Variational inference for deep Gaussian processes

Andreas Damianou

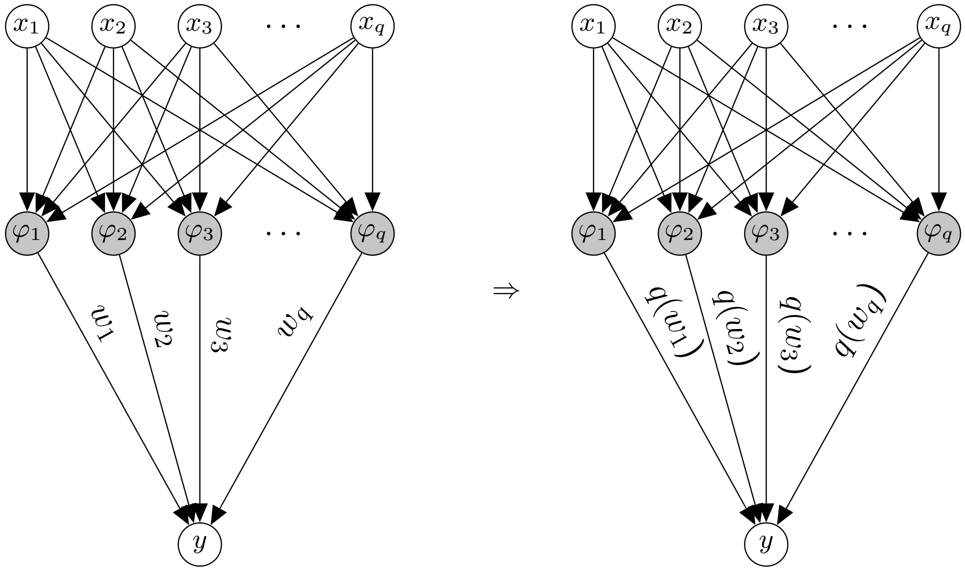
damianou@amazon.com

Amazon.com, Cambridge, UK

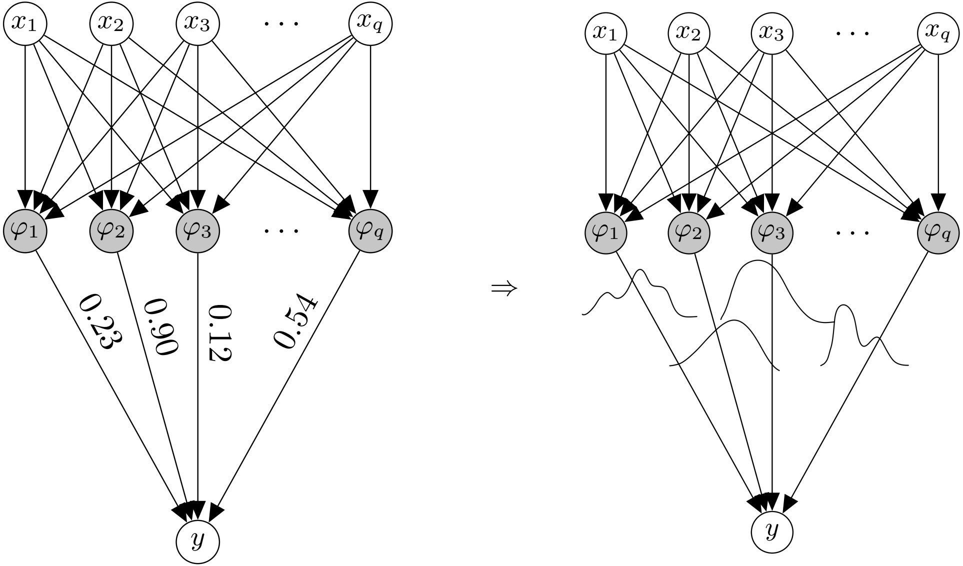
NIPS workshop on Advances in Approximate Bayesian Inference,
December 2017

