Bayesian non-parametrics and priors over functions

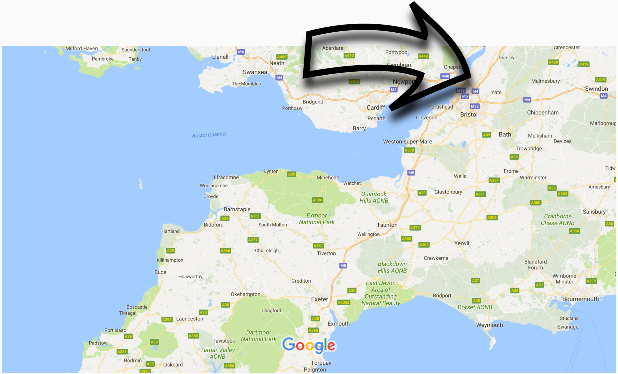
Carl Henrik Ek - `carlhenrik.ek@bristol.ac.uk`

November 22, 2017

`http://www.carlhenrik.com`

Introductions

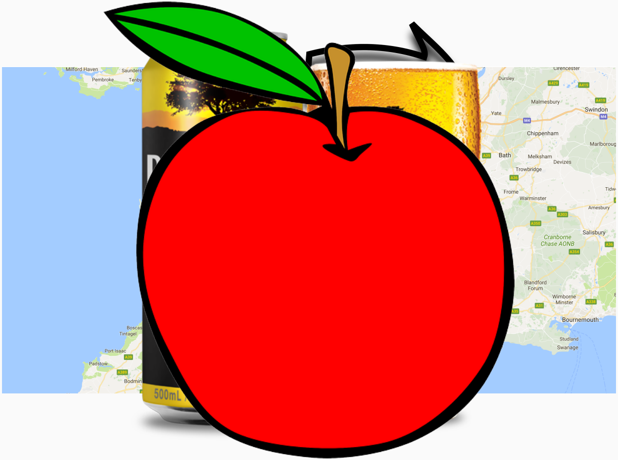
Me

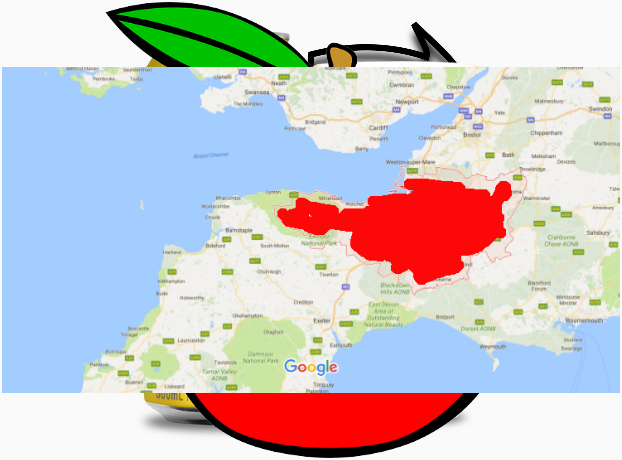


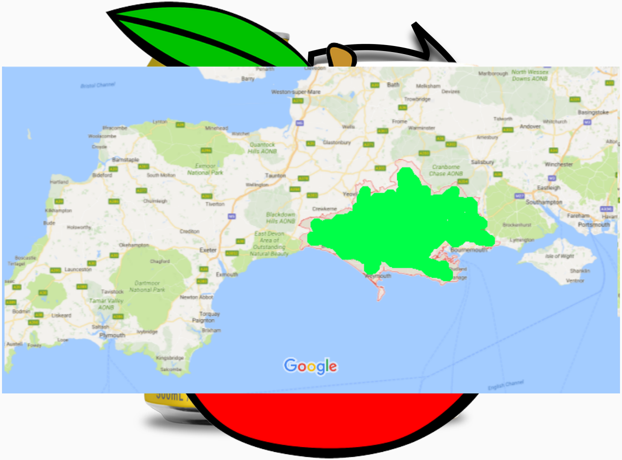
Milford Haven
Saundersfoot
Pembroke
Tenby
Llanelli
Swansea
Neath
The Mumbles
Aberdare
Pontypool
Cwmbran
Newport
Bridgend
Porthcawl
Cardiff
Penarth
Barry
Portishead
Clevedon
Bristol
Dursley
Cirencester
Malmesbury
Yate
Swindon
Chippenham
Marlborough
Bath
Melksham
Devizes
Trowbridge
Bristol Channel
Weston-super-Mare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

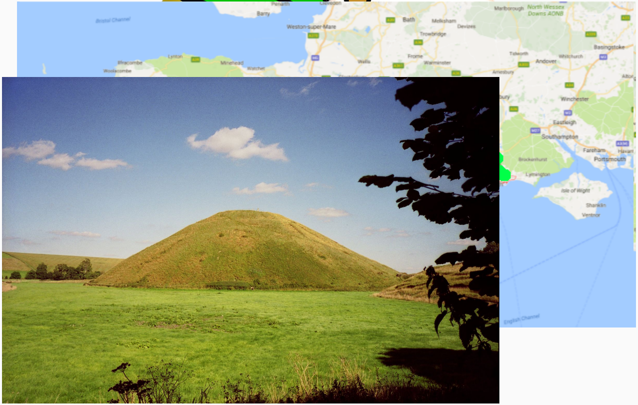

Milford Haven
Saundersfoot
Pembroke
Tenby
Llanelli
Neath
Swansea
The Mumbles
Aberdare
Pontypool
Cardiff
Penarth
Barry
Portishead
Clevedon
Bristol
Yate
Dursley
Cirencester
Malmesbury
Swindon
Chippenham
Bath
Melksham
Devizes
Trowbridge
Bristol Channel
Weston-super-Mare
Frome
Warminster
Amesbury
Salisbury
Cranborne Chase AONB
Ilfracombe
Woolacombe
Croyde
Hartland
Kilkhampton
Bude
Yeovil
Blandford Forum
Wimborne Minster
Bournemouth
Dorset AONB
Weymouth
Studland
Swanage
Boscastle
Tintagel
Port Isaac
Padstow
Launceston
Tavistock
Tamar Valley AONB
Bodmin
Liskeard
Google


BLACKTHORN
A DISTINCTIVE CRISP, DRY TASTE
FROM ENGLISH CIDER APPLES
IMPORTED
Fermented Hard Cider
PRODUCT OF IRELAND
500mL / 16.9 fl. oz. 6% ALC./VOL.
BLACKTHORN
Milford Haven
Pembroke
Tenby
Tintagel
Port Isaac
Padstow
Bodmin
Cirencester
Dursley
Malmesbury
Swindon
Yate
Chippenham
Bath
Melksham
Devizes
Trowbridge
Frome
Warminster
Amesbury
Salisbury
Cranborne Chase AONB
Blandford Forum
Wimborne Minster
Bournemouth
Weymouth
Studland
Swanage

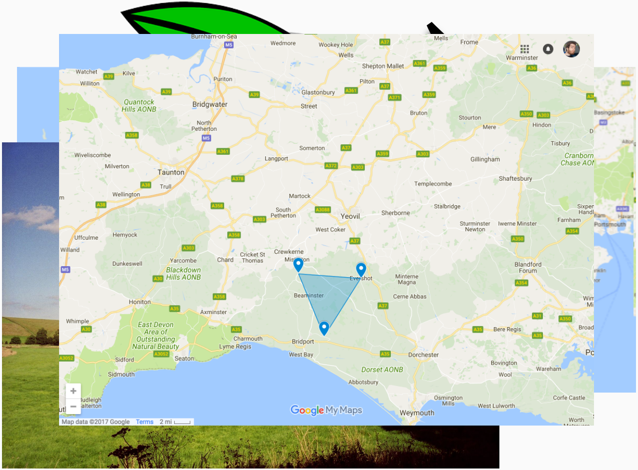

Milford Haven
Pembroke
Tenby
Tintagel
Port Isaac
Padstow
Bodmin
500mL
Dursley
Malmesbury
Swindon
Yate
Chippenham
Bath
Melksham
Devizes
Trowbridge
Frome
Warminster
Amesbury
Salisbury
Cranborne
Chase AONB
Blandford
Forum
Wimborne
Minster
Bournemouth
Studland
Swanage

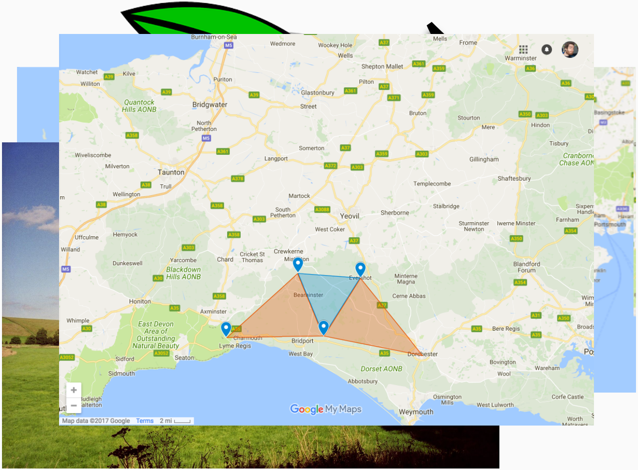

Swansea
Newport
Cardiff
Bristol
Bath
Chippenham
Swindon
Weston-super-Mare
Bristol Channel
Ilfracombe
Barnstaple
Bideford
Exmoor National Park
Tiverton
Salisbury
Cranborne Chase AONB
Blackdown Hills AONB
East Devon Area of Outstanding Natural Beauty
Exeter
Exmouth
Dorset AONB
Weymouth
Bournemouth
Dartmoor National Park
Launceston
Tavistock
Google


Google
Exeter
Plymouth
Taunton
Bath
Weston-super-Mare
Barnstaple
Salisbury
Southampton
Portsmouth
Bournemouth
Weymouth
Winchester
Andover
Dartmoor National Park
Exmoor National Park

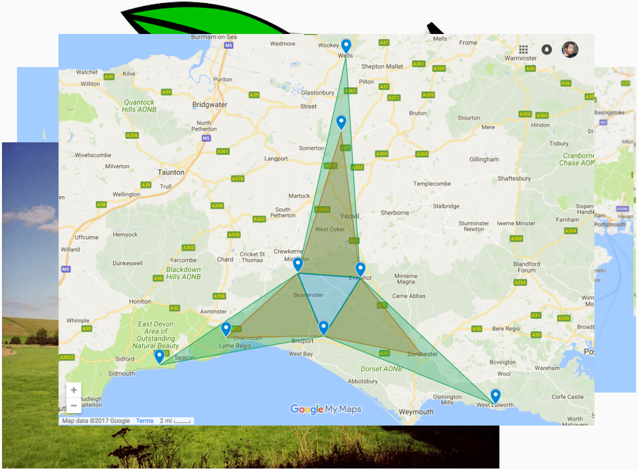


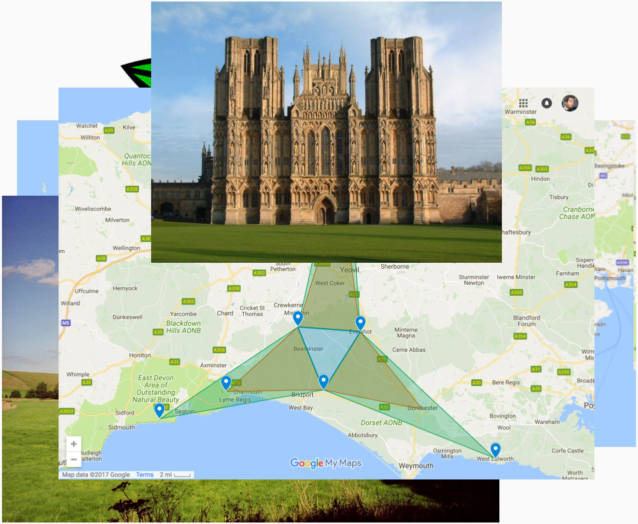
Bristol Channel
Barry
Weston-super-Mare
Bath
Melksham
Trowbridge
Devizes
Marlborough
North Wessex Downs AONB
Basingstoke
Andover
Frome
Warminster
Wells
Ilfracombe
Minehead
Winchester
Eastleigh
Southampton
Fareham
Portsmouth
Brockenhurst
Lymington
Isle of Wight
Shanklin
Ventnor

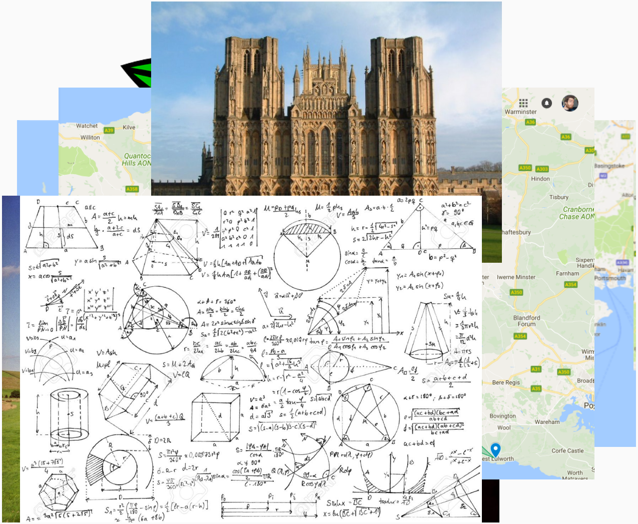

Burnham-on-Sea
Wedmore
Wookey Hole
Wells
Mells
Frome
Warminster
Watchet
Kilve
Williton
Puriton
Shepton Mallet
Pilton
Quantock Hills AONB
Bridgwater
Glastonbury
Street
Bruton
Stourton
Mere
Hindon
North Petherton
Tisbury
Wiveliscombe
Somerton
Langport
Gillingham
Cranborne Chase AONB
Milverton
Taunton
Trull
Templecombe
Shaftesbury
Wellington
Martock
South Petherton
Yeovil
Sherborne
Stalbridge
West Coker
Sturminster Newton
Iwerne Minster
Farnham
Uffculme
Hemyock
Willand
Crewkerne
Cricket St Thomas
Chard
Yarcombe
Dunkeswell
Blackdown Hills AONB
Evershot
Minterne Magna
Blandford Forum
Beaminster
Cerne Abbas
Honiton
Axminster
Whimple
East Devon Area of Outstanding Natural Beauty
Charmouth
Lyme Regis
Bridport
Bere Regis
West Bay
Dorchester
Sidford
Seaton
Bovington
Wareham
Sidmouth
Dorset AONB
Wool
Abbotsbury
Budleigh Salterton
Osmington Mills
West Lulworth
Corfe Castle
Weymouth
Worth
Google My Maps
Map data ©2017 Google
Terms
2 mi

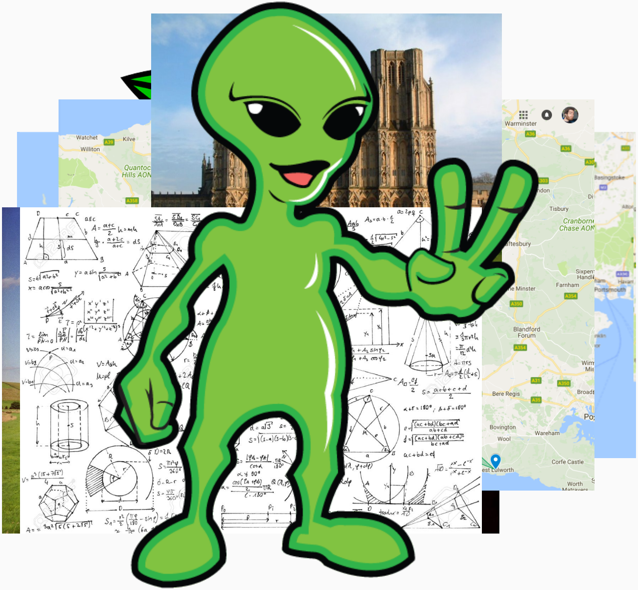

Burnham-on-Sea
Wedmore
Wookey Hole
Wells
Mells
Frome
Warminster
Shepton Mallet
Pilton
Watchet
Kilve
Williton
Puriton
Glastonbury
Street
Quantock Hills AONB
Bridgwater
North Petherton
Bruton
Stourton
Mere
Hindon
Tisbury
Cranborne Chase AONB
Wiveliscombe
Milverton
Taunton
Trull
Langport
Somerton
Gillingham
Shaftesbury
Templecombe
Wellington
Martock
South Petherton
Yeovil
Sherborne
Stalbridge
Sturminster Newton
Iwerne Minster
Farnham
West Coker
Uffculme
Hemyock
Willand
Dunkeswell
Yarcombe
Chard
Cricket St Thomas
Crewkerne
Blackdown Hills AONB
Evershot
Minterne Magna
Blandford Forum
Beaminster
Cerne Abbas
Honiton
Axminster
Whimple
East Devon Area of Outstanding Natural Beauty
Charmouth
Lyme Regis
Bridport
West Bay
Bere Regis
Dorchester
Sidford
Seaton
Sidmouth
Dorset AONB
Abbotsbury
Bovington
Wool
Wareham
Osmington Mills
West Lulworth
Corfe Castle
Weymouth
Worth Matravers
Google My Maps
Map data ©2017 Google
Terms
2 mi


Burnham-on-Sea
Wedmore
Wookey Hole
Wells
Mells
Frome
Warminster
Watchet
Kilve
Williton
Puriton
Quantock Hills AONB
Bridgwater
Glastonbury
Street
Shepton Mallet
Pilton
Bruton
Stourton
Mere
Hindon
North Petherton
Somerton
Langport
Tisbury
Wiveliscombe
Milverton
Taunton
Trull
Wellington
Gillingham
Templecombe
Shaftesbury
Martock
South Petherton
Yeovil
Sherborne
Stalbridge
West Coker
Sturminster Newton
Iwerne Minster
Farnham
Uffculme
Hemyock
Willand
Crewkerne
Cricket St Thomas
Chard
Yarcombe
Dunkeswell
Blackdown Hills AONB
Minterne Magna
Blandford Forum
Cerne Abbas
Honiton
Axminster
Whimple
East Devon Area of Outstanding Natural Beauty
Charmouth
Lyme Regis
Bridport
Seaton
West Bay
Bere Regis
Sidford
Sidmouth
Dorset AONB
Bovington
Wareham
Wool
Abbotsbury
Osmington Mills
West Lulworth
Corfe Castle
Weymouth
Google My Maps
Map data ©2017 Google
Terms
2 mi

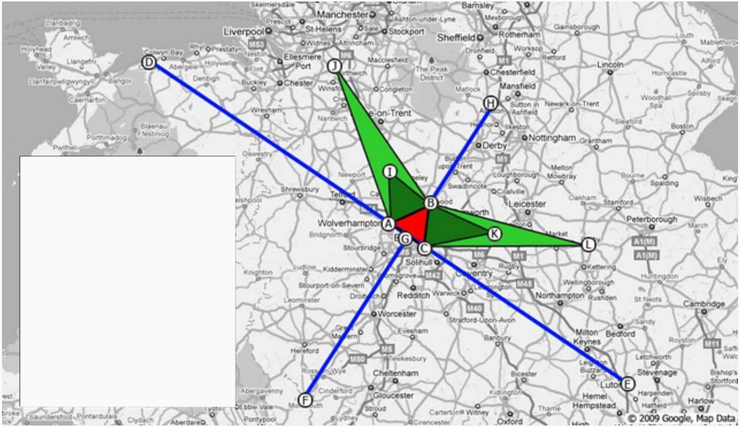

Burnham-on-Sea
Wedmore
Wookey
Wells
Mells
Frome
Warminster
Watchet
Kilve
Williton
Puriton
Shepton Mallet
Pilton
Glastonbury
Street
Quantock Hills AONB
Bridgwater
North Petherton
Bruton
Stourton
Mere
Hindon
Tisbury
Somerton
Langport
Wiveliscombe
Milverton
Taunton
Trull
Gillingham
Cranborne Chase AONB
Templecombe
Shaftesbury
Wellington
Martock
South Petherton
Yeovil
Sherborne
Stalbridge
West Coker
Sturminster Newton
Iwerne Minster
Farnham
Uffculme
Hemyock
Willand
Crewkerne
Chard
Cricket St Thomas
Yarcombe
Dunkeswell
Blackdown Hills AONB
Blandford Forum
Minterne Magna
Beaminster
Cerne Abbas
Honiton
Axminster
Whimple
East Devon Area of Outstanding Natural Beauty
Charmouth
Lyme Regis
Bridport
West Bay
Bere Regis
Seaton
Sidford
Sidmouth
Dorchester
Dorset AONB
Abbotsbury
Bovington
Wool
Wareham
Osmington Mills
West Lulworth
Corfe Castle
Weymouth
Worth Matravers
Google My Maps
Map data ©2017 Google
Terms
2 mi


Watchet
Kilve
Williton
Quantock Hills AON
Wiveliscombe
Milverton
Wellington
Uffculme
Hemyock
Willand
Dunkeswell
Yarcombe
Blackdown Hills AONB
Chard
Cricket St Thomas
Crewkerne
South Petherton
Yeovil
West Coker
Sherborne
Honiton
Axminster
Beaminster
Minterne Magna
Cerne Abbas
Whimple
East Devon Area of Outstanding Natural Beauty
Charmouth
Lyme Regis
Bridport
West Bay
Seaton
Sidford
Sidmouth
Dorchester
Dorset AONB
Abbotsbury
Osmington Mills
Weymouth
West Lulworth
Warminster
Hindon
Tisbury
Cranborne Chase AON
Sturminster Newton
Iwerne Minster
Farnham
Blandford Forum
Bere Regis
Bovington
Wool
Wareham
Corfe Castle
Google My Maps
Map data ©2017 Google Terms 2 mi


Watchet
Kilve
Williton
Quantock Hills AONB
Warminster
Hindon
Tisbury
Cranborne Chase AONB
Shaftesbury
Iwerne Minster
Farnham
Blandford Forum
Bere Regis
Bovington
Wool
Wareham
Corfe Castle
West Lulworth
Basingstoke
Portsmouth


Watchet
Kilve
Williton
Warminster
Tisbury
Farnham
Blandford Forum
Bere Regis
Bovington
Wool
Wareham
Corfe Castle

"Brooks has proved, he explains, that there were keen mathematicians here 5,000 years ago, millennia before the Greeks invented geometry: "Such is the mathematical precision, it is inconceivable that this work could have been carried out by the primitive indigenous culture we have always associated with such structures . . . all this suggests a culture existing in these islands in the past quite outside our expectation and experience today." He does not rule out extraterrestrial help." – The Guardian



[1] Bad Science Blog

"We know so little about the ancient Woolworths stores," he explains, "but we do still know their locations. I thought that if we analysed the sites we could learn more about what life was like in 2008 and how these people went about buying cheap kitchen accessories and discount CDs" – Matt Parker interviewed in The Guardian[1]

[1] Bad Science Blog


Lord Nelson

Laplace Demon [1]

Laplace's Demon [1]

An intelligence which at a given instant knew all the forces acting in nature and the position of every object in the universe - if endowed with a brain sufficiently vast to make all necessary calculations - could describe with a single formula the motions of the largest astronomical bodies and those of the smallest atoms. To such an intelligence, nothing would be uncertain; the future, like the past, would be an open book.

