University of BRISTOL

Unsupervised Learning with Gaussian Processes

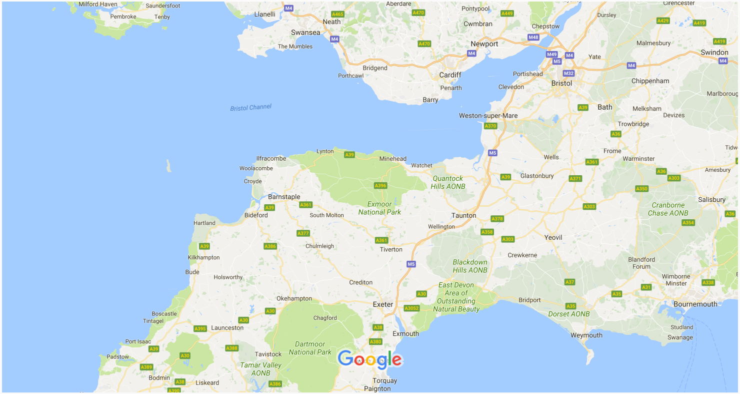
Carl Henrik Ek - carlhenrik.ek@bristol.ac.uk

September 5, 2018

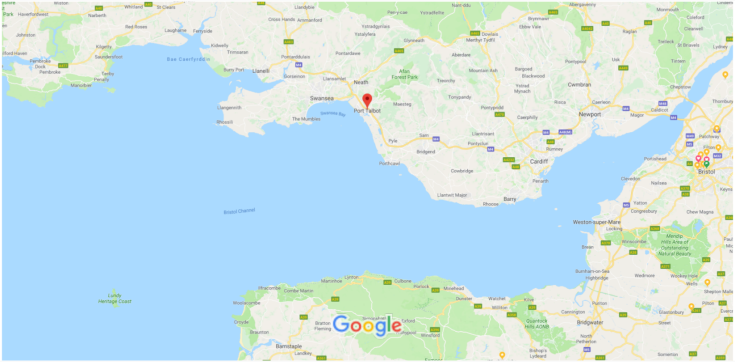
http://www.carlhenrik.com

Introductions

This where I live

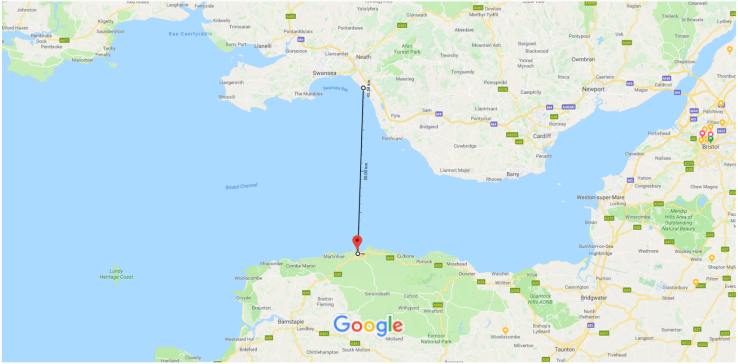


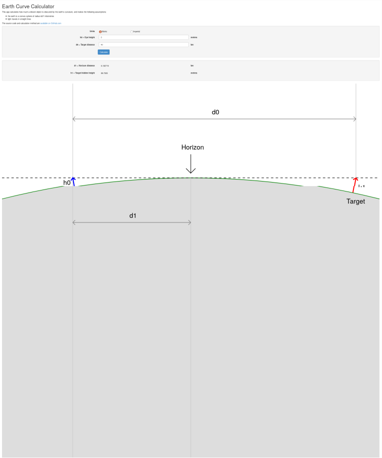
This is what I do



Swansea
Neath
Port Talbot
Llanelli
The Mumbles
Swansea Bay
Bae Caerfyrddin
Maesteg
Pyle
Bridgend
Porthcawl
Cowbridge
Llantwit Major
Barry
Rhoose
Penarth
Cardiff
Newport
Cwmbran
Caerphilly
Pontypridd
Tonypandy
Treorchy
Aberdare
Merthyr Tydfil
Pontypool
Afan Forest Park
Bristol Channel
Lundy Heritage Coast
Weston-super-Mare
Clevedon
Portishead
Bristol
Nailsea
Burnham-on-Sea
Bridgwater
Minehead
Lynton
Ilfracombe
Barnstaple
Pembroke
Tenby
Narberth
Haverfordwest
Kidwelly
Burry Port
Rhossili
Chepstow
Caldicot
Magor
Wells
Glastonbury
Street
Google

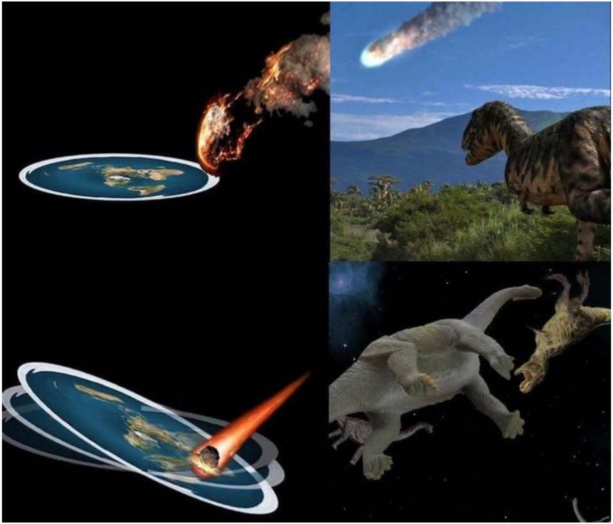

Bae Caerfyrddin
Llanelli
Swansea
Swansea Bay
The Mumbles
Neath
Port Talbot
Bridgend
Porthcawl
Cowbridge
Llantwit Major
Barry
Penarth
Cardiff
Caerphilly
Pontypridd
Newport
Cwmbran
Portishead
Clevedon
Bristol
Nailsea
Weston-super-Mare
Bristol Channel
Lundy Heritage Coast
Lynton
Martinhoe
Ilfracombe
Combe Martin
Woolacombe
Croyde
Braunton
Barnstaple
Bideford
Porlock
Minehead
Dunster
Watchet
Exmoor National Park
Quantock Hills AONB
Bridgwater
Burnham-on-Sea
Taunton
Mendip Hills Area of Outstanding Natural Beauty
Wells
Glastonbury
Google


Bae Caerfyrddin
Llanelli
Swansea
Neath
The Mumbles
Swansea Bay
Cwmbran
Newport
Cardiff
Bridgend
Barry
Penarth
Porthcawl
Cowbridge
Bristol Channel
Bristol
Clevedon
Portishead
Weston-super-Mare
Lundy
Heritage Coast
20.00 km
Ilfracombe
Combe Martin
Martinhoe
Porlock
Minehead
Woolacombe
Braunton
Barnstaple
Bideford
Exmoor
National Park
Bridgwater
Taunton
Quantock
Hills AONB
Burnham-on-Sea
Glastonbury
Wells
Google



Distance to horizon $6.2km$

Hidden height $125.6m$



Machine Learning

"In inductive inference, we go from the specific to the general. We make many observations, discern a pattern, make a generalization, and infer an explanation or a theory"
– Wassertheil-Smoller

Learning Theory

- $\mathcal{F}$ space of functions
- $\mathcal{A}$ learning algorithm
- $\mathcal{S} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$
- $\mathcal{S} \sim P(\mathcal{X} \times \mathcal{Y})$
- $\ell(\mathcal{A}_{\mathcal{F}}(\mathcal{S}), x, y)$ loss function

Statistical Learning

$$e(\mathcal{S}, \mathcal{A}, \mathcal{F}) = \mathbb{E}_{P(\{\mathcal{X}, \mathcal{Y}\})} \left[\ell(\mathcal{A}_{\mathcal{F}}(\mathcal{S}), x, y) \right]$$

No Free Lunch

We can come up with a combination of $\{\mathcal{S}, \mathcal{A}, \mathcal{F}\}$ that makes $e(\mathcal{S}, \mathcal{A}, \mathcal{F})$ take an arbitary value

Statistical Learning

$$\begin{aligned} e(\mathcal{S}, \mathcal{A}, \mathcal{F}) &= \mathbb{E}_{P(\{\mathcal{X}, \mathcal{Y}\})}\left[\ell(\mathcal{A}_{\mathcal{F}}(\mathcal{S}), x, y)\right] \\ &\approx \frac{1}{M}\sum_{n=1}^{M} \ell(\mathcal{A}_{\mathcal{F}}(\mathcal{S}), x_n, y_n) \end{aligned}$$

No Free Lunch

We can come up with a combination of $\{\mathcal{S}, \mathcal{A}, \mathcal{F}\}$ that makes $e(\mathcal{S}, \mathcal{A}, \mathcal{F})$ take an arbitary value

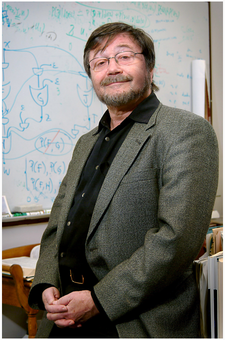
GPSS

IUDICIUM POSTERIUM DISCIPULUS EST PRIORIS

Today

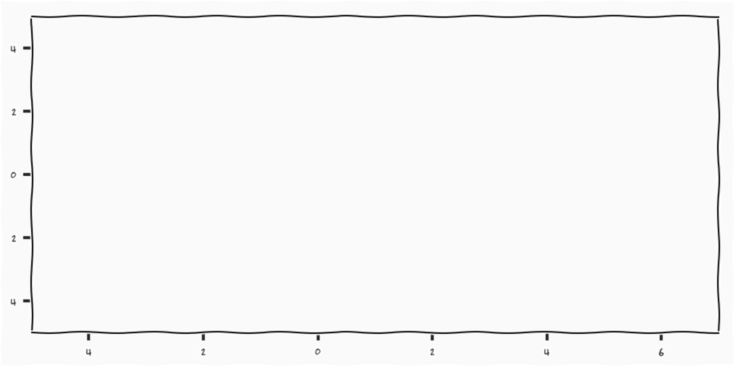
September 5, 2018

Learning

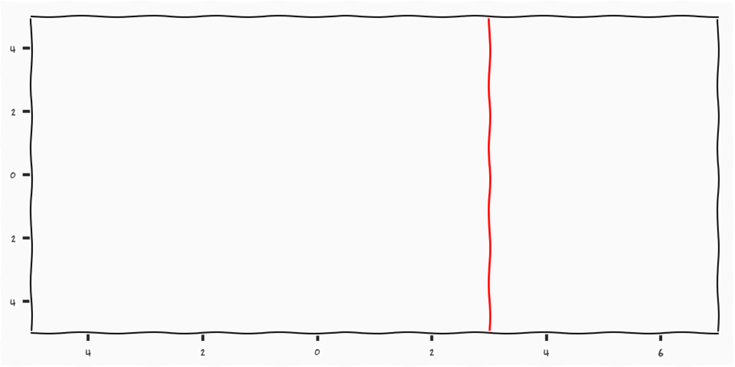


"Machine Learning is nothing but curve fitting, but its amazing what you can do by fitting curves."
– Judea Pearl

