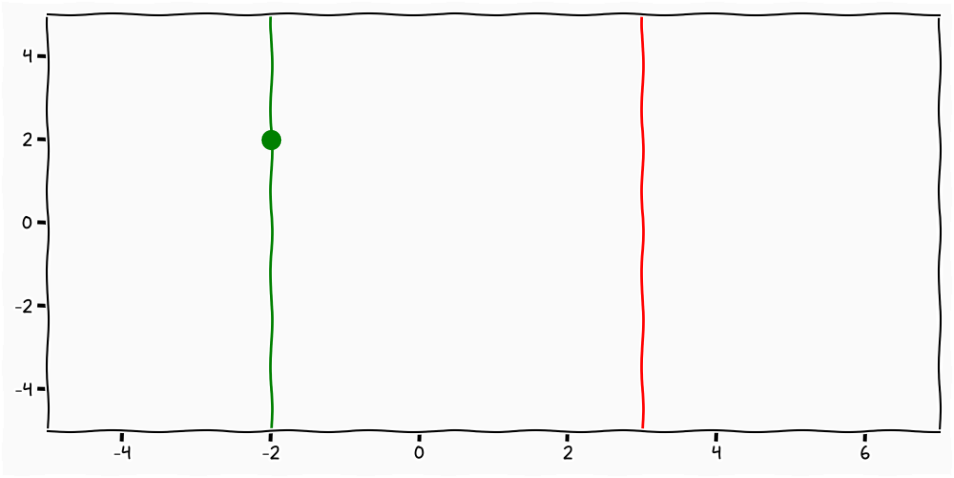


Gaussian function values

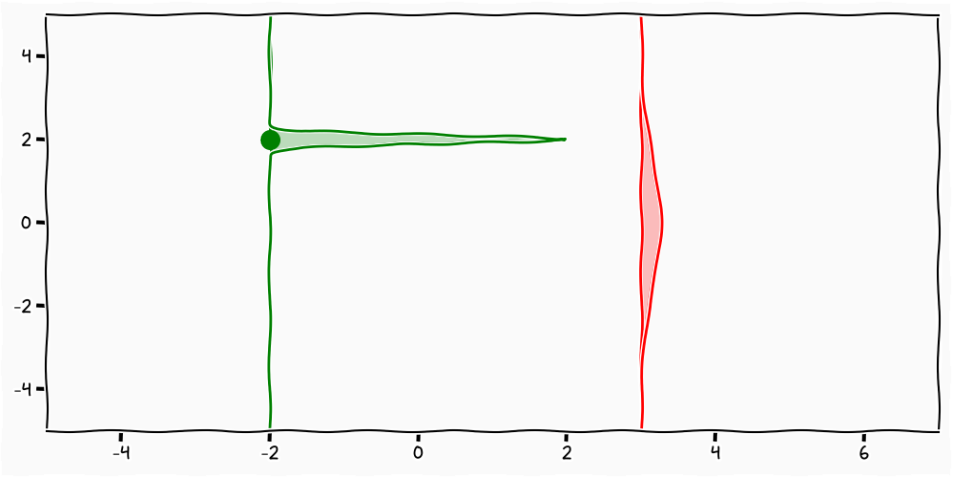
$$f_1 = \mathcal{N}(\mu_1, k_1)$$

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

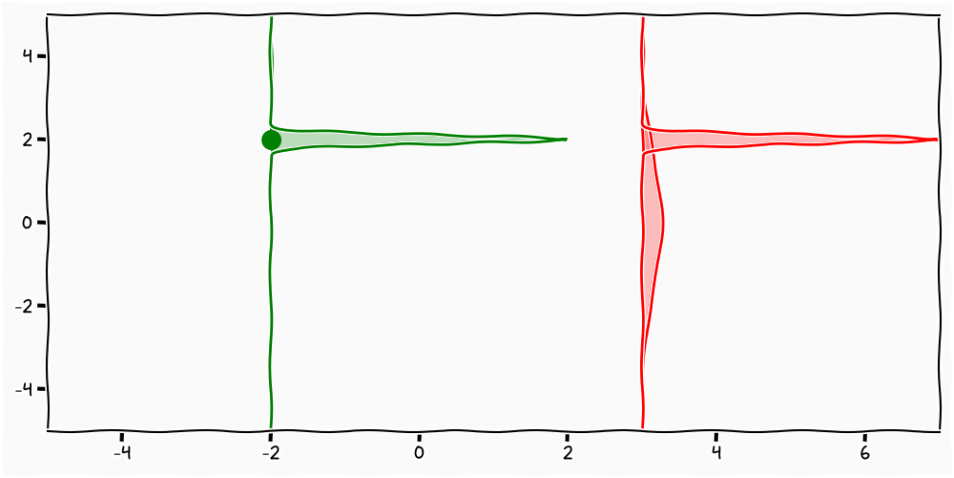
Non-parametric functions



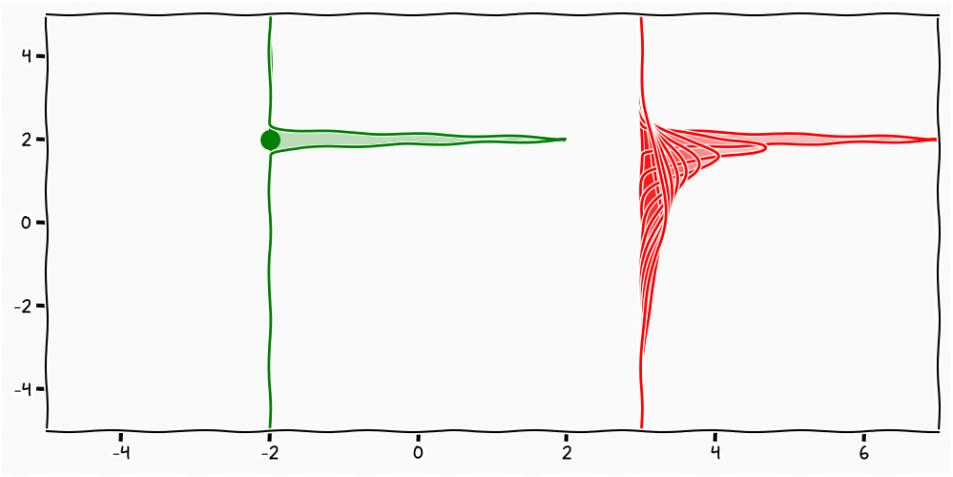
Non-parametric functions



Non-parametric functions



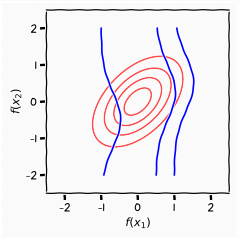
Non-parametric functions

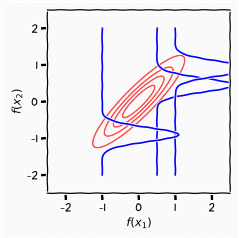


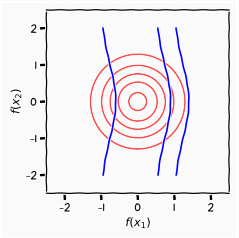
Jointly Gaussian function values

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

Conditional Gaussians





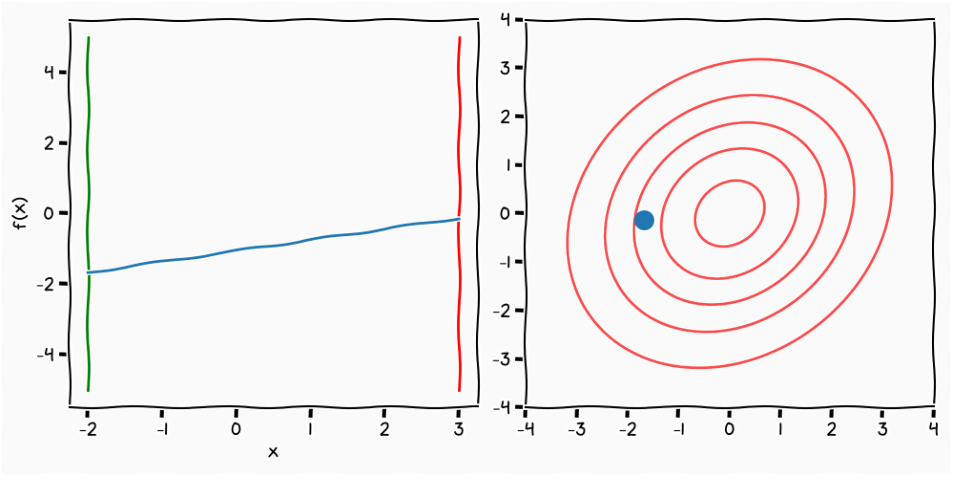


$$\mathsf{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0.5\\0.5 & 1\end{bmatrix}\right)$$

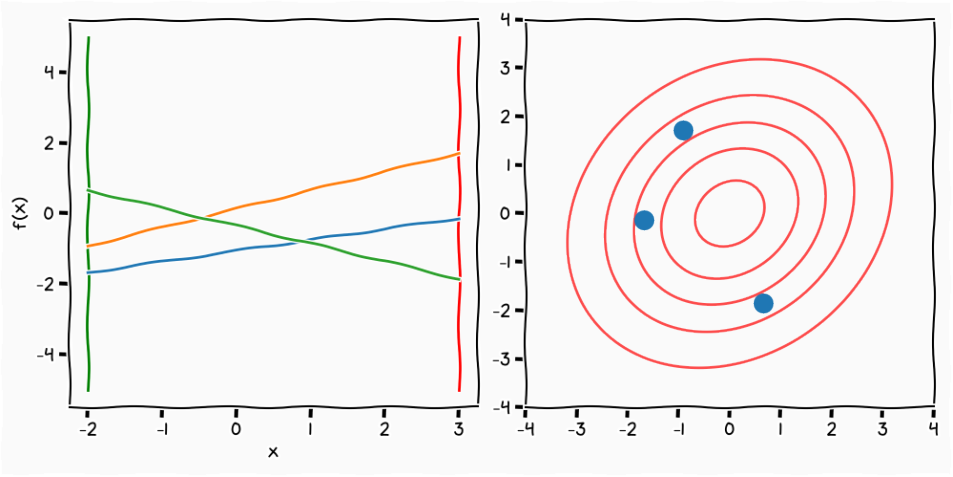
$$\mathsf{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0.9\\0.9 & 1\end{bmatrix}\right)$$

