Low Rank Gaussian Processes

Neil D. Lawrence

GPRS
13th February 2014

Outline

Parametric Bottleneck

Nonparametric Gaussian Processes

- This work takes us from parametric to non-parametric.
- The limit implies infinite dimensional $\mathbf{w}$.
- Gaussian processes are generally non-parametric: combine data with covariance function to get model.
- This representation *cannot* be summarized by a parameter vector of a fixed size.

The Parametric Bottleneck

- Parametric models have a representation that does not respond to increasing training set size.
- Bayesian posterior distributions over parameters contain the information about the training data.
  - Use Bayes' rule from training data, $p(\mathbf{w}|\mathbf{y}, \mathbf{X})$,
  - Make predictions on test data

$$p(y_*|\mathbf{X}_*, \mathbf{y}, \mathbf{X}) = \int p(y_*|\mathbf{w}, \mathbf{X}_*)\, p(\mathbf{w}|\mathbf{y}, \mathbf{X})\mathrm{d}\mathbf{w}) .$$

- $\mathbf{w}$ becomes a bottleneck for information about the training set to pass to the test set.
- Solution: increase m so that the bottleneck is so large that it no longer presents a problem.
- How big is big enough for m? Non-parametrics says $m \to \infty$.

The Parametric Bottleneck

- Now no longer possible to manipulate the model through the standard parametric form.

The Parametric Bottleneck

- Now no longer possible to manipulate the model through the standard parametric form.
- However, it *is* possible to express *parametric* as GPs:

$$k\left(\mathbf{x}_i, \mathbf{x}_j\right) = \phi_:\left(\mathbf{x}_i\right)^\top \phi_:\left(\mathbf{x}_j\right).$$

The Parametric Bottleneck

- Now no longer possible to manipulate the model through the standard parametric form.
- However, it *is* possible to express *parametric* as GPs:

$$k\left(\mathbf{x}_i, \mathbf{x}_j\right) = \phi_{:}\left(\mathbf{x}_i\right)^\top \phi_{:}\left(\mathbf{x}_j\right).$$

- These are known as degenerate covariance matrices.

The Parametric Bottleneck

- Now no longer possible to manipulate the model through the standard parametric form.
- However, it *is* possible to express *parametric* as GPs:

$$k\left(\mathbf{x}_i, \mathbf{x}_j\right) = \phi_:\left(\mathbf{x}_i\right)^\top \phi_:\left(\mathbf{x}_j\right).$$

- These are known as degenerate covariance matrices.
- Their rank is at most m, non-parametric models have full rank covariance matrices.

The Parametric Bottleneck

- Now no longer possible to manipulate the model through the standard parametric form.
- However, it *is* possible to express *parametric* as GPs:

$$k\left(\mathbf{x}_i, \mathbf{x}_j\right) = \phi_{:}\left(\mathbf{x}_i\right)^\top \phi_{:}\left(\mathbf{x}_j\right).$$

- These are known as degenerate covariance matrices.
- Their rank is at most m, non-parametric models have full rank covariance matrices.
- Most well known is the "linear kernel", $k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^\top \mathbf{x}_j$.

Making Predictions

- For non-parametrics prediction at new points $\mathbf{f}_*$ is made by conditioning on $\mathbf{f}$ in the joint distribution.

Making Predictions

- For non-parametrics prediction at new points $\mathbf{f}_*$ is made by conditioning on $\mathbf{f}$ in the joint distribution.
- In GPs this involves combining the training data with the covariance function and the mean function.

Making Predictions

- For non-parametrics prediction at new points $\mathbf{f}_*$ is made by conditioning on $\mathbf{f}$ in the joint distribution.
- In GPs this involves combining the training data with the covariance function and the mean function.
- Parametric is a special case when conditional prediction can be summarized in a *fixed* number of parameters.

Making Predictions

- For non-parametrics prediction at new points $\mathbf{f}_*$ is made by conditioning on $\mathbf{f}$ in the joint distribution.
- In GPs this involves combining the training data with the covariance function and the mean function.
- Parametric is a special case when conditional prediction can be summarized in a *fixed* number of parameters.
- Complexity of parametric model remains fixed regardless of the size of our training data set.

Making Predictions

- For non-parametrics prediction at new points $\mathbf{f}_*$ is made by conditioning on $\mathbf{f}$ in the joint distribution.
- In GPs this involves combining the training data with the covariance function and the mean function.
- Parametric is a special case when conditional prediction can be summarized in a *fixed* number of parameters.
- Complexity of parametric model remains fixed regardless of the size of our training data set.
- For a non-parametric model the required number of parameters grows with the size of the training data.

