FORMAL METHODS FOR SYSTEM VERIFICATION

Continuous Time Markov Chains

Sabina Rossi

DAIS
Università Ca' Foscari
Venezia

Stochastic Processes

Intuition

- Many physical systems, including computer and communication networks, evolve in time and often dynamic characteristics are important.
- For example, we may be interested in
  - the length of a queue
  - how long it takes for massages to be transmitted through a network
  - how much rain falls in successive years
- Stochastic processes provide models for such evolving systems by describing the sequences of states they enter.
- Formally, a stochastic model is one represented as a stochastic process.

Stochastic Processes

Definition

- A stochastic process is a set of random variables
$$\{X(t),\ t \in T\}\,.$$
- T is called the parameter space, its values are often referred to as times.
- Thus $(X(t_1), X(t_2), \ldots, X(t_n))$ has a known distribution for $t_1, t_2, \ldots t_n \in T$.
- Since we consider continuous time models $T = \mathbb{R}^{\geq 0}$, i.e., it is the set of non-negative real numbers.

State Space

- The state space $\mathcal{S}$ of a stochastic process is the set of all possible values that the random variables $X(t)$ can assume.
- Each of these values is called a state of the process.
- Any set of instances of $\{X(t),\ t \in T\}$ can be regarded as a path of a particle moving randomly in the state space $\mathcal{S}$, its position at time t being $X(t)$.
- These paths are called sample paths or realisations of the stochastic process.
- We will interpret each $X(t)$ as the state of some system at time t, and, in accordance with this interpretation, we say that the system is in state x_i at time t if $X(t) = x_i$.

Properties of Stochastic Processes

We will simply write $X(t)$ to denote the stochastic process $\{X(t),\ t \in T\}$ with $T = \mathbb{R}^{\geq 0}$.

Stationary stochastic process

- If for all $t_1, t_2, \ldots t_n \in T$ and for all $t_1 + \tau, t_2 + \tau, \ldots t_n + \tau \in T$ $(X(t_1), X(t_2), \ldots, X(t_n))$ has the same distribution as $(X(t_1 + \tau), X(t_2 + \tau), \ldots, X(t_n + \tau))$ then the stochastic process $X(t)$ is called stationary.
- In a stationary stochastic process the joint probability distribution does not change when shifted in time.

Properties of Stochastic Processes

Markov process

- The stochastic process $X(t)$ is a Markov process if for $t_0 < t_1 < \ldots < t_n < t_{n+1}$ the joint distribution of $(X(t_0), X(t_1), \ldots, X(t_n), X(t_{n+1}))$ is such that

$$Pr(X(t_{n+1}) = x_{n+1} \mid X(t_0) = x_0, X(t_1) = x_1, \ldots, X(t_n) = x_n)$$

$$= Pr(X(t_{n+1}) = x_{n+1} \mid X(t_n) = x_n)$$

- This is also called memoryless property, that is, given the value of $X(t_n)$ at some time $t_n \in T$, the future path $X(t_{n+1})$ for $t_{n+1} > t_n$ does not depend on knowledge of the past history $X(t)$ for $t < t_n$, i.e., for $t_0 < t_1 < \cdots < t_{n-1}$.

Properties of Stochastic Processes

Time Homogeneous Stochastic Processes

- A Markov process $X(t)$ is time homogeneous if

$$Pr(X(t+\tau) = x_i \mid X(t) = x_j) =$$

$$Pr(X(t'+\tau) = x_i \mid X(t') = x_j)$$

- The behaviour of the system does not depend on when it is observed. In particular, the transitions between states are independent of the time at which the transitions occur.

Properties of Stochastic Processes

Irreducible Stochastic Processes

- A Markov process is irreducible if all states in its state space $\mathcal{S}$ can be reached from all other states, by following the transitions of the process.
- If we draw a directed graph of the state space with a node for each state and an arc for each event, or transition, then for any pair of nodes there is a path connecting them, i.e., the graph is strongly connected.

Properties of Stochastic Processes

Positive recurrent Markov Processes

- A state in a Markov process, x_i, is called persistent or recurrent if the probability that the process will eventually return to x_i is 1. Otherwise the state is called transient.
- In terms of a system, the recurrent states correspond to the behaviour which is repeatedly exhibited by the system whereas transient states correspond to a behaviour which will be no longer exhibited after a certain time.
- A recurrent state x_i is termed positive-recurrent, or ergodic, if the expected number of steps until the process returns to x_i is less than infinity.
- A stochastic process is ergodic if all its states are positive-recurrent.

Assumptions

- We will consider Continuous Time Markov Processes (CTMC).
- We assume that any CTMC we deal with is
  - time homogeneous
  - irreducible
  - ergodic

Transition rate

- For a given a CTMC $X(t)$, the transition rate from state x_i to state x_j is defined by:

$$q_{ij} = lim_{\tau \to 0} \frac{Pr(X(t+\tau) = x_j \mid X(t) = x_i)}{\tau} \quad (i \neq j)$$

- The q_{ij} are also called the instantaneous transition rates.