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

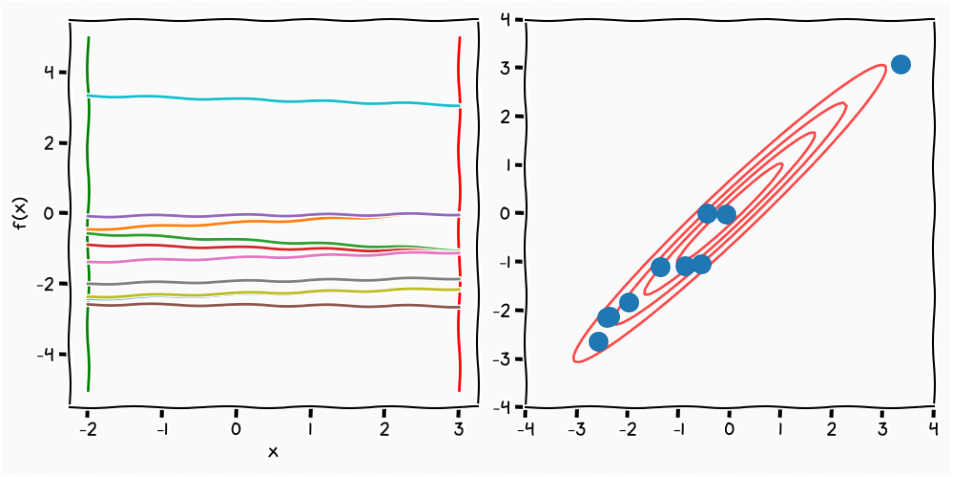
Gaussian Samples



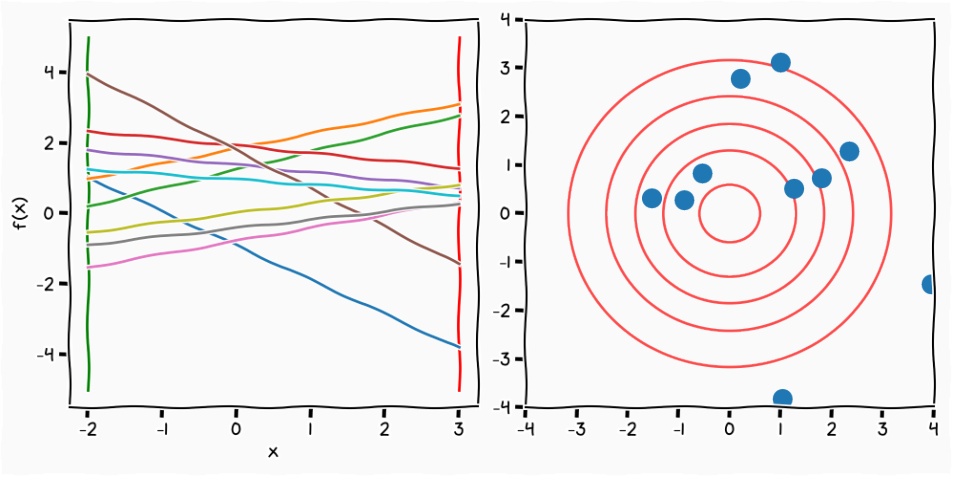
Gaussian Samples



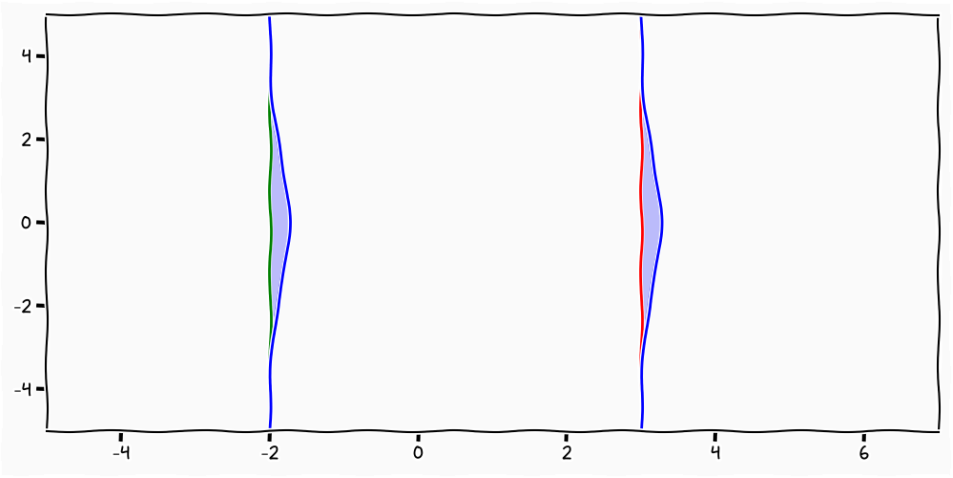
Gaussian Samples



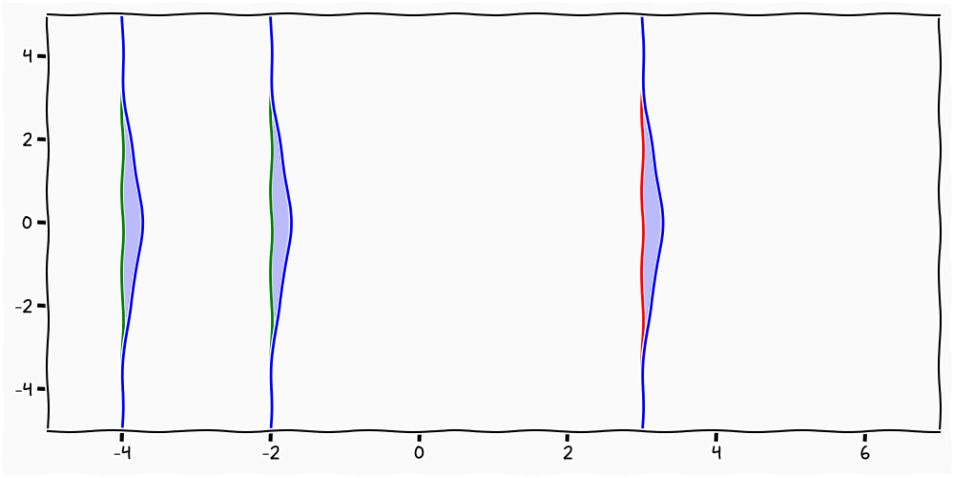
Gaussian Samples



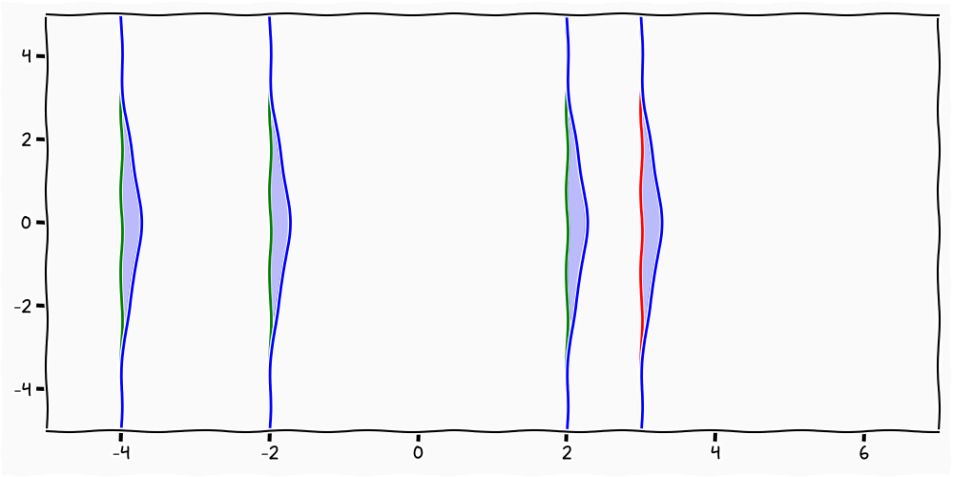
Lets talk about functions



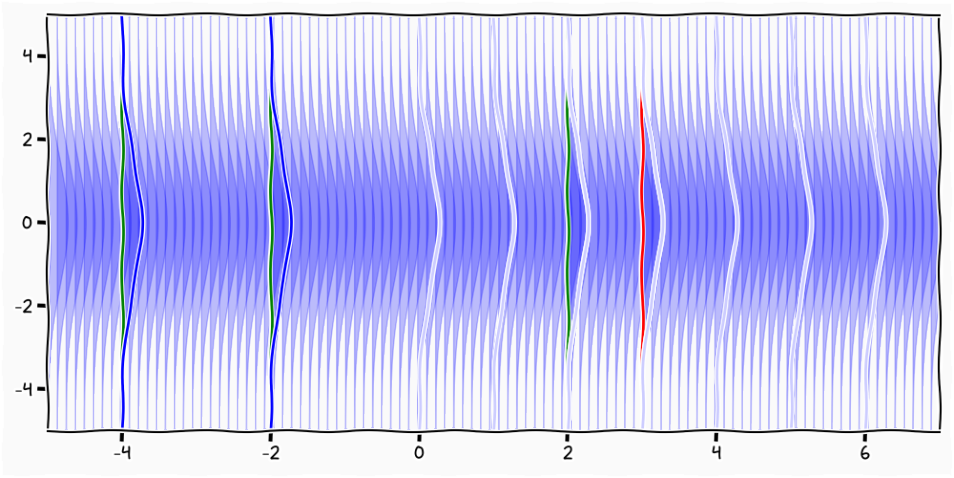
Non-parametric functions



Non-parametric functions



Non-parametric functions



Jointly Gaussian functions II

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

Gaussian Distribution - Marginal

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

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

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

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

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

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

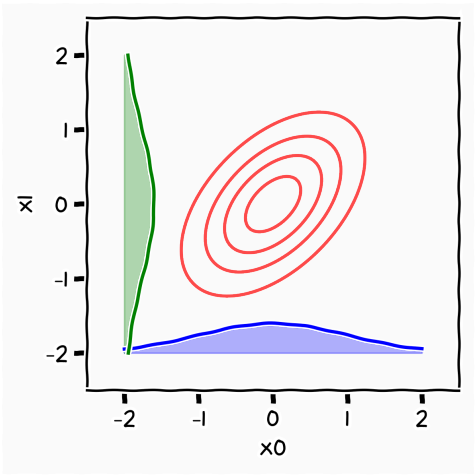
$$p(x_1, x_2) = \mathcal{N}\left(\begin{array}{c|cc} x_1 & \mu_1 & k_{11} \quad k_{12} \\ x_2 & \mu_2 & k_{21} \quad k_{22} \end{array}\right)$$

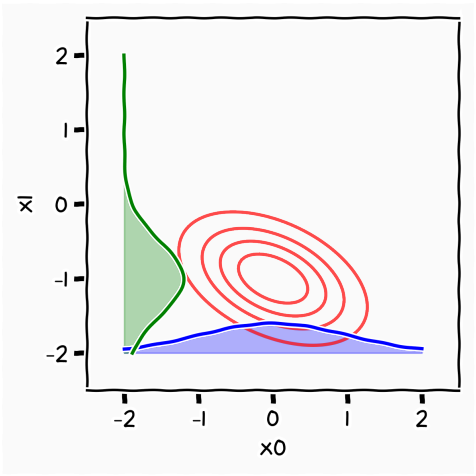
$$\Rightarrow p(x_1) = \int_{x_2} p(x_1, x_2) = \underline{\mathcal{N}\left(x_1 \mid \mu_1, k_{11}\right)}$$

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

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

Gaussian Distribution - Marginal



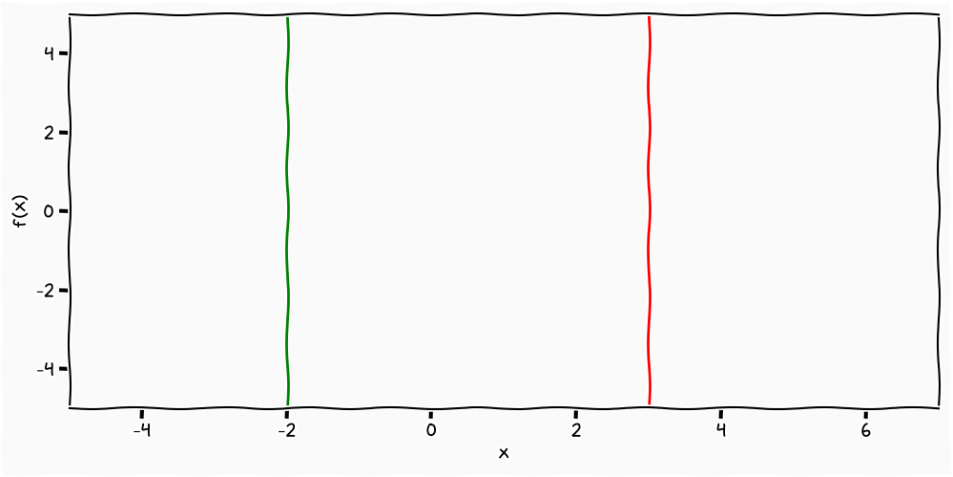


Marginal Property (Consistency)

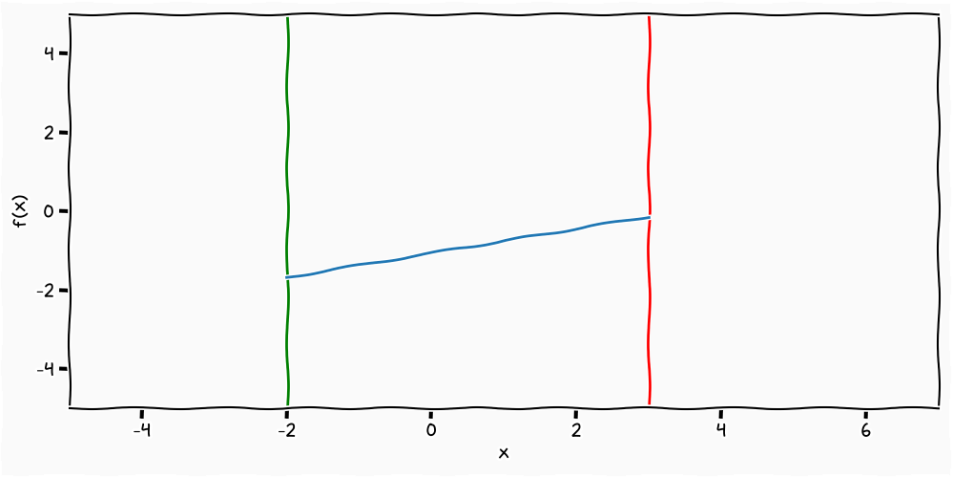
For all measurable sets $F_i \subseteq \mathbb{R}^n$ and probability measure $\mathcal{N}$

