Norges teknisk–
naturvitenskapelige universitet
Mathematical Sciences

Corresponding teacher:
Jo Eidsvik 90127472

EXAM IN TMA4300 COMPUTATIONAL STATISTICS

Thursday 16 May, 2013

Time: 09:00–13:00

Permitted assisting material:
Yellow A4-sheet with own handwritten notes.
Tabeller og formler i statistikk (Tapir Forlag).
K. Rottmann: *Matematisk formelsamling.*
Calculator without memory function.

ENGLISH

Results of exam: 11 June, 2013.

Problem 1 Across the lattice

a) Assume you have a routine for generating uniform numbers $U(0,1)$. Write and describe an algorithm for sampling a uniform integer between 1 and n, i.e. $X \sim U\{1,\ldots,n\}$.

Consider a regular lattice of size $m \times m$. A robot should walk from vertex $(1,1)$ to (m,m). It is only allowed to walk horizontal, vertical or diagonal edges. When the robot reaches a vertex, it takes a random direction, including the edge where it entered. For an internal vertex this means 8 equally likely directions. At the outermost vertices and corners there are less options (5 or 3), but these (5 or 3 directions) are still equally likely. Horizontal and vertical edges have distance 1, while diagonal edges have distance $\sqrt{2}$. Figure 1 illustrates this for $m = 4$.

b) Let T be the distance walked by the robot before reaching vertex (m,m). What is the probability of the event $T = (m-1)\sqrt{2}$?

The distribution of T may be explored by Monte Carlo sampling. Write the pseudo-code for drawing realizations of T.

Figure 2 shows 10000 independent Monte Carlo samples of T in the situation with $m = 4$. The sample distances are here sorted from smallest to largest.

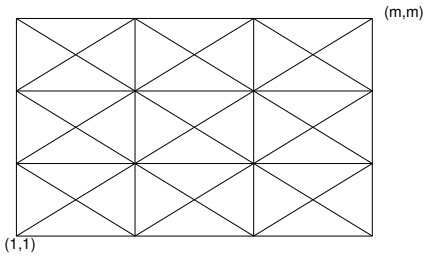

Figure 1: Illustration of vertices and edges. The robot walks from $(1, 1)$ to (m, m).

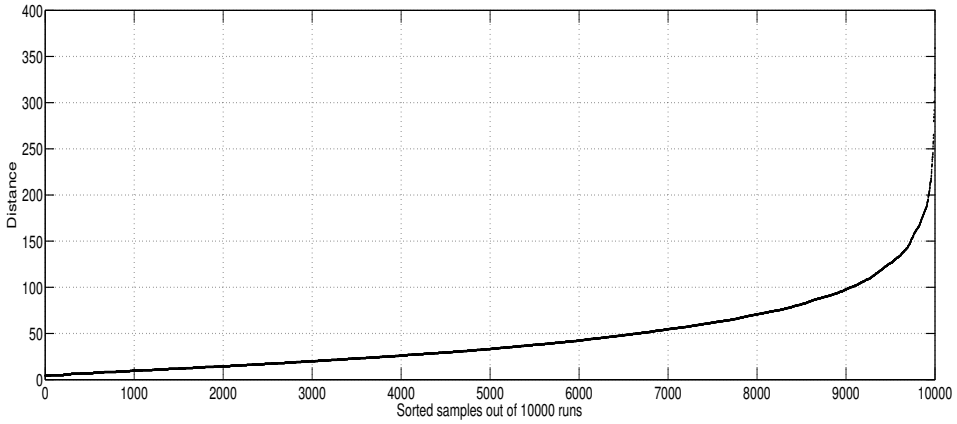

Figure 2: Results of 10000 independent Monte Carlo realization of the walk distance T for $m = 4$. Plot of sorted distances.

c) Use the plot to estimate the probability of $p = P(T > 100)$. What is the variance of the Monte Carlo estimate of p? Approximately how many Monte Carlo samples would ensure that a 90 percent confidence interval for p has length 0.001?

