Markovian Models and Stochastic Processes

Enrico Vicario
Lab. of Software and Data Science - Dept. of Information Engineering - University of Florence

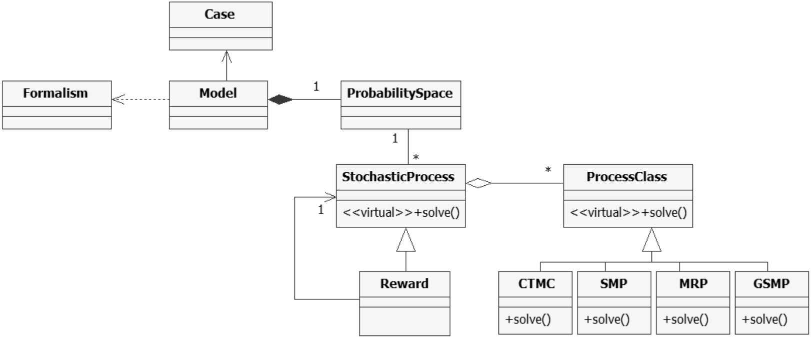
lectures at Wuhan University of Technology, January 2017 - Part 1

- this is about:
  - how models capture some case of reality into a probability space
  - and how solution of stochastic processes
    support quantitative evaluation of the case
  - ... in a Markovian setting
  - based on CTMCs and DTMCs underlying GSPN models

Outline

What are we talking when we talk about evaluating a model

- a model uses some formalism to capture a case in some reality
- the model identifies one single probability space
- ... on which we can define multiple stochastic processes and rewards
- ... amenable to solution techniques depending on the process class



... about the Probability Space

- a probability space characterizes an *experiment* as a triple $\langle \Omega, \mathcal{F}, \mathbb{P} \rangle$
- Ω is the space of *outcomes*,
  i.e. the set of possible behaviors that may result from the experiment
- $\mathcal{F}$ is the space of *events*, each made of a set of outcomes
  - $e \in \mathcal{F} \rightarrow ev \subseteq \Omega$
  - $\mathcal{F}$ is a sigma-algebra over Ω:
    $\Omega \in \mathcal{F}$
    if $e \in \Omega$ then $\Omega \setminus e \in \mathcal{F}$
    if $\{e_n\}_{n=1}^{\infty}$ is a countable set of events $e_n \in \mathcal{F}$, then $\bigcup_{n=1}^{\infty} e_n \in \mathcal{F}$
- $\mathbb{P} : \mathcal{F} \rightarrow [0, 1]$ is a *measure of probability* over the set of events
  - $\mathbb{P}(e)$ evaluates the probability that the outcome belongs to the event e
  - the measure evaluates to 1 over the entire set of outcomes, i.e. $\mathcal{F}(\Omega) = 1$,
  - ... and it is additive over countable union, i.e. if $\{e_n\}_{n=1}^{\infty}$ is a countable set of disjoint events $e_n \in \mathcal{F}$, then $\sum_{n=1}^{\infty} \mathbb{P}(e_n) = \mathbb{P}(\bigcup_{n=1}^{\infty} e_n)$

Probability Space: an example with discrete outcomes

- experiment: roll a die
  - an outcome is the value of the die
  - the set of outcomes is thus $\Omega = \{1, 2, 3, 4, 5, 6\}$
  - an event is any subset of Ω, and $\mathcal{F}$ is thus the set 2^{Ω} of the subsets of Ω
  - the measure of probability $\mathbb{P}$ descends from (fair die): $\mathbb{P}(\{1\}) = \mathbb{P}(\{2\}) = \mathbb{P}(\{3\}) = \mathbb{P}(\{4\}) = \mathbb{P}(\{5\}) = \mathbb{P}(\{6\}) = 1/6$
- another experiment: roll a die, until you get a 6
  - the number of values that make the outcome is not fixed, possibly infinite
  - but the set of outcomes is still countable

Probability Space: an example with continuous outcomes

- experiment: a travel schedule with guaranteed connection on the route Wuhan->Paris->Florence
  - an outcome is a 4-tuple $< \tau^d_{WUH}, \tau^a_{P-CDG}, \tau^d_{P-CDG}, \tau^a_{FLO} >$: departure/arrival times at Wuhan, Paris, Florence
  - the set of outcomes Ω is a subset of $\mathbb{R}^4$, with some constraints:

$$\begin{cases} a \leq & \tau^d_w & \leq A \\ b \leq & \tau^a_p - \tau^d_w & \leq B \\ c \leq & \tau^d_p & \leq C \\ d \leq & \tau^d_p - \tau^a_p & \leq D \\ e \leq & \tau^a_f - \tau^d_p & \leq E \end{cases}$$

  - the set of events $\mathcal{F}$ is made of the "reasonable" subsets of Ω, measurable by Lebesgue and "covering" Ω (... omissis)
  - the measure of probability $\mathbb{P}$ is derived through integrals over subsets of Ω weighted by distributions of delays and durations
- another experiment: a travel Wuhan->Paris->Florence where the Paris connection is not guaranteed but a protection flight is ensured
  - an outcome combines discrete and continuous quantities (flights and times)
- Remark: identifying a probability space may become difficult !

... about Random Variables

- a *random variable* X over a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a real valued function $X : \Omega \to \mathbb{R}$ such that $\forall x \in \mathbb{R}, \ \ \{\omega \in \Omega | X(\omega) \leq x\} \in \mathcal{F}$.
- $\{\omega \in \Omega | X(\omega) \leq x\} \in \mathcal{F}$ guarantees $\exists \mathbb{P}(\{\omega \in \Omega | X(\omega) \leq x\})$
- $\mathbb{P}(\{\omega \in \Omega | X(\omega) \leq x\})$ is the probability that X takes a value $\leq x$, and we thus write

$$Prob\{X \leq x\} := \mathbb{P}(\{\omega \in \Omega | X(\omega) \leq x\})$$

Conditional probabilities

- TBD: is this referred to the measure $\mathbb{P}$ of a Probability Space or to CDF of a Random Variable ?