amazon.com

Bayesian Neural Network



Bayesian Neural Network

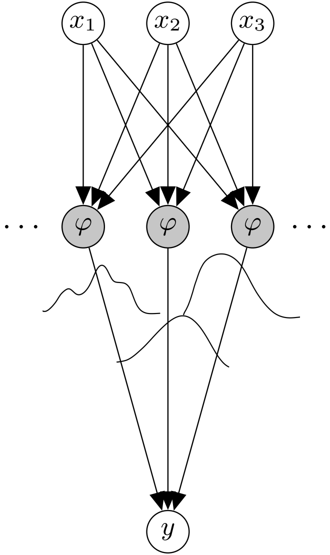


From NN to GP

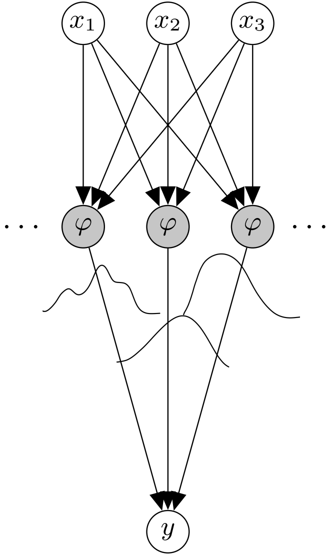
- In the limit of infinite units we obtain a GP*.
- Think of a function as an infinite dimensional vector.

$f \sim \mathcal{GP}(0, k(x, x'))$. f is *stochastic*!

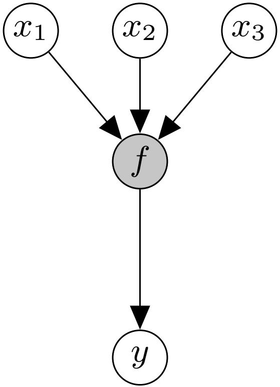
** Radford M Neal. Bayesian learning for neural networks. PhD thesis, 1995.*



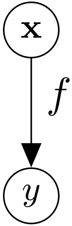
From NN to GP



From NN to DGP



From NN to GP



DeepGP

- Define a recursive stacked construction

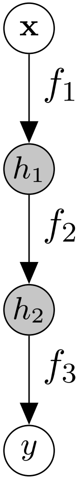
$$f(\mathbf{x}) \rightarrow \text{GP}$$

$$f_L(f_{L-1}(f_{L-2} \cdots f_1(\mathbf{x})))) \rightarrow \text{deep GP}$$

Compare to:

$$\varphi(\mathbf{x})^\top \mathbf{w} \rightarrow \text{NN}$$

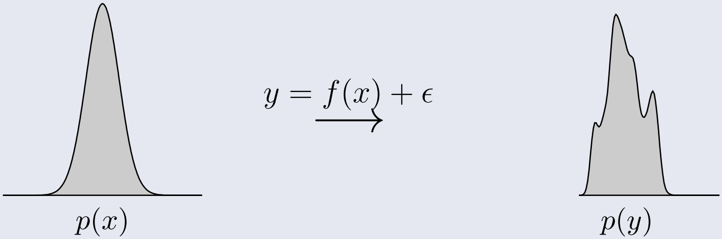
$$\varphi(\varphi(\varphi(\mathbf{x})^\top \mathbf{w}_1)^\top \ldots \mathbf{w}_{L-1})^\top \mathbf{w}_L \rightarrow \text{DNN}$$



Damianou & Lawrence, AISTATS 2013

Recap

Propagating uncertainty through non-linearities:



VI is challenging with propagation of uncertainty.

Direct marginalisation of h is intractable

- Objective: $p(y|x) = \int_{h_2} \left(p(y|h_2) \int_{h_1} p(h_2|h_1) p(h_1|x) \right)$

Direct marginalisation of h is intractable

- Objective: $p(y|x) = \int_{h_2} \left(p(y|h_2) \int_{h_1} p(h_2|h_1) p(h_1|x) \right)$

- $p(h_2|x) \quad = \int_{h_1, f_2} \; p(h_2|f_2) p(f_2|h_1) \; p(h_1|x)$

Direct marginalisation of h is intractable

- Objective: $p(y|x) = \int_{h_2} \left(p(y|h_2) \int_{h_1} p(h_2|h_1) p(h_1|x) \right)$
- $p(h_2|x) \quad = \int_{h_1, f_2} \quad p(h_2|f_2) p(f_2|h_1) \quad p(h_1|x)$

Direct marginalisation of h is intractable

- Objective: $p(y|x) = \int_{h_2} \left(p(y|h_2) \int_{h_1} p(h_2|h_1) p(h_1|x) \right)$
- $p(h_2|x) \quad = \int_{h_1, f_2} p(h_2|f_2) \underbrace{p(f_2|h_1)}_{\substack{\text{contains} \\ (k(h_1,h_1))^{-1}}} p(h_1|x)$

