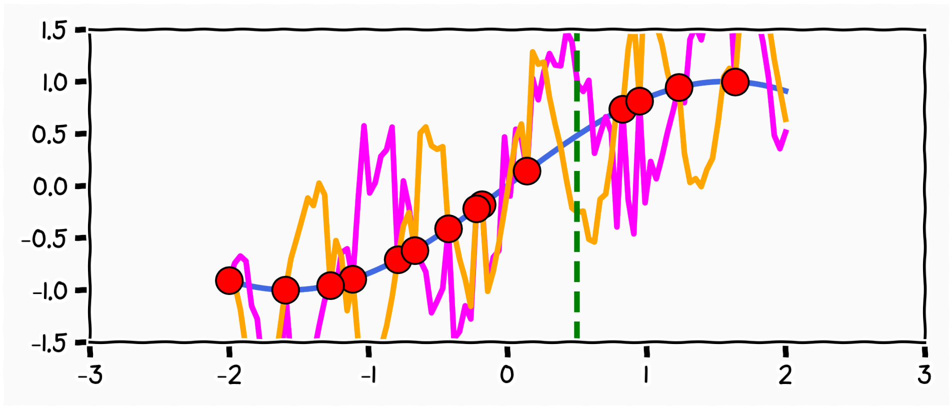
Approximate Bayesian Inference of Composite Functions

Carl Henrik Ek

September 9, 2025

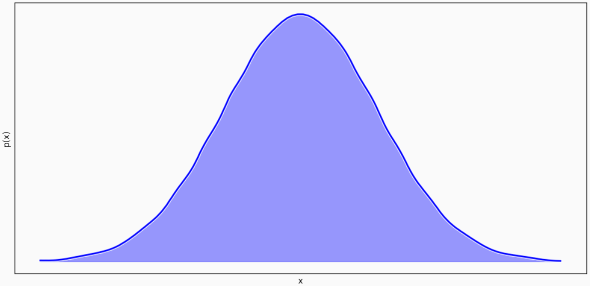
http://carlhenrik.com


1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-3
-2
-1
0
1
2
3

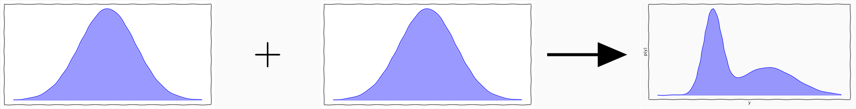
What is Machine Learning

$$\text{data} + \text{model} \rightarrow \text{prediction}$$

Quantification of Knowledge

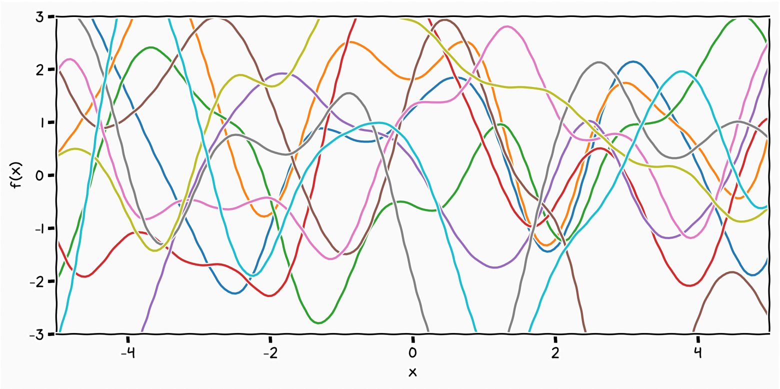


Knowledge (Uncertainty) Propagation

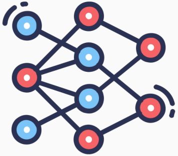


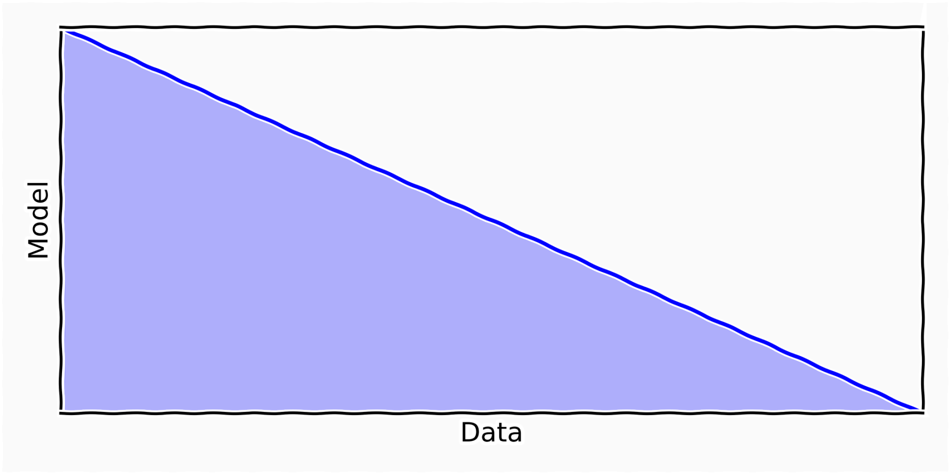
Data Model Prediction

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



$$p(f) = \mathcal{GP}(\mu(x), k(x, x'))$$

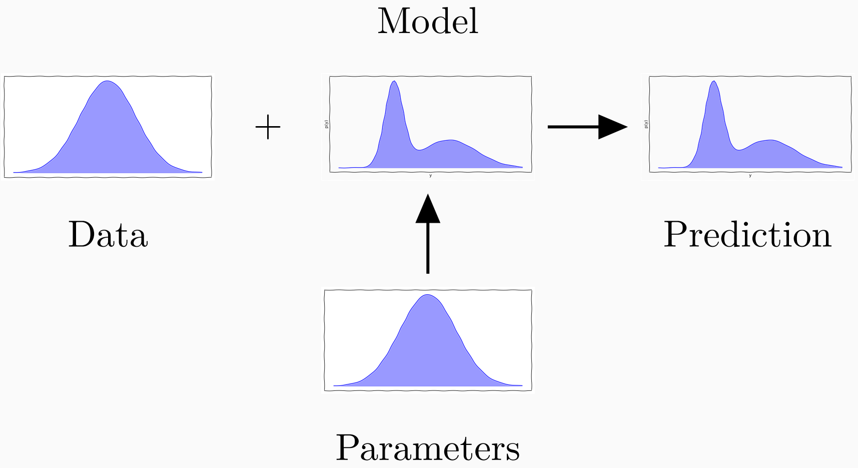


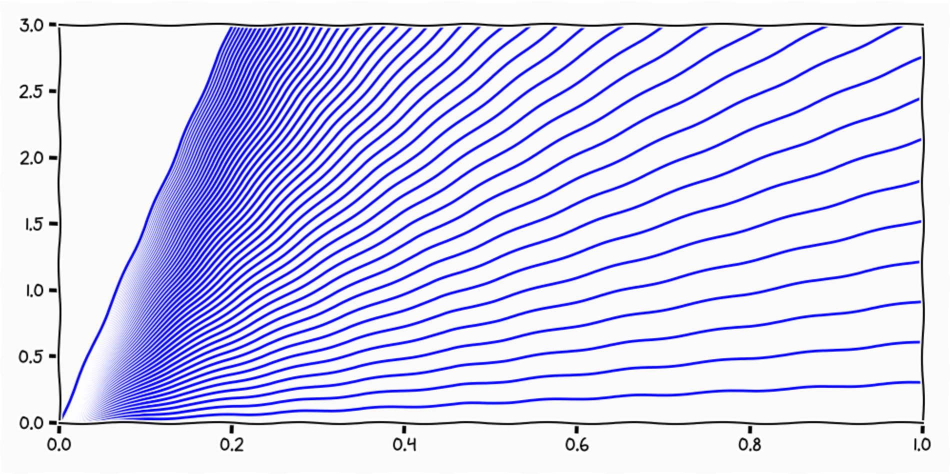


Model
Data

$$\mathbf{y} = \underbrace{f_N}_{w_N} \circ \underbrace{f_{N-1}}_{w_{N-1}} \circ \cdots \circ \underbrace{f_1}_{w_1} \circ \underbrace{f_0}_{w_0}(\mathbf{x})$$

Parametric Knowledge (Uncertainty) Propagation




3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.0
0.2
0.4
0.6
0.8
1.0

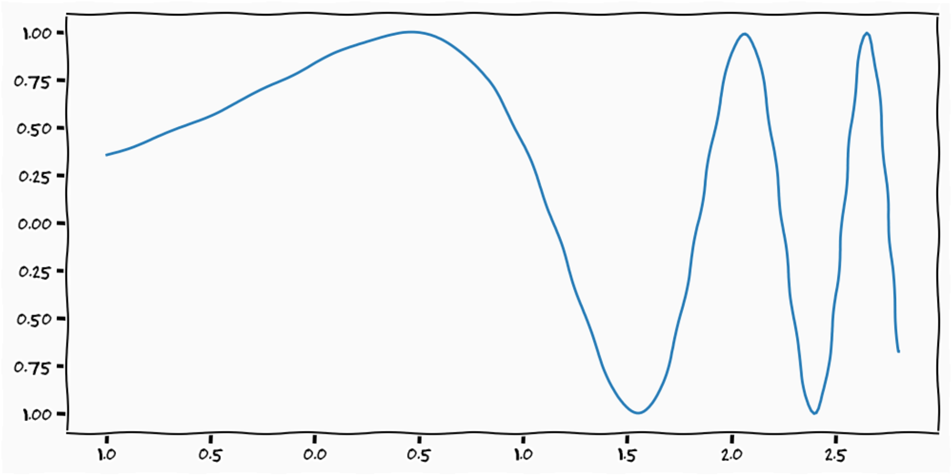
$$\mathbf{y} = \underbrace{f_N}_{w_N} \circ \underbrace{f_{N-1}}_{w_{N-1}} \circ \cdots \circ \underbrace{f_1}_{w_1} \circ \underbrace{f_0}_{w_0} (\mathbf{x})$$

When do I want Composite Functions

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

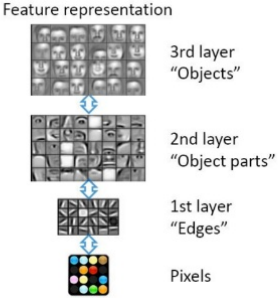
1. My generative process is composite
   - my prior knowledge is composite
2. I want to ”re-parametrise” my kernel in a learning setting
   - i have knowledge of the re-parametrisation

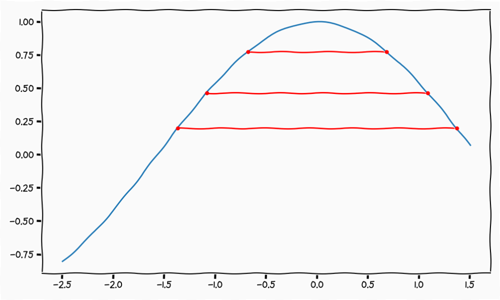
Because we lack ”models”?



