Sparse GPs

James Hensman
Gaussian Process Summer School
Sheffield
September 2014

Overview

Motivation

Posteriors over function values

Posteriors over inducing points

Distributed Computation and Stochastic optimization

Overview

Motivation

Posteriors over function values

Posteriors over inducing points

Distributed Computation and Stochastic optimization

Motivation

Inference in a GP has the following demands:

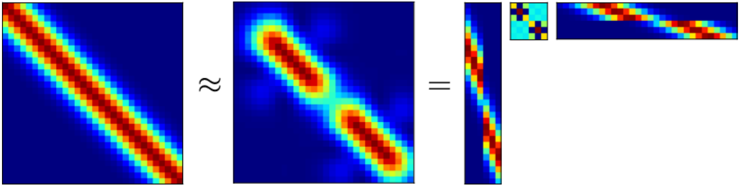
$$\begin{aligned} \text{Complexity:} \quad & \mathcal{O}(n^3) \\ \text{Storage:} \quad & \mathcal{O}(n^2) \end{aligned}$$

Inference in a *sparse* GP has the following demands:

$$\begin{aligned} \text{Complexity:} \quad & \mathcal{O}(nm^2) \\ \text{Storage:} \quad & \mathcal{O}(nm) \end{aligned}$$

where we get to pick m!

How to make computational savings



$$\mathbf{K}_{nn} \approx \mathbf{Q}_{nn} = \mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\mathbf{K}_{mn}$$

Instead of inverting $\mathbf{K}_{nn}$, we make a low rank (or Nyström) approximation, and invert $\mathbf{K}_{mm}$ instead.

Overview

Motivation

Posteriors over function values

Posteriors over inducing points

Distributed Computation and Stochastic optimization

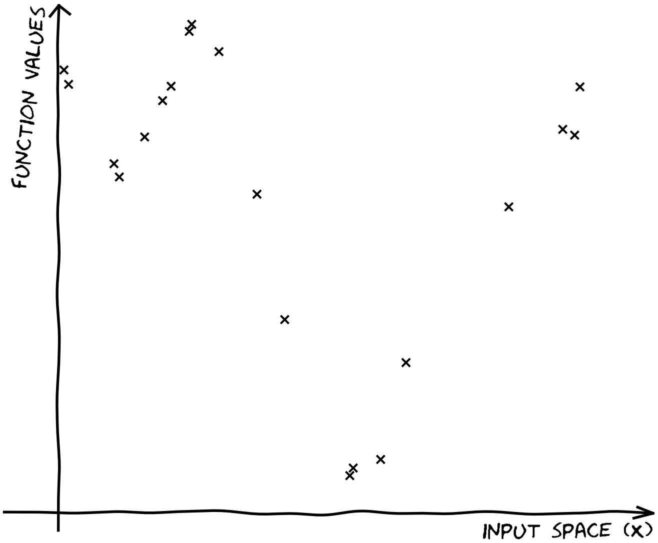
Posteriors over function values

Everything we want to do with a GP involves marginalising $\mathbf{f}$

- Predictions
- Marginal likelihood
- Estimating covariance parameters

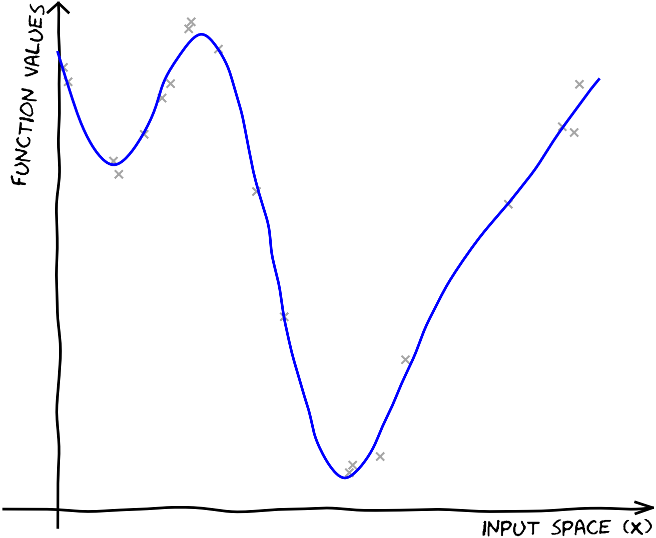
The posterior of $\mathbf{f}$ is the central object. This means inverting $\mathbf{K}_{nn}$.

