CUED - Engineering Tripos Part IIB 2019-2020 **Module Coursework**

Module	**4M24**	Title of report	**High-Dimensional MCMC**

Date submitted: **10/12/2020**	Assessment for this module is ☐ 100% / ☑ **25% coursework** of which this assignment forms **100** %

UNDERGRADUATE STUDENTS ONLY		**POST GRADUATE STUDENTS ONLY**			
Candidate number:	**5562E**	Name:		College:	

Feedback to the student

☐ **See also comments in the text**

		Very good	**Good**	Needs improvmt
CONTENT	**Completeness, quantity of content:** Has the report covered all aspects of the lab? Has the analysis been carried out thoroughly?			
	Correctness, quality of content Is the data correct? Is the analysis of the data correct? Are the conclusions correct?			
	Depth of understanding, quality of discussion Does the report show a good technical understanding? Have all the relevant conclusions been drawn?			
	Comments:			
PRESENTATION	**Attention to detail, typesetting and typographical errors** Is the report free of typographical errors? Are the figures/tables/references presented professionally?			
	Comments:			

Overall assessment (circle grade)		A*	**A**	B	C	D
Guideline standard		>75%	**65-75%**	55-65%	40-55%	<40%
Penalty for lateness:		*20% of marks per week or part week that the work is late.*				

Marker: Date:

4M24 CW - High-Dimensional MCMC

Candidate: 5562E

January 7, 2021

Abstract

This report outlines the result of the 4M24 coursework on high-dimensional Markov Chain Monte Carlo (MCMC).

Contents

1 Introduction

2 Simulation

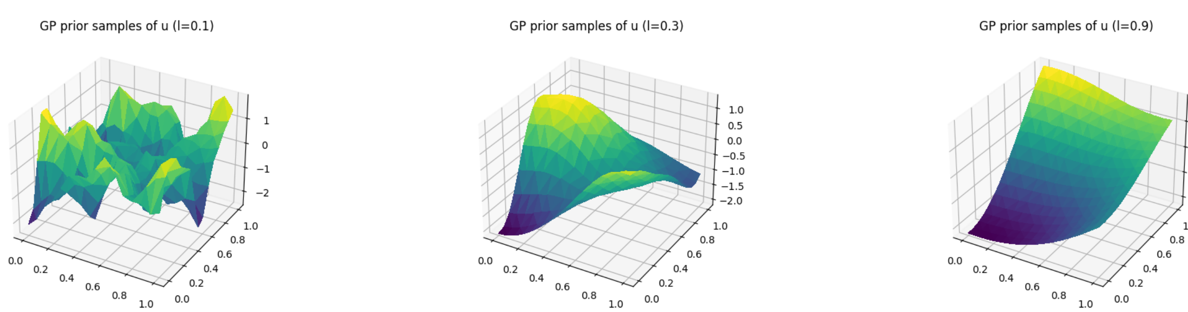
a Sampling from a Gaussian Process

We begin with a Gaussian Process (GP) defined on a 2D domain $\boldsymbol{x} \in [0,1]^2$. The realisations from this process are denoted $\boldsymbol{u} \sim \mathcal{N}(0, C)$ where $C_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ and k is the Squared Exponential (SE) covariance function with length parameter l:

$$k(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(\frac{-||\boldsymbol{x} - \boldsymbol{x}'||^2}{2l^2}\right) \tag{1}$$

If we specify the latent variables $\{\boldsymbol{x}_i\}_{i=1}^N$, then we can compute C and hence fully specify the prior on u. We choose to place $\{x_i\}_{i=1}^N$ on a $D \times D$ grid with equal spacing, starting at $(0,0)$ and ending at $(1,1)$. Obviously, we require $N = D^2$.

We can now plot the u-surface atop this grid by ensuring that for each $i \in \{1 \dots N\}$, u_i denotes the Z-position and $\boldsymbol{x}_i$ the X-Y-position. We can then investigate the effect of varying the length-scale parameter l. Three settings of l and the associated plots are given in figure 1 for $D = 16$.



