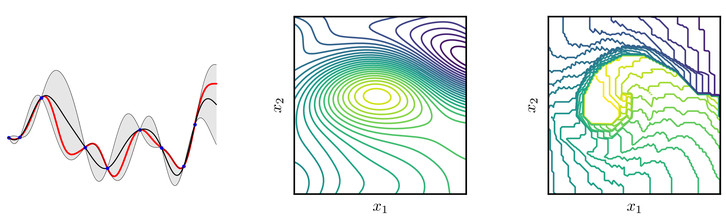
Approximate Inference in Deep Gaussian Processes by Minimizing Alpha Divergences

Carlos Villacampa-Calvo[1], Gonzalo Hernández-Muñoz, Daniel Hernández-Lobato[1]
(1) - Universidad Autónoma de Madrid

1 - Introduction to Gaussian Processes

GPs are flexible models that can output closed form predictive distributions:

GP Predictive Distribution **GP Fit** **Target Function**

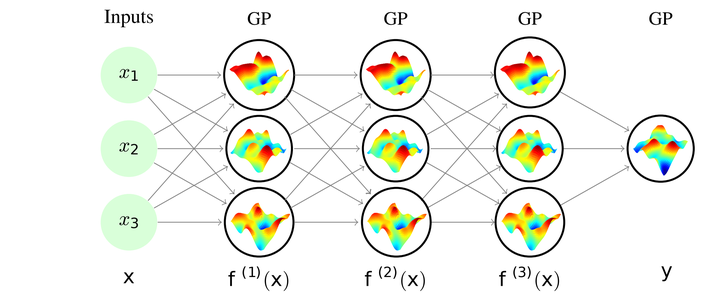


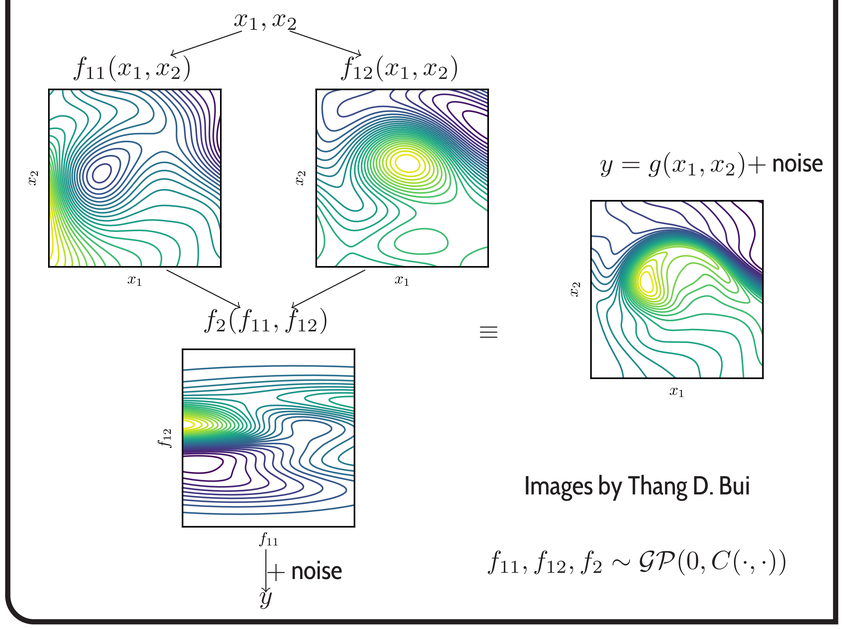
Limitations (Bui et al., 2016):

- Cubic training cost!
- Sparse approximations may limit model flexibility!
- Often the target is assumed to be stationary!