X, y



$\mathbf{X}, \mathbf{y}$

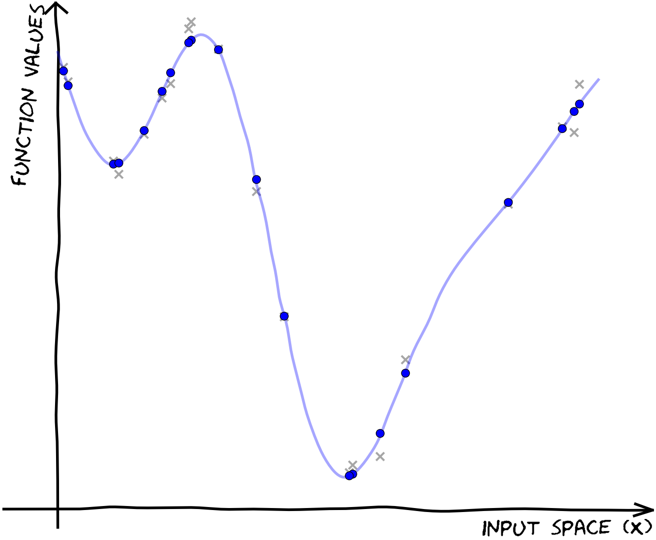
$f(\mathbf{x}) \sim \mathcal{GP}$



$\mathbf{X}, \mathbf{y}$

$f(\mathbf{x}) \sim \mathcal{GP}$

$p(\mathbf{f}) = \mathcal{N}(\mathbf{0}, \mathbf{K}_{nn})$

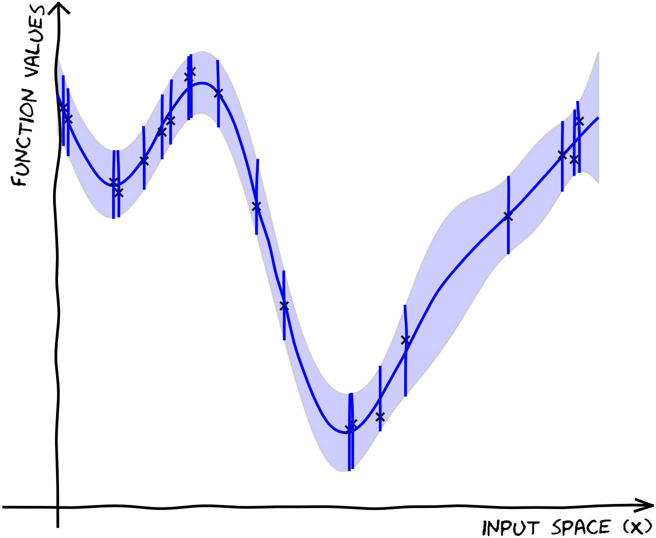


$\mathbf{X}, \mathbf{y}$

$f(\mathbf{x}) \sim \mathcal{GP}$

$p(\mathbf{f}) = \mathcal{N}(\mathbf{0}, \mathbf{K}_{nn})$

$p(\mathbf{f} \mid \mathbf{y}, \mathbf{X})$



Overview

Motivation

Posteriors over function values

Posteriors over inducing points

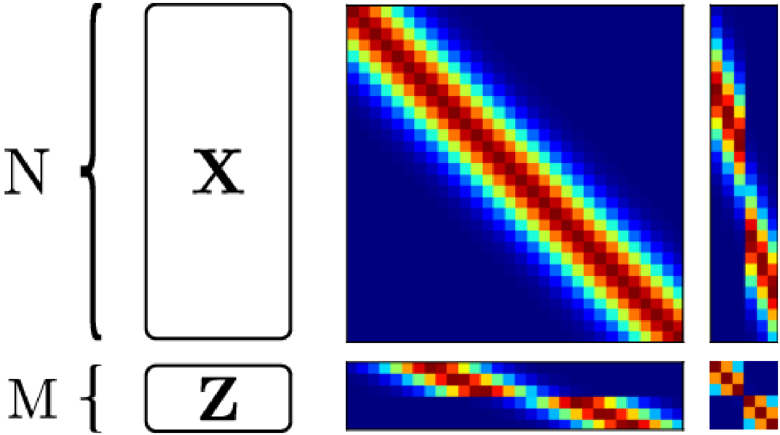
Distributed Computation and Stochastic optimization

Introducing $\mathbf{u}$

Take and extra M points on the function, $\mathbf{u} = f(\mathbf{Z})$.

$$p(\mathbf{y}, \mathbf{f}, \mathbf{u}) = p(\mathbf{y}\,|\,\mathbf{f})p(\mathbf{f}\,|\,\mathbf{u})p(\mathbf{u})$$

Introducing $\mathbf{u}$



Introducing $\mathbf{u}$

Take and extra M points on the function, $\mathbf{u} = f(\mathbf{Z})$.

$$p(\mathbf{y}, \mathbf{f}, \mathbf{u}) = p(\mathbf{y} \,|\, \mathbf{f}) p(\mathbf{f} \,|\, \mathbf{u}) p(\mathbf{u})$$

$$\begin{aligned} p(\mathbf{y} \,|\, \mathbf{f}) &= \mathcal{N}\left(\mathbf{y} | \mathbf{f}, \sigma^2 \mathbf{I}\right) \\ p(\mathbf{f} \,|\, \mathbf{u}) &= \mathcal{N}\left(\mathbf{f} | \mathbf{K}_{nm} \mathbf{K}_{mm}1 \mathbf{u}, \widetilde{\mathbf{K}}\right) \\ p(\mathbf{u}) &= \mathcal{N}\left(\mathbf{u} | \mathbf{0}, \mathbf{K}_{mm}\right) \end{aligned}$$

