Unsupervised and Composite Gaussian Processes

Carl Henrik Ek - che29@cam.ac.uk

September 15, 2020

http://carlhenrik.com

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

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

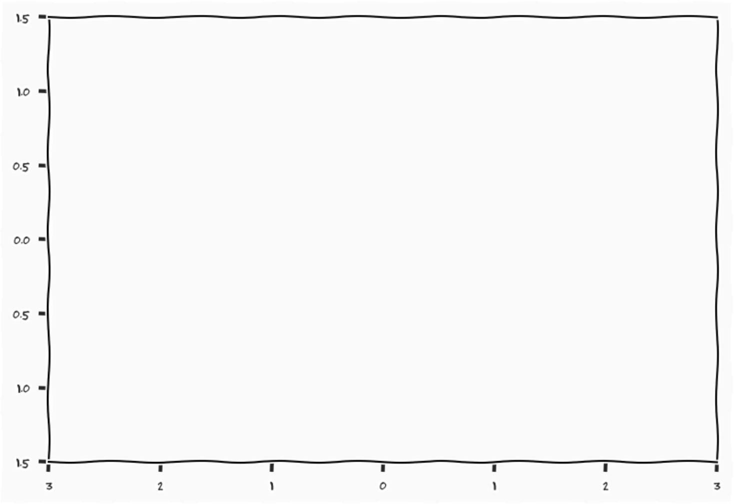
Statistical Learning

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

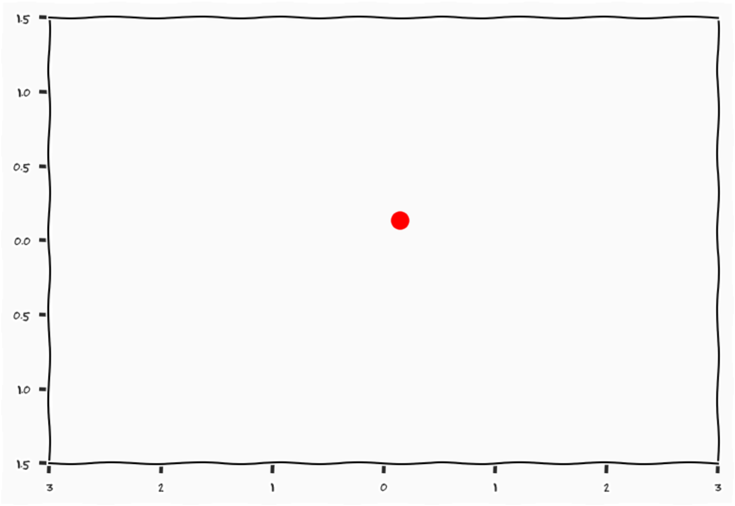
No Free Lunch

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

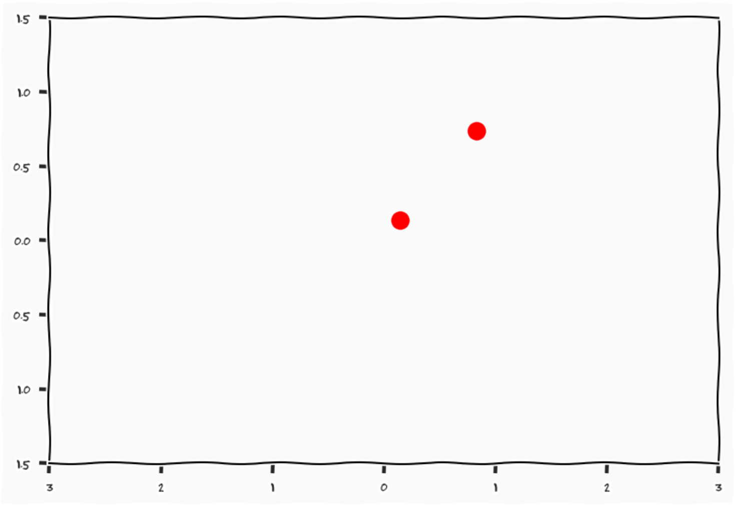
Example



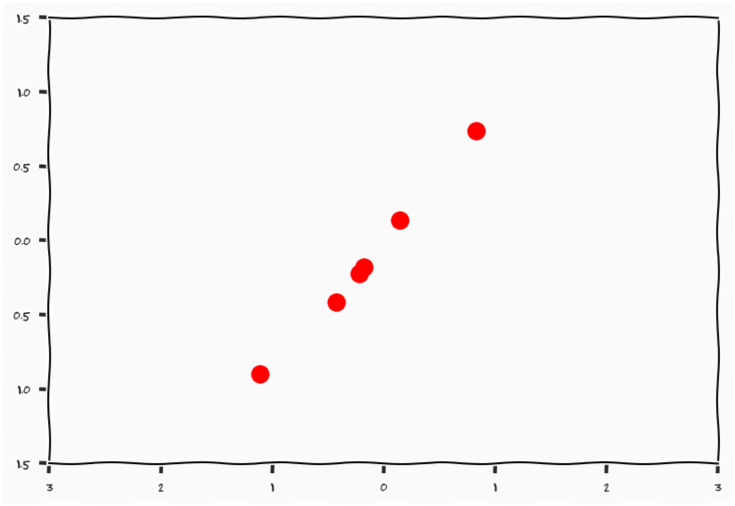
Example



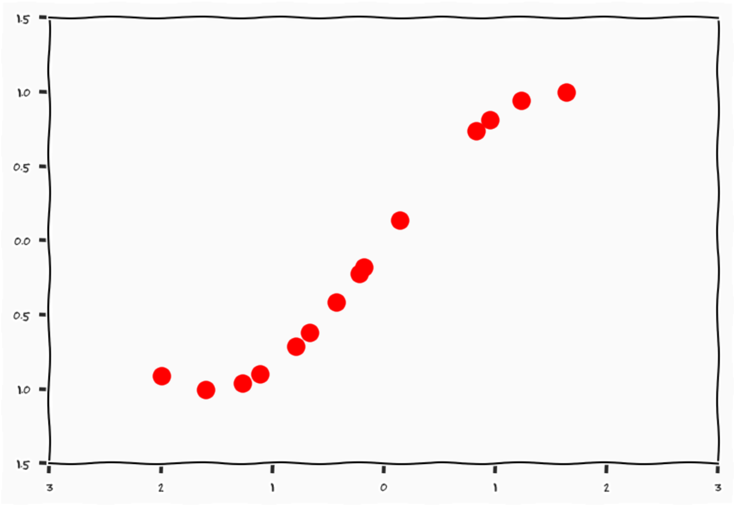
Example



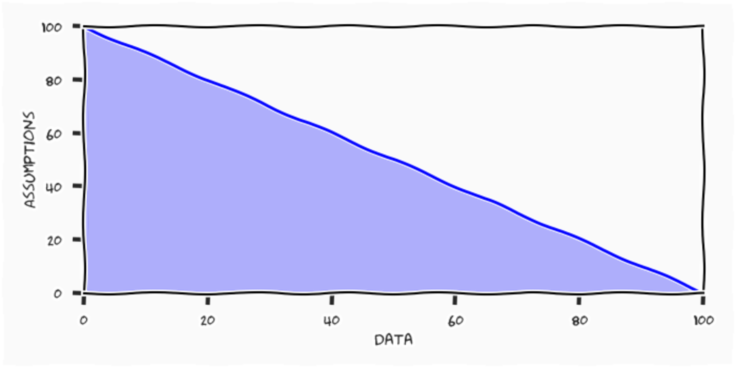
Example



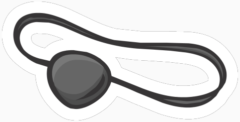
Example



Data and Knowledge



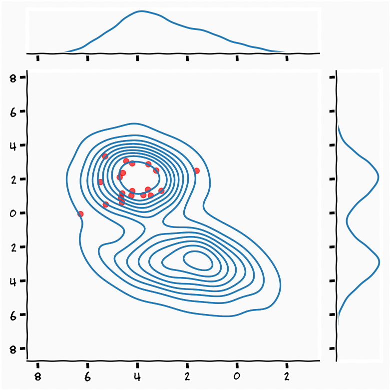
Assumptions: Algorithms



Statistical Learning

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

Assumptions: Biased Sample



Statistical Learning

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

Assumptions: Hypothesis space

Statistical Learning

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

The No Free Lunch

- There seems to be a narrative that the more *flexible* a model is the better it is

The No Free Lunch

- There seems to be a narrative that the more *flexible* a model is the better it is
  - This is not true

The No Free Lunch

- There seems to be a narrative that the more *flexible* a model is the better it is
  - This is not true
- The best possible model has infinite support (nothing is excluded) but very focused mass

The No Free Lunch

- There seems to be a narrative that the more *flexible* a model is the better it is
  - This is not true
- The best possible model has infinite support (nothing is excluded) but very focused mass
- *Your solution can only ever be interpreted in the light of your assumptions*

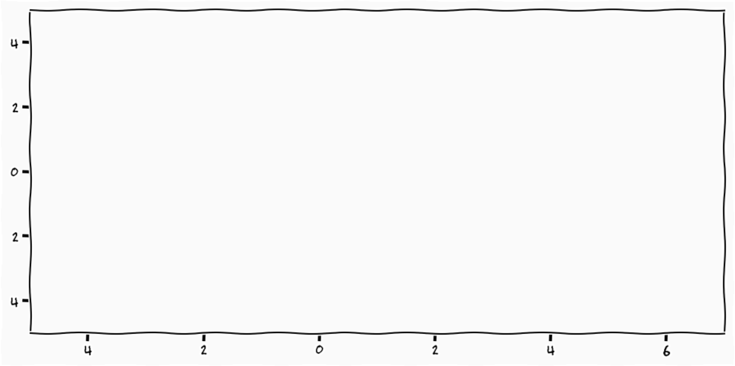
GPSS

Iudicium Posterium Discipulus Est Prioris[1]

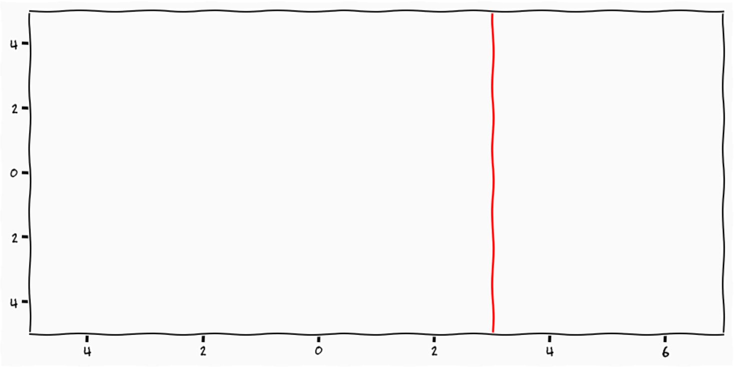
[1] The posterior is the student of the prior