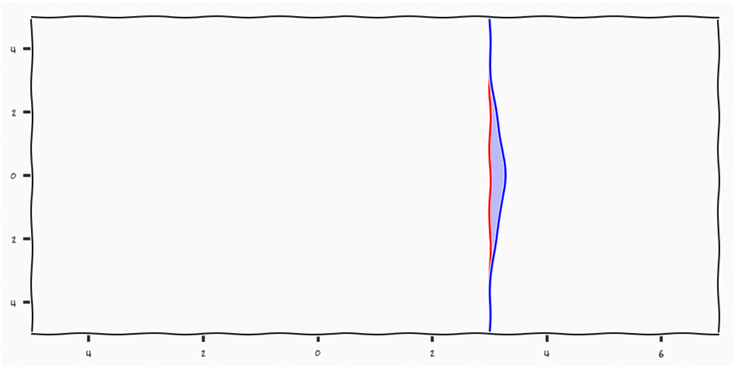
Gaussian Processes



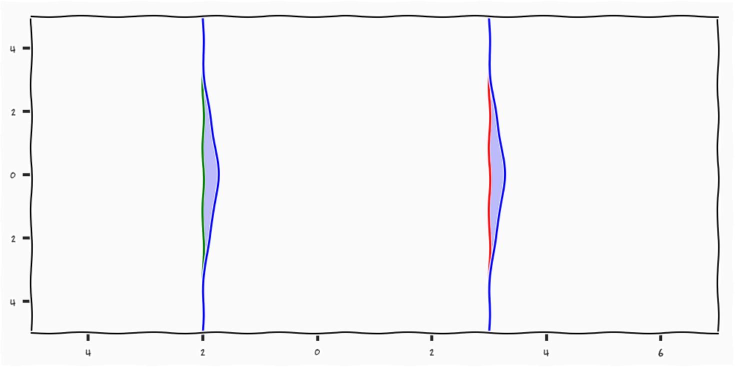
Gaussian Processes



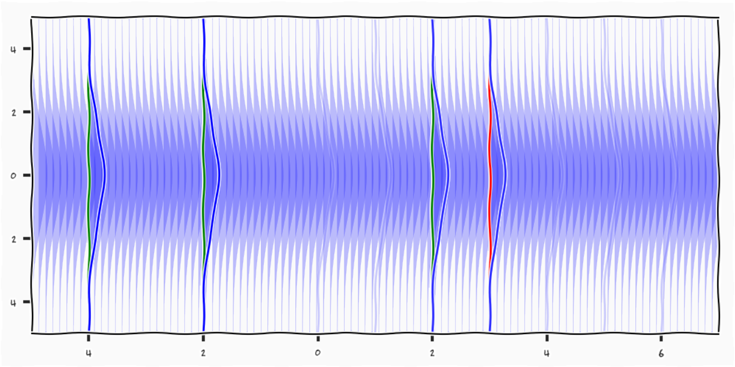
Gaussian Processes

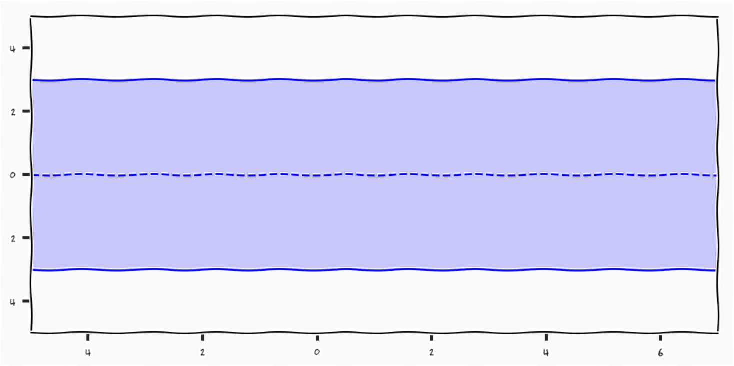


Gaussian Processes

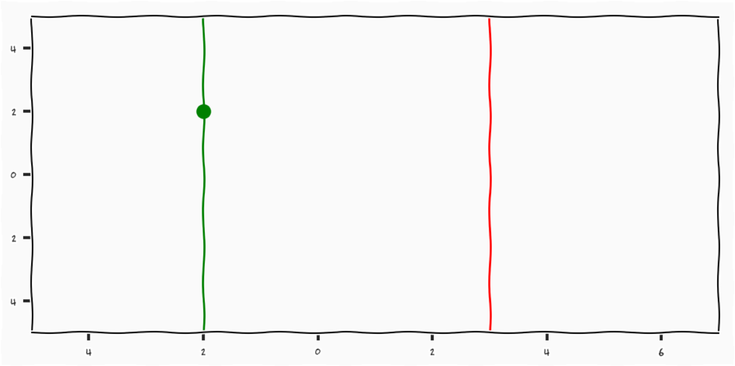


Gaussian Processes

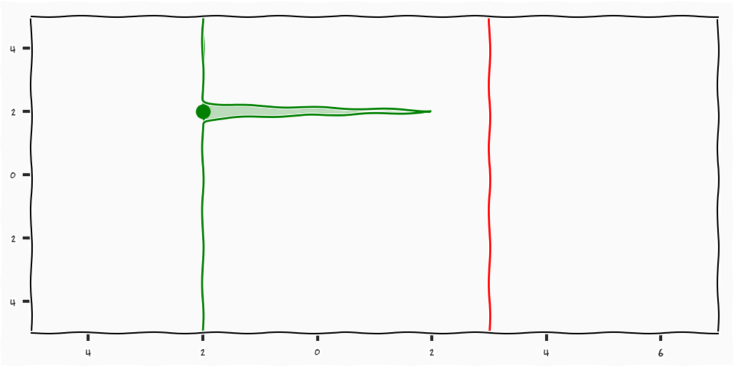



4
2
0
2
4
4
2
0
2
4
6

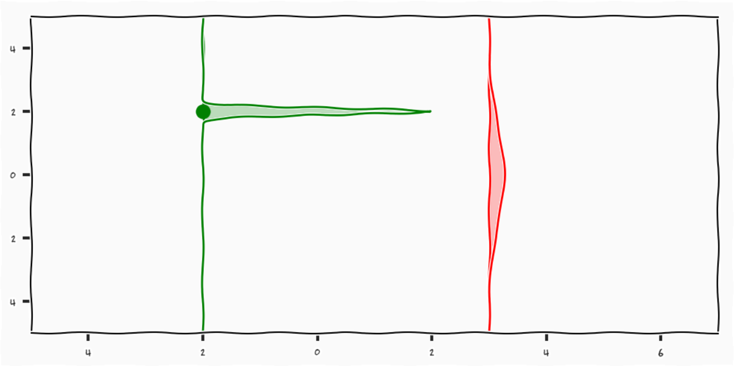
Gaussian Processes



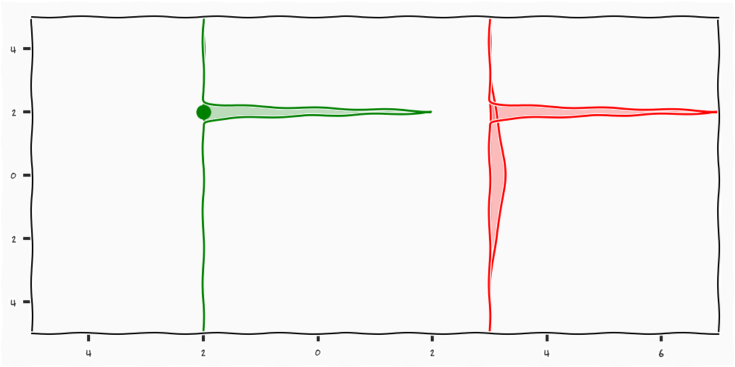
Gaussian Processes



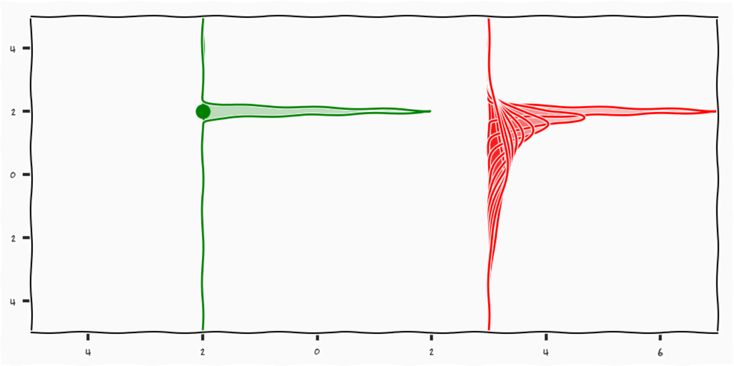
Gaussian Processes



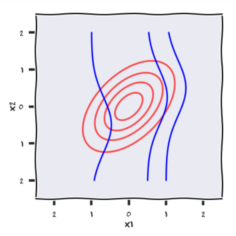
Gaussian Processes



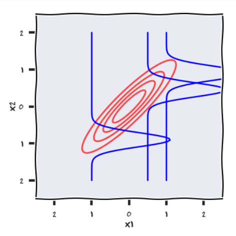
Gaussian Processes



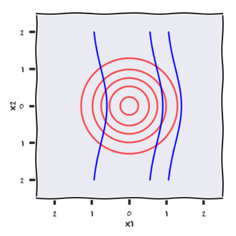
Conditional Gaussians



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

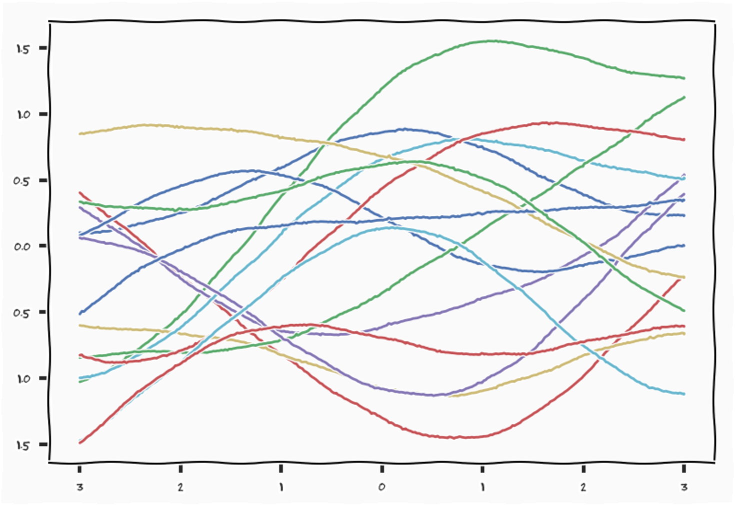


$$N\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}1 & 0.9\\0.9 & 1\end{bmatrix}\right)$$