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

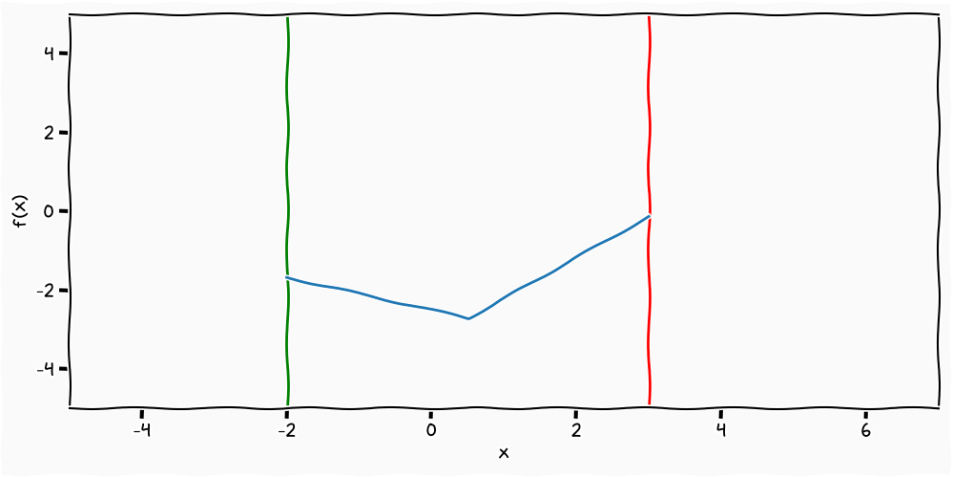
Gaussian Samples



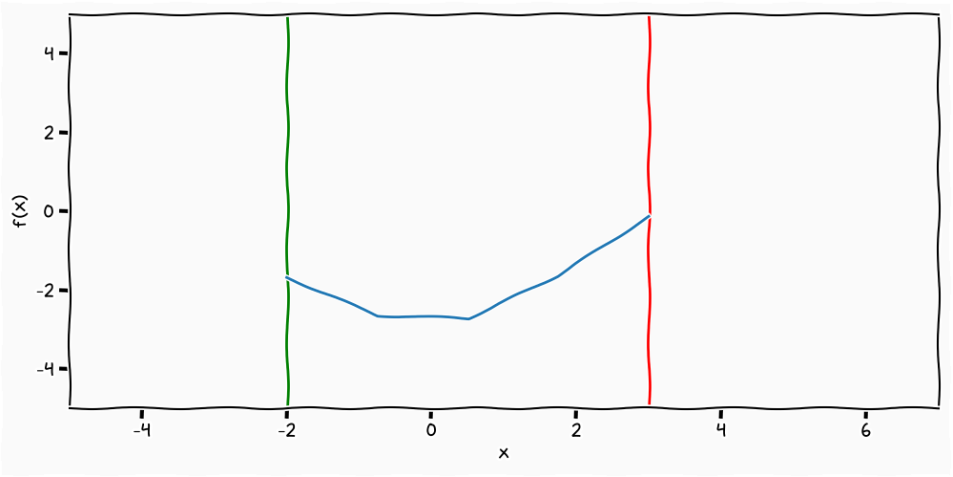
Gaussian Samples



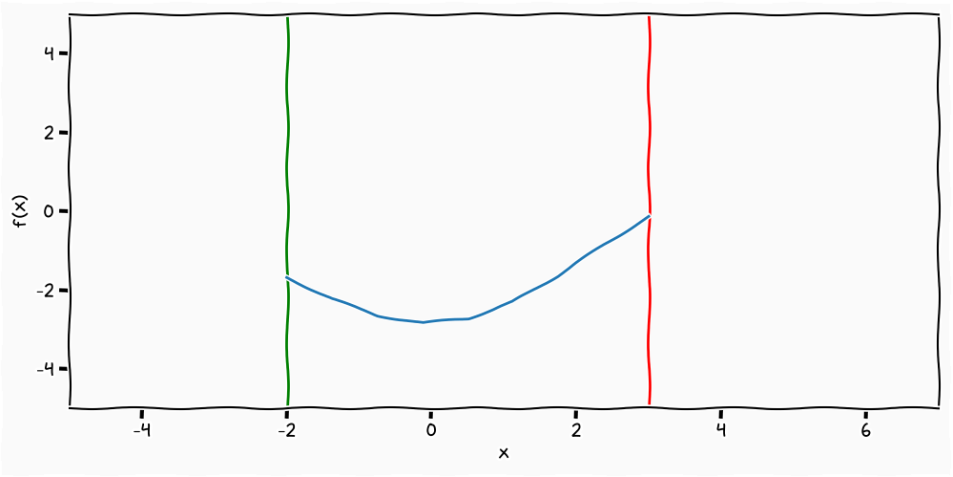
Gaussian Samples



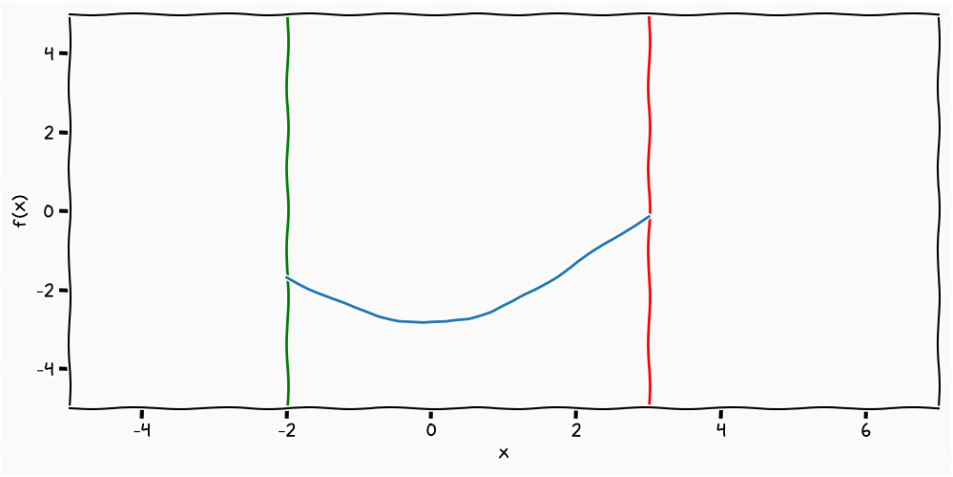
Gaussian Samples



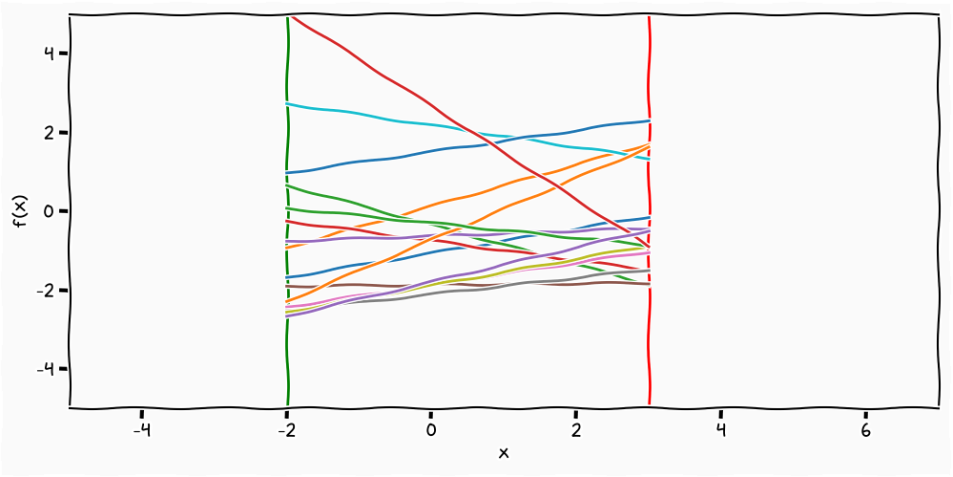
Gaussian Samples



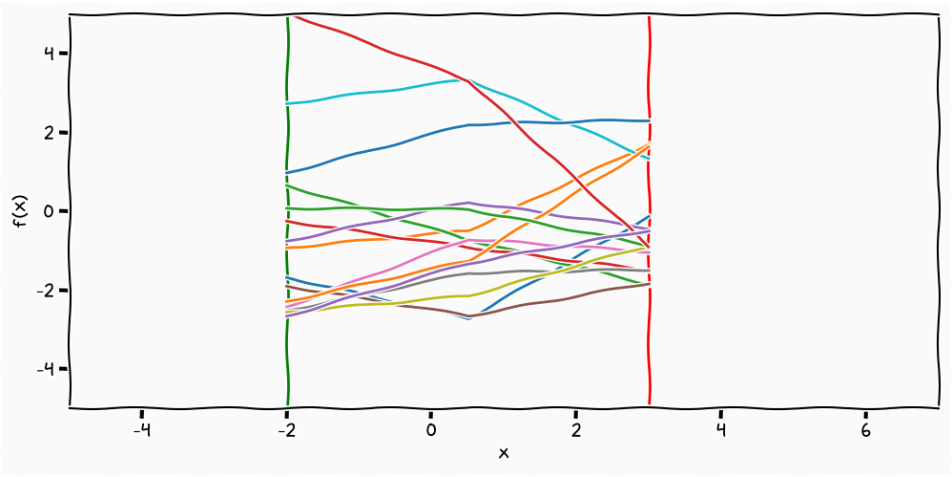
Gaussian Samples



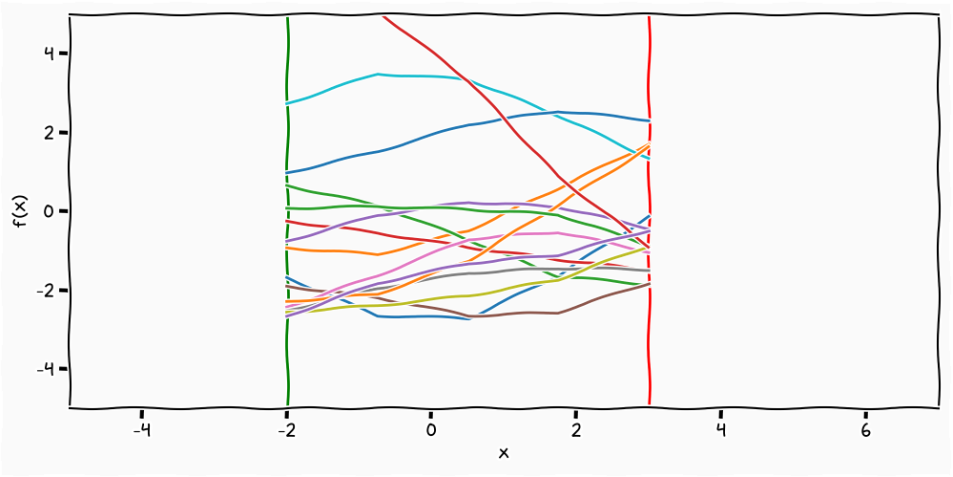
Gaussian Samples



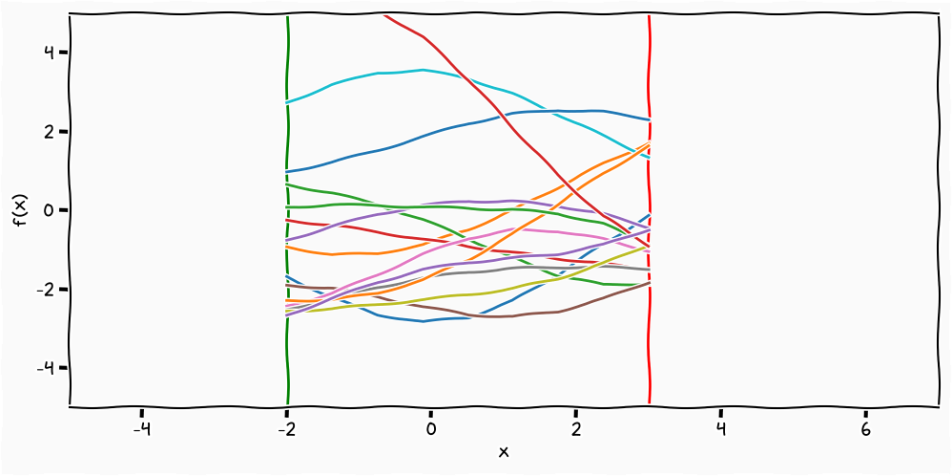
Gaussian Samples



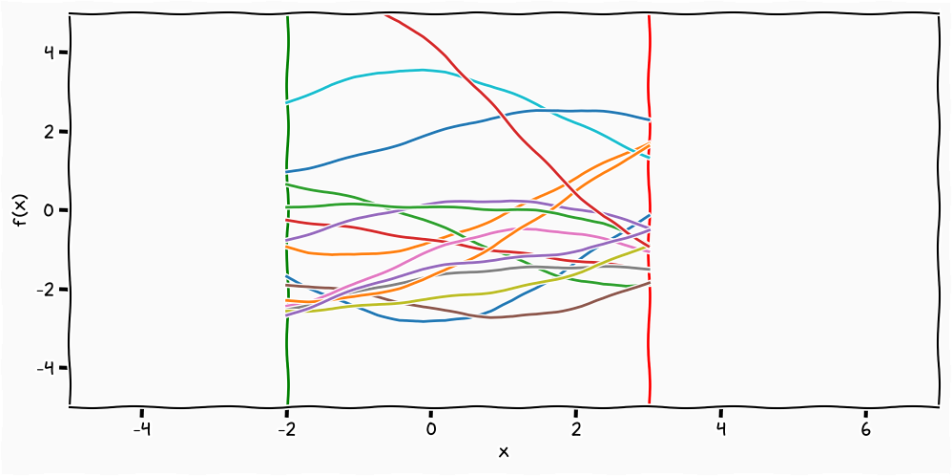
Gaussian Samples



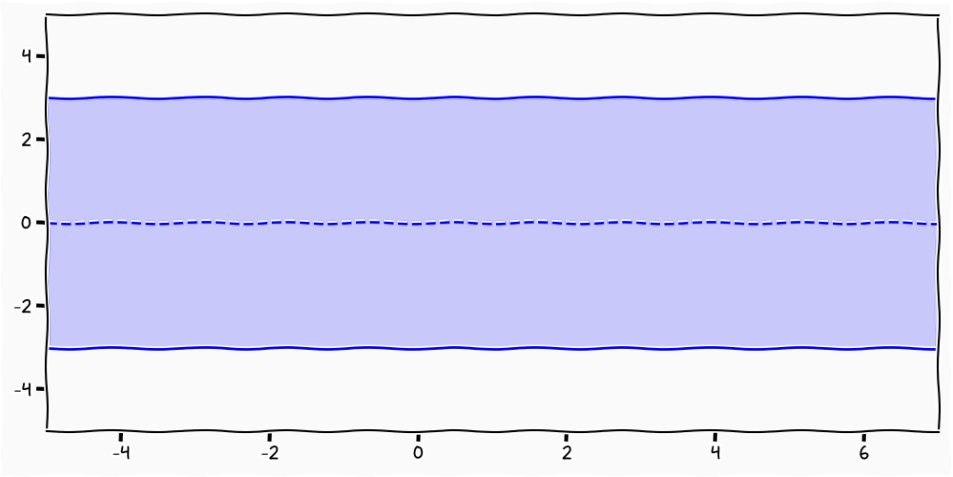
Gaussian Samples



Gaussian Samples



Gaussian Processes



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

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

$$\rightarrow$$

$$\infty \qquad\qquad N$$

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

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

"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"

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

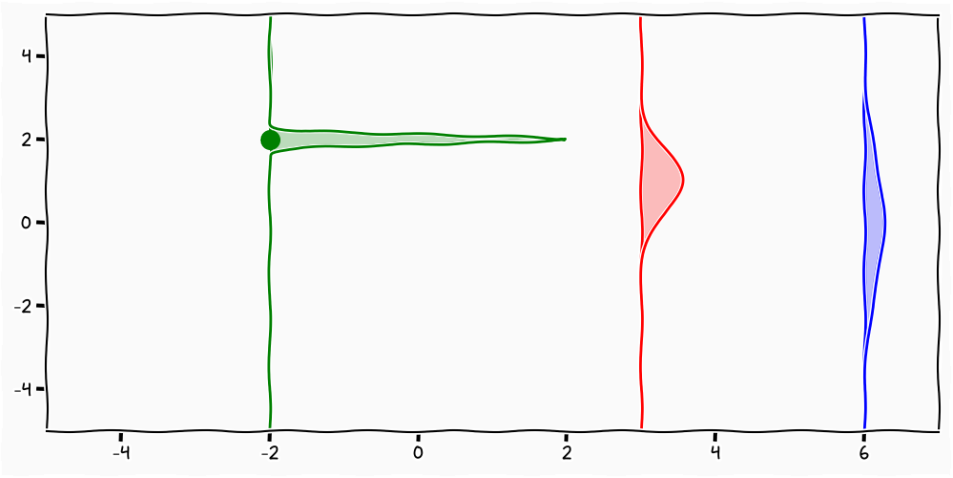
- We parameterise the covariance as a function of the input
- the index set of the measure is the uncountable infinity
- Your "handle" to input your knowledge into a GP is the covariance function
  - *you specify the degree of covariance between data-points*

"Parametrisation"