Sojourn time

- In any stochastic process the time spent in a state is called the sojourn time.
- By the Markov property, the sojourn times are memoryless, i.e., at any time point, the distribution of the time until the next change of state is independent of the time of the previous change of state.

Exit rate and Sojourn time

- Let $\mathcal{S}$ be the state space of a Markov process $X(t)$.
- The rate of leaving state x_i, denoted by q_i, is called exit rate, and it is the sum of all the rates of the individual transistions that leave state x_i, i.e.,

$$q_i = \sum_{x_j \in \mathcal{S},\ j \neq i} q_{ij}$$

- The sojourn time in state x_i is exponentially distributed with parameter q_i and its mean is $1/q_i$.

Transition probabilities

- The transition probability p_{ij} is the probability, given that a transition out of state x_i occurs, that it is the transition to state x_j.
- By the definition of conditional probability, this is

$$p_{ij} = q_{ij}/q_i \,.$$

Infinitesimal generator matrix

- Let S be the state space of a Markov process $X(t)$ and $|S| = N$.
- The infinitesimal generator matrix $\mathbf{Q}$ of $X(t)$ is the $N \times N$ matrix where
  - each entry q_{ij} is the transition rate of moving from state x_i to state x_j
  - the diagonal entries q_{ii} are the negative row sum for row i, i.e.,

$$q_{ii} = - \sum_{x_j \in \mathcal{S},\ j \neq i} q_{ij}$$

State transition diagram

- For small Markov processes the simplest way to represent the process is often in terms of its state transition diagram.
- The nodes in a graph correspond to the states of the process.
- The arcs are labelled by the transition rates between states.
- Since every transition is assumed to be governed by an exponential distribution, the rate of the transition will be the parameter of the distribution.
- Irreducibility implies that the state transition diagram must be strongly connected.

Example 1

- Consider a mobile phone antenna whose states are represented by the number of frequencies in use.
- Suppose that there are 6 frequencies, then the state space is $\{0, 1, 2, 3, 4, 5, 6\}$.
- The only possible transition from state 0 is to state 1, caused by a new connection.
- Similarly, from state 6 the only possible transition is to state 5, caused by a frequency being released.
- For any other state j, $1 \leq j \leq 5$ there are two possible transitions, to states $j + 1$ or $j - 1$, caused by a new connection or release of an old one, respectively.

Example 1

- Assume that calls arrive according to a random variable exponentially distributed with parameter λ, i.e., the rate at which $j \to j+1$ transistions occur is λ.
- Assume that the mean duration of a connection is $1/\mu$, so that the rate at which release transitions occur on each frequency is μ.
- The state transition diagram for this process is:

Continuous Time Markov Processes

Example 1

- Assume that calls arrive according to a random variable exponentially distributed with parameter λ, i.e., the rate at which $j \to j+1$ transistions occur is λ.
- Assume that the mean duration of a connection is $1/\mu$, so that the rate at which release transitions occur on each frequency is μ.
- The state transition diagram for this process is:

$$\begin{tikzcd}
0 \arrow[r, bend left, "\lambda"] & 1 \arrow[l, bend left, "\mu"] \arrow[r, bend left, "\lambda"] & 2 \arrow[l, bend left, "2\mu"] \arrow[r, bend left, "\lambda"] & 3 \arrow[l, bend left, "3\mu"] \arrow[r, bend left, "\lambda"] & 4 \arrow[l, bend left, "4\mu"] \arrow[r, bend left, "\lambda"] & 5 \arrow[l, bend left, "5\mu"] \arrow[r, bend left, "\lambda"] & 6 \arrow[l, bend left, "6\mu"]
\end{tikzcd}$$

Steady state probability distribution

- In performance modelling we are often interested in the probability distribution of the random variables $X(t)$ with $t \in \mathbb{R}^{\geq 0}$ over the state space $\mathcal{S}$, as the system settles into a regular pattern of behaviour.
- This is termed the steady state probability distribution.
- From this probability distribution we will derive performance measures based on subsets of states where some condition holds.

Existence of a steady state probability distribution

- For every time-homogeneous, finite, irreducible Markov process with state space S, there exists a steady state probability distribution

$$\{\pi_k \mid x_k \in S\}\,.$$

- This distribution is unique if it exists and it is the limiting or long term probability distribution

$$\pi_k = lim_{t\to\infty} Pr(X(t) = x_k \mid X(0) = x_0)\,.$$

- This distribution is reached when the initial state no longer has any influence.