Gaussian Processes

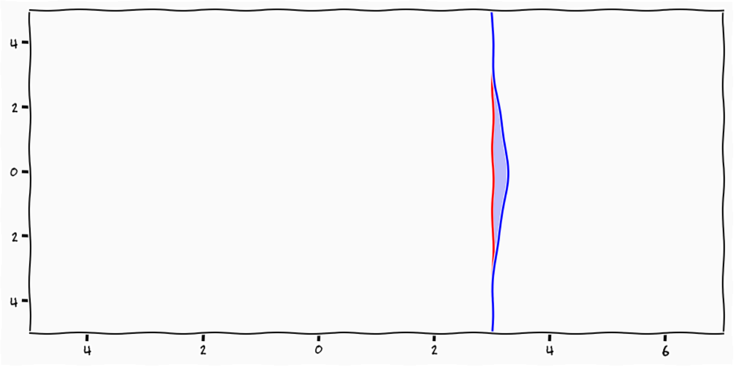
Gaussian Processes



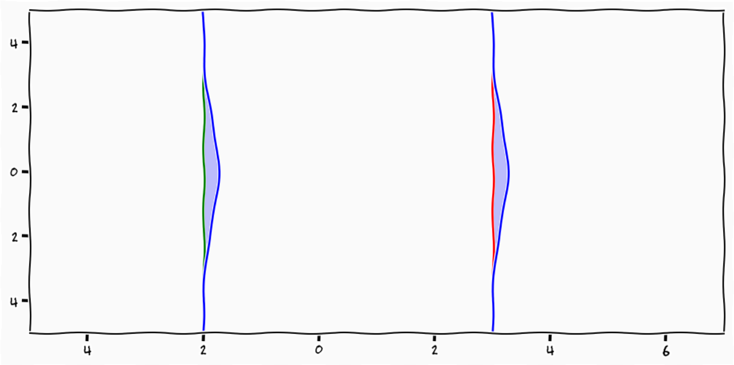
Gaussian Processes



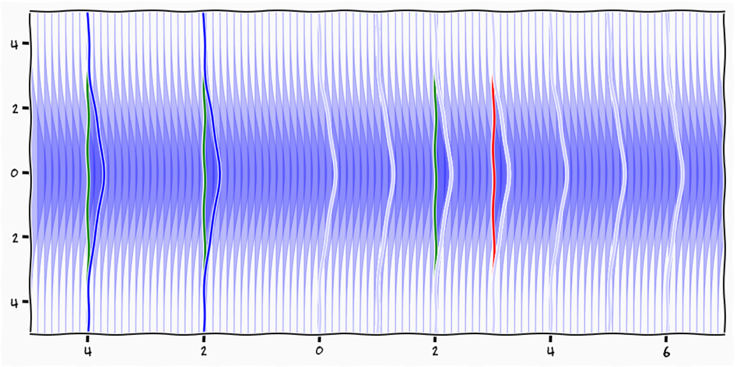
Gaussian Processes



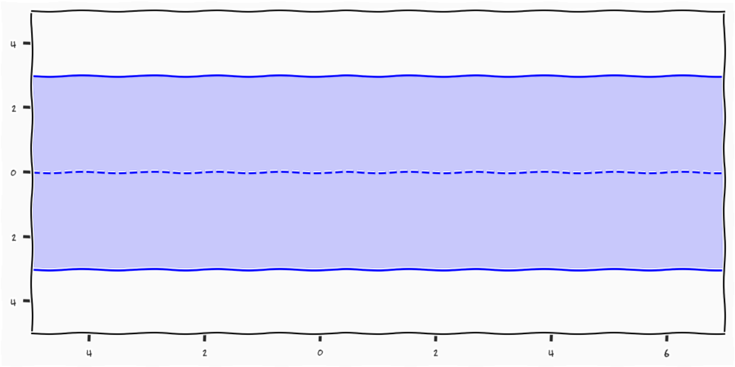
Gaussian Processes



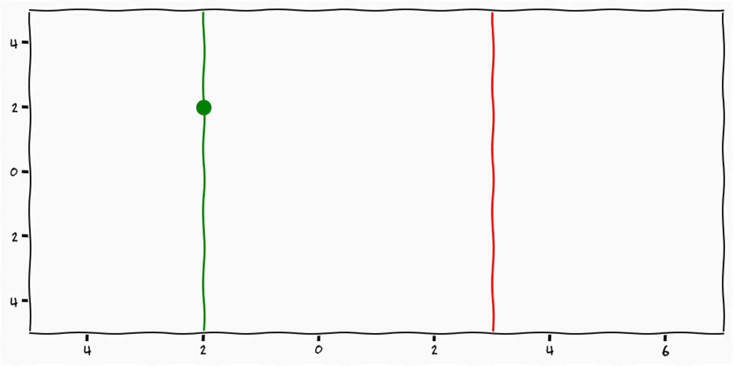
Gaussian Processes



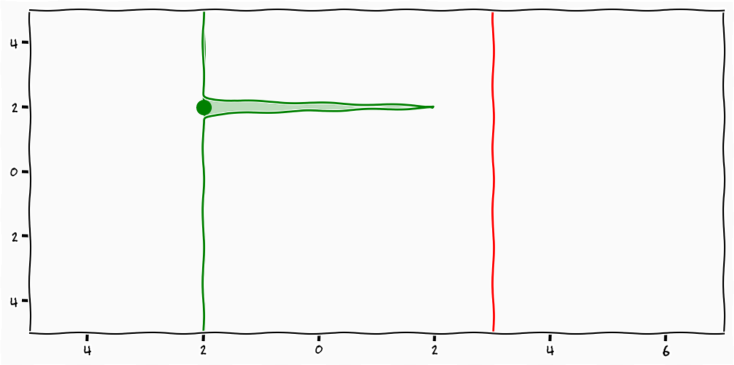
Gaussian Processes



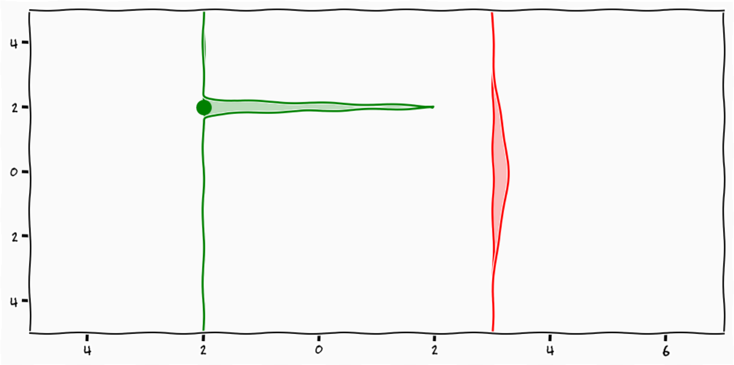
Gaussian Processes



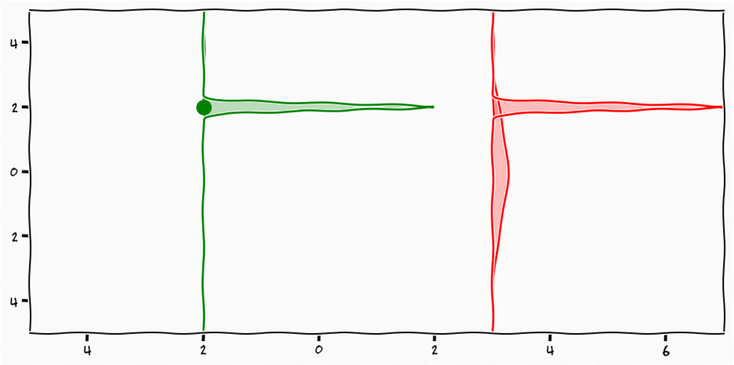
Gaussian Processes



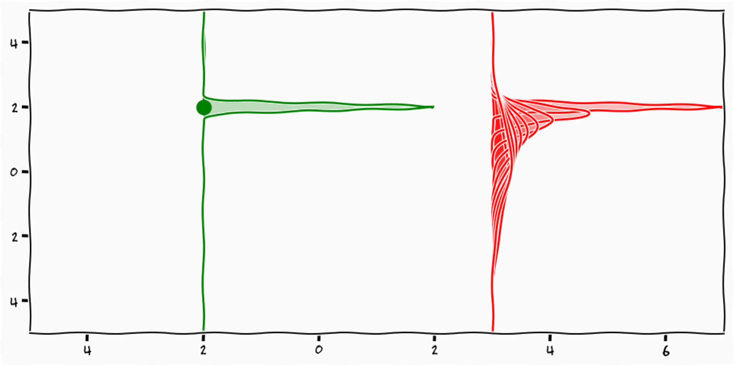
Gaussian Processes



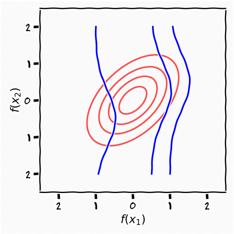
Gaussian Processes

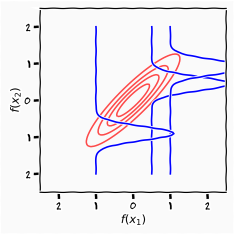


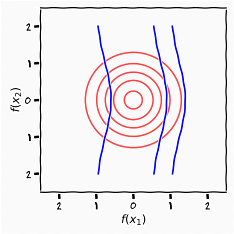
Gaussian Processes



Conditional Gaussians

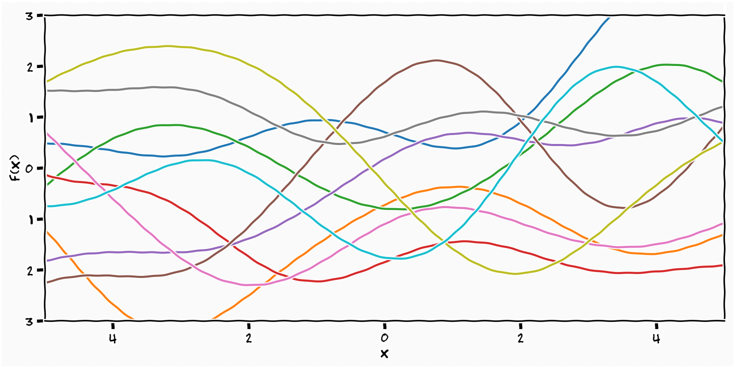




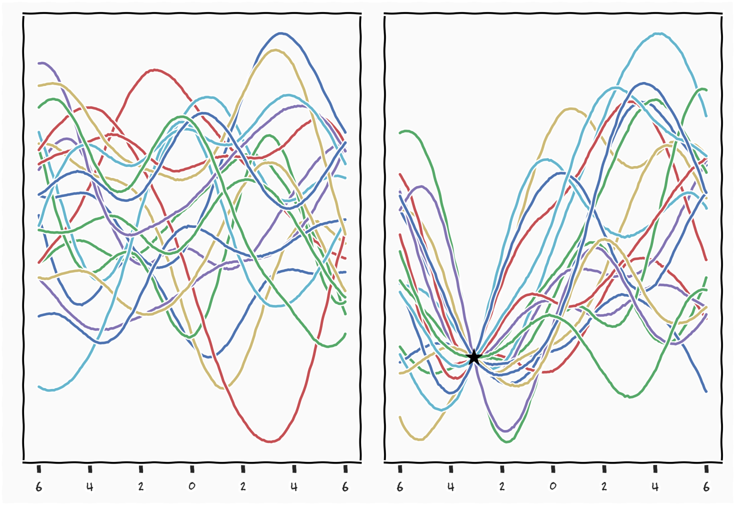


$$\mathsf{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}\right) \qquad \mathsf{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}\right) \qquad \mathsf{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix}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

Gaussian Processes



Gaussian Processes



The Gaussian Identities

$$p(x_1, x_2)$$

$$p(x_1) = \int p(x_1, x_2)\mathrm{d}x \quad p(x_1|x_2) = \frac{p(x_1, x_2)}{p(x_2)}$$

Gaussian Identities

Stochastic Processes

Kolmogrovs Excistence Theorem

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

1. Exchangeable

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

2. Marginal

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

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

Gaussian Distribution - Exchangeable

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

Gaussian Distribution - Exchangeable

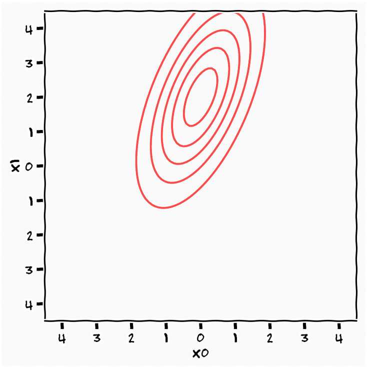
$$p(x_1, x_2) = \mathcal{N}\left(\begin{array}{c|c} x_1 & \mu_1 \\ x_2 & \mu_2 \end{array}, \begin{array}{cc} k_{11} & k_{12} \\ k_{21} & k_{22} \end{array} \right)$$

$$= p(x_2, x_1)$$

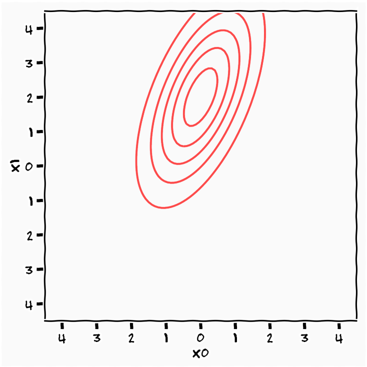
Gaussian Distribution - Exchangeable

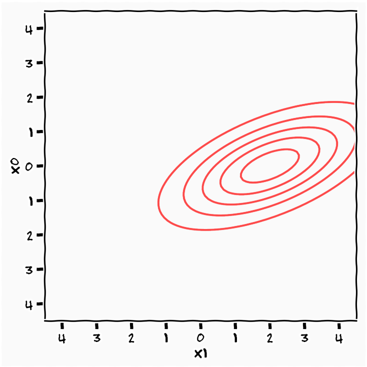
$$
\begin{aligned}
p(x_1, x_2) &= \mathcal{N}\left(\begin{array}{c} x_1 \\ x_2 \end{array} \middle| \begin{array}{c} \mu_1 \\ \mu_2 \end{array}, \begin{array}{cc} k_{11} & k_{12} \\ k_{21} & k_{22} \end{array} \right) \\
= p(x_2, x_1) &= \mathcal{N}\left(\begin{array}{c} x_2 \\ x_1 \end{array} \middle|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