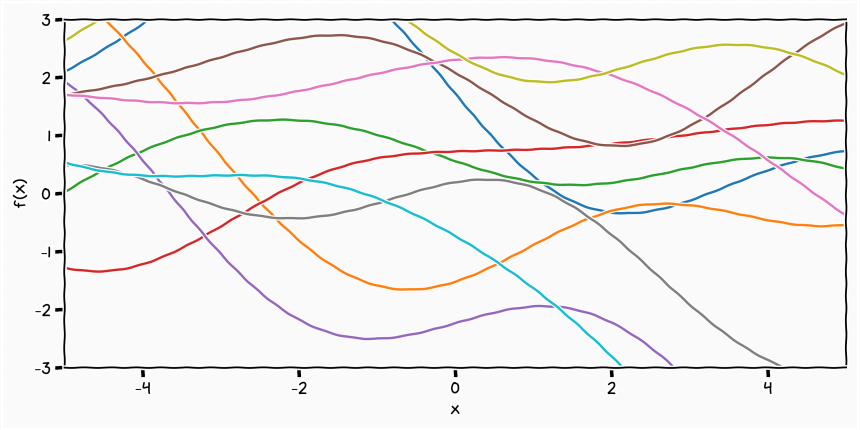
$$k_{ij} = k(x_i, x_j)$$

- 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

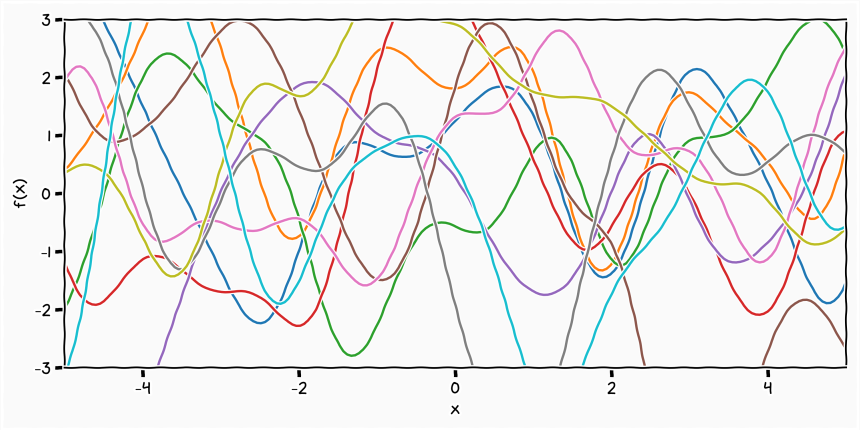


Gaussian Processes Samples



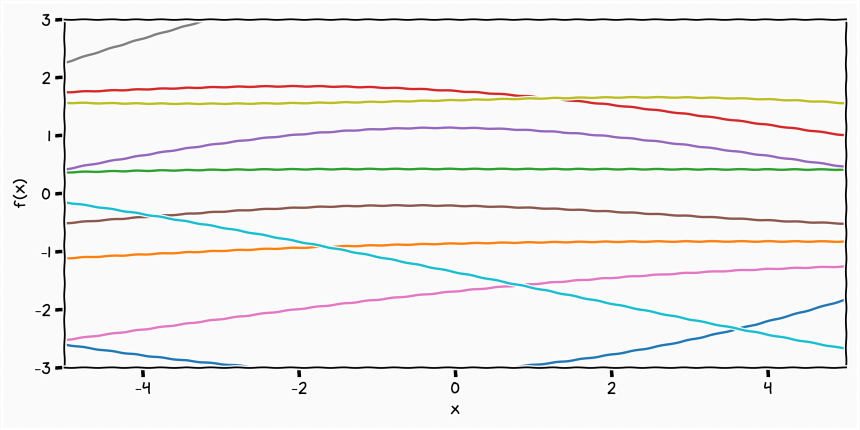
$$k(x_i, x_j) = 3 \cdot e^{-\frac{(x_i - x_j)^2}{15}}$$

Gaussian Processes Samples



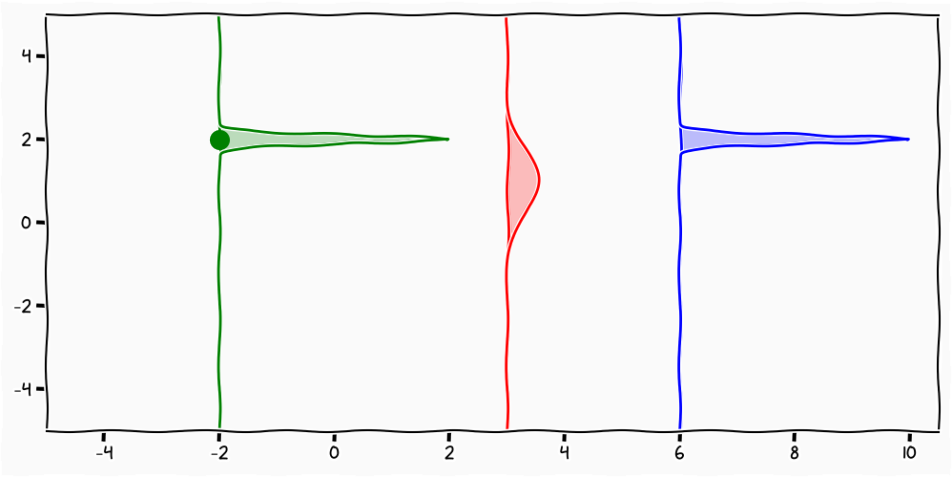
$$k(x_i, x_j) = 3 \cdot e^{-\frac{(x_i - x_j)^2}{1}}$$

Gaussian Processes Samples



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

Gaussian Processes



Inference

Bayes' Rule

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

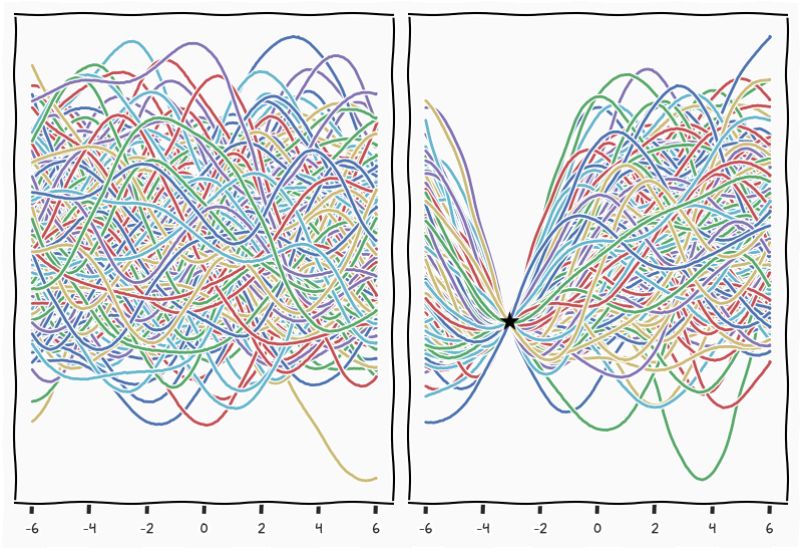
$$\int p(\mathbf{f}, \mathbf{f}_*) \mathrm{d}\mathbf{f}_* = \int p(\mathbf{f} \mid \mathbf{f}_*) p(\mathbf{f}_*) \mathrm{d}\mathbf{f}_*$$

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

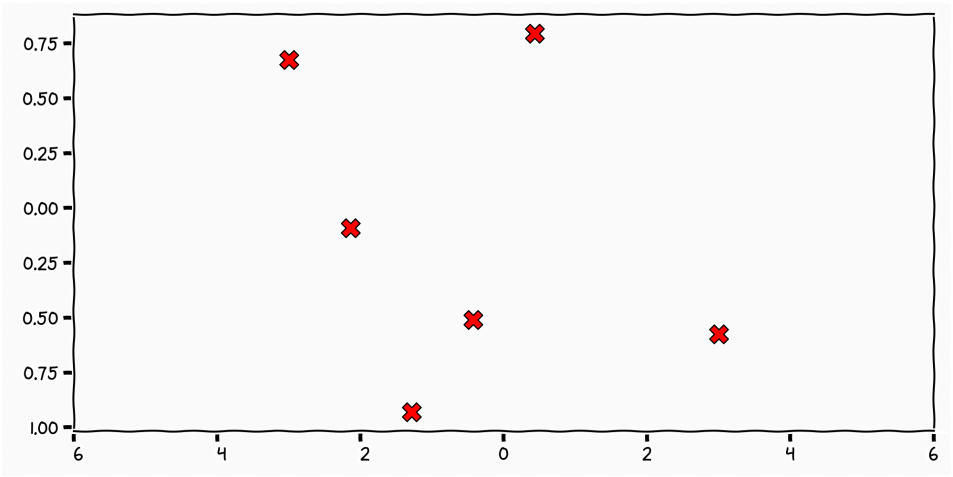
- All instantiations are jointly Gaussian

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

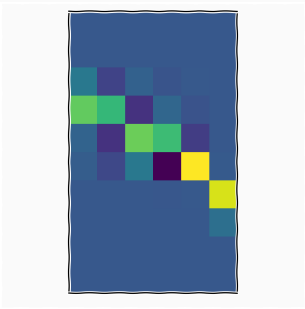
- Conditional Gaussian

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

Intuition

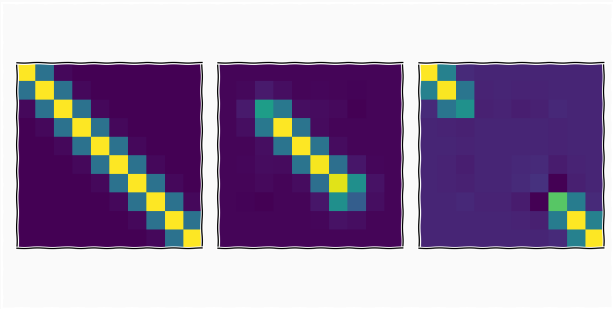


Does it make sense: Mean



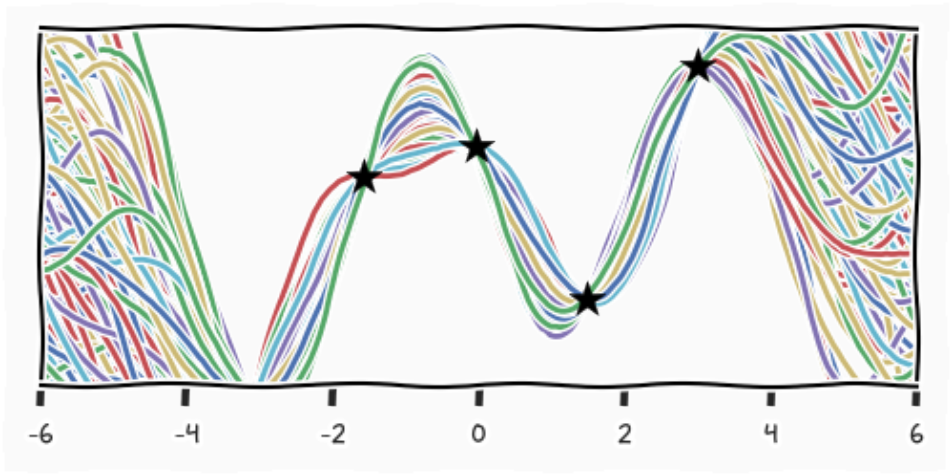
$$k(\mathbf{x}_*, \mathbf{X})^{\mathrm{T}} k(\mathbf{X}, \mathbf{X})^{-1} \mathbf{f}$$

Does it make sense: Covariance

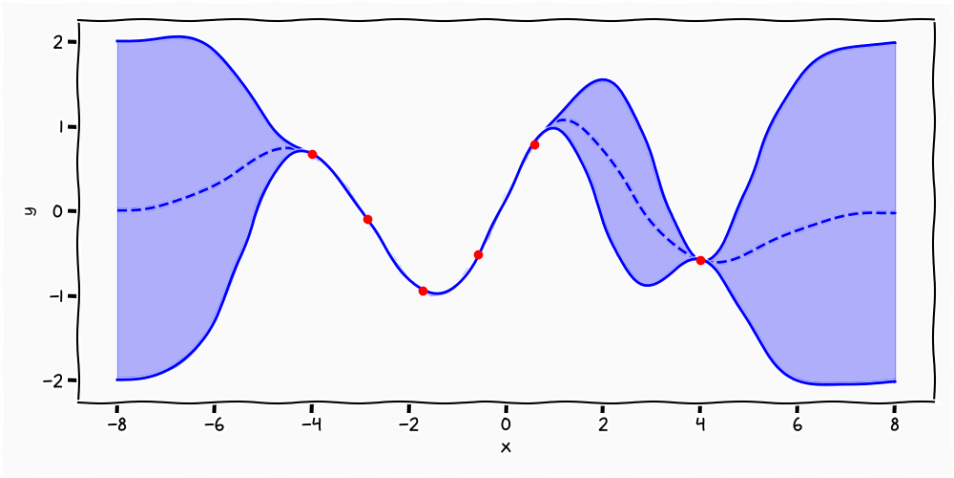


$$k(\mathbf{x}_*, \mathbf{x}_*) - k(\mathbf{x}_*, \mathbf{x})^{\mathrm{T}} k(\mathbf{x}, \mathbf{x})^{-1} k(\mathbf{x}, \mathbf{x}_*)$$

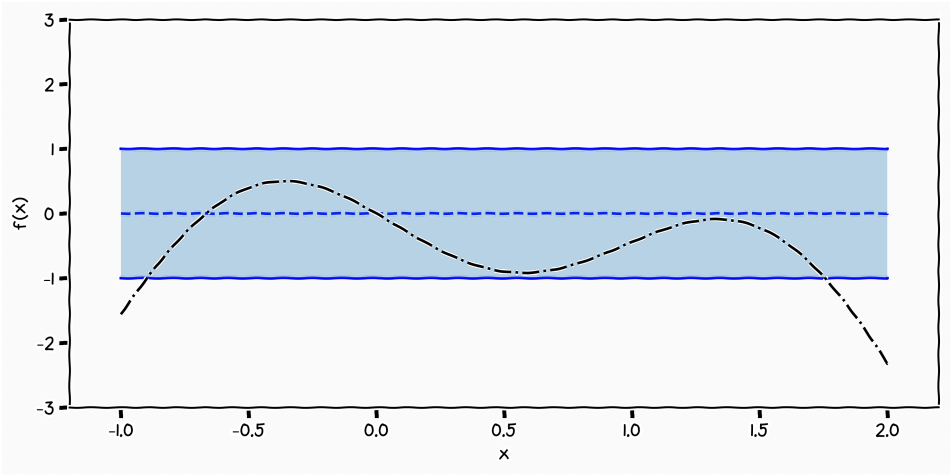
Gaussian Processes: "Predictive Posterior Samples"



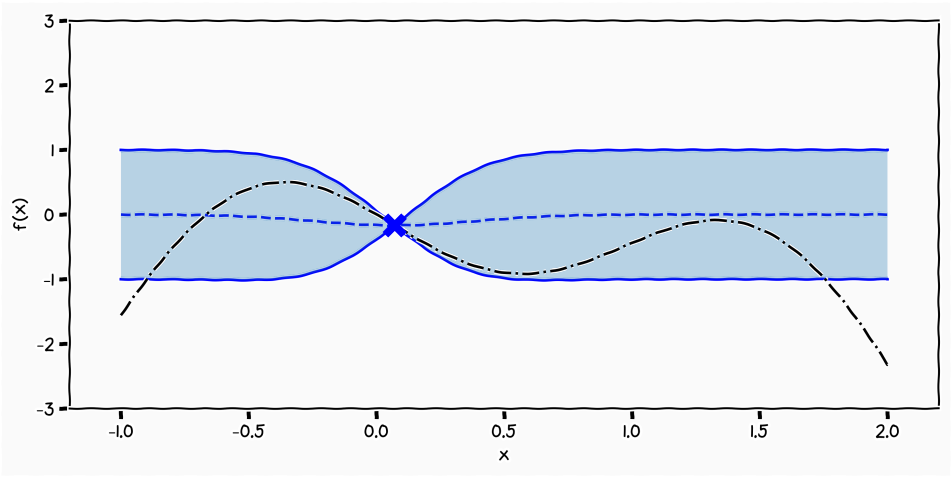
Gaussian Processes: "Predictive Posterior Process"