Direct marginalisation of h is intractable

- Objective: $p(y|x) = \int_{h_2} \left(p(y|h_2) \int_{h_1} p(h_2|h_1) p(h_1|x) \right)$
- $p(h_2|x) = \int_{h_1, f_2} p(h_2|f_2) p(f_2|h_1) p(h_1|x)$
- $p(h_2|x) = \int_{h_1, f_2, u_2} p(h_2|f_2) p(f_2|u_2, h_1) p(u_2) p(h_1|x)$

Direct marginalisation of h is intractable

- Objective: $p(y|x) = \int_{h_2} \left(p(y|h_2) \int_{h_1} p(h_2|h_1) p(h_1|x) \right)$
- $p(h_2|x) \quad = \int_{h_1, f_2} \quad p(h_2|f_2) p(f_2|h_1) \quad p(h_1|x)$
- $p(h_2|x) = \int_{h_1, f_2, u_2} p(h_2|f_2) p(f_2|u_2, h_1) p(u_2) p(h_1|x)$
- $\log p(h_2|x) \geq \int_{h_1, f_2, u_2} \mathcal{Q} \log \frac{p(h_2|f_2) p(f_2|u_2, h_1) p(u_2) p(h_1|x)}{\mathcal{Q}}$

Direct marginalisation of h is intractable

- Objective: $p(y|x) = \int_{h_2} \left(p(y|h_2) \int_{h_1} p(h_2|h_1) p(h_1|x) \right)$
- $p(h_2|x) \quad = \int_{h_1, f_2} \; p(h_2|f_2) p(f_2|h_1) \; p(h_1|x)$
- $p(h_2|x) = \int_{h_1, f_2, u_2} p(h_2|f_2) p(f_2|u_2, h_1) p(u_2) p(h_1|x)$
- $\log p(h_2|x) \geq \int_{h_1, f_2, u_2} \mathcal{Q} \log \frac{p(h_2|f_2) \cancel{p(f_2|u_2, h_1)} p(u_2) p(h_1|x)}{\mathcal{Q} = \cancel{p(f_2|u_2, h_1)} q(u_2) q(h_1)}$

Direct marginalisation of h is intractable

- Objective: $p(y|x) = \int_{h_2} \left(p(y|h_2) \int_{h_1} p(h_2|h_1) p(h_1|x) \right)$
- $p(h_2|x) \quad = \int_{h_1,f_2} \quad p(h_2|f_2) p(f_2|h_1) \quad p(h_1|x)$
- $p(h_2|x) = \int_{h_1,f_2,u_2} p(h_2|f_2) p(f_2|u_2, h_1) p(u_2) p(h_1|x)$
- $\log p(h_2|x) \geq \int_{h_1,f_2,u_2} \mathcal{Q} \log \frac{p(h_2|f_2) \cancel{p(f_2|u_2,h_1)} p(u_2) p(h_1|x)}{\mathcal{Q} = \cancel{p(f_2|u_2,h_1)} q(u_2) q(h_1)}$
- $\log p(h_2|x) \geq \int_{h_1,f_2,u_2} \mathcal{Q} \log \frac{p(h_2|f_2) p(u_2) p(h_1|x)}{q(u_2) q(h_1)}$

The information of f_2 was *compressed* in u_2, which is independent of h_1.

Direct marginalisation of h is intractable

- Objective: $p(y|x) = \int_{h_2} \left(p(y|h_2) \int_{h_1} p(h_2|h_1) p(h_1|x) \right)$
- $p(h_2|x) \quad = \int_{h_1, f_2} \quad p(h_2|f_2) p(f_2|h_1) \quad p(h_1|x)$
- $p(h_2|x) = \int_{h_1, f_2, u_2} p(h_2|f_2) p(f_2|u_2, h_1) p(u_2) p(h_1|x)$
- $\log p(h_2|x) \geq \int_{h_1, f_2, u_2} \mathcal{Q} \log \frac{p(h_2|f_2) \cancel{p(f_2|u_2, h_1)} p(u_2) p(h_1|x)}{\mathcal{Q} = \cancel{p(f_2|u_2, h_1)} q(u_2) q(h_1)}$
- $\log p(h_2|x) \geq \int_{h_1, f_2, u_2} \mathcal{Q} \log \frac{p(h_2|f_2) p(u_2) p(h_1|x)}{q(u_2) q(h_1)}$

The information of f_2 was *compressed* in u_2, which is independent of h_1.

Some extra work required for "linking" between layers: $q(h_l)$ involved in both layers l and $l+1$.

