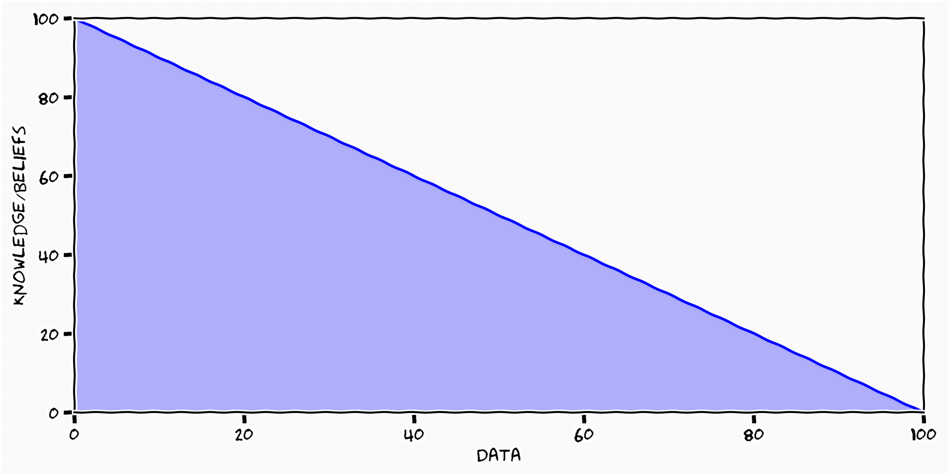
Deep Gaussian Processes as Point Estimates of Deep Neural Networks

Carl Henrik Ek - che29@cam.ac.uk

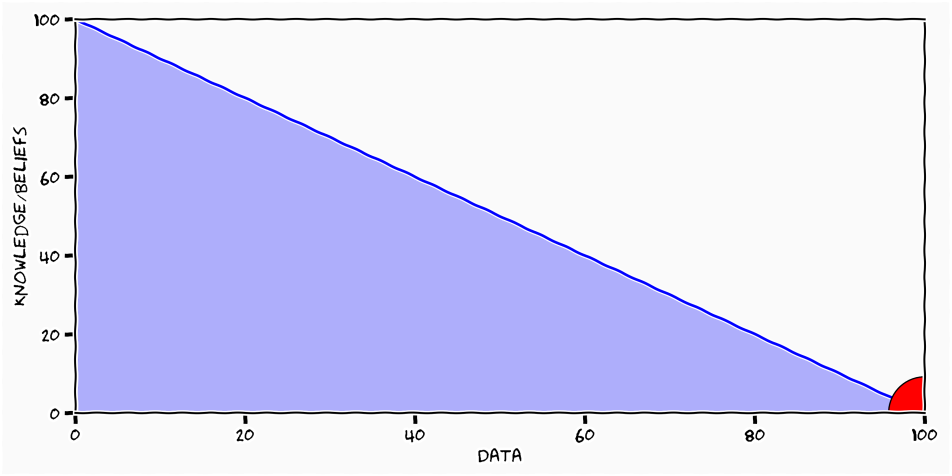
23rd of March, 2022

http://carlhenrik.com

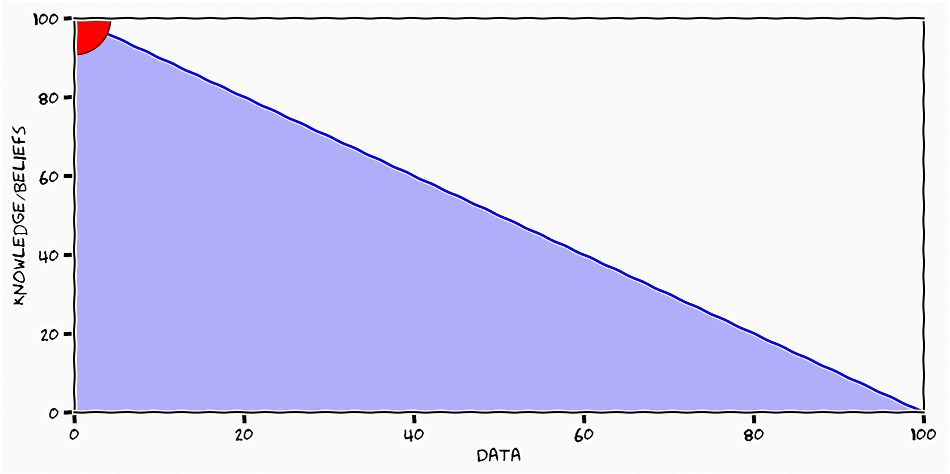
Data and Knowledge



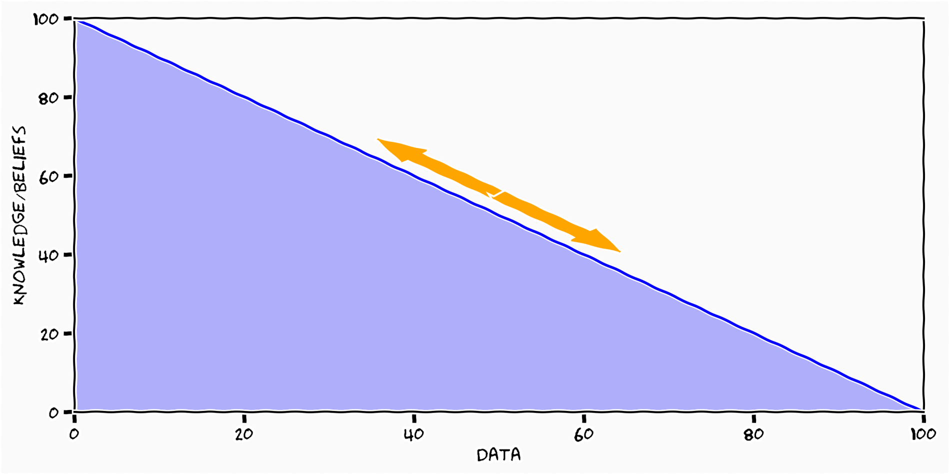
Data and Knowledge



Data and Knowledge



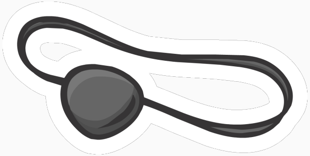
Machine Learning



Statistical Learning

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

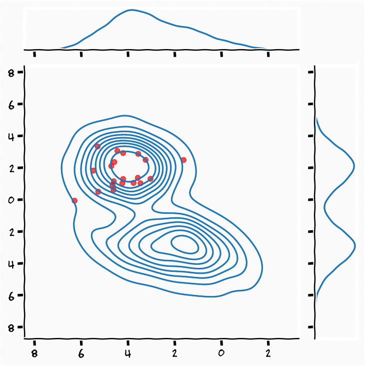
Knowledge: Algorithms



Statistical Learning

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

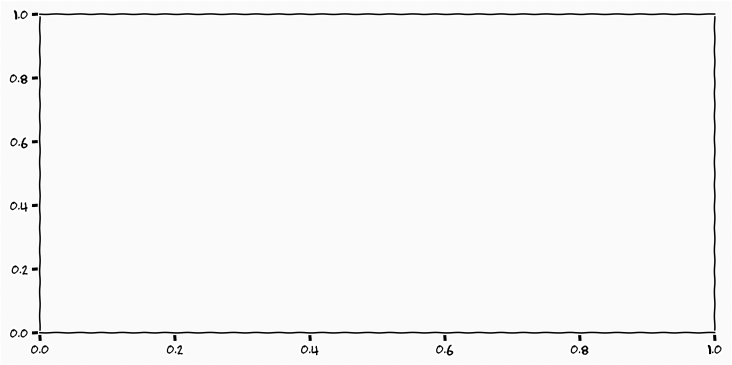
Knowledge: Biased Sample



Statistical Learning

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

Knowledge: Hypothesis space



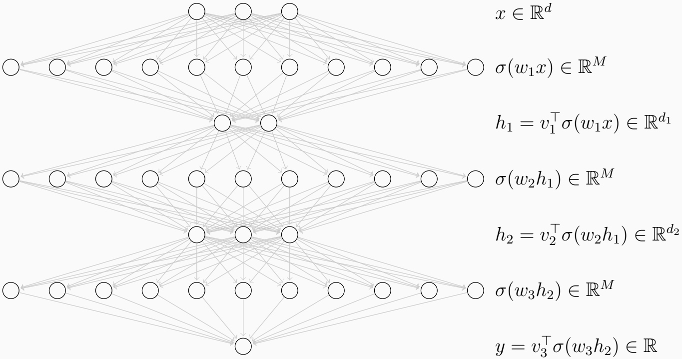
Statistical Learning

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

Composite Functions

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

Neural Networks[1]



$$\mathcal{F}_{\mathrm{DNN}}(\mathbf{x}) = \mathbf{V}_L^{\mathrm{T}} \left(\mathbf{W}_L \cdots \mathbf{V}_2^{\mathrm{T}} \sigma \left(\mathbf{W}_2 \mathbf{V}_1^{\mathrm{T}} \sigma \left(\mathbf{W}_1 \mathbf{x}\right)\right)\right)$$

[1] Recursive Generalised Linear Models

Neural Networks

- These methods shouldn't work

Neural Networks

- These methods shouldn't work
  - they do as predictive methods

- These methods shouldn't work
  - they do as predictive methods
- Why?