Laplace Demon [1]

All these efforts in the search for truth tend to lead the mind continously towards the intelligence we have just mentioned, although it will always remain infinetly distant from this intelligence.

Napoleon *"You have written this huge book on the system of the world without once mentioning the author of the universe."*

Napoleon *"You have written this huge book on the system of the world without once mentioning the author of the universe."*

Laplace *"I had no need for that assumption"*

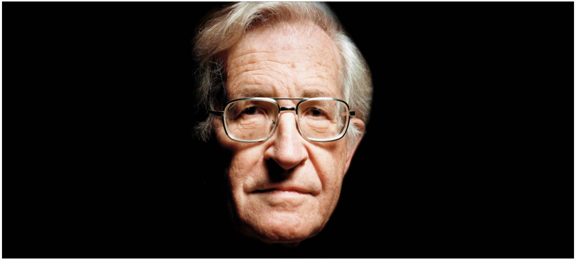
 

Napoleon *"You have written this huge book on the system of the world without once mentioning the author of the universe."*

Laplace *"I had no need for that assumption"*

Laplace *"Ah, but that is a fine hypothesis. It explains so many things"*

Inductivist Fallacy



2

[2] Chomsky, N. A., & Fodor, J. A. (1980). The inductivist fallacy. Language and Learning: The Debate between Jean Piaget and Noam Chomsky, (), .

Today

22/11/2017

Bayesian non-parametrics

Formalism

$$p(\mathbf{Y}|\theta)$$

$$\mathbf{Y} \in \mathcal{Y}$$

- Task of machine learning, describe models of data

Formalism

$$M \subset PM(\mathcal{Y})$$

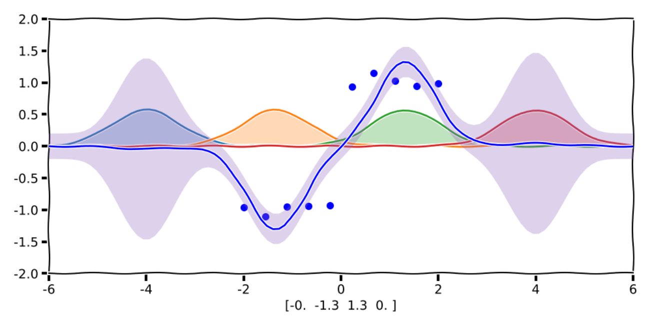
- all probability measures on the sample space $\mathcal{Y}$

Formalism

$$M = \{p(\mathbf{Y}|\theta)|\theta \in \mathcal{T}\}$$

- each model is indexed by θ from the parameter space $\mathcal{T}$

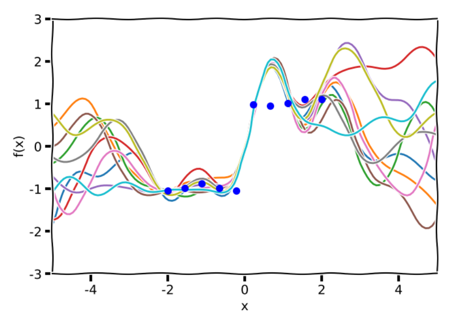
Linear Regression



$$p(\mathbf{w}|\mathbf{y},\mathbf{x}) = \frac{p(\mathbf{y}|\mathbf{x},\mathbf{w})p(\mathbf{w})}{p(\mathbf{y}|\mathbf{x})}$$

$$\mathcal{T} = \mathbb{R}^4$$

Non-Linear Regression



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

$$\mathcal{T} = \mathbb{R}^{\infty}$$

Parametric vs. Non-parametric

- If $\mathcal{T}$ is
  - infinite dimensional space we call this a non-parametric
  - finite dimensional space we call this a parametric

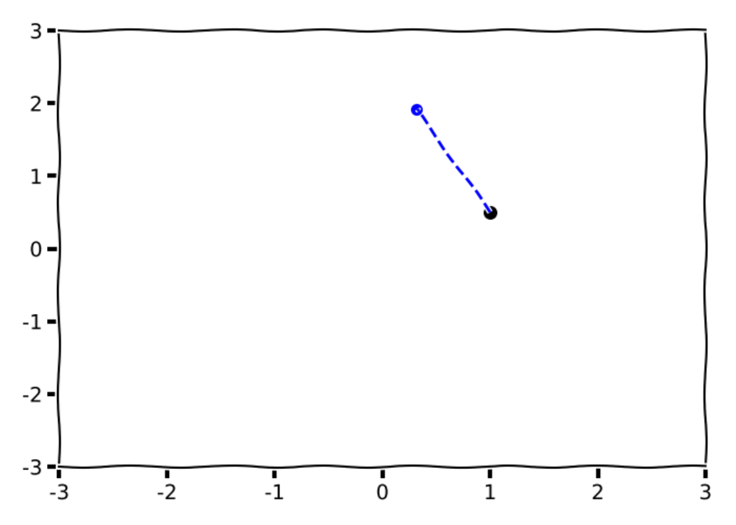
Nearest Neighbour

- Training data: $\{\mathbf{x}_i, y_i\}_{i=1}^{N}$
- Test data: $\{\mathbf{x}_i\}_{i=1}^{M}$
- Inference

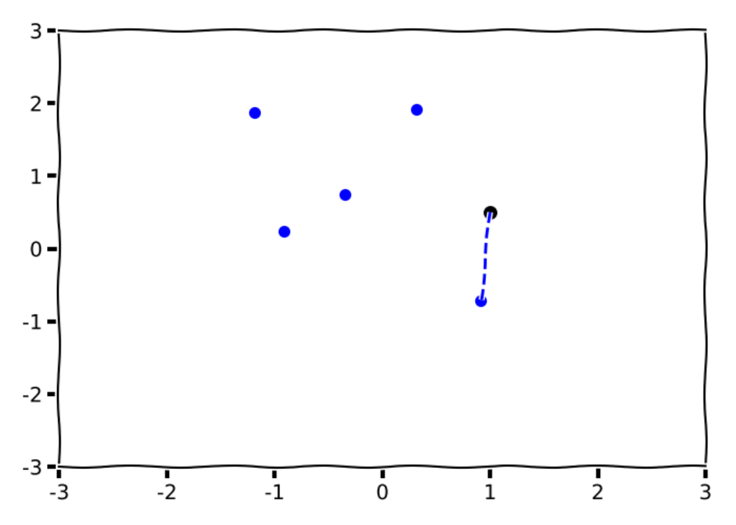
$$\hat{i} = \text{argmin}_i D(\mathbf{x}_*, \mathbf{x}_i)$$

- Complexity grows with number of training data
- Does not generalise at all

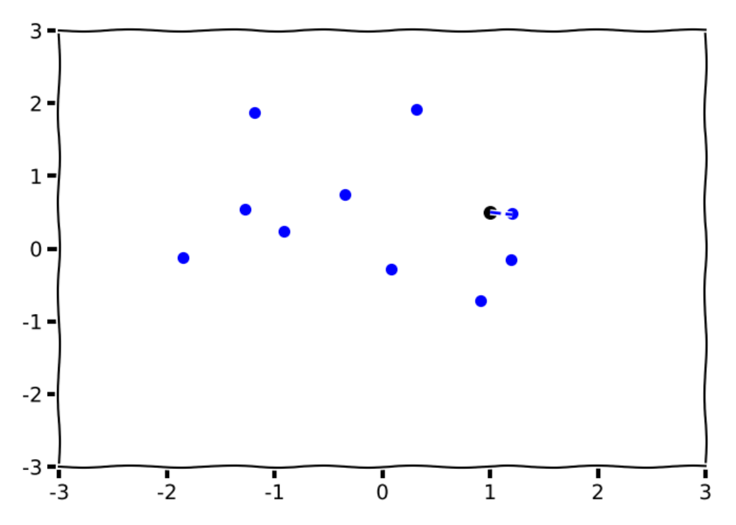
Nearest Neighbour



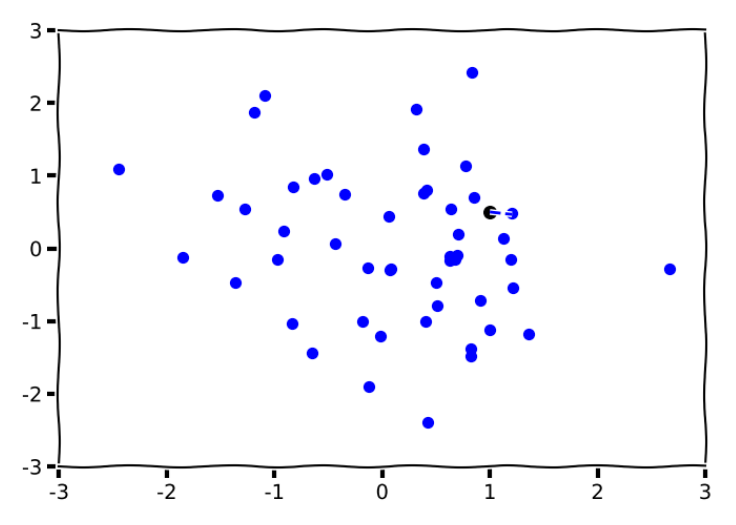
Nearest Neighbour



Nearest Neighbour



Nearest Neighbour



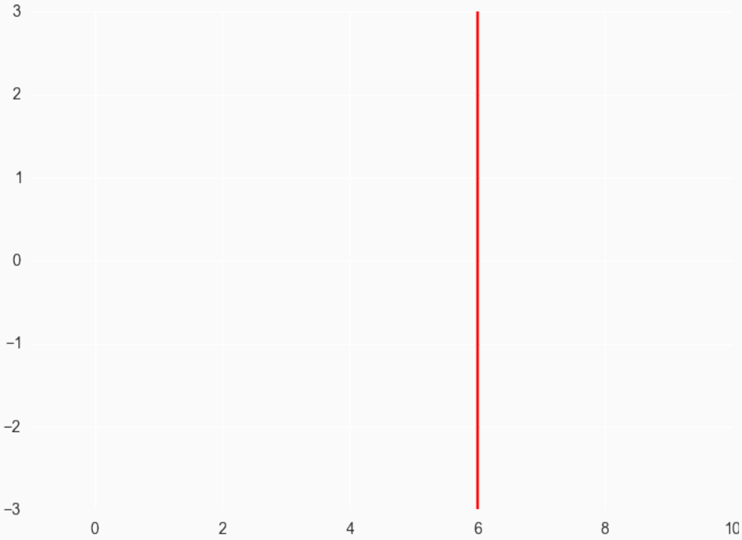
Bayesian Non-parametrics

$$\theta \sim Q$$

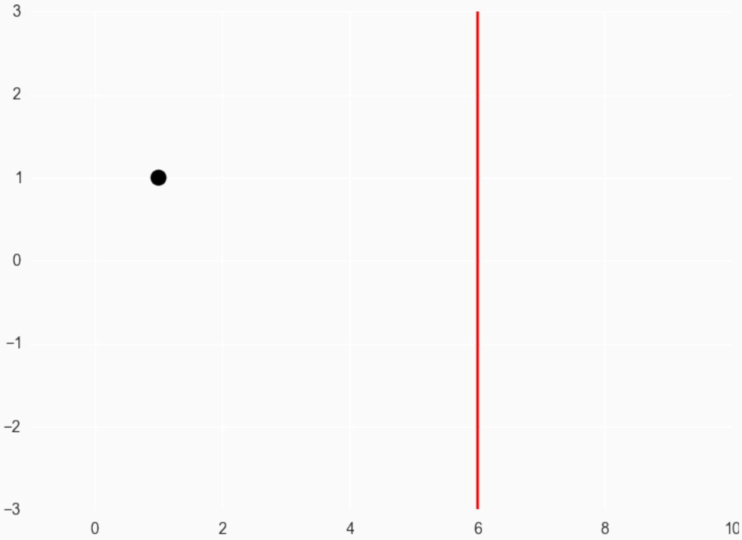
Treating the index into the parameter space as a random variable

