Sparse GPs

James Hensman

Disclaimers!

- Contributions from many people.
- Not in chronological order.
- Notation abuse ahead.

Motivation

Inference in a GP has the following demands:

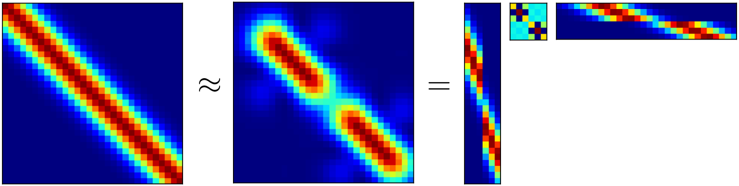
$$\begin{aligned}\text{Complexity:}\quad & \mathcal{O}(n^3)\\ \text{Storage:}\quad & \mathcal{O}(n^2)\end{aligned}$$

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!

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.

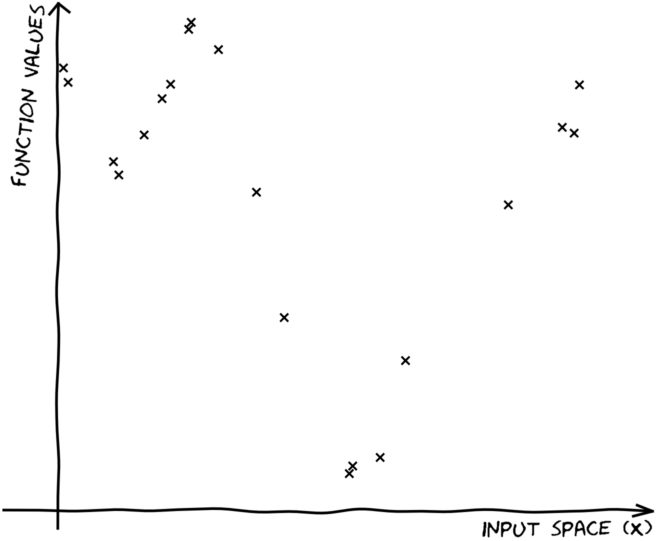
Information capture

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

- Predictions
- Marginal likelihood
- Estimating covariance parameters

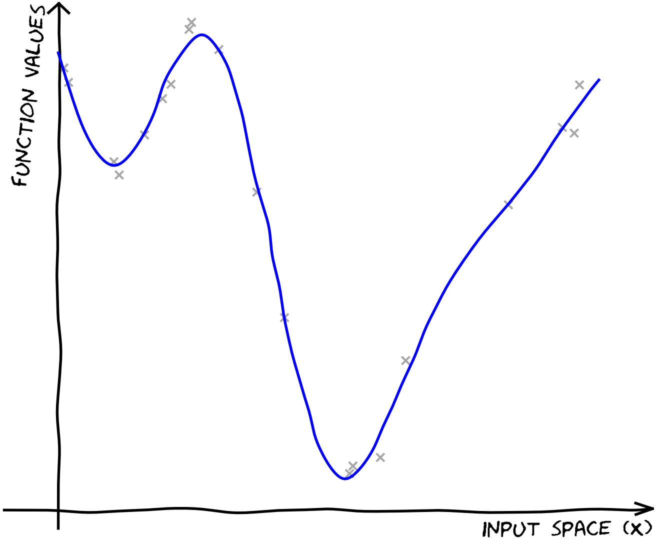
The posterior of $\mathbf{f}$ is the central object. This means inverting $\mathbf{K}_{nn}$.