Gaussian Distribution - Exchangeable



Gaussian Distribution - Exchangeable





Gaussian Distribution - Marginal

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

Gaussian Distribution - Marginal

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

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

Gaussian Distribution - Marginal

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

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

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

Gaussian Distribution - Marginal

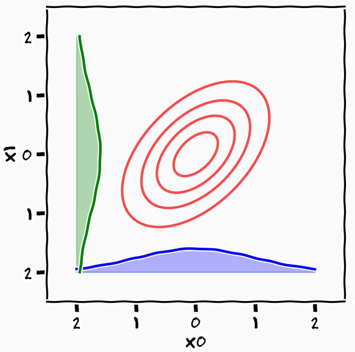
$$p(x_1, x_2) = \mathcal{N}\left(\begin{matrix} x_1 \\ x_2 \end{matrix} \,\middle|\, \begin{matrix} \mu_1 \\ \mu_2 \end{matrix}, \begin{matrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{matrix} \right)$$

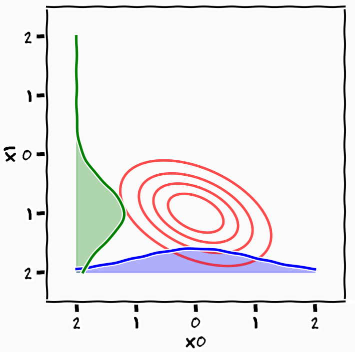
$$\Rightarrow p(x_1) = \int_{x_2} p(x_1, x_2) = \underline{\mathcal{N}\left(x_1 \mid \mu_1, k_{11}\right)}$$

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

$$\Rightarrow p(x_1) = \int_{x_2, \ldots, x_N} p(x_1, x_2, \ldots, x_N) = \underline{\mathcal{N}\left(x_1 \mid \mu_1, k_{1\ldots}\right.}$$

Gaussian Distribution - Marginal





Gaussian processes

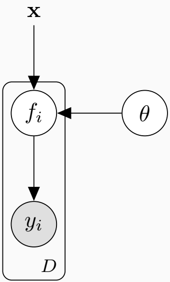
$$\mathcal{GP}(\cdot,\cdot) \qquad M \in \mathbb{R}^{\infty \times N} \qquad \mathcal{N}(\cdot,\cdot)$$

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

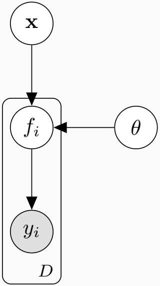
The Gaussian distribution is the projection of the infinite Gaussian process

