SPARSE GAUSSIAN PROCESSES

GPSS 2015, Sheffield

James Hensman

September 15, 2015

Lancaster University

MOTIVATION

Inference in a GP has the following demands:

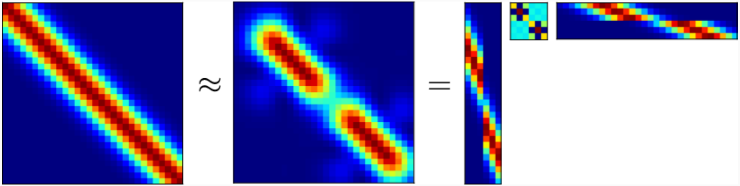
$$\begin{aligned} \text{Complexity:} \quad & \mathcal{O}(n^3) \\ \text{Storage:} \quad & \mathcal{O}(n^2) \end{aligned}$$

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



$$K_{nn} \approx Q_{nn} = K_{nm} K_{mm}^{-1} K_{mn}$$

Instead of inverting K_{nn}, we make a low rank (or Nyström) approximation, and invert K_{mm} instead.

A HISTORY LESSON

WHY ARE THEY CALLED SPARSE GPS?

Sparse (adj). From spagare, meaning “few and scattered”.

Subset of data

- Silverman 1985 (subset of regressors)
- Smola and Bartlett 2001 (greedy selection)

Pseudo-input approximaitons

- Snelson and Ghahramani (2005), Snelson (2007)

Variational approximations

- Titsias (2009) – derived a variational bound
- Matthews et al. (2015) – showed this minimised KL between processes

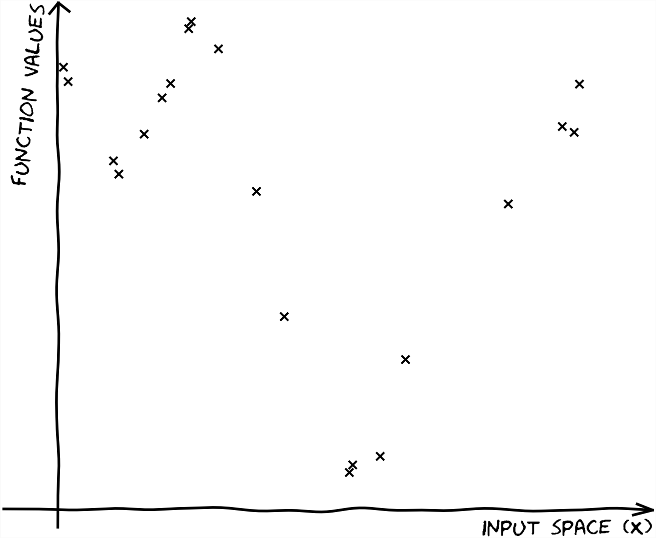
POSTERIORS OVER FUNCTIONS

Everything we want to do with a GP involves marginalising f

- Predictions
- Marginal likelihood
- Estimating covariance parameters

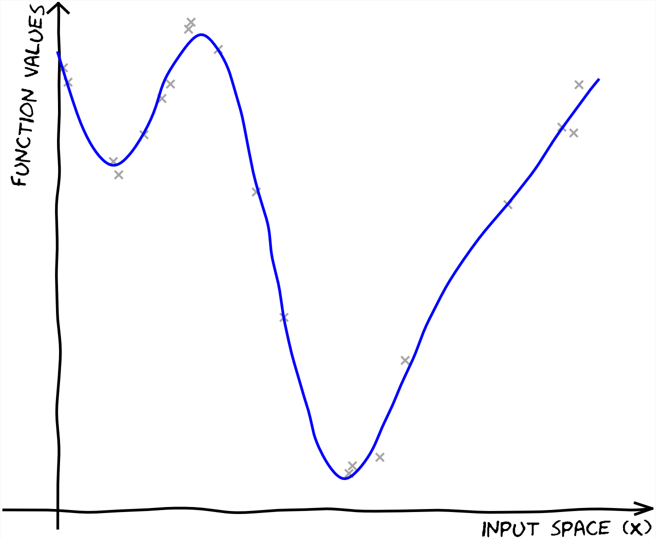
The posterior of $\mathbf{f}$ is the central object. This means inverting $\mathbf{K}_{nn}$.