Cumulative Distribution and Probability Density Functions

- a random variable X has a *Cumulative Distribution Function* (CDF) $F_X()$

  $$F_X(x) := Prob\{X \leq x\} := \mathcal{P}(\omega \in \Omega | X(\omega) \leq x\})$$

  $F_X(x)$ is monotonic non decreasing, with values in $[0, 1]$ and $\lim_{x\to\infty} F_X(x) = 1$.
- if $F_X()$ is absolutely continuous, then X also has a *Probability Density Function* (PDF)

  $$f_X(x) := \frac{dF_x(x)}{dx}$$

  $f_X(x)$ is non-negative, with $\lim_{x\to\infty} f_X(x) = 0$, and

  $$F_X(x) = \int_{-\infty}^{x} f_X(y)dy$$

- informally:

  $$f_X(x)dx = Prob\{X \in [x, x + dx]\}$$

  which subtends:

  $$f_X(x) = lim_{dx\to 0} \frac{Prob\{X \in [x, x + dx]\}}{dx}$$

Cumulative Distribution and Probability Density Functions

- while a random variable X identifies a unique CDF F_X and a PDF f_X, the viceversa is not true:
  - CDF and PDF are not sensitive to differences affecting the value of X over a null measure subset of Ω
  - relevant in the solution of *almost sure* problems, e.g. in the identification of events occurring *with probability 1* (w.p.1).
- the *support* S_X of a random variable X is the smallest closed interval $[a,b] \subseteq \mathbb{R}$, with $a \in \mathbb{R} \bigcup \{-\infty\}$ and $b \in \mathbb{R} \bigcup \{+\infty\}$, such that $Prob\{X < a\} = Prob\{X > b\} = 0$, i.e. $\int_a^b f_X(y)dy = 1$
  - $a = sup\{x \in \mathbb{R} | F_X(x) = 0\}$ and $b = inf\{x \in \mathbb{R} | F_X(x) = 1\}$
  - a PDF $f_X()$ is *Lebesgue equivalent* when for any subset of the support $[\alpha, \beta] \subseteq S_X$,

$$\int_\alpha^\beta f_X(y)dy = 0 \quad \text{if and only if} \quad \int_\alpha^\beta dy = 0$$

  - Lebesgue equivalence means that
    the PDF $f_X()$ does not have impulses or holes within its support

Discrete and continuous random variables

- a random variable is *continuous* if its CDF is absolutely continuous
  - this guarantees the existence of a PDF
  - e.g. arrival time in Florence
- a random variable taking values in a countable set is said to be *discrete*
  - e.g. summation of dice values before the first 6
  - the CDF becomes a stepwise function,
    the PDF can still be represented as a sequence of Dirac δ impulses
  - the distribution can be described by a *Probability Mass Function* (PMF):

$$PMF_t(\tau) := prob\{t = \tau\}$$

- a random variable is *mixed* if it takes values in a continuous space but still accepts non-null probabilities be concentrated in a countable set of points
- in the sequel, we refer to continuous random variables
  - yet, most can fit the case of discrete or mixed variables
    (by accepting Dirac distributions)

Moments and coefficients

- *moment* of order n of a random variable X

$$E[X^n] := \int_{-\infty}^{\infty} x^n f_X(x)dx$$

- some *coefficients* derived from moments characterize a distribution
- the *expected value* μ_X (aka mean or average) captures where the distribution is centered:

$$\mu_X := \int_{-\infty}^{\infty} x \cdot f_X(x)dx = E[X^1]$$

- the *variance* σ_X^2 captures how much the variable is dispersed:

$$\sigma_X^2 := \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x)dx = E[X^2] - (E[X])^2$$

- TBD: a picture with two distributions with the same mean but different variance

Moments and coefficients

- the *standard deviation* makes variance and mean comparable

$$\sigma_X := \sqrt{\sigma_X^2}$$

- the *variation coefficient* CV_X captures the relative dispersion

$$CV_X := \frac{\sigma_X}{\mu_X}$$

  - a low value of CV_X means that values of X are concentrated close to μ_X
  - since the Exponential distribution has $CV = 1$,
    distributions with $CV > 1$ are said *hyper-exponential*,
    while those with $CV < 1$ are said *hypo-exponential*

Moments and coefficients

- the *skewness* γ_X captures the asymmetry of the distribution of X with respect to its mean value:

$$\gamma_X := \int_{-\infty}^{\infty} (\frac{x - \mu_X}{\sigma_X})^3 f_X(x) dx$$

  in a mono-modal distribution, skewness is positive iff the mode comes before the mean
- moments of order higher than 2, and their corresponding coefficients, including skewness, are not represented in the Central Limit Theorem: their values are not relevant in the summation of infinite *Independent Identically Distributed* (IID) variables.
- a distribution is *heavy tailed*, if its moments do not exist beyond some order:
  e.g. for $f_X(x) = \frac{1}{x^{2.1}}$, the first order moment $E[X^1]$ exists finite, but $E[X^2]$ does not as $\int_0^{\infty} \frac{1}{x^{1.1}}$ exists finite, but $\int_0^{\infty} \frac{1}{x^{1.1}}$ does not

Multivariate random variables

- a *multivariate random variable* is a vector of random variables $\langle X_1, ...X_N\rangle$ on the same probability space
- a multivariate random variable has a *joint Cumulative Distribution Function*

$$F_{\langle X_1,...X_N\rangle}(x_1, ...x_N) := Prob\{X_1 \leq x_1, ...X_N \leq x_N\}$$

and a *joint Probability Density Function*

$$f_{\langle X_1,...X_N\rangle}(x_1, ...x_N) := \frac{\partial F_{\langle X_1,...X_N\rangle}(x_1, ...x_N)}{\partial x_1 ... \partial x_N}$$

- i.e., using integrals instead of derivatives,

$$F_{\langle X_1,...X_N\rangle}(x_1, ...x_N) = \int_{y_1 \leq x_1} \int_{y_2 \leq x_2} ... \int_{y_N \leq x_N} f_{\langle X_1,...X_N\rangle}(y_1, ...y_N) dy_1 dy_2 ... dy_N$$

- for any $S \subseteq \mathcal{R}^N$,

$$Prob\langle X_1, ...X_N\rangle \in S = \int_{\langle y_1, y_2, ... y_N\rangle \in S} f_{\langle X_1,...X_N\rangle}(y_1, ...y_N) dy_1 dy_2 ... dy_N$$

Independent random variables

- two random variables X_1 and X_2 on the same probability space are *independent* if the value taken by any of them does not condition the distribution of the values taken by the other one

$$\forall x_1, x_2 \; Prob\{X_1 \leq x_1 | X_2 \leq x_2\} = Prob\{X_1 \leq x_1\}$$

- if X_1 and X_2 are independent

$$Prob\{X_1 \leq x_1 \wedge X_2 \leq x_2\} = Prob\{X_1 \leq x_1\} \cdot Prob\{X_2 \leq x_2\}$$

  and their joint probability density function is thus a product form $f_{\langle X,Y \rangle}(x, y) = f_X(x) \cdot f_Y(y)$
- note that the concept of independence applies to components of a multivariate random variable, here captured saying that X_1 and X_2 are defined on the same probability space.
- the concept of independence can be extended to N variables by induction ...