Neural Networks

- These methods shouldn't work
  - they do as predictive methods
- Why?
  - Parameter space suitable for local optimisation

The Promise of Marginalisation

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

- build models using all the benefits that compositional parametrisations apparently gives

The Promise of Marginalisation

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

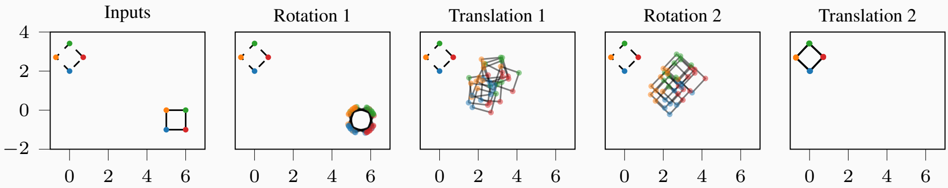
- build models using all the benefits that compositional parametrisations apparently gives
- interpretable biases

The Promise of Marginalisation

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

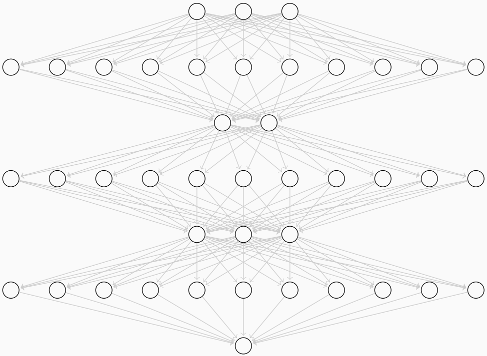
- build models using all the benefits that compositional parametrisations apparently gives
- interpretable biases
- compositional uncertainty structures

Compositional Uncertainty [2]



[2]Ustyuzhaninov et al., 2020

"Bayesian Neural Networks"[3]



$x \in \mathbb{R}^d$

$\sigma(w_1 x) \in \mathbb{R}^M$

$h_1 = v_1^\top \sigma(w_1 x) \in \mathbb{R}^{d_1}$

$\sigma(w_2 h_1) \in \mathbb{R}^M$

$h_2 = v_2^\top \sigma(w_2 h_1) \in \mathbb{R}^{d_2}$

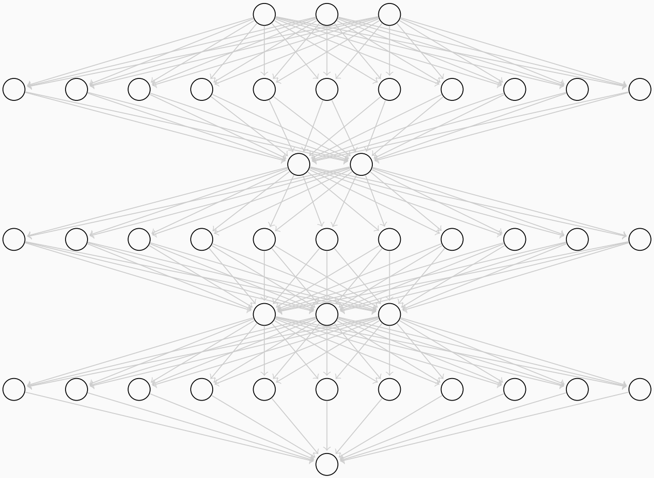
$\sigma(w_3 h_2) \in \mathbb{R}^M$

$y = v_3^\top \sigma(w_3 h_2) \in \mathbb{R}$

$$\mathbf{w} \sim \mathcal{N}(\cdot, \cdot)$$

[3] the worst of both worlds

Composite Gaussian Processes [4]



$x \in \mathbb{R}^d$

$k_{u_1}(x) \in \mathbb{R}^M$

$h_1 = b_1^\top k_{u_1}(x) \in \mathbb{R}^{d_1}$

$k_{u_2}(h_1) \in \mathbb{R}^M$

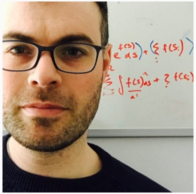
$h_2 = b_2^\top k_{u_2}(h_1) \in \mathbb{R}^{d_2}$

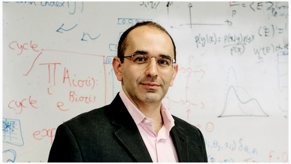
$k_{u_3}(h_2) \in \mathbb{R}^M$

$y = b_3^\top k_{u_3}(h_2) \in \mathbb{R}$

[4] Damianou et al., 2013

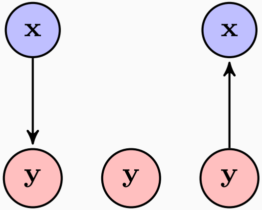
23/03/2022

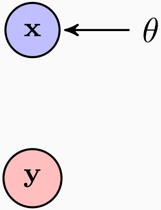




Sparse Gaussian Processes

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





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

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

Variational Bayes

$$p(y)$$

$$\log p(y)$$

Variational Bayes

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

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

Variational Bayes

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

Variational Bayes

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

Variational Bayes

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

Variational Bayes

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

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)
- $KL(q(x)||p(x|y)) = 0 \Leftrightarrow q(x) = p(x|y)$
- $KL(q(x)||p(x|y)) \geq 0$

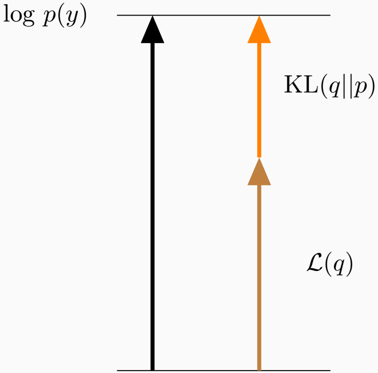
Variational Bayes

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

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

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

Deterministic Approximation



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[5]

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

[5] Damianou, 2015

Lower Bound[5]

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

[5] Damianou, 2015

Lower Bound[5]

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

[5] Damianou, 2015

Lower Bound[5]

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

[5] 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[6]

[6] Candela et al., 2005

Lower Bound [7]

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

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

[7] Titsias et al., 2010

Lower Bound [7]

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

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

[7] Titsias et al., 2010

Lower Bound [7]

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

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

[7] 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,

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

- 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

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

Lower Bound

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

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!