Functions

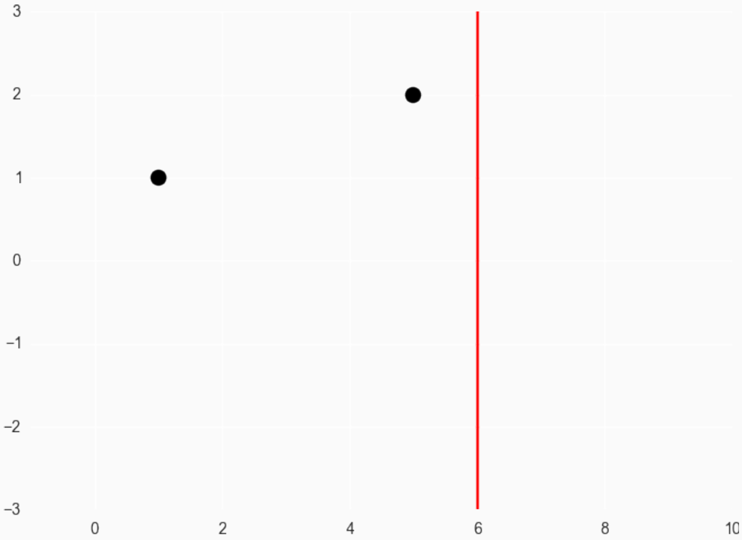
Functions



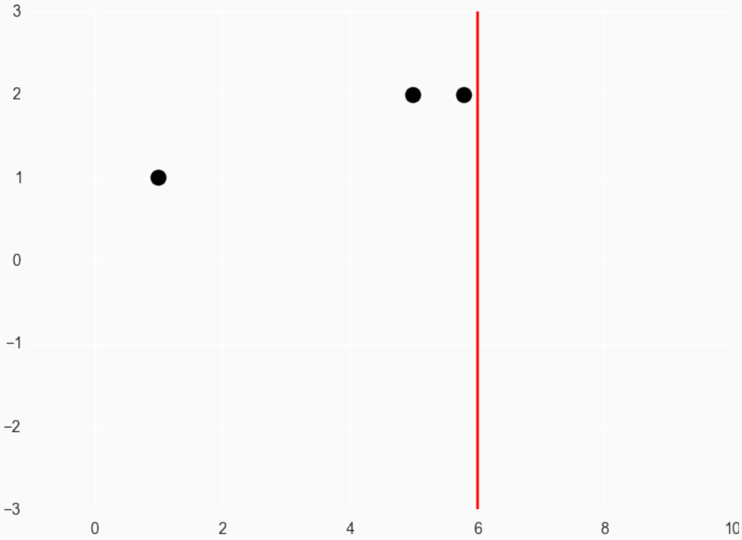
Functions



Functions



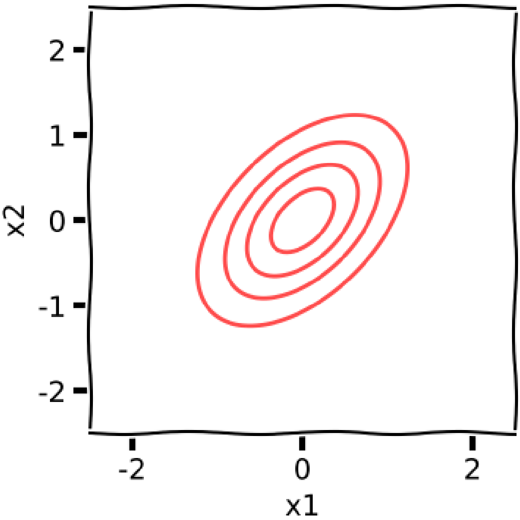
Functions



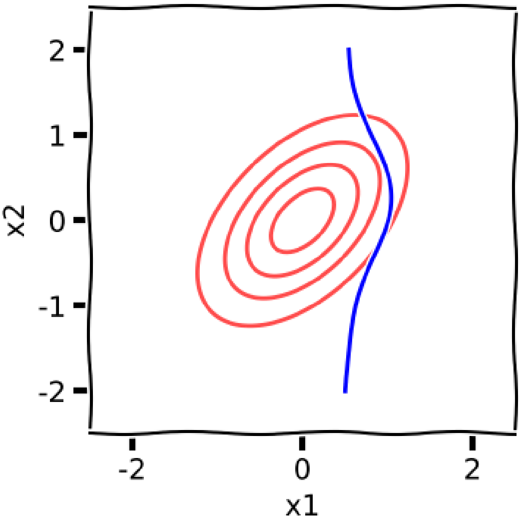
Conditional Gaussians

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

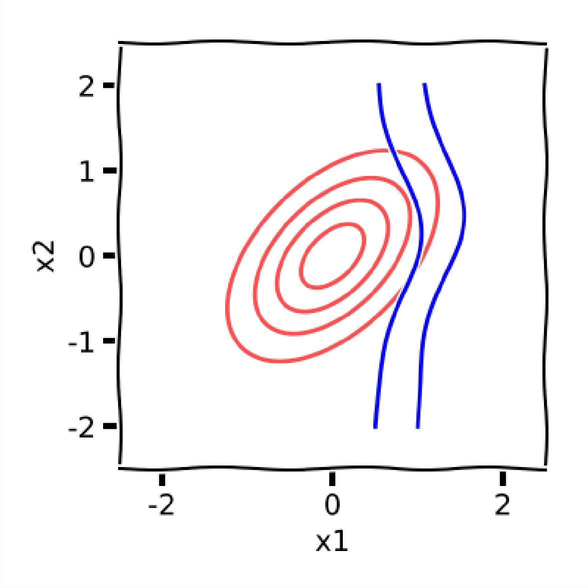
Conditional Gaussians



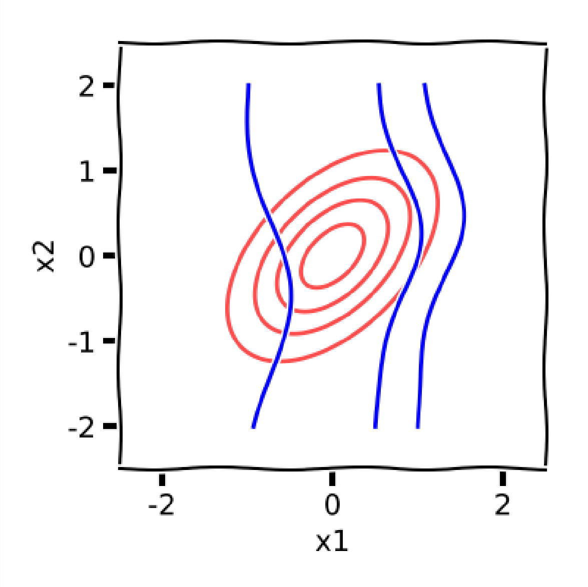
Conditional Gaussians



Conditional Gaussians



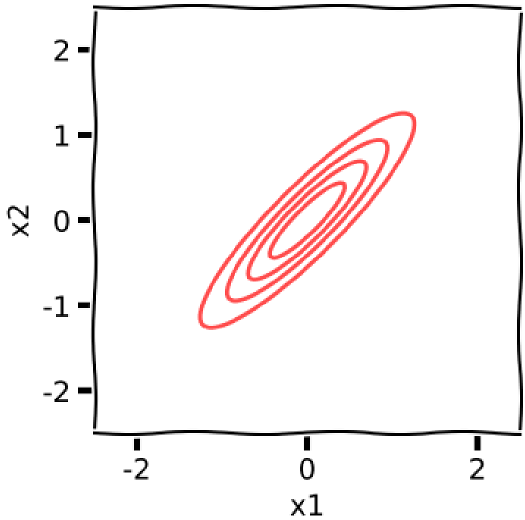
Conditional Gaussians



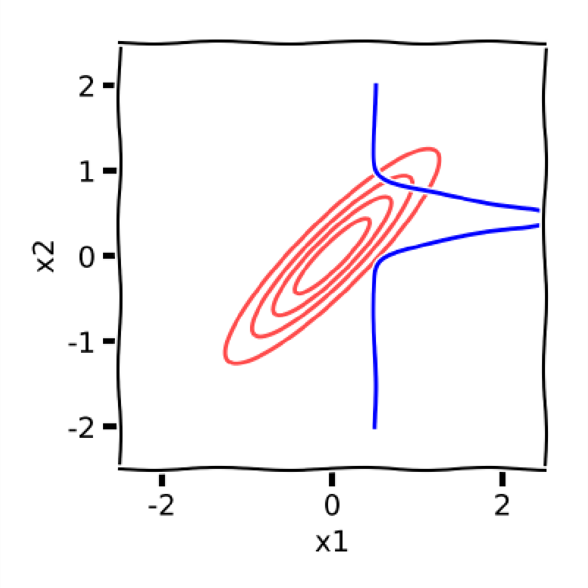
Conditional Gaussians

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

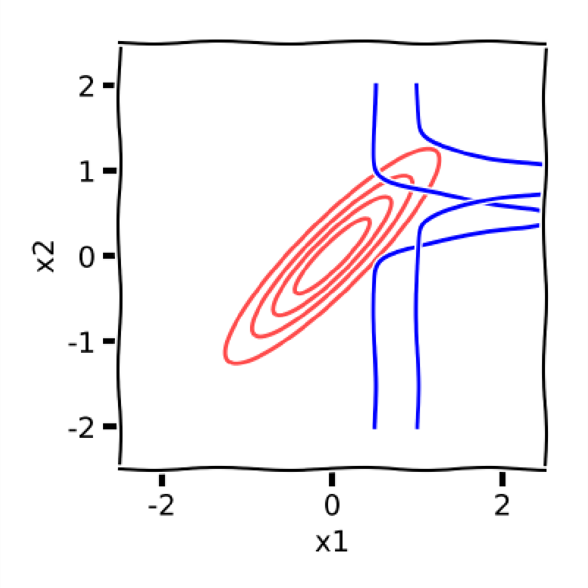
Conditional Gaussians



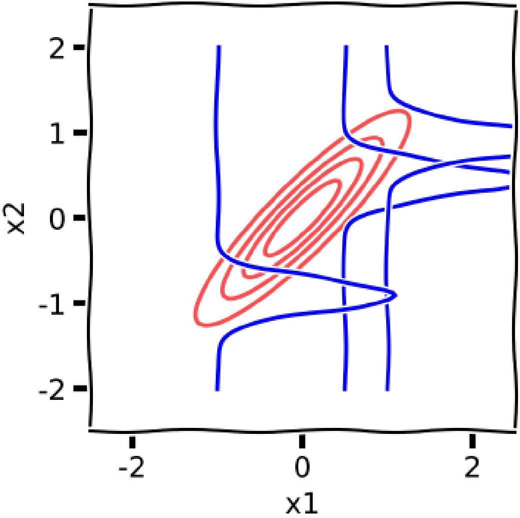
Conditional Gaussians



Conditional Gaussians



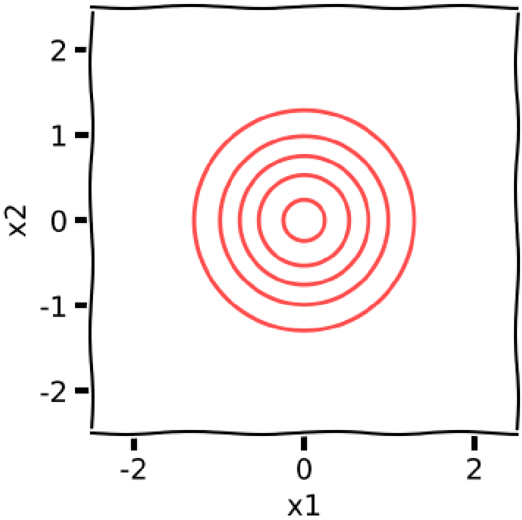
Conditional Gaussians



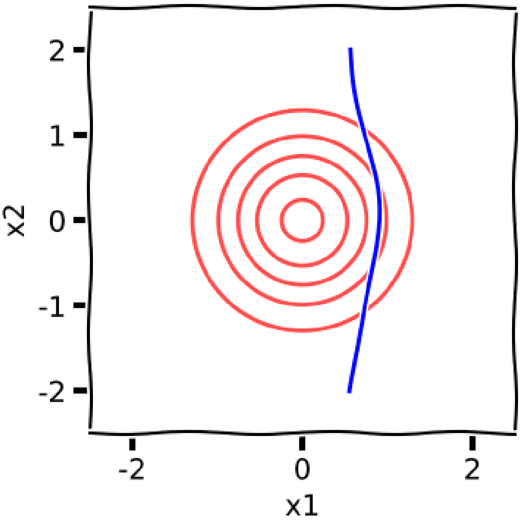
Conditional Gaussians

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

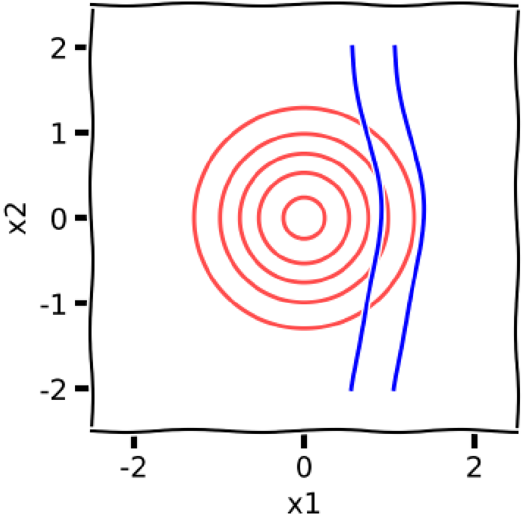
Conditional Gaussians



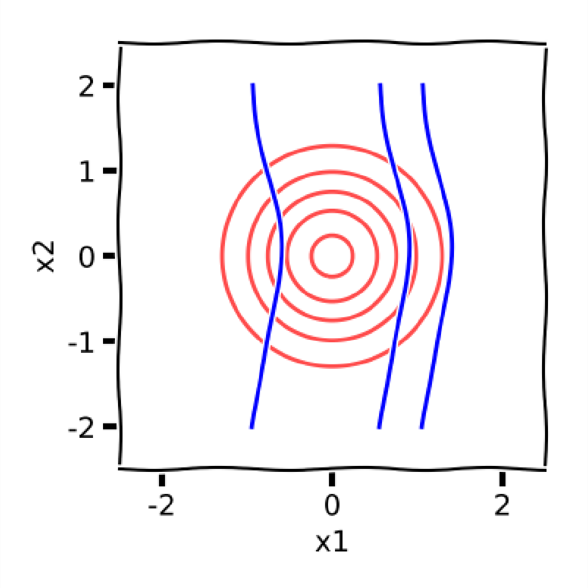
Conditional Gaussians



Conditional Gaussians

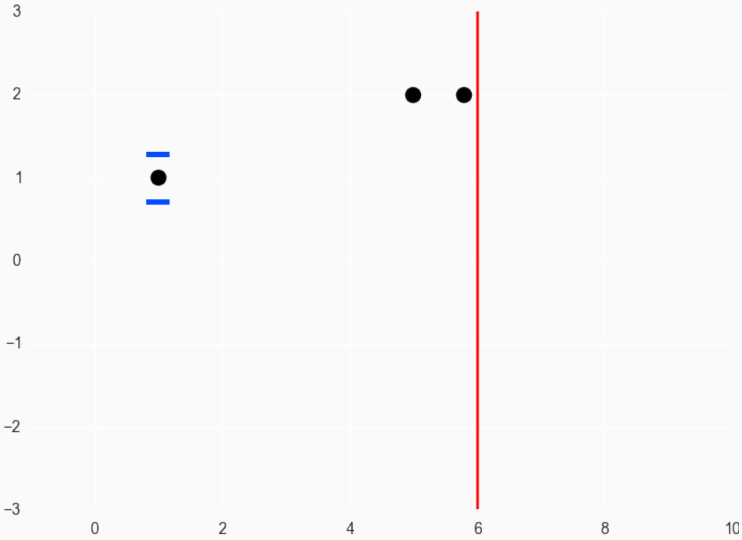


Conditional Gaussians

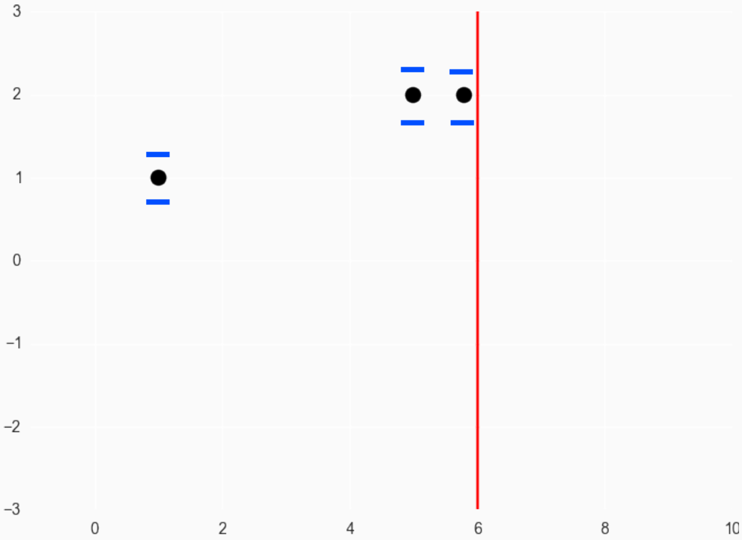


Eureka

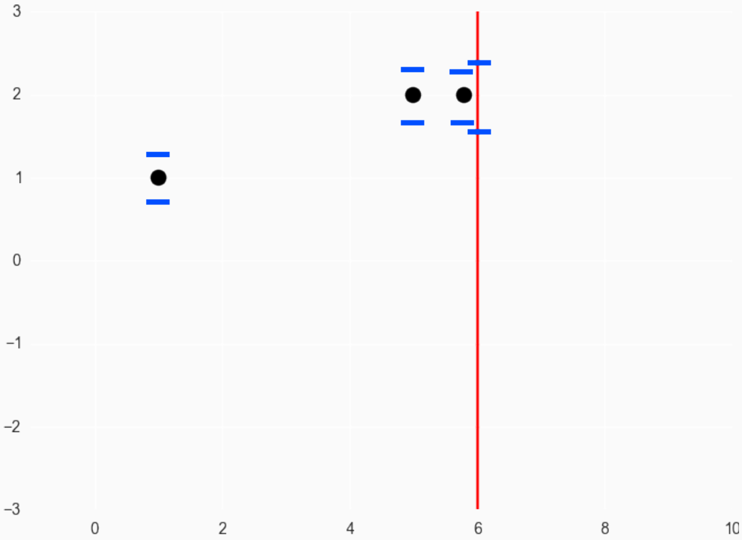
Functions



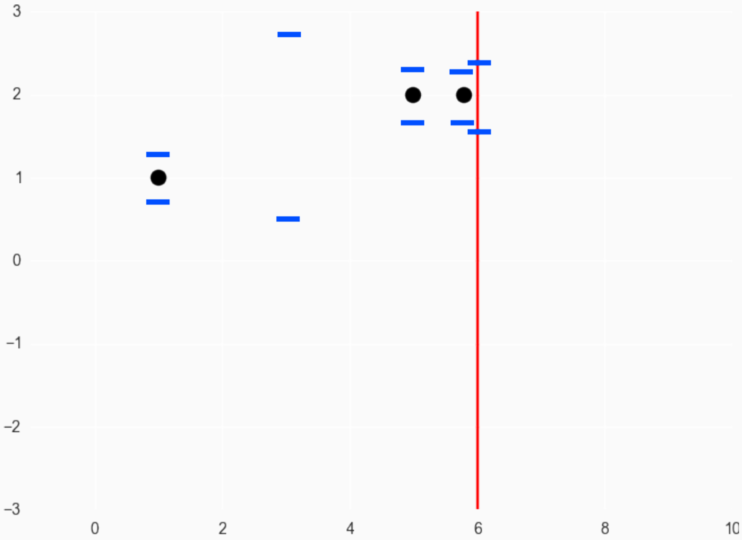
Functions



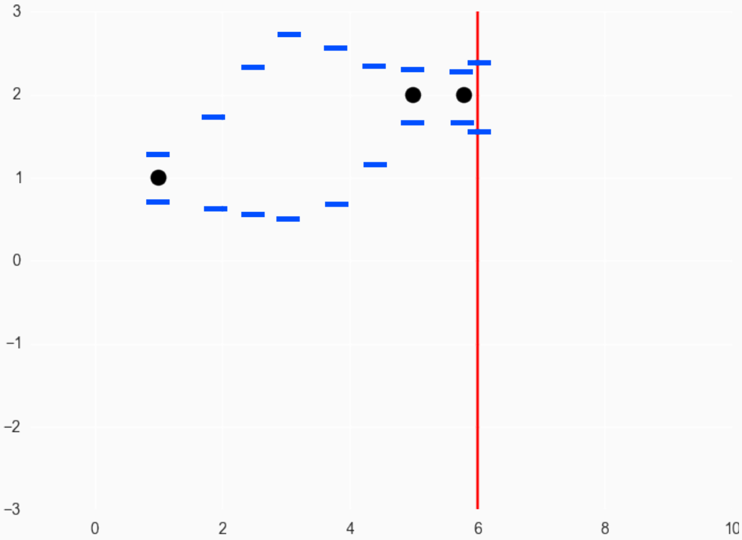
Functions



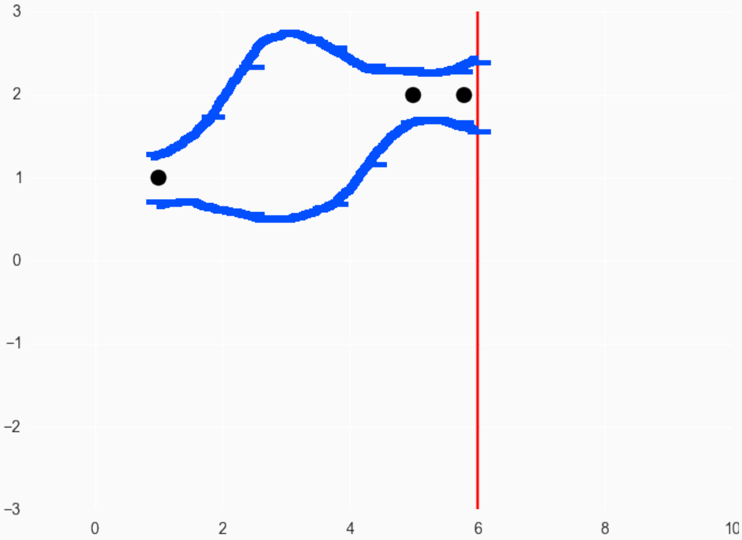
Functions



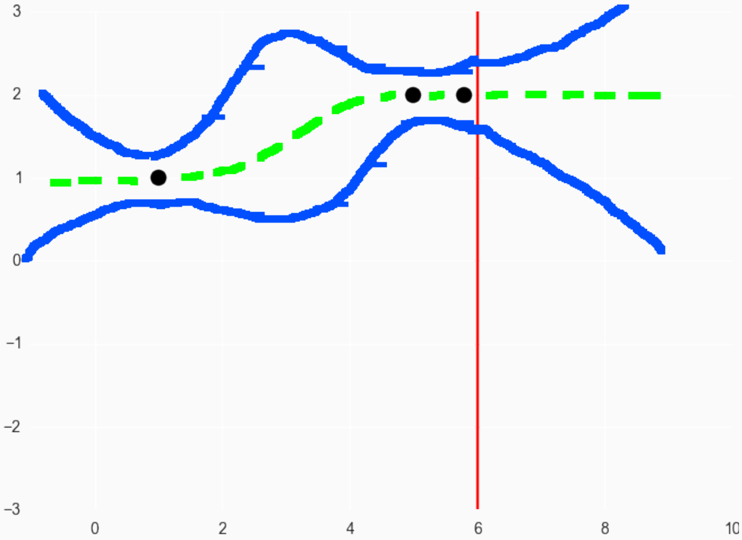
Functions



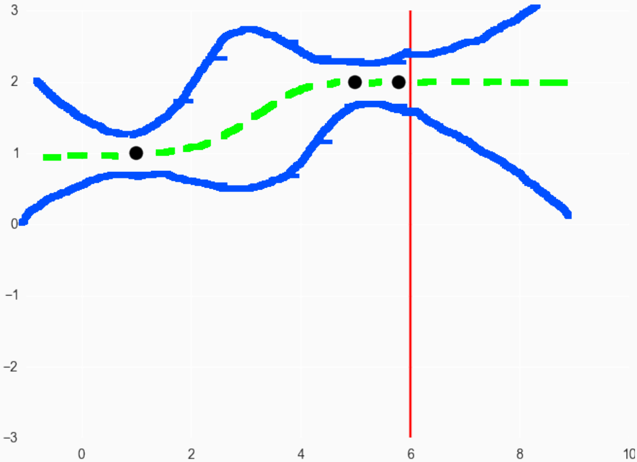
Functions



Functions



Functions



If all instantiations of the function is jointly Gaussian such that the co-variance structure depends on how much information an observation provides for the other we will get the curve above.

Uncertainty over functions

- Regression model,

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

- Introduce f_i as *instantiation* of function,

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

- as a new random variable.

Uncertainty over functions

- Regression model,

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

- Introduce f_i as *instantiation* of function,

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

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

Uncertainty over functions

Model,

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

Want to "push" $\mathbf{x}$ through a mapping f of which we are uncertain,

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

prior over instantiations of function.

Uncertainty over functions

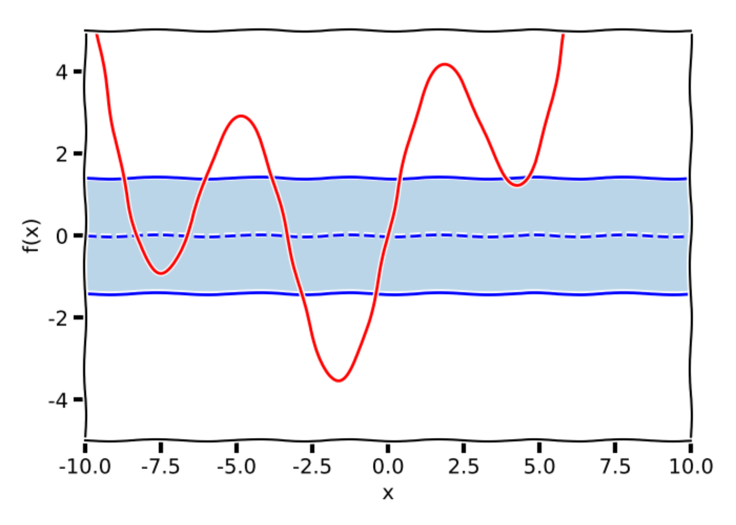
- As everything is gaussian both the marginal and predictive posterior are analytically tractable
- Marginal

