Theory

- Introduction to simulation
- Flow of simulation -- generating process realizations
- Random number generation from given distribution
- Collection and analysis of simulation data
  - Simulation of transient and stationary properties
  - Statistical analysis and confidence intervals
- Variance reduction techniques

Statistics collection

- The goal of simulation is evaluate the performance of the considered system. By simulation one tries to estimate the value of some parameter characterising system's performance
- This parameter may be related either to the system's **transient** behaviour
  - e.g. the average waiting time of 25 first customers entering an initially empty M/M/1 queue with a given load

  or to the **steady state**
  - e.g. the average waiting time of customers in an M/M/1 queue with a given load
- Further, this parameter may describe the situation from the point of view of the entering customers (discretely)
  - e.g. the average queue length seen by a customer entering an M/M/1

  or from the point of view of the system (continuously)
  - e.g. the average queue length in an M/M/1 queue with a given load
- One simulation run yields one observation about the parameter in question
- In order to make statistical inference we need several observations (preferably independent and identically distributed)

Simulation of transient properties (1)

- If we are studying a parameter related to the service level experienced by customers, then the simulation ends when a certain number of customers have been observed
  - e.g. if we are interested in the average waiting time of first k customers in M/M/1 queue, then the simulation is continued until the k*th* customer has arrived an entered service
- The observation X yielded by one simulation run is then the average of the waiting times W_i of the k first customers in this simulation run :

$$X = \frac{1}{k}\sum_{i=1}^{k} W_i$$

  - based on the central limit theorem, this average can be considered to be approximately normally distributed (the better the larger is k)
  - In order to make statistical inference we need several independent and identically distributed observations. These are obtained by making several simulation runs which are independent (realisations are generated with independent random number sequences).

Simulation of transient properties (2)

- If we are studying a quantity related to the system's performance, which can be observed continuously, a simulation run ends at a predefined time T
  - e.g. if we are interested in the average queue length in the interval [0,T], the (event driven) simulation is continued up to the first event occurring after time T
- The observation X yielded by one simulation run is now the time average of the queue length L(t) over the interval [0,T]

$$X = \frac{1}{T}\int_{0}^{T} L(t)dt$$

  - since the queue length is constant between events, the integral can easily be calculated as a sum of staircases (note the special handling of the last interevent interval)
  - In order to make statistical inference we need several independent and identically distributed observations. These are obtained by making several simulation runs which are independent (realisations are generated with independent random number sequences).

Simulation of steady state properties (1)

- Statistics collection is done much in the same way as in the simulation of transient properties.
- In the beginning of the simulation, however, we need a so called warm-up period (before the system reaches equilibrium), which must be discarded from the data.
- Repeated observations can now produced by at least three different ways:
  - independent repetitions
  - batch means method
  - regenerative method

Simulation of steady state properties (2)

- In the case of independent repetitions the data collection I each repetition is started after the warm-up period
  - a separate problem is to determine an appropriate length for the warm-up period
- In the batch means method one makes one long simulation run, which is (artificially) divided into pieces, each of which is considered as a separate run
  - one needs only one warm-up period, but the observations are not fully independent
- In the regenerative method, one requires that the process being simulated is regenerative. If this is the case, then one obtains independent and identically distribute observations from different regeneration periods
  - e.g. the G/G/1 queue regenerates itself always when a new customer enters an empty system
  - all Markov processes are regenerative
  - a problem is that the lengths of the regeneration periods can be very large

Transient removal

- Usually one is interested in the steady state properties of the system
- Then the initial part of the simulation, the transient, must not be included in the data collection
- The steady state is reached when the ”memory of the initial state of the system has been forgotten”
  - i.e. when the actual initial state of the system does not anymore have any impact on the distribution of the current state

- Transient removal can be done by one of the following ways
  - very long run
  - proper initialization
  - discarding warm-up period
  - batch means
  - regenerative simulation

Transient removal (continued)

- Long run
  - rough method
  - if the run is long enough, the effect of initial transient vanishes
  - a very long run may be needed -- resource wasting
  - difficult to know what is long enough

- Proper initialization
  - instead of starting the simulation from a artificial initial state (e.g. empty system), one may start the system from a state which is ”closer to an equilibrium”
    - initialize random variables in their long run average values
    - these may be approximately known from earlier simulation or on the basis of analytical considerations
  - this reduces the effect of the initial transient but does not completely remove it
  - if the equilibrium distributions of the state variables are known the initial transient can be completely removed by drawing the initial values from these distributions
    - independent drawing for each simulation run
    - usually, however, the distributions are not known (this is why simulation is needed in the first place)

Transient removal (continued)