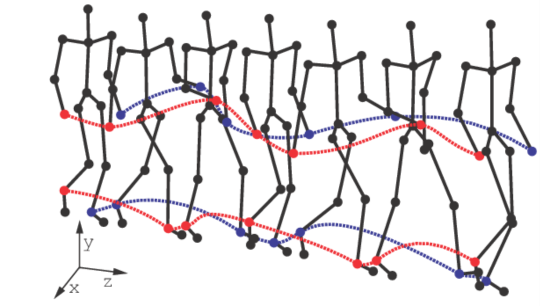
Automatic Relevance Determination[8]

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

- allows for learning the dimensionality of the latent space
- allows for latent space factorisations of latent space

[8]Damianou et al., 2021; Lawrence et al., 2019

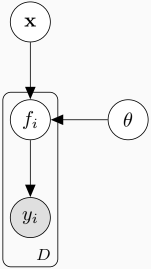
Dynamic Prior[9]

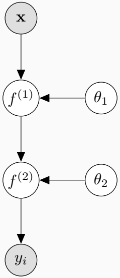


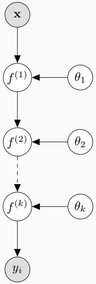
$$p(x \mid t) = \mathcal{N}(\mu_t, K_t)$$

[9] Damianou et al., 2011

Composite Gaussian Processes[10]



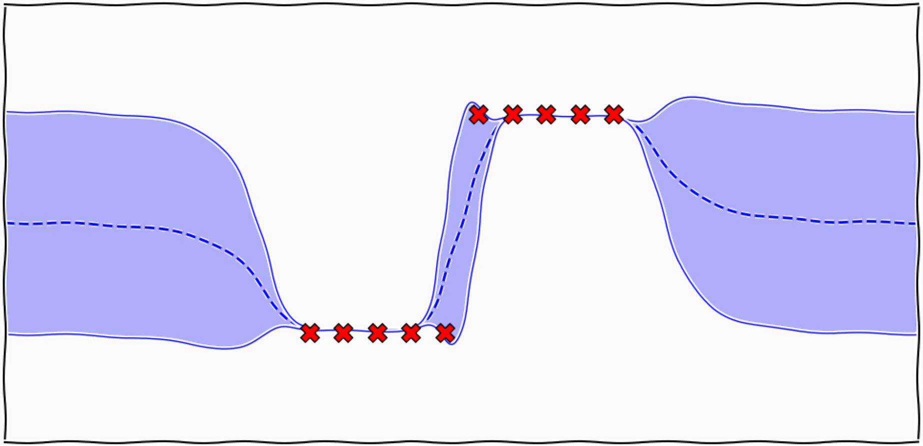




[10] Damianou et al., 2013; Titsias et al., 2010

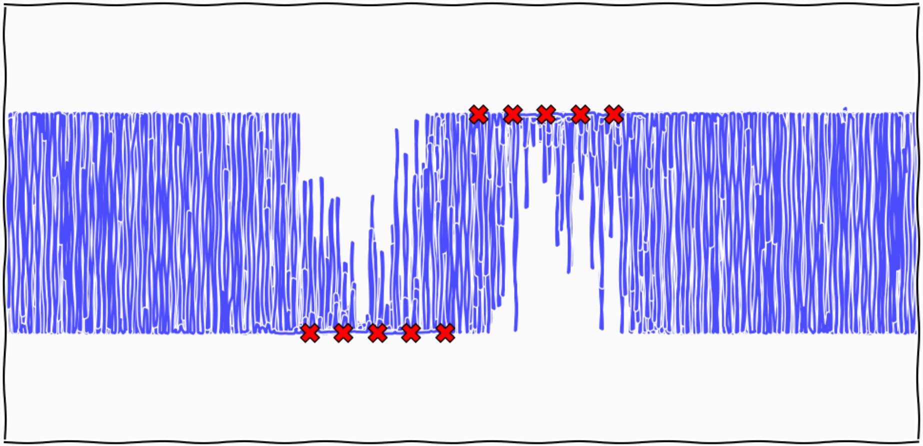
Composite Gaussian Processes

Composite Gaussian Processes [11]