Exponential random variable: distribution

- a random variable t is exponentially distributed (EXP) if its CDF $F_t()$ is:

$$F_t(\tau) := Prob\{t \leq \tau\} = \begin{cases} 0 & \text{if } \tau < 0 \\ 1 - e^{-\lambda_t \tau} & \text{if } \tau \geq 0 \end{cases}$$

- $F_t()$ is absolutely continuous and thus has a PDF $f_t()$ such that

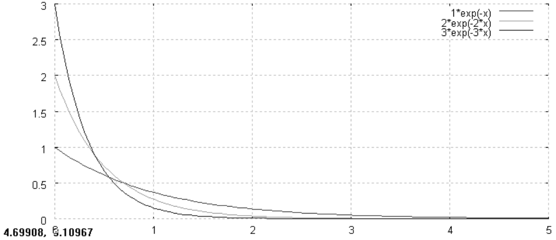
$$F_t(\tau) = \int_{-\infty}^{\tau} f_t(x)dx$$

with

$$f_t(\tau) = \frac{dF_t(\tau)}{d\tau} = \begin{cases} \lambda_t e^{-\lambda_t \tau} & \text{if } \tau \geq 0 \\ & \text{else} \end{cases}$$

Exponential random variable: moments

- (expected value) $\mu_t := \int_0^\infty \tau \cdot f_t(\tau) d\tau = \frac{1}{\lambda}$
- (variance) $\sigma_t^2 := \int_0^\infty (\tau - \mu_t)^2 \cdot f_t(\tau) d\tau = \frac{1}{\lambda^2}$
- (variation Coefficient) $CV_t := \frac{\sigma_t}{\mu_t} = 1$
- λ is the *rate* and accounts for how urgently an event is expected (in particular, $\frac{1}{\lambda}$ is the mean time to the event)



Exponential random variable: single parameter

- an EXP distribution is completely identified by its rate.
- when fitting some given statistics, having a single parameter prevents to independently set the expected value and standard deviation, which comprises one of the major limitations of the Exp variable ...
- the usual practice is that the rate is set so as to fit the expected value (a kind of rough application of Little's law) ...

Exponential random variable: minimum among variables

- the minimum $min := Min(t_1, t_2)$ among two EXP random variables t_1 and t_2 is an EXP random variable, with rate $\Lambda = \lambda_1 + \lambda_2$

$$\begin{aligned} f_{min}(x)dx &= Prob\{Min(t_1, t_2) \in [x, x+dx]\} \\ &= f_{t_1}(x)dx * Prob\{t_2 > x\} + f_{t_2}(x)dx * Prob\{t_1 > x\} \\ &= (\lambda_1 + \lambda_2)e^{-(\lambda_1+\lambda_2)x}dx \end{aligned}$$

- the minimum $min := Min(t_1, t_2, ...t_N)$ among N EXP random variables is an EXP random variable, with rate $\Lambda = \sum_{n=1}^{N} lambda_n$
  proof: by inductive hypothesis, the min among the first $N-1$ variables is an EXP with rate $\Lambda_{N-1} := \sum_{n=1}^{N-1} \lambda_n$; the min among the N variables is thus the minimum among this EXP and the N-th variable, and it has rate $\Lambda_N := \Lambda_{N-1} + \lambda_N$)
- the maximum between two EXP variables is not an EXP.
  This is unfortunate, as the maximum is what is computed in any synchronization, and it would be useful that it could assumed to maintain the properties of an EXP.

Exponential random variable: memoryless property

- in a queue where the waiting time is EXP, the time elapsed since when the queue was entered does not condition the expectance on the time to wait more
- formally: let t be EXP, let t' be the variable obtained by conditioning t to the assumption $t > k$, and let t'' be the variable obtained by reducing t' by k; we verify that t'' has the same distribution as that of t:

$$f_t(\tau) := \begin{cases} \lambda_t e^{-\lambda_t \tau} & \text{if } \tau \geq 0 \\ 0 & \text{else} \end{cases}$$

$$t' := t | t > k$$

$$f_{t'}(\tau) = \begin{cases} \dfrac{f_t(\tau)}{Prob\{t > k\}} & \text{if } \tau > k \\ 0 & \text{else} \end{cases} \quad \text{with } Prob\{t > k\} = 1 - F_t(k) = e^{-\lambda_t k}$$

$$t'' := t' - k$$

$$f_{t''}(\tau) = f_{t'-k}(\tau) = f_{t'}(\tau + k) = \begin{cases} \dfrac{f_t(\tau + k)}{Prob\{t > k\}} & \text{if } \tau > 0 \\ 0 & \text{else} \end{cases} = f_t(\tau)$$

Exponential random variable: memoryless property

- Viceversa, a continuous and memoryless distribution is exponential. A proof can be obtained by discretization ... TBD ... or it can be obtained using an argument that we will use later for CTMCs

Exponential random variable: random switch

- if X and Y are EXP variables with rates λ_X and λ_Y, respectively, then

$$Prob\{X < Y\} = Prob\{X \leq Y\} = \frac{\lambda_X}{\lambda_X + \lambda_Y}$$

  - proof: by butchery ... TBD
- the property is lifted to N concurrent variables by observing that the minimum among $X_2, ...X_N$ is an EXP with rate $\lambda_2 + ...\lambda_N$:
  - proof: if $X_1, ...X_N$ are EXP variables with rates $\lambda_1 ... \lambda_N$, respectively, then $Prob\{X_1 < X_n \forall n \in [2, N]\} = Prob\{X_1 \leq X_n \forall n \in [2, N]\} = Prob\{X_1 \leq Min_{n \in [2,N]}\{X_2 ... X_N\}\} = \frac{\lambda_1}{\lambda_1 + (\lambda_2 + ... \lambda_N)}$

Exponential random variable: independence

- if X and Y are EXP and τ is any stochastic variable, then the variables $X - \tau | X > \tau$ and $Y - \tau | Y > \tau$ are still independent and their joint distribution is in product form:

$$F_{\langle X-\tau, Y-\tau | X \geq \tau, Y \geq \tau \rangle}(x, y) = F_X(x) \cdot F_Y(y)$$

- Proof: ... use $f_{X-\tau}(x) = f_X(x + \tau)$ ...