What are the challenges when approximating $P(T > 300)$ here? Discuss and suggest specific Monte Carlo schemes for improved approximation of this probability.

Problem 2 Change point determination

Data $y_1, \ldots, y_n$ for $n = 100$ are plotted in Figure 3.

a) Assume first that data are modeled as follows: $y_i = \mu + \sigma\epsilon_i$, $i = 1, \ldots, 100$, where $\epsilon_i \sim N(0, 1)$ are Gaussian independent errors. The precision is defined by $q = 1/\sigma^2$.

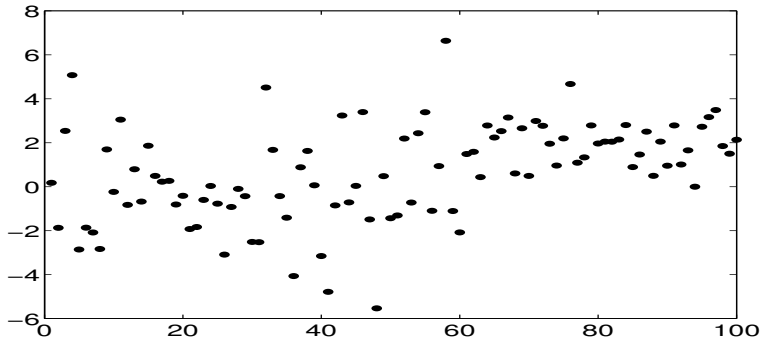

Figure 3: Plot of the 100 data values y_i, $i = 1, \ldots, 100$.

Let further $\mu \sim N(\eta, (1/r))$ and $q \sim \text{Gamma}(\alpha, \beta)$ be independent variables a priori, where $\eta = 0$, $r = 1$, $\alpha = 1$ and $\beta = 1$ are fixed. The Gamma density is defined by $p(q) \propto q^{\alpha-1}e^{-\beta q}$.

Show that the full conditional distribution of μ is Gaussian with variance $1/(r+qn)$ and mean $(r\eta + q\sum_{i=1}^{n} y_i)/(r+qn)$.

Show that the full conditional distribution of q is $\text{Gamma}(\alpha+n/2, \beta+(1/2)\sum_{i=1}^{n}(y_i-\mu)^2)$.

One suspects that the data gathering scheme was modified after $t < n$ of the data were acquired. An alternative model for the data is then as follows: $y_i = \mu_1 + \sigma_1\epsilon_i$, $i = 1, \ldots, t$, and $y_i = \mu_2 + \sigma_2\epsilon_i$, $i = t+1, \ldots, n$. Here $\epsilon_i \sim N(0,1)$ are Gaussian independent errors, and the precisions are defined by $q_1 = 1/\sigma_1^2$ and $q_2 = 1/\sigma_2^2$. Let further $\mu_1 \sim N(\eta, (1/r))$, $\mu_2 \sim N(\eta, (1/r))$, $q_1 \sim \text{Gamma}(\alpha, \beta)$ and $q_2 \sim \text{Gamma}(\alpha, \beta)$ be independent variables a priori, with η, r, α and β set as in point a).

b) Assume that $1 < t < n$ is known. Use the results from the previous point to:

Compute the full conditionals of μ_1 and μ_2.

Compute the full conditionals of q_1 and q_2.

We next assume that t is unknown. Let t have a uniform distribution among the integers 1 to n a priori.

c) The change point t can be updated using a Metropolis-Hastings step, keeping μ_1, μ_2, q_1 and q_2 fixed. Assume we use a proposal distribution which is uniform within $\{t(b) - h, \ldots, t(b) + h\}$, where $t(b)$ is the current value of t. Derive the associated acceptance rate of a proposed variable.

We implement a Markov chain Monte Carlo sampler to draw realizations from the posterior distribution of t, μ_1, μ_2, q_1 and q_2 given the data $y_1, \ldots, y_{100}$. One iteration of the Markov chain Monte Carlo consists of Gibbs-sampling from the full conditionals from point b) for μ_1, q_1, μ_2 and q_2, and a Metropolis-Hastings step for t, as described in point c). Figure 4 shows traceplots of 10000 updates.