$\mathbf{X}, \mathbf{y}$

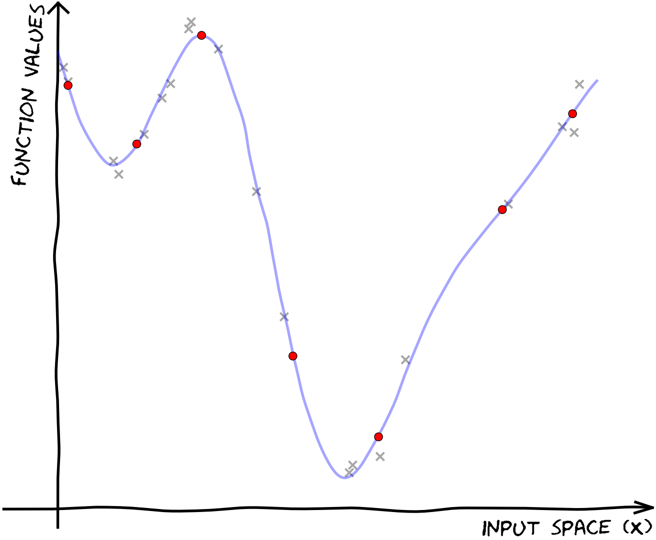
$f(\mathbf{x}) \sim \mathcal{GP}$

$p(\mathbf{f}) = \mathcal{N}(\mathbf{0}, \mathbf{K}_{nn})$

$p(\mathbf{f} \mid \mathbf{y}, \mathbf{X})$

$\mathbf{Z}, \mathbf{u}$

$p(\mathbf{u}) = \mathcal{N}(\mathbf{0}, \mathbf{K}_{mm}$



$\mathbf{X}, \mathbf{y}$

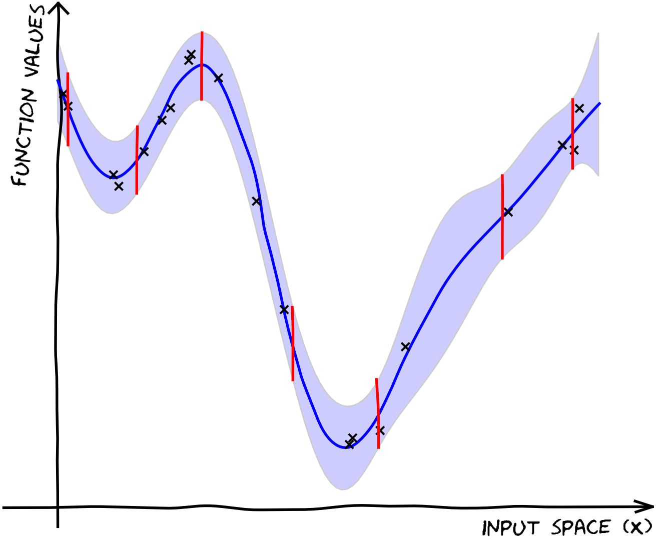
$f(\mathbf{x}) \sim \mathcal{GP}$

$p(\mathbf{f}) = \mathcal{N}(\mathbf{0}, \mathbf{K}_{nn})$

$p(\mathbf{f} \mid \mathbf{y}, \mathbf{X})$

$p(\mathbf{u}) = \mathcal{N}(\mathbf{0}, \mathbf{K}_{mm}$

$\tilde{p}(\mathbf{u} \mid \mathbf{y}, \mathbf{X})$



The alternative posterior

Instead of doing

$$p(\mathbf{f}\,|\,\mathbf{y},\mathbf{X}) = \frac{p(\mathbf{y}\,|\,\mathbf{f})p(\mathbf{f}\,|\,\mathbf{X})}{\int p(\mathbf{y}\,|\,\mathbf{f})p(\mathbf{f}\,|\,\mathbf{X})\mathrm{d}\mathbf{f}}$$

We'll do

$$p(\mathbf{u}\,|\,\mathbf{y},\mathbf{Z}) = \frac{p(\mathbf{y}\,|\,\mathbf{u})p(\mathbf{u}\,|\,\mathbf{Z})}{\int p(\mathbf{y}\,|\,\mathbf{u})p(\mathbf{u}\,|\,\mathbf{Z})\mathrm{d}\mathbf{u}}$$