Exponential random variable: invariance of the min random switch

- Given two EXP variables X and Y, the knowledge of the value of their minimum does not condition the probability of which of them realizes the minimum:

$$Prob\{X \leq Y \| Min\{X, Y\} \leq c\} = \frac{\lambda_X}{\lambda_X + \lambda_Y}$$

- A practical counterintuitive consequence:
  - i'm waiting for any of two events to occur;
  - I know that their occurrence times are distributed as Exp variables X and Y, with $\lambda_X = 1$ and $\lambda_Y = 0.1$, i.e. with expected values $\bar{X} = 1$ and $\bar{Y} = 10$;
  - I'm told that the first event occurs at time 10;
  - in a naive intuition, I might think that the event is more probably Y, yet, the probability that the event is Y is $\frac{\lambda_Y}{\lambda_X+\lambda_Y} = 0.00990099$, the same as I had before I knew the time of occurrence.
  - Proof: ... TBD
- if this sounds counter-intuitive, the reason can be that the property of being memoryless is not actually realized in most practical situations

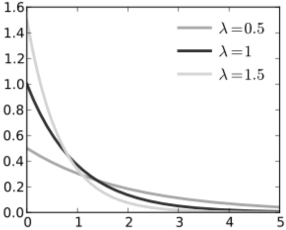
Exponential random variable: invariance of the min distribution

- Viceversa, Given two EXP variables X and Y, knowing which of them makes the minimum does not condition the value of the minimum:

$$Prob\{Min\{X,Y\} \leq c | X \leq Y\} = 1 - e^{(\lambda_X + \lambda_Y)c}$$

  Proof: ... TBD
- A practical consequence of the two invariance properties:
  - suppose you are simulating the race among a set of N Exp variables $X_1, ...X_N$ (e.g. the completion time of the first among of N concurrent operations), and you need to generate a sample that determines which variable will realize the minimum and at what time.
  - A direct way to generate the sample is to extract a sample for each of the N Exp variables and then select the minimum.
  - A different way producing the same statistical result is: evaluate the sum of rates $\Lambda = \lambda_1 + ...\lambda_N$; select which transition will fire first with a sample for the discrete distribution that assigns to X_N the probability $\frac{\lambda_n}{\Lambda}$; determine the time of this minimum through a further sample from the Exp distribution with rate Λ.

Exponential random variable: summary



- $$\begin{array}{l} \textit{support} \;\; x \in [0, \infty] \\ f_\tau(x) := \lambda e^{-\lambda x} \\ \bar{\tau} = \sigma_\tau = \frac{1}{\lambda} \end{array}$$
- an EXP variable is memoryless (sufficient and necessary for continuous distributions):

$$\begin{array}{l} \text{given } f_\tau(x) := \lambda e^{-\lambda x} \quad x \in [0, \infty] \\ \text{if } \tau' := \tau - x_0 | \tau > x_0 \text{ then } f_{\tau'}(x) = f_\tau(x) \end{array}$$

- the min of a set of EXP is EXP with rate equal to the sum of rates
- switching probability independent from the execution time $\frac{\lambda_0}{\sum_{n=0}^{N} \lambda_n}$
- samples can be easily generated by inversion
- ... good news for analysis, bad for expressivity
  - infinite support, single parameter, no memory

Geometric random variable

- the *discrete* random variable N is geometrically distributed if its MDF is:

$$MDF_N(n) := Prob\{N = n\} = p \cdot (1 - p)^{n-1}$$

- the geometric distribution results from a sequence of repeated Bernoulli experiments with *success probability p*
- the geometric distribution is the discrete analog of the EXP continuous distribution
    - a discrete random variable is memoryless iff it is geometrically distributed

Expolynomial distributions (aka exponomial)

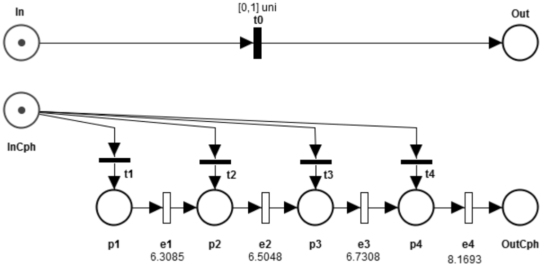
- $f_\tau(x) = \sum_{n=0}^{N} c_n x^{K_n} e^{-\lambda_n x} \qquad K_n \in \mathbb{N}, \lambda_n \in \mathbb{R}$
- closed form integration (note that $K_n \in \mathbb{N}$)

$$\int c_n x^{K_n} e^{-\lambda_n x} = c_n \lambda_n \sum_{k=0}^{K_n} \binom{K_n}{k} x^{K_n - k} e^{-\lambda_n x}$$

- closed wrt arithmetic operations, integration, derivation
- what about fitting other distributions through expolynomials?

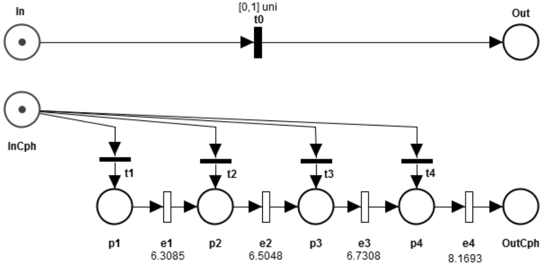
Continuous Phase Type distributions - CPH

- time to absorption in a Continuous Time Markov Chain (see later)

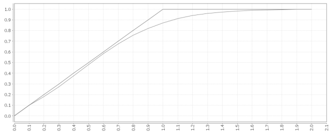


- by construction in the class of Expolynomial with support $[0, \infty]$ (for acyclic chains or allowing complex exponentials)

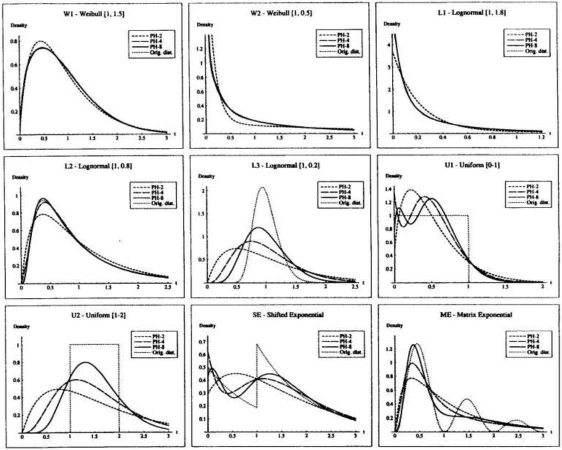
Continuous Phase Type distributions - CPH