X, y



X, y

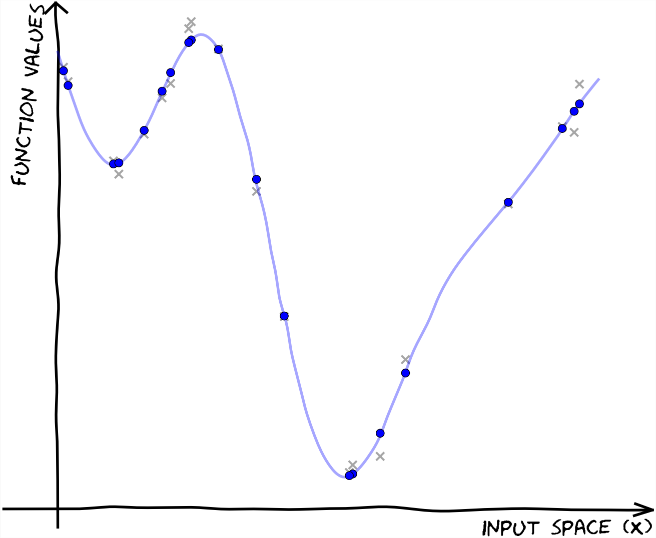
$f(x) \sim \mathcal{GP}$



X, y

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

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



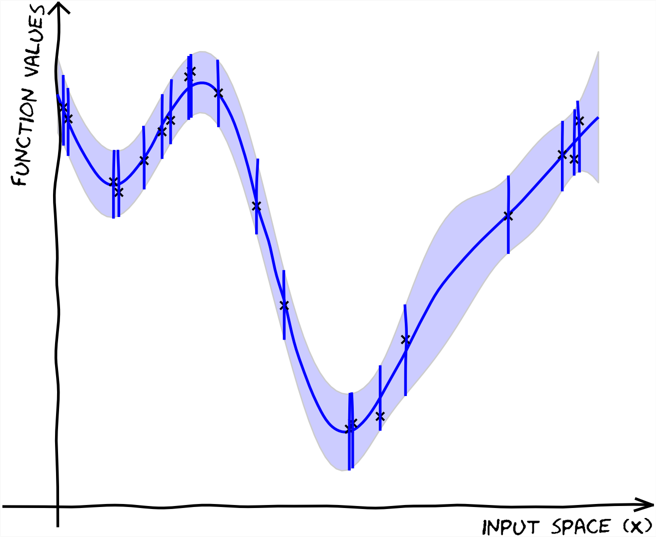
X, y

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

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

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

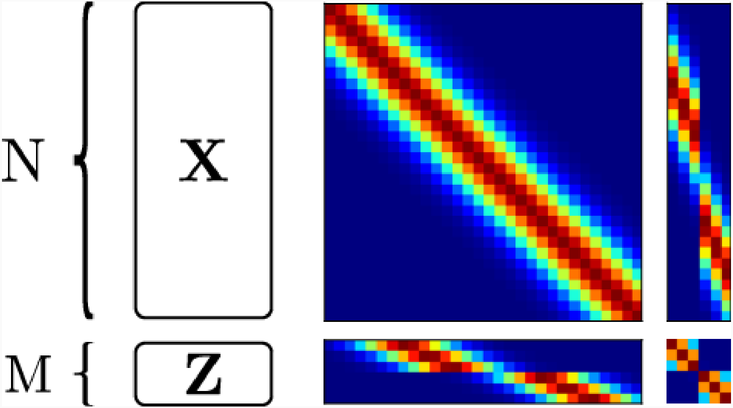
$p(\mathbf{f}^{\star} \mid \mathbf{f}, X, x^{\star})$



POSTERIORS OVER INDUCING POINTS

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

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


N
X
M
Z

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

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

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

X, y

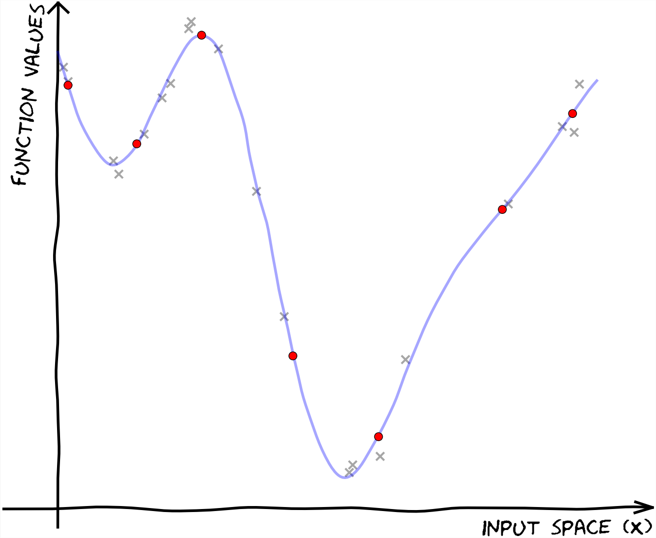
$f(x) \sim \mathcal{GP}$

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

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

Z, u

$p(u) = \mathcal{N}(0, K_{mm})$



X, y

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

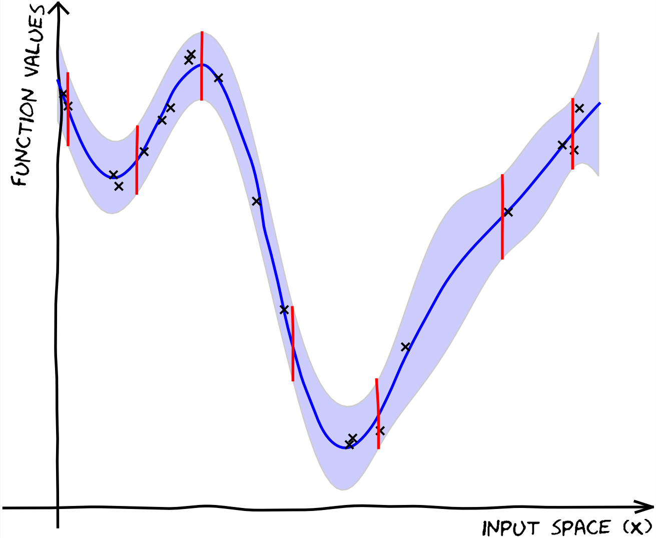
$p(\mathbf{f}) = \mathcal{N}(0, K_{nn})$