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



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

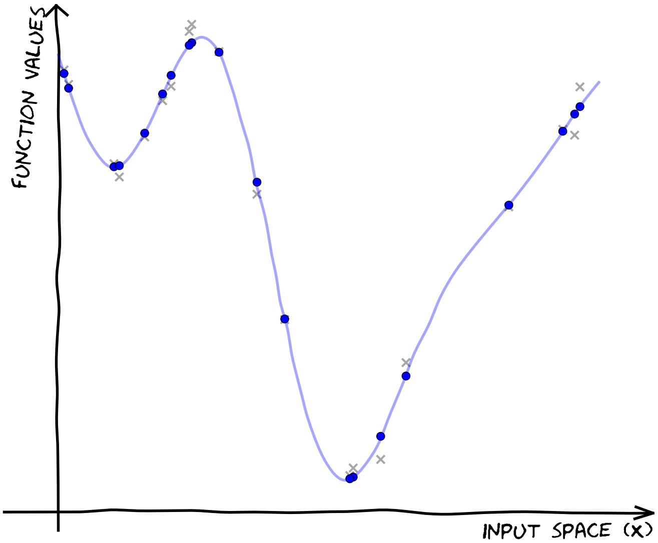
$f(\mathbf{x}) \sim \mathcal{GP}$



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

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

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

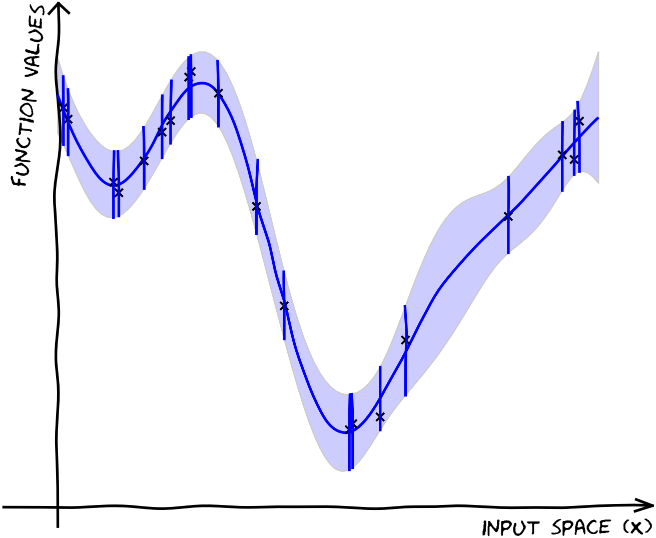


$\mathbf{X}, \mathbf{y}$

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

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

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

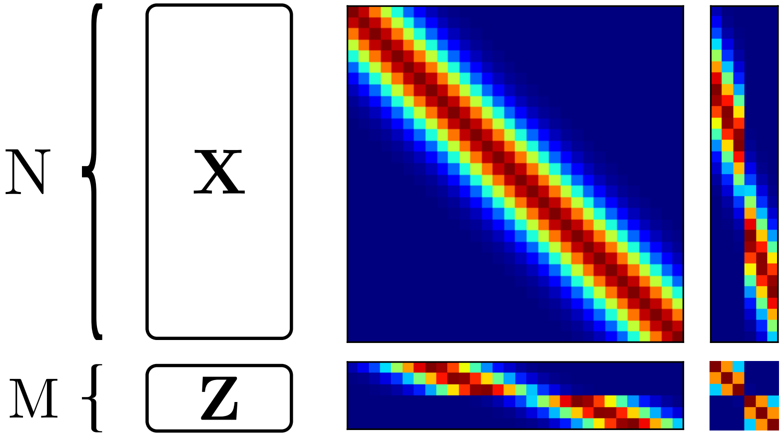


Introducing $\mathbf{u}$

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, $\mathbf{u} = f(\mathbf{Z})$.