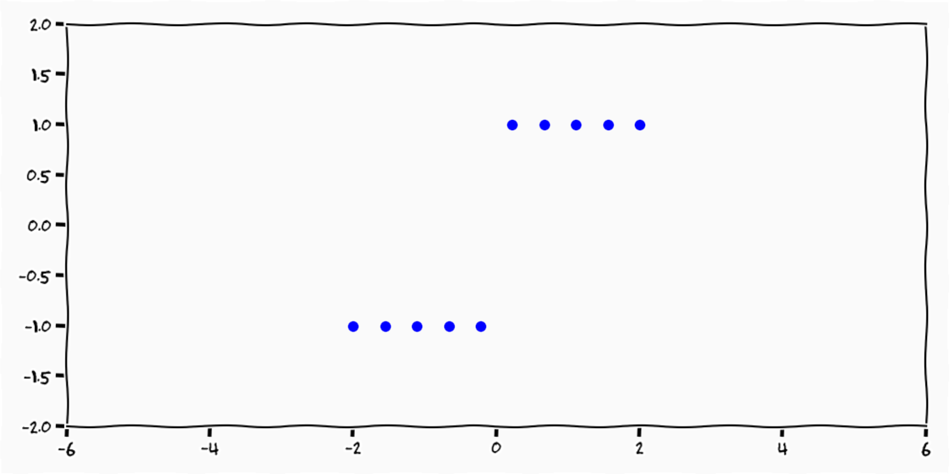


[11] Damianou et al., 2013

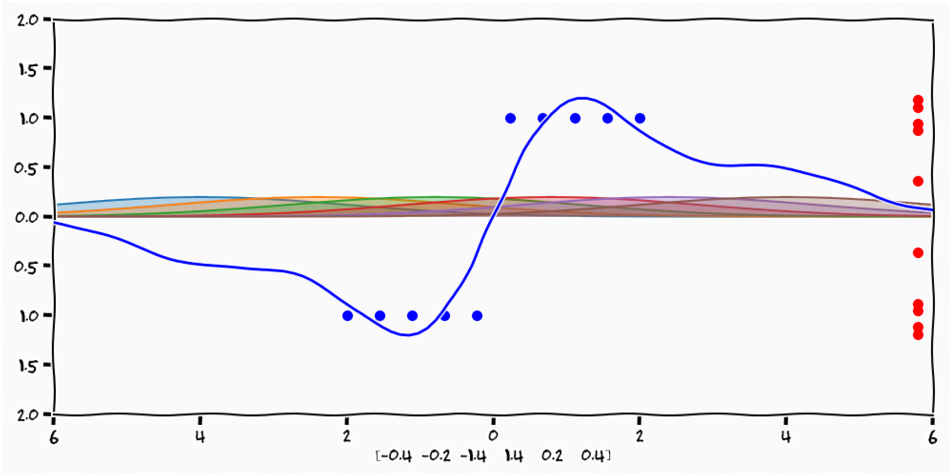
Composite Gaussian Processes



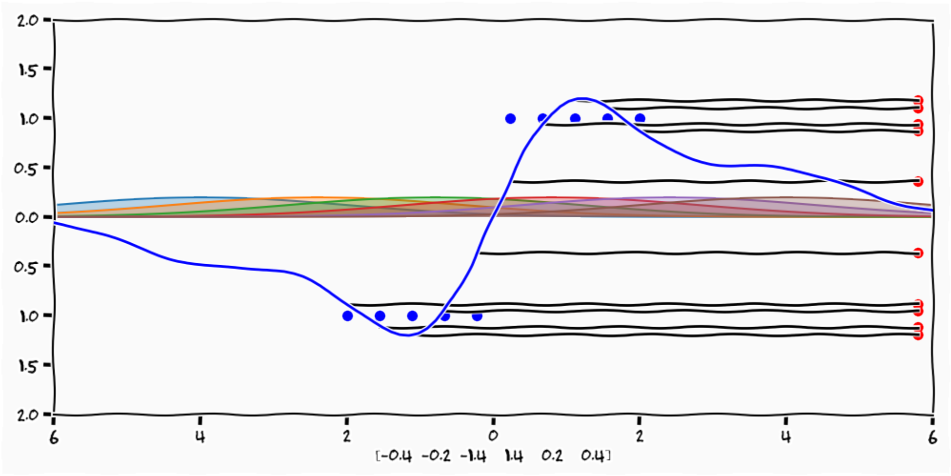
Composite Functions



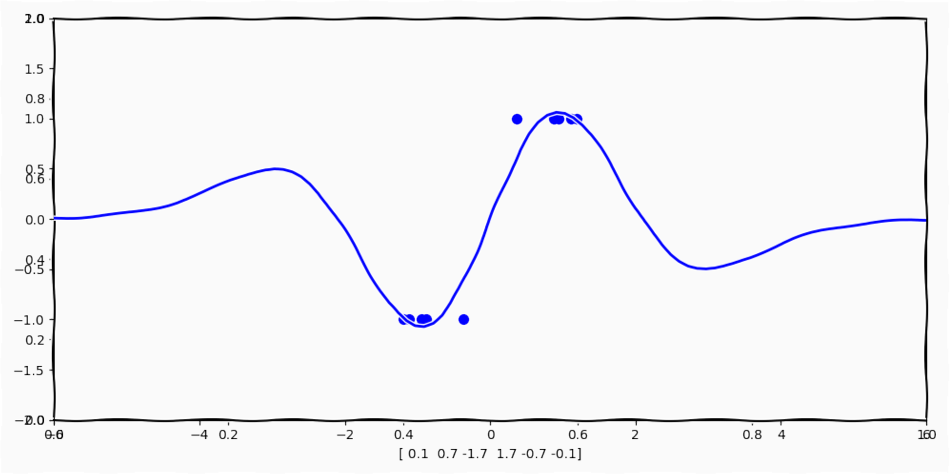
Composite Functions



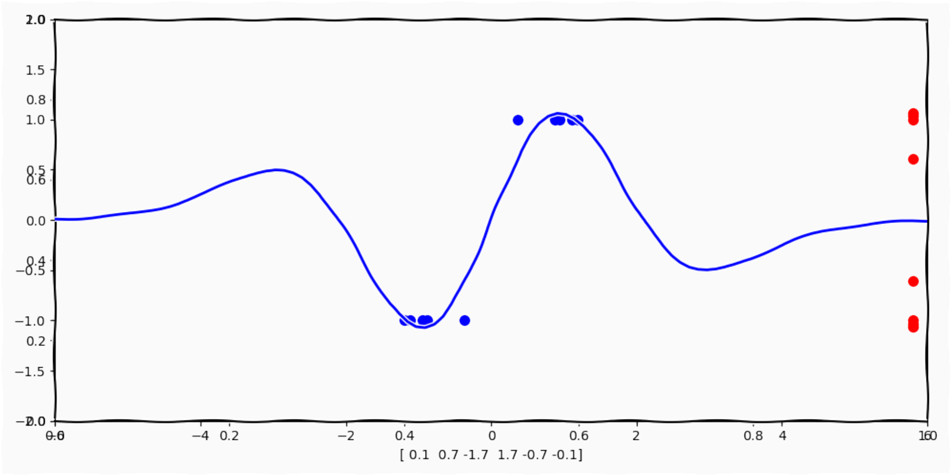
Composite Functions



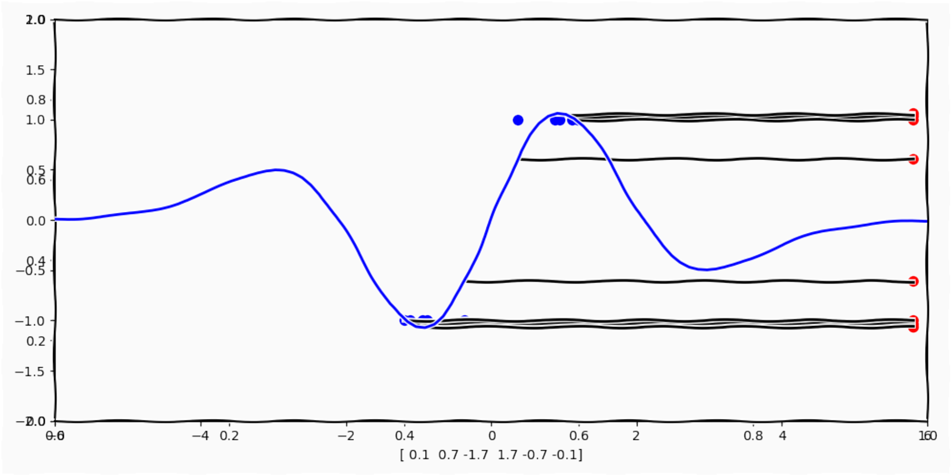
Composite Functions



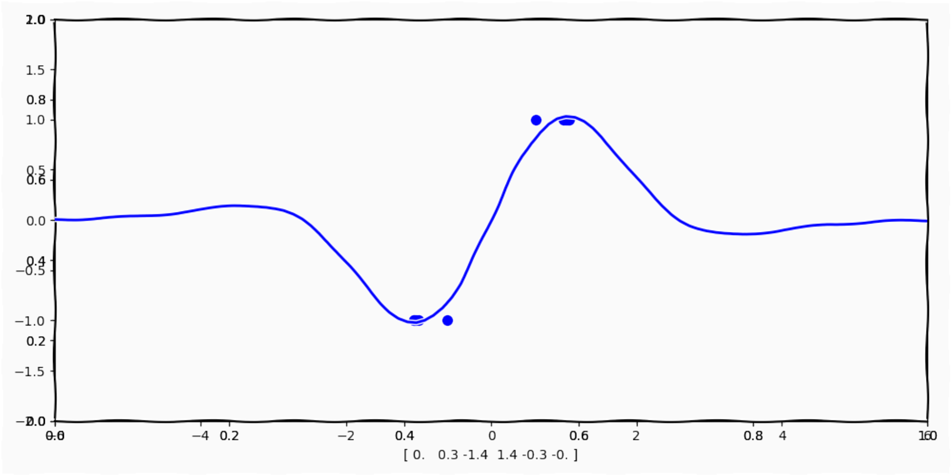
Composite Functions