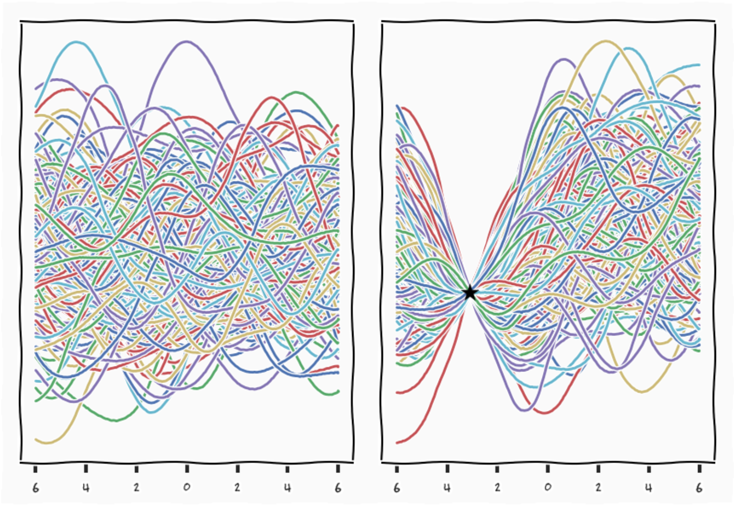


$$N\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right)$$

Gaussian Processes

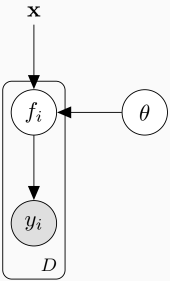


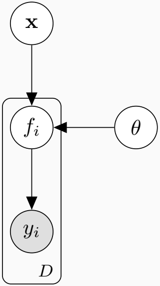
Gaussian Processes



Unsupervised Learning

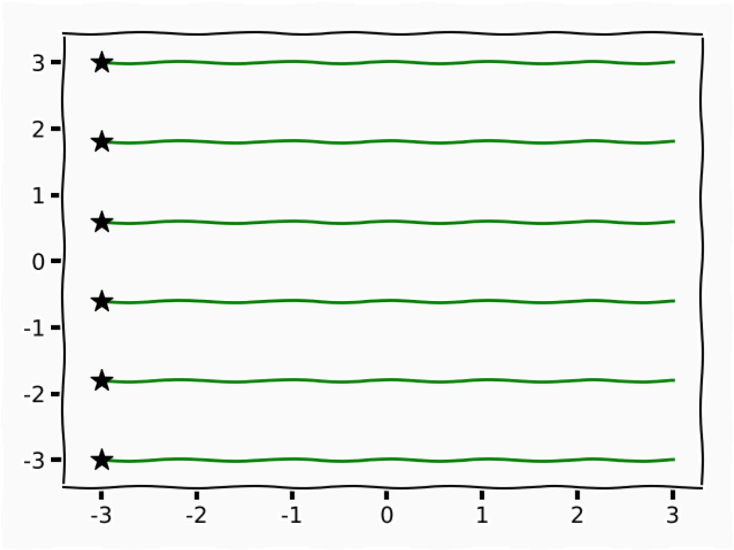
Unsupervised Learning



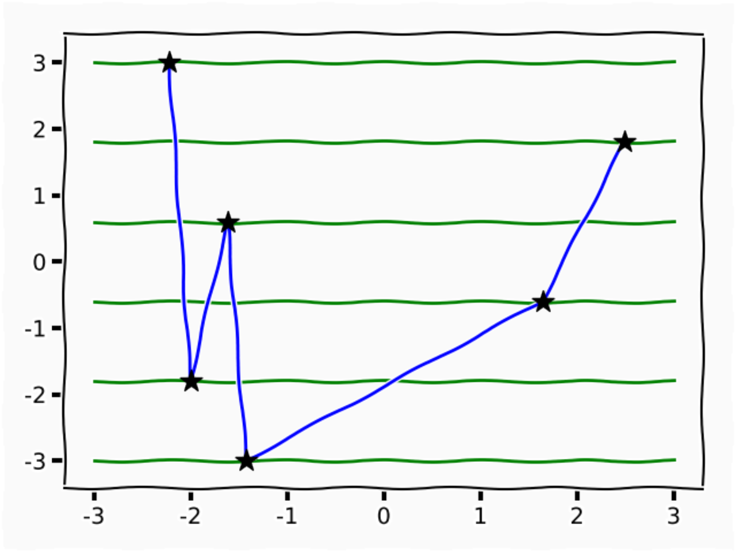


$p(y|x)$ $\qquad\qquad$ $p(y)$

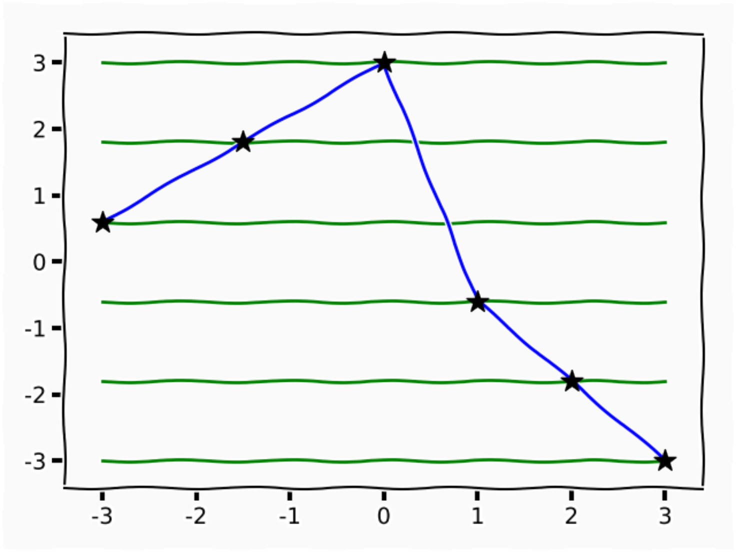
Unsupervised Learning

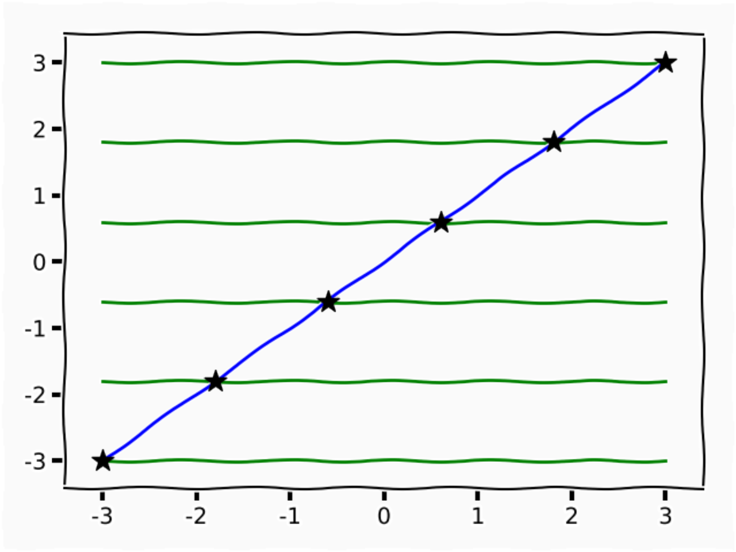


Unsupervised Learning

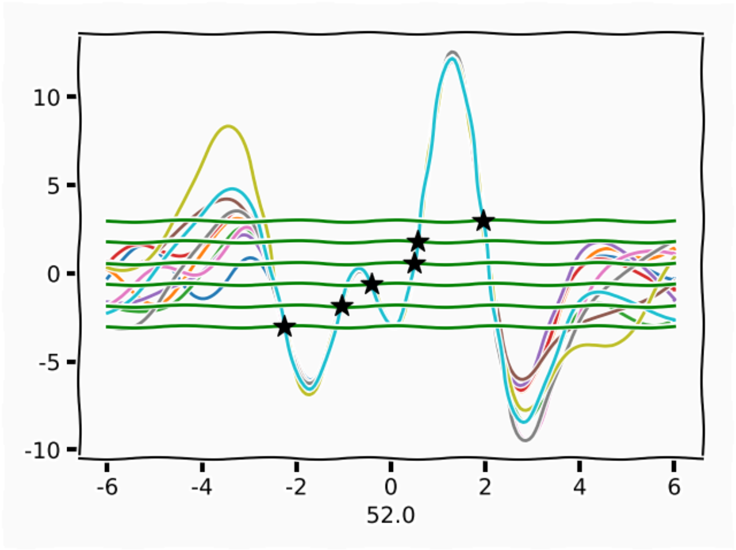


Unsupervised Learning

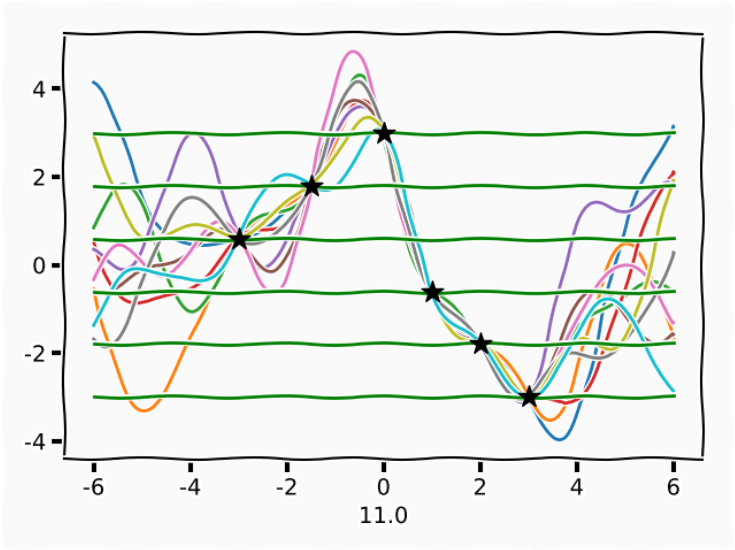



3
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3

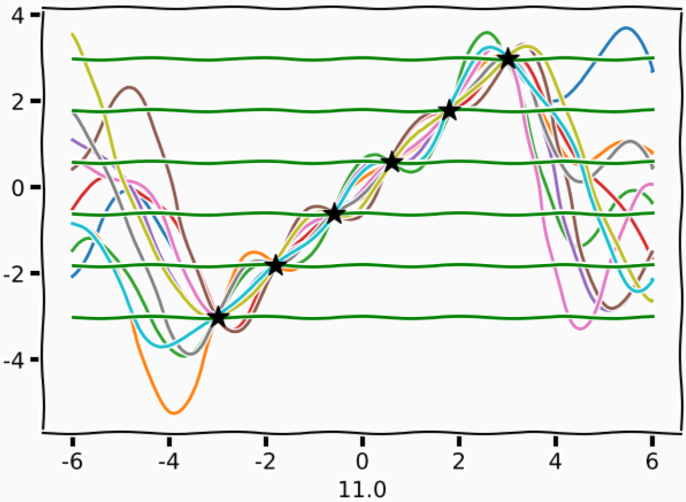
Unsupervised Learning



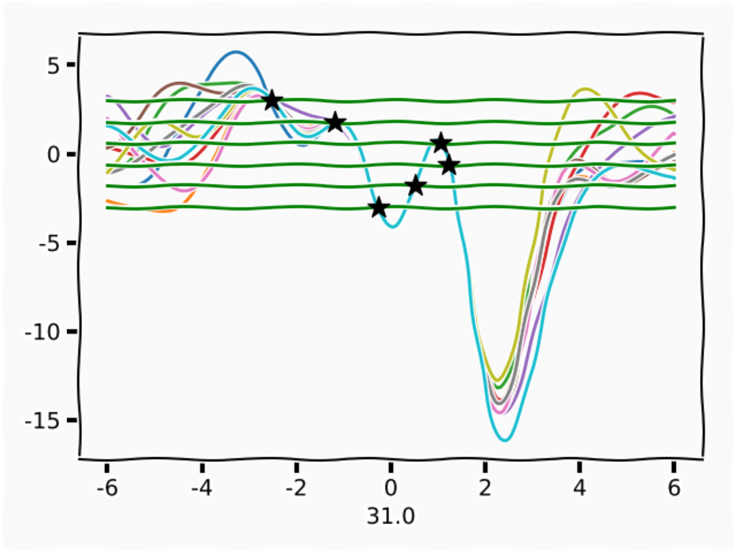
Unsupervised Learning



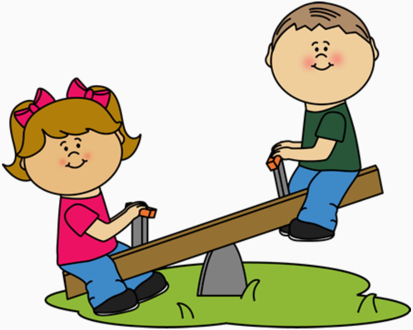
Unsupervised Learning



Unsupervised Learning



Priors



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. 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

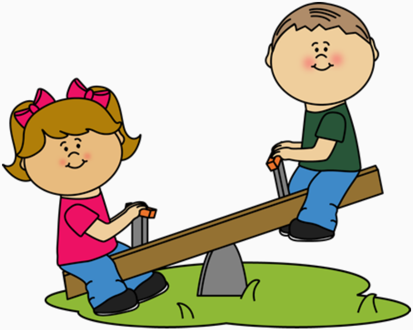
$$p(y|f) \sim N(y|f, \beta) \propto e^{-\frac{1}{2\beta}\mathrm{tr}(y-f)^{\mathrm{T}}(y-f)}$$

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

Laplace Integration

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

Priors



Being Bayesian[1]

[1]By Dieric Bouts (circa 1420-1475) - The Yorck Project: 10.000 Meisterwerke der Malerei, Public Domain, URL

Unnsupervised Learning with GPs

ML/MAP[2]

$$\hat{x} = \text{argmax}_x \int p(y|f)p(f|x)\mathrm{d}f p(x)$$

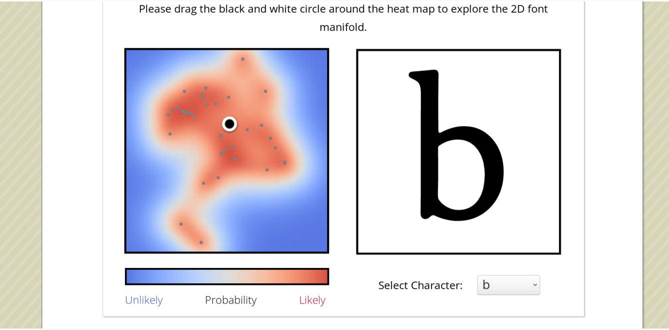
$$= \text{argmin}_x \frac{1}{2} y^{\mathrm{T}} \mathbf{K}^{-1} y + \frac{1}{2} |\mathbf{K}| - \log\, p(x)$$

[2]Lawrence, N. D. (2005). Probabilistic non-linear principal component analysis with Gaussian process latent variable models.

GP-LVM

- Li, W., Viola, F., Starck, J., Brostow, G. J., & Campbell, N. D. (2016). Roto++: accelerating professional rotoscoping using shape manifolds. (In proceeding of ACM SIGGRAPH'16)