- dense in the field of all positive-valued distributions
  - well developed fitting techniques (and tools)
  - suffering across discontinuities and finite supports



Continuous Phase Type distributions - CPH

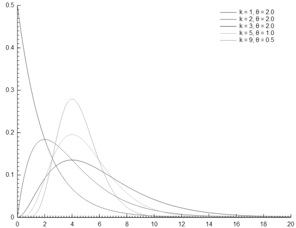


[1]

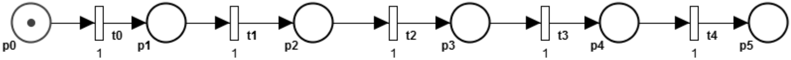
[1] A.Bobbio, M.Telek, "A benchmark for PH estimation algorithms: result for Acyclic-PH", Stochastic Models, 1994.

Erlang distribution

- $$\begin{array}{l} support: \quad x \in [0, \infty] \\ f_\tau(x) := \frac{\lambda^k x^{k-1} e^{-\lambda x}}{(k-1)!} \\ \bar{\tau} = \frac{k}{\lambda} \qquad \sigma_\tau = \frac{\bar{\tau}}{\sqrt{k}} \\ CV = \frac{1}{\sqrt{k}} \end{array}$$



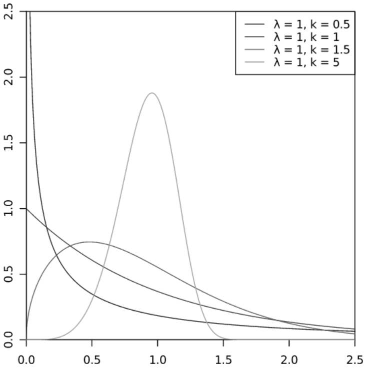
- sum of k identical independent EXP with rate λ (thus a Phase Type)



- when $k \in \mathbb{R}$, the Erlang becomes a Gamma distribution
  - integrals involve the γ function

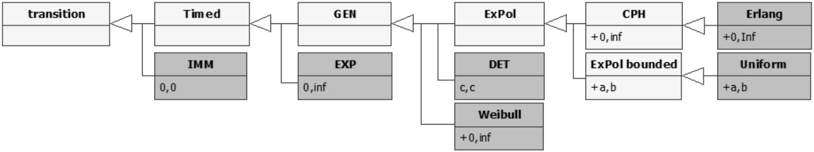
Weibull distribution

- *support* : $x \in [0, \infty]$
  $f_\tau(x) := \frac{k}{\lambda^k} x^{k-1} e^{-(\frac{x}{\lambda})^k}$
  $\bar{\tau} = \lambda\Gamma(1 + \frac{1}{k})$
  $\sigma_\tau^2 = \lambda^2(\Gamma(1 + \frac{2}{k}) - \Gamma(1 + \frac{1}{k})^2)$
  failure rate $= \frac{k}{\lambda}(\frac{x}{\lambda})^{k-1}$



- not in the class of expolynomials (and thus not even of CPH)
- if X is uniform, then $Y := \lambda(-ln(1 - X))^{\frac{1}{k}}$ is a Weibull (useful for generating samples in simulation)
- failure rate can be growing (an ageing component), decreasing (a component that consolidates) or constant (a memoryless component).

where do distributions come from?



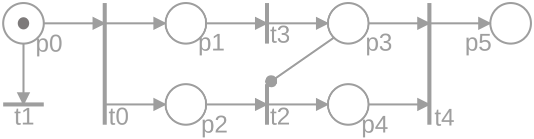
- EXP: no memory of time elapsed
- DET: timeout, watchdog, synchronous release
- UNI: jitter, delay wrt uncorrelated process, arrival time distribution in a Poisson process conditioned to n arrivals within t
- Continuous Phase Type: fit of a distribution over $[0, \infty]$
- Expolynomial: fit of a measured histogram, possibly over bounded support
- upper bounded support: timeout or watchdog truncation, guaranteed Worst Case Execution Time
- lower bounded support: min guaranteed intertime, Best Case Execution Time

... about Models, Formalisms, Random Variables, and Prob. Spaces

- a Model captures a Case using some Formalism
- a stochastic Model includes Random Variables
- a fully stochastic Model identifies a unique Probability Space

a Formalism for untimed concurrent activities

- Petri Nets (PN),
  a Formalism for concurrent untimed discrete event systems
  - a *place* (circle) encodes a state condition, true when containing *tokens*
  - a *transition* (bar) encodes a discrete event,
    enabled if all input conditions are true and all inhibiting conditions are false
  - an enabled transition is eventually *fired* or disabled
  - at firing, remove one token from each *input place*,
    and add a token to each *output place*



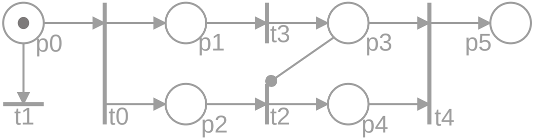
- abstraction
  - no timing: this is about eventually or never
  - no probabilities: this is about necessary or possible

... more details on Petri Nets

- syntax

$$PN = \langle P, T, A^+, A^-, A^\bullet, M_o \rangle$$

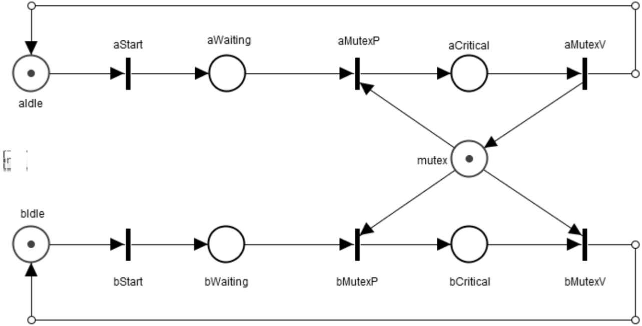
  - P and T are disjoint sets of places and transitions
  - $A^+ : T \to P$, $A^- : P \to T$, and $A^\bullet : P \to T$ are pre-conditions, post-conditions, and inhibitor-arcs
  - $M_0 : P \to \mathbb{N}^{||P||}$ is the initial marking