Unsupervised Gaussian Processes

Unsupervised Learning

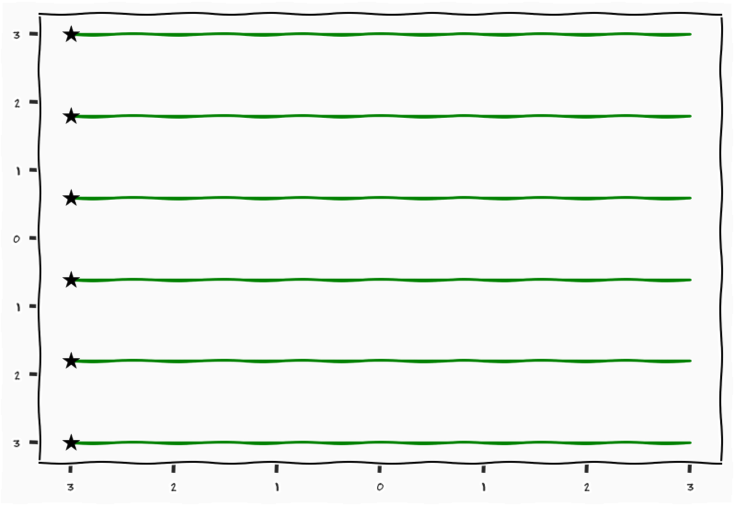


$p(y|x)$

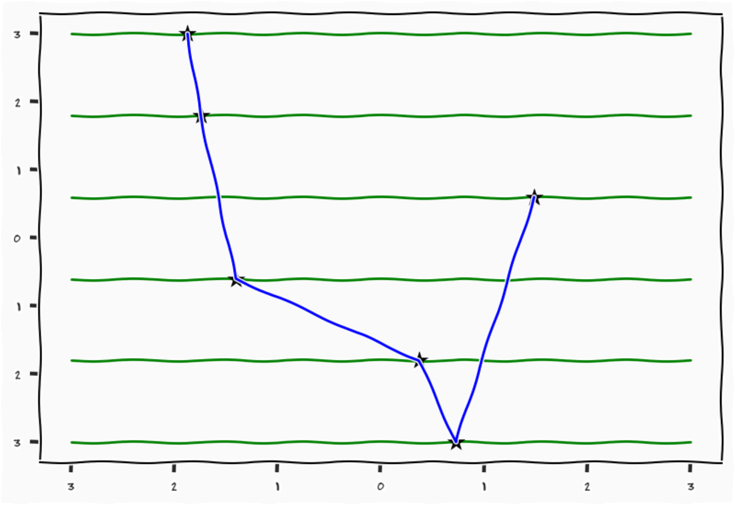


$p(y)$

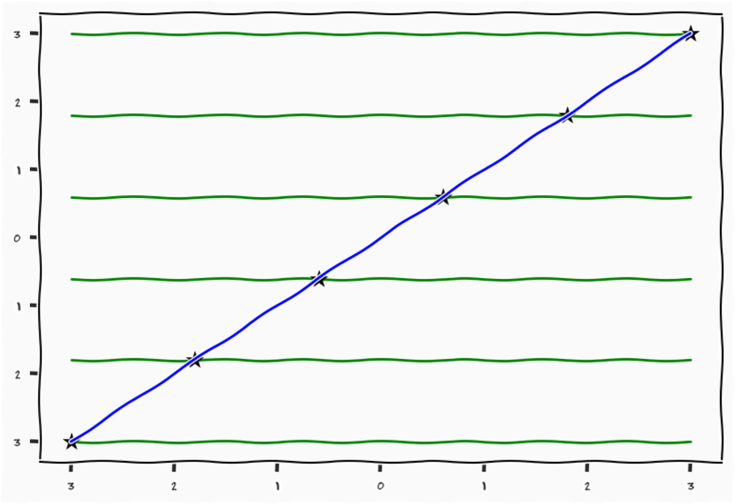
Unsupervised Learning



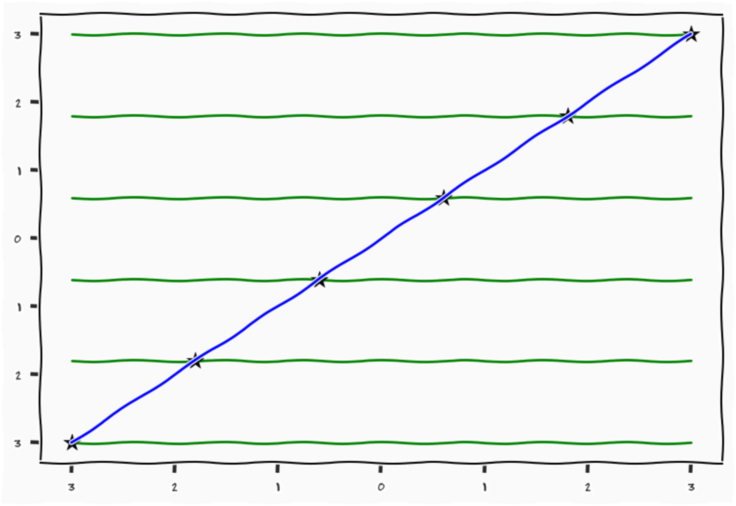
Unsupervised Learning



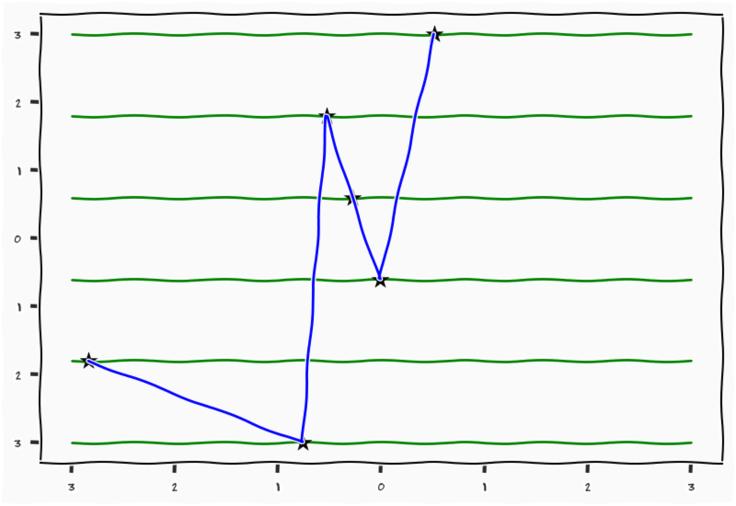
Unsupervised Learning



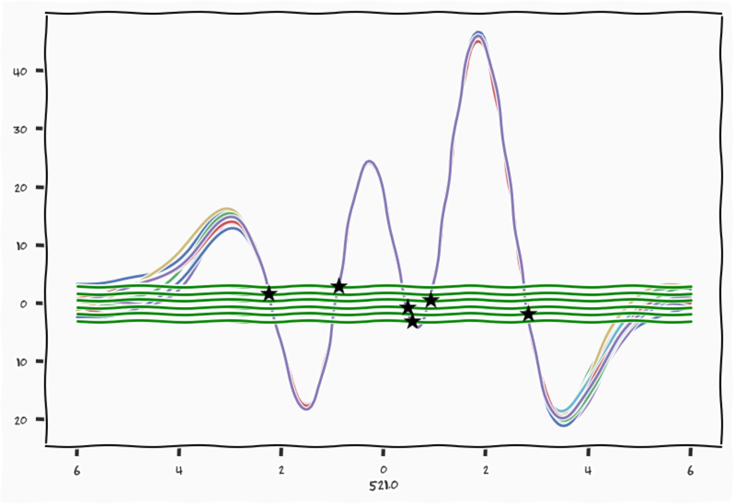
Unsupervised Learning



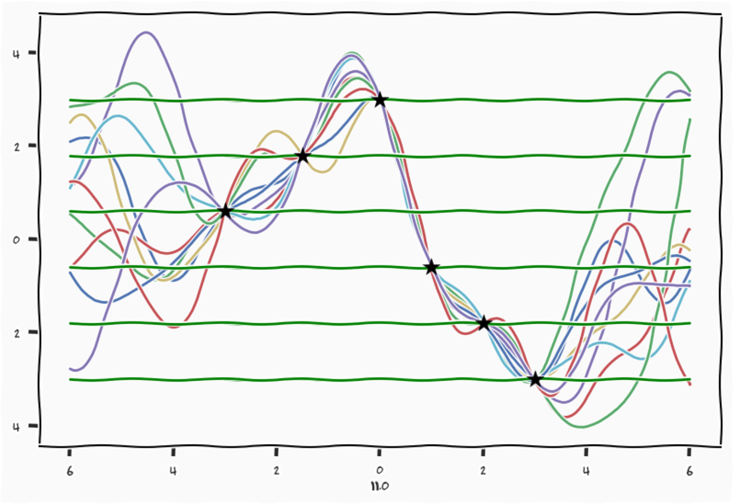
Unsupervised Learning

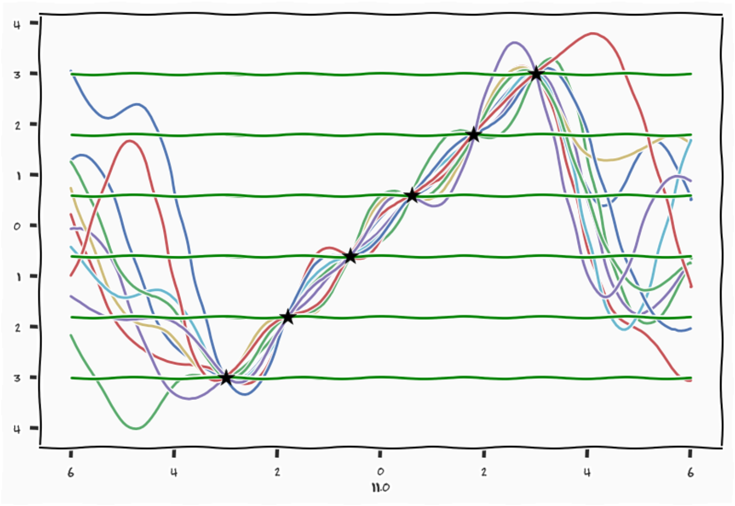


Unsupervised Learning

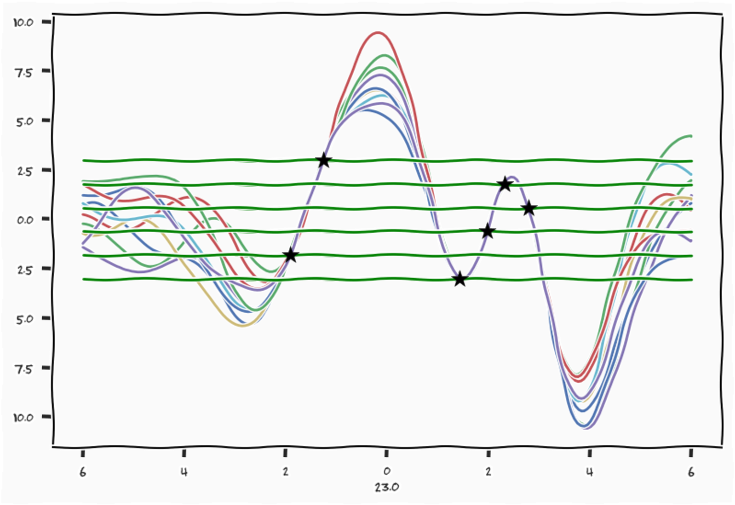


Unsupervised Learning

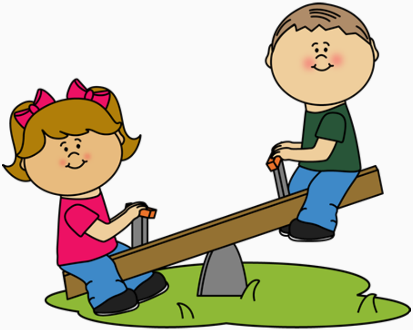



4
3
2
1
0
1
2
3
4
6
4
2
0
2
4
6
11.0

Unsupervised Learning



Priors



Priors

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

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

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

Laplace Integration

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

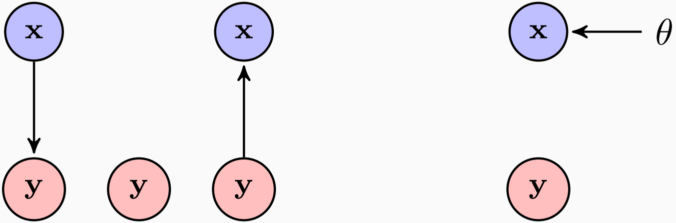
Approximate Inference

Machine Learning

$$p(y)$$

- Given some observed data y
- Find a probabilistic model such that the probability of the data is maximised
- Idea: find an approximate model q that we can integrate

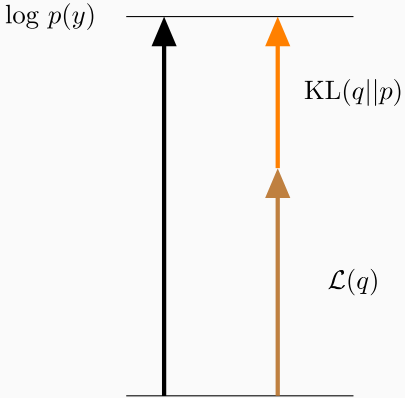
Lower Bound



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

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

Deterministic Approximation



Variational Bayes

$$p(y)$$

Variational Bayes

$$\log p(y)$$

Variational Bayes

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

Variational Bayes

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

Variational Bayes

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

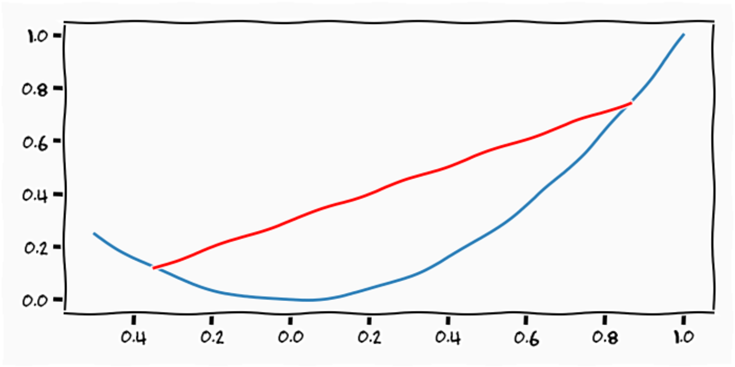
Variational Bayes

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

Variational Bayes

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

Jensen Inequality



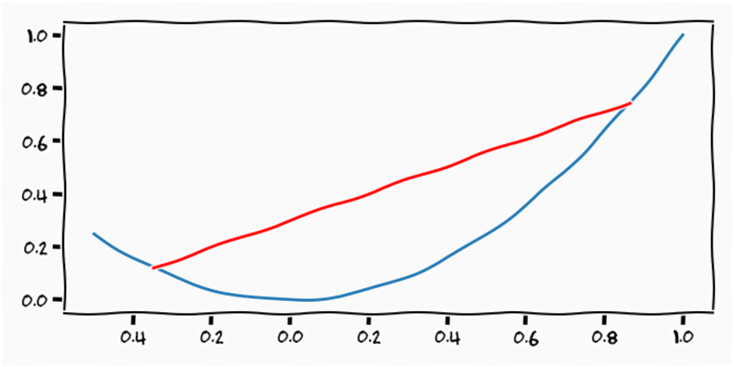
Convex Function

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

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

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

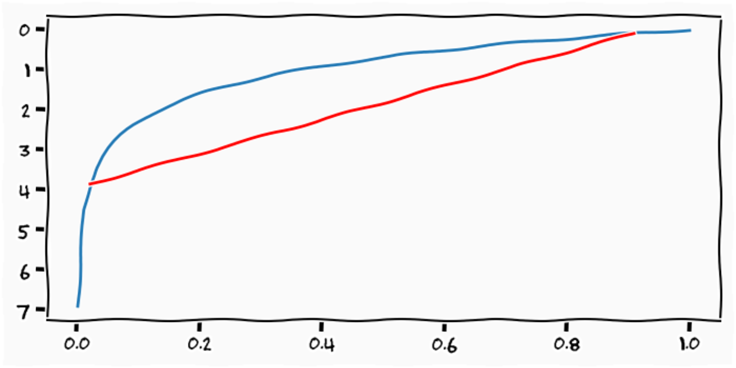
Jensen Inequality



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

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

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

The "posterior" term

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

The "posterior" term

$$
\begin{aligned}
\color{red}{KL(q(x)||p(x|y))} &= \int q(x)\,\log\,\frac{q(x)}{p(x|y)}\mathrm{d}x \\
&= -\int q(x)\,\log\,\frac{p(x|y)}{q(x)}\mathrm{d}x
\end{aligned}
$$

The "posterior" term

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

The "posterior" term

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

The "posterior" term

$$
\begin{aligned}
\color{red}{KL(q(x)||p(x|y))} &= \int q(x) \log \frac{q(x)}{p(x|y)} \mathrm{d}x \\
&= \{\text{Lets assume that } q(x) = p(x|y)\}
\end{aligned}
$$

The "posterior" term

$$
\begin{aligned}
{\color{red}KL(q(x)||p(x|y))} &= \int q(x)\,\log\frac{q(x)}{p(x|y)}\mathrm{d}x \\
&= \{\text{Lets assume that } q(x) = p(x|y)\} \\
&= \int p(x|y)\,\log\underbrace{\frac{p(x|y)}{p(x|y)}}_{=1}\mathrm{d}x
\end{aligned}
$$

The "posterior" term

$$
\begin{aligned}
\color{red}{KL(q(x)||p(x|y))} &= \int q(x) \log \frac{q(x)}{p(x|y)} \mathrm{d}x \\
&= \{\text{Lets assume that } q(x) = p(x|y)\} \\
&= \int p(x|y) \log \underbrace{\frac{p(x|y)}{p(x|y)}}_{=1} \mathrm{d}x \\
&= 0
\end{aligned}
$$

Kullback-Leibler Divergence

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

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

The "other terms"

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

The "other terms"

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

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

The "other terms"

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

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

$$= \{\text{Lets assume that } q(x) = p(x|y)\}$$

The "other terms"

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

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

$$= \{\text{Lets assume that } q(x) = p(x|y)\}$$

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

The "other terms"

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

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

$$=\{\text{Lets assume that } q(x)=p(x|y)\}$$

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

The "other terms"

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

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

$$= \{\text{Lets assume that } q(x) = p(x|y)\}$$

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

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

The "other terms"

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

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

$$= \{\text{Lets assume that } q(x) = p(x|y)\}$$

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

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

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

The "other terms"

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

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

$$= \{\text{Lets assume that } q(x) = p(x|y)\}$$

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

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

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

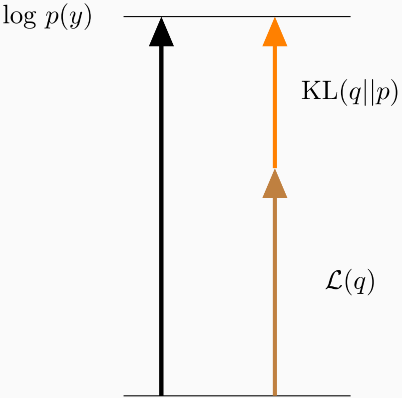
Variational Bayes

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

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

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

Deterministic Approximation

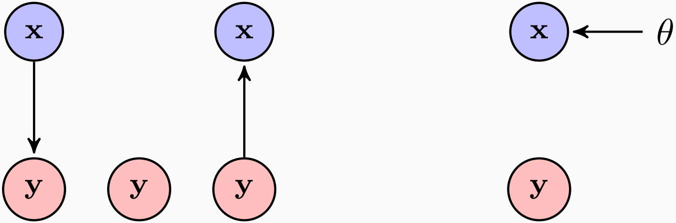


ELBO

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

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

Lower Bound



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

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

Why is this useful?

Why is this a sensible thing to do?

– Ryan Adams[2]

[2] Talking Machines Podcast

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[2]

[2] Talking Machines Podcast

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[2]

[2] Talking Machines Podcast

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[2]

[2] Talking Machines Podcast

How to choose Q?