- Grochow, K., Martin, S. L., Hertzmann, A., & Popovi\'c, Zoran (2004). Style-based inverse kinematics. SIGGRAPH '04: SIGGRAPH 2004
- Urtasun, R., Fleet, D. J., & Fua, P. (2006). 3D people tracking with Gaussian process dynamical models. Computer Vision and Pattern Recognition, 2006

Font Demo



URL

Bayesian GP-LVM[4]

- Challenges with ML estimation
  - How to initialise x?
  - What is the dimensionality q?
- *Our assumption on the latent space does not reach the data*

---

[3]Titsias, M. (2009). Variational learning of inducing variables in sparse Gaussian processes.

[4]Titsias, M., & Lawrence, N. D. (2010). Bayesian Gaussian Process Latent Variable Model

Bayesian GP-LVM[4]

- Challenges with ML estimation
  - How to initialise x?
  - What is the dimensionality q?
- *Our assumption on the latent space does not reach the data*
- Approximate integration![3]

[3] Titsias, M. (2009). Variational learning of inducing variables in sparse Gaussian processes.

[4] Titsias, M., & Lawrence, N. D. (2010). Bayesian Gaussian Process Latent Variable Model

Variational Bayes

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

Variational Bayes

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

Variational Bayes

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

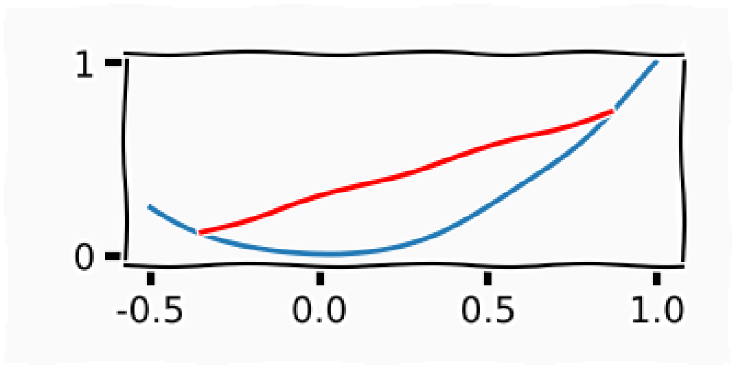
Variational Bayes

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

Variational Bayes

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

Jensen Inequality



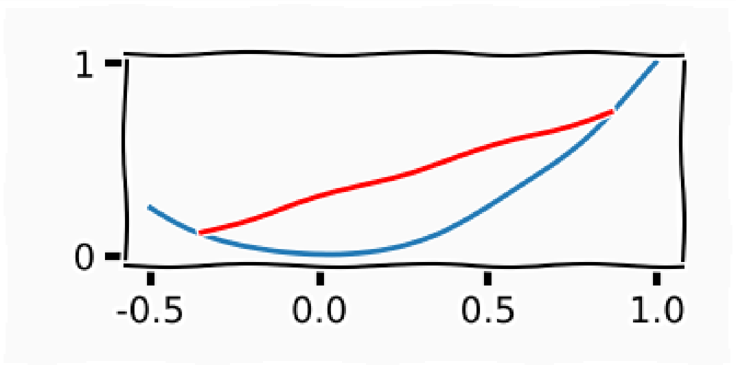
Convex Function

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

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

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

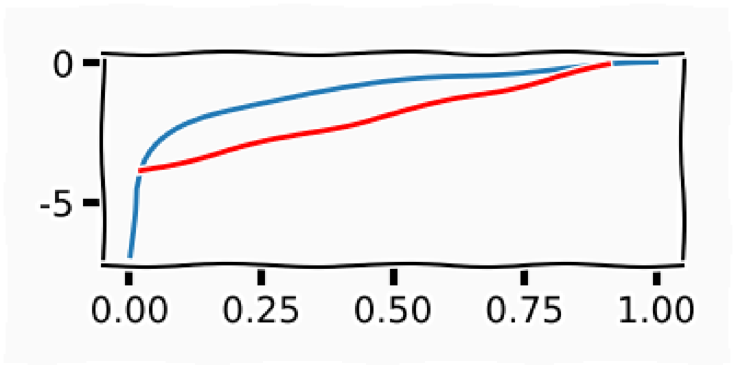
Jensen Inequality



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

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

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.