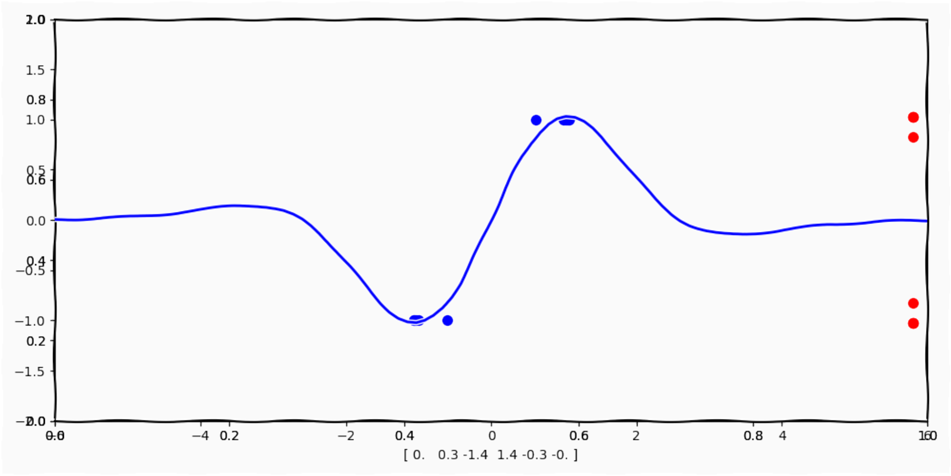
Composite Functions



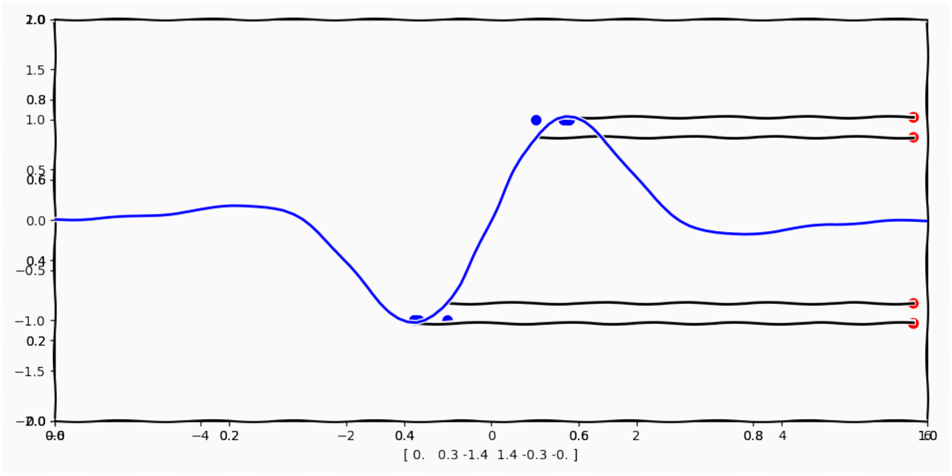
Composite Functions

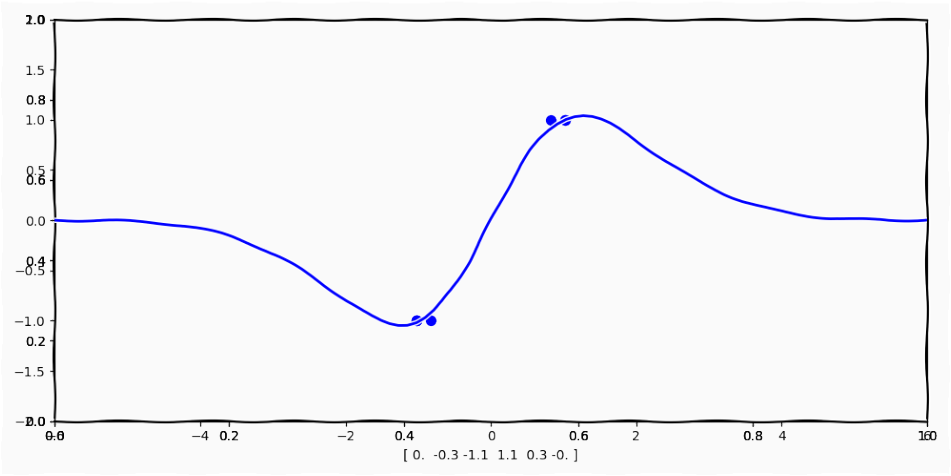


Composite Functions

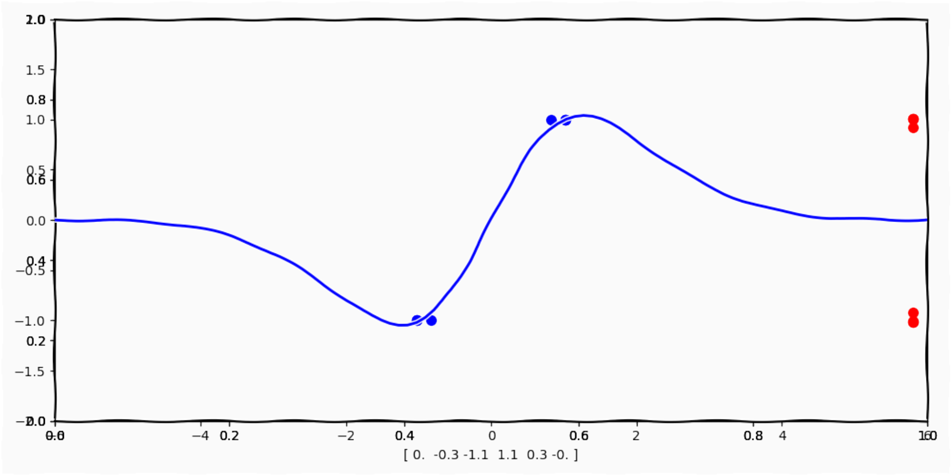


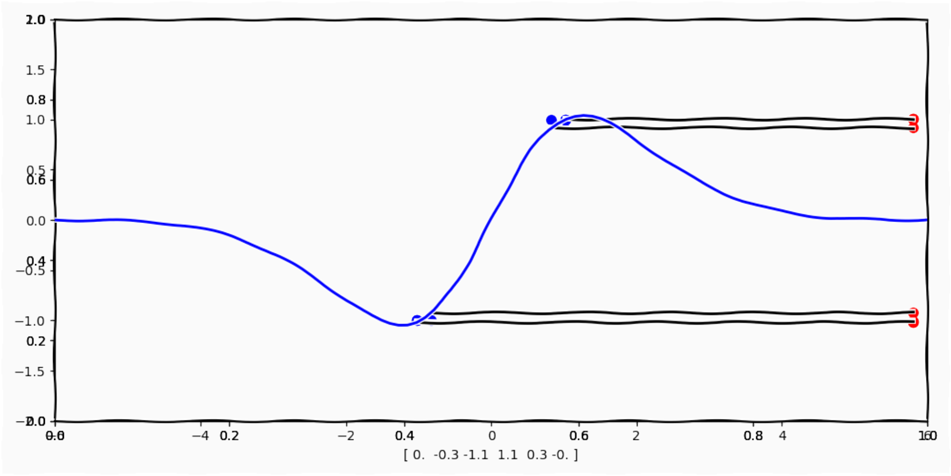
Composite Functions



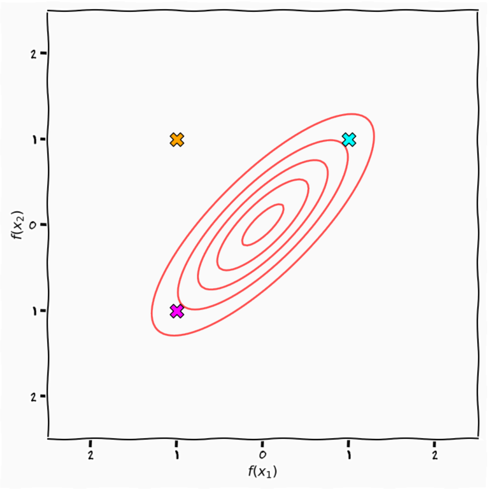

1.5
1.0
0.0
−1.0
−1.5
−4
−2
0
2
4
[0. -0.3 -1.1 1.1 0.3 -0.]

Composite Functions

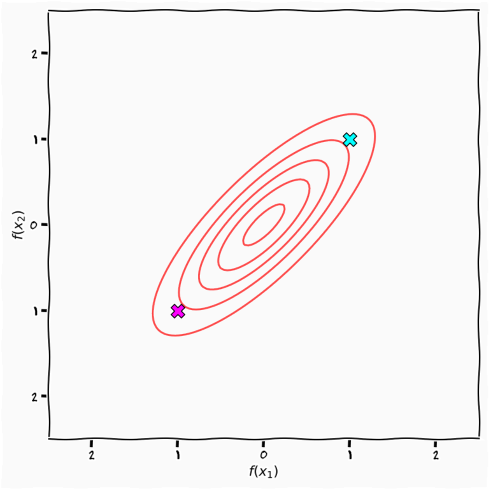



[0. -0.3 -1.1 1.1 0.3 -0.]