Probability flux

- In steady state, π_i is the proportion of time that the process spends in state x_i.
- Recall q_{ij} is the transition rate that the model makes a transition from state x_i to state x_j.
- Thus, in an instant of time, the probability that a transition will occur from state x_i to state x_j is the probability π_i that the model was in state x_i multiplied by the transition rate q_{ij}.
- This is called the probability flux from state x_i to state x_j.

Global balance equations

- In steady state, the equilibrium is maintained so that for any state the total probability flux out is equal to the total probability flux into the state.

$$\underbrace{\pi_i \sum_{x_j \in \mathcal{S},\ j \neq i} q_{ij}}_{\text{flux } \textbf{out} \text{ of } x_i} = \underbrace{\sum_{x_j \in S,\ j \neq i} \pi_j q_{ji}}_{\text{flux } \textbf{into } x_i}$$

Global balance equations

- Recall that the diagonal elements of the infinitesimal generator matrix $\mathbf{Q}$ are the negative sum of the other elements in the row, i.e., $q_{ii} = -\sum_{x_j \in \mathcal{S},\ j \neq i} q_{ij}$.
- We can use this to rearrange the flux balance equation to be:

$$\sum_{x_j \in \mathcal{S}} \pi_j q_{ji} = 0.$$

- Expressing the unknown values π_i as a row vector π, we can write this as a matrix equation:

$$\boldsymbol{\pi} \mathbf{Q} = 0\,.$$

Normalisation condition

- The π_i are unknowns - they are the values we wish to find.
- If there are N states in the state space, the global balance equations give us N equations in N unknowns.
- However this collection of equations is irreducible.
- Fortunately, since π is a probability distribution we also know that the normalisation condition holds:

$$\sum_{x_i \in \mathcal{S}} \pi_i = 1\,.$$

- With these $n + 1$ equations we can use standard linear algebra techniques to solve the equations and find the n unknowns $\{\pi_i\}$.

Example 2

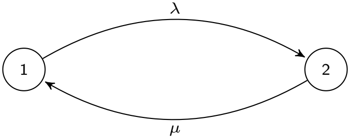
- Consider a system with multiple CPUs, each with its own private memory, and one common memory which can be accessed only by one processor at a time.
- The CPUs execute in private memory for a random time before issuing a common memory access request. Assume that this random time is exponentially distributed with parameter λ.
- The common memory access duration is also assumed to be exponentially distributed, with parameter μ (the average duration of a common memory access is $1/\mu$).

Example 2

If the system has only one processor, it has only two states:

1. The processor is executing in its private memory;
2. The processor is accessing common memory.

The system behaviour can be modelled by a 2-state Markov process whose state transition diagram and generator matrix are as shown below:



$$\mathbf{Q} = \begin{pmatrix} -\lambda & \lambda \\ \mu & -\mu \end{pmatrix}$$

Example 2

- If we consider the probability flux in and out of state 1 we obtain: $\pi_1 \lambda = \pi_2 \mu$. Similarly, for state 2: $\pi_2 \mu = \pi_1 \lambda$.
- We know from the normalisation condition that: $\pi_1 + \pi_2 = 1$.
- Thus the staedy state probability distribution is:

$$\boldsymbol{\pi} = \left(\frac{\mu}{\mu + \lambda}, \frac{\lambda}{\mu + \lambda} \right).$$

- From this we can deduce, for example, that the probability that the processor is executing in private memory is $\mu / (\mu + \lambda)$.

Solving the global balance equations

- In general the systems of equations will be too large to solve by hand.
- Instead we take advantage of linear algebra packages which can solve matrix equations of the form $\mathbf{Ax} = \mathbf{b}$
- where
  - $\mathbf{A}$ is an $N \times N$ matrix
  - $\mathbf{x}$ is a column vector of N unknowns
  - $\mathbf{b}$ is a column vector of N values.

Solving the global balance equations

First we must resolve two problems:

1. Our global balance equation is expressed in terms of a row vector of unknowns $\boldsymbol{\pi}$, $\boldsymbol{\pi}\mathbf{Q} = \mathbf{0}$.

   This problem is resolved by transposing the equation, i.e. $\mathbf{Q}^T\boldsymbol{\pi} = \mathbf{0}$, where the right hand side is now a column vector of zeros, rather than a row vector.

2 We must eliminate the redundancy in the global balance equations and add in the normalisation condition.

We replace one of the global balance equations by the normalisation condition. In $\mathbf{Q}^T$ we replace one row (usually the last) by a row of 1's. We denote the modified matrix $\mathbf{Q}_N^T$.

We must also make the corresponding change to the "solution" vector $\mathbf{0}$, to be a column vector with 1 in the last row, and zeros everywhere else. We denote this vector $\mathbf{e}_N$.