Diff Levels of Abstraction

- Hierarchical Learning
  - Natural progression from low level to high level structure as seen in natural complexity
  - Easier to monitor what is being learnt and to guide the machine to better subspaces
  - A good lower level representation can be used for many distinct tasks



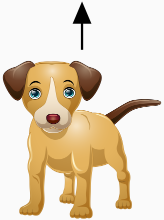


$$y = f_k(f_{k-1}(\dots f_0(x))) = f_k \circ f_{k-1} \circ \cdots \circ f_1(x)$$

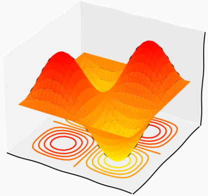
$$\mathrm{Kern}(f_1) \subseteq \mathrm{Kern}(f_{k-1} \circ \dots \circ f_2 \circ f_1) \subseteq \mathrm{Kern}(f_k \circ f_{k-1} \circ \dots \circ f_2 \circ f_1)$$

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

Dog

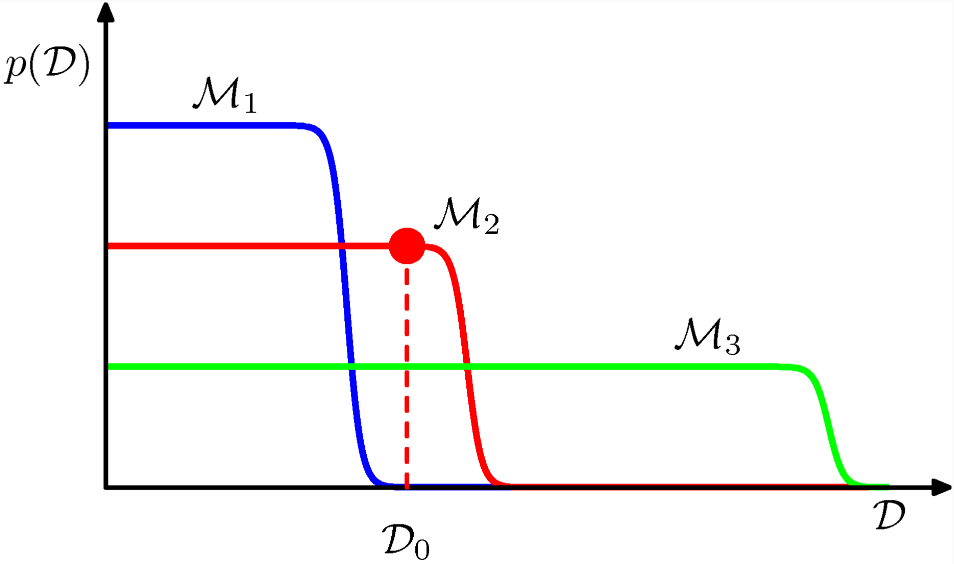


Cat

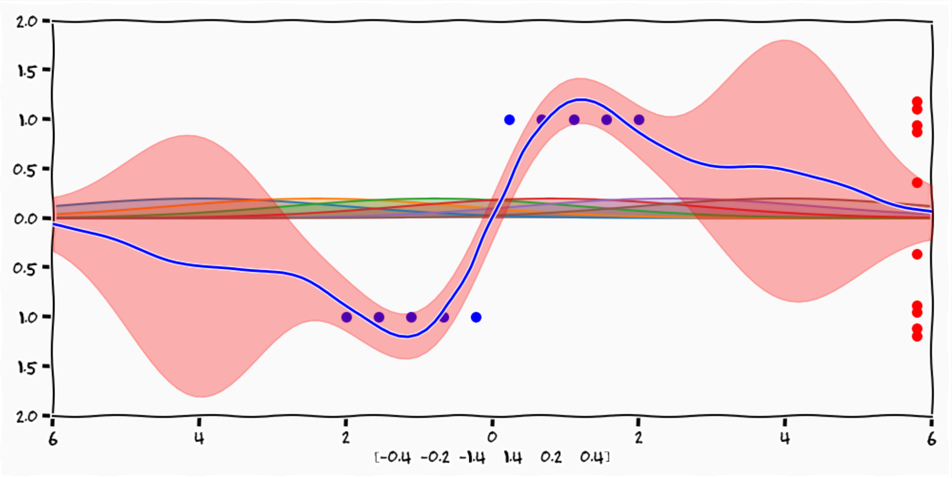



WE ARE THE BAYESIAN.
YOU WILL BE ASSIMILATED. YOUR TECHNOLOGICAL DISTINCTIVENESS
WILL BE CONSIDERED A SPECIAL CASE OF OUR OWN. RESISTANCE IS FUTILE.
imgflip.com

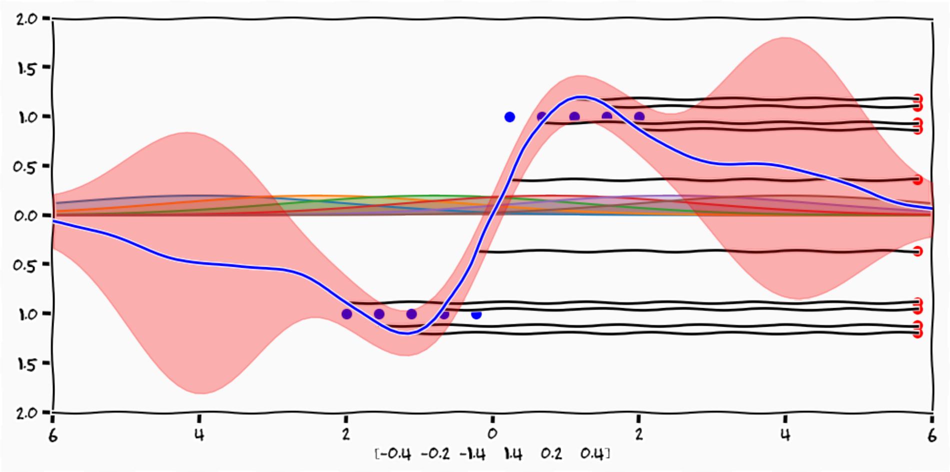
$$p(\mathbf{y}) = \int p(\mathbf{y} \mid w_N)p(w_N \mid w_{N-1}) \cdots p(w_1 \mid w_0)p(w_0)\mathrm{d}w_N\mathrm{d}w_{N-1}\mathrm{d}w_1\mathrm{d}w_0$$

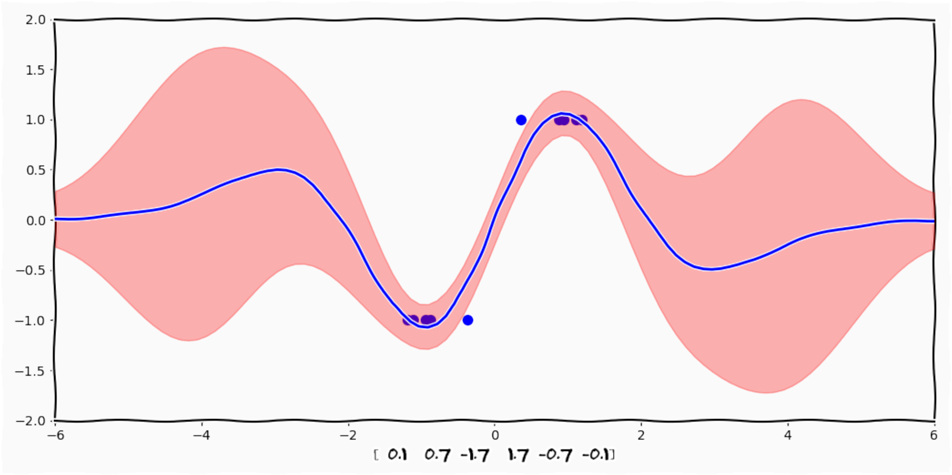

$p(\mathcal{D})$
$\mathcal{M}_1$
$\mathcal{M}_2$
$\mathcal{M}_3$
$\mathcal{D}$
$\mathcal{D}_0$

Composite Functions

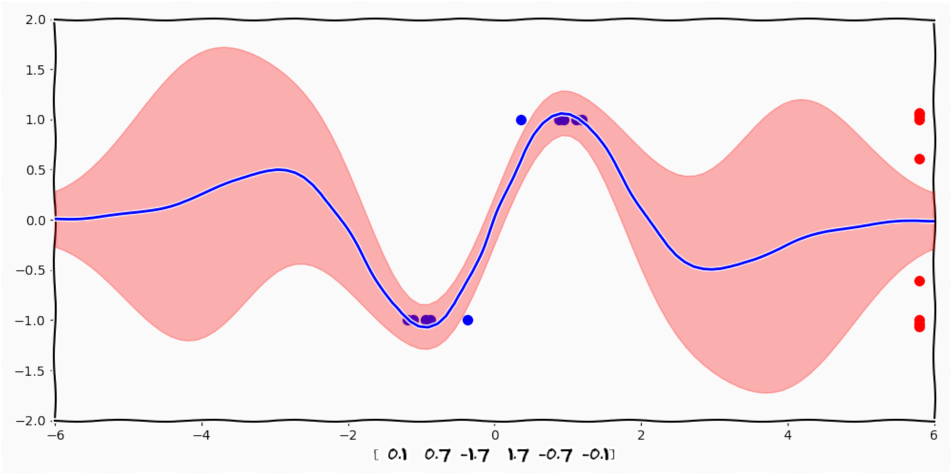


Composite Functions

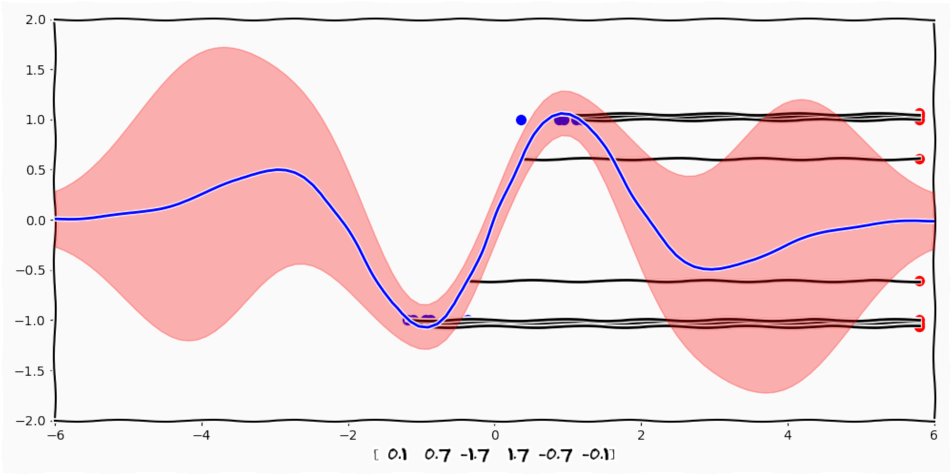



2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−6
−4
−2
0
2
4
6
[0.1 0.7 -1.7 1.7 -0.7 -0.1]