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

Variational Bayes cont.

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

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

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

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

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

ELBO

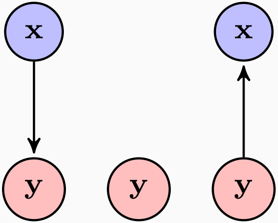
$$
\begin{aligned}
log\ p(\mathbf{Y}) &= \mathrm{KL}(q(\mathbf{X})||p(\mathbf{X}|\mathbf{Y})) + \underbrace{\mathbb{E}_{q(\mathbf{X})}\left[log\ p(\mathbf{X}, \mathbf{Y})\right] - H(q(\mathbf{X}))}_{\mathrm{ELBO}} \\
&\geq \mathbb{E}_{q(\mathbf{X})}\left[log\ p(\mathbf{X}, \mathbf{Y})\right] - H(q(\mathbf{X}))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

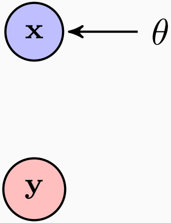
ELBO

$$
\begin{aligned}
\log\ p(\mathbf{Y}) &= \mathrm{KL}(q(\mathbf{X})||p(\mathbf{X}|\mathbf{Y})) + \underbrace{\mathbb{E}_{q(\mathbf{X})}\left[\log\ p(\mathbf{X},\mathbf{Y})\right] - H(q(\mathbf{X}))}_{\text{ELBO}} \\
&\geq \mathbb{E}_{q(\mathbf{X})}\left[\log\ p(\mathbf{X},\mathbf{Y})\right] - H(q(\mathbf{X})) = \mathcal{L}(q(\mathbf{X}))
\end{aligned}
$$

- 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





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

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

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[5]

[5] 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[5]

[5] 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[5]

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

Lower Bound[6]

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

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

Lower Bound[6]

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

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

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

Lower Bound[6]

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

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

Lower Bound[6]

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

[6] 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[7]

[7] 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,

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

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

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

- 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

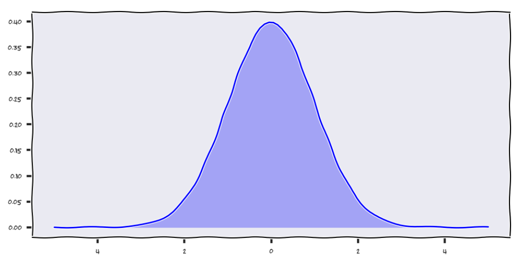
Latent space priors[8]

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

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

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

The Gaussian blob



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

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

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

Change of Variables

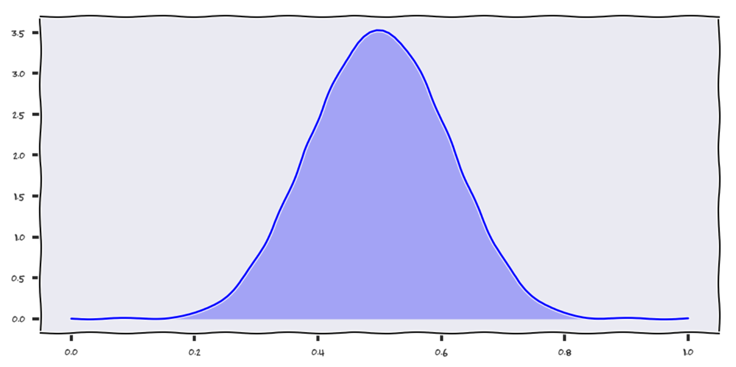
Theorem (Change of Variable)

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

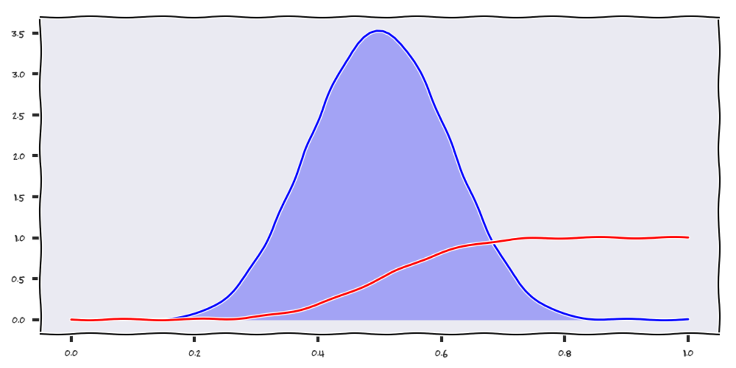
$$p_y(y) = p_x(\psi(y))| \bigtriangledown \psi(y)|$$

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

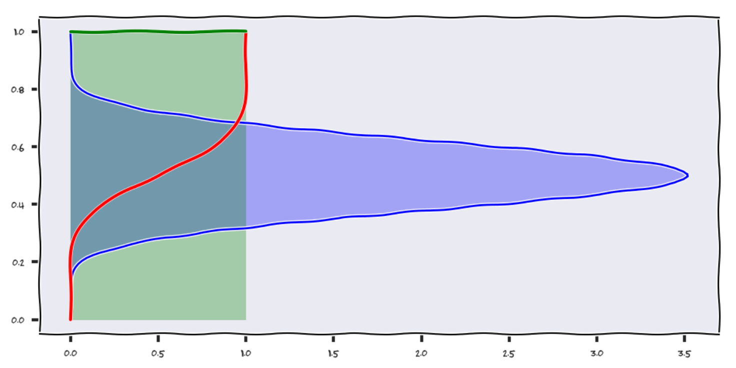
Sampling



Sampling



Sampling



Change of Variables

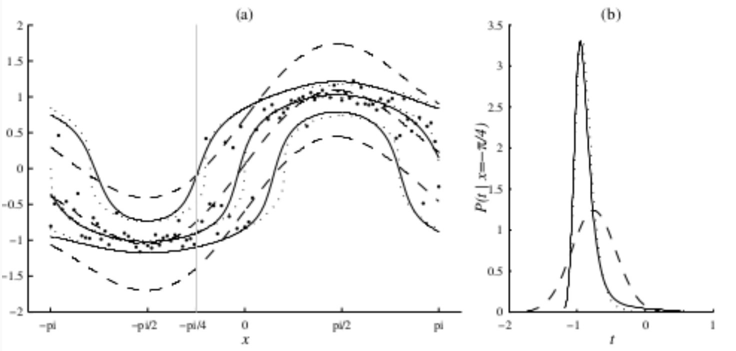
Theorem (Change of Variable)

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.

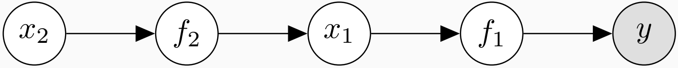
Warped Gaussian Processes[9, 10]



[9] Snelson, E., & Ghahramani, Z. (2004). Warped Gaussian Processes

[10] Lazaro-Gredilla, Miguel (2012). Bayesian Warped Gaussian Processes. In , Advances in Neural Information Processing Systems

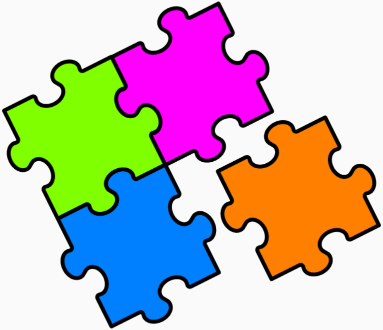
Deep Gaussian Processes[11]



- Place a GP as a warping function, that is warped, ...

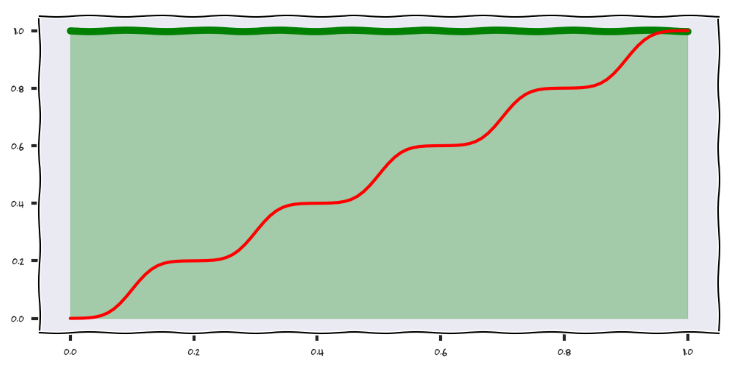
[11] Damianou, A. C., & Lawrence, N. D. (2013). Deep Gaussian Processes