(a) Small $l = 0.1$ (b) Medium $l = 0.3$ (c) Large $l = 0.9$

Figure 1: Samples from GP prior for varying length scales ($D = 16$)

We now proceed to make M random draws (denoted by the $M \times 1$ vector $\boldsymbol{v}$) from these samples $\boldsymbol{u}$ with additive Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$. The subsampling factor f is defined as $f \coloneqq N/M$. The draws can be computed as follows:

$$\boldsymbol{v} = G\boldsymbol{u} + \boldsymbol{\epsilon} \tag{2}$$

Where G is an $M \times N$ matrix with a single one in each row in a random location (without repetition) and rest zeros. The result is that the observations $\boldsymbol{v}$ are a jumbled subsample of $\boldsymbol{u}$ with additive noise $\boldsymbol{\epsilon}$. We can plot the data overlaid on the original prior samples by simply matching each entry of $\boldsymbol{v}$ back to the coordinate it was selected from. The result is plotted on figure 2.

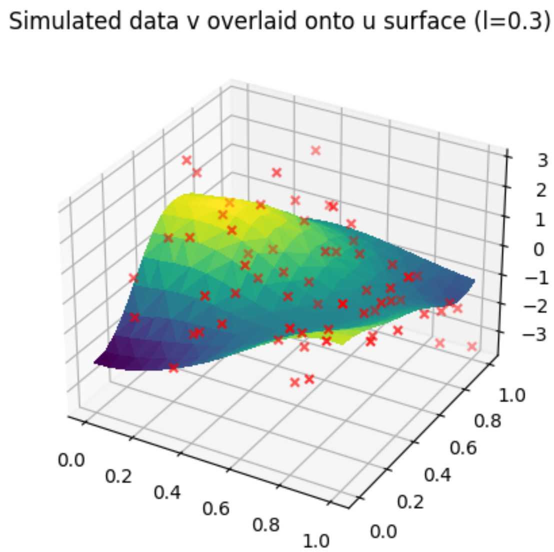


Figure 2: $\boldsymbol{v}$ (red crosses) overlaid on $\boldsymbol{u}$-surface ($f = 4, D = 16, l = 0.3$)

We observe $M = N/f = 16^2/4 = 64$ samples contained in the $\boldsymbol{v}$ vector. These are equally likely to appear above or below the $\boldsymbol{u}$-surface as the noise has zero mean. The noise variance for each data-point is of similar magnitude to the variation in the surface ($\sigma^2 = 1$) so some crosses appear relatively far away from the surface. Moreover, as the subsampling is random, the $\boldsymbol{v}$-points appear at randomly chosen (but distinct) locations X-Y plane.

b Log probabilities and MCMC

The log prior can be calculated simply, through manipulation of the Gaussian pdf:

$$\begin{aligned}
\ln p(\boldsymbol{u}) &= \ln \mathcal{N}(\boldsymbol{u}; 0, C) \\
&= \ln \frac{1}{(2\pi)^{N/2}|C|^{1/2}} \exp\left(-\frac{1}{2}\boldsymbol{u}^T C^{-1}\boldsymbol{u}\right) \\
&= -\left(\frac{N}{2}\ln 2\pi + \frac{1}{2}\ln|C| + \frac{1}{2}\boldsymbol{u}^T C^{-1}\boldsymbol{u}\right)
\end{aligned} \tag{3}$$

Likewise, $\boldsymbol{v}|\boldsymbol{u}$ is also a Gaussian such that $\boldsymbol{v}|\boldsymbol{u} \sim \mathcal{N}(G\boldsymbol{u}, I)$ (see equation 2). By comparison with the form of equation 3, we can jump straight to the log-likelihood, noting that $\ln|I| = 0$:

$$\ln p(\boldsymbol{v}|\boldsymbol{u}) = -\left(\frac{M}{2}\ln 2\pi + \frac{1}{2}(\boldsymbol{v} - G\boldsymbol{u})^T(\boldsymbol{v} - G\boldsymbol{u})\right) \tag{4}$$

Words: XX