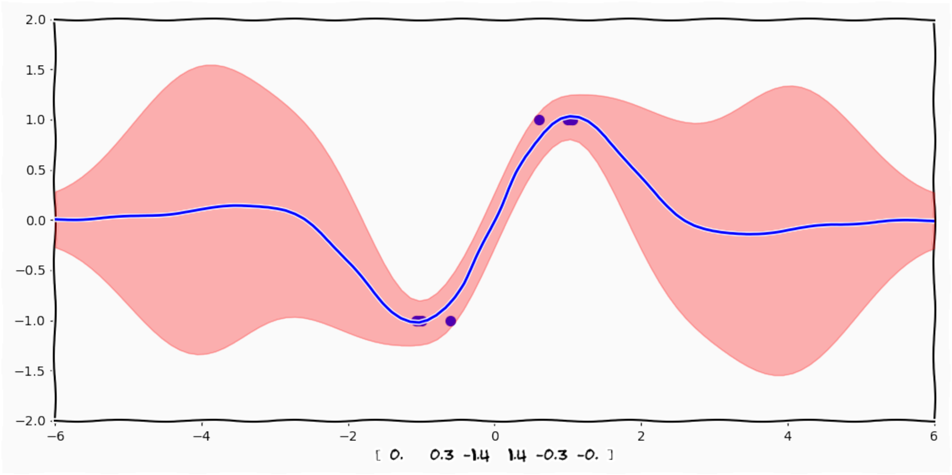
Composite Functions



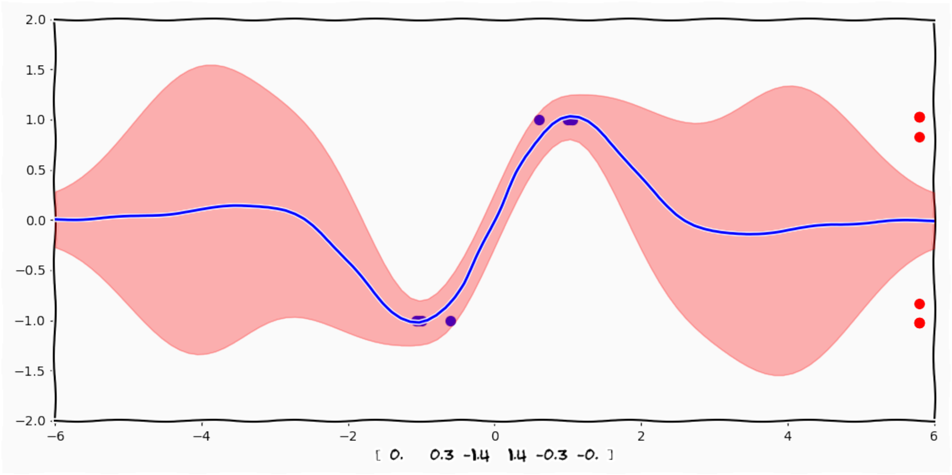
Composite Functions



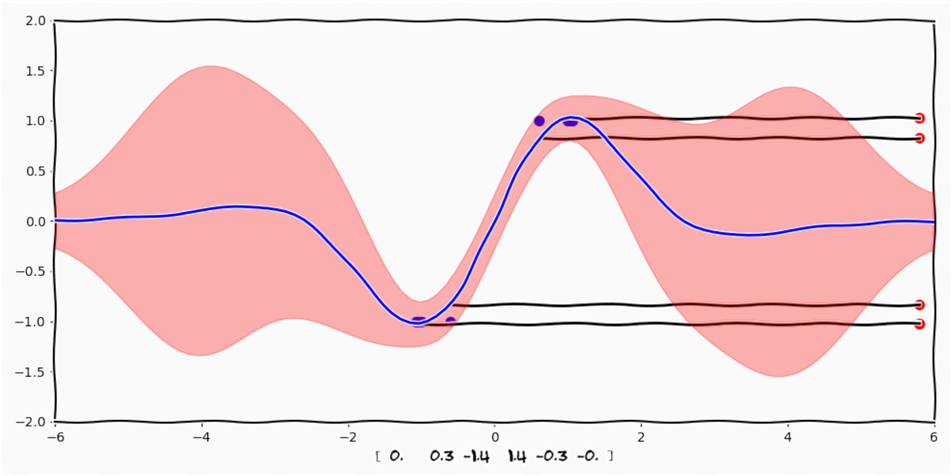
Composite Functions



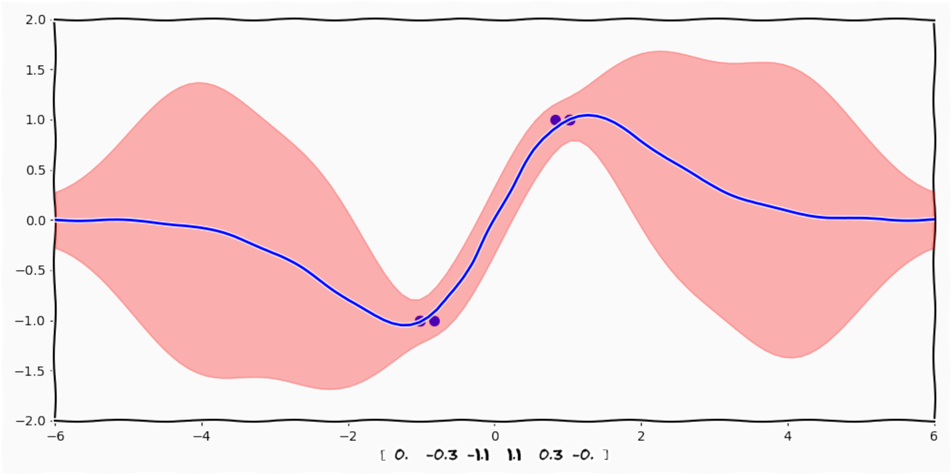
Composite Functions



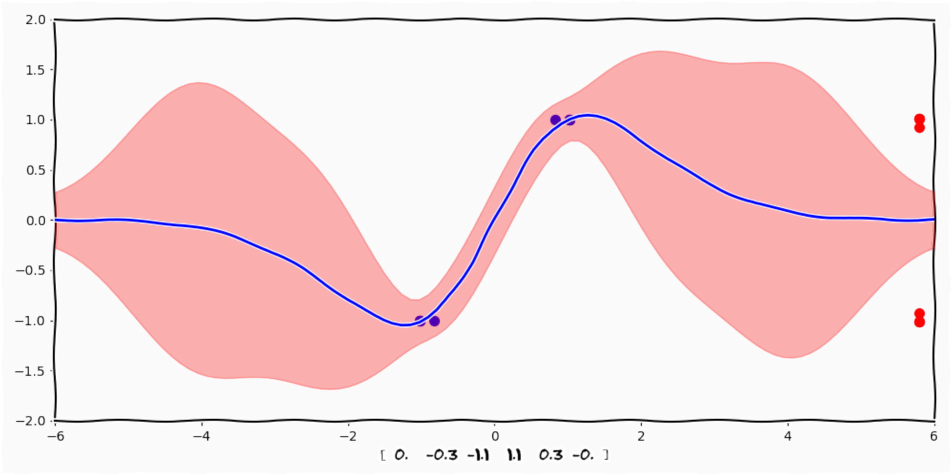
Composite Functions



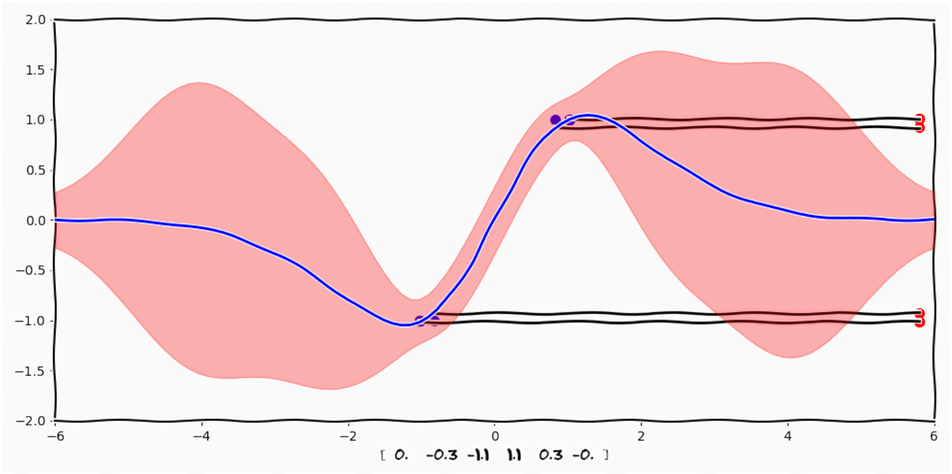
Composite Functions



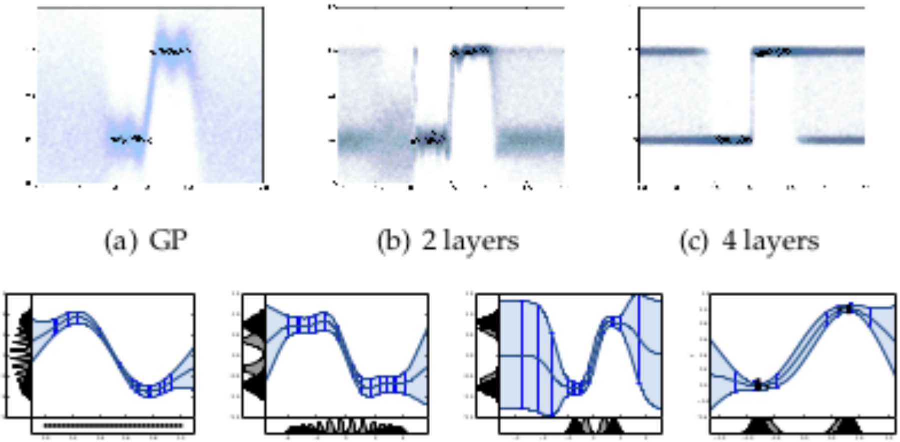
Composite Functions



Composite Functions



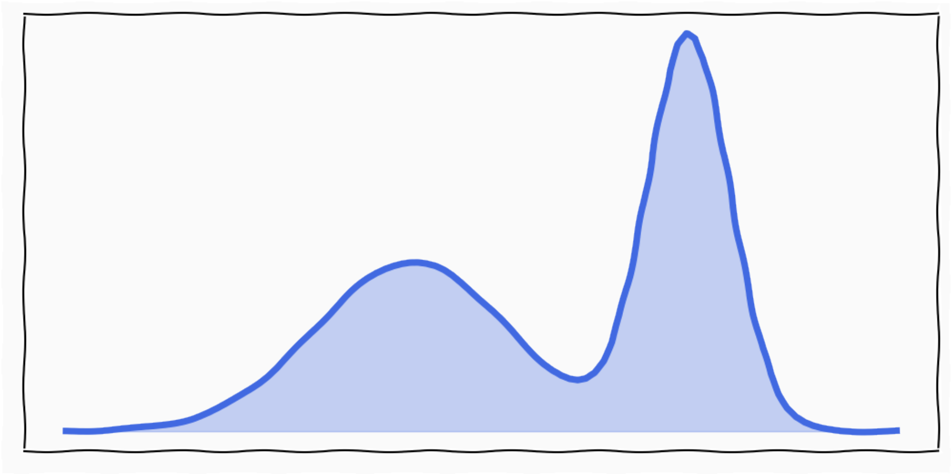
The Final Composition



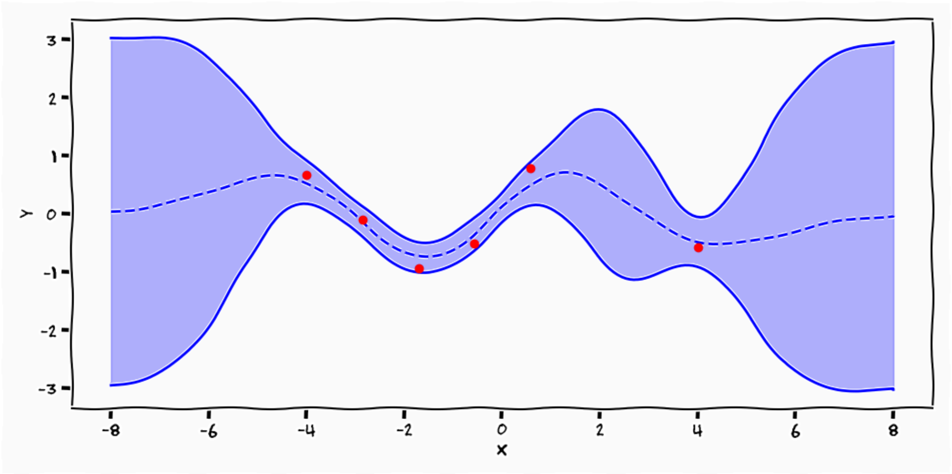
(a) GP (b) 2 layers (c) 4 layers

(d) Hidden spaces for 4 layer model

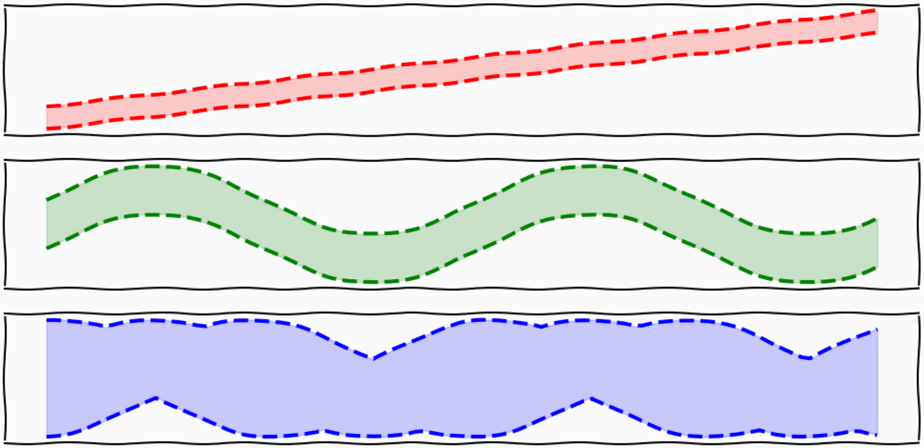
Remember why we did this in the first place



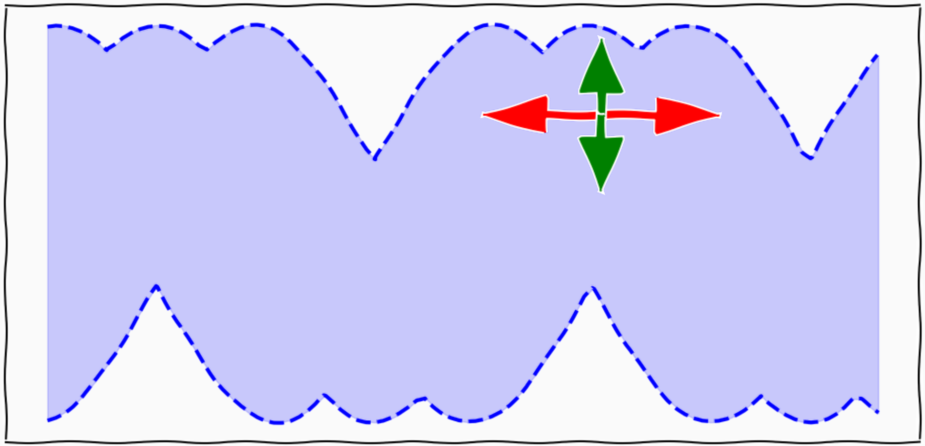
These damn plots



It gets worse



It gets even worse

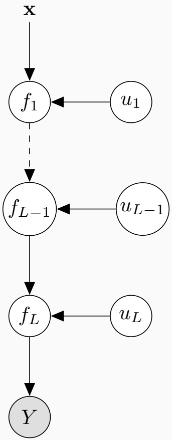


- Sufficient statistics

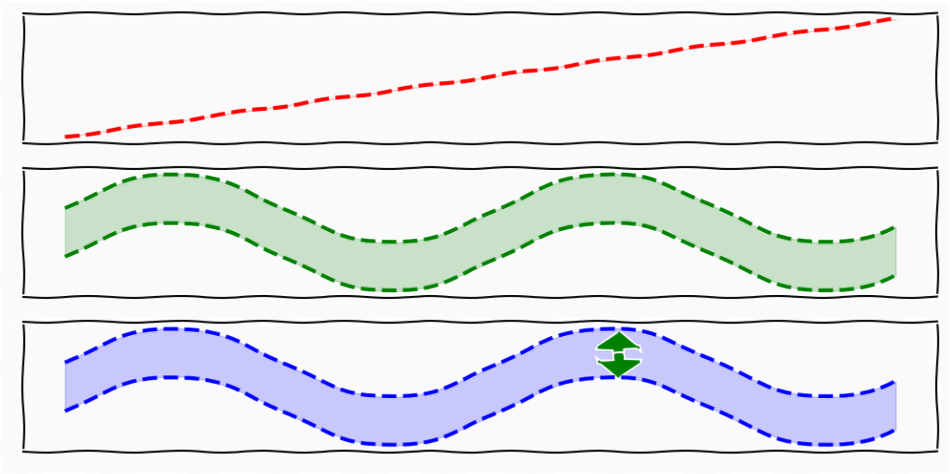
$$q(\mathbf{F})q(\mathbf{U})q(\mathbf{X}) = p(\mathbf{F}|\mathbf{Y},\mathbf{U},\mathbf{X},\mathbf{Z})q(\mathbf{U})q(\mathbf{X})$$
$$= p(\mathbf{F}|\mathbf{U},\mathbf{X},\mathbf{Z})q(\mathbf{U})q(\mathbf{X})$$

- 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

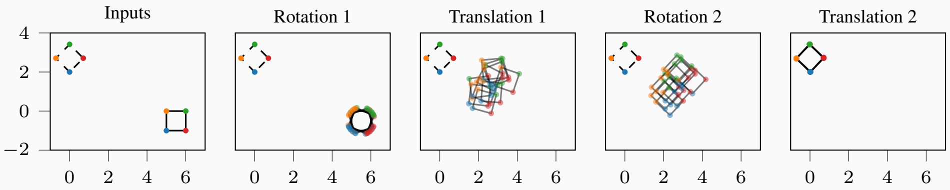
What have we lost

- 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

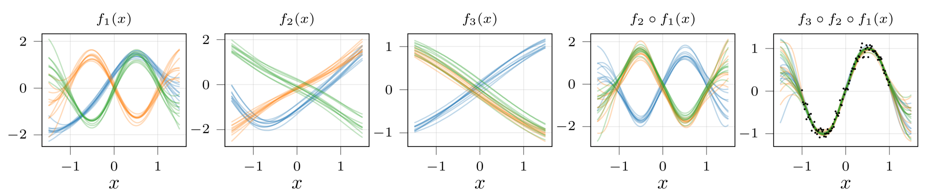
- Our priors are not reflected correctly
  - $\rightarrow$ we cannot interpret the results
- No intermediate uncertainties
  - $\rightarrow$ we cannot do sequential decision making
- We are performing a massive computational overhead for very little use
- *"...throwing out the baby with the bathwater..."*

What we really want[1]



[1] Ustyuzhaninov et al., 2020

What we really want[2]



*

[2]Ustyuzhaninov et al., 2020

State of the World

- The community have tried Bayesian principles for composite functions for a long time MacKay, 1991

State of the World

- The community have tried Bayesian principles for composite functions for a long time MacKay, 1991
- Empirically performance of Bayesian inference of composite functions is not impressive