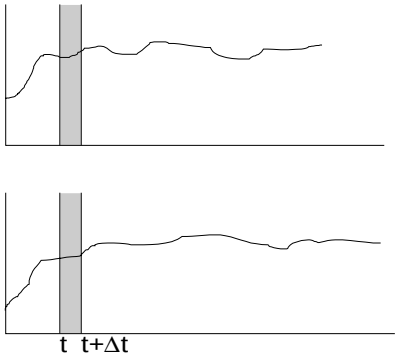
- Discarding initial data
  - a straight forward method
  - first a warm-up period is run and the data is collected first thereafter
  - problem: how long is the transient?
    a) in some cases it is known (at least approximately)
    - for instance, in an ordinary loss system the relaxation time is the same as the average holding time of a call
    - the time to be discarded is then n * holding time, where n is of the order 3...10
    - the effect of initial state has then been reduced by a factor $e^{-10} \approx 10^{-1.3...-4.3}$

    b) in general, the relaxation time is not known in advance
    - then one can use experiments
    - in repeated simulation the same length of data is discarded from all the runs
    - the length of the discard period is let grow from 0 upward and the average over the non-discarded data is drawn as a function of the length of the discard period
    - when the average does not anymore change with increasing discard period, the transient has been properly removed

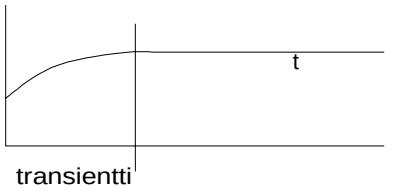
Transient removal (continued)

c) Average in a moving window
- this is another empirical method for determining the length of the transient
- in repeated simulations, the data is collected only within a relatively narrow window
- the average of the interesting quantity over the window is calculated as a function of the window location
- the function is averaged over the repetitions in order to suppress fluctuations (in a narrow window there is little data and statistical fluctuations are large)
- as a function of the window location, the considered quantity changes in the beginning and then tends to constant
- when the constant value is reached the transient is over
- often the transient is rather short and it is easy to be on the safe side: the period to be discarded can be chosen to be e.g. double the length of the empirically determined transient period

Estimating the length of the transient period by the moving window method



- One tries to get an idea of the instantaneous expectation of the interesting quantity as a function of time
- The expectation is determined as an sample average of repeated simulation runs
- The instantaneous value is estimated by the average value within a narrow window



- By inspecting the curve showing the average as the function of the window location one can assess when the transient is over

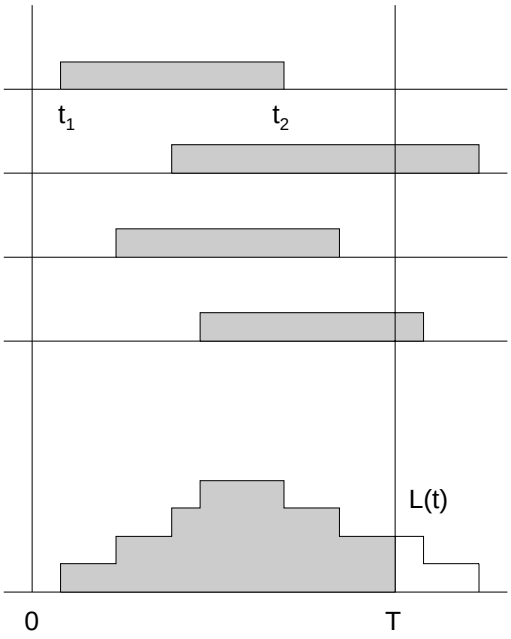
Stopping the simulation

- Simulation ends when the stopping condition is fulfilled
- One has to be careful in handling items that are still going on at the stopping time
- In considering an event based quantity (customer point of view)
  - blocking of offered calls, proportion of overflowing packets etc

  one should count only those events that have been fully handled
  - for instance, average waiting time = (sum of waiting times of those customers, who are no more waiting, i.e. who have entered service) / (the number of customers, who are no more waiting)
- In considering an time based quantity (system point of view)
  - queue length L, the proportion of time the system is in blocking state etc

  the calculation of the average has to be extended precisely to the ending time T of the simulation
  - e.g. the average queue length = $\frac{1}{T}\int_{o}^{T} L(t)dt$
  - the value of the integral can be collected by subtracting the time time of arrival upon the arrival and adding the time of departure upon the departure of a customer
  - those customers who are in the system at time T are handled as if they departed at that time

Calculating the time integral of the queue length



- The time integral of the queue length L(t) is composed of the times spent by individual customers in the system
- The time spent by a customer is the difference t_2-t_1 between the departure time t_2 and the arrival time t_1
- The integral of L(t) can accumulated by
  - subtracting t_1 from the integral upon the arrival of the customer
  - adding t_2 to it upon the departure
- When the simulation ends at time T, the integral will be evaluated correctly if all the customers in the system at that time are considered to depart at that time, i.e. T is added to the integral for each customer inside

Theory

- Introduction to simulation
- Flow of simulation -- generating process realizations
- Random number generation from given distribution
- Collection and analysis of simulation data
  - Simulation of transient and stationary properties
  - Statistical analysis and confidence intervals
- Variance reduction techniques