Low Rank Motivation

Inference in a GP has the following demands:

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

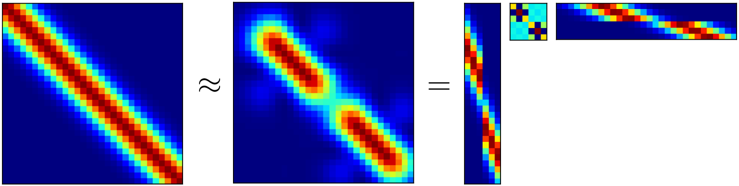
Inference in a low rank GP has the following demands:

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

where m is a user chosen parameter.

Computational Savings

(Smola and Bartlett, 2001; Csató and Opper, 2001, 2002; Csató, 2002; Seeger et al., 2003)

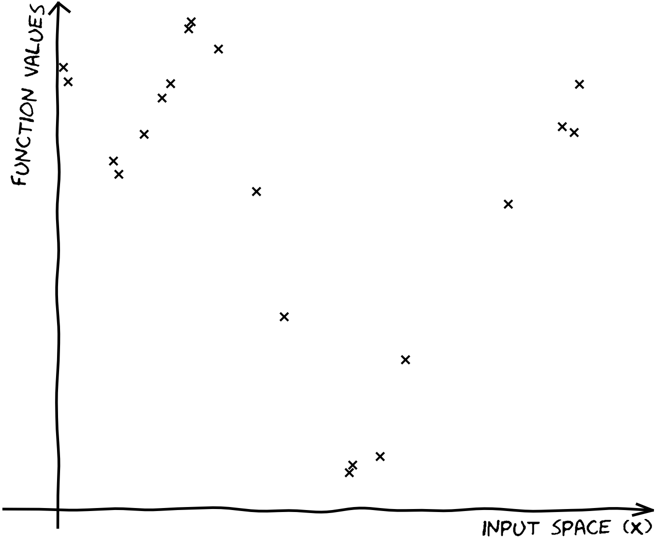


$$\mathbf{K_{ff}} \approx \mathbf{Q_{ff}} = \mathbf{K_{fu}} \mathbf{K_{uu}^{-1}} \mathbf{K_{uf}}$$

Instead of inverting $\mathbf{K_{ff}}$, we make a low rank (or Nyström) approximation, and invert $\mathbf{K_{uu}}$ instead.

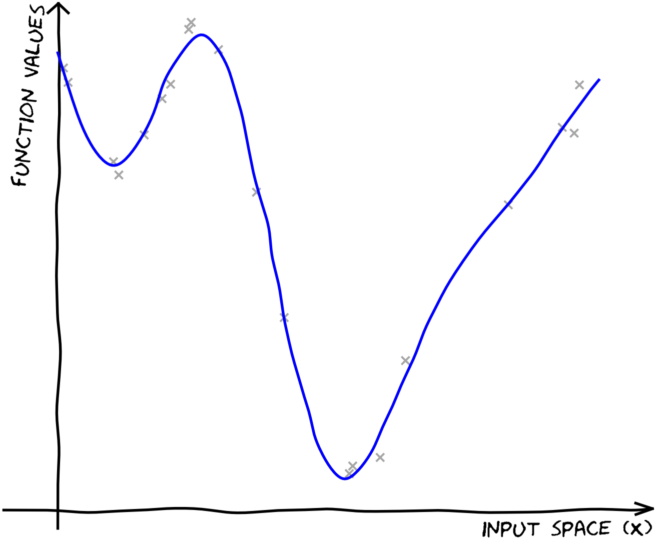
Figure originally from presentation by Ed Snelson at NIPS

