**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 O ɴLY** | | | **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 $N = D^2$.

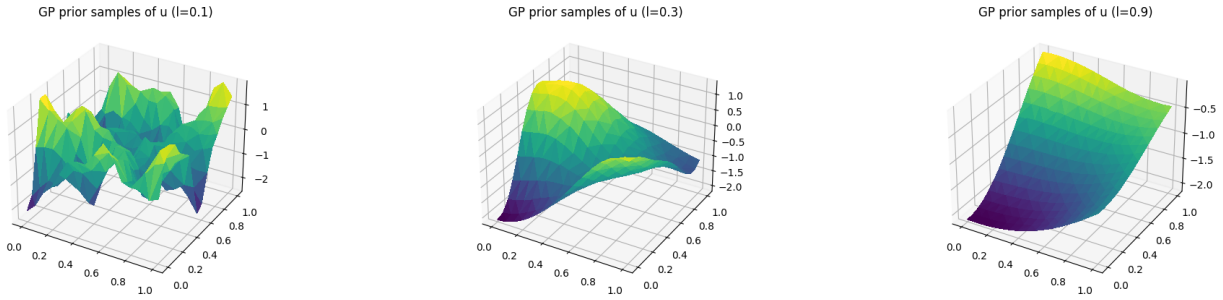
We can now plot the $u$-surface atop this grid by ensuring that for each $i \in \{1 \dots N\}$, $u_i$ denotes the Z-position and $\boldsymbol{x}_i$ the X-Y-position. We can then investigate the effect of varying the length-scale parameter $l$. Three settings of $l$ and the associated plots are given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

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.

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




(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 ($D = 16$)

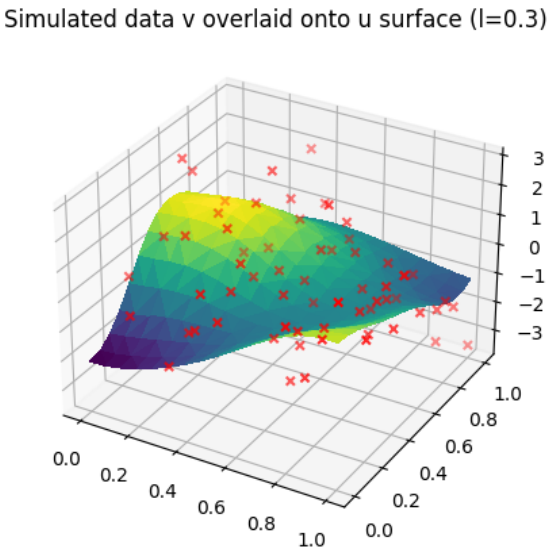



Figure 2: $\boldsymbol{v}$ (red crosses) overlaid on $\boldsymbol{u}$-surface ($f = 4, D = 16, l = 0.3$)

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

We assume that we have realised a set of observations $\boldsymbol{v}$ and it is now our job to determine probability distributions for $\boldsymbol{u}$ based on this information. As a matter of notation, we define the prior function $\pi(\cdot) \coloneqq p(\boldsymbol{u} = \cdot)$ and likelihood function $\mathbb{L}(\cdot) \coloneqq \ln p(\boldsymbol{v}|\boldsymbol{u} = \cdot)$. Likewise, we define the posterior $\rho(\cdot) \coloneqq p(\boldsymbol{u} = \cdot|\boldsymbol{v})$.

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

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

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 $\ln|I| = 0$:

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

From these it is trivial to determine the log-posterior from Bayes' rule:

$$\begin{aligned} \rho(\boldsymbol{w}) &\propto \pi(\boldsymbol{w}) \cdot \mathbb{L}(\boldsymbol{w}) \\ \therefore \ln \rho(\boldsymbol{w}) &= \ln \pi(\boldsymbol{w}) + \ln \mathbb{L}(\boldsymbol{w}) + \text{const} \end{aligned} \tag{5}$$

It is important to note that, $\boldsymbol{w}$ is simply a dummy variable.

### b.1 Gaussian Random Walk - Metropolis Hastings (GRW-MH)

Armed with the log-posterior (equation 5), observations $\boldsymbol{v}$ and observation matrix $G$, we can now apply the Gaussian Random Walk Metropolis-Hastings algorithm to draw samples of $\boldsymbol{u}$ from the posterior. We start with an initial estimate drawn from the prior $\boldsymbol{u}^{(0)} \sim p(\boldsymbol{u}) = \mathcal{N}(\boldsymbol{u}; 0, C)$. For simplicity we use the symbol $\boldsymbol{\zeta}^*$ to denote a fresh sample drawn from the prior $\pi \sim \mathcal{N}(0, C)$. It is computed by applying a Cholesky decomposition to the covariance matrix $C$ and multiplying a standard Gaussian vector $\boldsymbol{z} \sim \mathcal{N}(0, I)$. As such:

$$\boldsymbol{\zeta}^* = C^{1/2}\boldsymbol{z} \tag{6}$$

To emphasise, every time $\boldsymbol{\zeta}^*$ appears in an equation we draw a fresh sample according to equation 6. We then pick a symmetric proposal distribution to sequentially generate samples. Given a sample $\boldsymbol{u}^t$ we generate a proposal $\boldsymbol{w}^{(t)}$ as follows:

$$\boldsymbol{w}^{(t)} \coloneqq \boldsymbol{u}^{(t)} + \beta\boldsymbol{\zeta}^* \tag{7}$$