Composite Functions

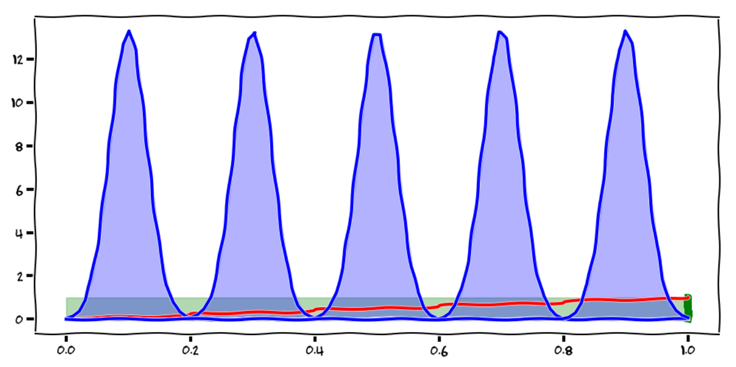


$$y = f_k(f_{k-1}(\dots f_0(x))) = f_k \circ f_{k-1} \circ \dots \circ f_1(x)$$

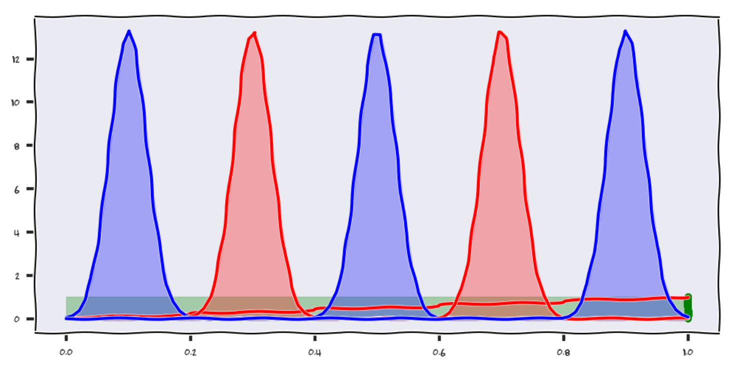
Change of Variables



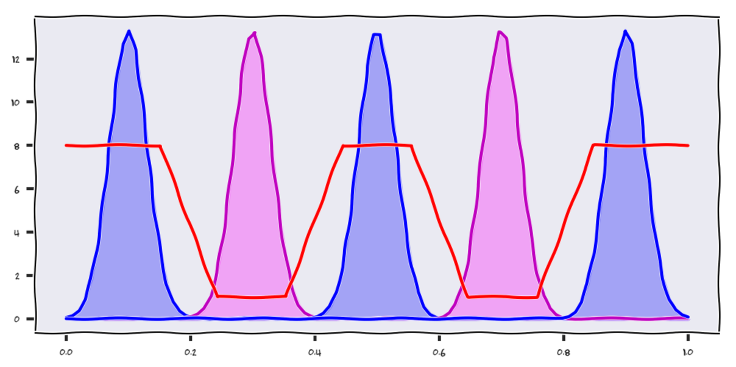
Change of Variables



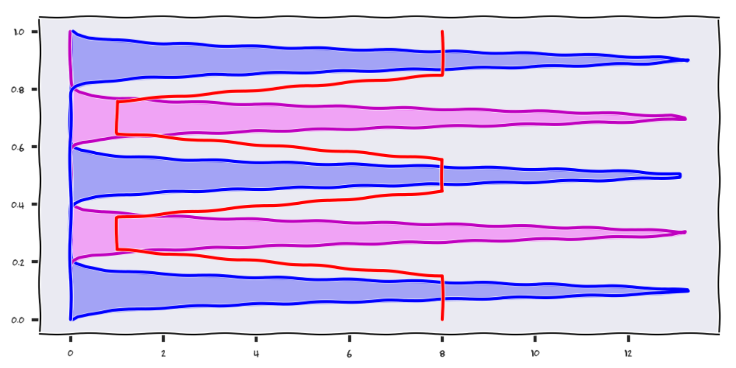
Change of Variables



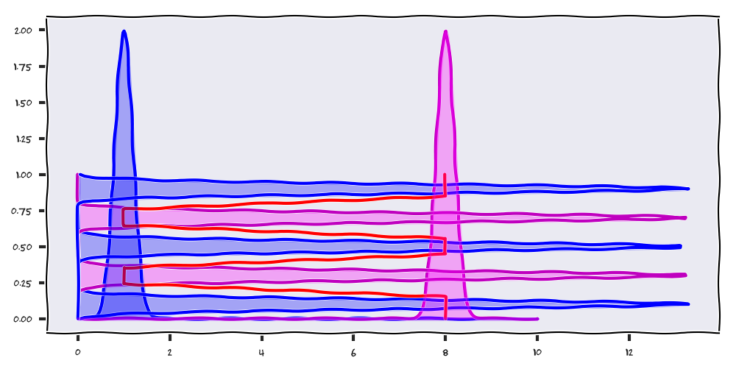
Change of Variables



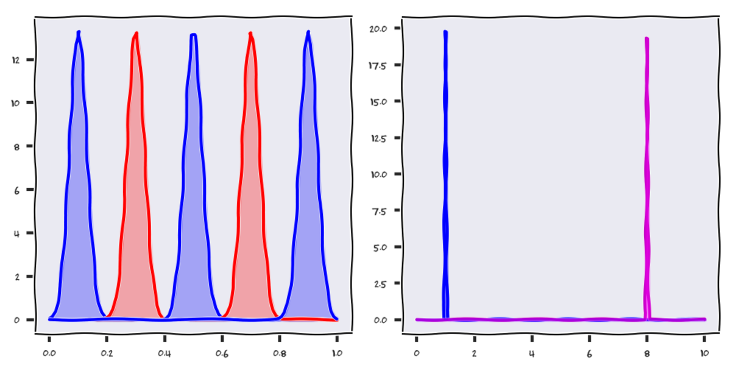
Change of Variables



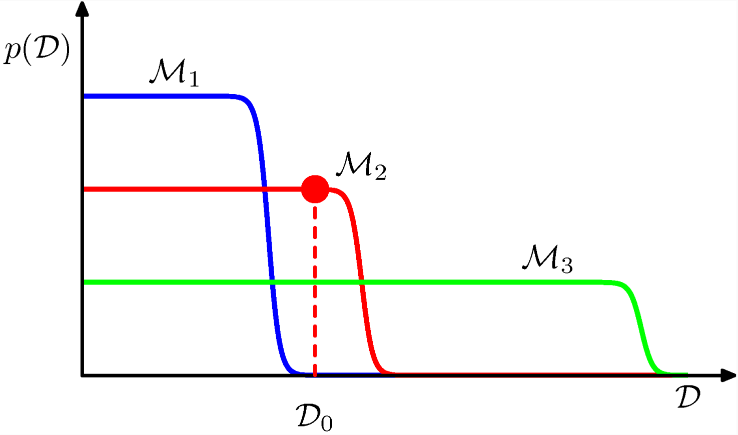
Change of Variables



Change of Variables

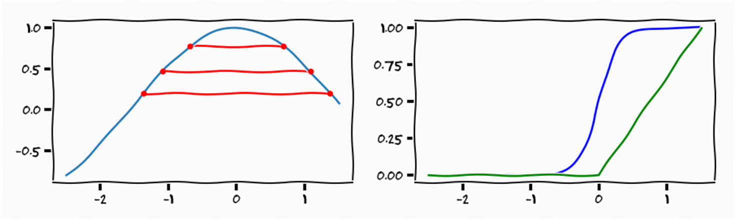


MacKay plot



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

Composition functions

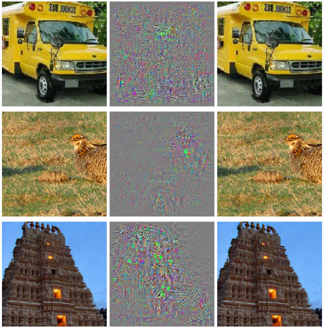


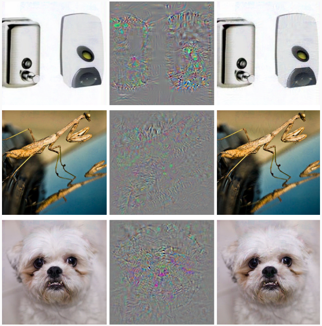
$$y = f_k(f_{k-1}(\dots f_0(x))) = f_k \circ f_{k-1} \circ \dots \circ f_1(x)$$

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

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

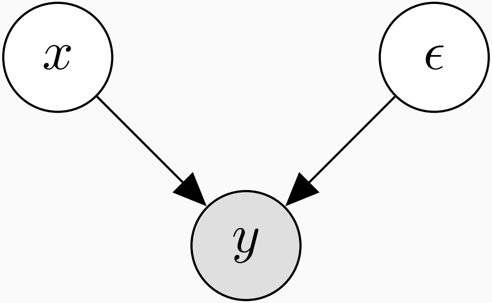
Data inefficiency[12]





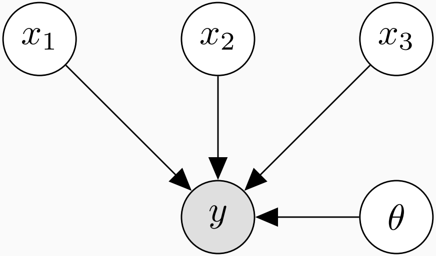
[12]Nguyen, A. M., Yosinski, J., & Clune, J., Deep neural networks are easily fooled: high confidence predictions for unrecognizable images, CoRR, abs/1412.1897(), (2014).