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

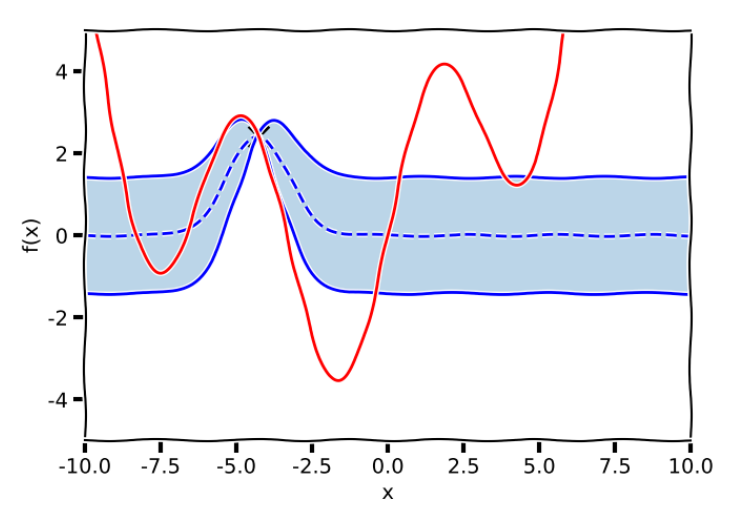
- Predictive posterior

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

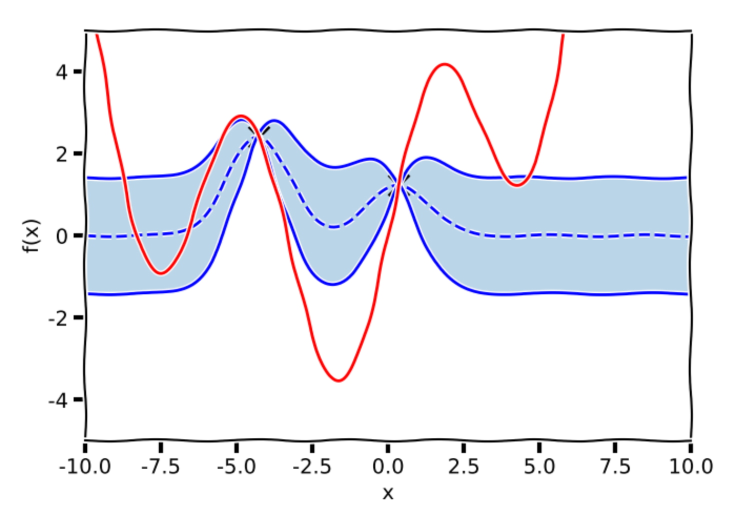
Gaussian Processes



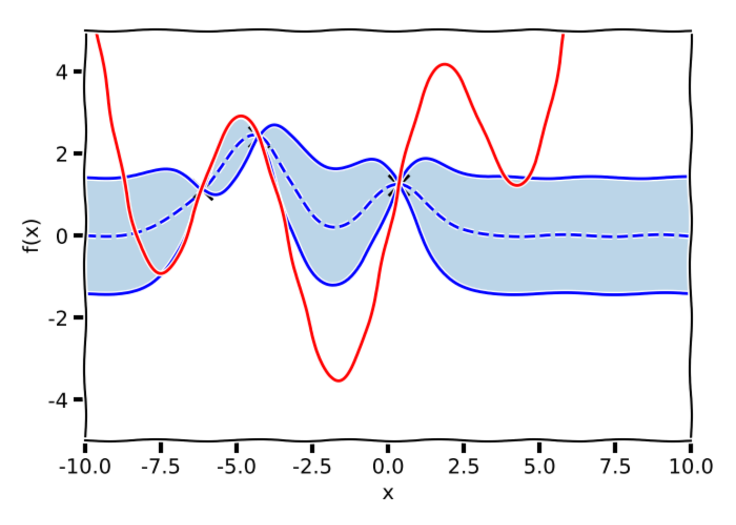
Gaussian Processes



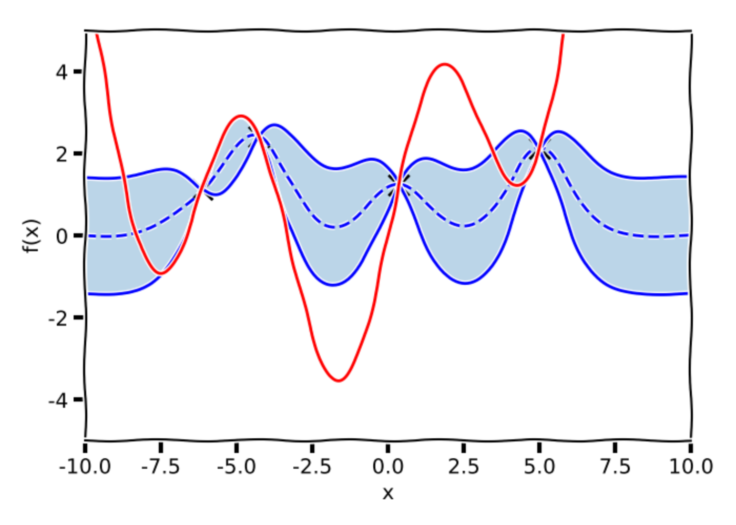
Gaussian Processes



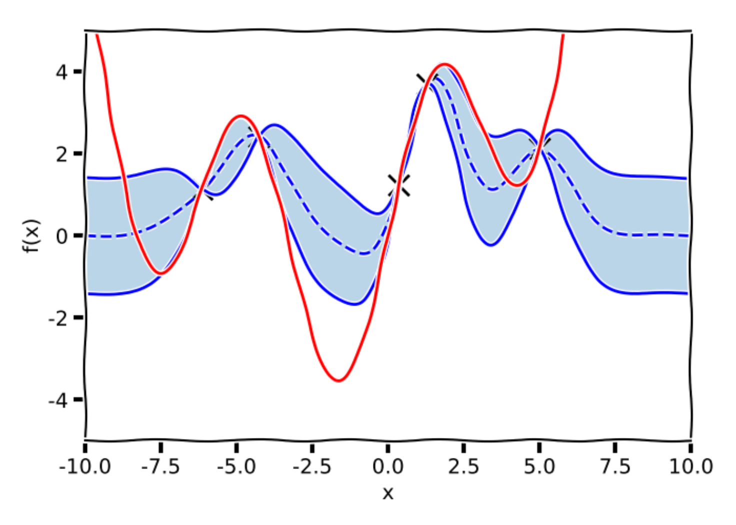
Gaussian Processes



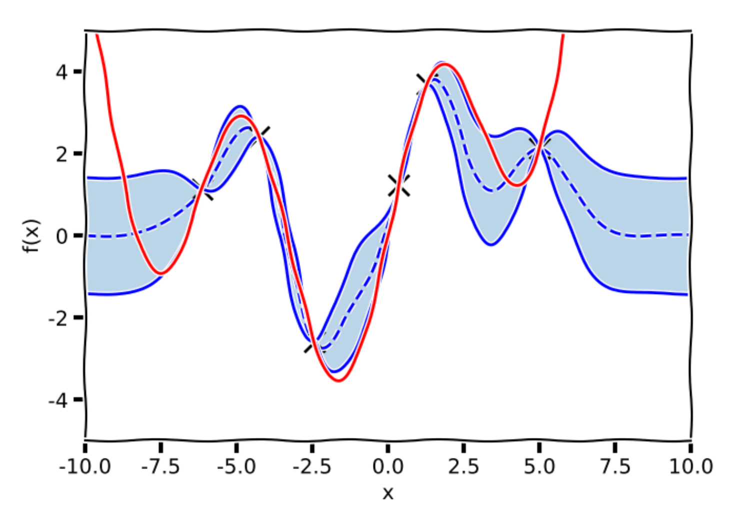
Gaussian Processes



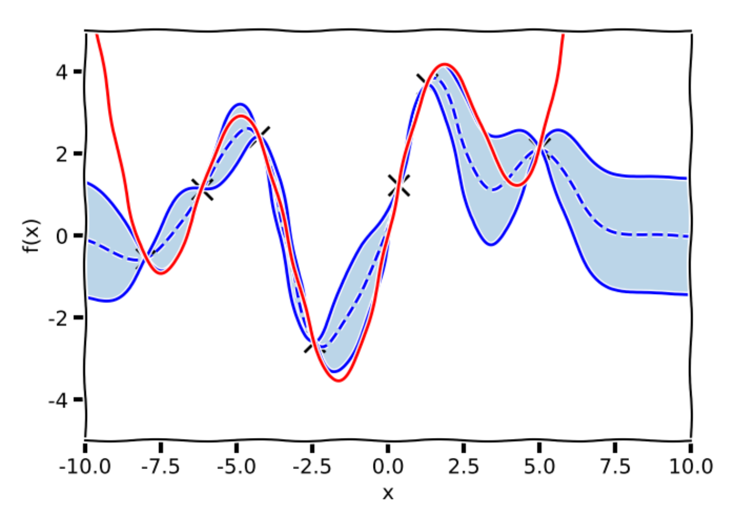
Gaussian Processes



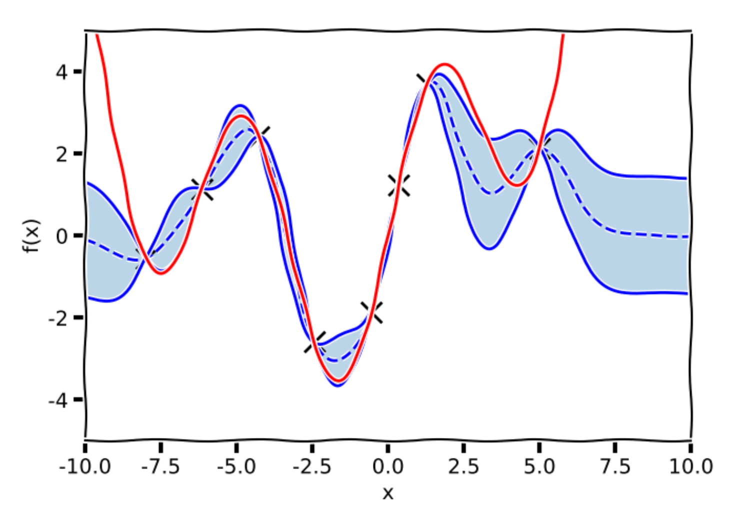
Gaussian Processes



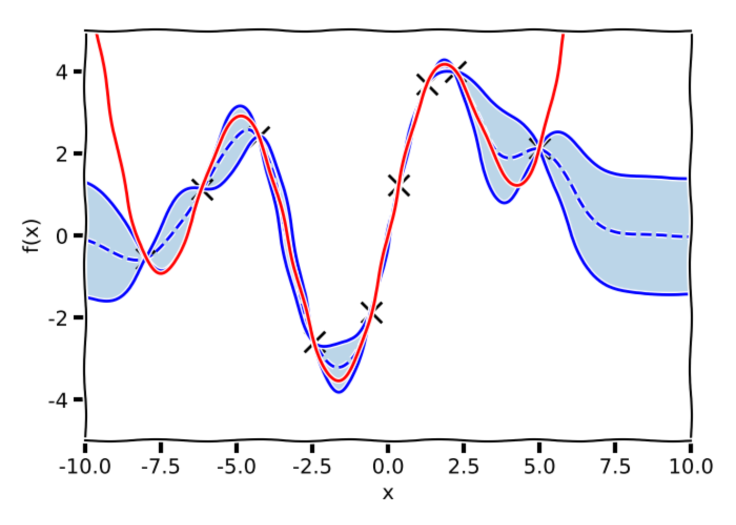
Gaussian Processes

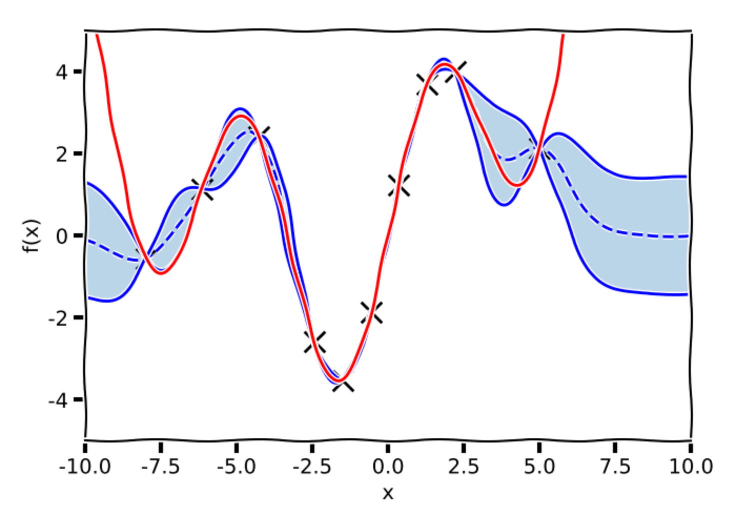


Gaussian Processes

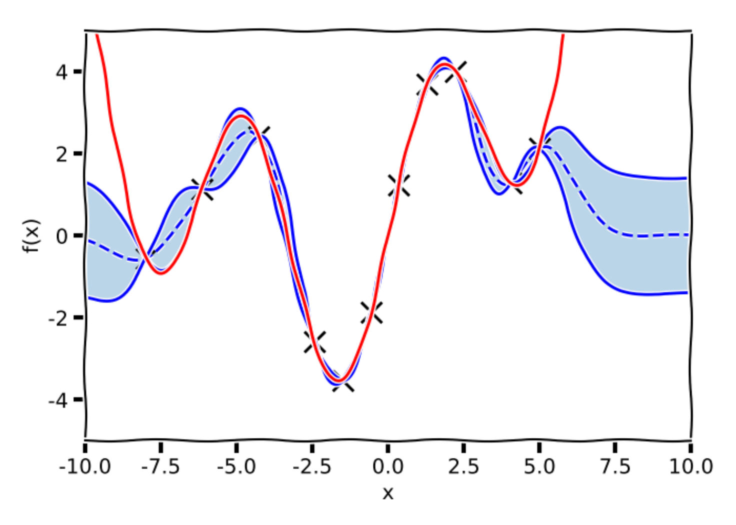


Gaussian Processes

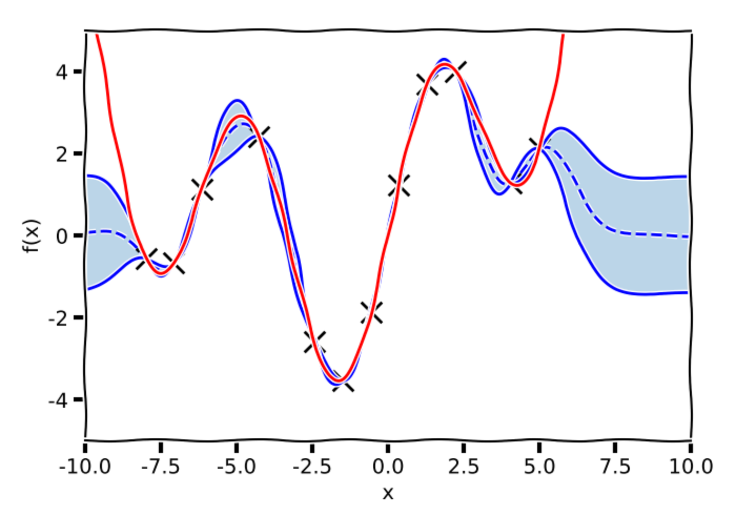


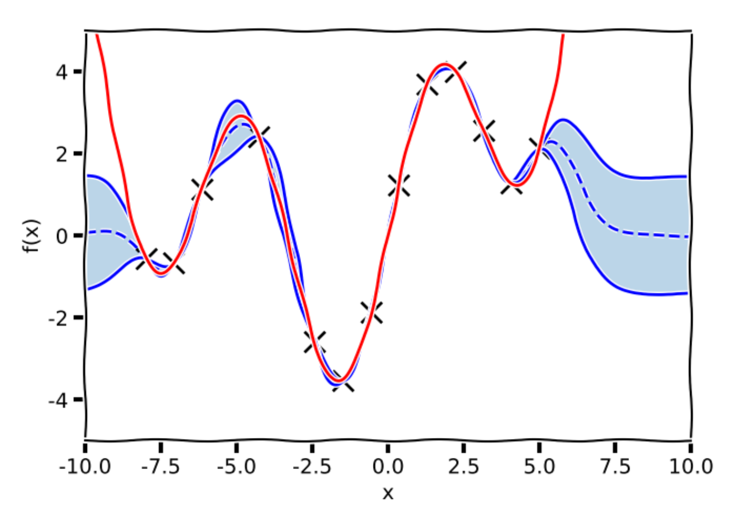

4
2
0
-2
-4
f(x)
-10.0
-7.5
-5.0
-2.5
0.0
2.5
5.0
7.5
10.0
x

Gaussian Processes



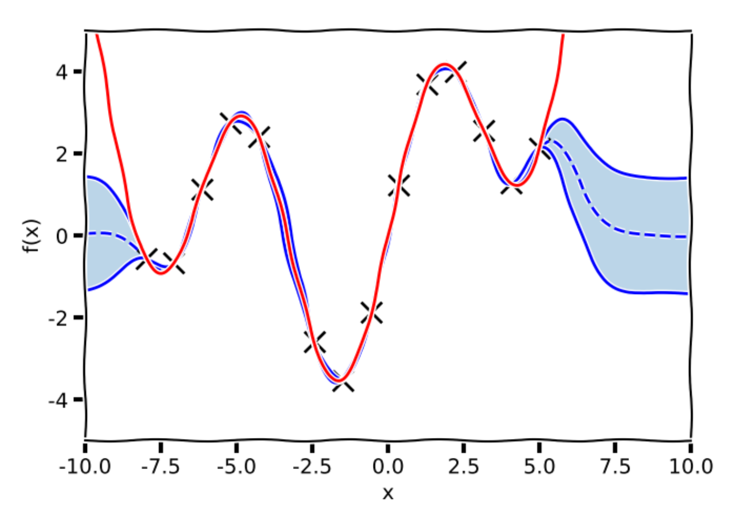
Gaussian Processes



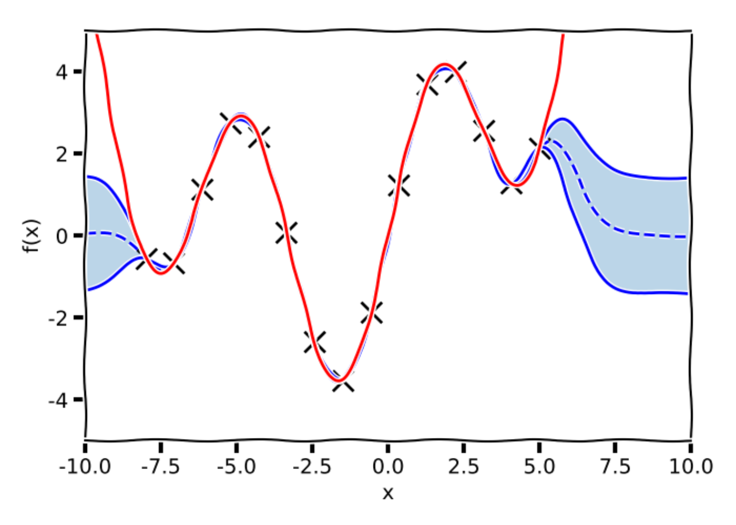


4
2
0
-2
-4
f(x)
-10.0
-7.5
-5.0
-2.5
0.0
2.5
5.0
7.5
10.0
x

Gaussian Processes



Gaussian Processes



Process $\rightarrow$ Distribution $\rightarrow$ value

- Each evaluation of a process is a distribution

$$\mathcal{N}(0, \Sigma) \sim \mathcal{N}(0, k(\mathbf{X}, \mathbf{X}))$$

Process $\rightarrow$ Distribution $\rightarrow$ value

- Each evaluation of a process is a distribution

$$\mathcal{N}(0, \Sigma) \sim \mathcal{N}(0, k(\mathbf{X}, \mathbf{X}))$$

- Each evaluation of a distribution is a value

$$y \sim \mathcal{N}(y|0, \Sigma)$$

Process $\rightarrow$ Distribution $\rightarrow$ value

- Each evaluation of a process is a distribution

$$\mathcal{N}(0, \Sigma) \sim \mathcal{N}(0, k(\mathbf{X}, \mathbf{X}))$$

- Each evaluation of a distribution is a value

$$y \sim \mathcal{N}(y|0, \Sigma)$$

- Kolmogrovs Existence Theorem defines which distributions have an infinite generalisation

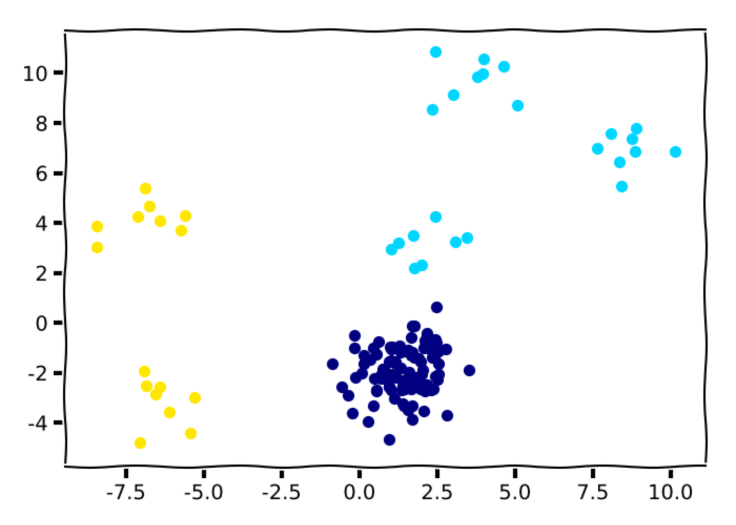
Gaussian Process

- Formulate process
- Evaluate process at specific location $x \rightarrow$ distribution
- Evaluate distribution at any location y
- GP is defined over uncountable infinite space

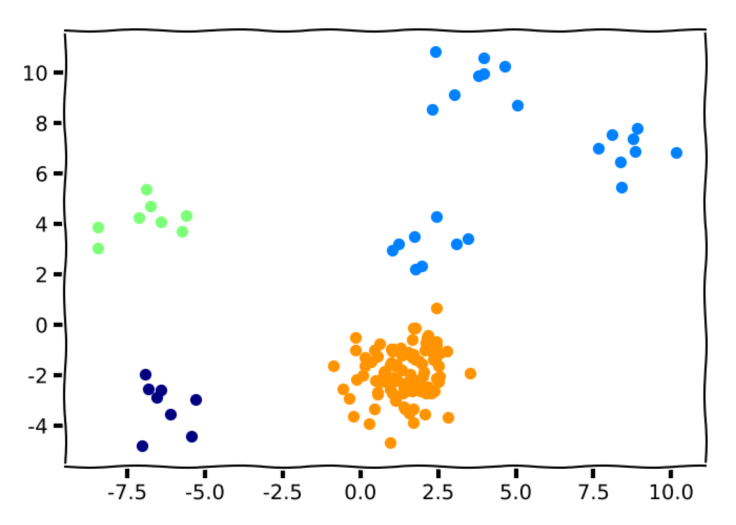
Gaussian Process

- Formulate process
- Evaluate process at specific location $x \rightarrow$ distribution
- Evaluate distribution at any location y
- GP is defined over uncountable infinite space
- *What about countable objects?*

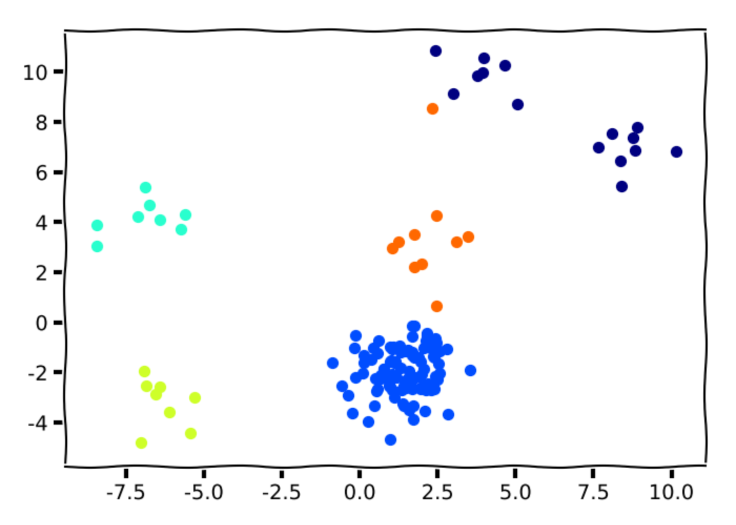
Gaussian Mixture Model



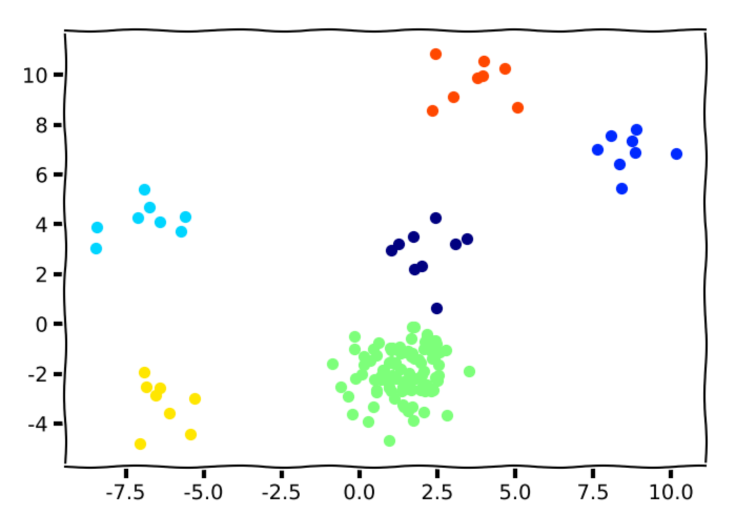
Gaussian Mixture Model



Gaussian Mixture Model



Gaussian Mixture Model

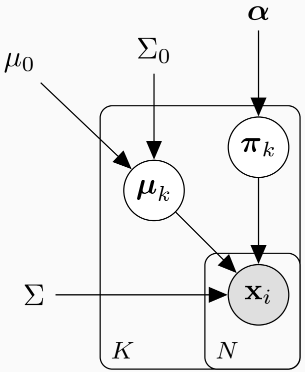


Gaussian Mixture Models

$$p(\mathbf{X}) = \sum_{k=1}^{K} p(\mathbf{X}|k)p(k) = \sum_{k=1}^{K} \mathcal{N}(\mathbf{X}|\boldsymbol{\mu}_k, \Sigma_k)p(k)$$

- Represent the probability of $\mathbf{X}$ as a combination or *mixture* of distributions
- What should K be?
- Can we make K infinite?

Gaussian Mixture Model



1. Sample proportions
2. Sample cluster id given proportions
3. Sample cluster mean
4. Sample data

$$p(\mathbf{X}) = \sum_{k=1}^{\infty} p(\mathbf{X}|k)p(k) = \sum_{k=1}^{\infty} \mathcal{N}(\mathbf{X}|\boldsymbol{\mu}_k, \Sigma_k)p(k)$$

Distributions over partitionings

- Multinomial

$$p(\mathsf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}$$

Distributions over partitionings

- Multinomial

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}$$

- Conjugate prior

$$p(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$$

Distributions over partitionings

- Multinomial

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}$$

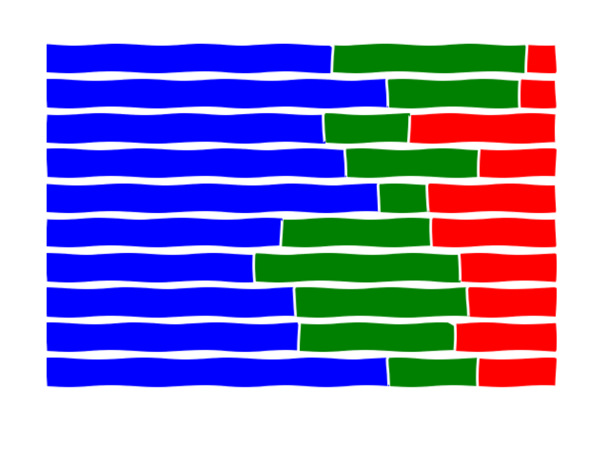
- Conjugate prior

$$p(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$$

- Dirichlet Distribution

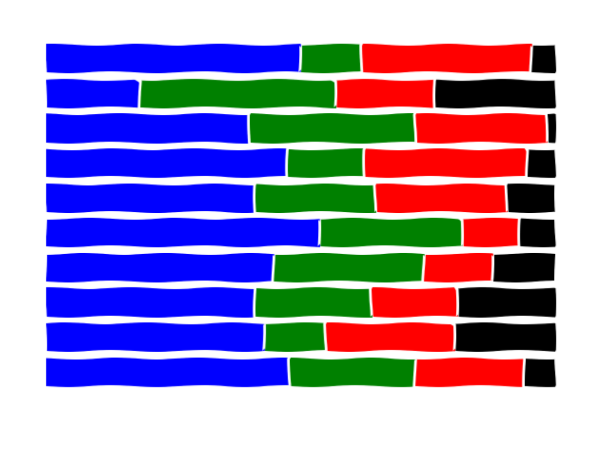
$$\mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \cdot \ldots \cdot \Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$$