Recap

- Introduce auxiliary variables: $p(f|h) = \int_u p(f|u, h)p(u)$
- Exact posterior factor in mean-field: $\mathcal{Q} = p(f|u, h)q(u)q(h)$

[Titsias & Lawrence, AISTATS 2010]

[Damianou, Titsias & Lawrence, JMLR 2016]

Recap

- Introduce auxiliary variables: $p(f|h) = \int_{u} p(f|u, h)p(u)$
- Exact posterior factor in mean-field: $\mathcal{Q} = p(f|u, h)q(u)q(h)$

[Titsias & Lawrence, AISTATS 2010]

[Damianou, Titsias & Lawrence, JMLR 2016]

Properties of the bound (unsupervised case)

$$\mathcal{F} = \overbrace{\sum_{l=2}^{L+1} \left\langle \sum_{n=1}^{N} \mathcal{L}(\mathbf{h}_l^{(n)}, \mathbf{u}_l) \right\rangle_{\mathcal{Q}}}^{\text{Data fit}} - \sum_{l=2}^{L+1} \mathsf{KL}\left(q(\mathbf{u}_l) \,\|\, p(\mathbf{u}_l)\right) \underbrace{- \mathsf{KL}\left(q(\mathbf{h}_1) \,\|\, p(\mathbf{h}_1)\right)}_{\text{Regularization}} + \sum_{l=2}^{L} \underbrace{\mathcal{H}\left(q(\mathbf{h}_l)\right)}_{\text{Regularization}}$$

- All terms factorize w.r.t data points *[Hensman et al 2013]*.

Recap

Bound has novel properties: factorization & interpretability.

Properties of the bound (unsupervised case)

$$\mathcal{F} = \overbrace{\sum_{l=2}^{L+1} \left\langle \sum_{n=1}^{N} \mathcal{L}(\mathbf{h}_l^{(n)}, \mathbf{u}_l) \right\rangle_{\mathcal{Q}}}^{\text{Data fit}} - \sum_{l=2}^{L+1} \mathsf{KL}\left(q(\mathbf{u}_l) \parallel p(\mathbf{u}_l)\right) \underbrace{- \mathsf{KL}\left(q(\mathbf{h}_1) \parallel p(\mathbf{h}_1)\right)}_{\text{Regularization}} + \sum_{l=2}^{L} \underbrace{\mathcal{H}\left(q(\mathbf{h}_l)\right)}_{\text{Regularization}}$$

- All terms factorize w.r.t data points *[Hensman et al 2013]*

Properties of the bound (unsupervised case)

$$\mathcal{F} = \overbrace{\sum_{l=2}^{L+1} \left\langle \sum_{n=1}^{N} \mathcal{L}(\mathbf{h}_l^{(n)}, \mathbf{u}_l) \right\rangle_{\mathcal{Q}}}^{\text{Data fit}} - \sum_{l=2}^{L+1} \mathrm{KL}\left(q(\mathbf{u}_l) \parallel p(\mathbf{u}_l)\right) \underbrace{-\mathrm{KL}\left(q(\mathbf{h}_1) \parallel p(\mathbf{h}_1)\right)}_{\text{Regularization}} + \sum_{l=2}^{L} \underbrace{\mathcal{H}\left(q(\mathbf{h}_l)\right)}_{\text{Regularization}}$$

- All terms factorize w.r.t data points *[Hensman et al 2013]*
- We can additionally collapse $q(\mathbf{u})$

"Collapse" $q(\mathbf{u})$

- Collapsing $q(\mathbf{u})$ eliminates many variational parameters and makes bound "tighter" *(Titsias & Lawrence 2010)*

- $q(\mathbf{u}) = \mathcal{G}(q(\mathbf{h}))$

- But this introduces coupling and breaks the factorisation.

- We can still distribute the computations efficiently (e.g. by extending the work of [1, 2])

[1] Y. Gal, M. van der Wilk, C. E. Rasmussen, NIPS 2014

[2] Z. Dai, A. Damianou, J. Hensman, N. Lawrence, NIPS workshops, 2014

- We're left with $q(\mathbf{h}_l{}^{(n)}) \sim \mathcal{N}(\boldsymbol{\mu}_l^{(n)}, \mathbf{S}_l^{(n)})$
- Difficult to initialize and optimize all these parameters!