State of the World

- The community have tried Bayesian principles for composite functions for a long time MacKay, 1991
- Empirically performance of Bayesian inference of composite functions is not impressive
- It is not just parametric models, non-parametric composite models doesn't work either

- The community have tried Bayesian principles for composite functions for a long time MacKay, 1991
- Empirically performance of Bayesian inference of composite functions is not impressive
- It is not just parametric models, non-parametric composite models doesn't work either
- *Why?*

- Roy, H., Miani, M., Ek, C. H., Hennig, P., Pförtner, M., Tatzel, L., & Hauberg, S., (2024). Reparameterization invariance in approximate bayesian inference. In Advances in Neural Information Processing Systems (NeurIPS)
- Fadel, S., Roy, H., Krämer, N., Zainchkovskyy, Y., Syrota, S., Mahou, A., Ek, C. H,. Hauberg, S. (2025). Deep variational inference with stochastic projections. In Submission

Statistical Models of Composite Functions

Parametric Composite Models

$$\mathbf{y} = f(\mathbf{x}) = \underbrace{f_N}_{w_N} \circ \underbrace{f_{N-1}}_{w_{N-1}} \circ \cdots \circ \underbrace{f_1}_{w_1} \circ \underbrace{f_0}_{w_0}(\mathbf{x})$$

Inference

$$p(\mathbf{w} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \mathbf{w}) p(\mathbf{w})}{p(\mathcal{D})}$$

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{0}, \alpha \mathbf{I})$$

Laplace Approximation MacKay, 1991

$$p(\mathbf{w} \mid \mathcal{D}) = \frac{1}{Z} p(\mathcal{D} \mid \mathbf{w}) p(\mathbf{w}) = \frac{1}{Z} \exp(-\mathcal{L}(\mathcal{D}; \mathbf{w}))$$

$$\hat{\mathbf{w}} = \operatorname{argmin}_{\mathbf{w}} \mathcal{L}(\mathcal{D}; \mathbf{w})$$

$$\mathcal{L}(\mathcal{D}; \mathbf{w}) \approx \mathcal{L}(\mathcal{D}; \hat{\mathbf{w}}) + \nabla \mathcal{L}(\mathcal{D}; \mathbf{w})\big|_{\mathbf{w}=\hat{\mathbf{w}}} (\mathbf{w} - \hat{\mathbf{w}}) \\ + \frac{1}{2}(\mathbf{w} - \hat{\mathbf{w}})^{\mathrm{T}} \nabla^2 \mathcal{L}(\mathcal{D}; \mathbf{w})\big|_{\mathbf{w}=\hat{\mathbf{w}}} (\mathbf{w} - \hat{\mathbf{w}})$$