2 - Deep Gaussian Processes

Deep neural network in which the activations functions are GPs:





Images by Thang D. Bui

$f_{11}, f_{12}, f_2 \sim \mathcal{GP}(0, C(\cdot, \cdot))$

3 - Bayesian Inference

Posterior over latent functions (typically at the observed data $\mathbf{X}$):

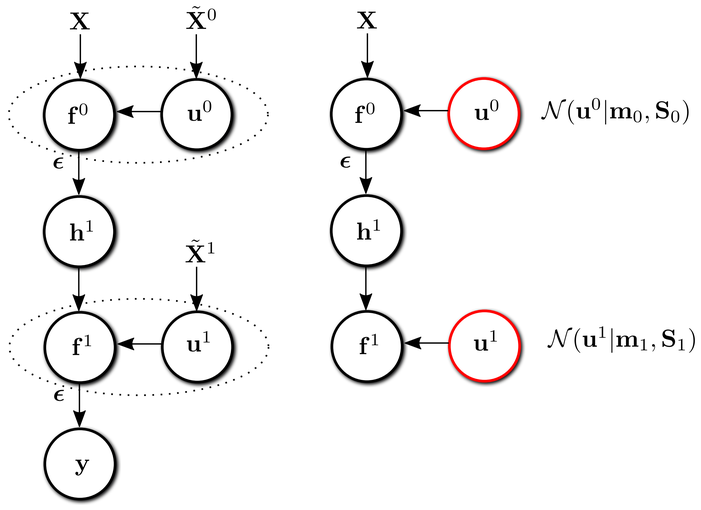
$$p(\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3|\mathbf{Y}) = \frac{p(\mathbf{f}_1)\mathbf{p}(f_2)\mathbf{p}(f_3)p(\mathbf{Y}|\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3, \mathbf{X})}{p(\mathbf{Y})}$$

- GP priors
- Likelihood function
- Marginal likelihood

But the posterior $p(\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3|\mathbf{Y})$ is **intractable**.

4 - Sparse GPs and Posterior Approximation

Inducing points $\tilde{\mathbf{X}}^i$ associated to each GP: $p(\mathbf{f}^i, \mathbf{u}^i|\mathbf{h}^i) = p(\mathbf{f}^i|\mathbf{u}^i, \mathbf{h}^i)p(\mathbf{u}^i)$



$$q(\mathbf{f}^0, \mathbf{u}^0, \mathbf{h}^1, \ldots, \mathbf{f}^L, \mathbf{u}^L) = q(\mathbf{u}^0)p(\mathbf{f}^0|\mathbf{u}^0)p(\mathbf{h}^1|\mathbf{f}^1)\cdots \cdots p(\mathbf{f}^L|\mathbf{u}^L, \mathbf{h}^L)q(\mathbf{u}^L)$$

5 - Alpha Divergence Minimization

$$D_\alpha(p||q) = \frac{\int_{\boldsymbol{\theta}} \left(\alpha p(\boldsymbol{\theta}) + (1-\alpha)q(\boldsymbol{\theta}) - p(\boldsymbol{\theta})^\alpha q(\boldsymbol{\theta})^{1-\alpha}\right) d\boldsymbol{\theta}}{\alpha(1-\alpha)}.$$

[Amari, 1985].

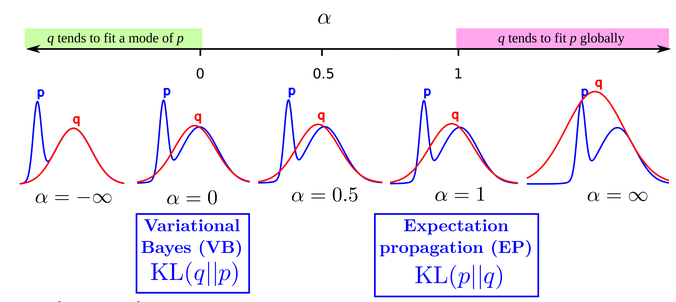


Figure source: [Minka, 2005].

5 - Local α-divergence Minimization via Power EP

Approximates $p(\boldsymbol{\theta}) \propto p_0(\boldsymbol{\theta}) \prod_{n=1}^N f_n(\boldsymbol{\theta})$ with $q(\boldsymbol{\theta}) \propto p_0(\boldsymbol{\theta}) \prod_{n=1}^N \tilde{f}_n(\boldsymbol{\theta})$

(Minka, 2004)

$p(\boldsymbol{\theta}) \propto$ $p_0(\boldsymbol{\theta})$ $f_1(\boldsymbol{\theta})$ $f_2(\boldsymbol{\theta})$ $f_3(\boldsymbol{\theta})$ $\approx$ $q(\boldsymbol{\theta}) \propto$ $p_0(\boldsymbol{\theta})$ $\tilde{f}_1(\boldsymbol{\theta})$ $\tilde{f}_2(\boldsymbol{\theta})$ $\tilde{f}_3(\boldsymbol{\theta})$