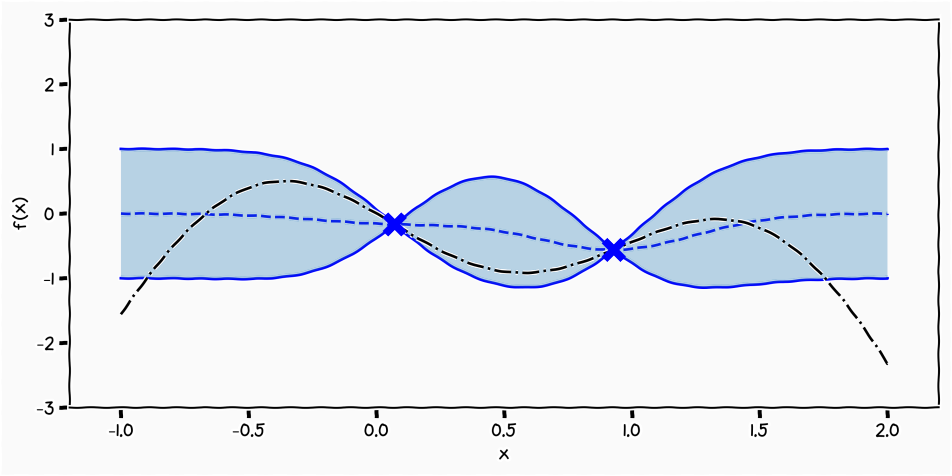
Posterior Processes



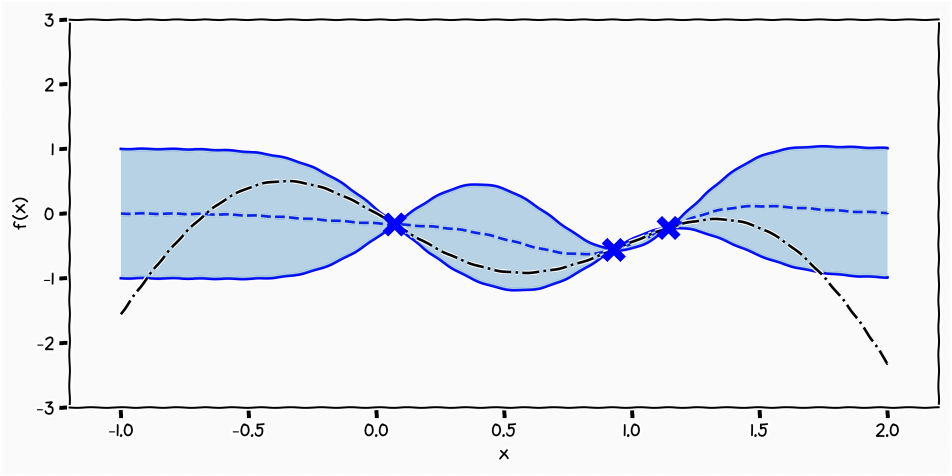
Posterior Processes



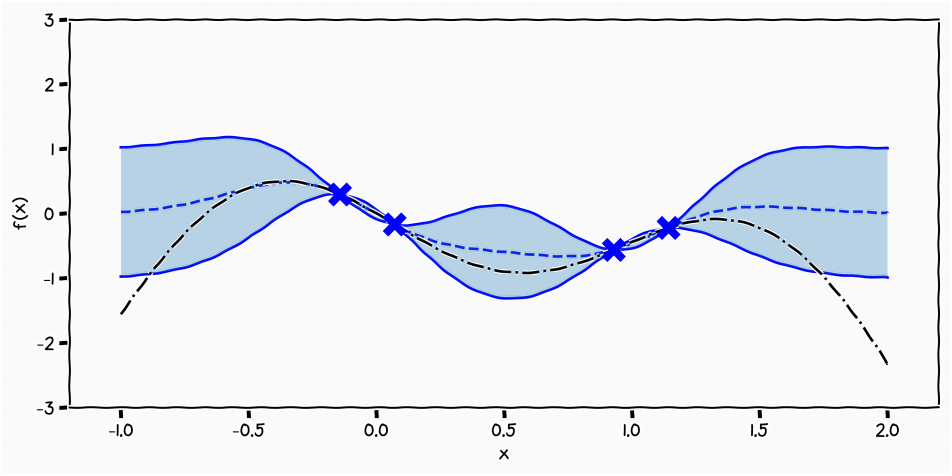
Posterior Processes



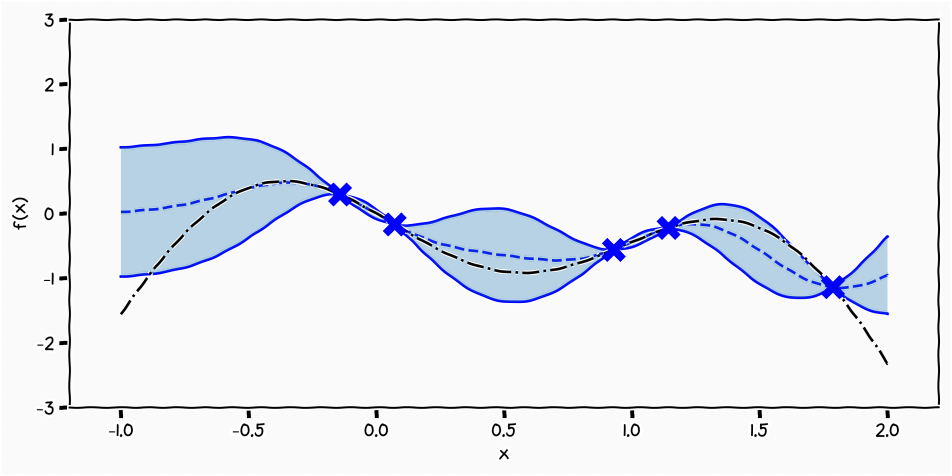
Posterior Processes



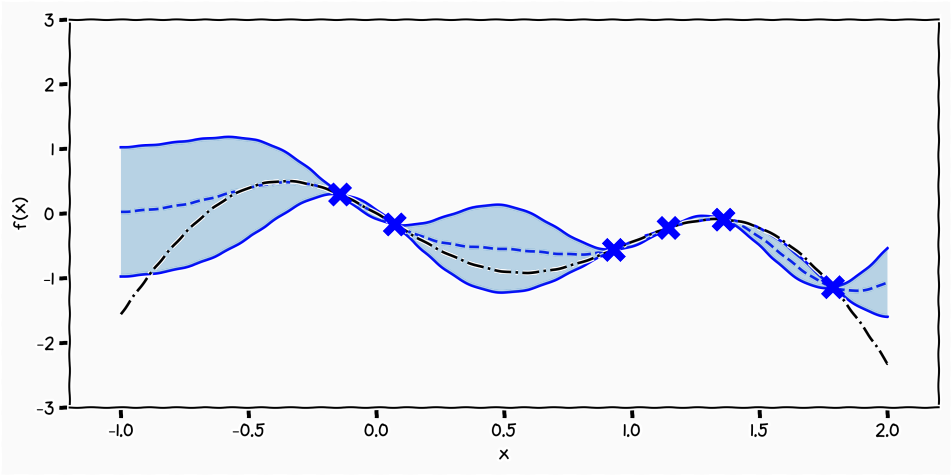
Posterior Processes



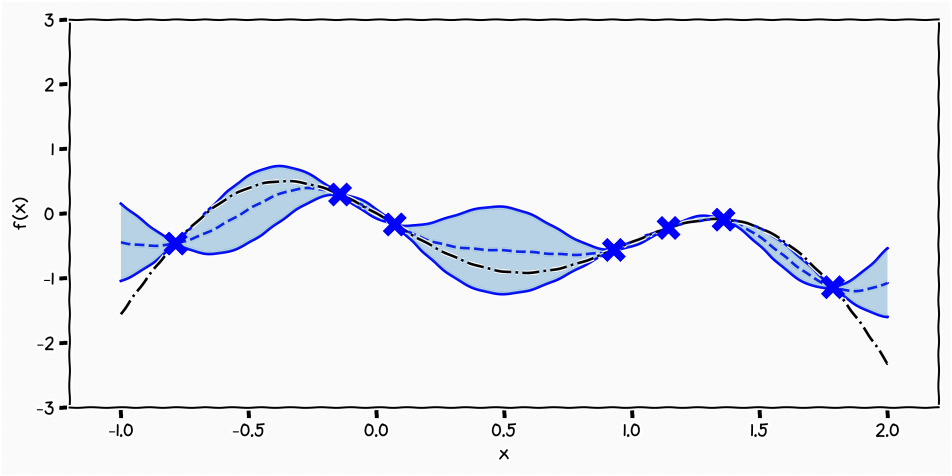
Posterior Processes



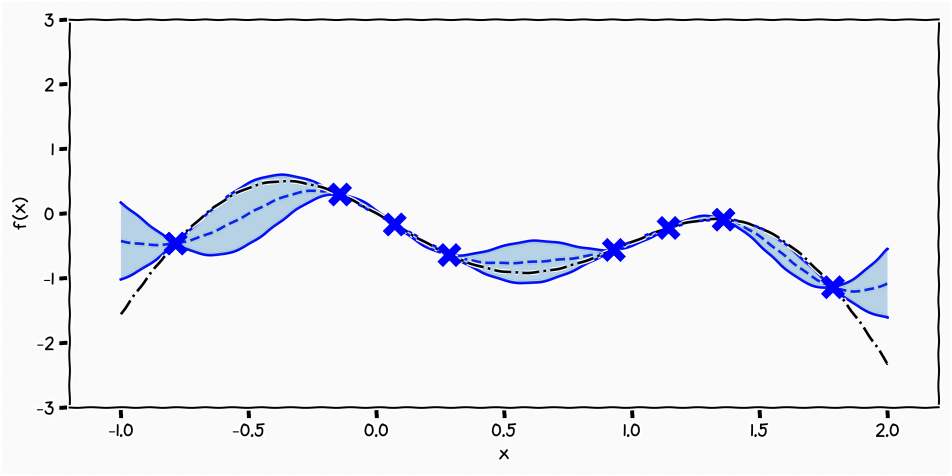
Posterior Processes



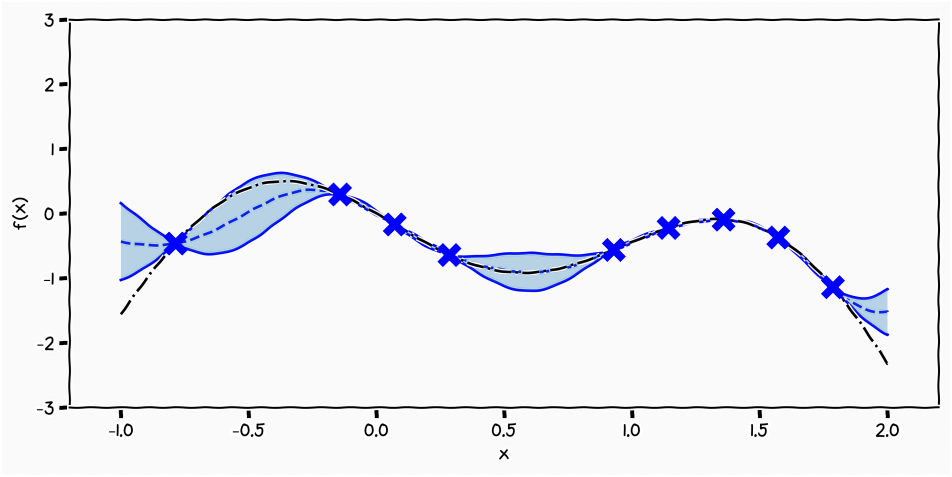
Posterior Processes



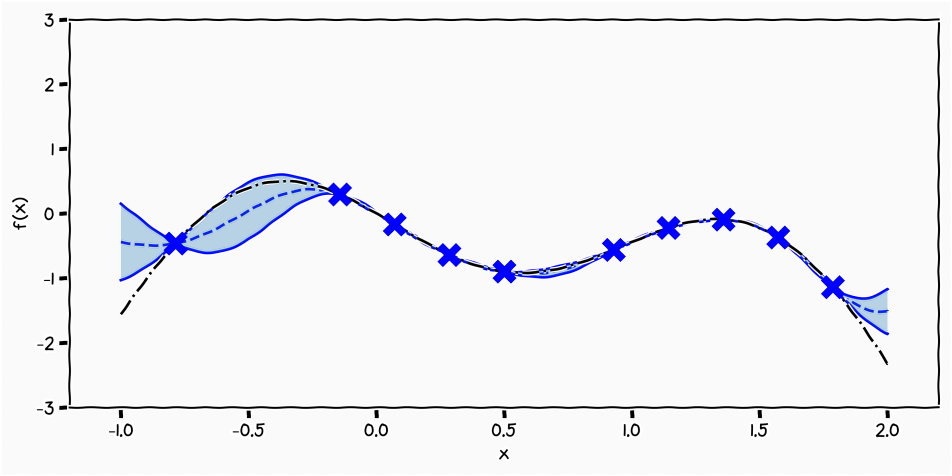
Posterior Processes



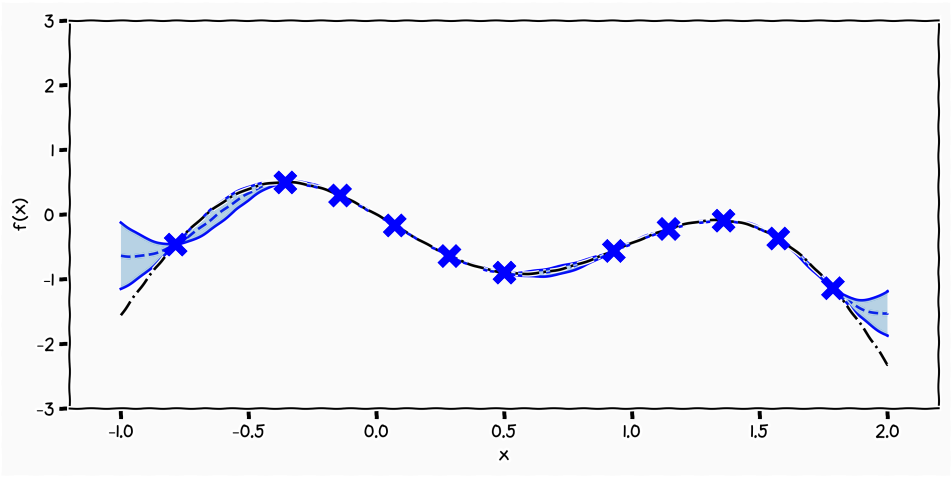
Posterior Processes



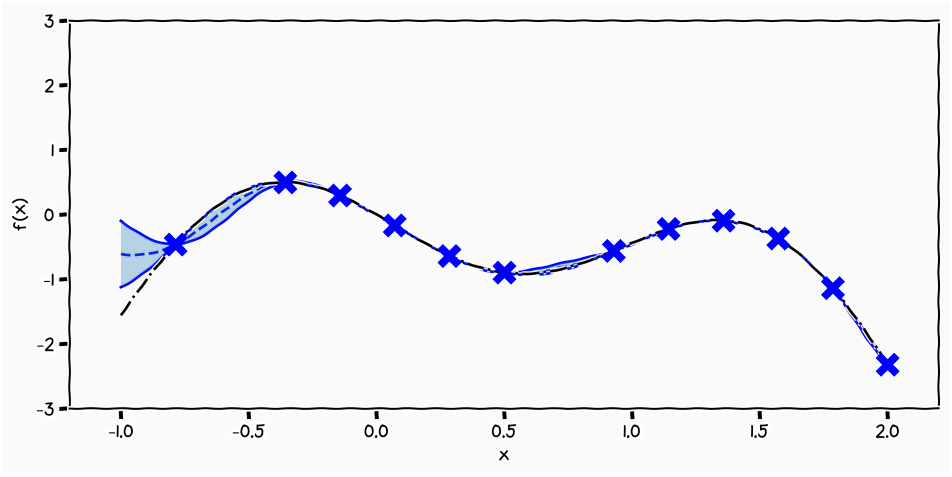
Posterior Processes



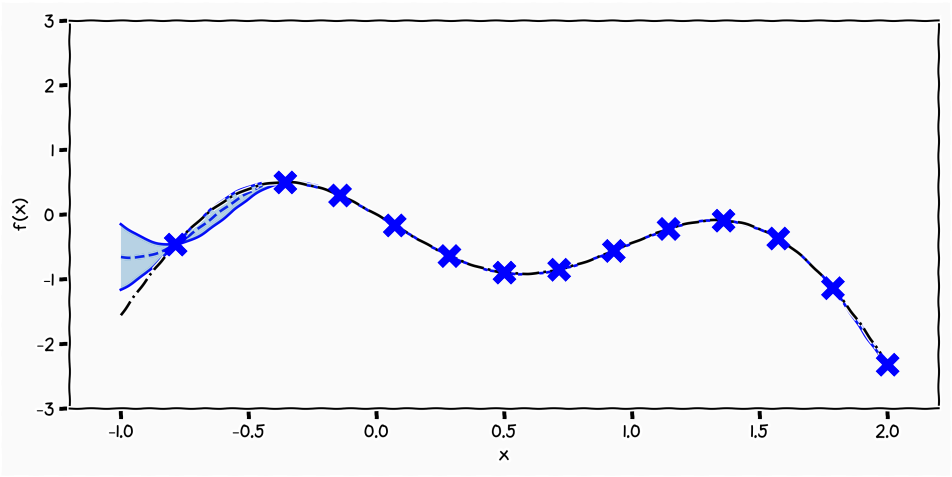
Posterior Processes



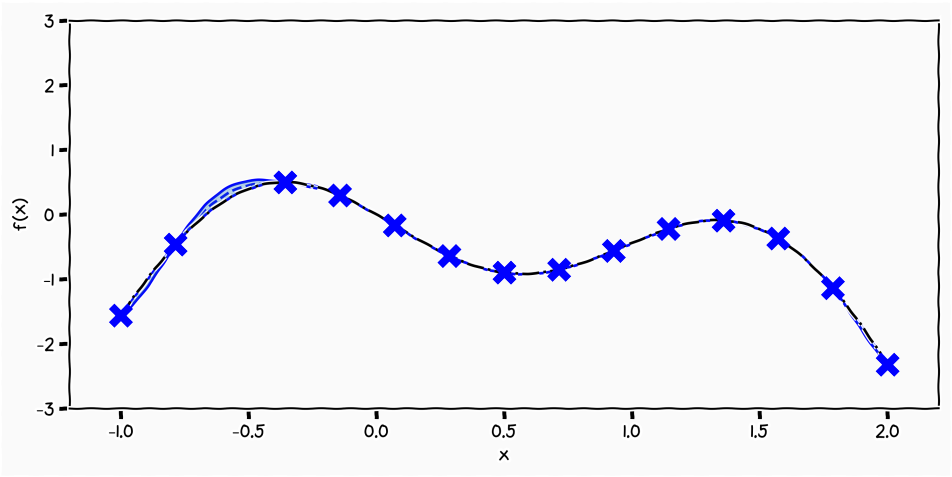
Posterior Processes



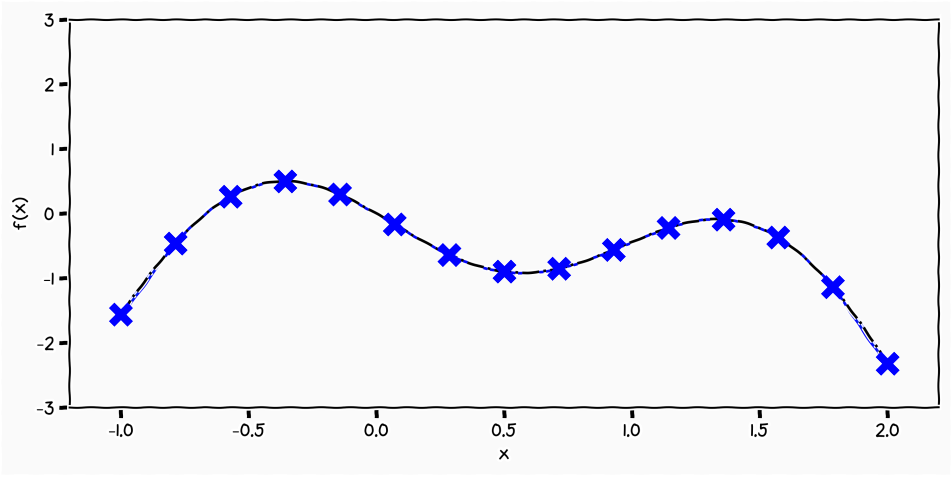
Posterior Processes