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



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

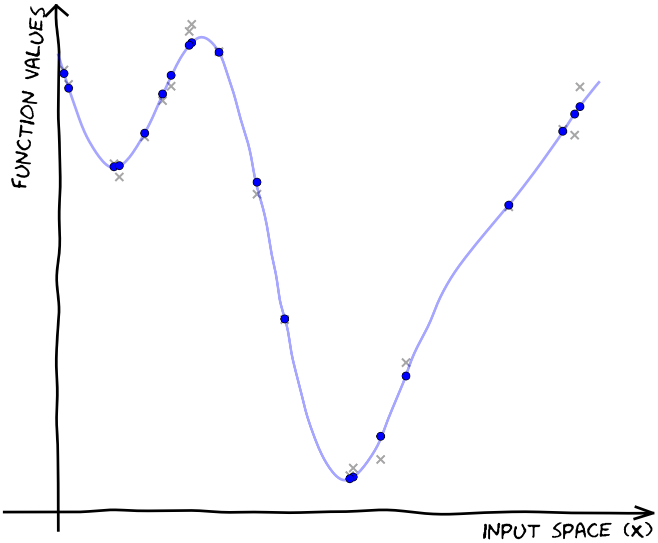
$f(\mathbf{x}) \sim \mathcal{GP}$



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

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

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

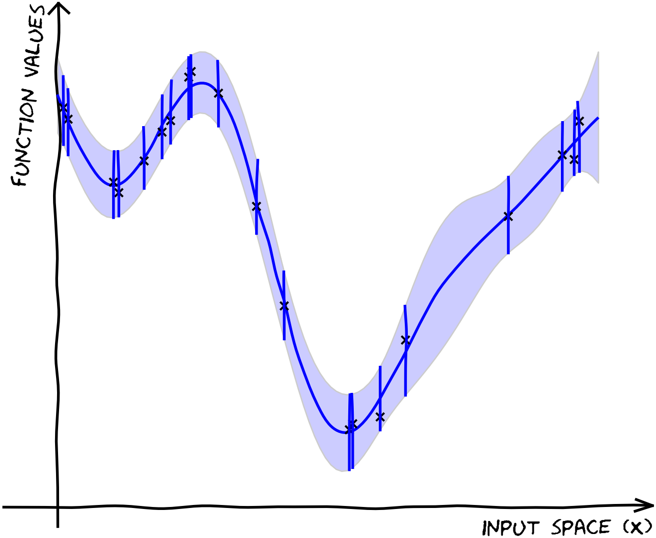


$\mathbf{X}, \mathbf{y}$

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

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

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

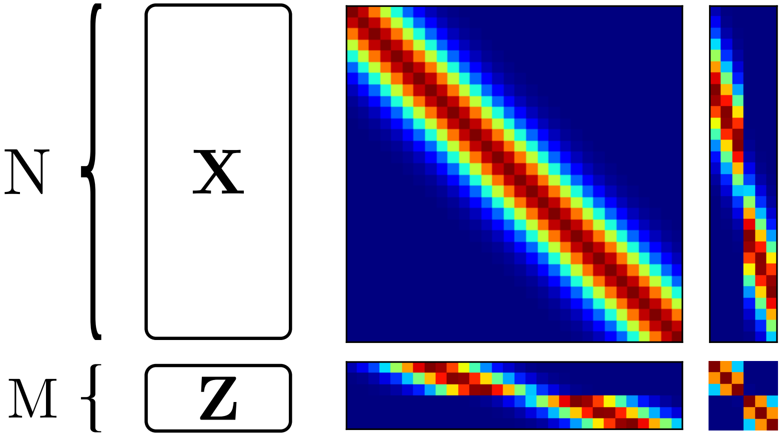


Introducing $\mathbf{u}$

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

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

Introducing $\mathbf{u}$



Introducing $\mathbf{u}$

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

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

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

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

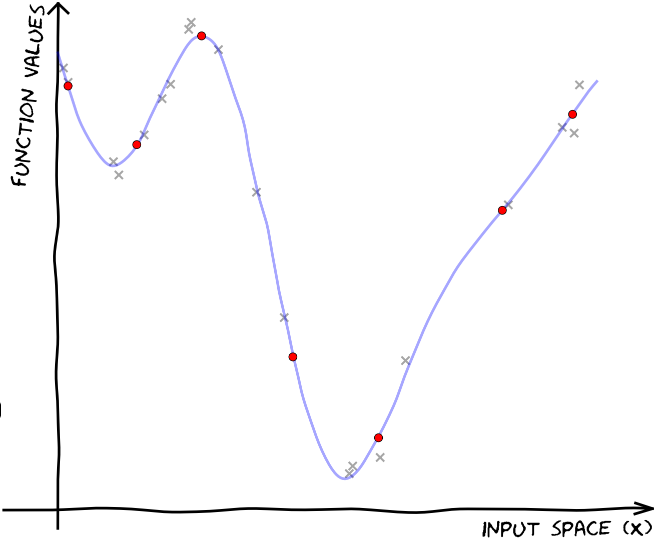
$f(\mathbf{x}) \sim \mathcal{GP}$

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

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

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

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



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

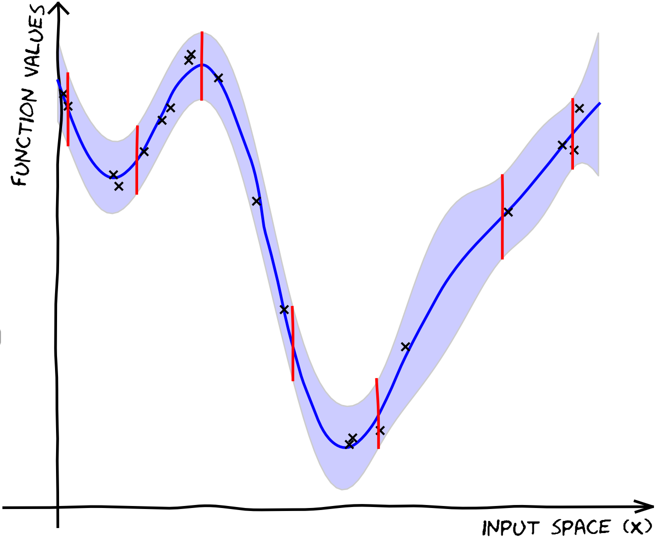
$f(\mathbf{x}) \sim \mathcal{GP}$

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

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

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

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



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 $\mathbf{K}_{\mathbf{ff}}$

Variational marginalisation of $\mathbf{f}$

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

Variational marginalisation of $\mathbf{f}$

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

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

Variational marginalisation of $\mathbf{f}$

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

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

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

Variational marginalisation of $\mathbf{f}$

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

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

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

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

