STA4502S (Monte Carlo Estimation), Winter 2013

Homework Assignment: worth 50% of final course grade.

Due: In class by 11:10 a.m. **sharp** on Friday February 8.

GENERAL NOTES:

- **Late homeworks, even by one minute, will be penalised!**
- Include at the top of the first page: Your name and student number and department and program and year and e-mail address.
- Homework assignments are to be solved by each student individually. You may discuss assignments in general terms with other students, but you must solve it on your own, including doing all of your own computing and writing.
- For full points you should provide very complete solutions, including explaining all of your reasoning clearly and neatly, performing detailed Monte Carlo investigations including multiple runs as appropriate, justifying all of the choices you make, etc.
- You may use results from lecture, but clearly state when you are doing so.
- When writing computer programs for homework assignments:
  - R is the "default" computer programming language and should normally be used for homework; it may be permitted to use other standard computer languages, but only with prior permission from the instructor.
  - You should include both the complete source code and the program output.
  - Programs should be clearly explained, with comments, so they are easy to follow.
  - You should always consider such issues as the accuracy of your answers, whether you should try running the program multiple times, etc.

THE ACTUAL ASSIGNMENT:

1. Write and run a computer program to compute a Monte Carlo estimate (including standard error) of $\mathbf{E}((Y+Z)/(1+|Z|))$, where $Y \sim$ Exponential(3) and $Z \sim$ Normal(0, 1) are independent.

2. Re-write the integral

$$I := \int_1^\infty \left(\int_{-\infty}^\infty (1+x^2+\sin(x))^{-|y|^3-2}\, dy \right) dx$$

as some expected value, and then estimate I using a Monte Carlo algorithm.

For the next four questions, let A, B, C, and D be the last four digits of your student number. (So, for example, if your student number were 840245070*, then $A = 5$, $B = 0$,

* (Historical note: this was the instructor's actual student number when he was a UofT undergraduate student in 1984–88.)

$C = 7$, and $D = 0$.) And, let $g : \mathbf{R}^5 \to [0, \infty)$ be the function defined by:

$$g(x_1, x_2, x_3, x_4, x_5) \;=\; (x_1+A+2)^{x_2+3}\Big(1+\cos[(B+3)x_3]\Big)(e^{(12-C)x_4})|x_4-3x_5|^{D+2}\prod_{i=1}^{5}\mathbf{1}_{0<x_i<2}\,,$$

and let $\pi(x_1, x_2, x_3, x_4, x_5) = c\, g(x_1, x_2, x_3, x_4, x_5)$ be the corresponding five-dimensional probability density function, with unknown normalising constant c. Finally, let f be the uniform density on $[0, 2]^5$.

3. Identify the values of A, B, C, and D. (This should be easy!)

4. Write and run a program to estimate $\mathbf{E}_\pi[(X_1 - X_2)/(1 + X_3 + X_4X_5)]$ by using an importance sampler with the above f. Discuss the extent to which this algorithm works well.

5. Write and run a program to estimate $\mathbf{E}_\pi[(X_1 - X_2)/(1 + X_3 + X_4X_5)]$ by using a rejection sampler with the above f. Discuss the extent to which this algorithm works well.

6. Write and run a program to estimate $\mathbf{E}_\pi[(X_1 - X_2)/(1 + X_3 + X_4X_5)]$ using an MCMC algorithm of your choice, and obtain the best estimate you can. Include some discussion of accuracy, uncertainty, standard errors, etc.

7. Consider an independence sampler algorithm on $\mathcal{X} = (1, \infty)$, where $\pi(x) = 5\,x^{-6}$ and $q(x) = r\,x^{-r-1}$ for some choice of $r > 0$, with identity functional $h(x) = x$.

(a) For what value of r will the algorithm provide i.i.d. samples?

(b) For what values of r will the sampler be geometrically ergodic?

(c) For $r = 1/20$, find a number n such that $D(x, n) < 0.01$ for all $x \in \mathcal{X}$.

(d) Write and run a computer program to estimate $\mathbf{E}_\pi(h)$ with this algorithm in the two cases $r = 1/20$ and $r = 10$, each with $M = 10^5$ and $B = 10^4$. Estimate the corresponding standard errors by two different methods: (i) using "varfact", and (ii) from repeated independent runs.

(e) Discuss and compare the standard errors estimated by each of the two methods in each of the two cases, including discussion of which method is "better" for assessing uncertainty, and which case is a "better" sampling algorithm.

8. Consider the standard variance components model described in lecture, with $K = 6$ and $J_i \equiv 5$, and $\{Y_{ij}\}$ the famous "dyestuff" data (from the file "Rdye"), with prior values $a_1 = a_2 = a_3 = b_1 = b_2 = b_3 = 100$. Estimate (as best as you can, together with a discussion of accuracy etc.) the posterior mean of W/V, in each of three ways:

(a) With a random-walk Metropolis algorithm.

(b) With a Metropolis-within-Gibbs algorithm.

(c) With a Gibbs sampler. [Note: first derive from scratch all of the conditional distributions, whether or not they were already described in lecture.]

(d) Finally, discuss the relative merits of all three algorithms for this example.