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

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

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

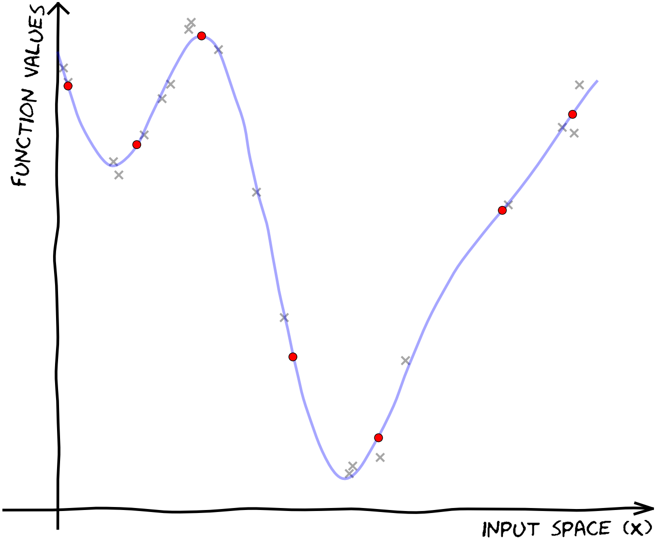
$f(\mathbf{x}) \sim \mathcal{GP}$

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

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

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

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



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

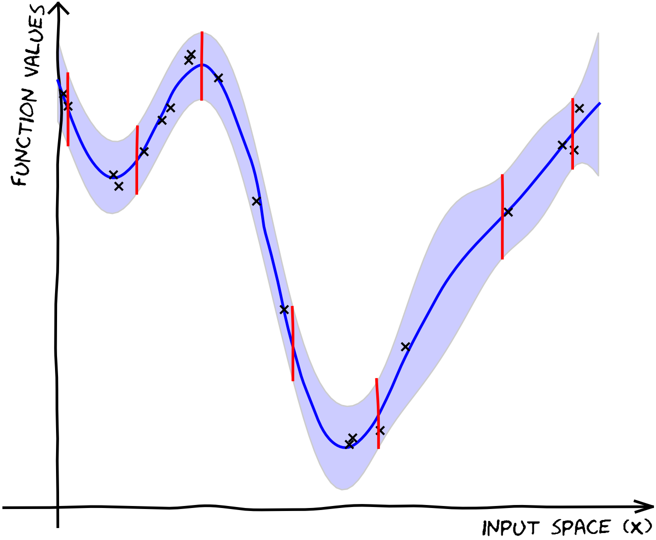
$f(\mathbf{x}) \sim \mathcal{GP}$

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

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

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

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



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 $\mathbf{K}_{nn}$

Variational marginalisation of $\mathbf{f}$

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

Variational marginalisation of $\mathbf{f}$

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

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

Variational marginalisation of $\mathbf{f}$

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

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

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

Variational marginalisation of $\mathbf{f}$

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

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

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

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

An approximate likelihood

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

A straightforward likelihood approximation, and a penalty term

Now we can marginalise $\mathbf{u}$

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

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

What does the penalty term do?

$$\sum_{i=1}^{n} -\frac{1}{2\sigma^2}\left(k_{nn} - \mathbf{k}_{mn}^{\top}\mathbf{K}_{mm}^{-1}\mathbf{k}_{mn}\right)$$

It doesn't affect the posterior

It appears on the top and bottom of Bayes' rule

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

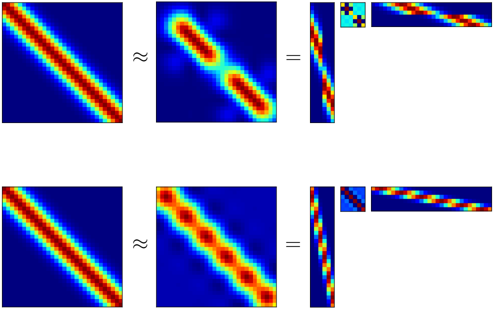
What does the penalty term do?

$$\sum_{i=1}^{n} -\frac{1}{2\sigma^2}\left(k_{nn} - \mathbf{k}_{mn}^{\top}\mathbf{K}_{mm}^{-1}\mathbf{k}_{mn}\right)$$

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 a sparse 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 a sparse 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 Z
- We can optimize Z along with the hyperparameters

Predictions

In a 'full' GP, we did

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

In a sparse GP, we do

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

Recap

So far we:

- introduced $\mathbf{Z}, \mathbf{u}$
- approximated the intergral over $\mathbf{f}$ variationally
- captured the information in $\widetilde{p}(\mathbf{u} \mid \mathbf{y})$
- 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 $\widetilde{p}(\mathbf{u} \mid \mathbf{y})$
- historical approximations

Other approximations

Subset selection

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

Selection is a combinatorial optimization problem!

Other approximations

Deterministic Training Conditional (DTC)

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

Optimization of $\mathbf{Z}$ is difficult

Other approximations

Fully independent training conditional

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

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