Example 3

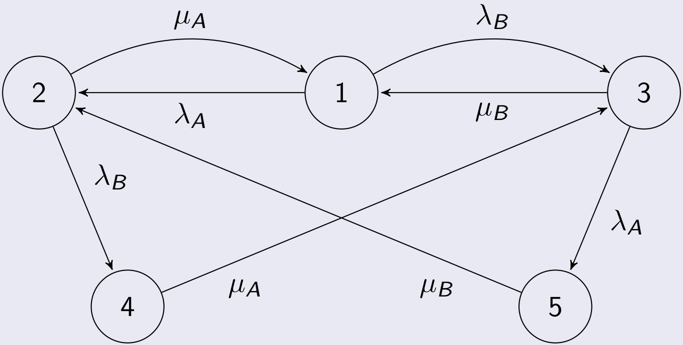
- Consider the two-processor version of the multiprocessor with processors A and B.
- We assume that the processors have different timing characteristics, the private memory access of A being governed by an exponential distribution with parameter λ_A, the common memory access of B being governed by an exponential distribution with parameter μ_B, etc.

Example 3: state space

Now the state space becomes:

1. A and B both executing in their private memories;
2. B executing in private memory, and A accessing common memory;
3. A executing in private memory, and B accessing common memory;
4. A accessing common memory, B waiting for common memory;
5. B accessing common memory, A waiting for common memory;

Example 3: state space



Example 3: generator matrix

$$\mathbf{Q} = \begin{pmatrix} -(\lambda_A + \lambda_B) & \lambda_A & \lambda_B & 0 & 0 \\ \mu_A & -(\mu_A + \lambda_B) & 0 & \lambda_B & 0 \\ \mu_B & 0 & -(\mu_B + \lambda_A) & 0 & \lambda_A \\ 0 & 0 & \mu_A & -\mu_A & 0 \\ 0 & \mu_B & 0 & 0 & -\mu_B \end{pmatrix}$$

Example 3: modified generator matrix

$$\mathbf{Q}_5^T = \begin{pmatrix} -(\lambda_A+\lambda_B) & \mu_A & \mu_B & 0 & 0 \\ \lambda_A & -(\mu_A+\lambda_B) & 0 & 0 & \mu_B \\ \lambda_B & 0 & -(\mu_B+\lambda_A) & \mu_A & 0 \\ 0 & \lambda_B & 0 & -\mu_A & 0 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

Example 3: steady state probability distribution

- If we choose the following values for the parameters:

$$\lambda_A = 0.05 \qquad \lambda_B = 0.1 \qquad \mu_A = 0.02 \qquad \mu_B = 0.05$$

solving the global balance equations, we obtain:

$$\pi = (0.0693,\ 0.0990,\ 0.1683,\ 0.4951,\ 0.1683)$$

Deriving Performance Measures

- Broadly speaking, there are three ways in which performance measures can be derived from the steady state distribution of a Markov process.
- These different methods can be thought of as corresponding to different types of measure:
  - state-based measures, e.g. utilisation;
  - rate-based measures, e.g. throughput;
  - other measures which fall outside the above categories, e.g. response time.

State-based measures

- State-based measures correspond to the probability that the model is in a state, or a subset of states, which satisfy some condition.
- For example, utilisation will correspond to those states where a resource is in use.
- If we consider the multiprocessor example, the utilisation of the common memory, U_{mem}, is the total probability that the model is in one of the states in which the common memory is in use:

$$U_{mem} = \pi_2 + \pi_3 + \pi_4 + \pi_5 = 0,9307 = 93.07\%$$

State-based measures

- Other examples of state-based measures are idle time, or the number of jobs in a system.
- Some measures such as the number of jobs will involve a weighted sum of steady state probabilities, weighted by the appropriate value (expectation).
- For example, if we consider jobs waiting for the common memory to be queued in that subsystem, then the average number of jobs in the common memory, N_{mem}, is:

$$N_{mem} = (1 \times \pi_2) + (1 \times \pi_3) + (2 \times \pi_4) + (2 \times \pi_5) = 1.594$$

Rate-based measures

- Rate-based measures are those which correspond to the predicted rate at which some event occurs.
- This will be the product of the rate of the event, and the probability that the event is enabled, i.e. the probability of being in one of the states from which the event can occur.

Example: Rate-based measures

- In order to calculate the throughput of the common memory, we need the average number of accesses from either processor which it satisfies in unit time.
- X_{mem} is thus calculated as:

$$X_{mem} = (\mu_A \times (\pi_2 + \pi_4)) + (\mu_B \times (\pi_3 + \pi_5)) = 0.0287$$

or, since $1/0.0287 \sim 35$, approximately one access every 35 milliseconds.

Other measures

- The other measures are those which are neither rate-based or state-based.
- For example, the response time of the common memory can be computed by

$$W_{mem} = N_{mem}/X_{mem} = 1.594/0.0287 = 55.54 \quad \text{milliseconds}$$