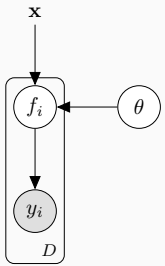


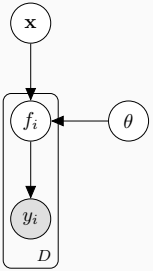
Posterior Processes



Posterior Processes

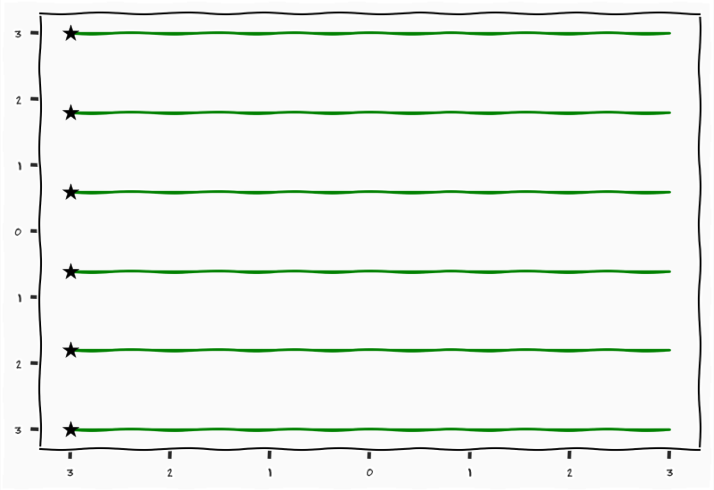


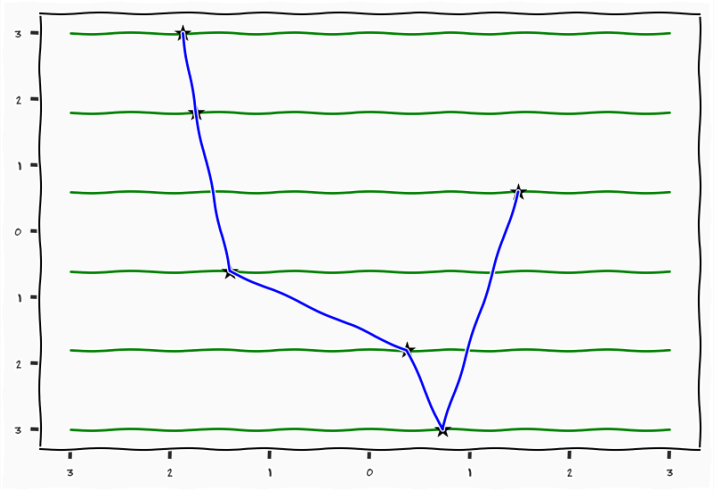




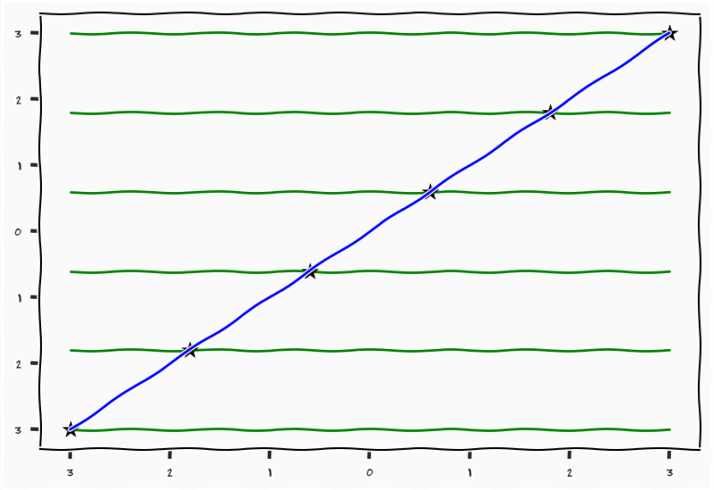
$$p(y|x) = \int p(y \mid f)p(f)\mathrm{d}f$$

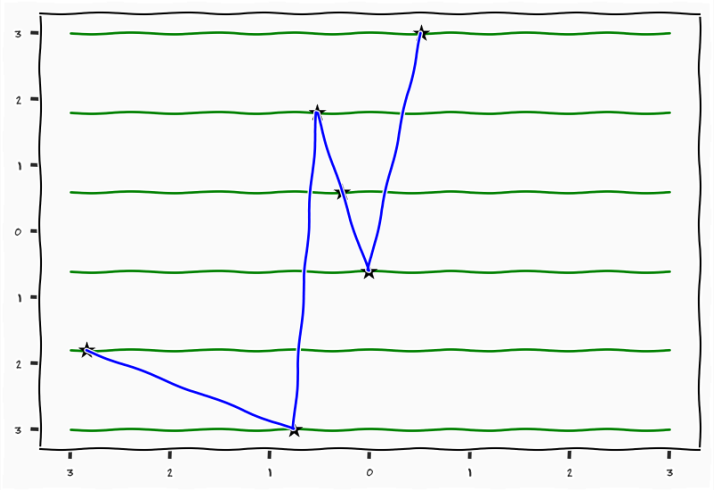
$$p(y) = \int p(y \mid f, x)p(f \mid x)p(x)\mathrm{d}f\mathrm{d}x$$


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

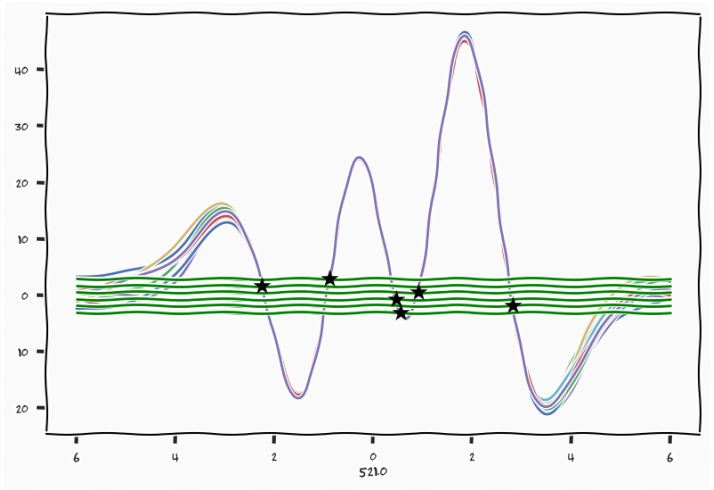

3
2
1
0
1
2
3
3
2
1
0
1
2
3

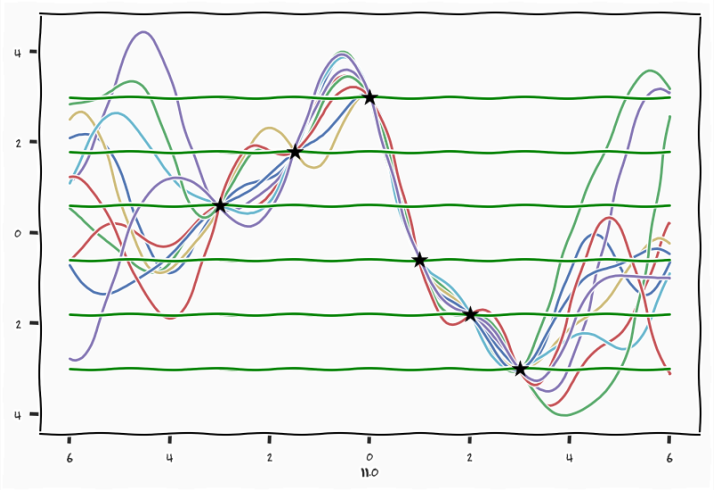
Unsupervised Learning

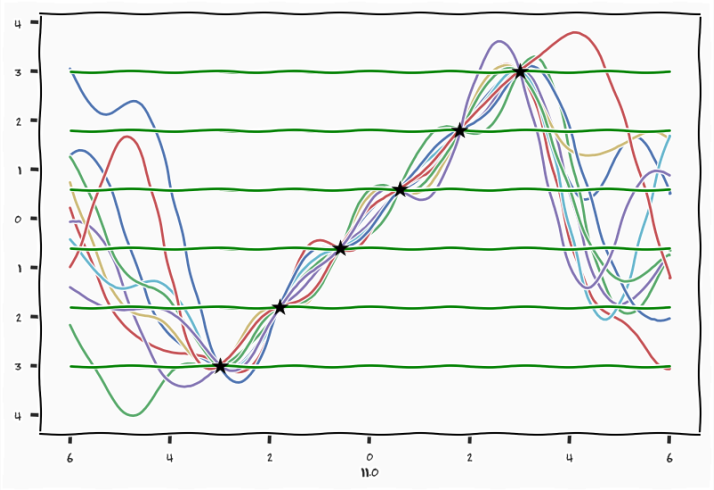


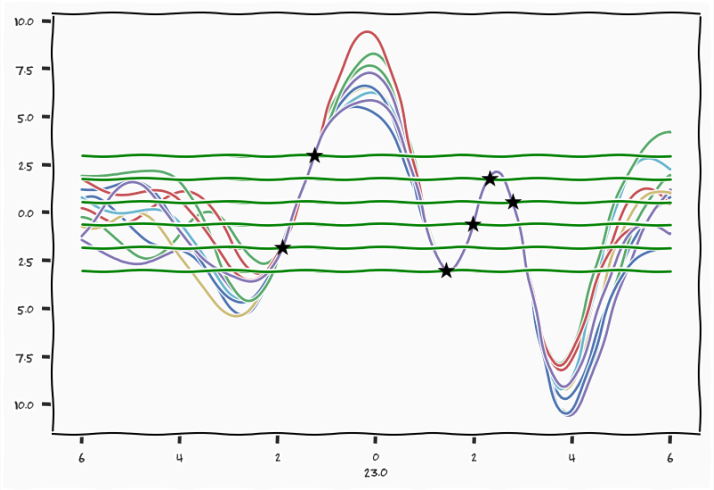

3
2
1
0
1
2
3
3
2
1
0
1
2
3

Unsupervised Learning




4
2
0
2
4
6
4
2
0
2
4
6
110


4
3
2
1
0
1
2
3
4
6
4
2
0
2
4
6
110


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

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".