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

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

Lower Bound[3]

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

[3] Damianou, 2015

Lower Bound[3]

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

[3] Damianou, 2015

Lower Bound[3]

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

[3] Damianou, 2015

Lower Bound[3]

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

[3] Damianou, 2015

Lower Bound

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

- Has not eliviate the problem at all, x still needs to go through f to reach the data
- Idea of sparse approximations[4]

[4]Candela et al., 2005

Lower Bound [5]

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

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

[5] Titsias et al., 2010

Lower Bound [5]

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

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

[5] Titsias et al., 2010

Lower Bound [5]

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

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

[5] Titsias et al., 2010

Lower Bound

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

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

Lower Bound

- Variational distributions are approximations to intractable posteriors,

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

Lower Bound

- Variational distributions are approximations to intractable posteriors,

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

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

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

Lower Bound

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

Lower Bound

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

- Assume that u is sufficient statistics of f

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

Lower Bound

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

Lower Bound

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

Lower Bound

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

Lower Bound

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

Lower Bound

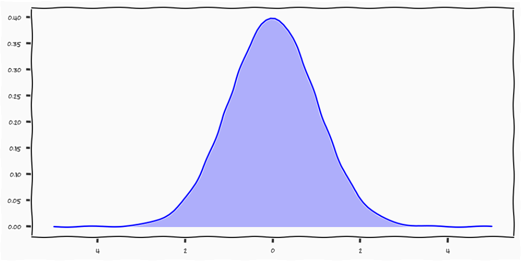
$$
\begin{aligned}
\tilde{\mathcal{L}} &= \int_{x,f,u} q(f)q(u)q(x)\log\frac{p(y \mid f)p(f \mid u,x,z)p(u \mid z)}{q(f)q(u)} \\
&= \int_{x,f,u} p(f \mid u,x,z)q(u)q(x)\log\frac{p(y \mid f){\color{red}p(f \mid u,x,z)}p(u \mid z)}{{\color{red}p(f \mid u,x,z)}q(u)} \\
&= \int_{x,f,u} p(f \mid u,x,z)q(u)q(x)\log\frac{p(y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

Lower Bound

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

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

Latent Space Priors



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

Automatic Relevance Determination

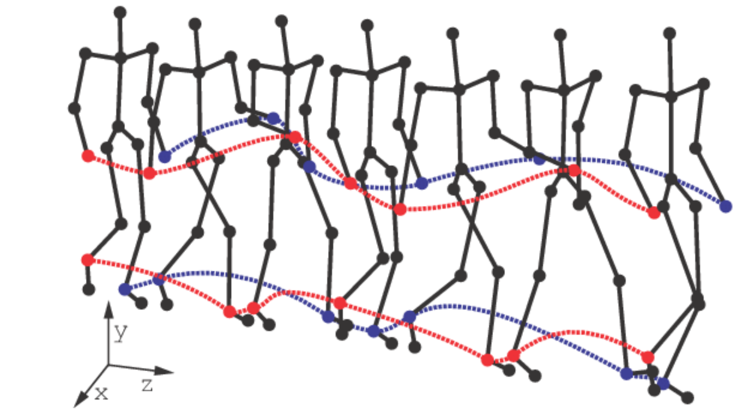
$$k(\mathbf{x}_i, \mathbf{x}_j) = \sigma e^{-\sum_d^D \alpha_d \cdot (x_{i,d} - x_{j,d})^2}$$

GPy

Code

[]python RBF(...,ARD=True) Matern32(...,ARD=True)

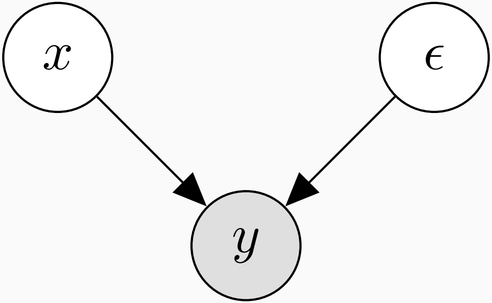
Dynamic Prior



$$p(x \mid t) = \mathcal{N}(\mu_t, K_t)$$

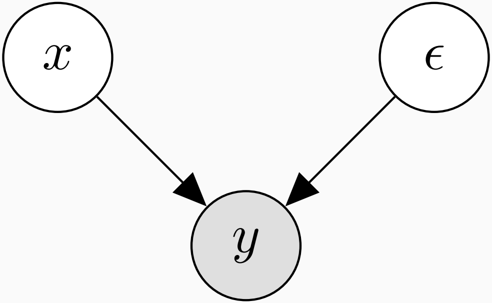
Structured Latent Spaces

Explaining Away



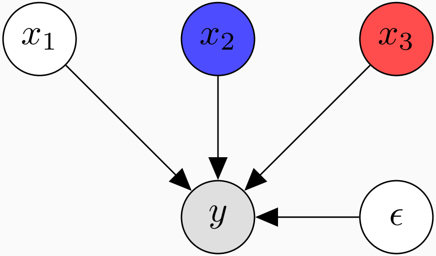
$$y = f(x) + \epsilon$$

Explaining Away



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

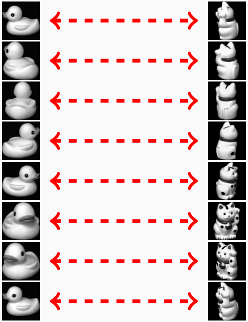
Factor Analysis



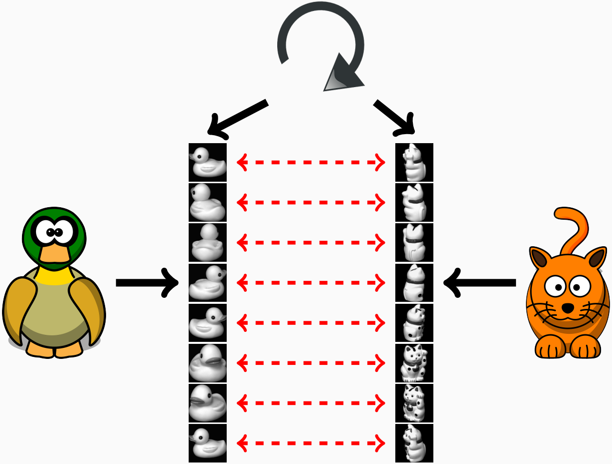
$$y = f(x_1, {\color{blue}x_2}, {\color{red}x_3}) + \epsilon$$

Alignments

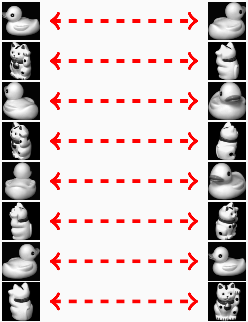
Alignments



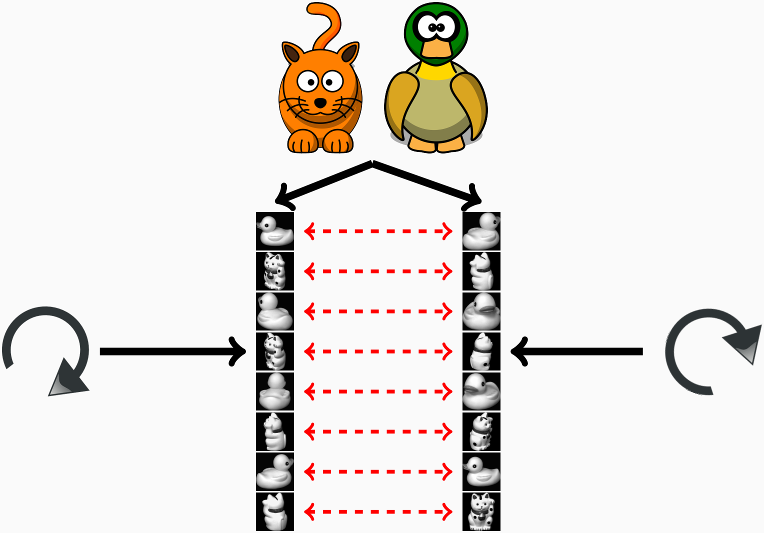
Alignments



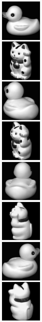
Alignments



Alignments



Alignments



$\{duck\}$

$\{cat\}$

$\{duck\}$

$\{cat\}$

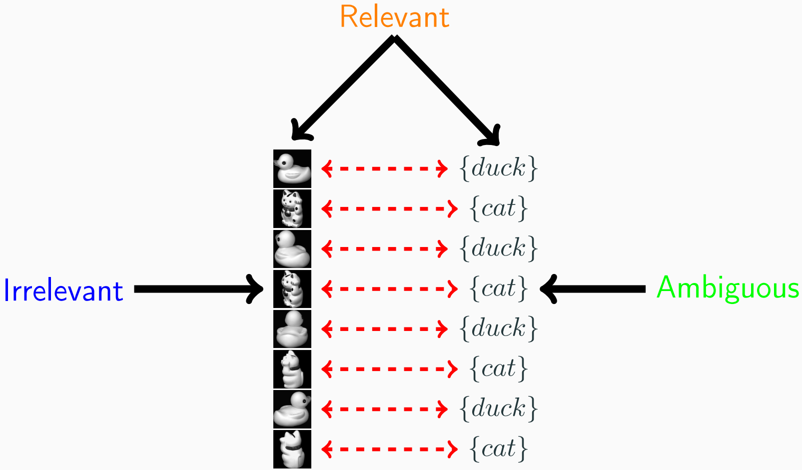
$\{duck\}$

$\{cat\}$

$\{duck\}$

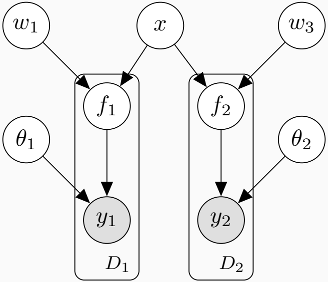
$\{cat\}$

Alignments



Relevant
Irrelevant
Ambiguous
$\{duck\}$
$\{cat\}$
$\{duck\}$
$\{cat\}$
$\{duck\}$
$\{cat\}$
$\{duck\}$
$\{cat\}$

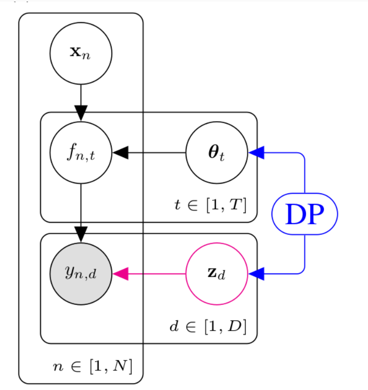
IBFA with GP-LVM[6]

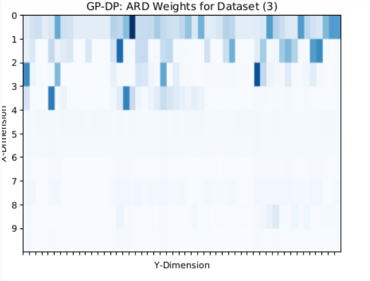


$$y_1 = f(w_1^{\mathrm{T}} x) \quad y_2 = f(w_2^{\mathrm{T}} x)$$

[6] Damianou et al., 2016

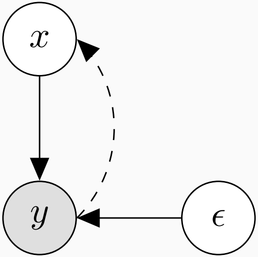
GP-DP[7]





[7] Lawrence et al., 2019

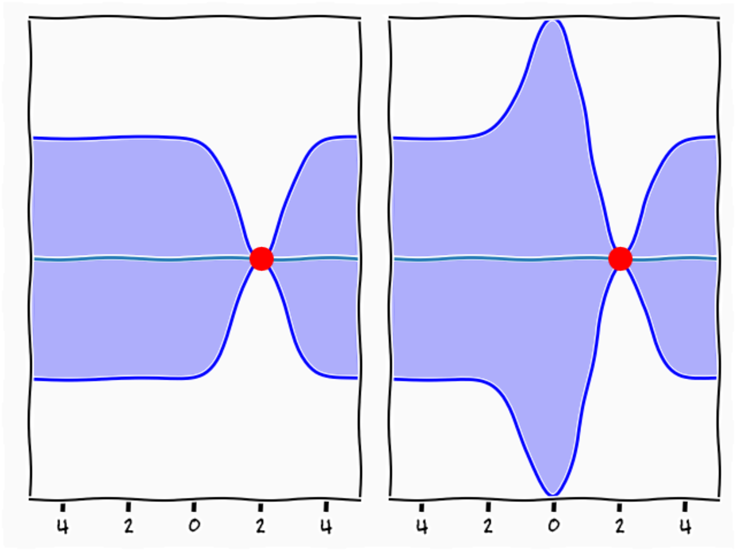
Constrained Latent Space[8]



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

[8] Lawrence et al., 2006

Geometry

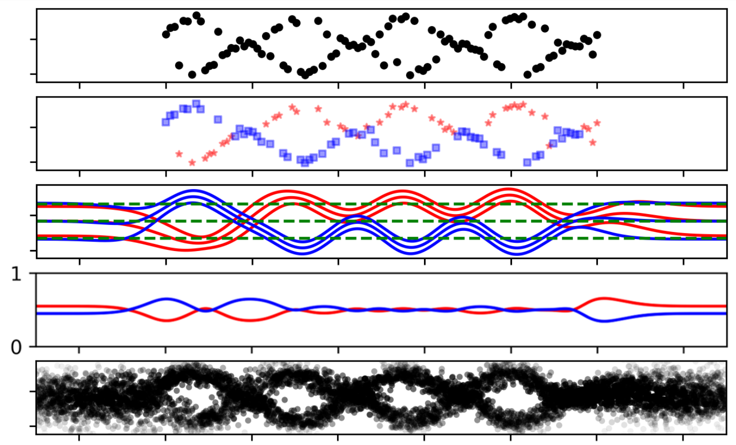


Latent GP-Regression[9]

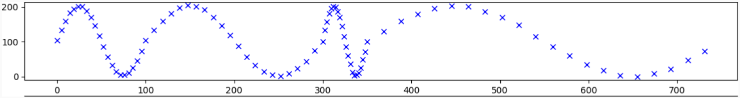
$$p(\mathbf{Y}|\mathbf{X}) = \int p(\mathbf{Y}|\mathbf{F})\, p(\mathbf{F}|\mathbf{X}, \mathbf{X}^{(C)})\, p(\mathbf{X}^{(C)})\, \mathrm{d}\mathbf{F}\, \mathrm{d}\mathbf{X}^{(C)}.$$