Where $\beta$ is a hyperparameter that controls the step-size of each iteration. As our proposal distribution is symmetric, the acceptance probability $\alpha^{(t)} \coloneqq \alpha(\boldsymbol{u}^{(t)}, \boldsymbol{w}^{(t)})$ simplifies to the ratio of posteriors (with an upper bound of 1):

$$\alpha(\boldsymbol{u}, \boldsymbol{w}) \coloneqq \min\left(\frac{\rho(\boldsymbol{w})}{\rho(\boldsymbol{u})}, 1\right) \tag{8}$$

It may be more natural to deal in log of this value:

$$\ln \alpha^{(t)} = \min\Big(\ln \rho(\boldsymbol{w}^{(t)}) - \ln \rho(\boldsymbol{u}^{(t)}), 0\Big) \tag{9}$$

Naturally, we can sample a uniform random variable $U \sim \mathcal{U}(0, 1)$ and compare to $\alpha^{(t)}$:

$$\begin{aligned} p(U < \alpha) &= \alpha \\ \therefore p(\ln U < \ln \alpha) &= \alpha \end{aligned} \tag{10}$$

As such the algorithm goes as follows:

**Algorithm 1** Gaussian Random Walk - Metropolis Hastings

$\boldsymbol{u}^{(0)} \leftarrow \boldsymbol{\zeta}^*$
**for** $t \in \{0, 1 \cdots T-1\}$ **do**
  $\boldsymbol{w}^{(t)} \leftarrow \boldsymbol{u}^{(t)} + \beta\boldsymbol{\zeta}^*$ ▷ Generate proposal
  $\ln \alpha^{(t)} \leftarrow \min\big(\ln \rho(\boldsymbol{w}^{(t)}) - \ln \rho(\boldsymbol{u}^{(t)}), 0\big)$
  $U^{(t)} \leftarrow\sim \mathcal{U}(0, 1)$

  **if** $\ln U^{(t)} \leq \ln \alpha^{(t)}$ **then**
    $\boldsymbol{u}^{(t+1)} \leftarrow \boldsymbol{w}^{(t)}$ ▷ Proposal accepted
  **else**
    $\boldsymbol{u}^{(t+1)} \leftarrow \boldsymbol{u}^{(t)}$ ▷ Proposal rejected
  **end if**
**end for**

As our algorithm depends only on the difference of the posteriors, there is no need to compute the constant term. This massively speeds up computation.

**Algorithm 2** preconditioned Crank-Nicholsons

$\boldsymbol{u}^{(0)} \leftarrow \boldsymbol{\zeta}^*$
**for** $t \in \{0, 1 \cdots T-1\}$ **do**
  $\boldsymbol{w}^{(t)} \leftarrow \sqrt{1-\beta^2}\boldsymbol{u}^{(t)} + \beta\boldsymbol{\zeta}^*$ ▷ Generate proposal
  $\ln \alpha^{(t)} \leftarrow \min\left(\ln \mathbb{L}(\boldsymbol{w}^{(t)}) - \ln \mathbb{L}(\boldsymbol{u}^{(t)}), 0\right)$
  $U^{(t)} \leftarrow\sim \mathcal{U}(0,1)$

  **if** $\ln U^{(t)} \leq \ln \alpha^{(t)}$ **then**
    $\boldsymbol{u}^{(t+1)} \leftarrow \boldsymbol{w}^{(t)}$ ▷ Proposal accepted
  **else**
    $\boldsymbol{u}^{(t+1)} \leftarrow \boldsymbol{u}^{(t)}$ ▷ Proposal rejected
  **end if**
**end for**

### b.2 preconditioned Crank-Nicholson (pCN)

Preconditioned Crank-Nicholson is rather similar, except we choose a subtly different proposal distribution:

$$\boldsymbol{\omega}^{(t)} \coloneqq \sqrt{1-\beta^2}\boldsymbol{u}^{(t)} + \beta\boldsymbol{\zeta}^* \tag{11}$$

This changes the form of our acceptance probability subtly by exchanging the posterior $\rho$ for the likelihood $\mathbb{L}$. The rest of the algorithm is broadly unchanged:

One advantage is that we do not need to compute the prior probability for each proposal, speeding up computation.

## c Probit classification

## d Length scale inference

# 3 Spatial Data

## a Poisson Likelihood

## b Monte Carlo Estimation

**Words**: XX