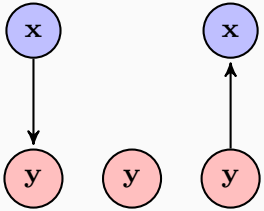
Regression there are infinite number of possible functions that connects the data equally well. A GP provides a measure over these solutions that makes the problem "well-posed".

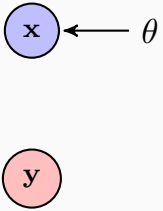
Unsupervised Learning there are infinite number of possible combinations of input locations and functions that generate the data equally well. A GP and a latent space prior jointly provides a measure over these solutions to make the problem "well-posed"

Approximate Inference

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

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





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

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

$$p(y)$$

$$\log p(y)$$

$$\log p(y) = \log p(y) + \int \log \frac{p(x|y)}{p(x|y)} \mathrm{d}x$$

$$\log p(y) = \log p(y) + \int \log \frac{p(x|y)}{p(x|y)} \mathrm{d}x$$
$$= \int q(x) \log p(y) \mathrm{d}x + \int q(x) \log \frac{p(x|y)}{p(x|y)} \mathrm{d}x$$

$$\log p(y) = \log p(y) + \int \log \frac{p(x|y)}{p(x|y)} \mathrm{d}x$$
$$= \int q(x) \log p(y) \mathrm{d}x + \int q(x) \log \frac{p(x|y)}{p(x|y)} \mathrm{d}x$$
$$= \int q(x) \log \frac{p(x|y)p(y)}{p(x|y)} \mathrm{d}x$$

$$\log p(y) = \log p(y) + \int \log \frac{p(x|y)}{p(x|y)} \mathrm{d}x$$
$$= \int q(x) \log p(y) \mathrm{d}x + \int q(x) \log \frac{p(x|y)}{p(x|y)} \mathrm{d}x$$
$$= \int q(x) \log \frac{p(x|y)p(y)}{p(x|y)} \mathrm{d}x = \int q(x) \log \frac{p(x,y)}{p(x|y)} \mathrm{d}x$$

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

$$\log p(y) = \log p(y) + \int \log \frac{p(x|y)}{p(x|y)} \mathrm{d}x$$

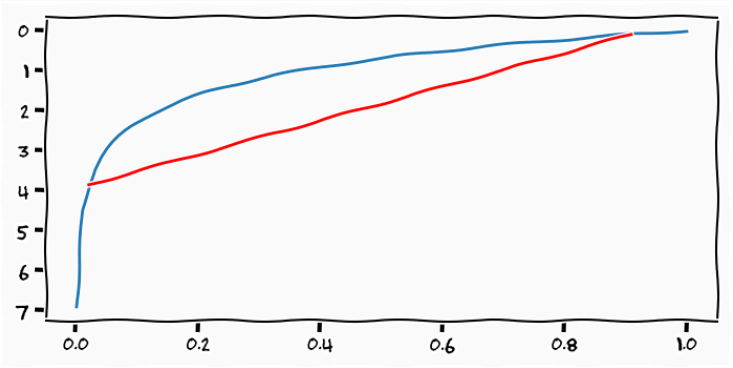
$$= \int q(x) \log p(y) \mathrm{d}x + \int q(x) \log \frac{p(x|y)}{p(x|y)} \mathrm{d}x$$

$$= \int q(x) \log \frac{p(x|y)p(y)}{p(x|y)} \mathrm{d}x = \int q(x) \log \frac{p(x,y)}{p(x|y)} \mathrm{d}x$$

$$= \int q(x) \log \frac{q(x)}{q(x)} \mathrm{d}x + \int q(x) \log p(x,y) \mathrm{d}x + \int q(x) \log \frac{1}{p(x|y)} \mathrm{d}x$$

$$= \int q(x) \log \frac{1}{q(x)} \mathrm{d}x + \int q(x) \log p(x,y) \mathrm{d}x + \int q(x) \log \frac{q(x)}{p(x|y)} \mathrm{d}x$$

Jensen Inequality



The "posterior" term

$$\int q(x) \log \frac{q(x)}{p(x|y)} \mathrm{d}x$$

The "posterior" term

$$\int q(x) \log \frac{q(x)}{p(x|y)} \mathrm{d}x = -\int q(x) \log \frac{p(x|y)}{q(x)} \mathrm{d}x$$

The "posterior" term

$$\begin{aligned}\int q(x)\,\log\,\frac{q(x)}{p(x|y)}\mathrm{d}x &= -\int q(x)\,\log\,\frac{p(x|y)}{q(x)}\mathrm{d}x \\ &\geq \log\int p(x|y)\mathrm{d}x \\ &= \log 1 = 0\end{aligned}$$

The "posterior" term

$$\int q(x) \log \frac{q(x)}{p(x|y)} \mathrm{d}x$$

The "posterior" term

$$\int q(x) \log \frac{q(x)}{p(x|y)} \mathrm{d}x = \{\text{Lets assume that } q(x) = p(x|y)\}$$

The "posterior" term

$$\int q(x) \log \frac{q(x)}{p(x|y)} \mathrm{d}x = \{\text{Lets assume that } q(x) = p(x|y)\}$$

$$= \int p(x|y) \log \underbrace{\frac{p(x|y)}{p(x|y)}}_{=1} \mathrm{d}x$$

$$\int q(x) \log \frac{q(x)}{p(x|y)} \mathrm{d}x = \{\text{Lets assume that } q(x) = p(x|y)\}$$

$$= \int p(x|y) \log \underbrace{\frac{p(x|y)}{p(x|y)}}_{=1} \mathrm{d}x$$

$$= 0$$

Kullback-Leibler Divergence

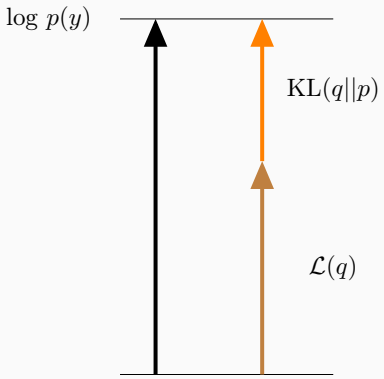
$$KL(q(x)||p(x|y)) = \int q(x) \log \frac{q(x)}{p(x|y)} \mathrm{d}x$$

- Measure of divergence between distributions
- Not a metric (not symmetric)
- $KL(q(x)||p(x|y)) = 0 \Leftrightarrow q(x) = p(x|y)$
- $KL(q(x)||p(x|y)) \geq 0$

$$\log p(y) = \int q(x) \log \frac{1}{q(x)} \mathrm{d}x + \int q(x) \log p(x, y) \mathrm{d}x + \int q(x) \log \frac{q(x)}{p(x|y)} \mathrm{d}x$$

$$\geq - \int q(x) \log q(x) \mathrm{d}x + \int q(x) \log p(x, y) \mathrm{d}x$$

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


log $p(y)$
KL$(q||p)$
$\mathcal{L}(q)$

$$\log p(y) \geq -\int q(x) \log q(x) \mathrm{d}x + \int q(x) \log p(x, y) \mathrm{d}x$$
$$= \mathbb{E}_{q(x)} \left[\log p(x, y)\right] - H(q(x)) = \mathcal{L}(q(x))$$

- if we maximise the ELBO we,
  - find an approximate posterior
  - lower bound the marginal likelihood
- *maximising* $p(y)$ is learning
- finding $q(x) \approx p(x|y)$ is prediction

How to choose Q?

$$\mathcal{L}(q(x)) = \mathbb{E}_{q(x)}\left[\log\, p(x, y)\right] - H(q(x))$$

- We have to be able to compute an expectation over the joint distribution
- The second term should be trivial

Lower Bound[1]

$$\mathcal{L} = \int_x q(x) \log \left(\frac{p(y, f, x)}{q(x)} \right)$$

[1] Damianou, 2015

Lower Bound[1]

$$\begin{aligned}\mathcal{L} &= \int_x q(x) \log \left(\frac{p(y, f, x)}{q(x)} \right) \\ &= \int_x q(x) \log \left(\frac{p(y \mid f) p(f \mid x) p(x))}{q(x)} \right)\end{aligned}$$

[1] Damianou, 2015

Lower Bound[1]

$$
\begin{aligned}
\mathcal{L} &= \int_x q(x) \log \left(\frac{p(y, f, x)}{q(x)} \right) \\
&= \int_x q(x) \log \left(\frac{p(y \mid f) p(f \mid x) p(x))}{q(x)} \right) \\
&= \int_x q(x) \log p(y \mid f) p(f \mid x) - \int_x q(x) \log \frac{q(x)}{p(x)}
\end{aligned}
$$

[1] Damianou, 2015

Lower Bound[1]

$$
\begin{aligned}
\mathcal{L} &= \int_x q(x) \log \left(\frac{p(y, f, x)}{q(x)} \right) \\
&= \int_x q(x) \log \left(\frac{p(y \mid f) p(f \mid x) p(x))}{q(x)} \right) \\
&= \int_x q(x) \log p(y \mid f) p(f \mid x) - \int_x q(x) \log \frac{q(x)}{p(x)} \\
&= \tilde{\mathcal{L}} - \mathrm{KL}(q(x) \parallel p(x))
\end{aligned}
$$

[1] Damianou, 2015

$$\tilde{\mathcal{L}} = \int q(x) \log p(y|f) p(f|x) \mathrm{d}f \mathrm{d}x$$

- Has not eliviate the problem at all, x still needs to go through f to reach the data
- Idea of sparse approximations[2]

[2] Candela et al., 2005

Lower Bound [3]

$$p(f, u \mid x, z)$$

- Add another set of samples from the same prior
- Conditional distribution

[3] Titsias et al., 2010

Lower Bound [3]

$$p(f, u \mid x, z) = p(f \mid u, x, z)p(u \mid z)$$

- Add another set of samples from the same prior
- Conditional distribution

[3] Titsias et al., 2010

Lower Bound [3]

$$
\begin{aligned}
p(f, u \mid x, z) &= p(f \mid u, x, z) p(u \mid z) \\
&= \mathcal{N}(f \mid K_{fu} K_{uu}^{-1} u, K_{ff} - K_{fu} K_{uu}^{-1} K_{uf}) \mathcal{N}(u \mid \mathbf{0}, K_{uu})
\end{aligned}
$$

- Add another set of samples from the same prior
- Conditional distribution

[3] Titsias et al., 2010

$$p(y, f, u, x \mid z) = p(y \mid f)p(f \mid u, x)p(u \mid z)p(x)$$

- we have done nothing to the model, just project an additional set of marginals from the GP
- *however* we will now interpret u and z not as random variables but variational parameters
- i.e. the variational distribution $q(\cdot)$ is parametrised by these

- Variational distributions are approximations to intractable posteriors,

$$q(u) \approx p(u \mid y, x, z, f)$$
$$q(f) \approx p(f \mid u, x, z, y)$$
$$q(x) \approx p(x \mid y)$$

Lower Bound

- Variational distributions are approximations to intractable posteriors,

$$
\begin{aligned}
q(u) &\approx p(u \mid y, x, z, f) \\
q(f) &\approx p(f \mid u, x, z, y) \\
q(x) &\approx p(x \mid y)
\end{aligned}
$$

- Bound is tight if u completely represents f i.e. u is sufficient statistics for f

$$
q(f) \approx p(f \mid u, x, z, y) = p(f \mid u, x, z)
$$

Lower Bound

$$\tilde{\mathcal{L}} = \int_{x,f,u} q(f)q(u)q(x)\log\frac{p(y,f,y \mid x,z)}{q(f)q(u)}$$

$$\tilde{\mathcal{L}} = \int_{x,f,u} q(f)q(u)q(x)\log\frac{p(y,f,y \mid x,z)}{q(f)q(u)}$$
$$= \int_{x,f,u} q(f)q(u)q(x)\log\frac{p(y \mid f)p(f \mid u,x,z)p(u \mid z)}{q(f)q(u)}$$

- Assume that u is sufficient statistics of f

$$q(f) = p(f \mid u,x,z)$$

Lower Bound

$$\tilde{\mathcal{L}} = \int_{x,f,u} q(f)q(u)q(x)\log\frac{p(y \mid f)p(f \mid u,x,z)p(u \mid z)}{q(f)q(u)}$$

Lower Bound

$$\begin{aligned}\tilde{\mathcal{L}} &= \int_{x,f,u} q(f)q(u)q(x)\log\frac{p(y \mid f)p(f \mid u,x,z)p(u \mid z)}{q(f)q(u)} \\ &= \int_{x,f,u} p(f \mid u,x,z)q(u)q(x)\log\frac{p(y \mid f)p(f \mid u,x,z)p(u \mid z)}{p(f \mid u,x,z)q(u)}\end{aligned}$$