Laplace Approximation

$$\begin{aligned}\mathcal{L}(\mathcal{D};\mathbf{w}) &\approx \mathcal{L}(\mathcal{D};\hat{\mathbf{w}}) + \nabla\mathcal{L}(\mathcal{D};\mathbf{w})\big|_{\mathbf{w}=\hat{\mathbf{w}}}(\mathbf{w}-\hat{\mathbf{w}}) \\ &\quad + \frac{1}{2}(\mathbf{w}-\hat{\mathbf{w}})^{\mathrm{T}}\nabla^2\,\mathcal{L}(\mathcal{D};\mathbf{w})\big|_{\mathbf{w}=\hat{\mathbf{w}}}(\mathbf{w}-\hat{\mathbf{w}}) \\ &= \mathcal{L}(\mathcal{D};\hat{\mathbf{w}}) + \frac{1}{2}(\mathbf{w}-\hat{\mathbf{w}})^{\mathrm{T}}\nabla^2\,\mathcal{L}(\mathcal{D};\mathbf{w})\big|_{\mathbf{w}=\hat{\mathbf{w}}}(\mathbf{w}-\hat{\mathbf{w}})\end{aligned}$$

Laplace Approximation

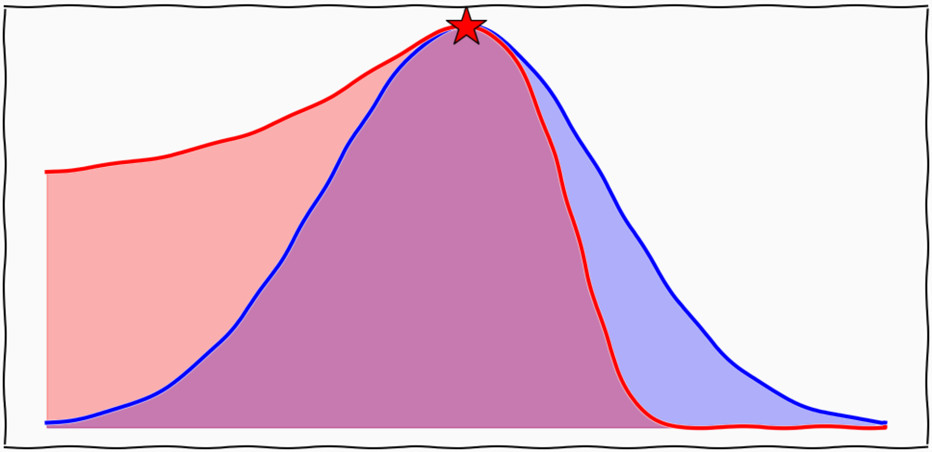
$$p(\mathbf{w} \mid \mathcal{D}) = \frac{1}{Z} p(\mathcal{D} \mid \mathbf{w}) p(\mathbf{w}) = \frac{1}{Z} \exp(-\mathcal{L}(\mathcal{D}; \mathbf{w}))$$

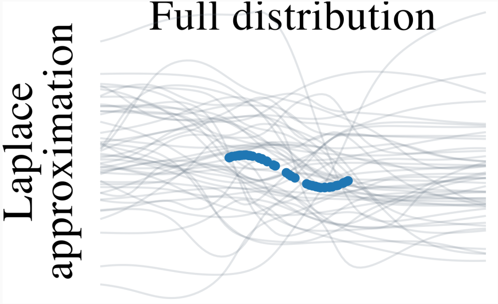
Laplace Approximation

$$
\begin{aligned}
p(\mathbf{w} \mid \mathcal{D}) &= \frac{1}{Z} p(\mathcal{D} \mid \mathbf{w}) p(\mathbf{w}) = \frac{1}{Z} \exp(-\mathcal{L}(\mathcal{D}; \mathbf{w})) \\
&= \frac{1}{Z} \exp(-\mathcal{L}(\mathcal{D}; \hat{\mathbf{w}})) \exp\left(\frac{1}{2} (\mathbf{w} - \hat{\mathbf{w}})^{\mathrm{T}} \nabla^2 \left.\mathcal{L}(\mathcal{D}; \mathbf{w})\right|_{\mathbf{w}=\hat{\mathbf{w}}} (\mathbf{w} - \hat{\mathbf{w}})\right)
\end{aligned}
$$

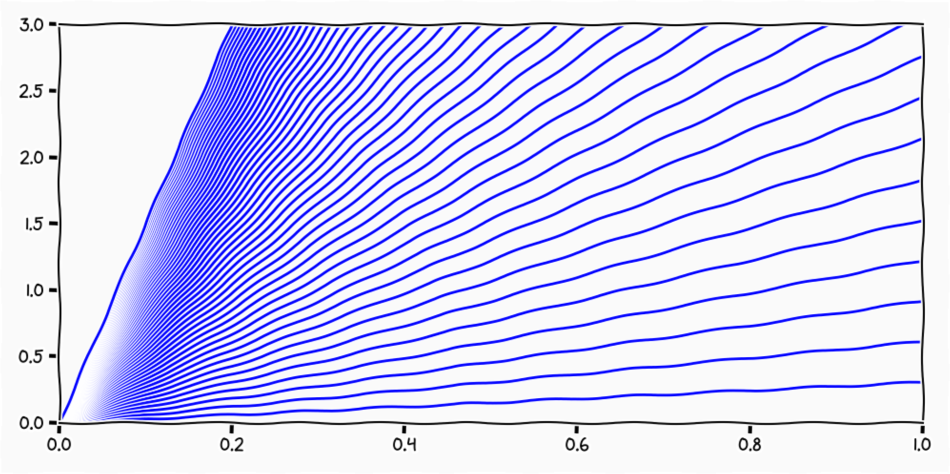
Laplace Approximation

$$
\begin{aligned}
p(\mathbf{w} \mid \mathcal{D}) &= \frac{1}{Z} p(\mathcal{D} \mid \mathbf{w}) p(\mathbf{w}) = \frac{1}{Z} \exp(-\mathcal{L}(\mathcal{D}; \mathbf{w})) \\
&= \frac{1}{Z} \exp(-\mathcal{L}(\mathcal{D}; \hat{\mathbf{w}})) \exp\left(\frac{1}{2} (\mathbf{w} - \hat{\mathbf{w}})^{\mathrm{T}} \nabla^2 \left.\mathcal{L}(\mathcal{D}; \mathbf{w})\right|_{\mathbf{w}=\hat{\mathbf{w}}} (\mathbf{w} - \hat{\mathbf{w}}) \right) \\
&= \mathcal{N}\left(\mathbf{w} \mid \hat{\mathbf{w}}, \left(\nabla^2 \left.\mathcal{L}(\mathcal{D}; \mathbf{w})\right|_{\mathbf{w}=\hat{\mathbf{w}}} \right)^{-1} \right)
\end{aligned}
$$

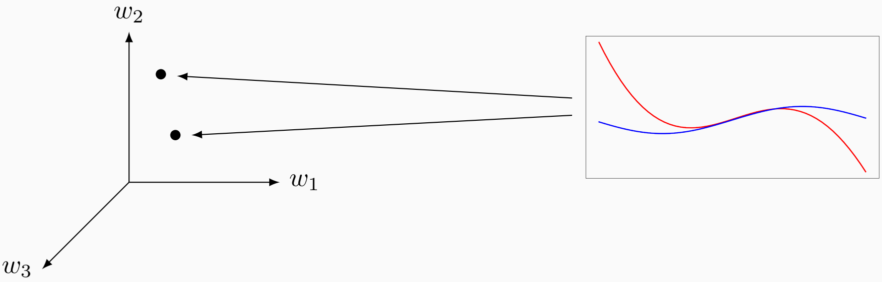




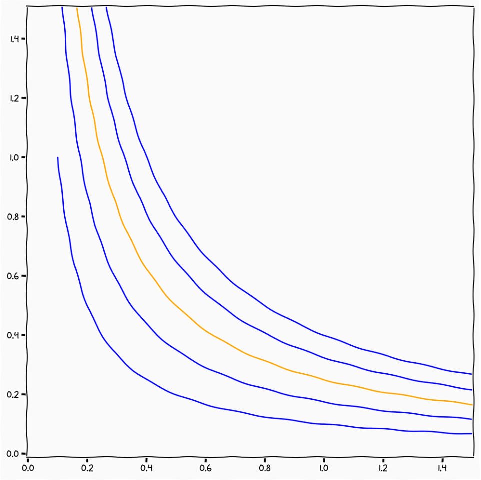
Full distribution
Laplace approximation



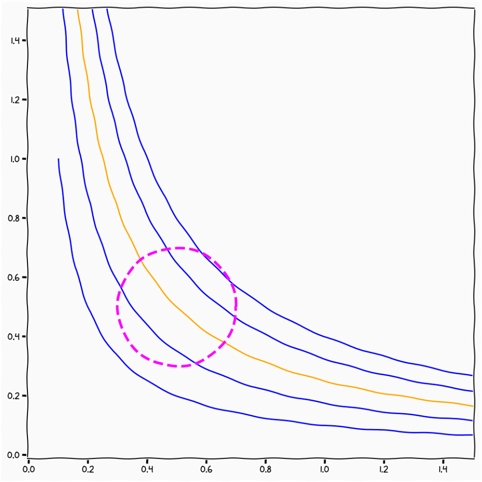
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.0
0.2
0.4
0.6
0.8
1.0