The $\tilde{f}_n$ are tuned by minimizing the local α-divergences

$$D_\alpha[p_n||q] \quad \text{for } n = 1, \ldots, N, \quad \text{where} \quad \begin{matrix} p_n(\boldsymbol{\theta}) & \propto & f_n(\boldsymbol{\theta}) \prod_{j \neq n} \tilde{f}_j(\boldsymbol{\theta}) \\ q(\boldsymbol{\theta}) & \propto & \tilde{f}_n(\boldsymbol{\theta}) \prod_{j \neq n} \tilde{f}_j(\boldsymbol{\theta}) \end{matrix}.$$

Power EP steps to refine f_n:

1. Compute cavity: $q^{\backslash \alpha n} \propto q / \tilde{f}_n^\alpha$.
2. Minimize $\mathrm{KL}(Z_n^{-1} f_n^\alpha q^{\backslash \alpha n}||q)$ to find q^{new}.
3. Update factor: $f_n^{\text{new}} = (Z_n q^{\text{new}} / q^{\backslash \alpha n})^{\frac{1}{\alpha}}$.

At convergence the moments of $p_n = Z_n^{-1} f_n^\alpha q^{\backslash \alpha n}$ and q match!

$$\nabla_{\eta_q} D_\alpha[p_n||q] = \frac{Z_{\tilde{p}}}{\alpha} \left(\mathbb{E}_q[s(\boldsymbol{\theta})] - \mathbb{E}_{\tilde{p}}[s(\boldsymbol{\theta})]\right) \propto \nabla_{\eta_q} \mathrm{KL}[\tilde{p}||q]$$

where $\tilde{p} \propto (f_n q^{\backslash n})^\alpha q^{1-\alpha} = f_n^\alpha q^{\backslash \alpha n}$.

The likelihood factors are simply $p(y_i|\mathbf{f}_i^L)$!

KL minimization by adding approx. factors that depend only on $\mathbf{u}^0, \ldots, \mathbf{u}^L$!

6 - Approximate Power Expectation Propagation

The Power-EP approximation to the **evidence** is given by

$$\log Z_{\text{PEP}} = \log Z_q - \log Z_{\text{prior}} + \sum_{n=1}^N \frac{1}{\alpha} \log \mathbf{E}_q \left[\left(\frac{f_n(\boldsymbol{\theta})}{\tilde{f}_n(\boldsymbol{\theta})}\right)^\alpha\right],$$

The power-EP solution for q can be obtained by solving

$$\max_q \min_{\tilde{f}_1, \ldots, \tilde{f}_N} \log Z_{\text{PEP}} \quad \text{subject to} \quad q(\boldsymbol{\theta}) = p_0(\boldsymbol{\theta}) \prod_{n=1}^N \tilde{f}_n(\boldsymbol{\theta}).$$

Solved with **double-loop** algorithm (Heskes, 2002). **Too slow in practice!**

By following (Li et al., 2015):

$p(\boldsymbol{\theta}) \propto$ $p_0(\boldsymbol{\theta})$ $f_1(\boldsymbol{\theta})$ $f_2(\boldsymbol{\theta})$ $f_3(\boldsymbol{\theta})$ $\approx$ $q(\boldsymbol{\theta}) \propto$ $p_0(\boldsymbol{\theta})$ $\tilde{f}_1(\boldsymbol{\theta})$ $\tilde{f}_2(\boldsymbol{\theta})$ $\tilde{f}_3(\boldsymbol{\theta})$

We tie the factor approximations

$p(\boldsymbol{\theta}) \propto$ $p_0(\boldsymbol{\theta})$ $f_1(\boldsymbol{\theta})$ $f_2(\boldsymbol{\theta})$ $f_3(\boldsymbol{\theta})$ $\approx$ $q(\boldsymbol{\theta}) \propto$ $p_0(\boldsymbol{\theta})$ $\tilde{f}(\boldsymbol{\theta})^N$

- $\max_q \min_{\tilde{f}_1, \ldots, \tilde{f}_N}$ problem $\rightarrow \max_q$ problem, **no double-loop needed!**

The expectations can be approximated via Monte Carlo by sampling!