Introduction

TMA4300: Computer Intensive Statistical Methods
(Spring 2016)

Andrea Riebler

1

[1] Slides are based on lecture notes kindly provided by Håkon Tjelmeland.

General information

Lectures:

- Lectures: Andrea Riebler, room 1242
  andrea.riebler@math.ntnu.no
- Tuesday 10:15-12:00 S5,
  Thursday: 16:15-18:00 R10

Computer exercises:

- Exercises: Xin Luo, room 1026, xinlu@math.ntnu.no
- Tuesday 10:15-12:00 Nullrommet 380A,
  Thursday: 16:15-18:00 Nullrommet 380A
- Extra hour: Monday 14:15-15:00 EL4

See course webpage regularly for time plan!

Access to computer lab Nullrommet 380A

For those who do not have access to Nullrommet 380A: please send me (andrea.riebler@math.ntnu.no) as soon as possible your name, student number, NTNU username/e-mail address and study programme.

TMA4300: Course webpage

https://wiki.math.ntnu.no/tma4300/2015v/start

Please check this website regularly!

- Messages
- Course information
- Curriculum
- Lecture plan
- Exercise classes
- Statistical software
- Reference group
- Exam

Quality ensurance: Reference group

Three members preferably from different study programmes, such as Industrial Mathematics, Erasmus Programme, others.

Duties:

- Stay in dialogue with all students.
- Participate in three meetings distributed over the semester.
- Give feedback to lecturer and teaching assistant about lectures and exercises, and provide suggestions for improvement.
- Write a joint final report providing constructive feedback and evaluation of the course. This report will be published unedited in course evaluation.

A certificate will prove participation in the reference group.

Volunteers?

Course outline

Reference book:

Givens, G.H., Hoeting, J.A., 2013, *Computational Statistics*, 2nd edition, John Wiley & Sons.

The book is freely available as e-book within the NTNU network from the library.

Extra references might be used.

The course is divided in three topic blocks:

Part 1: Algorithms for stochastic simulation

Part 2: Markov chains Monte Carlo methods and INLA

Part 3: Expectation-maximisation algorithms, bootstrap and classification methods

Exercises

- Each lecture block is followed by an exercise block.
- Exercises are obligatory. If you did the exercises in a previous year and you were admitted to the exam, you will still be admitted to the exam this year. However, the points you obtained will not be transferred. Thus, you get credit only for exercise work made during this semester.
- The exercises account for 30% of the final mark. The final exam counts for 70%. You must pass the final exam to pass the course (exercises + final exam).

Exercises

Each lecture block is followed by an exercise block

- The exercises MUST be done in groups of two persons.

Register your group until **26 January** by sending an email to Xin (`xinlu@math.ntnu.no`). If you need help to find a team partner please send also an email to Xin.

- The exercises have to be done using the statistical package `R`.
- The exercises account for 30% of the final mark.
- In an oral presentation each group will once present their finding on a selected part of the exercises.

Statistical software R

R is available for free download at The Comprehensive R Archive Network (Windows, Linux and Mac).

- Rstudio `http://www.rstudio.org` is an integrated development environment (system where you have your script, your run-window, overview of objects and a plotting window).
- A nice introduction to R is the book P. Dalgaard: Introductory statistics with R, 2nd edition, Springer which is also available freely to NTNU students as an ebook.

Exam

- The exam will be on 01.06.2016 at 09:00.
- Examination aids: C (to be decided on)
- The exam counts for 70% of the final mark and must be passed in order to pass the course.

TMA4300: Topics of the course

- Simulation & Bayesian inference:
  - Generation from standard parametric families
  - Inverse cumulative distribution function
  - Rejection sampling
  - ...
- Markov chain Monte Carlo
  - Metropolis-Hastings algorithm
  - Gibbs-Sampling
  - Implementation & Output analysis
  - Approximate Bayesian inference
- Expectation-Maximisation (EM) algorithms, bootstrap, classification

TMA4300: Learning outcome, Knowledge

- The student knows computational intensive methods for doing statistical inference.
- This includes direct and iterative Monte Carlo simulations, as well as the expectation-maximisation algorithm and the bootstrap.
- The student has basic knowledge in how hierarchical Bayesian models can be used to formulate and solve complex statistical problems.
- Finally, the student understands a number of classification techniques.

TMA4300: Learning outcome, Skills

- The student can apply computational methods, such as Monte Carlo simulations, the expectation-maximisation algorithm and the bootstrap, on simple applied problems.
- General competence. The student is able to give an oral presentation where he or she communicate his or her findings in a project.

Do you have experience with ...

- Markov chains
- The computer language R
- Basic Bayesian inference

The word simulation ...

... refers to the treatment of a real problem through reproduction in an environment controlled by the experimenter.