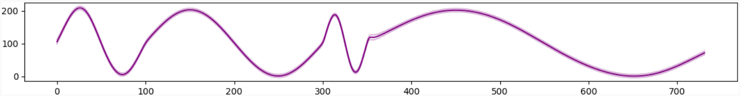
[9] Bodin et al., 2017, Yousefi et al., 2016

Discrete



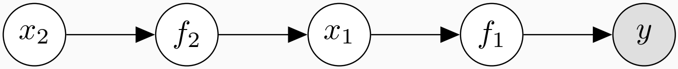
Continous


200
100
0
0
100
200
300
400
500
600
700


200
100
0
0
100
200
300
400
500
600
700

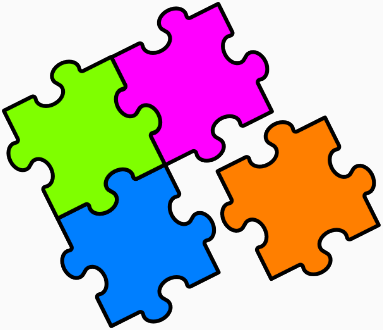
Composite Gaussian Processes

Composite Gaussian Processes [10]



[10] Damianou et al., 2013

Composite Functions



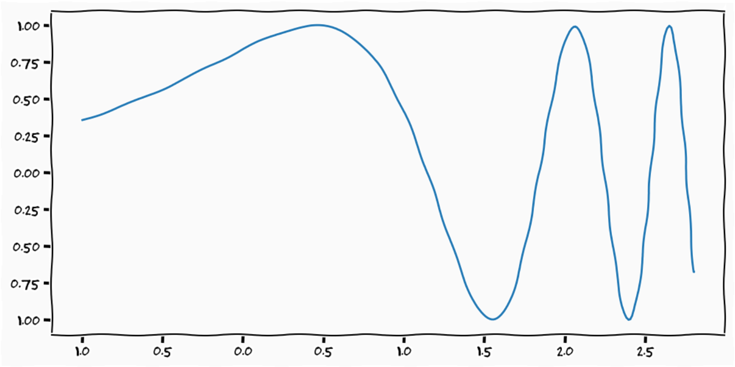
$$y = f_k(f_{k-1}(\dots f_0(x))) = f_k \circ f_{k-1} \circ \dots \circ f_1(x)$$

When do I want Composite Functions

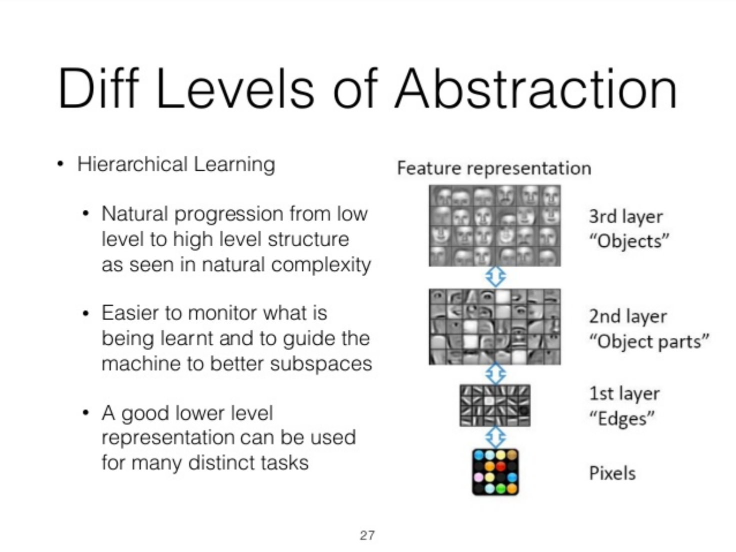
$$y = f_k \circ f_{k-1} \circ \cdots \circ f_1(x)$$

1. My generative process is composite
    - my prior knowledge is composite
2. I want to "re-parametrise" my kernel in a learning setting
    - i have knowledge of the re-parametrisation

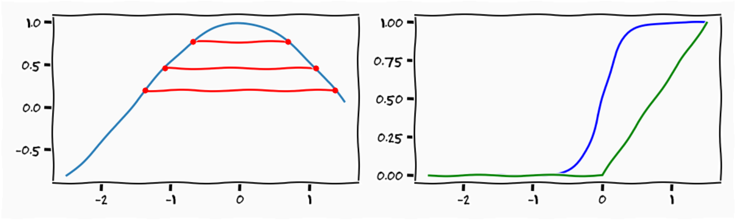
Because we lack "models"?



Composite Functions



Composite functions

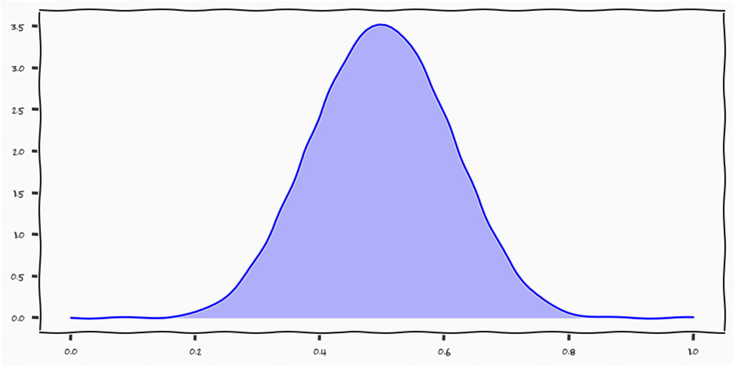


$$y = f_k(f_{k-1}(\dots f_0(x))) = f_k \circ f_{k-1} \circ \dots \circ f_1(x)$$

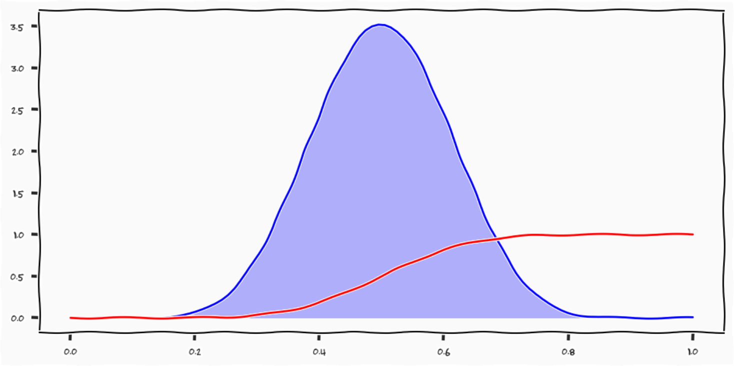
$$\mathsf{Kern}(f_1) \subseteq \mathsf{Kern}(f_{k-1} \circ \dots \circ f_2 \circ f_1) \subseteq \mathsf{Kern}(f_k \circ f_{k-1} \circ \dots \circ f_2 \circ f_1)$$

