STA4502S (Monte Carlo Estimation), Winter 2016

Homework Assignment: worth 60% of final course grade.

Due: In class at 11:15 a.m. **sharp** on Friday February 26.

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 and error estimates as appropriate, justifying all of the choices you make, etc.
- You may use results from lecture, but clearly indicate when you do so.
- When writing computer programs for homework assignments:
  - R is the "default" computer programming language and should normally be used for homework; you may perhaps use other standard computer languages with prior permission from the instructor.
  - You should include both complete source code and program output.
  - Programs should be clearly explained, with comments, so they are easy to follow.
  - You should always consider such issues as the accuracy and consistency and so on of the answers you obtain.

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 $Y \sim \text{Exponential}(3)$ and $Z \sim \text{Normal}(0,1)$ 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 $A = 5$, $B = 0$, $C = 7$, 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.

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

(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. Let $\mathcal{X} = \mathbf{R}$, and let $\pi(x) = c\, g(x)$, where $g(x) = e^{-|x|/10}(1 + \cos(x)\sin(x^3))$, and let $h(x) = x + x^2$. With appropriate choice of M and B and σ and starting distribution $\mathcal{L}(X_0)$, estimate $\mathbf{E}_\pi(h)$ in each of two different ways:

(a) With a usual random-walk Metropolis algorithm for π, with the usual proposal distributions $Y_n \sim N(X_{n-1}, \sigma^2)$.

(b) With a Langevin (Metropolis-Hastings) algorithm with proposals $Y_n \sim N(X_{n-1} + \frac{1}{2}\sigma^2\, g'(X_{n-1}) \,/\, g(X_{n-1}),\, \sigma^2)$. [Note: Here $g'(x)$ is the usual derivative of g, and should be computed analytically by you in advance (show your work) and entered into your program.]

(c) Compare the two algorithms and discuss which one is "better".

9. 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.

10. (Mini-project) This final question is actually a mini-project, and may be summarised as follows: find an interesting and challenging quantity to compute, and conduct a Monte Carlo investigation to compute it.

More specifically, your assignment for this question is as follows:

You should begin by finding an interesting and challenging quantity to compute. The quantity could be inspired by a research paper that you read, or an application related to your own field of research, or a topic of general interest to you. You could focus on anything from statistical inference to bioinformatics applications to artificial intelligence to card shuffling to game playing to astronomy – try to be creative. The quantity does not have to be completely original, i.e. it can be related to topics discussed elsewhere, as you long as you cite this in your answer. Above all, try to choose a quantity which is challenging to compute, i.e. that simple direct computation is infeasible. (Your project should not directly repeat material from another course or project, but it could be related. If you do make any use of results or programs or ideas from other sources or other courses, then this should be clearly explained.)

Then, you should attempt to compute this quantity using various Monte Carlo methods (perhaps those developed in class, perhaps others). Ideally, for each method you should investigate its success or failure, with as much computational evidence as possible, together with whatever theoretical analysis (e.g. standard errors, or geometric ergodicity, or ...) you can manage. (It is okay if some or all of the methods fail.)

Finally, you should state your conclusions, regarding the value(s) you were trying to compute, and also regarding which Monte Carlo methods did or did not work well.

In the main part of your mini-project, your topic, motivation, methodology, and results, should all be **very clearly explained** in a brief and reasonable brief manner. Then, supplementary materials and source code and additional explanations should be put in a separate Appendix.

It is intended that students will complete this mini-project individually. However, if you wish, with advance permission, you may work in a group of 2–3 students on a correspondingly larger project; contact the instructor if you are considering this option. (This option applies only to the mini-project, not to the rest of the homework which should still be completed individually.)

Reminder: There will be a sit-down test on Wed Feb 24 during class time (location TBA), worth 40% of your final grade. No aids allowed.