Gamerman & Lopes, Markov Chain Monte Carlo, 2nd Edition, Page 9

Motivation: Queueing problem

M/G/1 - queue:

- Customers arrive to a queueing system according to a Poisson process, i.e. interarrival times are exponentially distributed.
- One server
- Independent service times distributed according to $f(t)$.
- Queue system empty at time $t = 0$

$X(t)$ customers in queueing system at time t.

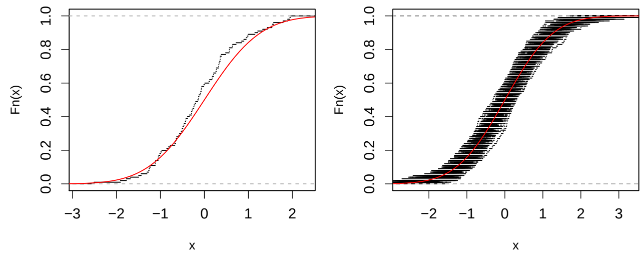
Motivation: Queueing problem (II)

If service times are exponentially distributed, $X(t)$ is a Markov process and an explicit analytical formula for the limiting distribution is available.
For general f, analytical solutions might not be available.
$\Rightarrow$ Idea: Simulate the queueing process on a computer and return the quality of interest, e.g. $\min\{t > 0 : X(t) \geq 7\}$.

Simulation, why do we need it?

Necessity to produce chance on the computer:

- Evaluation of the behaviour of a complex system (Epidemics, weather forecast, networks, economic actions, etc)
- Determine probabilistic properties of a novel statistical procedure or under an unknown distribution.



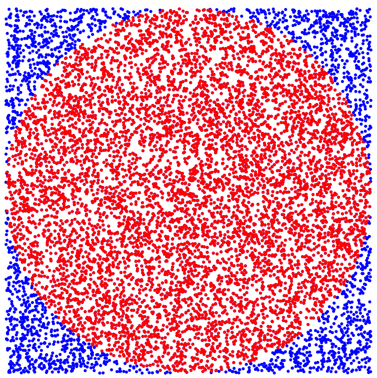
(Left: Estimation of CDF from a normal sample of 100 points,

Right: Variation of the estimation over 200 samples.)

Simulation, why do we need it?

- Approximation of an integral/area

$n = 1000$ ▷ (# of simulations)
$m = 0$ ▷ (# points in circle)
$i = 1$
while $i < n$ **do**
  $x = Rand(1), y = Rand(1)$
  if $x^2 + y^2 < 1$ **then**
    $m \leftarrow m + 1$
  end if
  $i = i + 1$
end while
return $4 \cdot m/n$



$\hat{\pi} = 3.1353.$

Pseudo-random generator

Central building block of simulation: Always requires availability of uniform $\mathcal{U}(0, 1)$ random variables.

```
[1] 0.97258602 0.66461211 0.27355502 0.67593875 0.05145283 0.90977330
[7] 0.96999184 0.93834818 0.32973574 0.01587857
```

Pseudo-random generator

A pseudo-random generator is a deterministic function f which takes a uniform random bit string as input and outputs a bit string which cannot be distinguished from a uniform random string.

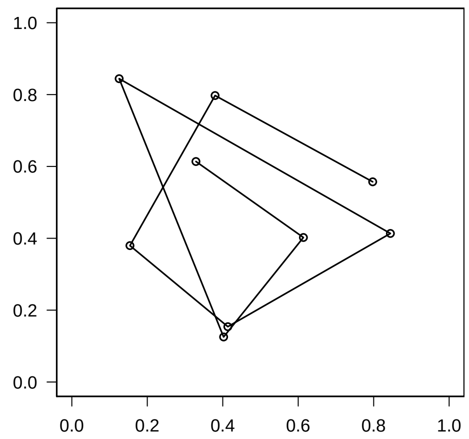
In more detail, this means that for starting value u_0 and any n, the sequence

$$\{u_0, f(u_0), f(f(u_0)), f(f(f(u_0))), \ldots, f^n(u_0)\}$$

behaves statistically like an $\mathcal{U}(0,1)$ sequence (when appropriately scaled).

Pseudo-random generator

Illustration of first 10 (u_t, u_{t+1}) steps



A standard uniform generator

The congruential random generator on $\{0, 1, \ldots, M-1\}$

$$f(x) = (a \cdot x + b) \mod M$$

has a period equal to M for proper choices (a, b) and becomes a generator on $[0, 1)$ when dividing by M.

Example

Take

$$f(x) = (69069069 \cdot x + 12345) \mod (2^{32})$$

and produce

$\ldots, 69081414, 2406887111, 1109307232, 2802677792, 3651430880, 806776992, \ldots$

i.e.

$\ldots, 0.01608427, 0.56039708, 0.25828072, 0.65254927, 0.85016500, 0.18784241, \ldots$