The alternative posterior

Instead of doing

$$p(\mathbf{f}\,|\,\mathbf{y}, \mathbf{X}) = \frac{p(\mathbf{y}\,|\,\mathbf{f})p(\mathbf{f}\,|\,\mathbf{X})}{\int p(\mathbf{y}\,|\,\mathbf{f})p(\mathbf{f}\,|\,\mathbf{X})\mathrm{d}\mathbf{f}}$$

We'll do

$$p(\mathbf{u}\,|\,\mathbf{y}, \mathbf{Z}) = \frac{p(\mathbf{y}\,|\,\mathbf{u})p(\mathbf{u}\,|\,\mathbf{Z})}{\int p(\mathbf{y}\,|\,\mathbf{u})p(\mathbf{u}\,|\,\mathbf{Z})\mathrm{d}\mathbf{u}}$$

but $p(\mathbf{y}\,|\,\mathbf{u})$ involves inverting $\mathbf{K}_{nn}$

Variational marginalisation of $\mathbf{f}$

$$p(\mathbf{y}|\mathbf{u}) = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u})}{p(\mathbf{f}|\mathbf{y},\mathbf{u})}$$

Variational marginalisation of $\mathbf{f}$

$$p(\mathbf{y}|\mathbf{u}) = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u})}{p(\mathbf{f}|\mathbf{y},\mathbf{u})}$$

$$\ln p(\mathbf{y}|\mathbf{u}) = \ln p(\mathbf{y}|\mathbf{f}) + \ln \frac{p(\mathbf{f}|\mathbf{u})}{p(\mathbf{f}|\mathbf{y},\mathbf{u})}$$