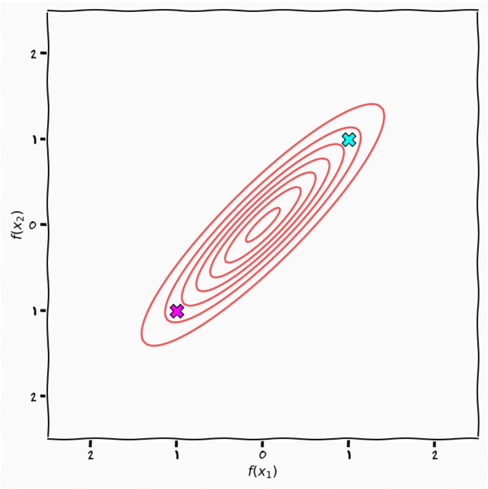
Composite Gaussian Processes

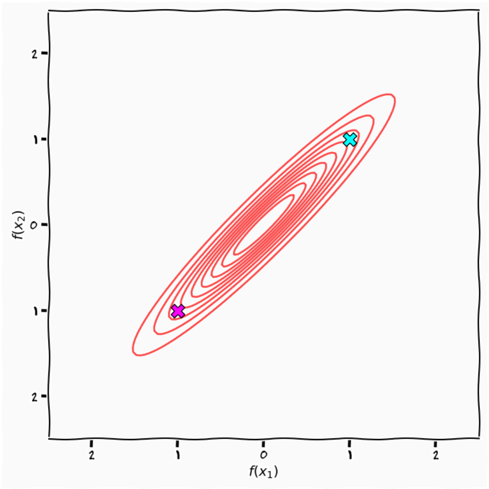


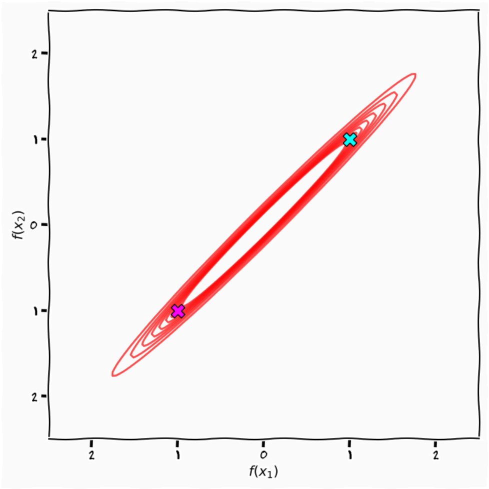
Composite Gaussian Processes

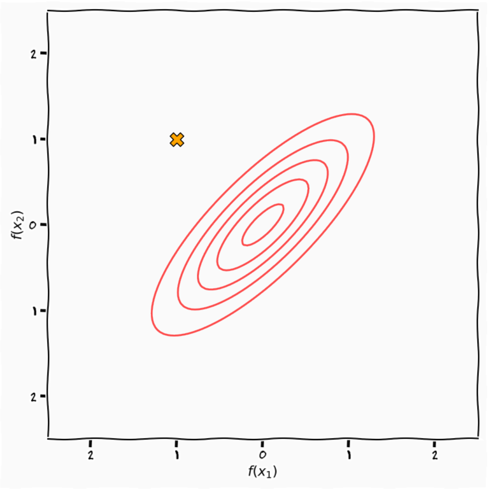


Composite Gaussian Processes

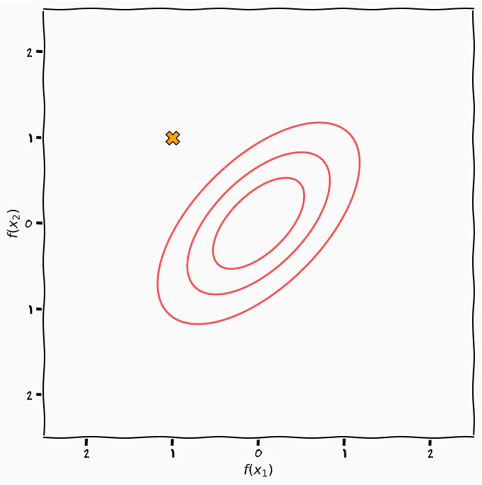



2
1
0
1
2
$f(x_2)$
2
1
0
1
2
$f(x_1)$

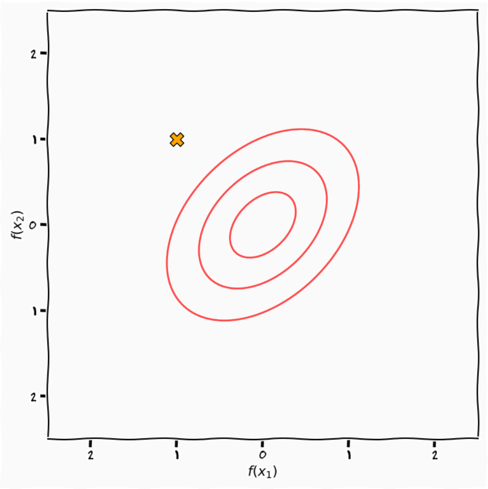

$f(x_1)$
$f(x_2)$


2
1
0
1
2
2
1
0
1
2
$f(x_1)$
$f(x_2)$

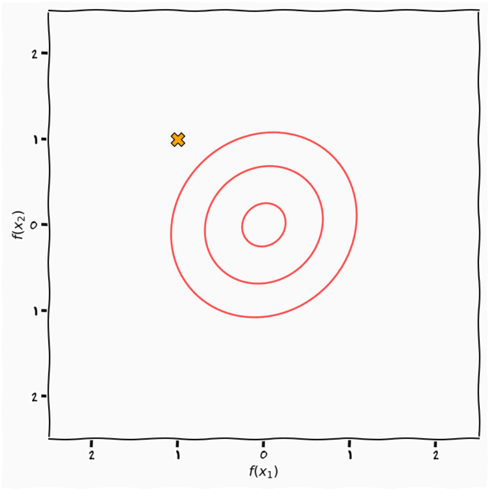
Composite Gaussian Processes



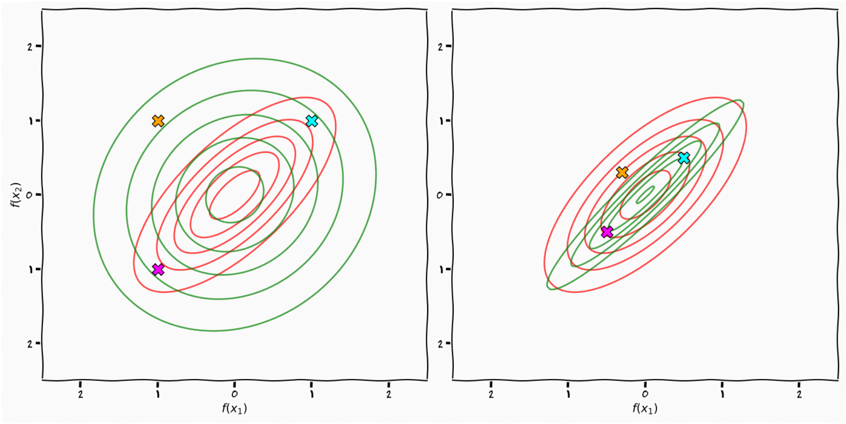
Composite Gaussian Processes



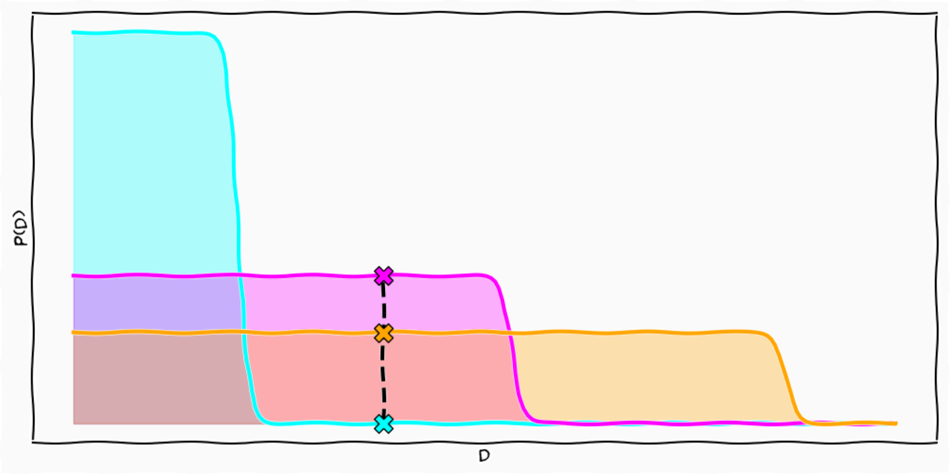
Composite Gaussian Processes



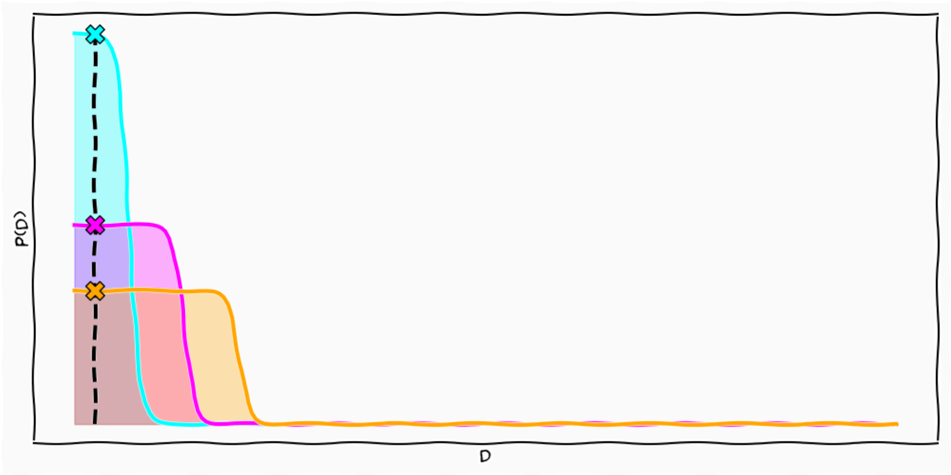
Composite Gaussian Processes

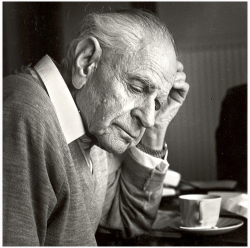


Learning



Learning

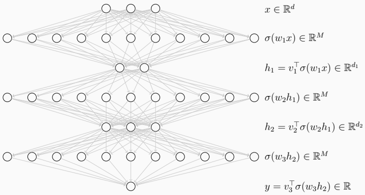


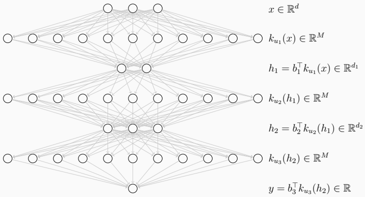


"A theory that explains everything, explains nothing"
– Karl Popper The Logic of Scientific Discovery