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

(Titsias, 2009)

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

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

(Titsias, 2009)

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

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

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

(Titsias, 2009)

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

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

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

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

(Titsias, 2009)

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

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

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

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

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

A Lower Bound on the Likelihood

$$\widetilde{p}(\mathbf{y}\,|\,\mathbf{u}) = \prod_{i=1}^{n} \widetilde{p}(y_i\,|\,\mathbf{u})$$

$$\widetilde{p}(y\,|\,\mathbf{u}) = \mathcal{N}\left(y|\mathbf{k}_{f\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{u}, \sigma^2\right) \exp\left\{-\frac{1}{2\sigma^2}\left(k_{ff} - \mathbf{k}_{f\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{k}_{\mathbf{u}f}\right)\right\}$$

A straightforward likelihood approximation, and a penalty term

Now we can marginalise $\mathbf{u}$

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

- Computing the posterior costs $\mathcal{O}(nm^2)$
- We also get a lower bound of the marginal likelihood

What does the penalty term do?

$$\sum_{i=1}^{n} -\frac{1}{2\sigma^2}\left(k_{ff} - \mathbf{k}_{f\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{k}_{\mathbf{u}f}\right)$$

It doesn't affect the posterior

It appears on the top and bottom of Bayes' rule

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

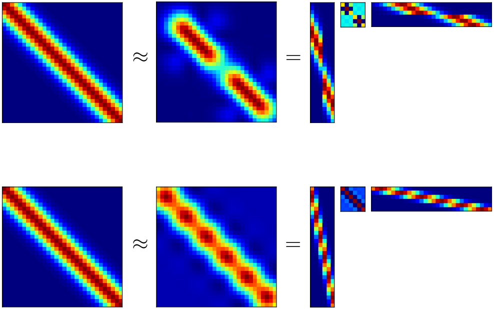
What does the penalty term do?

$$\sum_{i=1}^{n} -\frac{1}{2\sigma^2}\left(k_{ff} - \mathbf{k}_{f\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{k}_{\mathbf{u}f}\right)$$

It affects the marginal likelihood

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

What does the penalty term do?



How good is the inducing approximation?

It's easy to show that as $\mathbf{Z} \to \mathbf{X}$:

- $\mathbf{u} \to \mathbf{f}$ (and the posterior is exact)
- The penalty term is zero.
- The cost returns to $\mathcal{O}(n^3)$

How good is the inducing approximation?