Variational marginalisation of $\mathbf{f}$

$$p(\mathbf{y}|\mathbf{u}) = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u})}{p(\mathbf{f}|\mathbf{y},\mathbf{u})}$$

$$\ln p(\mathbf{y}|\mathbf{u}) = \ln p(\mathbf{y}|\mathbf{f}) + \ln \frac{p(\mathbf{f}|\mathbf{u})}{p(\mathbf{f}|\mathbf{y},\mathbf{u})}$$

$$\ln p(\mathbf{y}|\mathbf{u}) = \mathbb{E}_{p(\mathbf{f}|\mathbf{u})}\Big[\ln p(\mathbf{y}|\mathbf{f})\Big] + \mathbb{E}_{p(\mathbf{f}|\mathbf{u})}\Big[\ln \frac{p(\mathbf{f}|\mathbf{u})}{p(\mathbf{f}|\mathbf{y},\mathbf{u})}\Big]$$

Variational marginalisation of $\mathbf{f}$

$$p(\mathbf{y}\,|\,\mathbf{u}) = \frac{p(\mathbf{y}\,|\,\mathbf{f})p(\mathbf{f}\,|\,\mathbf{u})}{p(\mathbf{f}\,|\,\mathbf{y},\mathbf{u})}$$

$$\ln p(\mathbf{y}\,|\,\mathbf{u}) = \ln p(\mathbf{y}\,|\,\mathbf{f}) + \ln \frac{p(\mathbf{f}\,|\,\mathbf{u})}{p(\mathbf{f}\,|\,\mathbf{y},\mathbf{u})}$$

$$\ln p(\mathbf{y}\,|\,\mathbf{u}) = \mathbb{E}_{p(\mathbf{f}\,|\,\mathbf{u})}\Big[\ln p(\mathbf{y}\,|\,\mathbf{f})\Big] + \mathbb{E}_{p(\mathbf{f}\,|\,\mathbf{u})}\Big[\ln \frac{p(\mathbf{f}\,|\,\mathbf{u})}{p(\mathbf{f}\,|\,\mathbf{y},\mathbf{u})}\Big]$$

$$\ln p(\mathbf{y}\,|\,\mathbf{u}) = \tilde{p}(\mathbf{y}\,|\,\mathbf{u}) + \mathrm{KL}[p(\mathbf{f}|\mathbf{u})\|p(\mathbf{f}\,|\,\mathbf{y},\mathbf{u})]$$

No inversion of $\mathbf{K}_{nn}$ required

Variational marginalisation of $\mathbf{f}$ (another way)

$$\ln p(\mathbf{y}|\mathbf{u}) = \ln \int p(\mathbf{y}|\mathbf{f}) p(\mathbf{f}|\mathbf{u}, \mathbf{X}) \mathrm{d}\mathbf{f}$$

Variational marginalisation of $\mathbf{f}$ (another way)

$$\ln p(\mathbf{y}|\mathbf{u}) = \ln \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u}, \mathbf{X})\mathrm{d}\mathbf{f}$$

$$\ln p(\mathbf{y}|\mathbf{u}) = \ln \mathbb{E}_{p(\mathbf{f}|\mathbf{u},\mathbf{X})}\left[p(\mathbf{y}|\mathbf{f})\right]$$

Variational marginalisation of $\mathbf{f}$ (another way)

$$\ln p(\mathbf{y}\,|\,\mathbf{u}) = \ln \int p(\mathbf{y}\,|\,\mathbf{f}) p(\mathbf{f}\,|\,\mathbf{u}, \mathbf{X}) \mathrm{d}\mathbf{f}$$

$$\ln p(\mathbf{y}\,|\,\mathbf{u}) = \ln \mathbb{E}_{p(\mathbf{f}\,|\,\mathbf{u}, \mathbf{X})} \left[p(\mathbf{y}\,|\,\mathbf{f}) \right]$$

$$\ln p(\mathbf{y}\,|\,\mathbf{u}) \geq \mathbb{E}_{p(\mathbf{f}\,|\,\mathbf{u}, \mathbf{X})} \left[\ln p(\mathbf{y}\,|\,\mathbf{f}) \right] \triangleq \ln \tilde{p}(\mathbf{y}\,|\,\mathbf{u})$$