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

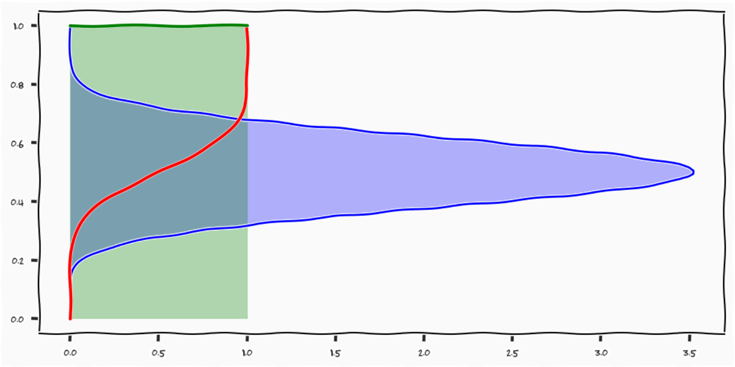
Sampling



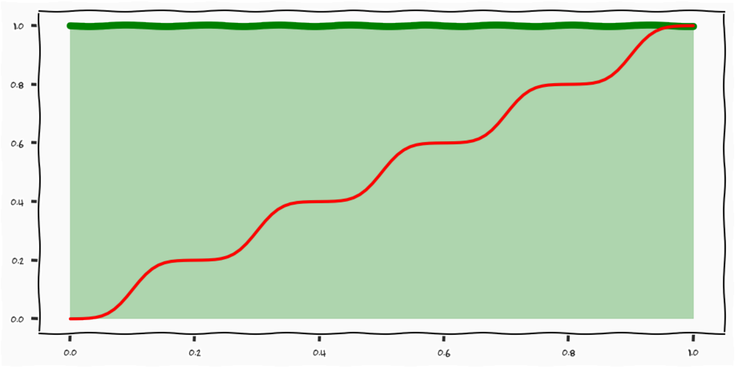
Sampling



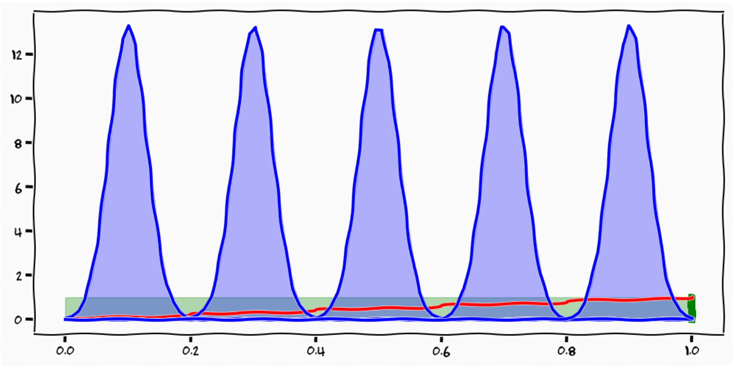
Sampling



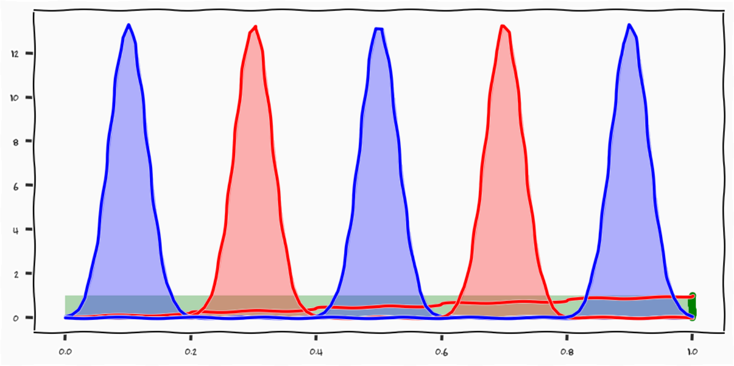
Change of Variables



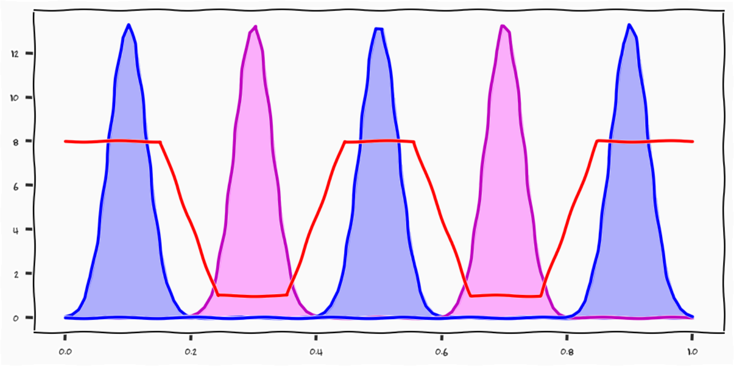
Change of Variables



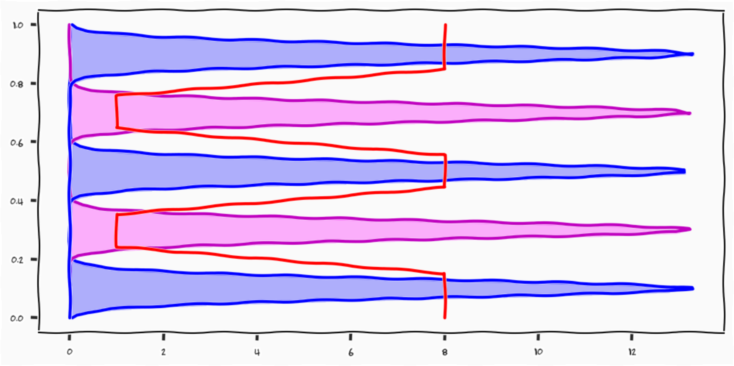
Change of Variables



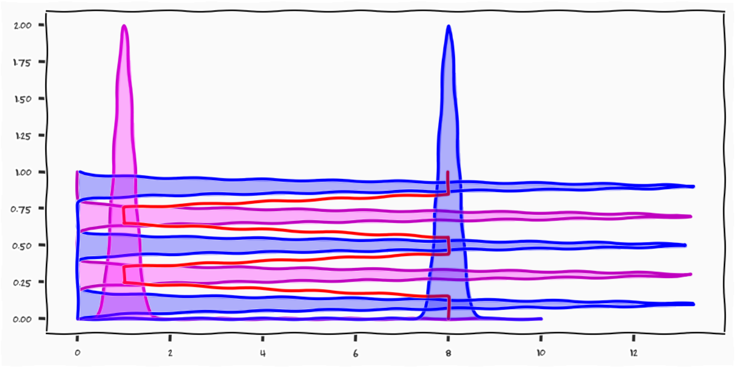
Change of Variables



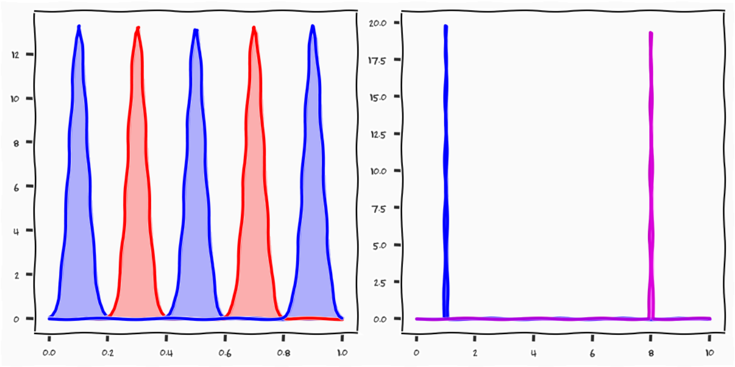
Change of Variables



Change of Variables

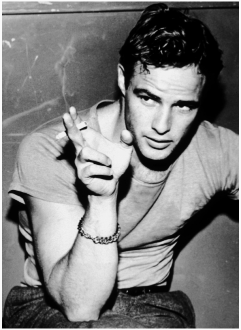


Change of Variables



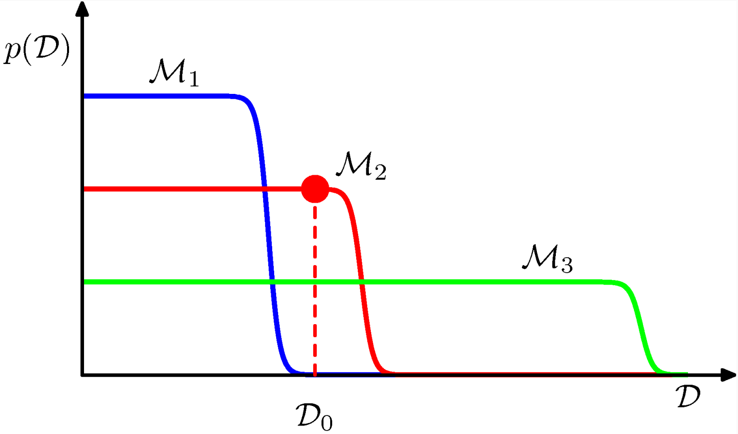
"I'm Bayesian therefore I am superior"

Because we want to hang out with the cool kids



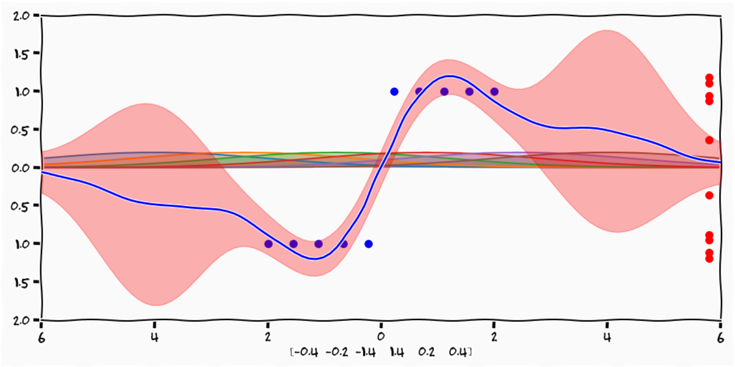
Deep Learning is a bit like smoking, you know that its wrong but you do it anyway because you want to look cool.
– Fantomens Djungelordspråk

MacKay plot

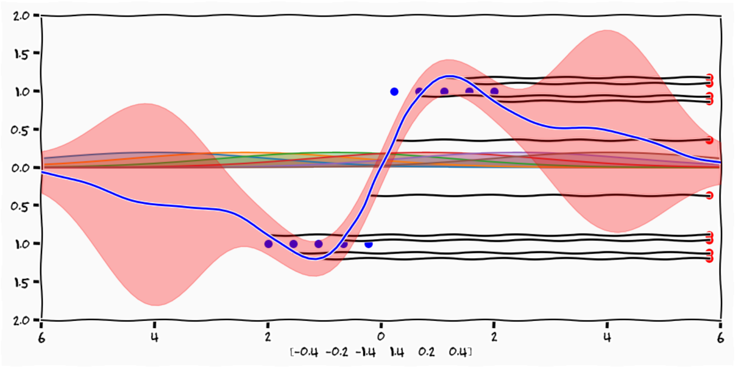


$p(\mathcal{D})$
$\mathcal{M}_1$
$\mathcal{M}_2$
$\mathcal{M}_3$
$\mathcal{D}$
$\mathcal{D}_0$