- semantics
  - state := marking ($m : P \to \mathbb{N}$)
  - t_o is enabled if there is at least 1 token in each input place and 0 tokens in each inhibitor place
  - at the firing of t_0, remove one token from each input place of t_0, and then add one to each output place of t_0
- ... plus convenience shorthands
  - enabling functions, update functions ... (as in Minski machines)

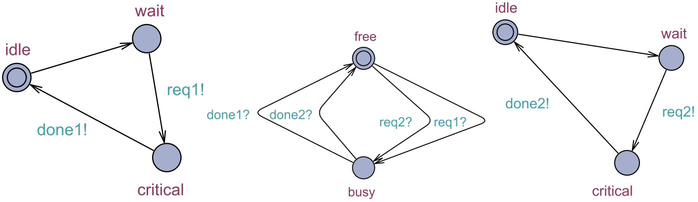
a PN Model of mutual exclusion

- synchronization of two threads on a binary semaphore *mutex*
  - each thread cycles through the states *idle*,*waiting*, and *critical*;
  - mutual exclusion enforced by *v* and *p* operations on the *mutex*



a Model of mutual exclusion based on the Formalism of Automata

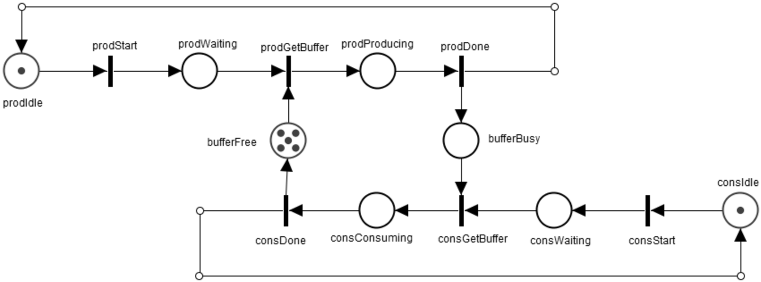
- The same model could be specified using communicating finite state machines instead of Petri Nets
  - 3 communicating automata: one for each thread and one for the mutex; synchronized two-by-two on *p* and *v* events



- specification in a single Finite State Machine is possible as well
  - a flat automaton omo-morphic to the reachability graph,
  - each location encodes the states of each thread and the mutex
- the (flat) PN model natively supports decomposition of the state
  - multiple state components represented by different places
  - enables"local" reasoning on pre- and post-conditions of each event

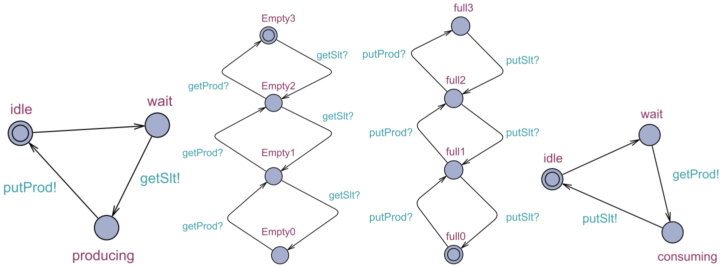
a PN Model of a producer consumer system

- a producer and a consumer
  exchanging products over a buffer with capacity 5



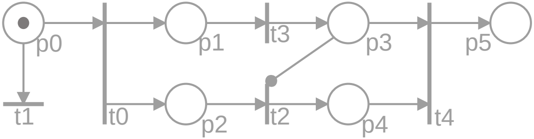
- exercise: add failures on the buffer when busy,
  and repair restoring the buffer in the free state

a Model of producer consumer based on the Formalism of Automata



... about the Formalism of Petri Nets and Probability Spaces

- a Petri Net Model can be used to specify an *experiment*
  - a possible outcome: the marking when the PN stops (i.e. $\Omega :=$ set of reachable stopping markings)
- another kind of outcome might be: the sequence of transitions fired before stopping
  - ... on state properties vs path properties



- ... in any case,
  - outcomes do not observe time
  - ... and no measure of probability is provided

two reasons for stochastic models

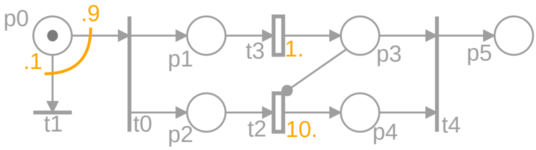
- non-functional requirements may involve probability
  - various classes of quality figures natively subtend a probabilistic characterization: performance, reliability availability and maintainability (RAM), performability and recoverability.
  - safety requirements are usually expressed with certainty, but they are often recast into a probabilistic form
    e.g.: a protection mechanism is triggered whenever the system is not able to guarantee with certainty that some safety requirement is satisfied, but the corresponding reduction of availability within a given time period is restricted in probabilistic form.
- (the abstraction of) system behavior may be probabilistic
  - by inherent consequence of non controllable facts
    (e.g. the Execution Time of a computation chunk with random duration)
  - or by design (e.g. randomization in the Ethernet truncated exponential backoff)
  - or by abstraction or epistemic origin, due to the inability to observe the details of a deterministic system

... about Models, Formalisms, and Random Variables [step 2]

- a Model captures a Case using some Formalism
- a stochastic Model includes Random Variables

Generalized Stochastic Petri Nets (GSPN)

- Generalized Stochastic Petri Nets
  augment Petri Nets with stochastic durations and probabilistic switches
  - transitions are either immediate (IMM, thin bar)
    or their duration is an exponential (EXP, thick bar) random variable
  - IMM transitions have priority over EXP transitions
  - choices among IMM transitions are resolved by random switches
  - choices among EXP transitions are resolved by race semantics,
    relying on the minimum among exponentially distributed transitions
  - IMM random switches and EXP rates may depend on the marking



... more details on GSPN: syntax

- Generalized Stochastic Petri Net (GSPN)

$$GSPN = \langle P, T, A^-, A^+, A^\cdot, M_0, R \rangle$$

  - the set of transition T is partitioned in two sets of exponentially distributed transitions (EXP) and immediate transitions (IMM):

$$T = T^{IMM} \bigcup T^{EXP}$$
$$T^{IMM} \bigcap T^{EXP} = \emptyset$$

  - R is a positive real valued function associating each transition with a value that depends on the current marking:

$$R : T \times \mathbb{N}^{\#P} \rightarrow \mathbb{R}$$

- remark:
  - for $t_e \in T^{EXP}$, $R(t_e, m)$ encodes the *rate* of the EXP duration of t_e;
  - for $t_i \in T^{IMM}$, $R(t_i, m)$ encodes the probabilistic *weight* of t_i in a random switch.