Variational marginalisation of $\mathbf{f}$ (another way)

$$\ln p(\mathbf{y}\,|\,\mathbf{u}) = \ln \int p(\mathbf{y}\,|\,\mathbf{f})p(\mathbf{f}\,|\,\mathbf{u}, \mathbf{X})\mathrm{d}\mathbf{f}$$

$$\ln p(\mathbf{y}\,|\,\mathbf{u}) = \ln \mathbb{E}_{p(\mathbf{f}\,|\,\mathbf{u},\mathbf{X})}\left[p(\mathbf{y}\,|\,\mathbf{f})\right]$$

$$\ln p(\mathbf{y}\,|\,\mathbf{u}) \geq \mathbb{E}_{p(\mathbf{f}\,|\,\mathbf{u},\mathbf{X})}\left[\ln p(\mathbf{y}\,|\,\mathbf{f})\right] \triangleq \ln \tilde{p}(\mathbf{y}\,|\,\mathbf{u})$$

No inversion of $\mathbf{K}_{nn}$ required

An approximate likelihood

$$\tilde{p}(\mathbf{y} \,|\, \mathbf{u}) = \prod_{i=1}^{n} \mathcal{N}\left(\mathbf{y}_i | \mathbf{k}_{mn}^{\top} \mathbf{K}_{mm}^{-1} \mathbf{u}, \sigma^2\right) \exp\left\{-\frac{1}{2\sigma^2}\left(k_{nn} - \mathbf{k}_{mn}^{\top} \mathbf{K}_{mm}^{-1} \mathbf{k}_{mn}\right)\right\}$$

A straightforward likelihood approximation, and a penalty term

Now we can marginalise $\mathbf{u}$

$$\tilde{p}(\mathbf{u} \,|\, \mathbf{y}, \mathbf{Z}) = \frac{\tilde{p}(\mathbf{y} \,|\, \mathbf{u}) p(\mathbf{u} \,|\, \mathbf{Z})}{\int \tilde{p}(\mathbf{y} \,|\, \mathbf{u}) p(\mathbf{u} \,|\, \mathbf{Z}) \mathrm{d}\mathbf{u}}$$

- Computing the (approximate) posterior costs $\mathcal{O}(nm^2)$
- We also get a lower bound of the marginal likelihood
- This is the standard variational sparse GP (**?**).

The marginal likelihood lower bound

$$
\begin{aligned}
\tilde{p}(\mathbf{y}) &= \int \tilde{p}(\mathbf{y} \mid \mathbf{u}) p(\mathbf{u} \mid \mathbf{Z}) \mathrm{d}\mathbf{u} \\
&= \mathcal{N}\left(\mathbf{y}|\mathbf{0}, \mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\mathbf{K}_{mn} + \sigma^2\mathbf{I}\right) \exp \sum_i \left\{ -\tfrac{1}{2\sigma^2}\left(k_{nn} - \mathbf{k}_{mn}^{\top}\mathbf{K}_{mm}^{-1}\mathbf{k}_{mn}\right)\right\}
\end{aligned}
$$

Optimisation

The variational objective $\ln \tilde{p}(\mathbf{y})$ is a function of

- the parameters of the covariance function $\boldsymbol{\theta}$
- the inducing inputs, $\mathbf{Z}$

Strategy: jointly optimize $\boldsymbol{\theta}$ and $\mathbf{Z}$.