Lower Bound

$$\tilde{\mathcal{L}} = \int_{x,f,u} q(f)q(u)q(x)\log\frac{p(y \mid f)p(f \mid u,x,z)p(u \mid z)}{q(f)q(u)}$$
$$= \int_{x,f,u} p(f \mid u,x,z)q(u)q(x)\log\frac{p(y \mid f)p(f \mid u,x,z)p(u \mid z)}{p(f \mid u,x,z)q(u)}$$

$$\begin{aligned}
\tilde{\mathcal{L}} &= \int_{x,f,u} q(f)q(u)q(x)\log\frac{p(y \mid f)p(f \mid u,x,z)p(u \mid z)}{q(f)q(u)} \\
&= \int_{x,f,u} p(f \mid u,x,z)q(u)q(x)\log\frac{p(y \mid f)p(f \mid u,x,z)p(u \mid z)}{p(f \mid u,x,z)q(u)} \\
&= \int_{x,f,u} p(f \mid u,x,z)q(u)q(x)\log\frac{p(y \mid f)p(u \mid z)}{q(u)}
\end{aligned}$$

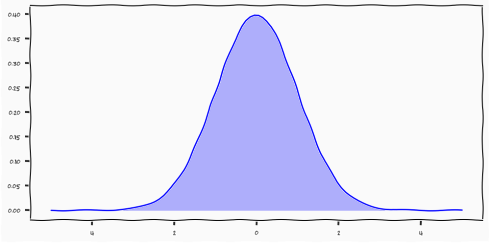
$$
\begin{aligned}
\tilde{\mathcal{L}} &= \int_{x,f,u} q(f)q(u)q(x)\log\frac{p(y \mid f)p(f \mid u,x,z)p(u \mid z)}{q(f)q(u)} \\
&= \int_{x,f,u} p(f \mid u,x,z)q(u)q(x)\log\frac{p(y \mid f)p(f \mid u,x,z)p(u \mid z)}{p(f \mid u,x,z)q(u)} \\
&= \int_{x,f,u} p(f \mid u,x,z)q(u)q(x)\log\frac{p(y \mid f)p(u \mid z)}{q(u)} \\
&= \mathbb{E}_{p(f|u,x,z)}[p(y \mid f)] - \mathrm{KL}(q(u) \parallel p(u \mid z))
\end{aligned}
$$

Lower Bound

$$\mathcal{L} = \mathbb{E}_{p(f|u,x,z)}[p(y \mid f)] - \mathrm{KL}(q(u) \parallel p(u \mid z)) - \mathrm{KL}(q(x) \parallel p(x))$$

- Expectation tractable (for some co-variances)
- Allows us to place priors and not "regularisers" over the latent representation
- Stochastic inference Hensman et al., 2013
- Importantly $p(x)$ only appears in $\mathrm{KL}(\cdot \parallel \cdot)$ term!

Latent Space Priors



$$p(x) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$$

Automatic Relevance Determination

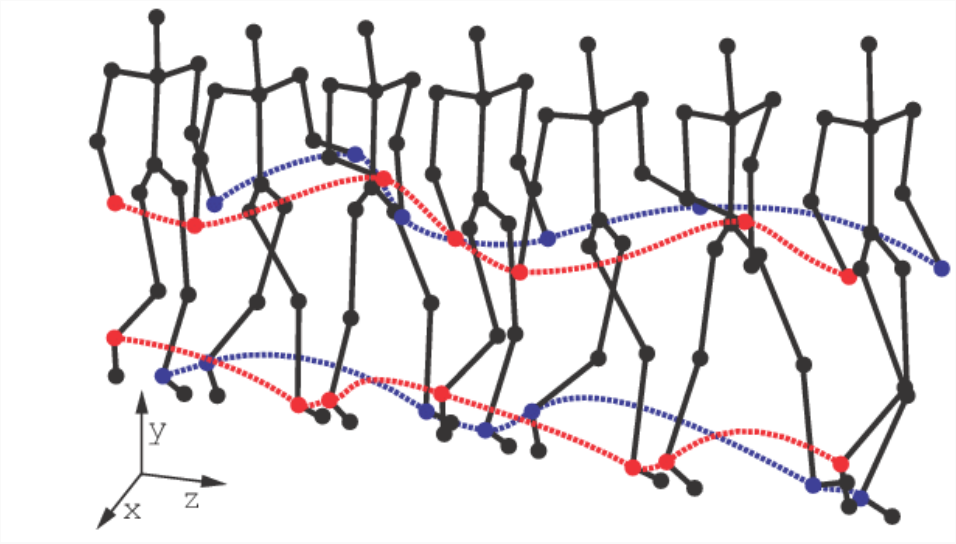
$$k(\mathbf{x}_i, \mathbf{x}_j) = \sigma e^{-\sum_d^D \alpha_d \cdot (x_{i,d} - x_{j,d})^2}$$

GPy

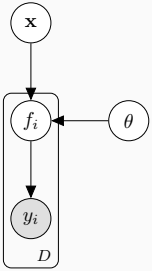
Code

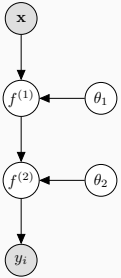
```
RBF(...,ARD=True)
Matern32(...,ARD=True)
```

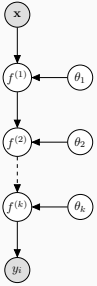
Dynamic Prior



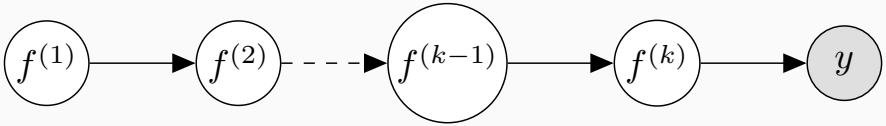
Composite Gaussian Processes







Composite Gaussian Processes [4]



$$y = f^{(k)}(f^{(k-1)}(\cdots f^{(2)}(f^{(1)}(x))))$$

[4] Damianou et al., 2013

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

- 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

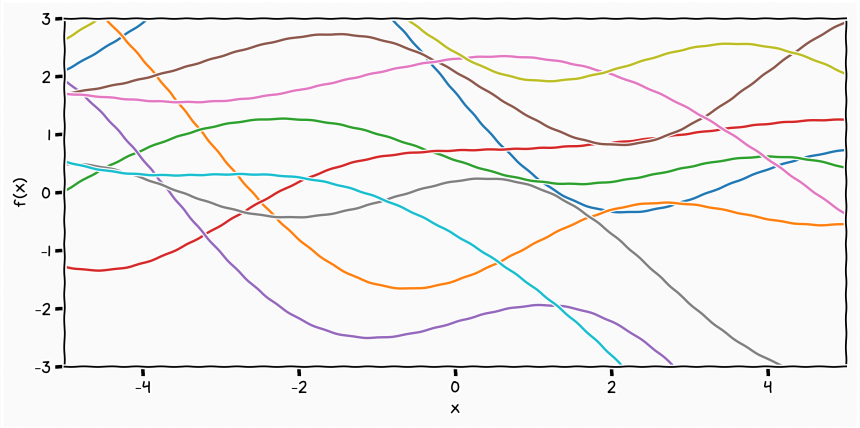
$$\nu_{t_{\pi(1)} \cdots t_{\pi(k)}} \left(F_{\pi(1)} \times \cdot \times F_{\pi(k)} \right) = \nu_{t_1 \cdots t_k} \left(F_1 \times \cdots \times F_k \right)$$

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

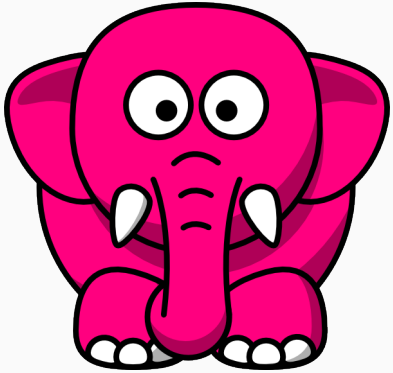
Yes their parametrisation is very well aligned to the knowledge we have of many problems, most complex knowledge (like beer) is relative

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

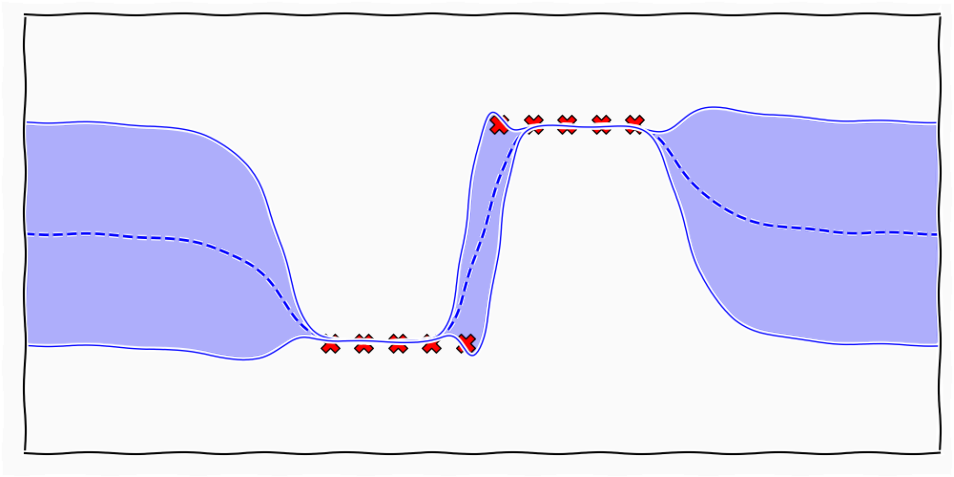
Yes they are incredibly "narrow" but have infinite coverage



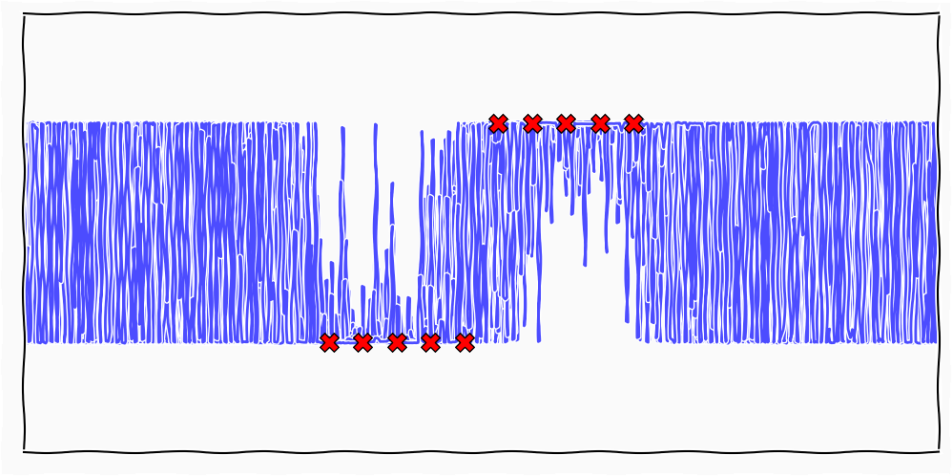
Composite Functions

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

Composite GP Step



Composite GP Step



Best of both worlds?

Composite GPs potentially interesting but inference is a huge issue and

Best of both worlds?

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

Best of both worlds?

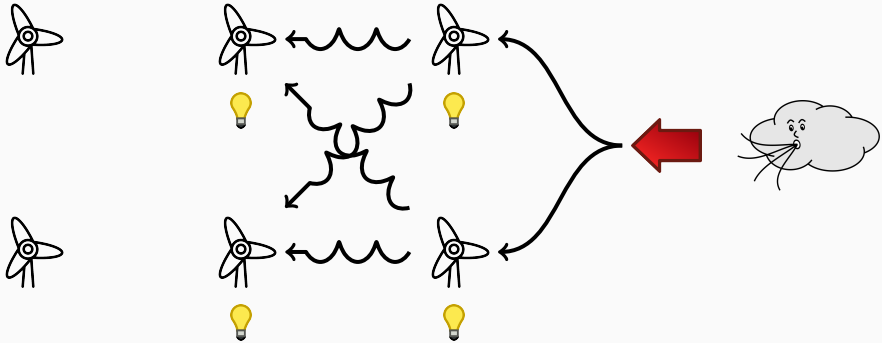
Composite GPs potentially interesting but inference is a huge issue and

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

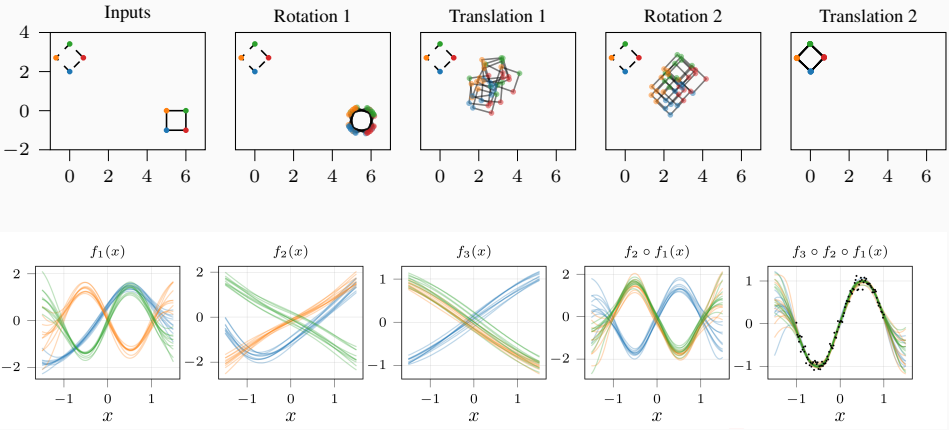
When should we use composite models when our knowledge is composite

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

Windfarms Kaiser et al., 2018



Composite Model[5]



[5] Ustyuzhaninov et al., 2020

eof

References

- Candela, Joaquin Quiñonero and Carl Edward Rasmussen (2005). "A Unifying View of Sparse Approximate Gaussian Process Regression." In: *Journal of Machine Learning Research* 6, pp. 1939–1959.
- Damianou, Andreas C (Feb. 2015). "Deep Gaussian Processes and Variational Propagation of Uncertainty." PhD thesis. University of Sheffield.
- Damianou, Andreas C and Neil D Lawrence (2013). "Deep Gaussian Processes." In: *International Conference on Airtificial Inteligence and Statistical Learning*, pp. 207–215.

- Hensman, James, N Fusi, and Neil D Lawrence (2013). "Gaussian Processes for Big Data." In: *Uncertainty in Artificial Intelligence*.
- 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]*.
- Shalev-Shwartz, Shai and Shai Ben-David (2014). *Understanding Machine Learning: From Theory to Algorithms*. New York, NY, USA: Cambridge University Press.

- Titsias, Michalis and Neil D Lawrence (2010). "Bayesian Gaussian Process Latent Variable Model." In: *International Conference on Airtificial Inteligence and Statistical Learning*, pp. 844–851.
- 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.