... more details on GSPN: semantics

- state := marking ($m : P \rightarrow \mathbb{N}$)
- a transition is *enabled* if each input place contains one token at least and each inhibiting place does not contain any token
  $T^{IMM,en}(m)$:= set of IMM transitions enabled by m
  $T^{EXP,en}(m)$:= set of EXP transitions enabled by m
- if $T^{IMM,en}(m) \neq \emptyset$, the next firing occurs after a null delay, and $t_i \in T^{IMM,enab}$ fires with probability

$$Prob\{t_i fires\} := \frac{R(t_i, m)}{\sum_{t_h \in T^{IMM,en}(m)} R(t_h, m)}$$

- if $T^{IMM,en}(m) = \emptyset$, the next firing occurs after an EXP delay with rate $\Lambda := \sum_{t_k \in T^{IMM,enab}} R(t_k, m)$, and $t_e \in T^{EXP,en}(m)$ fires with probability

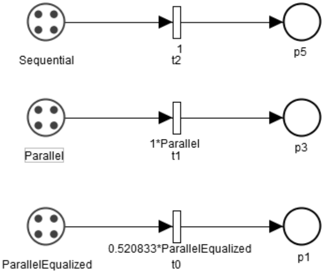
$$Prob\{t_e fires\} := \frac{R(t_e, m)}{\sum_{t_k \in T^{EXP,en}(m)} R(t_k, m)}$$

  (race semantics and minimum among EXP random variables)
- when $t_o \in T$ fires, the marking m is updated "as usual" (resampling of times to fire and EXP memoryless property)

GSPN modeling pragmatics

- IMM transitions for (relatively) negligible durations,
  EXP for anything else (not good news for expressivity)
- Exp rates equal to the inverse of the mean value of the duration
- marking dependent rates for actions performed in parallel or series

GSPN modeling pragmatics: single vs parallel service - 1/2

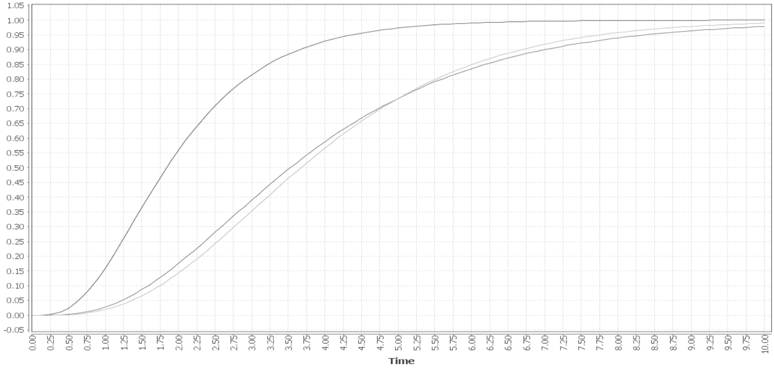


- Sequential: 4 jobs with rate 1, served sequentially (single server semantics)
  (completion time is an ERL(1.,4) random variable)
- Parallel: jobs are served in parallel (multiple server semantics)
- ParallelEqualized: jobs are served in parallel, but rate λ_p is set so as to have the same mean value μ_s of Sequential:

$$\frac{1}{4\lambda_p} + \frac{1}{3\lambda_p} + \frac{1}{2\lambda_p} + \frac{1}{1\lambda_p} = \frac{4}{\mu_s}$$

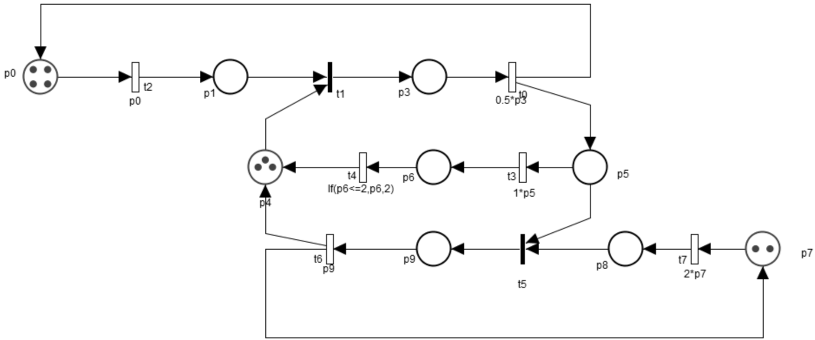
GSPN modeling pragmatics: single vs parallel server - 2/2

- Cumulative Distribution Function of the completion time
  - Parallel (upper plot) completes much faster
  - Sequential and ParallelEqualized have comparable completion times, with ParallelEqualized starting faster and then becoming slower.



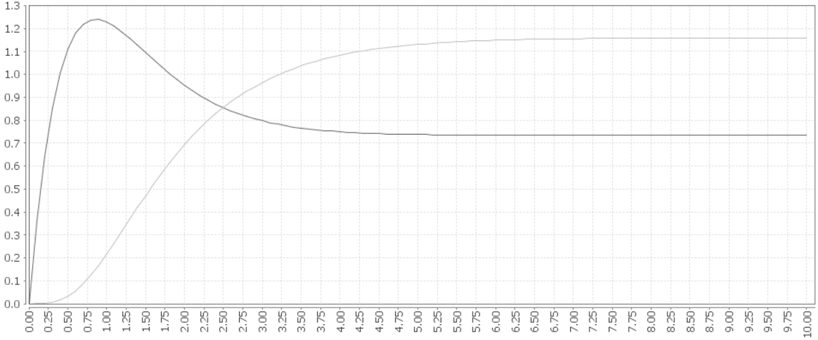
a GSPN Model for a producer consumer with buffer breakdowns - 1/2

- a producer/consumer system,
  with bounded buffer size, with buffer breakdowns and repair
- production, service and breakdown occur in parallel,
  repair is parallel but bounded to a maximum value