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

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

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

$p(\mathbf{f}^{\star} \mid \mathbf{u}, Z, x^{\star})$



Instead of doing

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

We'll do

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

THE ALTERNATIVE POSTERIOR

Instead of doing

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

We'll do

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

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

VARIATIONAL MARGINALISATION OF $\mathbf{f}$

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

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

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

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

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

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

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

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

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

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

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

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

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

variational marginalisation of f (another way)

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

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

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

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

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

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

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

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

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

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

$$\tilde{p}(y \mid u) = \prod_{i=1}^{n} \mathcal{N}\left(y_i | k_{mn}^{\top} K_{mm}^{-1} u, \sigma^2\right) \; \exp\left\{-\frac{1}{2\sigma^2}\left(k_{nn} - k_{mn}^{\top} K_{mm}^{-1} k_{mn}\right)\right\}$$

A straightforward likelihood approximation, and a penalty term

NOW WE CAN MARGINALISE U

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

- 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 as Titsias 2009
- looks like a low rank approximation.

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

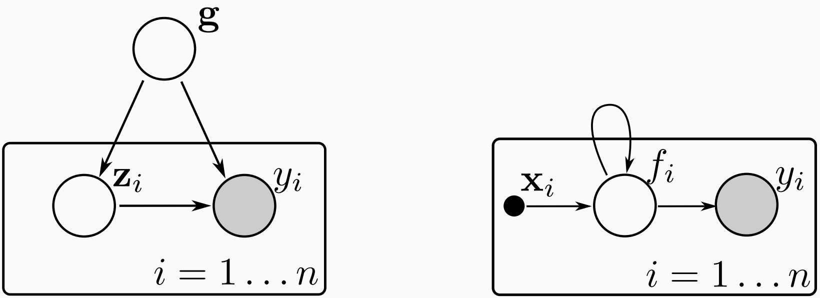
The variational objective $\ln \tilde{p}(\mathbf{y})$ is a function of

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

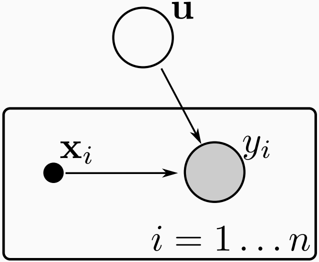
Strategy: jointly optimize $\boldsymbol{\theta}$ and $\mathbf{Z}$.

DISTRIBUTED COMPUTATION AND STOCHASTIC OPTIMIZATION

- 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?



- 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 $\mathbf{u}$



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

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

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

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

Original 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 $q(\mathbf{u})$.

New strategy: optimize everything jointly with AdaDelta.

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

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

PREDICTION AND THE KL BETWEEN PROCESSES

We have minimised the KL divergence

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

We have minimised the KL divergence

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

but this turns out to be equivalent to

$$\mathrm{KL}[p(\mathbf{f}^\star \mid \mathbf{u})\tilde{p}(\mathbf{u})||p(\mathbf{f}^\star \mid \mathbf{y})p(\mathbf{f} \mid \mathbf{y})]$$

To predict, just compute the required quantity of the variational stochastic process.

$$\mathrm{p}(\mathsf{f}^\star \mid \mathsf{y}) \approx \int \mathrm{p}(\mathsf{f}^\star \mid \mathsf{u})\tilde{\mathrm{p}}(\mathsf{u} \mid \mathsf{y}) \, \mathrm{d}\mathsf{u}$$

SUMMARY AND DEMO

- I have guided you through the variational sparse GP method (for regression).

SUMMARY

- I have guided you through the variational sparse GP method (for regression).
- Great framework for extensibility
  - Non Gaussian likelihoods
  - Multiple outputs
  - Stochastic optimization
  - ...

SUMMARY

- I have guided you through the variational sparse GP method (for regression).
- Great framework for extensibility
  - Non Gaussian likelihoods
  - Multiple outputs
  - Stochastic optimization
  - ...
- Move away from thinking of a model approximation: separate model and inference

SUMMARY

- I have guided you through the variational sparse GP method (for regression).
- Great framework for extensibility
  - Non Gaussian likelihoods
  - Multiple outputs
  - Stochastic optimization
  - ...
- Move away from thinking of a model approximation: separate model and inference
- Work of many authors (406,000 scholar hits!). Apologies for the 405,995 omissions