Dirichlet Distribution



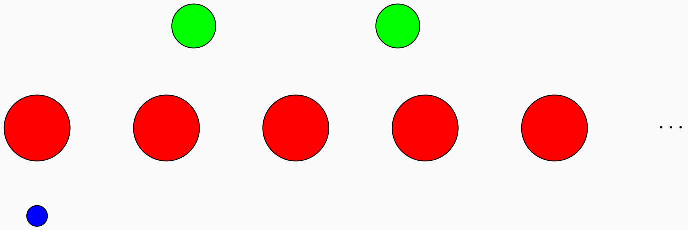
$\mathrm{Dir}(10, 5, 3)$

Dirichlet Distribution

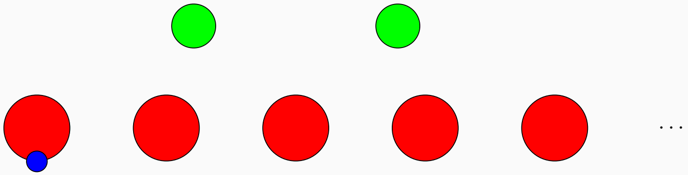


$\mathrm{Dir}(7, 5, 3, 2)$

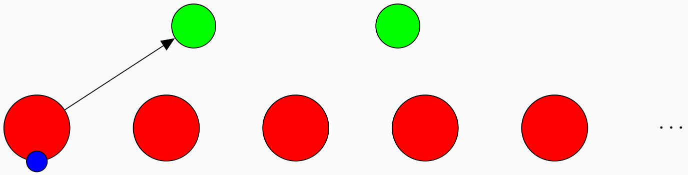
Chinese Resturant Process



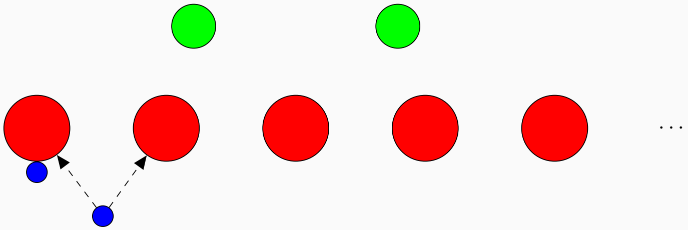
Chinese Resturant Process



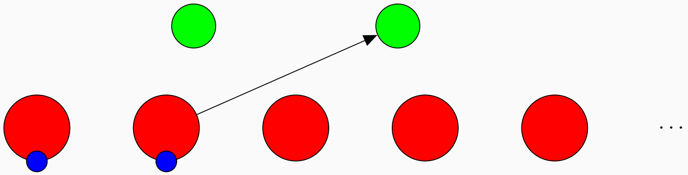
Chinese Resturant Process



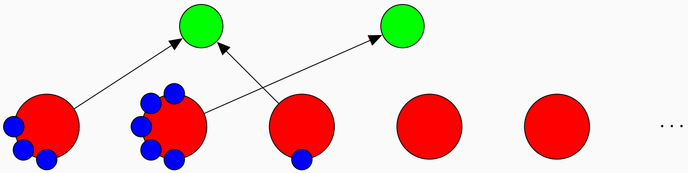
Chinese Resturant Process



Chinese Resturant Process

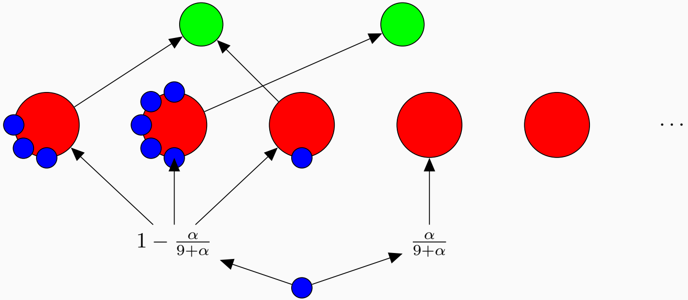


Chinese Resturant Process

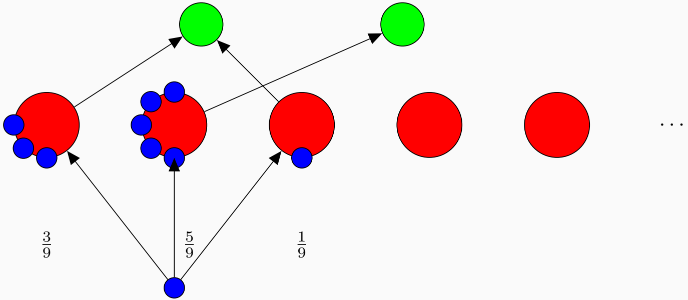


- Go to new table $\frac{\alpha}{N-1+\alpha}$
- If not choose table as $\frac{n_i}{N}$ where n_i number of diners at table *in*

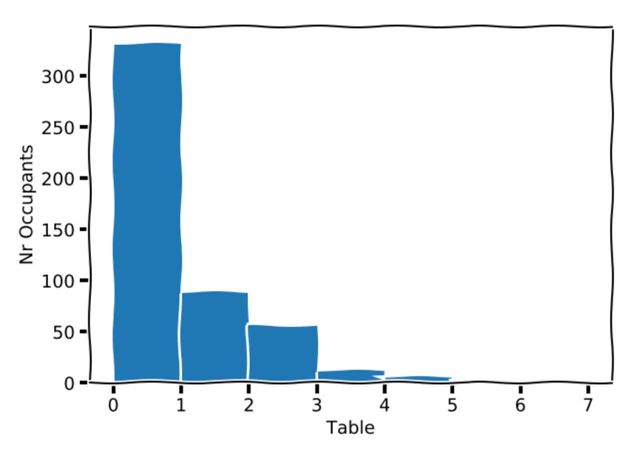
Chinese Resturant Process



Chinese Resturant Process

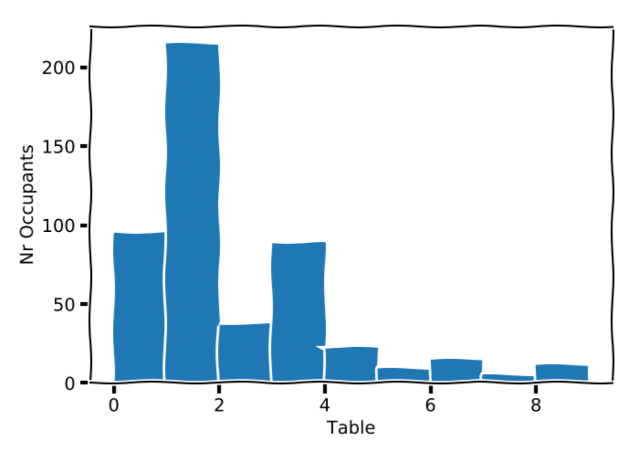


Chinese Resturant Process



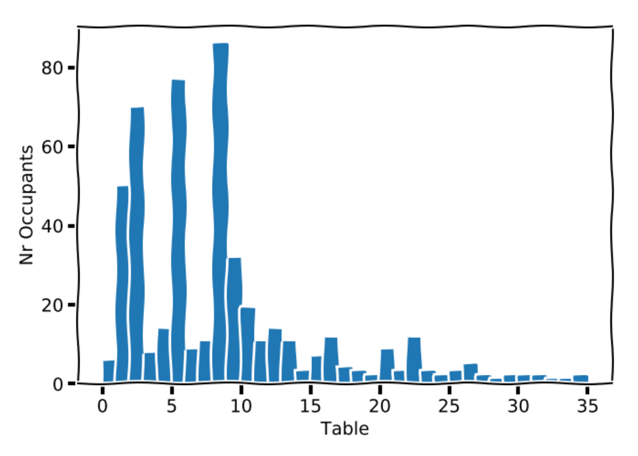
$$N = 500 \quad \alpha = 1.0$$

Chinese Resturant Process



$$N = 500 \quad \alpha = 2.0$$

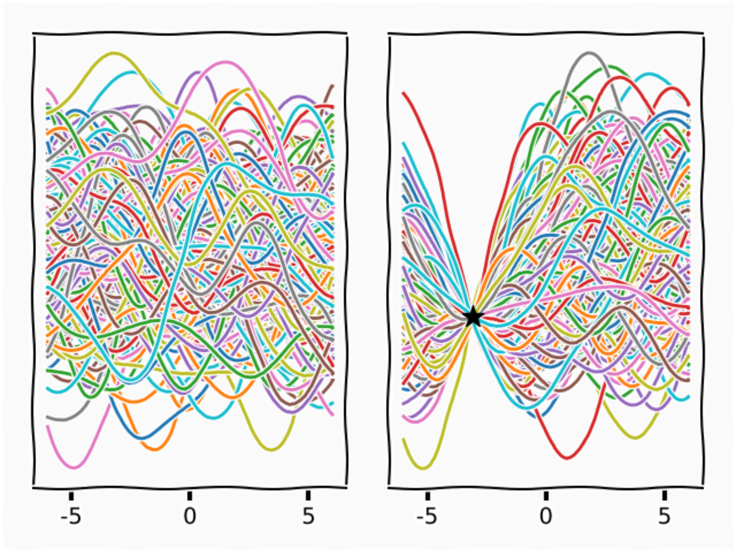
Chinese Resturant Process

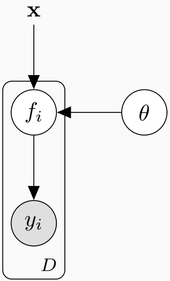


$$N = 500 \quad \alpha = 10.0$$

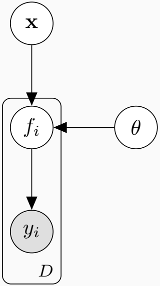
Unsupervised Learning

Gaussian Processes



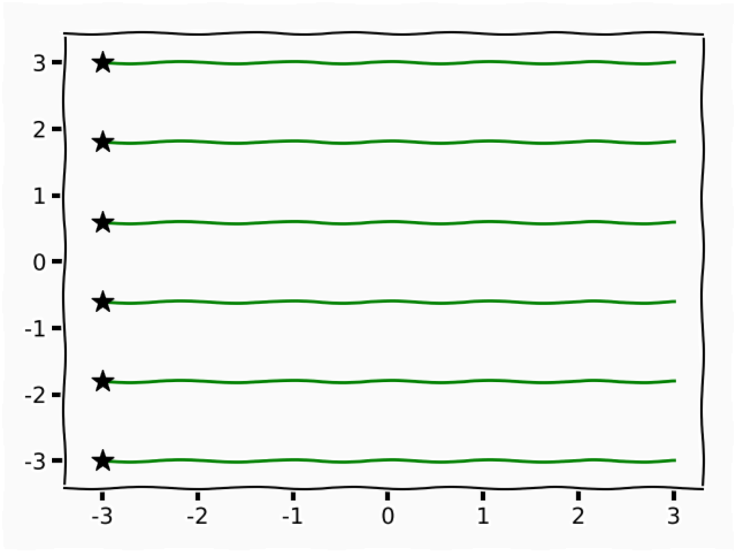


$p(y|x)$

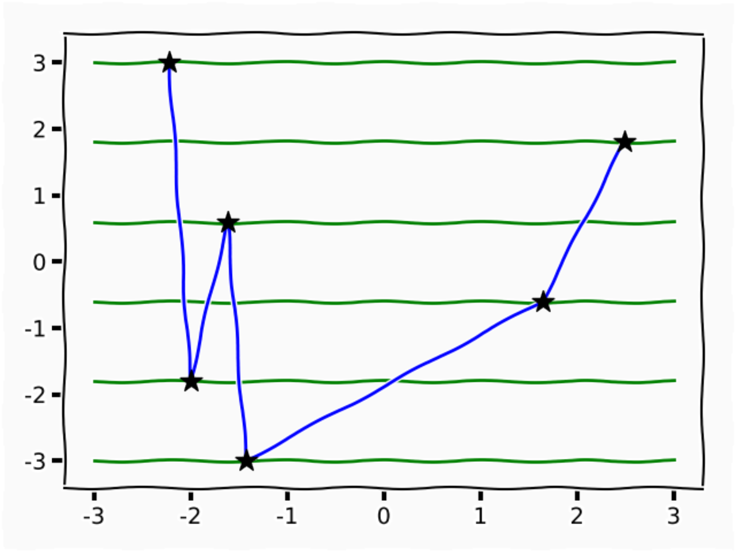


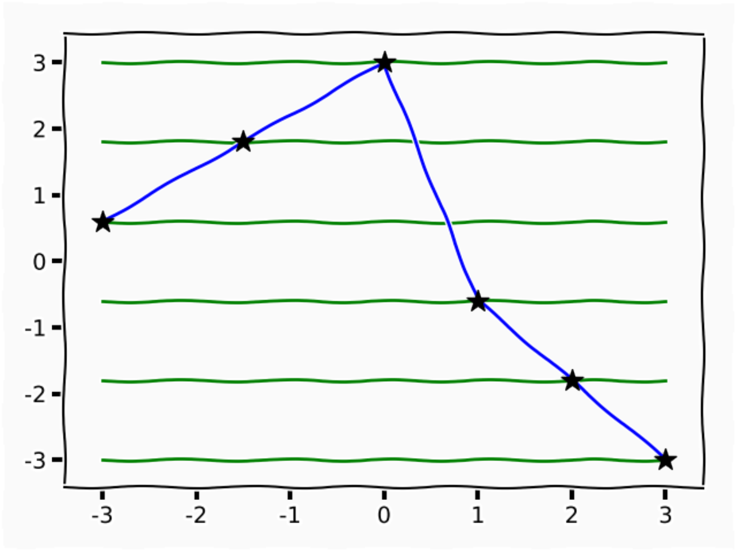
$p(y)$

Unsupervised Learning

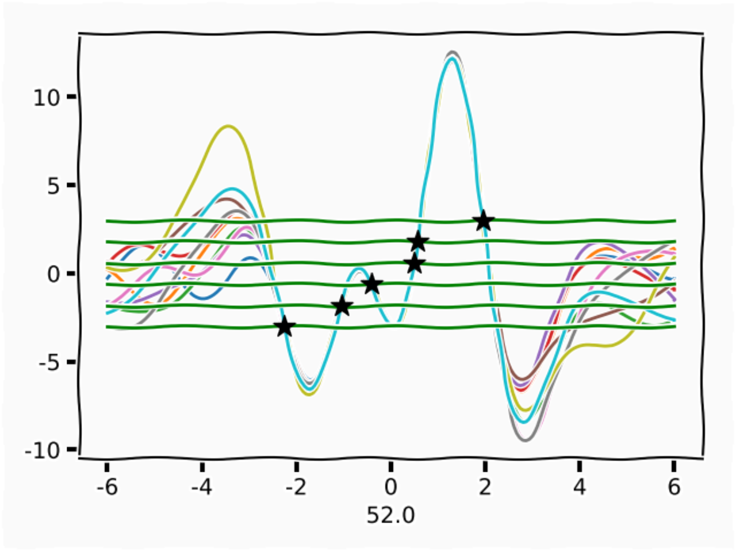


Unsupervised Learning

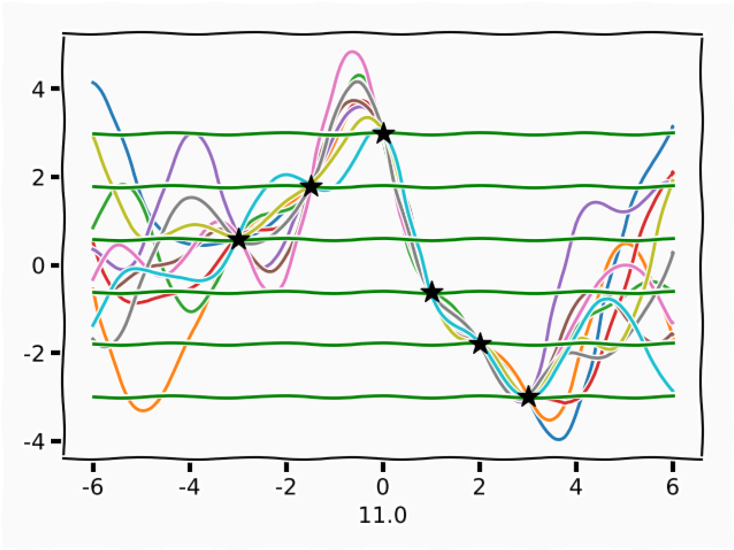


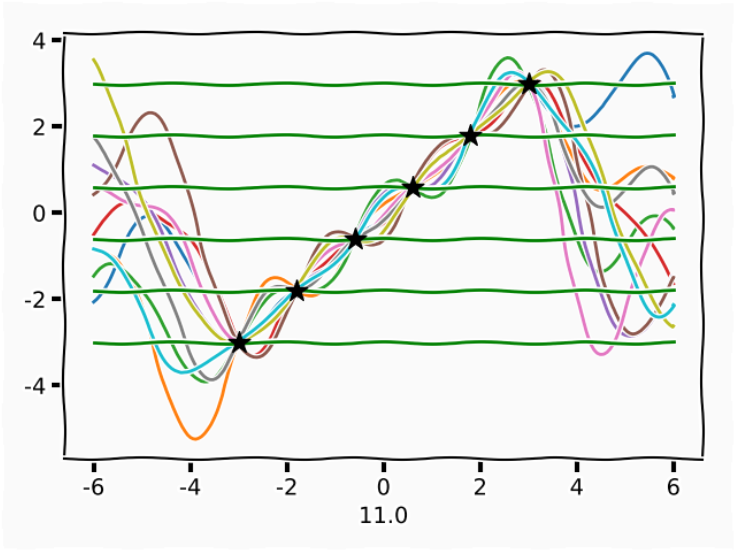

3
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3

Unsupervised Learning

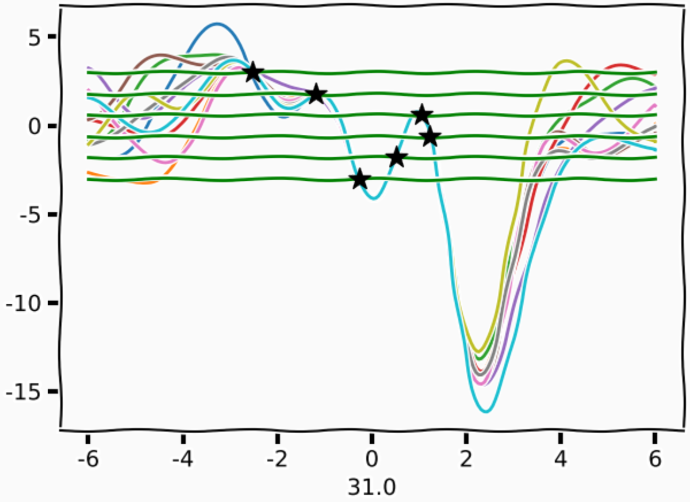


Unsupervised Learning




4
2
0
-2
-4
-6
-4
-2
0
2
4
6
11.0

Unsupervised Learning



Priors

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

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

1. Priors that makes sense

   p(f) describes our belief/assumptions and defines our notion of complexity in the function

   p(x) expresses our belief/assumptions and defines our notion of complexity in the latent space

2. The priors are *"balanced"*
3. Now lets churn the handle

Relationship between x and data

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

- GP prior

$$p(f|x) \sim \mathcal{N}(0, K) \propto e^{-\frac{1}{2}(f^{\mathrm{T}}K^{-1}f)}$$

$$K_{ij} = e^{-(x_i - x_j)^{\mathrm{T}}M^{\mathrm{T}}M(x_i - x_j)}$$

Relationship between x and data

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

- GP prior

$$p(f|x) \sim \mathcal{N}(0, K) \propto e^{-\frac{1}{2}(f^{\mathrm{T}}K^{-1}f)}$$

$$K_{ij} = e^{-(x_i - x_j)^{\mathrm{T}}M^{\mathrm{T}}M(x_i - x_j)}$$

- Likelihood

$$p(y|f) \sim N(y|f, \beta) \propto e^{-\frac{1}{2\beta}\mathrm{tr}(y-f)^{\mathrm{T}}(y-f)}$$

Relationship between x and data

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

- GP prior

$$p(f|x) \sim \mathcal{N}(0, K) \propto e^{-\frac{1}{2}(f^{\mathrm{T}}K^{-1}f)}$$

$$K_{ij} = e^{-(x_i - x_j)^{\mathrm{T}}M^{\mathrm{T}}M(x_i - x_j)}$$

- Likelihood

$$p(y|f) \sim N(y|f, \beta) \propto e^{-\frac{1}{2\beta}\mathrm{tr}(y-f)^{\mathrm{T}}(y-f)}$$

- Analytically intractable (`Non Elementary Integral`) and infinitely differientiable

Laplace Integration

"Nature laughs at the difficulties of integrations"
– Simon Laplace

Unsupervised Learning with GPs

Variational Bayes

$$p(\mathbf{Y})$$

Variational Bayes

$$\log p(\mathbf{Y})$$

Variational Bayes

$$\log p(\mathbf{Y}) = \log \int p(\mathbf{Y}, \mathbf{X}) \mathrm{d}\mathbf{X}$$

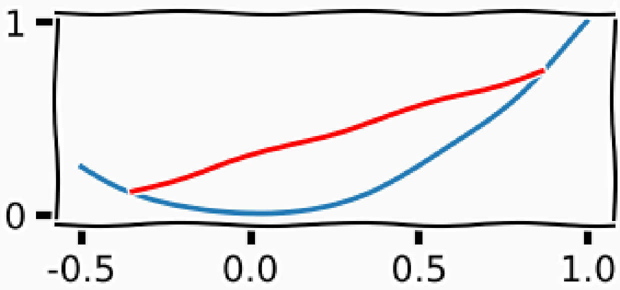
Variational Bayes

$$\log\, p(\mathbf{Y}) = \log \int p(\mathbf{Y}, \mathbf{X}) \mathrm{d}\mathbf{X} = \log \int p(\mathbf{X}|\mathbf{Y}) p(\mathbf{Y}) \mathrm{d}\mathbf{X}$$

Variational Bayes

$$
\begin{aligned}
\log\, p(\mathbf{Y}) &= \log \int p(\mathbf{Y}, \mathbf{X}) \mathrm{d}\mathbf{X} = \log \int p(\mathbf{X}|\mathbf{Y}) p(\mathbf{Y}) \mathrm{d}\mathbf{X} \\
&= \log \int \frac{q(\mathbf{X})}{q(\mathbf{X})} p(\mathbf{X}|\mathbf{Y}) p(\mathbf{Y}) \mathrm{d}\mathbf{X}
\end{aligned}
$$

Jensen Inequality



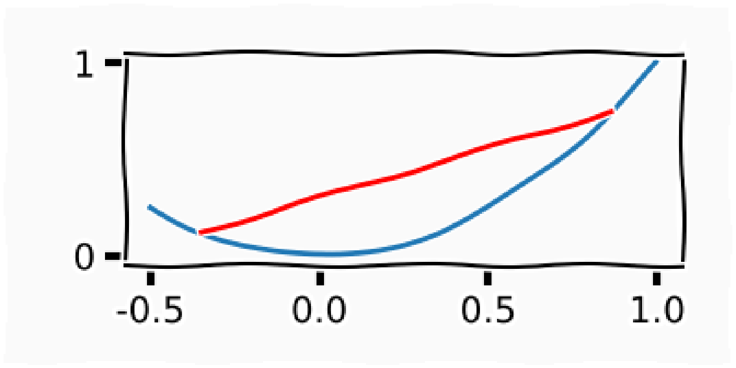
Convex Function

$$\lambda f(x_0) + (1-\lambda) f(x_1) \geq f(\lambda x_0 + (1-\lambda) x_1)$$

$$x \in [x_{min}, x_{max}]$$

$$\lambda \in [0, 1]]$$

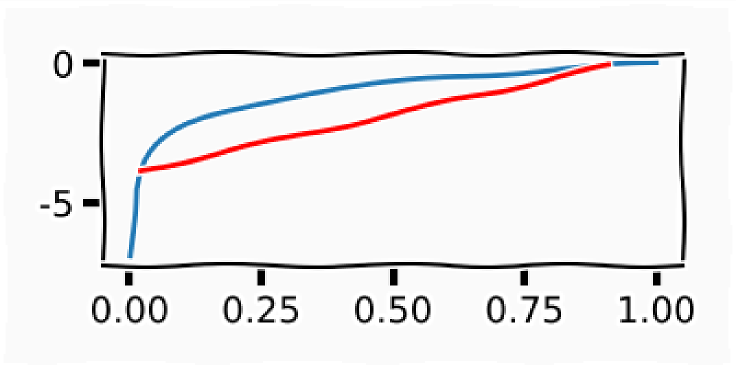
Jensen Inequality



$$\mathbb{E}[f(x)] \geq f(\mathbb{E}[x])$$

$$\int f(x)p(x)\mathrm{d}x \geq f\left(\int xp(x)\mathrm{d}x\right)$$

Jensen Inequality in Variational Bayes



$$\int \log(x)p(x)\mathrm{d}x \leq \log\left(\int xp(x)\mathrm{d}x\right)$$

moving the log inside the the integral is a lower-bound on the integral

Variational Bayes cont.