The Power of Scale

Equivalence class





$$\mathcal{F}_{\mathrm{DNN}}(\mathbf{x}) = \mathbf{V}_L^{\mathrm{T}} \left(\mathbf{W}_L \cdots \mathbf{V}_2^{\mathrm{T}} \sigma \left(\mathbf{W}_2 \mathbf{V}_1^{\mathrm{T}} \sigma \left(\mathbf{W}_1 \mathbf{x} \right) \right) \right)$$

$$\mathbb{E}\left[\mathcal{F}_{\mathrm{DGP}}(\mathbf{x})\right] = \mathbf{B}_L^{\mathrm{T}} \mathbf{c}_{\mathbf{u}_L} \left(\cdots \mathbf{B}_2^{\mathrm{T}} \mathbf{c}_{\mathbf{u}_2} \left(\mathbf{B}_1^{\mathrm{T}} \mathbf{c}_{\mathbf{u}_1}(\mathbf{x}) \right) \right)$$

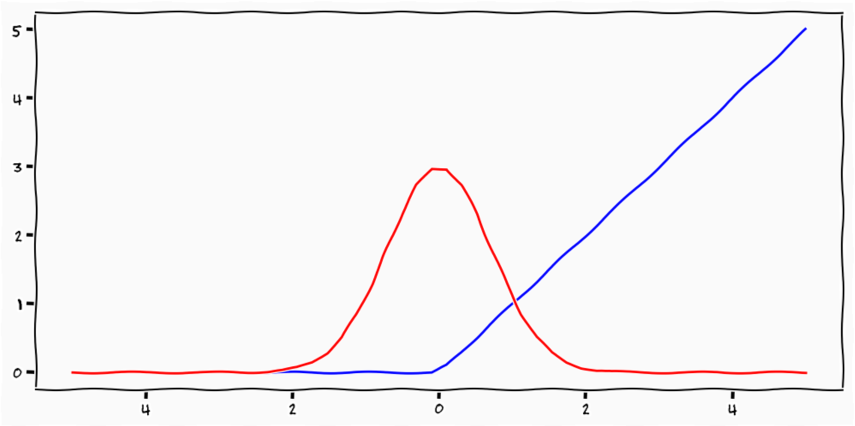
Outline

$$\mathbb{E}\left[\mathcal{F}_{\mathrm{DGP}}(\mathbf{x})\right] = \mathbf{B}_L^{\mathrm{T}}\mathbf{c}_{\mathbf{u}_L}\left(\cdots\mathbf{B}_2^{\mathrm{T}}\mathbf{c}_{\mathbf{u}_2}\left(\mathbf{B}_1^{\mathrm{T}}\mathbf{c}_{\mathbf{u}_1}(\mathbf{x})\right)\right)$$

- $\left[\mathbf{c}_{\mathbf{u}}(\cdot) \in \mathbb{R}^M\right]_m = cov\left(f(\cdot), u_m\right) = k(\cdot, \mathbf{w}_m)$
- $\left[\mathbf{C}_{\mathbf{uu}} \in \mathbb{R}^{M\times M}\right]_{m,m'} = cov(u_m, u_{m'}) = k(\mathbf{w}_m, \mathbf{w}_{m'})$
- $\left[\mathbf{B} \in \mathbb{R}^{M\times P}\right]_p = \mathbf{C}_{\mathbf{uu}}^{-1}\boldsymbol{\mu}_p$
- $q(\mathbf{u}_p) = \mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$

Activated Gaussian Processes

Basis Functions



$$u_m = f(w_m)$$

Interdomain Gaussian Processes[12]

- Sparse Gaussian Process

$$
\begin{aligned}
u_m &= f(\mathbf{w}_m) \\
[\mathbf{c}_\mathbf{u}(\cdot)]_m &= k(\mathbf{w}_m, \cdot)
\end{aligned}
$$

[12] Lázaro-Gredilla et al., 2009

Interdomain Gaussian Processes[12]

- Sparse Gaussian Process

$$u_m = f(\mathbf{w}_m)$$
$$[\mathbf{c}_\mathbf{u}(\cdot)]_m = k(\mathbf{w}_m, \cdot)$$

- Interdomain Gaussian Process

$$u_m = \mathcal{L}_m\left(f(\cdot)\right)$$
$$u_m = \int f(\mathbf{x}) g_m(\mathbf{x}) \mathrm{d}\mathbf{x}$$
$$[\mathbf{c}_\mathbf{u}(\cdot)]_m = cov(u_m, f(\cdot))$$

[12] Lázaro-Gredilla et al., 2009

Zonal Kernels

$$\sigma(\mathbf{x}^{\mathrm{T}}\mathbf{w}) \underbrace{=}_{\text{Homogeneity}} \|\mathbf{x}\|\|\mathbf{w}\|\sigma\left(\frac{\mathbf{x}^{\mathrm{T}}}{\|\mathbf{x}\|}\frac{\mathbf{w}}{\|\mathbf{w}\|}\right) = \|\mathbf{x}\|\|\mathbf{w}\|\sigma(\cos(\theta_{\mathbf{x},\mathbf{w}}))$$

- Inner-product of two points on the unit-sphere

Zonal Kernels

$$\sigma(\mathbf{x}^{\mathrm{T}}\mathbf{w}) \underbrace{=}_{\text{Homogeneity}} \|\mathbf{x}\|\|\mathbf{w}\|\sigma\left(\frac{\mathbf{x}^{\mathrm{T}}}{\|\mathbf{x}\|}\frac{\mathbf{w}}{\|\mathbf{w}\|}\right) = \|\mathbf{x}\|\|\mathbf{w}\|\sigma(\cos(\theta_{\mathbf{x},\mathbf{w}}))$$

- Inner-product of two points on the unit-sphere
- Zonal kernels

Zonal Kernels

$$\sigma(\mathbf{x}^{\mathrm{T}}\mathbf{w}) \underbrace{=}_{\text{Homogeneity}} \|\mathbf{x}\|\|\mathbf{w}\|\sigma\left(\frac{\mathbf{x}^{\mathrm{T}}}{\|\mathbf{x}\|}\frac{\mathbf{w}}{\|\mathbf{w}\|}\right) = \|\mathbf{x}\|\|\mathbf{w}\|\sigma(\cos(\theta_{\mathbf{x},\mathbf{w}}))$$

- Inner-product of two points on the unit-sphere
- Zonal kernels
  - rotationally invariant

Zonal Kernels

$$\sigma(\mathbf{x}^{\mathrm{T}}\mathbf{w}) \underbrace{=}_{\text{Homogeneity}} \|\mathbf{x}\|\|\mathbf{w}\|\sigma\left(\frac{\mathbf{x}^{\mathrm{T}}}{\|\mathbf{x}\|}\frac{\mathbf{w}}{\|\mathbf{w}\|}\right) = \|\mathbf{x}\|\|\mathbf{w}\|\sigma(\cos(\theta_{\mathbf{x},\mathbf{w}}))$$

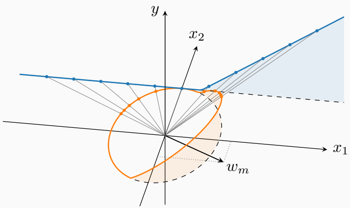
- Inner-product of two points on the unit-sphere
- Zonal kernels
  - rotationally invariant
  - we can construct the RKHS