Estimation of a parameter

- As noted, the aim of a simulation is to give an estimate of a performance related parameter α (for instance, average time in system spent by customers or average queue length in an M/M/1 queue)
- One simulation yields about the parameter one observation X_i, which is a random variable. Observation X_i is called **unbiased** if $E[X_i] = \alpha$.
- Assume that we obtained by repeated simulation runs n n independent and identically distributed observations. Then their **average**

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is an **unbiased** and **consistent** estimator for parameter α, since

$$E[\overline{X}_n] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \alpha$$

$$D^2[\overline{X}_n] = \frac{1}{n^2}\sum_{i=1}^{n} D^2[X_i] = \frac{1}{n} D^2[X] = \frac{\sigma^2}{n} \xrightarrow{n\to\infty} 0$$

Example

- Our task is to estimate by simulations the average waiting time of 25 first customers in an M/M/1 queue with load $\rho = 0.9$, when the system at time 0 is empty.
- Analytically one can calculate the exact value: $\alpha = 2.124$
- Ten simulation runs have produced the following observations X_i (i.e. the average waiting times in the runs):
  - 1.051, 6.438, 2.646, 0.805, 1.505, 0.546, 2.281, 2.822, 0.414 and 1.307
- The average value of these

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{10}(1.051 + 6.438 + \ldots + 1.307) = 1.982$$

is the the estimate for the average waiting time (of 25 first customers) given by our simulation.

Confidence interval of an estimator (1)

- The observations X_i are approximately normally distributed
- If the observations X_i obeyed exactly the normal distribution $N(\alpha,\sigma^2)$ and if the variance $\sigma^2 = D^2[X]$ of a single observation were known, then the average of the n repeated observations had the distribution $N(\alpha,\sigma^2/n)$
- This gives the confidence interval of the estimator (the average of the observations) at confidence level 1 - β:

$$\overline{X}_n \pm z_{1-\beta/2} \cdot \frac{\sigma}{\sqrt{n}}$$

where the coefficient z_p stands for the p-fractile of the standard normal distribution N(0,1), i.e. $P\{Z \leq z_p\} = p$, where $Z \sim N(0,1)$

- Interpretation: the parameter α lies in the shown interval with probability 1 - β
- For instance, for the confidence level 95%:n we have the coefficient $z_{0.975} \approx 1.960$

Confidence interval of an estimator (2)

- In general, the variance of a single observation$\sigma^2 = D^2[X]$ is not known
- It can, however be estimated by the sample variance

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2$$

which is (in the case of independent and identically distributed observations) an unbiased estimator for the variance

- Sample standard deviation is the square root of the sample variance :

$$S_n = \sqrt{S_n^2}$$

Confidence interval of an estimator (3)

- If the observations X_i obeyed exactly the normal distribution $N(\alpha,\sigma^2)$, then the sample average, suitably normed with the sample standard deviation, would obey the Student t-distribution with n-1 degrees of freedom.
- This gives the confidence interval of the the estimator (sample average) at the confidence level 1 - β:

$$\overline{X}_n \pm t_{n-1,1-\beta/2} \cdot \frac{S_n}{\sqrt{n}}$$

where the coefficient $t_{n-1,p}$ stands for the p-fractile of the t-distribution (with n-1 degrees of freedom)

$P\{T \leq t_{n-1,p}\} = p$, where T obeys the t-distribution

- Interpretation: the parameter α lies in the shown interval with probability 1 - β
- For example, a confidence level of 95%:n corresponds to
  - the coefficient $t_{9,0.975} \approx 2.262$ in case of 10 observations
  - the coefficient $t_{100,0.975} \approx 1.984$ in case of 101 observations

Example (continued)

- The task is to estimate the average waiting time of 25 first customers entering an initially empty M/M/1 queue with load $\rho = 0.9$. Theoretical value: $\alpha = 2.124$
- Ten simulation runs have given the following observations X_i (i.e. Average waiting times in respective runs) :
  - 1.051, 6.438, 2.646, 0.805, 1.505, 0.546, 2.281, 2.822, 0.414 and 1.307
- Sample average was calculated to be 1.982 and the sample variance is

$$S_n = \sqrt{\frac{1}{9}((1.051-1.982)^2+\ldots+(1.307-1.982)^2)} = 1.781$$

- The confidence interval of our point estimator for the average waiting time of the 25 first customers with confidence level 95% is thus

$$\overline{X}_n \pm t_{n-1,1-\beta/2} \cdot \frac{S_n}{\sqrt{n}} = 1.982 \pm 2.262 \cdot \frac{1.781}{\sqrt{10}} = 1.982 \pm 1.274$$

Observations

- The estimator becomes more accurate (i.e. the confidence interval of the point estimator becomes smaller), when
  - the number n of independent simulation runs is increased
  - the variance of a single observation is decreased (e.g. by making longer individual runs or by using some variance reduction method)
- If the targeted relative accuracy of the estimator is given (i.e. the ratio of the half of the confidence interval to the sample average), one can dynamically monitor how many independent runs have to be done in order to achieve the targeted accuracy