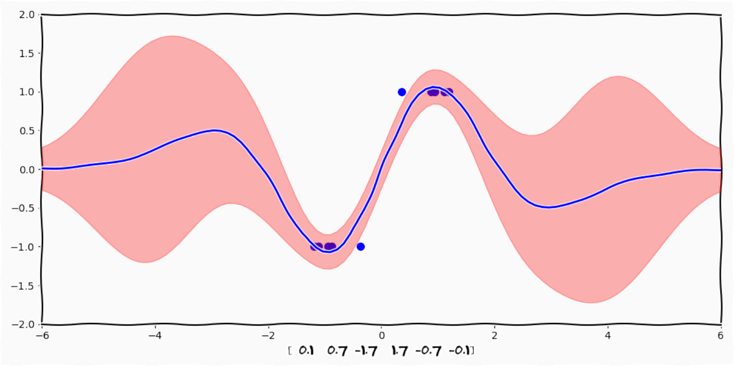
Composite Functions



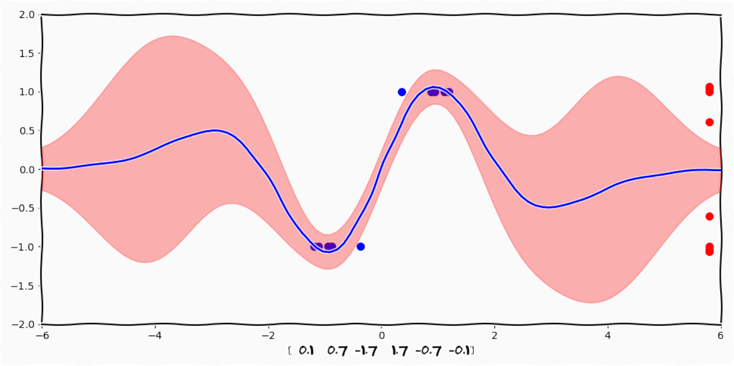
Composite Functions



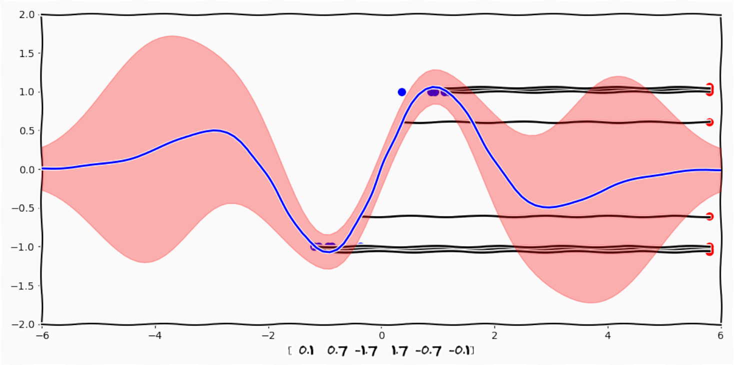
Composite Functions



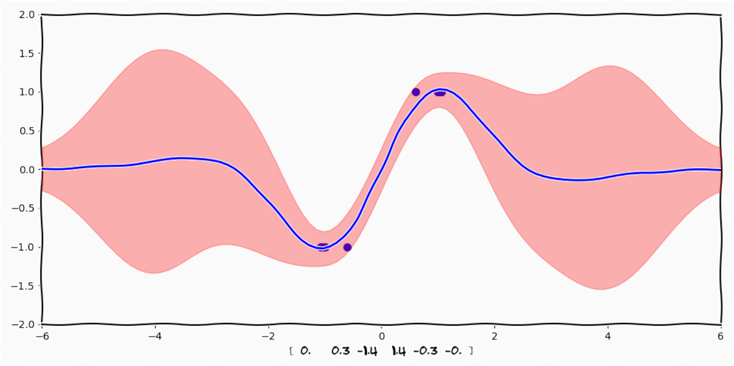
Composite Functions



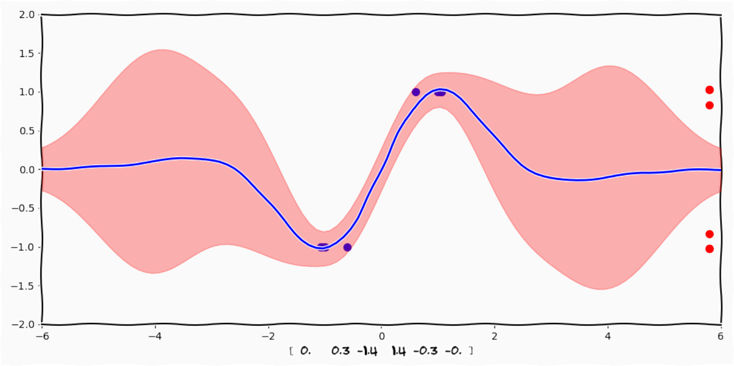
Composite Functions



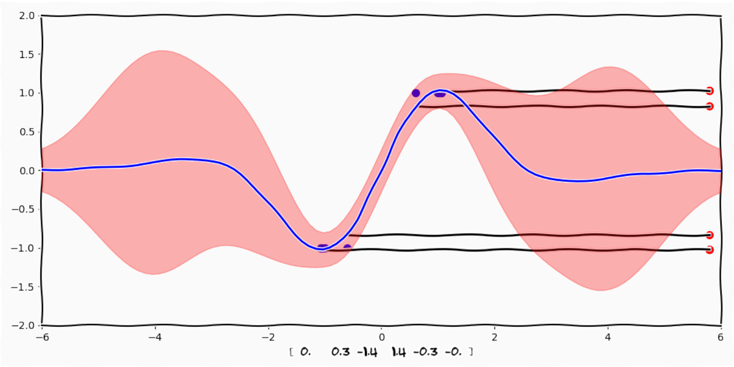
Composite Functions



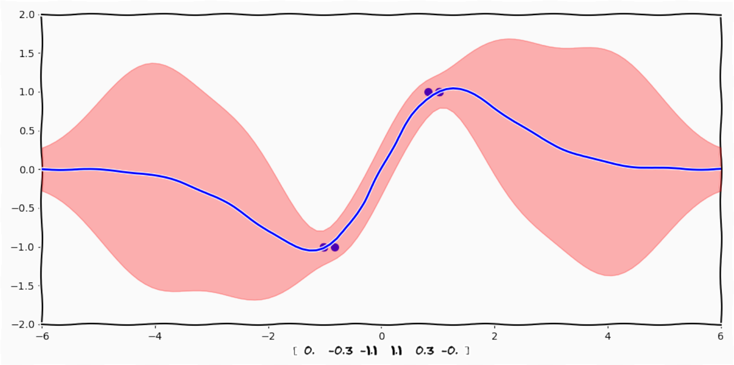
Composite Functions



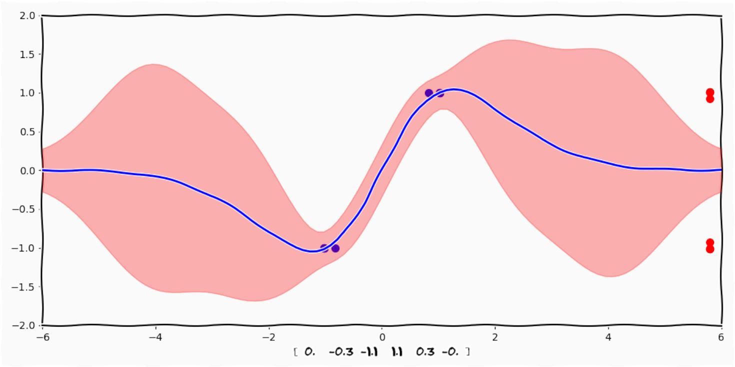
Composite Functions



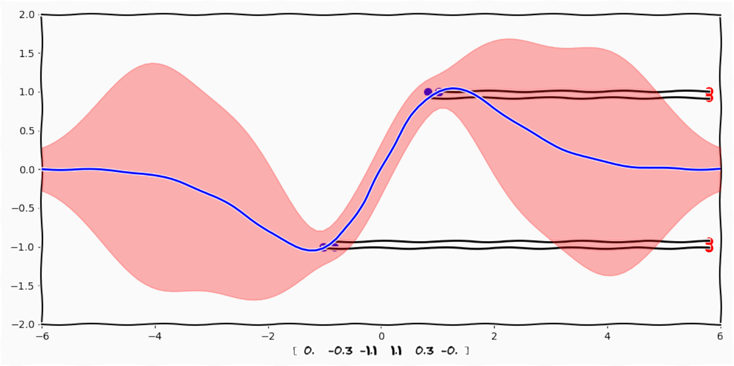
Composite Functions



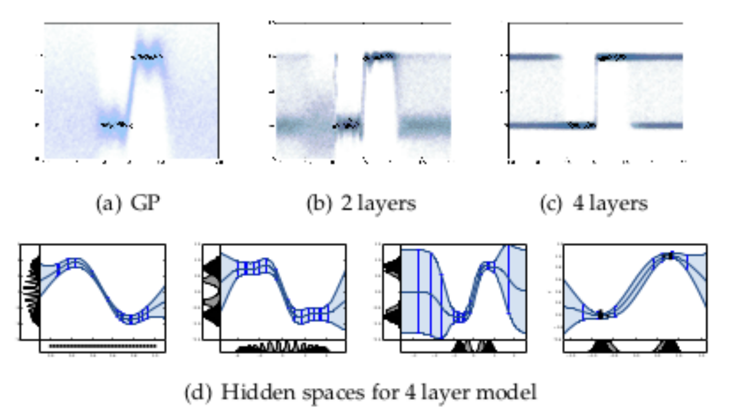
Composite Functions



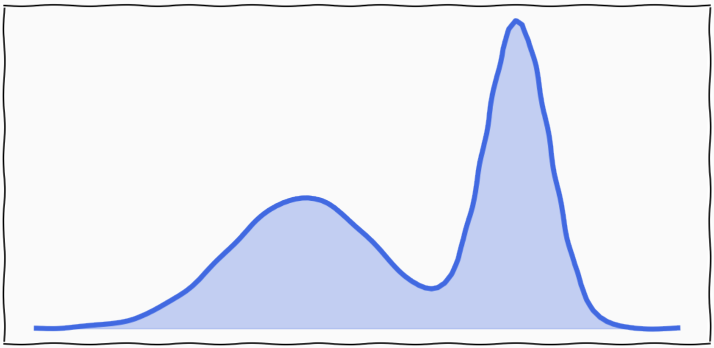
Composite Functions



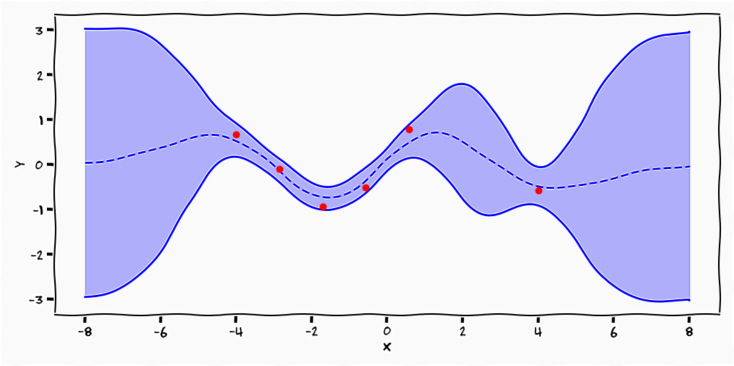
The Final Composition



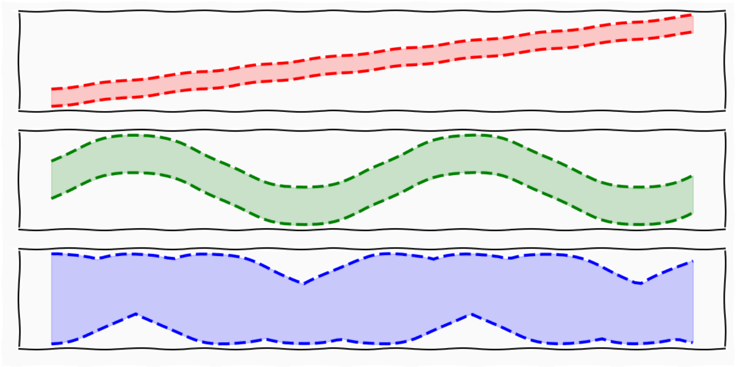
Remember why we did this in the first place



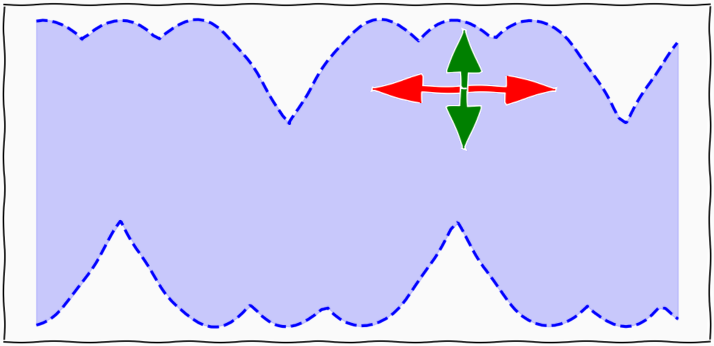
These damn plots



It gets worse



It gets even worse



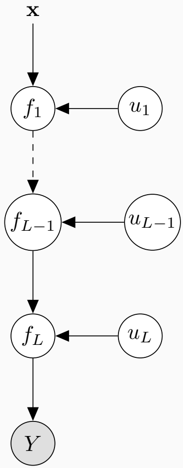
- Sufficient statistics

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

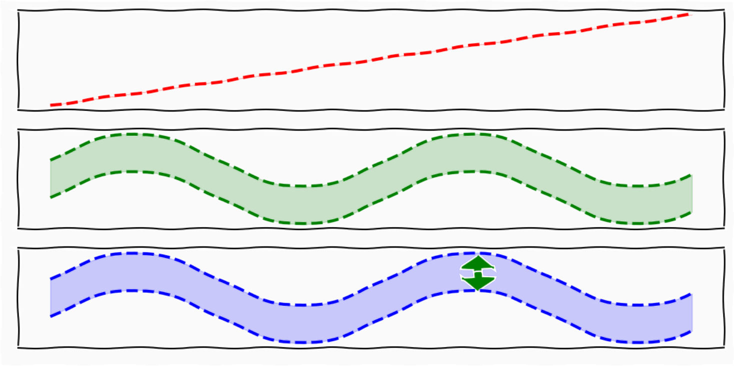
$$= p(\mathbf{F}|\mathbf{U},\mathbf{X},\mathbf{Z})q(\mathbf{U})q(\mathbf{X})$$

- Mean-Field

$$q(\mathbf{U}) = \prod_i^L q(\mathbf{U}_i)$$



The effect



What have we lost

- Our priors are not reflected correctly
  - $\rightarrow$ we cannot interpret the results
- No intermediate uncertainties
  - $\rightarrow$ we cannot do sequential decision making

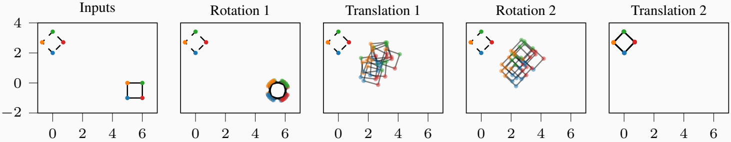
What have we lost

- Our priors are not reflected correctly
  - $\rightarrow$ we cannot interpret the results
- No intermediate uncertainties
  - $\rightarrow$ we cannot do sequential decision making
- We are performing a massive computational overhead for very little use

What have we lost

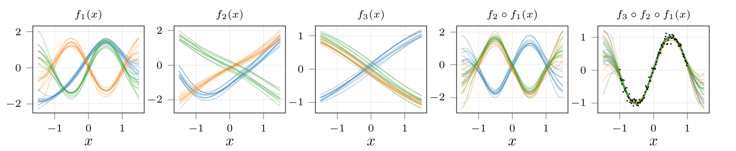
- Our priors are not reflected correctly
  - $\rightarrow$ we cannot interpret the results
- No intermediate uncertainties
  - $\rightarrow$ we cannot do sequential decision making
- We are performing a massive computational overhead for very little use
- *"... throwing out the baby with the bathwater... "*

What we really want[11]



[11] Ustyuzhaninov et al., 2019

What we really want[12]



[12]Ustyuzhaninov et al., 2019

Summary

Summary

- Unsupervised learning[13] is very hard.

[13] I would argue that there is no such thing

Summary

- Unsupervised learning[13] is very hard.
    - *Its actually not, its really really easy.*

[13] I would argue that there is no such thing

Summary

- Unsupervised learning[13] is very hard.
    - *Its actually not, its really really easy.*
- Relevant assumptions needed to learn anything useful

[13] I would argue that there is no such thing

Summary

- Unsupervised learning[13] 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

[13] I would argue that there is no such thing

Summary

- Unsupervised learning[13] 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

[13] I would argue that there is no such thing

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
- We need to think about correlated uncertainty, not marginals

eof

Reference

References

Bodin, Erik, Neill D. F. Campbell, and Carl Henrik Ek (2017). *Latent Gaussian Process Regression*.

Candela, Joaquin Quiñonero and Carl Edward Rasmussen (2005). "A Unifying View of Sparse Approximate Gaussian Process Regression". In: *Journal of Machine Learning Research* 6, pp. 1939–1959.

Damianou, Andreas, Neil D Lawrence, and Carl Henrik Ek (2016). "Multi-view Learning as a Nonparametric Nonlinear Inter-Battery Factor Analysis". In: *arXiv preprint arXiv:1604.04939*.

Damianou, Andreas C (Feb. 2015). "Deep Gaussian Processes and Variational Propagation of Uncertainty". PhD thesis. University of Sheffield.

Damianou, Andreas C and Neil D Lawrence (2013). "Deep Gaussian Processes". In: *International Conference on Airtificial Inteligence and Statistical Learning*, pp. 207–215.

Hensman, James, N Fusi, and Neil D Lawrence (2013). "Gaussian Processes for Big Data". In: *Uncertainty in Artificial Intelligence*.

Lawrence, Andrew R., Carl Henrik Ek, and Neill W. Campbell (2019). "DP-GP-LVM: A Bayesian Non-Parametric Model for Learning Multivariate Dependency Structures". In: *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pp. 3682–3691.

Lawrence, Neil D. and Joaquin Quiñonero Candela (2006). "Local Distance Preservation in the GP-LVM Through Back Constraints". In: *Proceedings of the 23rd International Conference on Machine Learning*. ICML '06. Pittsburgh, Pennsylvania, USA: ACM, pp. 513–520.

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

Ustyuzhaninov, Ivan et al. (2019). "Compositional Uncertainty in Deep Gaussian Processes". In: *CoRR*.

Yousefi, Fariba, Zhenwen Dai, Carl Henrik Ek, and Neil Lawrence (2016). "Unsupervised Learning With Imbalanced Data Via Structure Consolidation Latent Variable Model". In: *CoRR*.