Overview

Motivation

Posteriors over function values

Posteriors over inducing points

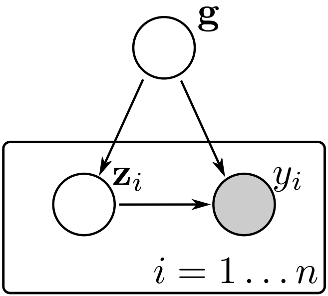
Distributed Computation and Stochastic optimization

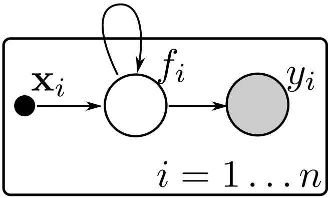
Variational Bayes

- Approximate the true posterior distribution with a simpler one.
- Usually assume factorisation in the approximation
- Iterative 'update' procedure (like EM)
- Can be seen as a coordinate-wise steepest ascent method

Stochastic Variational Inference

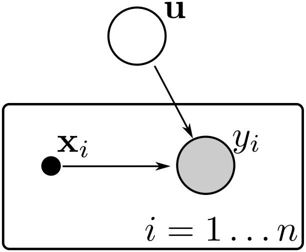
- Combine the ideas of stochastic optimisation with Variational inference
- example: apply Latent Dirichlet allocation to project Gutenberg
- Can apply variational techniques to Big Data
- How could this work in GPs?





Maintain the factorisation!

- The variational marginalisation of $\mathbf{f}$ introduced factorisation across the datapoints (conditioned on $\mathbf{u}$)
- Marginalising $\mathbf{u}$ re-introdcuced dependencies between the data
- Solution: a variational treatment of $\mathbf{u}$



$$\log p(\mathbf{y} \,|\, \mathbf{X}) \geq \left\langle \log \tilde{p}(\mathbf{y} \,|\, \mathbf{u}) + \log p(\mathbf{u}) - \log q(\mathbf{u}) \right\rangle_{q(\mathbf{u})} \triangleq \mathcal{L}. \tag{1}$$

$$\begin{aligned} \mathcal{L} = \sum_{i=1}^{n} \Bigg\{ & \log \mathcal{N}\left(y_i | \mathbf{k}_{mn}^{\top} \mathbf{K}_{mm}^{-1} \mathbf{m}, \beta^{-1}\right) \\ & - \frac{1}{2} \beta \tilde{k}_{i,i} - \frac{1}{2} \operatorname{tr}\left(\mathbf{S} \boldsymbol{\Lambda}_i\right) \Bigg\} \\ & - \mathrm{KL}\left(q(\mathbf{u}) \,\|\, p(\mathbf{u})\right) \end{aligned} \tag{2}$$

Optimisation

The variational objective $\mathcal{L}$ is a function of

- the parameters of the covariance function
- the parameters of $q(\mathbf{u})$
- the inducing inputs, $\mathbf{Z}$

Strategy: set $\mathbf{Z}$. Take the data in small minibatches, take stochastic gradient steps in the covariance function parameters, stochastic *natural* gradient steps in the parameters of $q(\mathbf{u})$.

Natural Gradients

$$\tilde{\mathbf{g}}(\boldsymbol{\theta}) = G(\boldsymbol{\theta})^{-1}\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}} = \frac{\partial \mathcal{L}}{\partial \boldsymbol{\eta}}.$$

$$\begin{aligned}
\boldsymbol{\theta}_{2(t+1)} &= -\frac{1}{2}\mathbf{S}\mathrm{1}_{(t+1)} \\
&= -\frac{1}{2}\mathbf{S}\mathrm{1}_{(t)} + \ell\left(-\frac{1}{2}\boldsymbol{\Lambda} + \frac{1}{2}\mathbf{S}\mathrm{1}_{(t)}\right), \\
\boldsymbol{\theta}_{1(t+1)} &= \mathbf{S}\mathrm{1}_{(t+1)}\mathbf{m}_{(t+1)} \\
&= \mathbf{S}\mathrm{1}_{(t)}\mathbf{m}_{(t)} + \ell\left(\beta\mathbf{K}_{mm}\mathrm{1}\mathbf{K}_{mn}\mathbf{y} - \mathbf{S}\mathrm{1}_{(t)}\mathbf{m}_{(t)}\right),
\end{aligned}$$