Assignment 2, Question 1 MAST90125: Bayesian Statistical Learning

Due: Friday 20 September 2019

There are places in this assignment where R code will be required. Therefore set the random seed so assignment is reproducible.

```
set.seed(123456)  #Please change random seed to your student id number.
```

Question One (12 marks)

In generalised linear models, rather than estimating effects from the response data directly, we model through a link function, $\eta(\boldsymbol{\theta})$, and assume $\eta(\boldsymbol{\theta})_i = \mathbf{x}_i'\boldsymbol{\beta}$. The link function can be determined by re-arranging the likelihood of interest into the exponential family format,

$$p(y|\boldsymbol{\theta}) = f(y)g(\boldsymbol{\theta})e^{\eta(\boldsymbol{\theta})'u(y)}. \tag{1}$$

a) Re-arrange the Poisson probability mass function into the exponential family format to determine the canonical link function. The Poisson pmf is

$$Pr(y|\lambda) = \frac{\lambda^y e^{-\lambda}}{y!}.$$

To explore some properties of Metropolis sampling, consider the dataset `Warpbreaks.csv`, which is on LMS. This dataset contains information of the number of breaks in a consignment of wool. In addition, Wool type (A or B) and tension level (L, M or H) was recorded.

b) Fit a Poisson regression to the warpbreak data, with Wool type and tension treated as factors using the function glm in R. Report co-efficient estimates and the variance-covariance matrix.

c) Fit a Bayesian Poisson regression using Metropolis sampling. Assume flat priors for all coefficients. Extract the design matrix $\mathbf{X}$ from the `glm` fitted in a). For the proposal distribution, use a Normal distribution with mean θ^{t-1} and variance-covariance matrix $c^2\hat{\boldsymbol{\Sigma}}$ where $\boldsymbol{\Sigma}$ is the variance-covariance matrix from the glm fit. Consider three candidates for c, $1.6/\sqrt{p}, 2.4/\sqrt{p}, 3.2\sqrt{p}$, where p is the number of parameters estimated. Run the Metropolis algorithm for 10,000 iterations, and discard the first 5,000. Report the following:

- Check, using graphs and appropriate statistics, that each chain converges to the same distribution. To do this, you may find installing the R package `coda` helpful.
- The proportion of candidate draws that were accepted.
- The effective sample size for each chain.
- What do you think is the best choice for c. Does this match the results stated in class on efficiency and optimal acceptance rate?