Explaining Away



$$y = f(x) + \epsilon$$

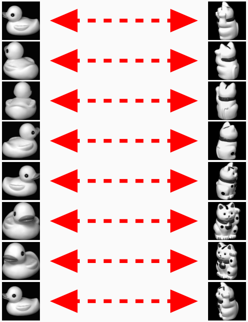
Factor Analysis



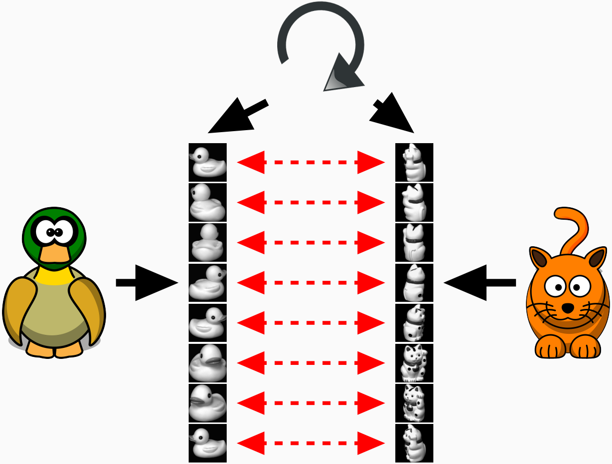
$$y = f(x_1, x_2, x_3) + \epsilon$$

Alignments

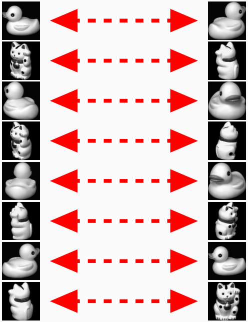
Alignments



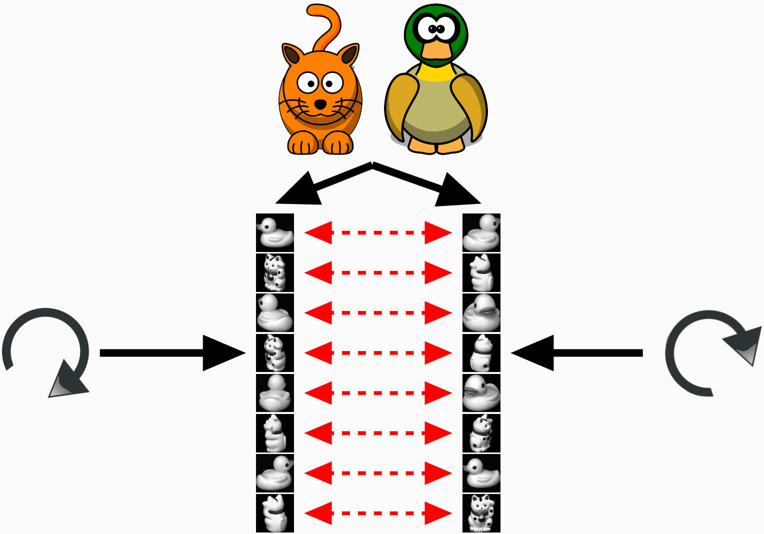
Alignments



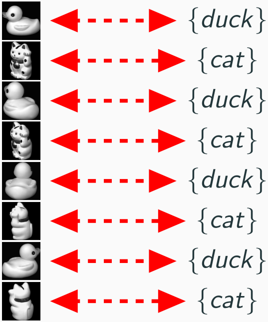
Alignments



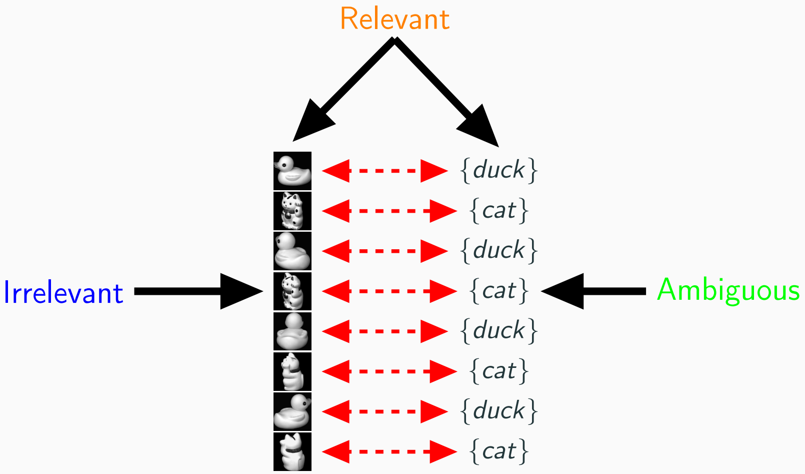
Alignments



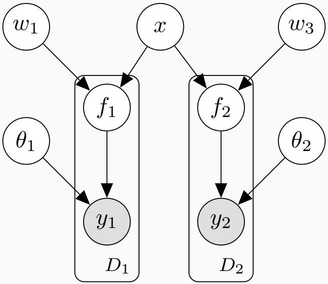
Alignments



Alignments


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

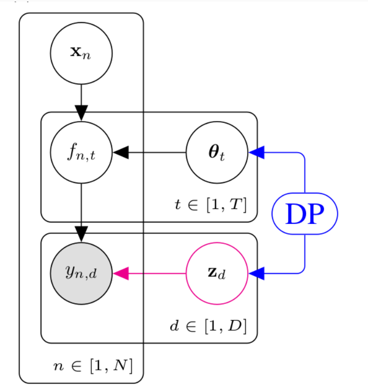
IBFA with GP-LVM[13]

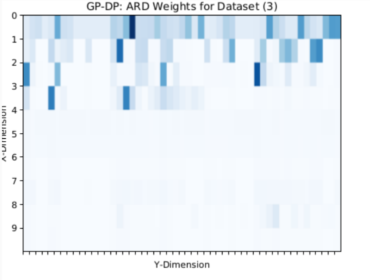


$$y_1 = f(w_1^{\mathrm{T}} x) \quad y_2 = f(w_2^{\mathrm{T}} x)$$

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

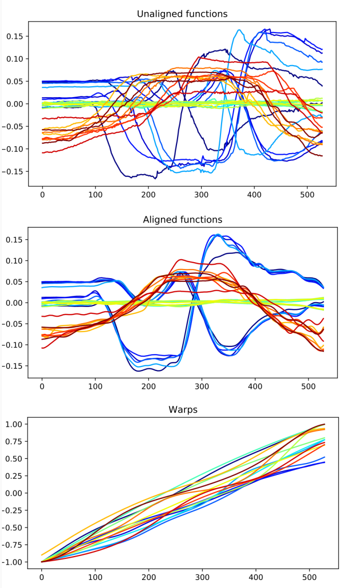
GP-DP[14]



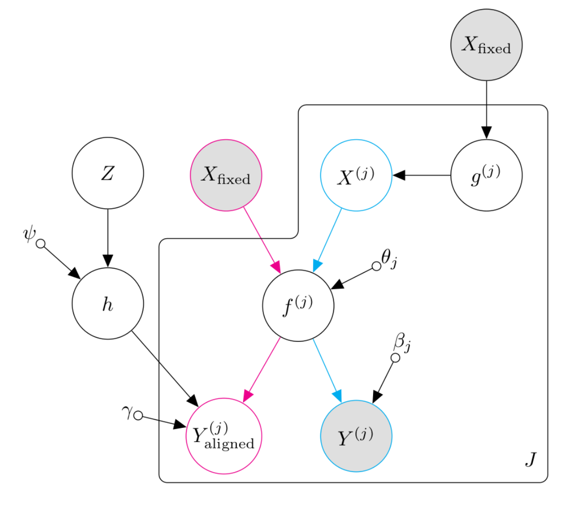


[14]Lawrence, A. R., Ek, C. H., & Campbell, N. D. F., Dp-gp-lvm: a bayesian non-parametric model for learning multivariate dependency structures, CoRR, (), (2018).

Alignment Learning

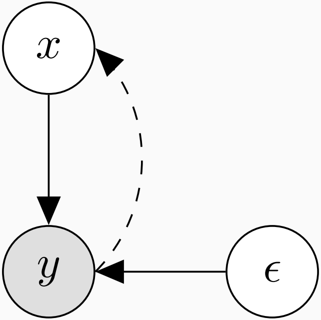


Alignment Learning[15]



[15]Kazlauskaite, I., Ek, C. H., & Campbell, N. D. F., Gaussian Process Latent Variable Alignment Learning, CoRR, (), (2018).

Constrained Latent Space



$$y = f(g(y)) + \epsilon$$

Constrained Latent Space

- Dai, Z., Damianou, A., Gonzalez, Javier, & Lawrence, N., Variational auto-encoded deep Gaussian processes, International Conference on Learning Representations (ICLR), (2016).
- Snoek, J., Adams, R. P., & Larochelle, H., Nonparametric guidance of autoencoder representations using label information, Journal of Machine Learning Research, 13(), 2567–2588 (2012).
- Ek, C. H., Torr, P. H. S., & Lawrence, N. D., Gaussian process latent variable models for human pose estimation, International conference on Machine learning for multimodal interaction, (), 132–143 (2007).

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 such as 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