It's easy to show that as $\mathbf{Z} \to \mathbf{X}$:

- $\mathbf{u} \to \mathbf{f}$ (and the posterior is exact)
- The penalty term is zero.
- The cost returns to $\mathcal{O}(n^3)$

- We're okay if we have sufficient coverage with $\mathbf{Z}$
- We can optimize $\mathbf{Z}$ along with the hyperparameters

Predictions

In a 'full' GP, we did

$$p(f_\star \mid \mathbf{y}) = \int p(f_\star \mid \mathbf{f}) p(\mathbf{f} \mid \mathbf{y}) \mathrm{d}\mathbf{f}$$

In a induced GP, we do

$$p(f_\star \mid \mathbf{y}) = \int p(f_\star \mid \mathbf{u}) \widetilde{p}(\mathbf{u} \mid \mathbf{y}) \mathrm{d}\mathbf{u}$$

Recap

So far we:

- introduced $\mathbf{Z}, \mathbf{u}$
- approximated the intergral over $\mathbf{f}$ variationally
- captured the information in $\widetilde{p}(\mathbf{u} \,|\, \mathbf{y})$
- obtained a lower bound on the marginal likeihood
- saw the effect of the penalty term
- prediction for new points

Omitted details:

- optimization of the covariance parameters using the bound
- optimization of Z (simultaneously)
- the form of $\widetilde{p}(\mathbf{u} \,|\, \mathbf{y})$
- historical approximations

Other approximations

Subset selection (Lawrence et al., 2003)

- Random or systematic
- Set $\mathbf{Z}$ to subset of $\mathbf{X}$
- Set $\mathbf{u}$ to subset of $\mathbf{f}$
- Approximation to $p(\mathbf{y}|\mathbf{u})$:
  - $p(\mathbf{y}_i|\mathbf{u}) = p(\mathbf{y}_i|\mathbf{f}_i) \qquad i \in \text{selection}$
  - $p(\mathbf{y}_i|\mathbf{u}) = 1 \qquad i \notin \text{selection}$

Other approximations

(Quiñonero Candela and Rasmussen, 2005) Deterministic Training Conditional (DTC)

- Approximation to $p(\mathbf{y} \mid \mathbf{u})$:
  - $\widetilde{p}(\mathbf{y}_i \mid \mathbf{u}) = \delta(\mathbf{y}_i, \mathbb{E}[\mathbf{f}_i \mid \mathbf{u}])$
- As our variational formulation, but without penalty

Optimization of $\mathbf{Z}$ is difficult

Other approximations

Fully Independent Training Conditional (Snelson and Ghahramani, 2006)

- Approximation to $p(\mathbf{y} \mid \mathbf{u})$:
- $p(\mathbf{y} \mid \mathbf{u}) = \prod_i p(\mathbf{y}_i \mid \mathbf{u})$

Optimization of $\mathbf{Z}$ is still difficult, and there are some weird heteroscedatic effects

References I

L. Csató. *Gaussian Processes — Iterative Sparse Approximations*. PhD thesis, Aston University, 2002.

L. Csató and M. Opper. Sparse representation for Gaussian process models. In Leen et al. (2001), pages 444–450.

L. Csató and M. Opper. Sparse on-line Gaussian processes. *Neural Computation*, 14(3):641–668, 2002.

N. D. Lawrence, M. Seeger, and R. Herbrich. Fast sparse Gaussian process methods: The informative vector machine. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems*, volume 15, pages 625–632, Cambridge, MA, 2003. MIT Press.

T. K. Leen, T. G. Dietterich, and V. Tresp, editors. *Advances in Neural Information Processing Systems*, volume 13, Cambridge, MA, 2001. MIT Press.

J. Quiñonero Candela and C. E. Rasmussen. A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, 6:1939–1959, 2005.

M. Seeger, C. K. I. Williams, and N. D. Lawrence. Fast forward selection to speed up sparse Gaussian process regression. In C. M. Bishop and B. J. Frey, editors, *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, Key West, FL, 3–6 Jan 2003.

A. J. Smola and P. L. Bartlett. Sparse greedy Gaussian process regression. In Leen et al. (2001), pages 619–625.

E. Snelson and Z. Ghahramani. Sparse Gaussian processes using pseudo-inputs. In Y. Weiss, B. Schölkopf, and J. C. Platt, editors, *Advances in Neural Information Processing Systems*, volume 18, Cambridge, MA, 2006. MIT Press.

M. K. Titsias. Variational learning of inducing variables in sparse Gaussian processes. In D. van Dyk and M. Welling, editors, *Proceedings of the Twelfth International Workshop on Artificial Intelligence and Statistics*, volume 5, pages 567–574, Clearwater Beach, FL, 16-18 April 2009. JMLR W&CP 5.