$$\log p(\mathbf{Y}) = \log \int \frac{q(\mathbf{X})}{q(\mathbf{X})} p(\mathbf{X}|\mathbf{Y}) p(\mathbf{Y}) \mathrm{d}\mathbf{X} =$$

Variational Bayes cont.

$$\log p(\mathbf{Y}) = \log \int \frac{q(\mathbf{X})}{q(\mathbf{X})} p(\mathbf{X}|\mathbf{Y}) p(\mathbf{Y}) \mathrm{d}\mathbf{X} =$$

$$\geq \int q(\mathbf{X}) \log \frac{p(\mathbf{X}|\mathbf{Y}) p(\mathbf{Y})}{q(\mathbf{X})} \mathrm{d}\mathbf{X}$$

Variational Bayes cont.

$$
\begin{aligned}
\log p(\mathbf{Y}) &= \log \int \frac{q(\mathbf{X})}{q(\mathbf{X})} p(\mathbf{X}|\mathbf{Y}) p(\mathbf{Y}) \mathrm{d}\mathbf{X} = \\
&\geq \int q(\mathbf{X}) \log \frac{p(\mathbf{X}|\mathbf{Y}) p(\mathbf{Y})}{q(\mathbf{X})} \mathrm{d}\mathbf{X} \\
&= \int q(\mathbf{X}) \log \frac{p(\mathbf{X}|\mathbf{Y})}{q(\mathbf{X})} \mathrm{d}\mathbf{X} + \int q(\mathbf{X}) \mathrm{d}\mathbf{X} \log\, p(\mathbf{Y})
\end{aligned}
$$

Variational Bayes cont.

$$
\begin{aligned}
\log p(\mathbf{Y}) &= \log \int \frac{q(\mathbf{X})}{q(\mathbf{X})} p(\mathbf{X}|\mathbf{Y}) p(\mathbf{Y}) \mathrm{d}\mathbf{X} = \\
&\geq \int q(\mathbf{X}) \log \frac{p(\mathbf{X}|\mathbf{Y}) p(\mathbf{Y})}{q(\mathbf{X})} \mathrm{d}\mathbf{X} \\
&= \int q(\mathbf{X}) \log \frac{p(\mathbf{X}|\mathbf{Y})}{q(\mathbf{X})} \mathrm{d}\mathbf{X} + \int q(\mathbf{X}) \mathrm{d}\mathbf{X} \log\, p(\mathbf{Y}) \\
&= -\mathrm{KL}\left(q(\mathbf{X}) || p(\mathbf{X}|\mathbf{Y})\right) + \log\, p(\mathbf{Y})
\end{aligned}
$$

Variational Bayes cont.

$$
\begin{aligned}
\log p(\mathbf{Y}) &= \log \int \frac{q(\mathbf{X})}{q(\mathbf{X})} p(\mathbf{X}|\mathbf{Y}) p(\mathbf{Y}) \mathrm{d}\mathbf{X} = \\
&\geq \int q(\mathbf{X}) \log \frac{p(\mathbf{X}|\mathbf{Y}) p(\mathbf{Y})}{q(\mathbf{X})} \mathrm{d}\mathbf{X} \\
&= \int q(\mathbf{X}) \log \frac{p(\mathbf{X}|\mathbf{Y})}{q(\mathbf{X})} \mathrm{d}\mathbf{X} + \int q(\mathbf{X}) \mathrm{d}\mathbf{X} \log\, p(\mathbf{Y}) \\
&= -\mathrm{KL}\left(q(\mathbf{X}) || p(\mathbf{X}|\mathbf{Y})\right) + \log\, p(\mathbf{Y})
\end{aligned}
$$

- if $q(\mathbf{X})$ is the true posterior we have an equality, therefore match the distributions

Variational Bayes cont.

$$
\begin{aligned}
\log p(\mathbf{Y}) &= \log \int \frac{q(\mathbf{X})}{q(\mathbf{X})} p(\mathbf{X}|\mathbf{Y}) p(\mathbf{Y}) \mathrm{d}\mathbf{X} = \\
&\geq \int q(\mathbf{X}) \log \frac{p(\mathbf{X}|\mathbf{Y}) p(\mathbf{Y})}{q(\mathbf{X})} \mathrm{d}\mathbf{X} \\
&= \int q(\mathbf{X}) \log \frac{p(\mathbf{X}|\mathbf{Y})}{q(\mathbf{X})} \mathrm{d}\mathbf{X} + \int q(\mathbf{X}) \mathrm{d}\mathbf{X} \log\, p(\mathbf{Y}) \\
&= -\mathrm{KL}\left(q(\mathbf{X}) || p(\mathbf{X}|\mathbf{Y})\right) + \log\, p(\mathbf{Y})
\end{aligned}
$$

- if $q(\mathbf{X})$ is the true posterior we have an equality, therefore match the distributions
- i.e. $\mathrm{argmin}_q\ \mathrm{KL}\left(q(\mathbf{X}) || p(\mathbf{X}|\mathbf{Y})\right)$

  $\Rightarrow$ variational distributions are approximations to intractable posteriors

ELBO

$$\mathrm{KL}(q(\mathbf{X})||p(\mathbf{X}|\mathbf{Y}))$$

ELBO

$$\mathrm{KL}(q(\mathbf{X})||p(\mathbf{X}|\mathbf{Y})) = \int q(\mathbf{X}) log \frac{q(\mathbf{X})}{p(\mathbf{X}|\mathbf{Y})} \mathrm{d}\mathbf{X}$$

ELBO

$$\begin{aligned}\mathrm{KL}(q(\mathbf{X})||p(\mathbf{X}|\mathbf{Y})) &= \int q(\mathbf{X}) log \frac{q(\mathbf{X})}{p(\mathbf{X}|\mathbf{Y})} \mathrm{d}\mathbf{X} \\ &= \int q(\mathbf{X}) log \frac{q(\mathbf{X})}{p(\mathbf{X},\mathbf{Y})} \mathrm{d}\mathbf{X} + log \ \ p(\mathbf{Y})\end{aligned}$$

ELBO

$$
\begin{aligned}
\mathrm{KL}(q(\mathbf{X}) || p(\mathbf{X}|\mathbf{Y})) &= \int q(\mathbf{X}) log \frac{q(\mathbf{X})}{p(\mathbf{X}|\mathbf{Y})} \mathrm{d}\mathbf{X} \\
&= \int q(\mathbf{X}) log \frac{q(\mathbf{X})}{p(\mathbf{X}, \mathbf{Y})} \mathrm{d}\mathbf{X} + log \ p(\mathbf{Y}) \\
&= H(q(\mathbf{X})) - \mathbb{E}_{q(\mathbf{X})} \left[log \ p(\mathbf{X}, \mathbf{Y}) \right] + log \ p(\mathbf{Y})
\end{aligned}
$$

ELBO

$$\mathit{log}\ p(\mathbf{Y}) = \mathrm{KL}(q(\mathbf{X})||p(\mathbf{X}|\mathbf{Y})) + \underbrace{\mathbb{E}_{q(\mathbf{X})}\left[\mathit{log}\ p(\mathbf{X},\mathbf{Y})\right] - H(q(\mathbf{X}))}_{\mathrm{ELBO}}$$

ELBO

$$\log\ p(\mathbf{Y}) = \mathrm{KL}(q(\mathbf{X})||p(\mathbf{X}|\mathbf{Y})) + \underbrace{\mathbb{E}_{q(\mathbf{X})}\left[\log\ p(\mathbf{X},\mathbf{Y})\right] - H(q(\mathbf{X}))}_{\mathrm{ELBO}}$$

$$\geq \mathbb{E}_{q(\mathbf{X})}\left[\log\ p(\mathbf{X},\mathbf{Y})\right] - H(q(\mathbf{X})) = \mathcal{L}(q(\mathbf{X}))$$

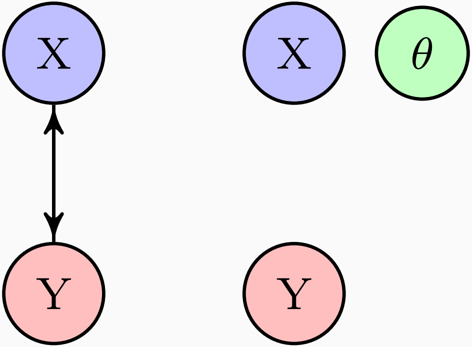
ELBO

$$\mathit{log}\ p(\mathbf{Y}) = \mathrm{KL}(q(\mathbf{X})||p(\mathbf{X}|\mathbf{Y})) + \underbrace{\mathbb{E}_{q(\mathbf{X})}\left[\mathit{log}\ p(\mathbf{X},\mathbf{Y})\right] - H(q(\mathbf{X}))}_{\mathrm{ELBO}}$$

$$\geq \mathbb{E}_{q(\mathbf{X})}\left[\mathit{log}\ p(\mathbf{X},\mathbf{Y})\right] - H(q(\mathbf{X})) = \mathcal{L}(q(\mathbf{X}))$$

- if we maximise the ELBO we,
    - find an approximate posterior
    - get an approximation to the marginal likelihood
- *maximising* $p(\mathbf{Y})$ is learning
- finding $p(\mathbf{X}|\mathbf{Y}) \approx q(\mathbf{X})$ is prediction

ELBO



Why is this useful?

Why is this a sensible thing to do?

- If we can't formulate the joint distribution there isn't much we can do

– Ryan Adams[3]

[3] `Talking Machines Season 2, Episode 5`

Why is this useful?

Why is this a sensible thing to do?

- If we can't formulate the joint distribution there isn't much we can do
- Taking the expectation of a log is usually easier than the expectation

– Ryan Adams[3]

[3] Talking Machines Season 2, Episode 5

Why is this useful?

Why is this a sensible thing to do?

- If we can't formulate the joint distribution there isn't much we can do
- Taking the expectation of a log is usually easier than the expectation
- We are allowed to choose the distribution to take the expectation over

– Ryan Adams[3]

[3] `Talking Machines Season 2, Episode 5`

Lower Bound[4]

$$\mathcal{L} = \int_{\mathbf{X},\mathbf{F}} q(\mathbf{X}) \log \left(\frac{p(\mathbf{Y}, \mathbf{F}, \mathbf{X})}{q(\mathbf{X})} \right)$$

[4] Damianou, A. C. (2015). Deep Gaussian Processes and Variational Propagation of Uncertainty (Doctoral dissertation)

Lower Bound[4]

$$\mathcal{L} = \int_{\mathbf{X},\mathbf{F}} q(\mathbf{X}) \log \left(\frac{p(\mathbf{Y}, \mathbf{F}, \mathbf{X})}{q(\mathbf{X})} \right)$$

$$\int_{\mathbf{X},\mathbf{F}} q(\mathbf{X}) \log \left(\frac{p(\mathbf{Y}|\mathbf{F}) p(\mathbf{F}|\mathbf{X}) p(\mathbf{X})}{q(\mathbf{X})} \right)$$

[4] Damianou, A. C. (2015). Deep Gaussian Processes and Variational Propagation of Uncertainty (Doctoral dissertation)

Lower Bound[4]

$$
\begin{aligned}
\mathcal{L} = & \int_{\mathbf{X},\mathbf{F}} q(\mathbf{X}) \log\left(\frac{p(\mathbf{Y},\mathbf{F},\mathbf{X})}{q(\mathbf{X})}\right) \\
& \int_{\mathbf{X},\mathbf{F}} q(\mathbf{X}) \log\left(\frac{p(\mathbf{Y}|\mathbf{F})p(\mathbf{F}|\mathbf{X})p(\mathbf{X})}{q(\mathbf{X})}\right) \\
& = \int_{\mathbf{F},\mathbf{X}} q(\mathbf{X}) \log p(\mathbf{Y}|\mathbf{F})p(\mathbf{F}|\mathbf{X}) - \int_{\mathbf{X}} q(\mathbf{X}) \log \frac{q(\mathbf{X})}{p(\mathbf{X})}
\end{aligned}
$$

[4] Damianou, A. C. (2015). Deep Gaussian Processes and Variational Propagation of Uncertainty (Doctoral dissertation)

Lower Bound[4]

$$
\begin{aligned}
\mathcal{L} = \int_{\mathbf{X},\mathbf{F}} q(\mathbf{X}) \log \left(\frac{p(\mathbf{Y}, \mathbf{F}, \mathbf{X})}{q(\mathbf{X})} \right) \\
\int_{\mathbf{X},\mathbf{F}} q(\mathbf{X}) \log \left(\frac{p(\mathbf{Y}|\mathbf{F}) p(\mathbf{F}|\mathbf{X}) p(\mathbf{X})}{q(\mathbf{X})} \right) \\
= \int_{\mathbf{F},\mathbf{X}} q(\mathbf{X}) \log p(\mathbf{Y}|\mathbf{F}) p(\mathbf{F}|\mathbf{X}) - \int_{\mathbf{X}} q(\mathbf{X}) \log \frac{q(\mathbf{X})}{p(\mathbf{X})} \\
= \tilde{\mathcal{L}} - \mathrm{KL}\left(q(\mathbf{X}) \parallel p(\mathbf{X}) \right)
\end{aligned}
$$

[4] Damianou, A. C. (2015). Deep Gaussian Processes and Variational Propagation of Uncertainty (Doctoral dissertation)

Lower Bound

$$\tilde{\mathcal{L}} = \int_{\mathbf{F},\mathbf{X}} q(\mathbf{X}) \log p(\mathbf{Y}|\mathbf{F})p(\mathbf{F}|\mathbf{X})$$

- Has not eliviate the problem at all, X still needs to go through F to reach the data
- Idea of sparse approximations[5]

[5] Quinonero-Candela, Joaquin, & Rasmussen, C. E. (2005). A unifying view of sparse approximate Gaussian process regression & Snelson, E., & Ghahramani, Z. (2006). Sparse Gaussian processes using pseudo-inputs

Lower Bound

- Add another set of samples from the same prior

$$p(\mathbf{U}|\mathbf{Z}) = \prod_{j=1}^{d} \mathcal{N}(\mathbf{u}_{:,j}|\mathbf{0}, \mathbf{K})$$

Lower Bound

- Add another set of samples from the same prior

$$p(\mathbf{U}|\mathbf{Z}) = \prod_{j=1}^{d} \mathcal{N}(\mathbf{u}_{:,j}|\mathbf{0}, \mathbf{K})$$

- Conditional distribution

$$p(\mathbf{f}_{:,j}, \mathbf{u}_{:,j}|\mathbf{X}, \mathbf{Z}) = p(\mathbf{f}_{:,j}|\mathbf{u}_{:,j}, \mathbf{X}, \mathbf{Z})p(\mathbf{u}_{:,j}|\mathbf{Z})$$

$$= \mathcal{N}\left(\mathbf{f}_{:,j}|\mathbf{K}_{fu}(\mathbf{K}_{uu})^{-1}\mathbf{u}_{:,j}, \mathbf{K}_{ff} - \mathbf{K}_{fu}(\mathbf{K}_{uu})^{-1}\mathbf{K}_{uf}\right)\mathcal{N}\left(\mathbf{u}_{:,j}|\mathbf{0}, \mathbf{K}_{uu}\right),$$

Lower Bound

$$p(\mathbf{Y}, \mathbf{F}, \mathbf{U}, \mathbf{X}|\mathbf{Z}) = p(\mathbf{X}) \prod_{j=1}^{d} p(\mathbf{y}_{:,j}|\mathbf{f}_{:,j}) p(\mathbf{f}_{:,j}|\mathbf{u}_{:,j}, \mathbf{X}) p(\mathbf{u}_{:,j}|\mathbf{Z})$$

- we have done nothing to the model, just added *halucinated* observations

Lower Bound

$$p(\mathbf{Y}, \mathbf{F}, \mathbf{U}, \mathbf{X}|\mathbf{Z}) = p(\mathbf{X}) \prod_{j=1}^{d} p(\mathbf{y}_{:,j}|\mathbf{f}_{:,j}) p(\mathbf{f}_{:,j}|\mathbf{u}_{:,j}, \mathbf{X}) p(\mathbf{u}_{:,j}|\mathbf{Z})$$

- we have done nothing to the model, just added *halucinated* observations
- however, we will now interpret $\mathbf{U}$ and $\mathbf{X}_u$ not as random variables but variational parameters

Lower Bound

$$p(\mathbf{Y},\mathbf{F},\mathbf{U},\mathbf{X}|\mathbf{Z}) = p(\mathbf{X})\prod_{j=1}^{d} p(\mathbf{y}_{:,j}|\mathbf{f}_{:,j})p(\mathbf{f}_{:,j}|\mathbf{u}_{:,j},\mathbf{X})p(\mathbf{u}_{:,j}|\mathbf{Z})$$

- we have done nothing to the model, just added *halucinated* observations
- however, we will now interpret $\mathbf{U}$ and $\mathbf{X}_u$ not as random variables but variational parameters
- i.e. parametrise approximate posterior using these parameters (remember sparse motivation)

Lower Bound

- Variational distributions are approximations to intractable posteriors,

$$
\begin{aligned}
q(\mathbf{U}) &\approx p(\mathbf{U}|\mathbf{Y}, \mathbf{X}, \mathbf{Z}, \mathbf{F}) \\
q(\mathbf{F}) &\approx p(\mathbf{F}|\mathbf{U}, \mathbf{X}, \mathbf{Z}, \mathbf{Y}) \\
q(\mathbf{X}) &\approx p(\mathbf{X}|\mathbf{Y})
\end{aligned}
$$

Lower Bound

- Variational distributions are approximations to intractable posteriors,

$$
\begin{aligned}
q(\mathbf{U}) &\approx p(\mathbf{U}|\mathbf{Y}, \mathbf{X}, \mathbf{Z}, \mathbf{F}) \\
q(\mathbf{F}) &\approx p(\mathbf{F}|\mathbf{U}, \mathbf{X}, \mathbf{Z}, \mathbf{Y}) \\
q(\mathbf{X}) &\approx p(\mathbf{X}|\mathbf{Y})
\end{aligned}
$$

- Assume that we can *find* $\mathbf{U}$ that completely represents $\mathbf{F}$, i.e. $\mathbf{U}$ is sufficient statistics of $\mathbf{F}$,

$$
q(\mathbf{F}) \approx p(\mathbf{F}|\mathbf{U}, \mathbf{X}, \mathbf{Z}, \mathbf{Y}) = p(\mathbf{F}|\mathbf{U}, \mathbf{X}, \mathbf{Z})
$$

Lower Bound

$$\tilde{\mathcal{L}} = \int_{\mathbf{X},\mathbf{F},\mathbf{U}} q(\mathbf{F})q(\mathbf{U})q(\mathbf{X}) \log \frac{p(\mathbf{Y},\mathbf{F},\mathbf{U}|\mathbf{X},\mathbf{Z})}{q(\mathbf{F})q(\mathbf{U})}$$

Lower Bound

$$\tilde{\mathcal{L}} = \int_{\mathbf{X},\mathbf{F},\mathbf{U}} q(\mathbf{F})q(\mathbf{U})q(\mathbf{X}) \log \frac{p(\mathbf{Y},\mathbf{F},\mathbf{U}|\mathbf{X},\mathbf{Z})}{q(\mathbf{F})q(\mathbf{U})}$$

$$= \int_{\mathbf{X},\mathbf{F},\mathbf{U}} q(\mathbf{F})q(\mathbf{U})q(\mathbf{X}) \log \frac{\prod_{j=1}^{d} p(\mathbf{y}_{:,j}|\mathbf{f}_{:,j})p(\mathbf{f}_{:,j}|\mathbf{u}_{:,j},\mathbf{X},\mathbf{Z})p(\mathbf{u}_{:,j}|\mathbf{Z})}{q(\mathbf{F})q(\mathbf{U})}$$

Lower Bound

$$
\begin{aligned}
\tilde{\mathcal{L}} &= \int_{\mathbf{X},\mathbf{F},\mathbf{U}} q(\mathbf{F})q(\mathbf{U})q(\mathbf{X}) \log \frac{p(\mathbf{Y},\mathbf{F},\mathbf{U}|\mathbf{X},\mathbf{Z})}{q(\mathbf{F})q(\mathbf{U})} \\
&= \int_{\mathbf{X},\mathbf{F},\mathbf{U}} q(\mathbf{F})q(\mathbf{U})q(\mathbf{X}) \log \frac{\prod_{j=1}^{d} p(\mathbf{y}_{:,j}|\mathbf{f}_{:,j})p(\mathbf{f}_{:,j}|\mathbf{u}_{:,j},\mathbf{X},\mathbf{Z})p(\mathbf{u}_{:,j}|\mathbf{Z})}{q(\mathbf{F})q(\mathbf{U})}
\end{aligned}
$$

- Assume that $\mathbf{U}$ is sufficient statistics for $\mathbf{F}$

$$
q(\mathbf{F})q(\mathbf{U})q(\mathbf{X}) = p(\mathbf{F}|\mathbf{U},\mathbf{X},\mathbf{Z})q(\mathbf{U})q(\mathbf{X})
$$

Lower Bound

$$\tilde{\mathcal{L}} = \int_{\mathbf{X},\mathbf{F},\mathbf{U}} \prod_{j=1}^{d} p(\mathbf{f}_{:,j}|\mathbf{u}_{:,j}, \mathbf{X}, \mathbf{Z}) q(\mathbf{u}_{:,j}) q(\mathbf{X})$$

$$\log \frac{\prod_{j=1}^{d} p(\mathbf{y}_{:,j}|\mathbf{f}_{:,j}) p(\mathbf{f}_{:,j}|\mathbf{u}_{:,j}, \mathbf{X}, \mathbf{Z}) p(\mathbf{u}_{:,j}|\mathbf{Z})}{\prod_{j=1}^{d} p(\mathbf{f}_{:,j}|\mathbf{u}_{:,j}, \mathbf{X}, \mathbf{Z}) q(\mathbf{u}_{:,j})} =$$

Lower Bound

$$\tilde{\mathcal{L}} = \int_{\mathbf{X},\mathbf{F},\mathbf{U}} \prod_{j=1}^{d} p(\mathbf{f}_{:,j}|\mathbf{u}_{:,j},\mathbf{X},\mathbf{Z})q(\mathbf{u}_{:,j})q(\mathbf{X})$$

$$\log \frac{\prod_{j=1}^{d} p(\mathbf{y}_{:,j}|\mathbf{f}_{:,j})p(\mathbf{f}_{:,j}|\mathbf{u}_{:,j},\mathbf{X},\mathbf{Z})p(\mathbf{u}_{:,j}|\mathbf{Z})}{\prod_{j=1}^{d} p(\mathbf{f}_{:,j}|\mathbf{u}_{:,j},\mathbf{X},\mathbf{Z})q(\mathbf{u}_{:,j})} =$$

$$= \int_{\mathbf{X},\mathbf{F},\mathbf{U}} \prod_{j=1}^{p} p(\mathbf{f}_{:,j}|\mathbf{u}_{:,j},\mathbf{X},\mathbf{Z})q(\mathbf{u}_{:,j})q(\mathbf{X}) \log \frac{\prod_{j=1}^{p} p(\mathbf{y}_{:,j}|\mathbf{f}_{:,j})p(\mathbf{u}_{:,j}|\mathbf{Z})}{\prod_{j=1}^{p} q(\mathbf{u}_{:,j})}$$

$$= \mathbb{E}_{q(\mathbf{F}),q(\mathbf{X}),q(\mathbf{U})}\left[p(\mathbf{Y}|\mathbf{F})\right] - \mathrm{KL}\left(q(\mathbf{U})||p(\mathbf{U}|\mathbf{Z})\right)$$