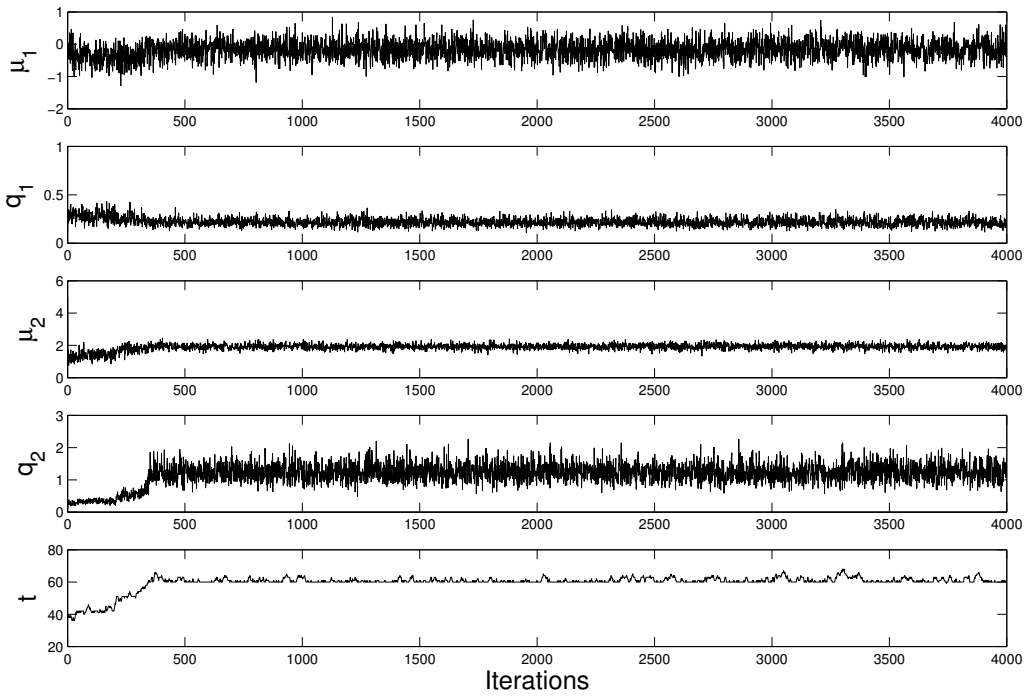

Figure 4: Plot of variables as a function of the Markov chain Monte Carlo iterations.

d) What is the burn-in time for this Markov chain Monte Carlo algorithm?

The implementation uses $h = 1$ in the uniform proposal distribution for t. What effects would a larger h give?

e) Construct a useful *joint* proposal distribution for t, μ_1, μ_2, q_1 og q_2. Compute the associated accept probability?

Problem 3 Bootstrap

Data $y_1, \ldots, y_{11}$ are available from an experiment. The empirical mean is $\bar{y} = (1/11)\sum_{i=1}^{11} y_i = 9.46$.

a) Explain the boostrap idea to approximate the sampling distribution of $\bar{y}$.

Write pseudo-code for drawing B bootstrap replicates to approximate this sampling distribution.

Using $B = 100$ we get the following sorted bootstrap replicates of the empirical mean
$3.11, 3.37, 3.82, 4.37, 4.41, 4.62, 4.64, \ldots, 14.60, 14.89, 15.88, 16.08, 16.28, 17.89, 17.90$.

b) Use these bootstrap results directly to compute an approximate 90 percent confidence interval for the mean.

Classical confidence intervals for the mean are $(\bar{y} \pm t_{10,0.95} s/\sqrt{n})$ or $(\bar{y} \pm z_{0.95} s/\sqrt{n})$, where $t_{v,\alpha}$ and z_α are the α percentiles of the t-distribution with v degrees of freedom and the standard normal, respectively. Further, the empirical variance is $s^2 = (1/10)\sum_{i=1}^{11}(y_i - \bar{y})^2$.

What are the assumptions underlying these classical results compared to those of bootstrapping?