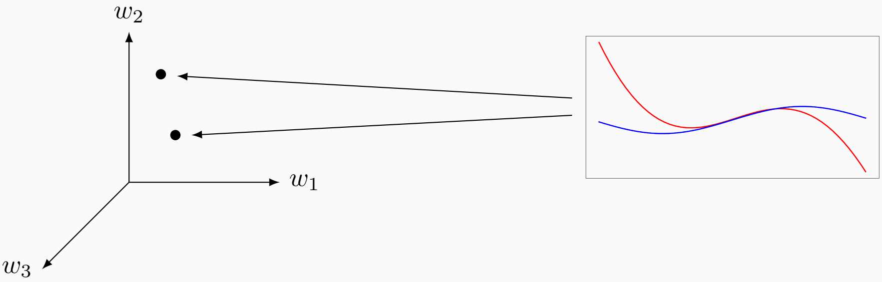
w_2
w_1
w_3



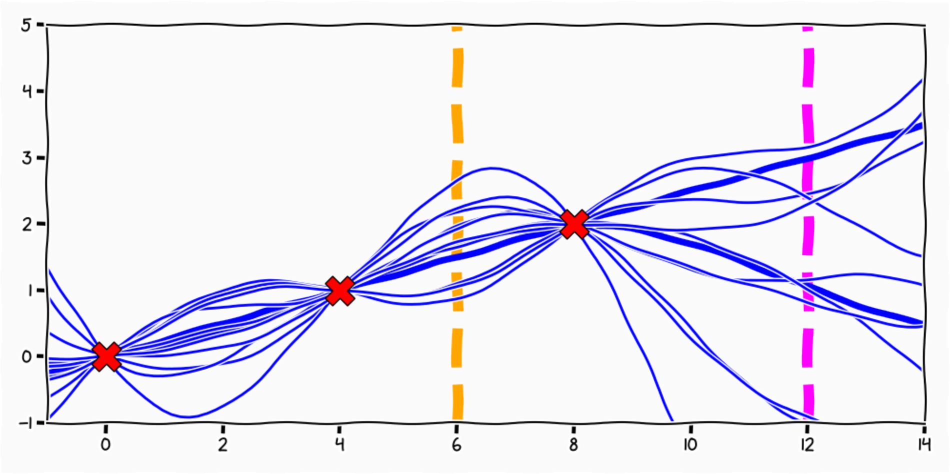
$$f(x) = w_1 \cdot w_2 \cdot x = (\alpha \cdot w_1) \cdot \left(\frac{1}{\alpha} \cdot w_2\right) \cdot x$$



$$f(x) = w_1 \cdot w_2 \cdot x = (\alpha \cdot w_1) \cdot \left(\frac{1}{\alpha} \cdot w_2\right) \cdot x$$



w_2
w_1
w_3


5
4
3
2
1
0
−1
0
2
4
6
8
10
12
14

Definition (x-reparametrisations)

Given a datapoint $\mathbf{x} \in \mathbb{R}^I$, for any $\mathbf{w} \in \mathbb{R}^D$ we define the $\mathbf{x}$-*reparameterizations* as the set $\mathcal{R}^f_{\mathbf{x}}(\mathbf{w}) = \{\mathbf{w}' \text{ such that } f(\mathbf{w}', \mathbf{x}) = f(\mathbf{w}, \mathbf{x})\}$. Consistently, given a collection of points $\mathcal{X} \subseteq \mathbb{R}^I$, we call the intersection $\mathcal{R}^f_{\mathcal{X}}(\mathbf{w}) = \bigcap_{\mathbf{x}\in\mathcal{X}} \mathcal{R}^f_{\mathbf{x}}(\mathbf{w})$ $\mathcal{X}$-*reparameterizations*.

The effective parameter quotient group

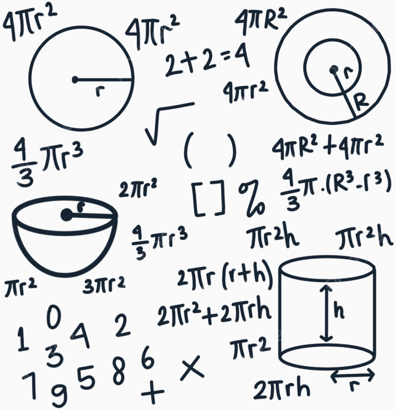
- We define the relation $\sim$ over $\mathbb{R}^D$ as $\mathbf{w} \sim \mathbf{w}'$ if $\mathbf{w}' \in \bar{\mathcal{R}}^f_{\mathcal{X}}(\mathbf{w})$.

The effective parameter quotient group

- We define the relation $\sim$ over $\mathbb{R}^D$ as $\mathbf{w} \sim \mathbf{w}'$ if $\mathbf{w}' \in \bar{\mathcal{R}}^f_{\mathcal{X}}(\mathbf{w})$.
- Quotient space of effective parameters $\mathcal{P} = \mathbb{R}^D / \sim$

The effective parameter quotient group

- We define the relation $\sim$ over $\mathbb{R}^D$ as $\mathbf{w} {\sim} \mathbf{w}'$ if $\mathbf{w}' \in \bar{\mathcal{R}}^f_{\mathcal{X}}(\mathbf{w})$.
- Quotient space of effective parameters $\mathcal{P} = \mathbb{R}^D / \sim$
- $[\mathbf{w}_1], [\mathbf{w}_2] \in \mathcal{P}$ are the same point if and only if $\mathbf{w}_1 \sim \mathbf{w}_2$.


4πr²
4πr²
4πR²
2+2=4
r
4πr²
r
R
4/3 πr³
( )
4πR²+4πr²
2πr²
[ ] %
4/3 π.(R³-r³)
4/3 πr³
πr²h
πr²h
2πr (r+h)
πr²
3πr²
h
2πr²+2πrh
1 0 3 4 2
πr²
7 9 5 8 6 + ×
2πrh
r

$$\mathrm{dist}(\mathbf{w}_1, \mathbf{w}_2) = 0 \leftrightarrow \mathbf{w}_1 \sim \mathbf{w}_2$$

$$\text{dist}(\mathbf{w}_1, \mathbf{w}_2) = 0 \leftrightarrow \mathbf{w}_1 \sim \mathbf{w}_2$$

$$\text{dist}^2(\mathbf{w}, \mathbf{w} + \boldsymbol{\epsilon}) = \sum_{n=1}^{N} ||f(\mathbf{w}, \mathbf{x}_n) - f(\mathbf{w} + \boldsymbol{\epsilon}, \mathbf{x}_n)||^2$$

$$\begin{aligned}\operatorname{dist}(\mathbf{w}_1, \mathbf{w}_2) &= 0 \leftrightarrow \mathbf{w}_1 \sim \mathbf{w}_2 \\ \operatorname{dist}^2(\mathbf{w}, \mathbf{w} + \boldsymbol{\epsilon}) &= \sum_{n=1}^{N} ||f(\mathbf{w}, \mathbf{x}_n) - f(\mathbf{w} + \boldsymbol{\epsilon}, \mathbf{x}_n)||^2 \\ &= \sum_{n=1}^{N} ||f(\mathbf{w}, \mathbf{x}_n) - f(\mathbf{w}, \mathbf{x}_n) - \nabla f(\mathbf{w}, \mathbf{x}_n)\boldsymbol{\epsilon}||^2\end{aligned}$$

$$
\begin{aligned}
\mathrm{dist}(\mathbf{w}_1, \mathbf{w}_2) &= 0 \leftrightarrow \mathbf{w}_1 \sim \mathbf{w}_2 \\
\mathrm{dist}^2(\mathbf{w}, \mathbf{w} + \boldsymbol{\epsilon}) &= \sum_{n=1}^{N} ||f(\mathbf{w}, \mathbf{x}_n) - f(\mathbf{w} + \boldsymbol{\epsilon}, \mathbf{x}_n)||^2 \\
&= \sum_{n=1}^{N} ||f(\mathbf{w}, \mathbf{x}_n) - f(\mathbf{w}, \mathbf{x}_n) - \nabla f(\mathbf{w}, \mathbf{x}_n)\boldsymbol{\epsilon}||^2 \\
&= \boldsymbol{\epsilon}^{\mathrm{T}} \mathbf{J}^{\mathrm{T}} \mathbf{J} \boldsymbol{\epsilon}
\end{aligned}
$$

Geometric Aware Laplace Approximation

$$q(\mathbf{w}) = \mathcal{N}\left(\mathbf{w} \mid \hat{\mathbf{w}}, \left(\underbrace{\mathbf{J}^{\mathrm{T}} \nabla^2 \mathcal{L}(\mathcal{D}; \mathbf{w})\big|_{\mathbf{w}=\hat{\mathbf{w}}} \mathbf{J}}_{\mathbf{GGN}_{\mathbf{w}}}\right)^{-1}\right)$$

- This is also known as the Generalised Gauss Newton Approximation

Geometric Aware Laplace Approximation

$$q(\mathbf{w}) = \mathcal{N}\left(\mathbf{w} \mid \hat{\mathbf{w}}, \left(\underbrace{\mathbf{J}^{\mathrm{T}} \nabla^2 \mathcal{L}(\mathcal{D}; \mathbf{w})\big|_{\mathbf{w}=\hat{\mathbf{w}}} \mathbf{J}}_{\mathbf{GGN}_{\mathbf{w}}}\right)^{-1}\right)$$

- This is also known as the Generalised Gauss Newton Approximation
- Linearised Laplace Approximation Immer et al., 2021

Geometric Aware Laplace Approximation

$$q(\mathbf{w}) = \mathcal{N}\left(\mathbf{w} \mid \hat{\mathbf{w}}, \left(\underbrace{\mathbf{J}^{\mathrm{T}} \nabla^2 \mathcal{L}(\mathcal{D}; \mathbf{w})\big|_{\mathbf{w}=\hat{\mathbf{w}}} \mathbf{J}}_{\mathbf{GGN_w}}\right)^{-1}\right)$$

- This is also known as the Generalised Gauss Newton Approximation
- Linearised Laplace Approximation Immer et al., 2021
- Interpreted as a Riemannian metric it is called the Fisher-Rao metric

Geometric Aware Laplace Approximation

$$q(\mathbf{w}) = \mathcal{N}\left(\mathbf{w} \mid \hat{\mathbf{w}}, \left(\underbrace{\mathbf{J}^{\mathrm{T}} \nabla^2 \mathcal{L}(\mathcal{D}; \mathbf{w})\big|_{\mathbf{w}=\hat{\mathbf{w}}} \mathbf{J}}_{\mathbf{GGN}_{\mathbf{w}}}\right)^{-1}\right)$$

- This is also known as the Generalised Gauss Newton Approximation
- Linearised Laplace Approximation Immer et al., 2021
- Interpreted as a Riemannian metric it is called the Fisher-Rao metric
- It is a *pseudo-metric*

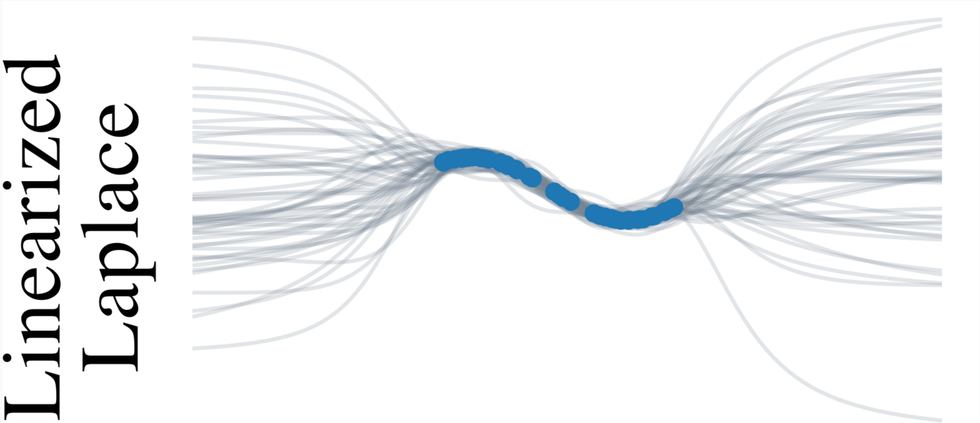
Theorem (Topological Equivivalence)