Summary

$$\mathbb{E}_{q(\mathbf{F}),q(\mathbf{X}),q(\mathbf{U})}\left[p(\mathbf{Y}|\mathbf{F})\right] - \mathrm{KL}\left(q(\mathbf{U})||p(\mathbf{U}|\mathbf{Z})\right) - \mathrm{KL}\left(q(\mathbf{X})||p(\mathbf{X})\right)$$

- Expectation tractable (for some co-variances)
- Reduces to expectations over co-variance functions know as Ψ statistics
- Allows us to place priors and not "regularisers" over the latent representation

Latent space priors

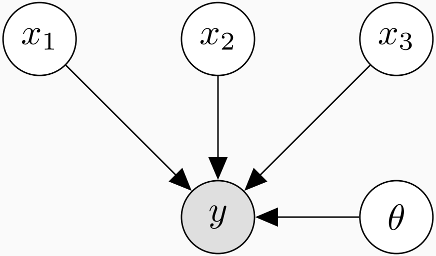
Latent space priors[6]

$$\mathbb{E}_{q(\mathbf{F}),q(\mathbf{X}),q(\mathbf{U})}\left[p(\mathbf{Y}|\mathbf{F})\right] - \mathrm{KL}\left(q(\mathbf{U})||p(\mathbf{U}|\mathbf{Z})\right) - \mathrm{KL}\left(q(\mathbf{X})||p(\mathbf{X})\right)$$

- Importantly $p(\mathbf{X})$ appears only in KL term
- Allows us to express stronger assumptions about the model

[6] Damianou, A. C., Titsias, M., & Lawrence, Neil D, Variational Inference for Uncertainty on the Inputs of Gaussian Process Models (2014)

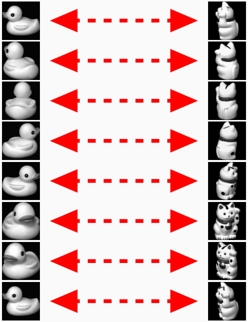
Factor Analysis



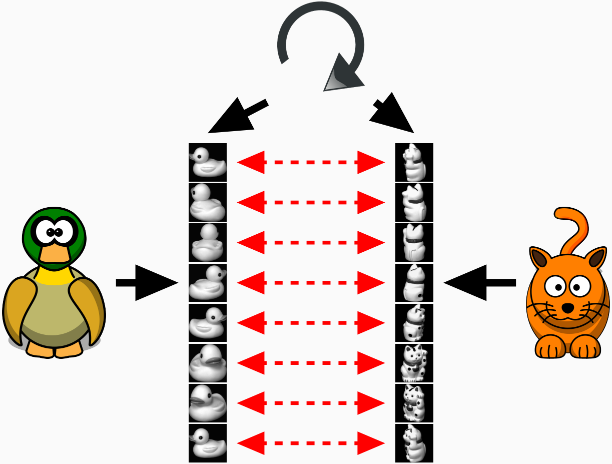
$$y = f(x_1, x_2, x_3) + \epsilon$$

Alignments

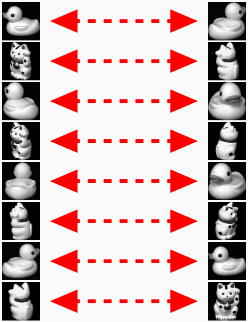
Alignments



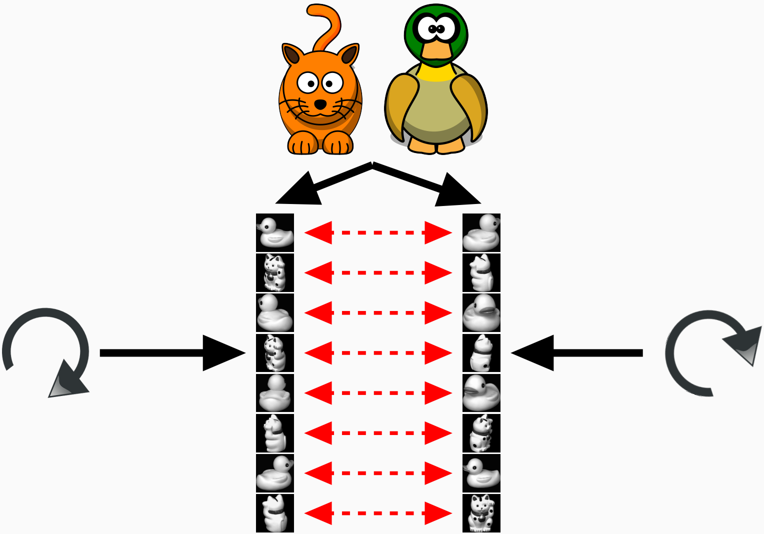
Alignments



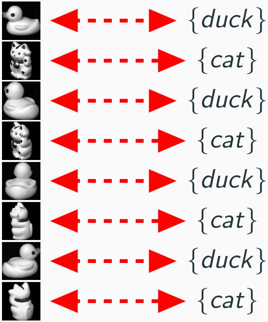
Alignments



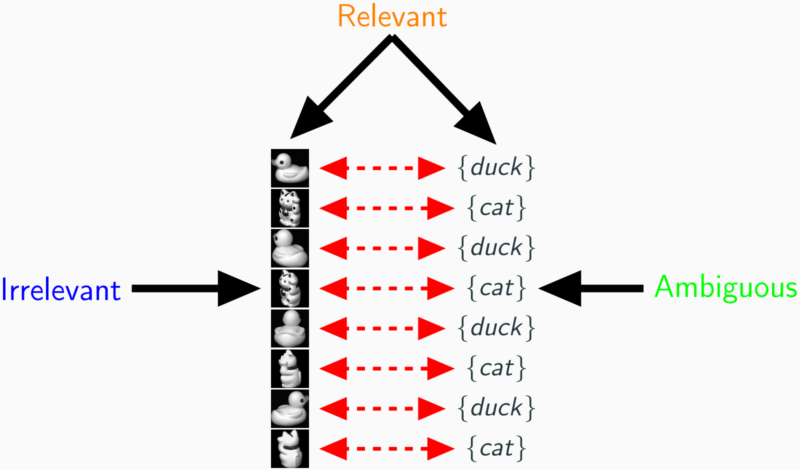
Alignments



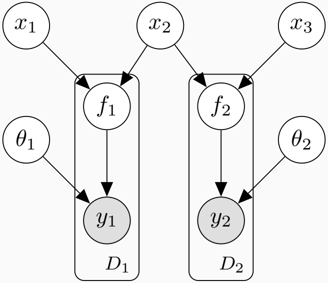
Alignments



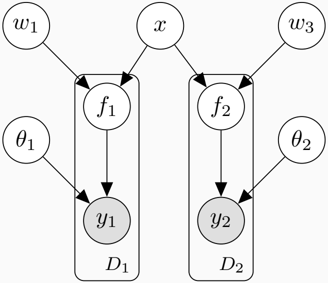
Alignments


Relevant
{duck}
{cat}
{duck}
Irrelevant
{cat}
Ambiguous
{duck}
{cat}
{duck}
{cat}

Explaining Away cont.

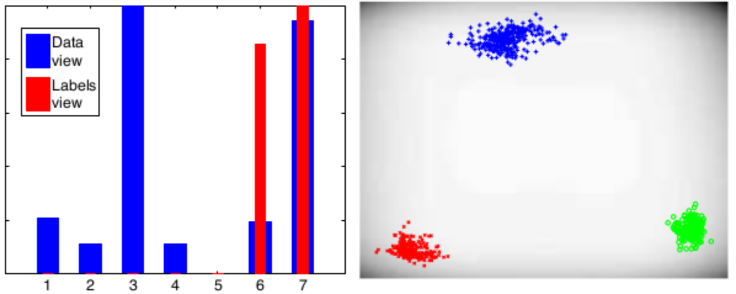


IBFA with GP-LVM[7]

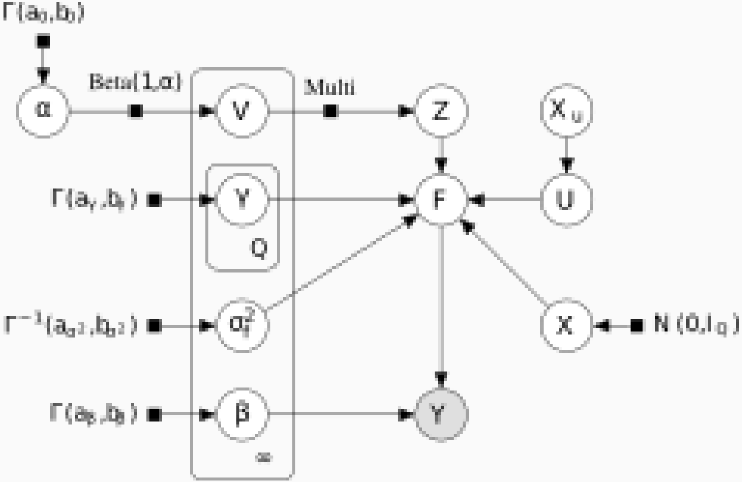


$$y_1 = f(w_1^{\mathrm{T}} x) \quad y_2 = f(w_2^{\mathrm{T}} x)$$

[7]Damianou, A., Lawrence, N. D., & Ek, C. H. (2016). Multi-view learning as a nonparametric nonlinear inter-battery factor analysis

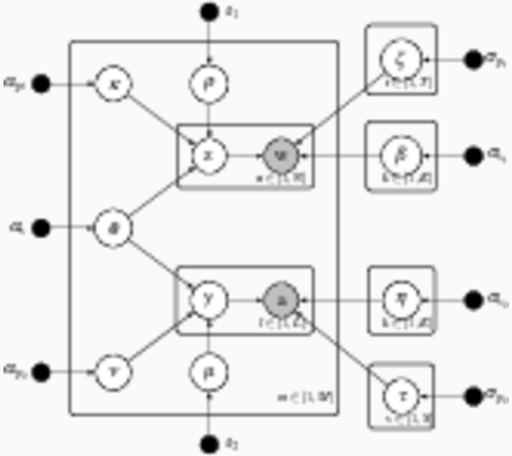

Data view
Labels view
1
2
3
4
5
6
7

GPDP [8]

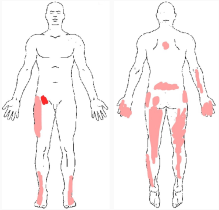


[8] Joint work with Andrew Lawrence and Neill Cambpell at University of Bath, Will be presented at *Advances in Modeling and Learning Interactions from Complex Data* NIPS 2017

IBTM [9]



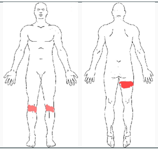
[9] Zhang, C., Kjellstr\"om, Hedvig, & Ek, C. H., Inter-battery topic representation learning, In ECCV 2016

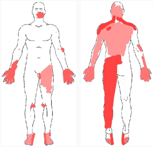


Symptom diagnoses: Interscapular discomfort; R arm discomfort; B hands discomfort; Lumbago; B crest of the ilium discomfort; L side thigh discomfort; B back thigh discomfort; B calf discomfort; B achilles tendinitis; B shin discomfort; R inguinal discomfort;

Pattern diagnoses B L5 Radiculopathy; B S1 Radiculopathy; B C7 Radiculopathy;

Pathophysiological diagnoses DLI L4-L5; DLI S1-S2; DLI C6-C7

	**6 Prd:** R back thigh dcf; L PFS (Patellofemoral pain syndrome); R PFS; L L5 Rdc; R L5 Rdc; DLI L4-L5; **6 GT:** R back thigh dcf; L PFS; R PFS; L L5 Rdc; R L5 Rdc; DLI L4-L5;

	**50 Prd:** Headache; L neck dcf; R neck dcf; Neck dcf; L upper trapezius dcf; R upper trapezius dcf; L shoulder dcf; L hand dcf; R hand dcf; Interscapular dcf; Lumbago; Lateral abdominal dcf; L groin dcf; L side thigh dcf; R side thigh dcf; L calf dcf; L back thigh dcf; L crest of the ilium dcf; R crest of the ilium dcf; R foot arch dcf; L toe joint dcf; R toe joint dcf; L medial elbow dcf; L ankle dcf; R ankle dcf; L foot arch dcf; L PFS; L dorsal knee dcf; R medial knee dcf; L C2 Rdc; R C2 Rdc; L C3 Rdc; R C3 Rdc; L C4 Rdc; R C4 Rdc; R C6 Rdc; L C6 Rdc; L C7 Rdc; R C7 Rdc; L L5 Rdc; R L5 Rdc; L S1 Rdc; R S1 Rdc; DLI C2-C3; DLI C3-C4; DLI C5-C6; DLI C6-C7; DLI L4-L5; DLI L5-S1; OB; **50 GT:** L back headache; R back headache; Neck dcf; L jaw dcf; L upper trapezius dcf; R upper trapezius dcf; L arm dcf; R arm dcf; L lateral elbow dcf; R lateral elbow dcf; L hand joint dcf; R hand joint dcf; L hand dcf; R hand dcf; L thumb dcf; R thumb dcf; L finger dcf; R finger dcf; Lumbago; L groin dcf; L back thigh dcf; L calf dcf; L medial knee dcf; L ankle dcf; R ankle dcf; R medial knee dcf; R big toe dcf; L big toe dcf; L C2 Rdc; R C2 Rdc; L C3 Rdc; R C3 Rdc; L C4 Rdc; R C4 Rdc; L C5 Rdc; R C5 Rdc; L C6 Rdc; R C6 Rdc; L C7 Rdc; R C7 Rdc; L L4 Rdc; L L5 Rdc; R L5 Rdc; L S1 Rdc; R S1 Rdc; Craniocervical joint injury; DLI C4-C5; DLI L3-L4; DLI L4-L5; DLI L5-S1;

Convolutional Deep GPs

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

"Scientific modelling is a scientific activity, the aim of which is to make a particular part or feature of the world easier to understand, define, quantify, visualize, or simulate by referencing it to existing and usually commonly accepted knowledge." [10]

[10] Wikipedia

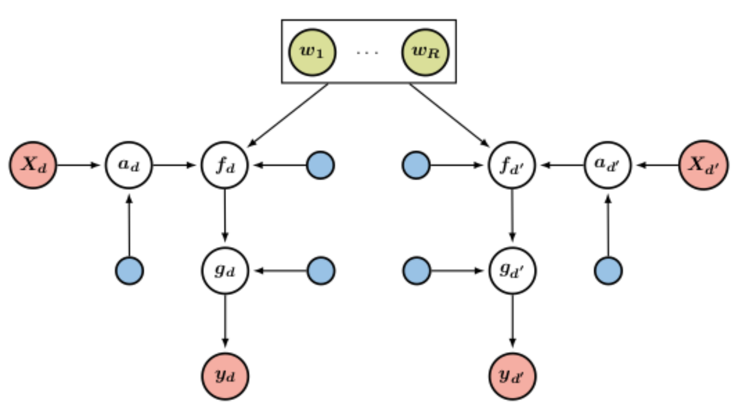
Hierarchies

$$\mathbf{y}_d = g_d(f_d(a_d(\mathbf{x}))) + \epsilon$$

$$f_d(\mathbf{x}) = \sum_{r=1}^{R} \int T_{d,r}(\mathbf{x} - \mathbf{z}) w_r(\mathbf{z}) \mathrm{d}\mathbf{z}$$

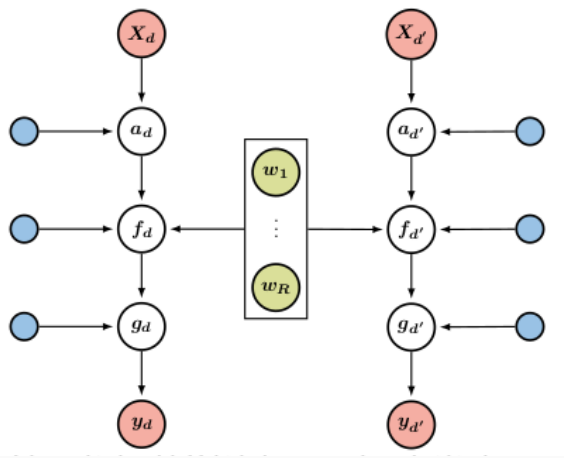
- Hierarchical set of function
- Convolution process with shared kernel

Hierarchies[11]

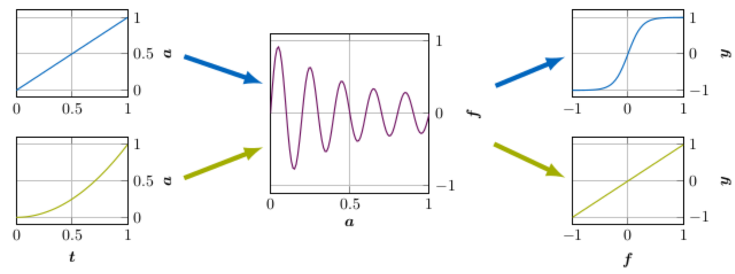


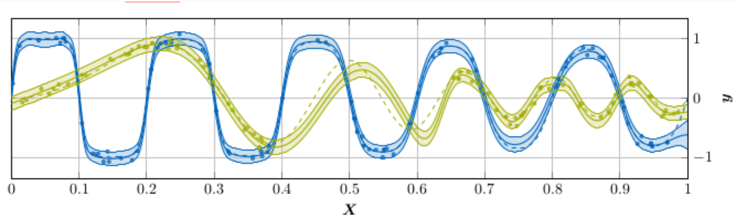
[11] Kaiser, M., Otte, C., Runkler, T., & Ek, C. H. , Bayesian alignments of warped multi-output gaussian processes, ArXiv e-prints, (), (2017).

Hierarchies[11]



Hierarchies[11]





Composite Functions

- Learning Composite functions have become very popular
- Composite functions are not as intuitive as one might think

Composite Functions

Why are composite functions attractive?

$$\mathbf{y} = g(\mathbf{x}) = f_K(f_{K-1}(f_{K-2}(\ldots f_1(\mathbf{x})\ldots)))$$

Functions

- Kernel of a function

$$\mathrm{Kern}(f_k) = \left\{(\mathsf{x}, \mathsf{x}') | f_k(\mathsf{x}) = f_k(\mathsf{x}')\right\}$$

Functions

- Kernel of a function

$$\mathrm{Kern}(f_k) = \left\{(\mathbf{x}, \mathbf{x}') | f_k(\mathbf{x}) = f_k(\mathbf{x}')\right\}$$

- Image of a function

$$\mathrm{Im}(f_k(\mathbf{x})) = \{\mathbf{y} \in Y | \mathbf{y} = f_k(\mathbf{x}), \mathbf{x} \in X\}$$

Composite Functions

- Kernel of function

  $\text{Kern}(f_1) \subseteq \text{Kern}(f_{k-1} \circ \ldots \circ f_2 \circ f_1) \subseteq \text{Kern}(f_k \circ f_{k-1} \circ \ldots \circ f_2 \circ f_1)$

Composite Functions

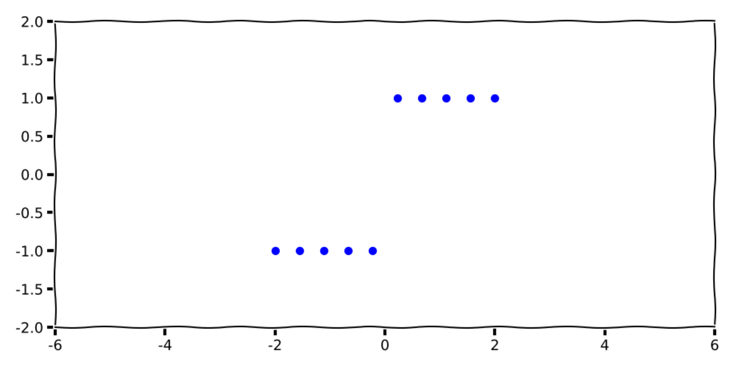
- Kernel of function

  $$\mathrm{Kern}(f_1) \subseteq \mathrm{Kern}(f_{k-1} \circ \ldots \circ f_2 \circ f_1) \subseteq \mathrm{Kern}(f_k \circ f_{k-1} \circ \ldots \circ f_2 \circ f_1)$$

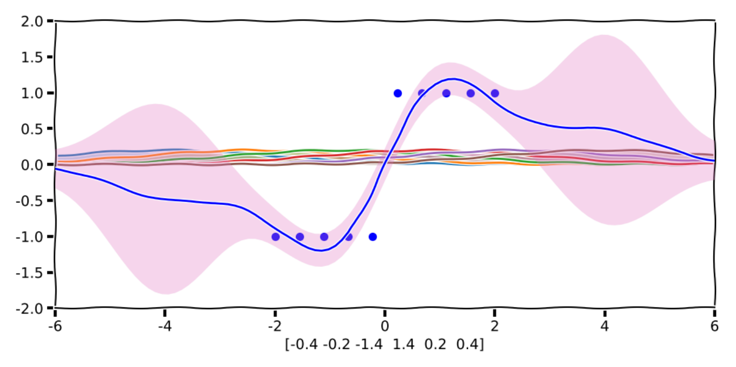
- Image of a function

  $$\mathrm{Im}(f_k \circ f_{k-1} \circ \ldots \circ f_2 \circ f_1) \subseteq \mathrm{Im}(f_k \circ f_{k-1} \circ \ldots \circ f_2) \subseteq \ldots \subseteq \mathrm{Im}(f_k)$$