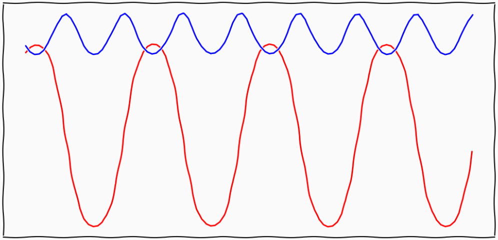
Summary II

- 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

Appendix

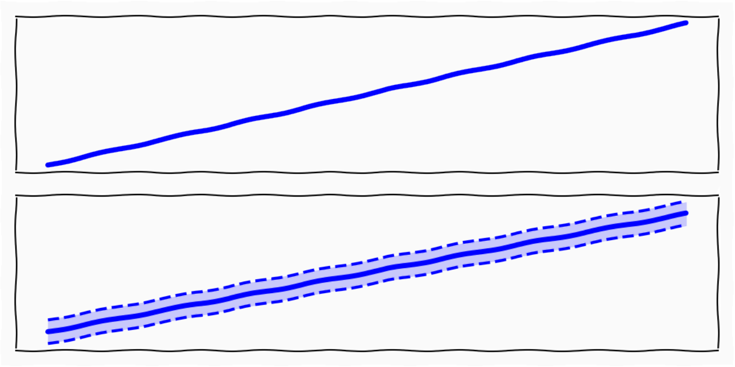
Composition: priors



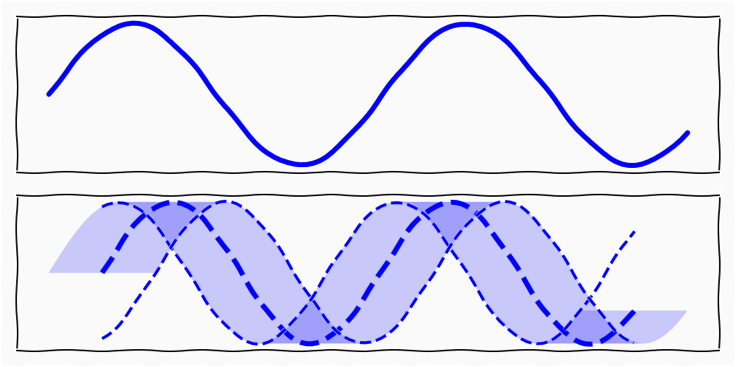
Composition: priors

./bin/composition2.png

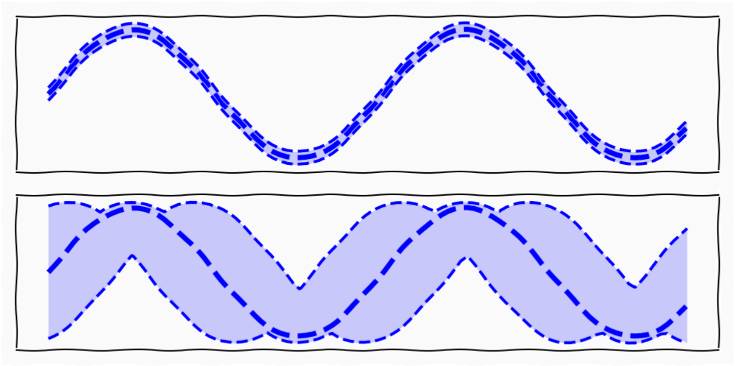
Composition: uncertainty



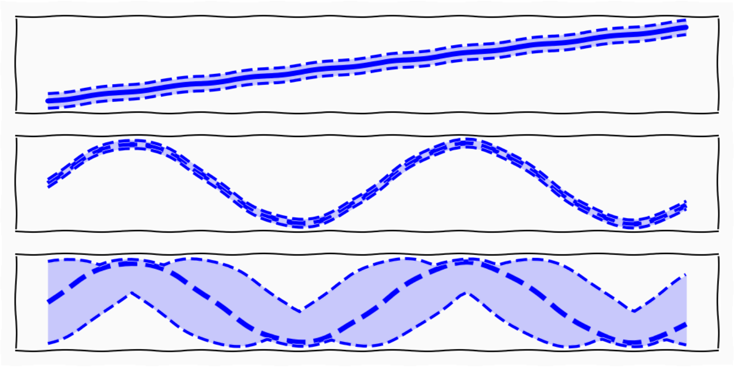
Composition: uncertainty



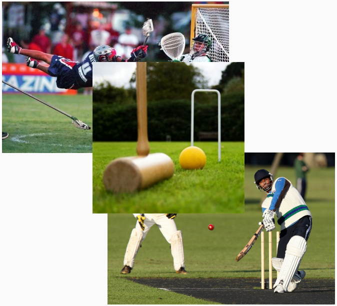
Composition: uncertainty



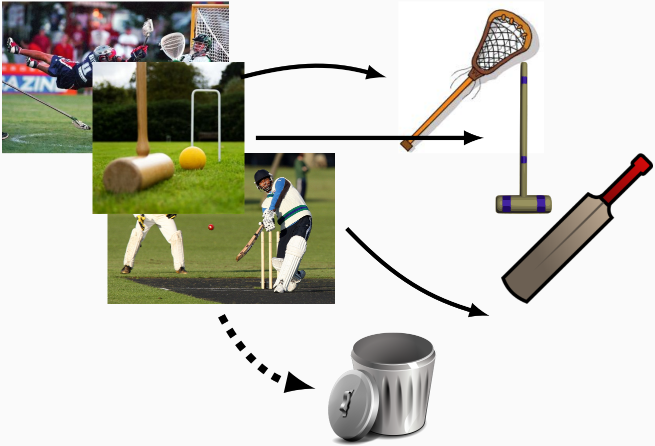
Composition: uncertainty



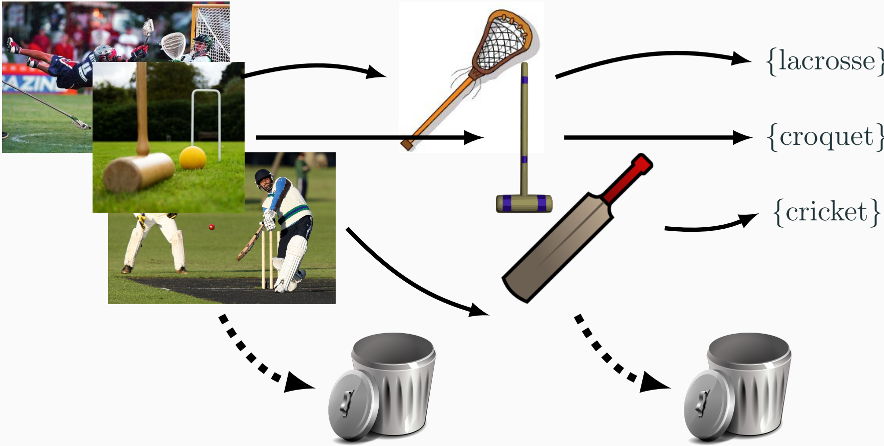
Latent Structure



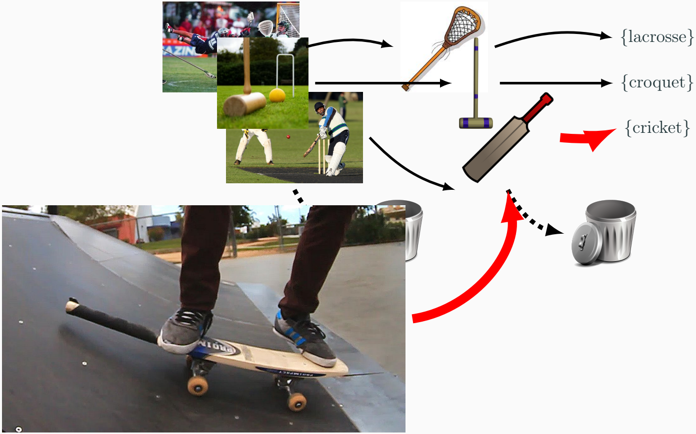
Latent Structure



Latent Structure

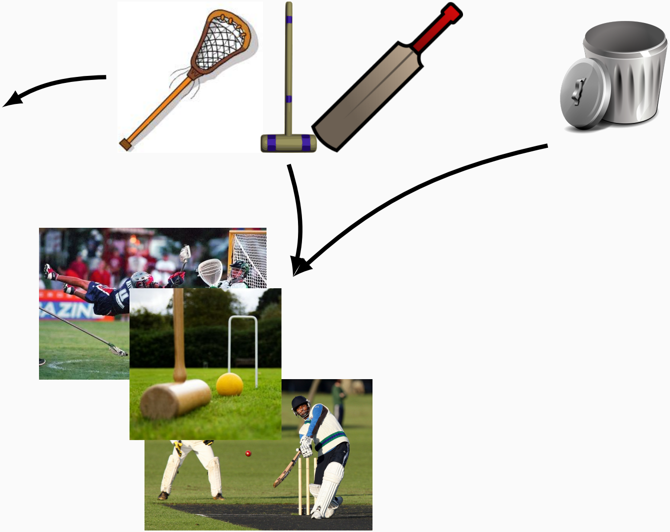


Latent Structure



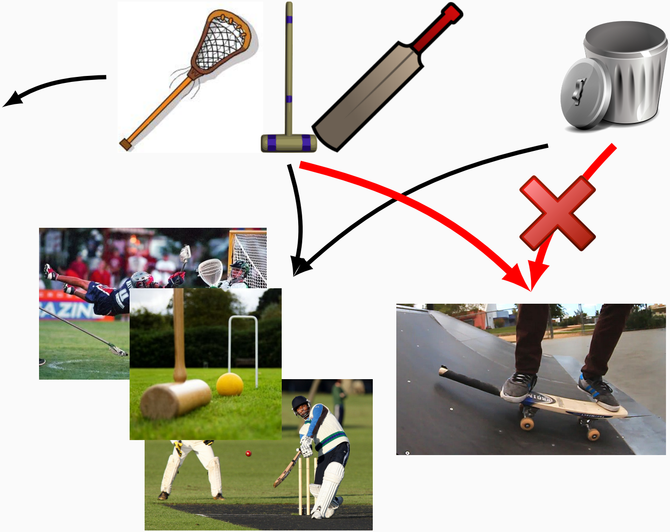
Latent Structure

{lacrosse}
{croquet}
{cricket}



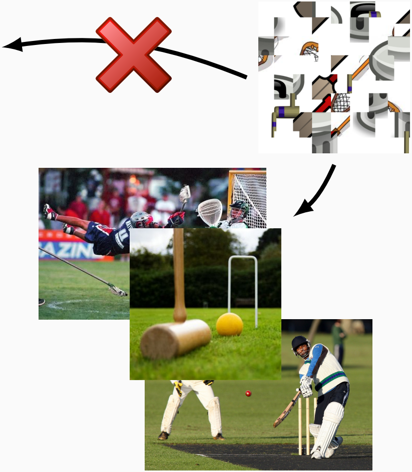
Latent Structure

{lacrosse}
{croquet}
{cricket}

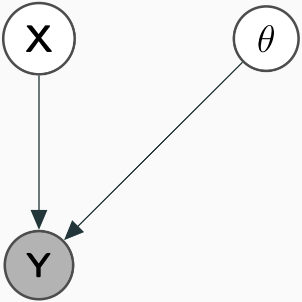


Latent Structure

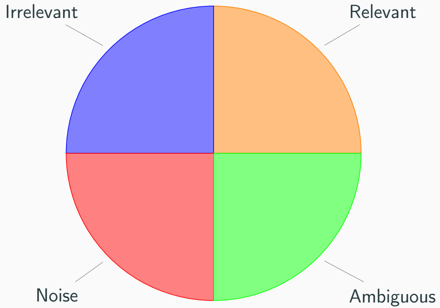
{lacrosse}
{croquet}
{cricket}



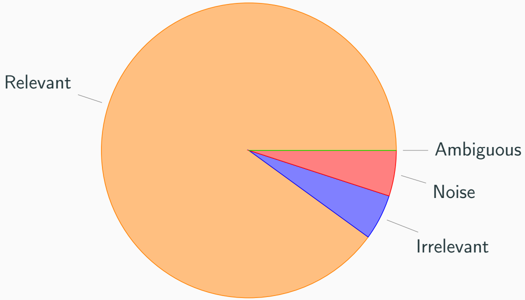
Latent Structure



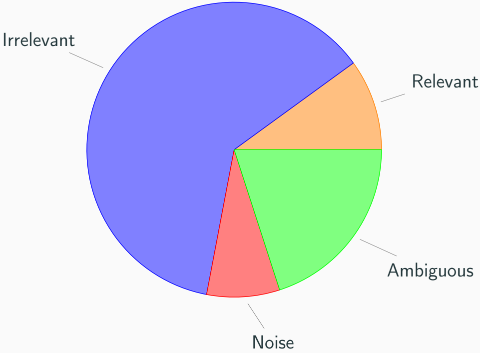
Latent Structure



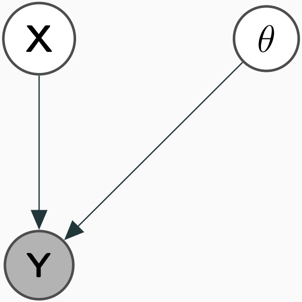
Latent Structure



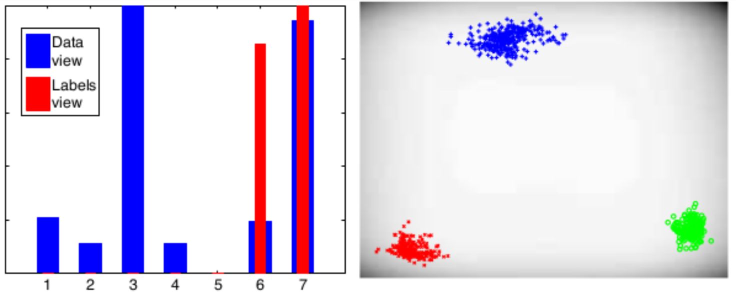
Latent Structure



Latent Structure



IBFA with GP-LVM



IBFA with GP-LVM