The pseudo-Riemannian manifold obtained with the pullback pseudo-metric $(\mathbb{R}^D, GGN_{\mathbf{w}})$ *is homeomorphic to the quotient group* $(\mathcal{P}, d_{\mathcal{P}})$

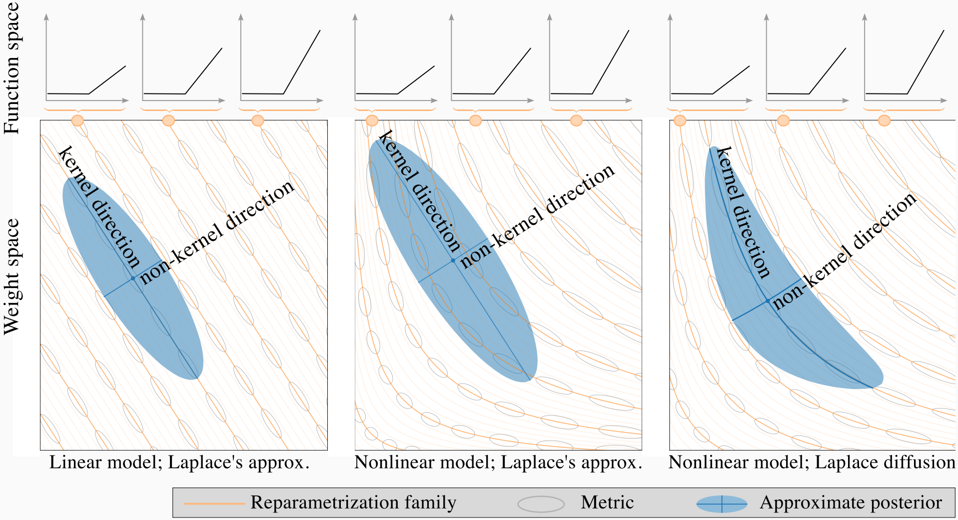
For any $\mathbf{w}_0, \mathbf{w}_1 \in \mathbf{R}^D$ it holds

$$d_{f^*H}(\mathbf{w}_0, \mathbf{w}_1) = 0 \quad \Longleftrightarrow \quad [\mathbf{w}_0] = [\mathbf{w}_1] \in \mathcal{P}$$

Results



Intuition



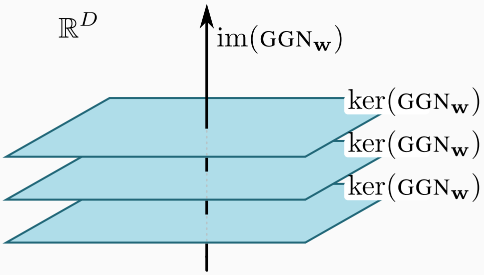
Function space
Weight space
kernel direction
non-kernel direction
kernel direction
non-kernel direction
kernel direction
non-kernel direction
Linear model; Laplace's approx.
Nonlinear model; Laplace's approx.
Nonlinear model; Laplace diffusion
Reparametrization family
Metric
Approximate posterior

- Linear Function

$$f(\mathbf{w}) = \mathbf{A}\mathbf{w} + b$$
$$g : \mathbb{R}^D \to \mathbb{R}^D$$
$$\text{s.t } \mathbf{A}(g(\mathbf{w}) - \mathbf{w}) = \mathbf{0}$$

- Nullspace of $\mathbf{A}$

$$(g(\mathbf{w}) - \mathbf{w}) \in \ker(\mathbf{A})$$



$$f_w(x) \approx f_{\hat{w}} + \mathbf{J}_w(x)(w - \hat{w})$$

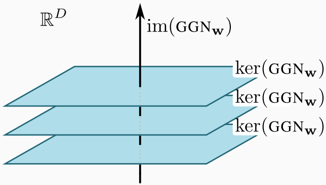
- re-parametrisations are characterised by the $\text{kern}(\mathbf{J_w})$
- By construction

$$\text{kern}(\mathbf{J_w}) = \text{kern}(\mathbf{J_w^T J_w})$$

- Neural Tangent Kernel Jacot et al., 2018

$$\text{NTK} = \mathbf{J_w J_w^T}$$

Orthogonal Subspaces



$$\mathrm{im}(\mathrm{GGN}_{\mathbf{w}}) \oplus \mathrm{kern}(\mathrm{GGN}_{\mathbf{w}}) = \mathbb{R}^D$$

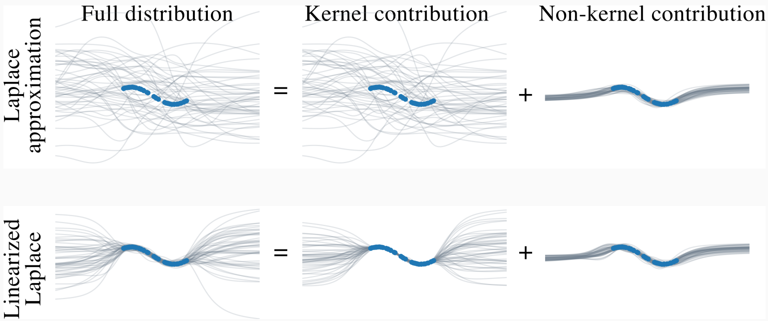
- $\mathrm{im}(\mathrm{GGN}_{\mathbf{w}})$ - spans the *effective* parameters of the model
- $\mathrm{kern}(\mathrm{GGN}_{\mathbf{w}})$ - parameters leading to the same function

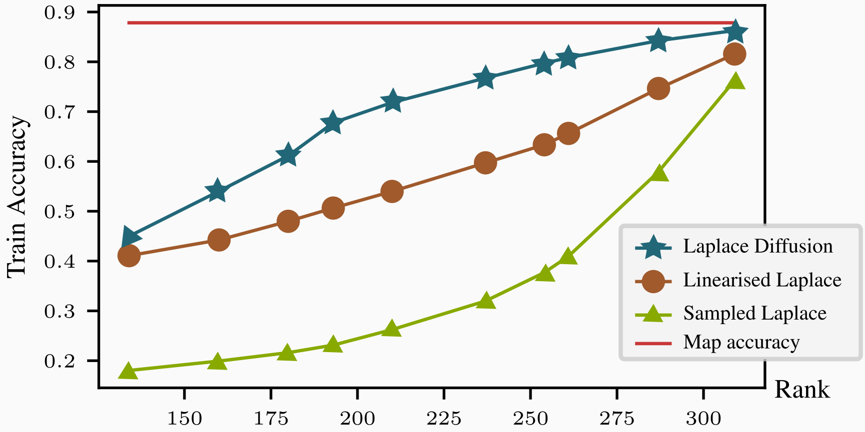
Laplace Covariance

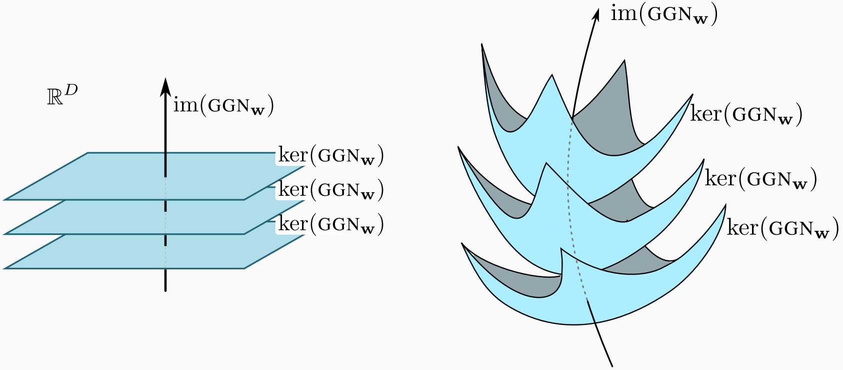
$$\Sigma = \left(\left[\begin{array}{c} U_1 \\ \hline U_2 \end{array} \right]^T \left[\begin{array}{c|c} \tilde{\Lambda} & 0 \\ \hline 0 & 0 \end{array} \right] \left[\begin{array}{c} U_1 \\ \hline U_2 \end{array} \right] + \alpha I \right)^{-1} = U_1^{\mathrm{T}} (\tilde{\Lambda} + \alpha I_k)^{-1} U_1 + \alpha^{-1} U_2^{\mathrm{T}} U_2.$$

- Decomposition of parameter space

$$\mathbf{w} = \hat{\mathbf{w}} + \mathbf{w}_{\mathrm{ker}} + \mathbf{w}_{\mathrm{im}}$$


Full distribution
Kernel contribution
Non-kernel contribution
Laplace approximation
Linearized Laplace
=
+
=
+


Train Accuracy
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
150
175
200
225
250
275
300
Rank
Laplace Diffusion
Linearised Laplace
Sampled Laplace
Map accuracy



$\mathbb{R}^D$
im(GGN$_\mathbf{w}$)
ker(GGN$_\mathbf{w}$)
ker(GGN$_\mathbf{w}$)
ker(GGN$_\mathbf{w}$)
im(GGN$_\mathbf{w}$)
ker(GGN$_\mathbf{w}$)
ker(GGN$_\mathbf{w}$)
ker(GGN$_\mathbf{w}$)

Diffusion process

- Riemannian Diffusion

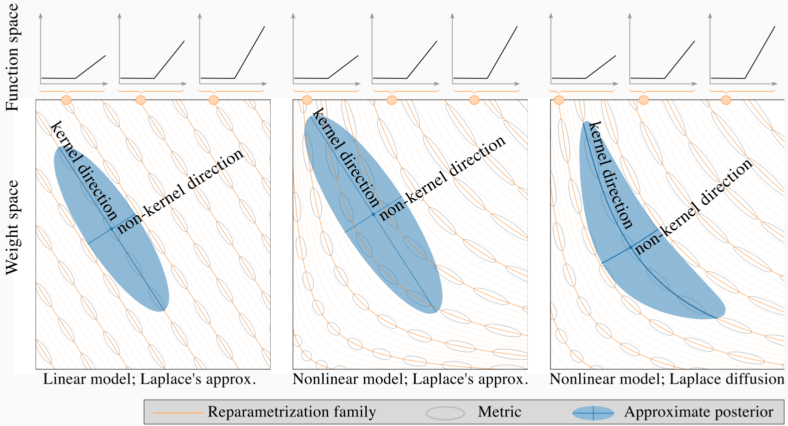
$$\mathbf{w} = \sqrt{2\tau}\,G(\mathbf{w})^{-\frac{1}{2}}\,W + \tau\Gamma t$$

$$\text{where} \quad \Gamma_i(\mathbf{w}) = \sum_{j=1}^{D} \frac{\partial}{\partial \mathbf{w}_j}(G(\mathbf{w})^{-1})_{ij}.$$

- Update rule

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \sqrt{2h_t}\,G(\mathbf{w}_t)^{-\frac{1}{2}}\epsilon$$

Intuition



Function space
Weight space
kernel direction
non-kernel direction
kernel direction
non-kernel direction
kernel direction
non-kernel direction
Linear model; Laplace's approx.
Nonlinear model; Laplace's approx.
Nonlinear model; Laplace diffusion
Reparametrization family
Metric
Approximate posterior