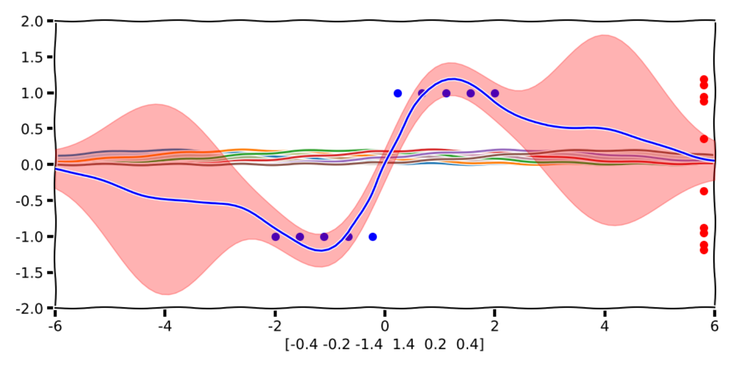
Composite Functions



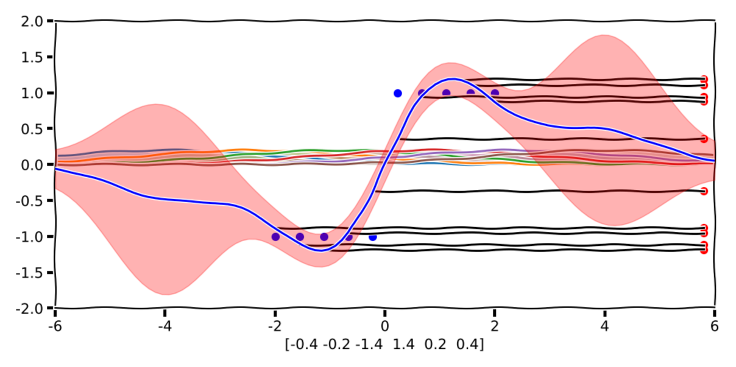
Composite Functions



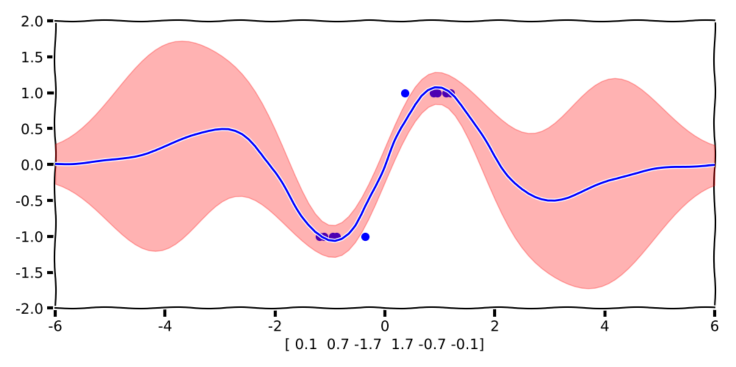
Composite Functions



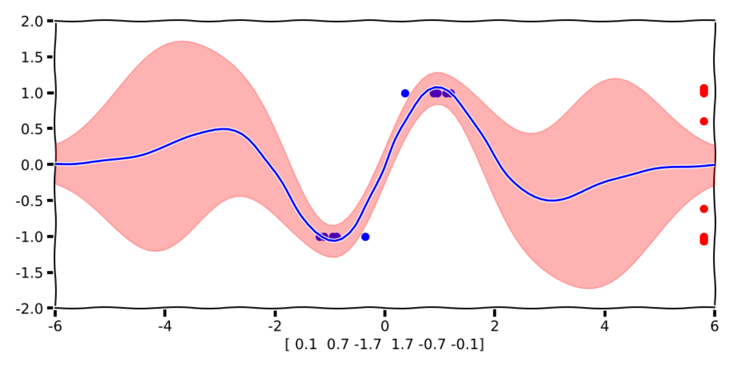
Composite Functions



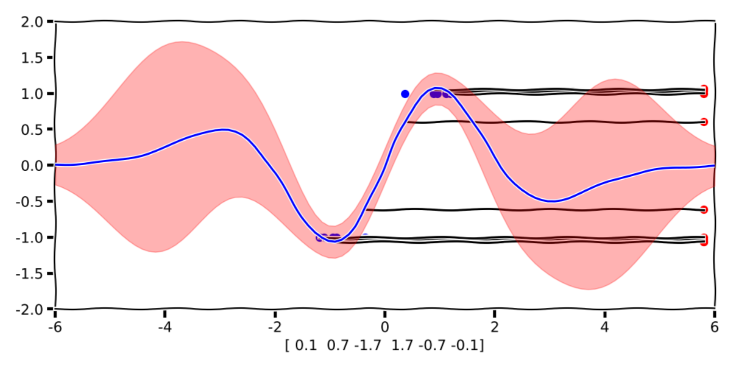
Composite Functions



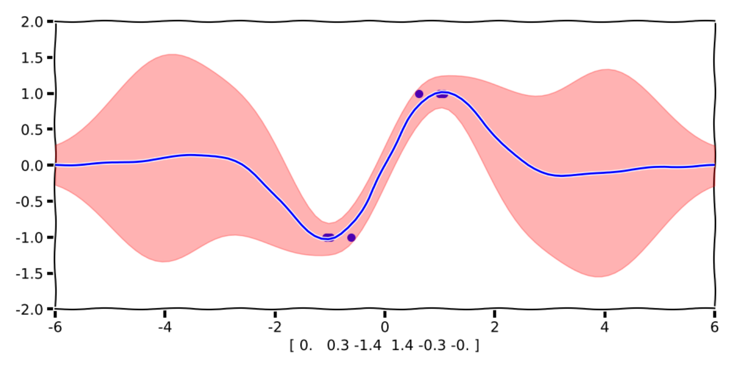
Composite Functions



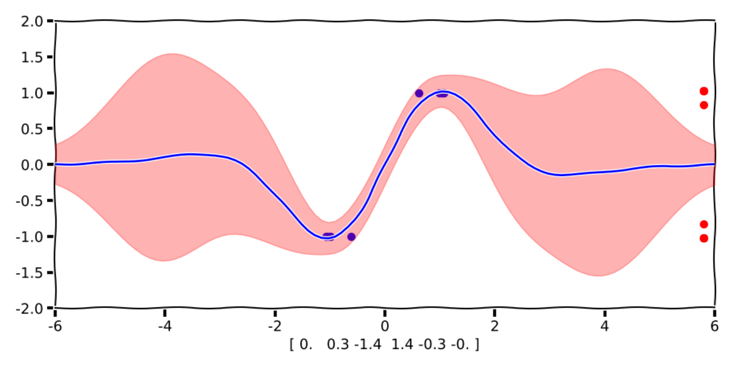
Composite Functions



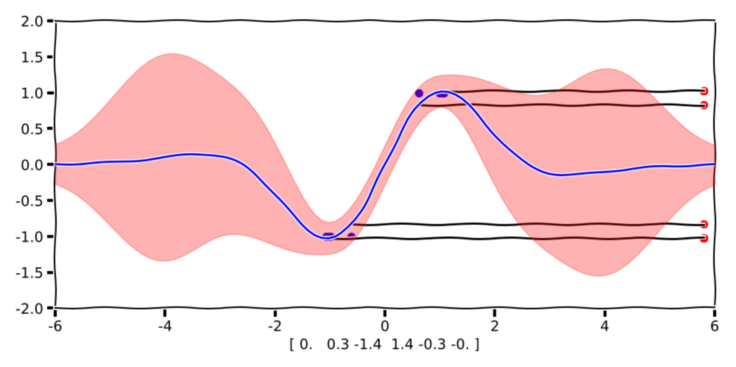
Composite Functions



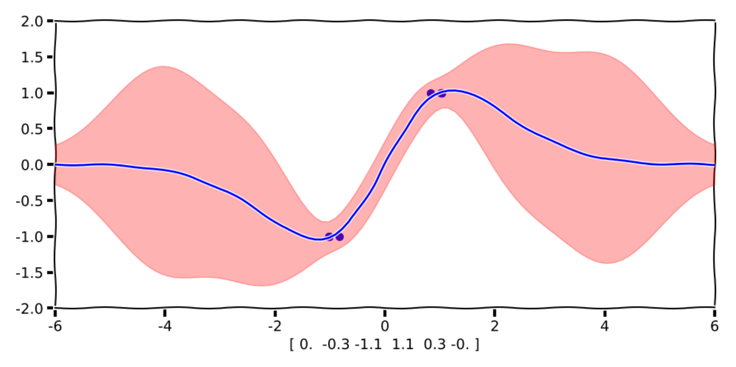
Composite Functions



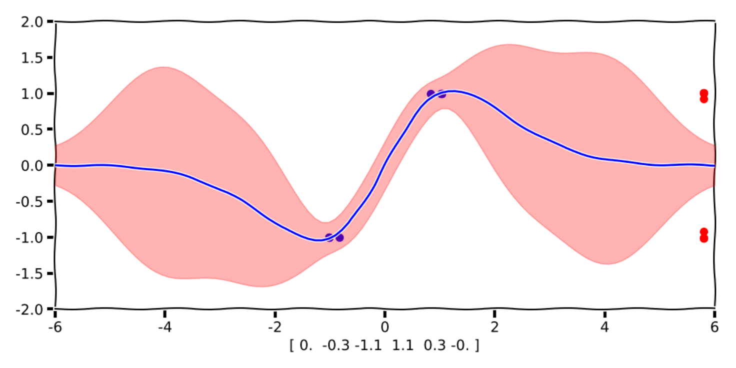
Composite Functions



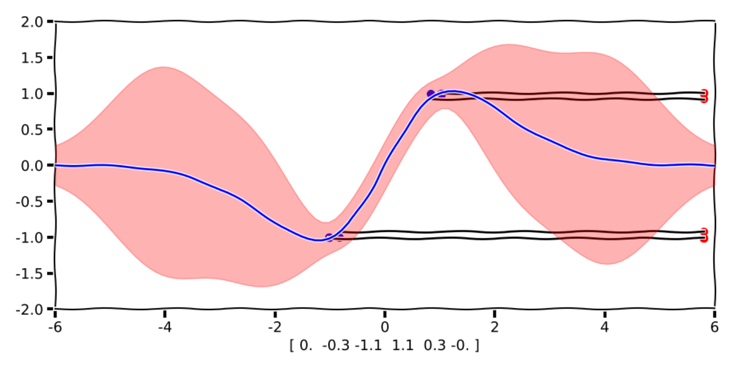
Composite Functions



Composite Functions



Composite Functions

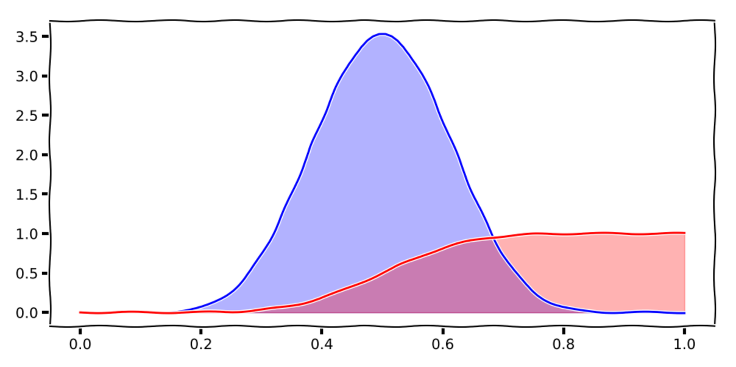


Theorem (Change of Variable)
Let $x \in \mathcal{X} \subseteq \mathbb{R}^n$ be a random vector with a probability density function given by $p_x(x)$, and let $y \in \mathcal{Y} \subseteq \mathbb{R}^n$ be a random vector such that $\psi(y) = x$, where the function $\psi : \mathcal{Y} \to \mathcal{X}$ is bijective of class of $\mathcal{C}^1$ and $| \bigtriangledown \psi(y)| > 0, \forall y \in \mathcal{Y}$. Then, the probability density function $p_y(\cdot)$ induced in $\mathcal{Y}$ is given by

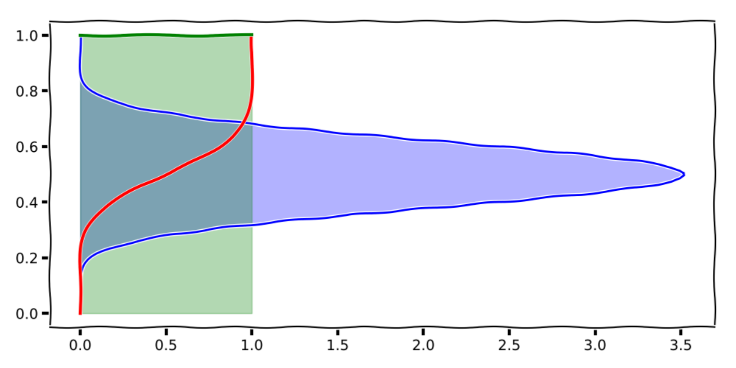
$$p_y(y) = p_x(\psi(y))| \bigtriangledown \psi(y)|$$

where $\bigtriangledown\psi(\cdot)$ denotes the Jacobian of $\psi(\cdot)$, and $| \cdot |$ denotes the determinant operator.

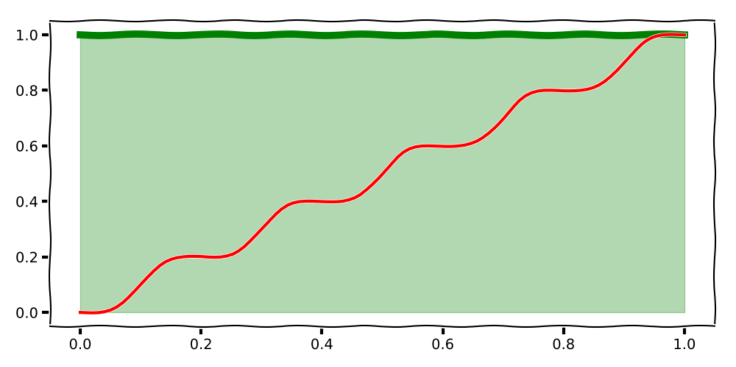
Sampling



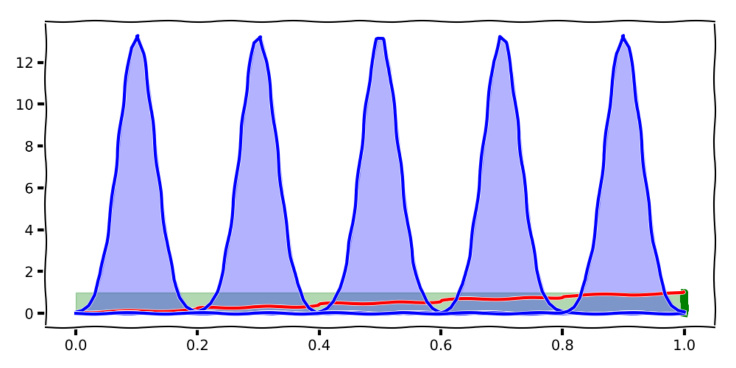
Sampling



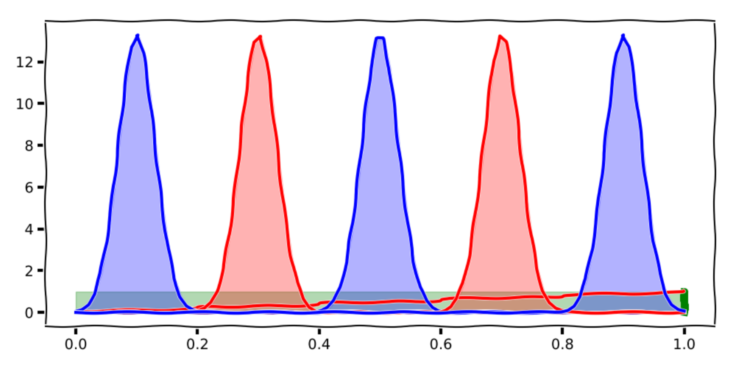
Change of Variables



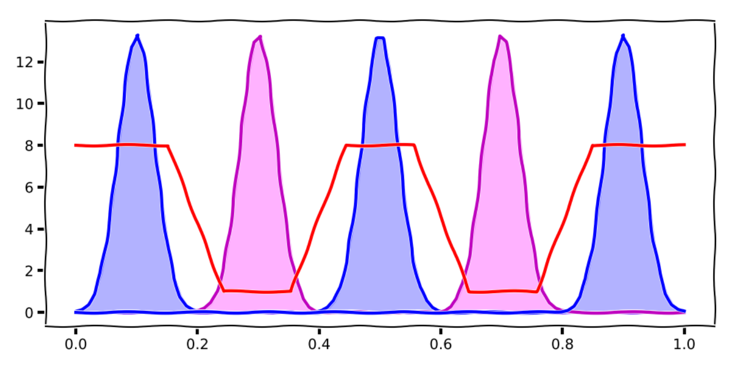
Change of Variables



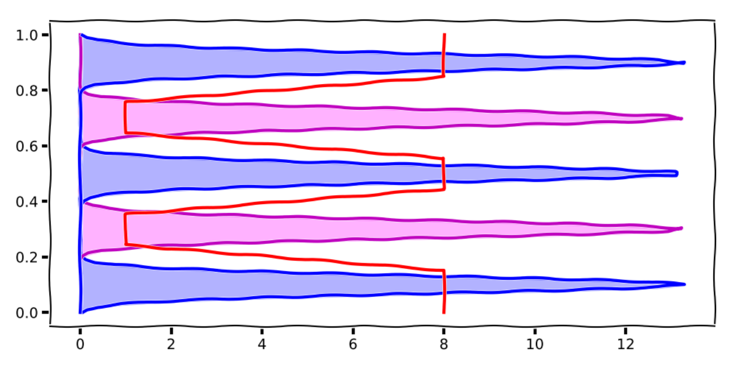
Change of Variables



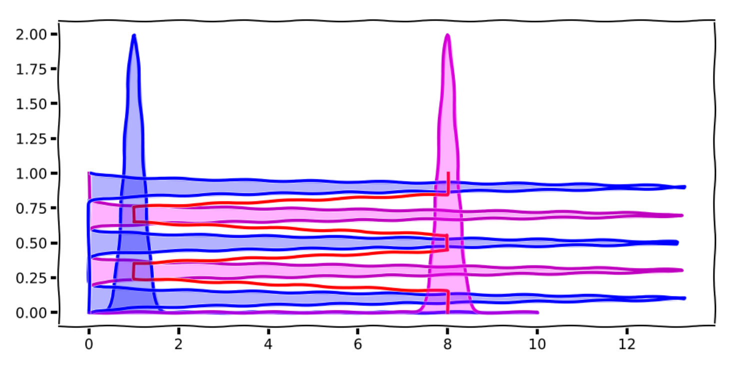
Change of Variables



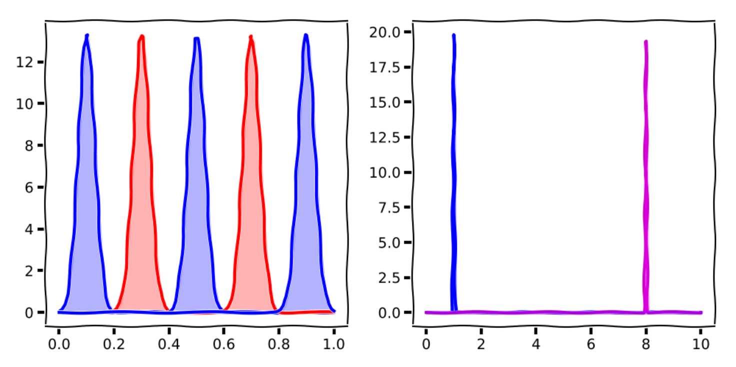
Change of Variables



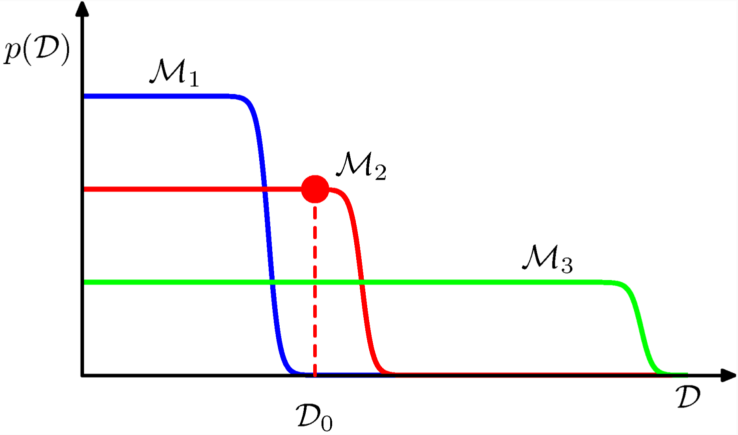
Change of Variables



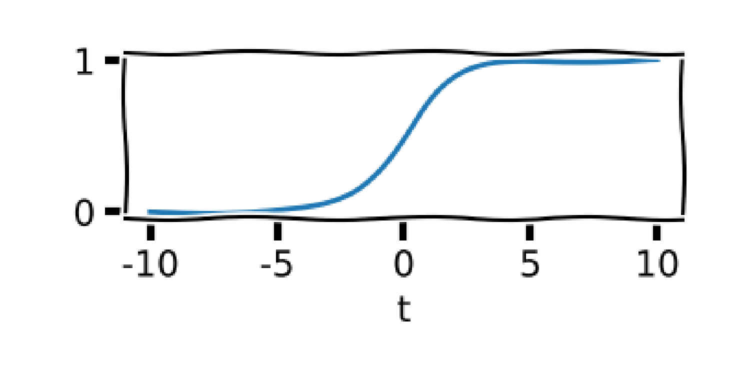
Change of Variables



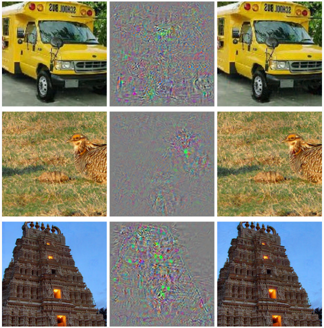
MacKay plot

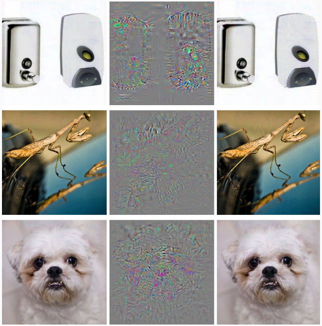

$p(\mathcal{D})$
$\mathcal{M}_1$
$\mathcal{M}_2$
$\mathcal{M}_3$
$\mathcal{D}$
$\mathcal{D}_0$

Activation functions



Data inefficiency[12]





[12] Nguyen, A. M., Yosinski, J., & Clune, J., Deep neural networks are easily fooled: high confidence predictions for unrecognizable images, CoRR, abs/1412.1897(), (2014).

Summary

Summary

- Unsupervised learning is very hard

Summary

- Unsupervised learning is very hard
  - *Its actually not, its really really easy.*

Summary

- Unsupervised learning is very hard
    - *Its actually not, its really really easy.*
- Relevant assumptions needed to learn anything useful

Summary

- Unsupervised learning is very hard
    - *Its actually not, its really really easy.*
- Relevant assumptions needed to learn anything useful
- Strong assumptions needed to learn anything from "sensible" amounts of data

Summary

- Unsupervised learning is very hard
  - *Its actually not, its really really easy.*
- Relevant assumptions needed to learn anything useful
- Strong assumptions needed to learn anything from "sensible" amounts of data
- Stochastic processes (DPs,GPs) provide strong, interpretative assumptions that aligns well to our intuitions allowing us to make relevant assumptions

Summary II

- Composite functions cannot model more things

Summary II

- Composite functions cannot model more things
- However, they can easily warp the input space to model less things

Summary II

- Composite functions cannot model more things
- However, they can easily warp the input space to model less things
- This leads to high requirments on data

Summary II

- Composite functions cannot model more things
- However, they can easily warp the input space to model less things
- This leads to high requirments on data
- Even bigger need for uncertainty propagation, we cannot assume noiseless data

Summary II

- Composite functions cannot model more things
- However, they can easily warp the input space to model less things
- This leads to high requirments on data
- Even bigger need for uncertainty propagation, we cannot assume noiseless data
- Intuitions needs to change, we need to think of priors over hierarchies

eof

References

Pierre Simon Laplace.
A philosophical essay on probabilities, 1814.