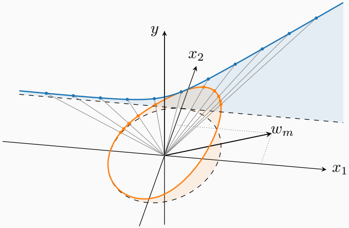
Zonal Kernels

$$\sigma(\mathbf{x}^\mathrm{T}\mathbf{w}) \underbrace{=}_{\text{Homogeneity}} \|\mathbf{x}\|\|\mathbf{w}\|\sigma\left(\frac{\mathbf{x}^\mathrm{T}}{\|\mathbf{x}\|}\frac{\mathbf{w}}{\|\mathbf{w}\|}\right) = \|\mathbf{x}\|\|\mathbf{w}\|\sigma(\cos(\theta_{\mathbf{x},\mathbf{w}}))$$

- Inner-product of two points on the unit-sphere
- Zonal kernels
  - rotationally invariant
  - we can construct the RKHS
  - eigenfunctions are the spherical harmonics

Example





- Inducing Function

$$g_m(\cdot) : \mathbb{S}^{d-1} \to \mathbb{R}$$
$$g_m(\mathbf{x}) = \sigma(\mathbf{w}_m^{\mathrm{T}}\mathbf{x})$$

- Inducing output

$$u_m = \int f(\mathbf{x}) g_m(\mathbf{x}) \mathrm{d}x = \langle f(\cdot), g_m(\cdot) \rangle_{\mathcal{H}}$$

- Reproducing Property

$$\langle f, K_x(\cdot)\rangle_{\mathcal{H}} = f(x)$$

Covariances

- Reproducing Property

$$\langle f, K_x(\cdot)\rangle_{\mathcal{H}} = f(x)$$

- Reproducing property for covariances

$$cov(u_m, f(\mathbf{x})) = \langle k(\mathbf{x}, \cdot), g_m\rangle_{\mathcal{H}} = g_m(\mathbf{x})$$
$$cov(u_m, u_{m'}) = cov(\langle f, g_m\rangle_{\mathcal{H}}, \langle f, g_{m'}\rangle_{\mathcal{H}}) = \langle g_m, g_{m'}\rangle_{\mathcal{H}}$$

Activated Gaussian Processes

$$\mathbb{E}\left[\mathcal{F}_{\mathrm{DGP}}(\mathbf{x})\right] = \mathbf{B}_L^{\mathrm{T}}\mathbf{c}_{\mathbf{u}_L}\left(\cdots\mathbf{B}_2^{\mathrm{T}}\mathbf{c}_{\mathbf{u}_2}\left(\mathbf{B}_1^{\mathrm{T}}\mathbf{c}_{\mathbf{u}_1}(\mathbf{x})\right)\right)$$

- $\left[\mathbf{c}_{\mathbf{u}}(\cdot) \in \mathbb{R}^M\right]_m = cov\left(f(\cdot), u_m\right) = g_m(\cdot)$
- $\left[\mathbf{C}_{\mathbf{uu}} \in \mathbb{R}^{M \times M}\right]_{m,m'} = cov(u_m, u_{m'}) = \langle g_m, g_{m'} \rangle_{\mathcal{H}}$
- $\left[\mathbf{B} \in \mathbb{R}^{M \times P}\right]_p = \mathbf{C}_{\mathbf{uu}}^{-1}\boldsymbol{\mu}_p$

Interdomain GPs allows for inducing *functions* rather than points (global)

Interdomain GPs allows for inducing *functions* rather than points (global)

Basis decomposition ReLU functions are projections of spherical functions (matching activations)

Interdomain GPs allows for inducing *functions* rather than points (global)

Basis decomposition ReLU functions are projections of spherical functions (matching activations)

Access to RKHS allows for efficient computations using orthonormal basis (spherical harmonics)

Interdomain GPs allows for inducing *functions* rather than points (global)

Basis decomposition ReLU functions are projections of spherical functions (matching activations)

Access to RKHS allows for efficient computations using orthonormal basis (spherical harmonics)

Combination Specify GP on sphere, use interdomain function to generate activations

Experiments

Classification

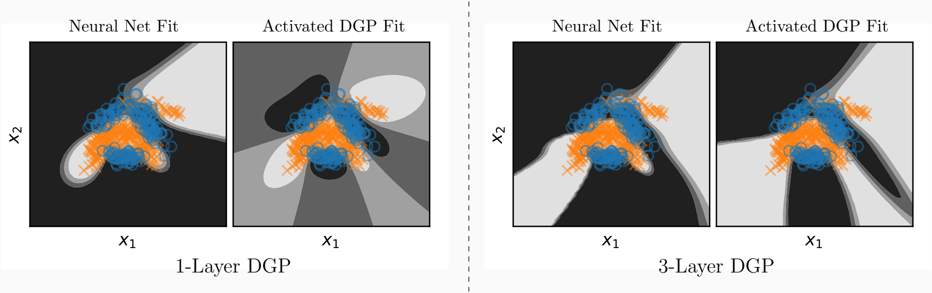
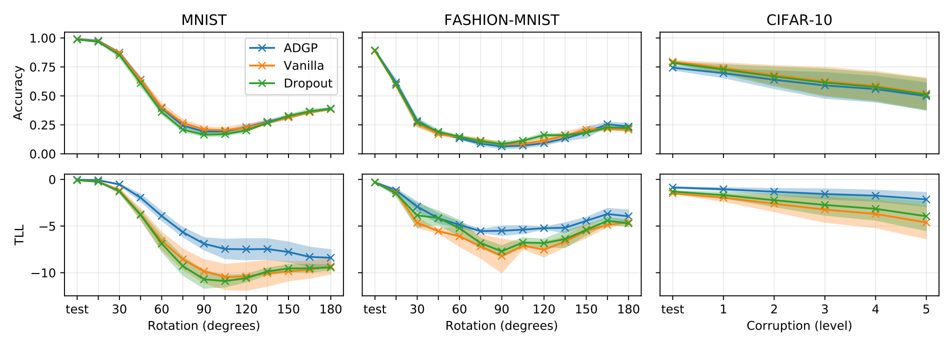


Image Classification



Summary

Summary

Prior long time established connection between NN and GPs

Summary

Prior long time established connection between NN and GPs

Posterior finite basis function models

Summary

Prior long time established connection between NN and GPs

Posterior finite basis function models

Activated GPs allows equivalence between composite GPs and GLM

Summary

Prior long time established connection between NN and GPs

Posterior finite basis function models

Activated GPs allows equivalence between composite GPs and GLM

Benefit exploit all the heuristics from NN in a model

Paper

Vincent Dutordoir et al. (2021). "Deep Neural Networks As Point Estimates for Deep Gaussian Processes". In: *CoRR*

References

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

Damianou, Andreas, Neil D. Lawrence, and Carl Henrik Ek (2021). "Multi-view Learning as a Nonparametric Nonlinear Inter-Battery Factor Analysis". In: *Journal of Machine Learning Research* 22.86, pp. 1–51.

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.

Damianou, Andreas C., Michalis Titsias, and Neil D. Lawrence (2011). "Variational Gaussian Process Dynamical Systems". In: *Advances in Neural*

Information Processing Systems 24. Ed. by J. Shawe-Taylor et al., pp. 2510–2518.

Dutordoir, Vincent et al. (2021). "Deep Neural Networks As Point Estimates for Deep Gaussian Processes". In: *CoRR*.

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.

Lázaro-Gredilla, Miguel and Aníbal Figueiras-Vidal (2009). "Inter-domain Gaussian Processes for Sparse Inference using Inducing Features". In:

Advances in Neural Information Processing Systems. Ed. by Y. Bengio et al. Vol. 22. Curran Associates, Inc.

Popper, K.R. (1959). *The Logic of Scientific Discovery*. ISSR library. Routledge.

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

Ustyuzhaninov, Ivan et al. (2020). "Compositional uncertainty in deep Gaussian processes". In: *Proceedings of the Thirty-Sixth Conference on Uncertainty in Artificial Intelligence, UAI 2020, virtual online, August 3-6, 2020*. Ed. by Ryan P. Adams and Vibhav Gogate. Vol. 124. Proceedings of Machine Learning Research. AUAI Press, pp. 480–489.