Amortized inference

Solution: Reparameterization through recognition model g:

$$\boldsymbol{\mu}_1^n = g_1(\mathbf{y}^{(n)})$$
$$\boldsymbol{\mu}_l^{(n)} = g_l(\boldsymbol{\mu}_{l-1}^{(n)})$$
$$g_l = \mathsf{MLP}(\boldsymbol{\theta}_l)$$

g_l deterministic $\Rightarrow \boldsymbol{\mu}_l^{(n)} = g_l(\dots g_1(\mathbf{y}^{(n)}))$

Dai, Damianou, Gonzalez, Lawrence, ICLR 2016

Structured VI for dynamical systems

Reparameterization through recurrent recognition model g:

$$\boldsymbol{\mu}_l^{(n)} = g_l(\boldsymbol{\mu}_{l-1}^{(n)}, \boldsymbol{\mu}_{l-1}^{(n-1)}, \cdots, \boldsymbol{\mu}_{l-1}^{(n-K)})$$

Mattos, Dai, Damianou, Forth, Barreto, Lawrence, ICLR 2016

The variational Gaussian approximation re-re-visited

- So far we considered $q(\mathbf{H}) = \prod_n \prod_d \mathcal{N}(h^{(n,d)} | \mu^{(n,d)}, s^{(n,d)})$

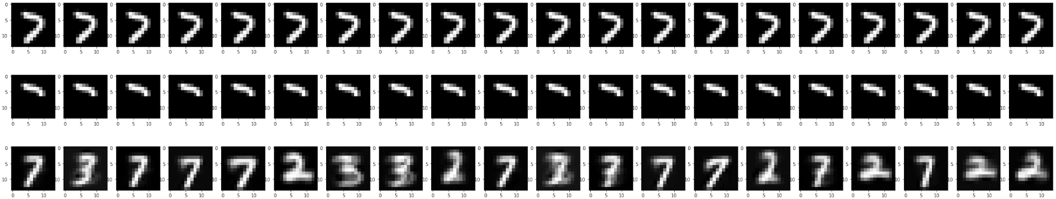
- To model correlations: $q(\mathbf{H}) = \prod_d \mathcal{N}(\mathbf{h}^{(:,d)} | \boldsymbol{\mu}^{(:,d)}, \boldsymbol{\Sigma}^{(:,d)})$

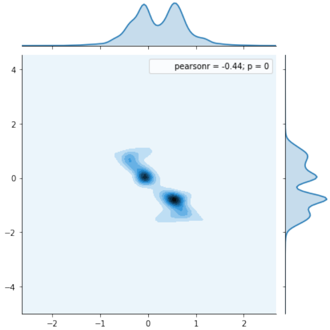
- Re-parameterization for GPs + Gaussian approximation:
$$\underbrace{\boldsymbol{\Sigma}^{(:,d)}}_{O(N^2)} = (\mathbf{K}^{-1} + \underbrace{\text{diag}(\boldsymbol{\lambda}^{(d)})}_{O(N)} \mathbf{I})^{-1}$$

Opper & Archambeau, Neural Computation, 2009

Damianou et al., NIPS, 2011

Normalizing flows for GP-LVMs





- $q(h)$ is rendered more expressive by being composed as a series of invertible transforms on a simpler density $q_0(h)$

  (Rezende and Mohamed, 2015), (Louizos and Welling, 2017)

(ongoing work: N. Knudde, M. Bauer)

Recap

Dealing with (many) variational params:

- *Collapse* a factor: $\hat{\mathcal{Q}}(q(h)) \geq \mathcal{Q}(q(h), q(u))$
- Amortized inference: $q(h^{(n)}; \theta^{(n)})$ with $\theta^{(n)} = g(\cdot; \boldsymbol{\phi})$
- Re-parameterization: $q(h) \sim \mathcal{N}(\mu, \mathbf{\Sigma})$ with $\mathbf{\Sigma} = g(\boldsymbol{\lambda})$
- Normalizing flows: $q_0(h) \xrightarrow{f_0, f_1, \cdots, f_K} q_K(h)$

Other DeepGP approximations

- Mean-field, amortized, re-parameterized *[Damianou & Lawrence '13, Damianou '15, Dai et al. '14]*
- Approximate scalable EP *[Bui et al. '16]*
- Projected $q(h)$ distribution in nested variational inference. *[Hensman & Lawrence '14]*
- Sample through the $q(f_{1:L})$ chain to maintain layer coupling *[Salimbeni & Deisenroth '17]*
- Sampling + FITC + MAP for inducing variables *[Vafa '16]*
- Approximate kernel's spectral density + VI *[Cutajar et al. '17]*
- DeepGPs & NN regularization connections *[Gal & Ghahramani '15; Louizos & Welling '16]*