Diffusion Process

- Diffusion in Linearised Laplace

$$\left(\mathbb{R}^D, \mathbf{GGN}_{\hat{\mathbf{w}}} + \alpha \mathbf{I}\right)$$

- Diffusion in Linearised Laplace

$$\left(\mathbb{R}^D, \mathbf{GGN}_{\hat{\mathbf{w}}} + \alpha \mathbf{I}\right)$$

- Kernel Manifold

$$\left(\mathcal{P}_{\mathbf{w}}^{\perp}, \alpha \mathbf{I}\right)$$

Diffusion Process

- Diffusion in Linearised Laplace

$$\left(\mathbb{R}^D, \mathbf{GGN}_{\hat{\mathbf{w}}} + \alpha \mathbf{I}\right)$$

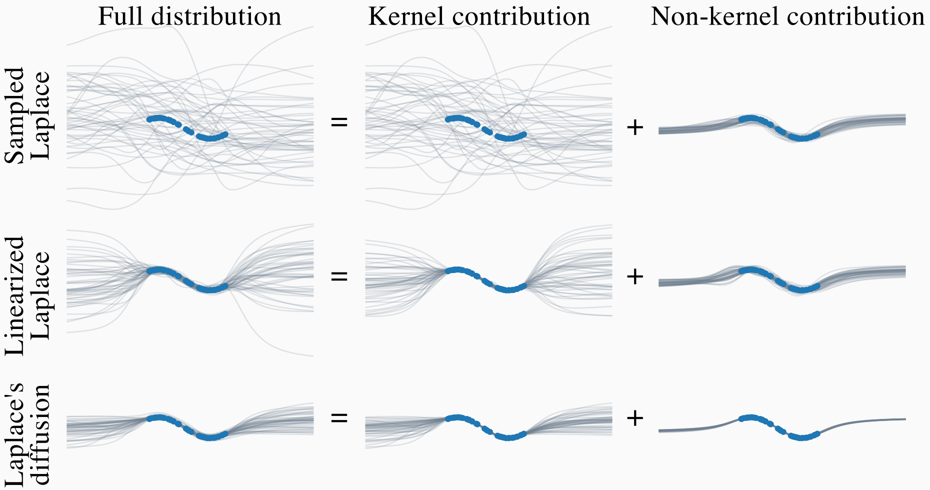
- Kernel Manifold

$$\left(\mathcal{P}_{\mathbf{w}}^{\perp}, \alpha \mathbf{I}\right)$$

- Non-kernel Manifold

$$\left(\mathcal{P}_{\mathbf{w}}, \mathbf{GGN}_{\mathbf{w}}^{\not\perp}\right)$$

Results


Full distribution
Kernel contribution
Non-kernel contribution
Sampled Laplace
Linearized Laplace
Laplace's diffusion
=
+
=
+
=
+

Taking Stock

- we have characterised the geometry of reparametrisations

Taking Stock

- we have characterised the geometry of reparametrisations
- the geometry explains why linearised laplace approximation works

Taking Stock

- we have characterised the geometry of reparametrisations
- the geometry explains why linearised laplace approximation works
- we have derived a simple random walk to sample from manifold

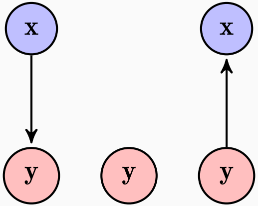
$$p(y) = \int p(y \mid \theta) p(\theta) \mathrm{d}\theta$$

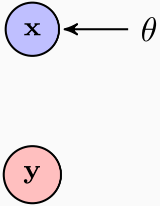
$$
\begin{aligned}
p(y) &= \int p(y \mid \theta) p(\theta) \mathrm{d}\theta \\
&= \text{``}\tilde{p}(y_{\perp}) + \tilde{p}(y_{\not\perp})\text{''}
\end{aligned}
$$

Variational Inference

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

Variational Inference





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

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

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

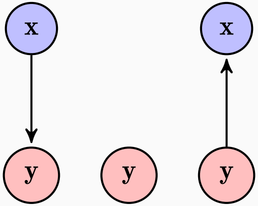
$$=$$

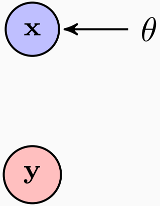
$$\log p(y) = \int q(x) \log \frac{p(x, y)}{p(x \mid y)} \mathrm{d}x$$

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

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

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





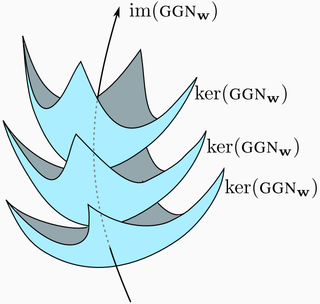
$$p(y) = \int_x p(y|x)p(x) = \frac{p(y|x)p(x)}{p(x|y)}$$

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

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



$$\mathbf{J}^{\mathrm{T}} \nabla^2 \mathcal{L}(\mathcal{D}; \mathbf{w})\big|_{\mathbf{w}=\hat{\mathbf{w}}} \mathbf{J}$$

- Can we propose an approximate distribution that reflects the geometry of the parametrisation?

Gemetrically Aware Variational Distribution

$$q(\theta) = \mathcal{N}(\theta|\hat{\theta}, \Sigma)$$
$$\Sigma = \sigma_{\mathrm{ker}}^2 \mathbf{U}\mathbf{U}^{\mathrm{T}} + \sigma^2(\mathbb{I} - \mathbf{U}\mathbf{U}^{\mathrm{T}})$$

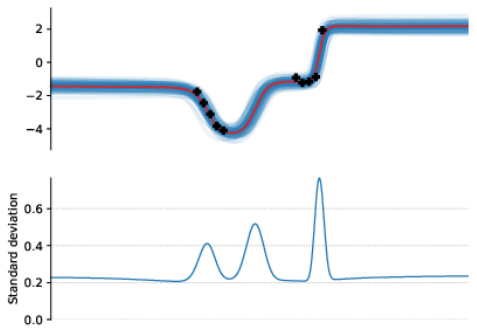
Lower Bound

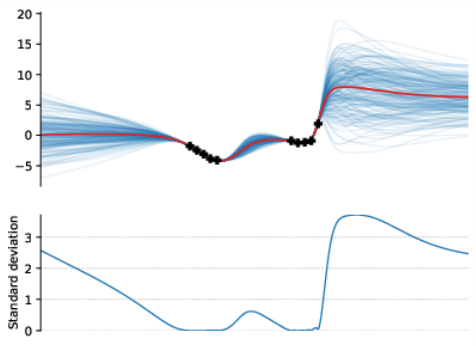
$$\mathcal{L} = \mathbb{E}_{\theta \sim q}\left[\log p(\mathbf{y}|\theta, \mathbf{x})\right] - \mathrm{KL}(q(\theta) \| p(\theta))$$

Computing the Lower Bound using Stochastic Projections

$$\begin{aligned}
& \epsilon^{(s)} \sim \mathcal{N}(0, \mathbb{I}) && \in \mathbb{R}^D, \\
\text{(projection onto kernel space)} \quad & \epsilon_{\text{ker}}^{(s)} = \mathbf{U}\mathbf{U}^{\text{T}}\epsilon^{(s)} && \in \mathbb{R}^D, \\
\text{(image space is orthogonal)} \quad & \epsilon_{\text{im}}^{(s)} = (\mathbb{I} -)\epsilon^{(s)} = \epsilon^{(s)} - \epsilon_{\text{ker}}^{(s)} && \in \mathbb{R}^D, \\
& \theta^{(s)} = \hat{\theta} + \sigma_{\text{ker}}\epsilon_{\text{ker}}^{(s)} + \sigma_{\text{im}}\epsilon^{(s)} && \in \mathbb{R}^D, \\
\mathbb{E}_{\theta \sim q}\left[\log p(|\theta, \mathbf{x})\right] \approx & \frac{1}{S}\sum_{s=1}^{S} \log p(|\theta^{(s)}, \mathbf{x}) && \in \mathbb{R} \quad .
\end{aligned}$$

Results


2
0
−2
−4
Standard deviation
0.6
0.4
0.2
0.0


20
15
10
5
0
−5
Standard deviation
3
2
1
0

Summary

Summary

- Symmetries are great learning in ”deterministic models”

Summary

- Symmetries are great learning in ”deterministic models”
- Symmetries are very problematic for statistical models

Summary

- Symmetries are great learning in ”deterministic models”
- Symmetries are very problematic for statistical models
- ”Under parametrised” approximate posteriors leads to pathological measures

Summary

- The Laplace approximation severely underfits because it does not reflect the re-parametrisations of functions

Summary

- The Laplace approximation severely underfits because it does not reflect the re-parametrisations of functions
- The Linearised Laplace approximation is infinitesimally invariant to re-parametrisations

Summary

- The Laplace approximation severely underfits because it does not reflect the re-parametrisations of functions
- The Linearised Laplace approximation is infinitesimally invariant to re-parametrisations
- The covariance of the Linearised Laplace Approximation defines a Riemannian metric on the Manifold of effective parameters

Summary

- The factorisation tells what degrees of freedom are connected to data and to the prior

Summary

- The factorisation tells what degrees of freedom are connected to data and to the prior
- What parametrisations are useful for algorithms?

Summary

- The factorisation tells what degrees of freedom are connected to data and to the prior
- What parametrisations are useful for algorithms?
- We can formulate approximate distributions that reflect the geometry of the parametrisation

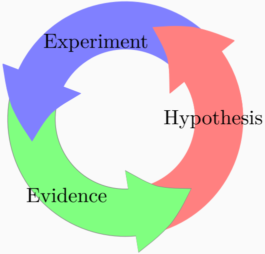
Summary

- The factorisation tells what degrees of freedom are connected to data and to the prior
- What parametrisations are useful for algorithms?
- We can formulate approximate distributions that reflect the geometry of the parametrisation
- Factorisation of measures

Summary

- The factorisation tells what degrees of freedom are connected to data and to the prior
- What parametrisations are useful for algorithms?
- We can formulate approximate distributions that reflect the geometry of the parametrisation
- Factorisation of measures
- Matrix free algebra allows us to approximate parameter spaces with millions of parameters

Summary



$$p(w) = \mathcal{N}(0, \alpha\mathbf{I})$$

eof

References

- Immer, Alexander, Maciej Korzepa, and Matthias Bauer (2021). **"Improving predictions of Bayesian neural nets via local linearization".** In: *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 703–711.
- Jacot, Arthur, Franck Gabriel, and Clément Hongler (2018). **"Neural Tangent Kernel: Convergence and Generalization in Neural Networks".** In: *CoRR*.
- MacKay, D. J. C. (1991). **"Bayesian Methods for Adaptive Models".** PhD thesis. California Institute of Technology.

Ustyuzhaninov, Ivan, Ieva Kazlauskaite, Markus Kaiser, Erik Bodin, Neill D. F. Campbell, and Carl Henrik Ek (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.