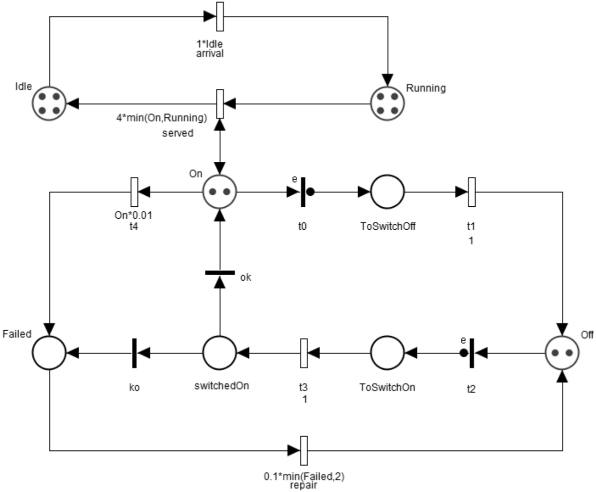
a GSPN Model for a producer consumer with buffer breakdowns - 2/2

- probability that producer threads are blocked waiting for free slots on the buffer, evaluated as the mean number of tokens in p1 when p4 is empty (If(p4==0,p1,0)) (the curve tending to 1.159);
- mean number of consumer threads blocked waiting for a busy slot, evaluated as the mean number of tokens in p8 when p5 is empty (If(p5==0,p8,0)) (the curve tending to 0.735)



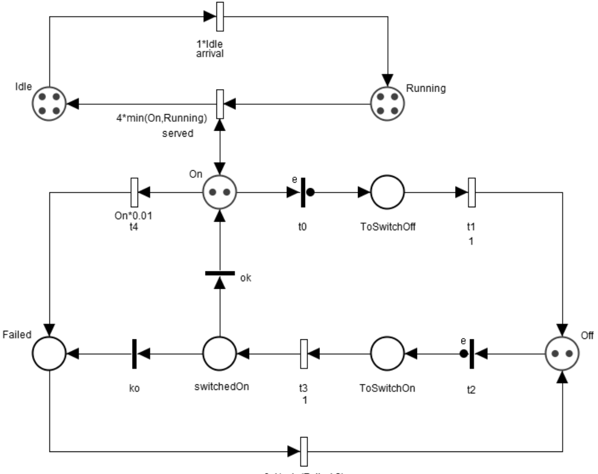
a GSPN Model for on/off control in a server cluster - 1/4

- `Idle` and `Running` account for the level of load
- `On`, `Off`, and `Failed` are the number of on, off, and failed servers



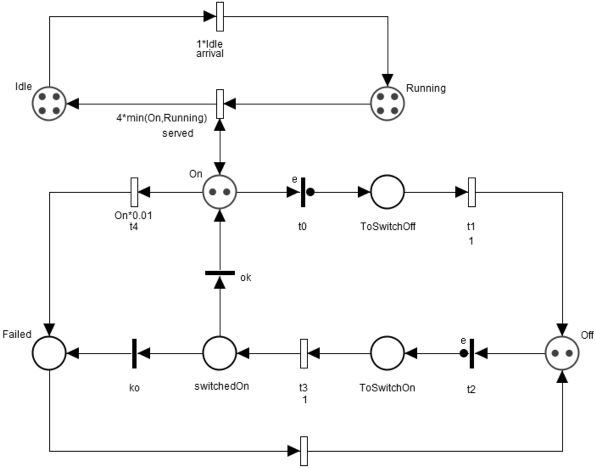
a GSPN Model for on/off control in a server cluster - 2/4

- `arrival` rate proportional to number of tokens in `Idle` (in parallel)
- `served` rate proportional to min between On servers and Running jobs: service in parallel with degree of concurrence equal to the number of active servers



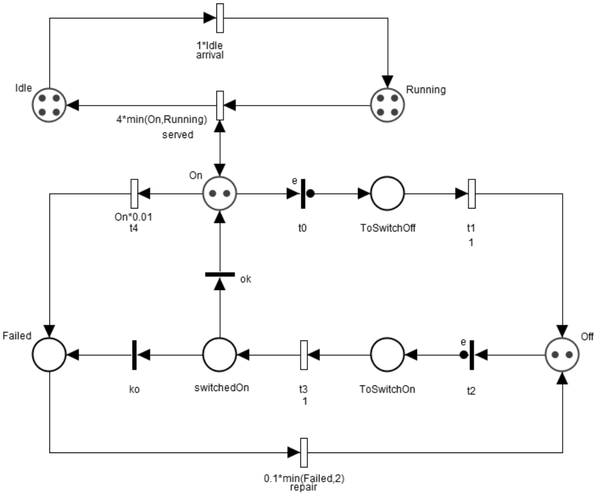
a GSPN Model for on/off control in a server cluster - 2/4

- switch-on/off started by IMM transitions `t0` and `t2`, under enabling functions (not shown) `if(Running<(On*2-2))` and `if(Running>(On*2+2))` (hysteresis)
- switch actuation delay modeled by `t1` and `t3` (preselection)

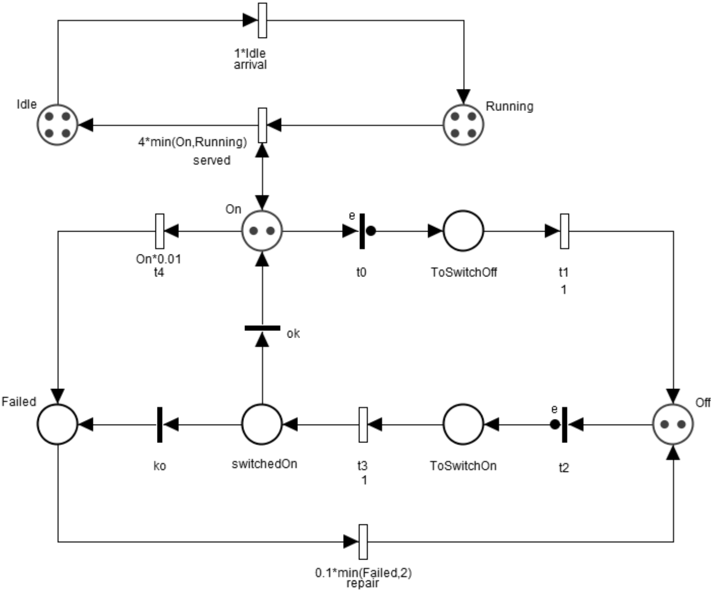


a GSPN Model for on/off control in a server cluster - 2/4

- a server can fail in use (`fail`) or at switch-on (`ko`)
- switch-on failure probability $\frac{5}{5+95}$ for weights 5 and 95 of `ko` and `ok`
- `repair` has a maximum degree of concurrence 2



a GSPN Model for on/off control in a server cluster - 2/4



a GSPN Model for on/off control in a server cluster - 2/4

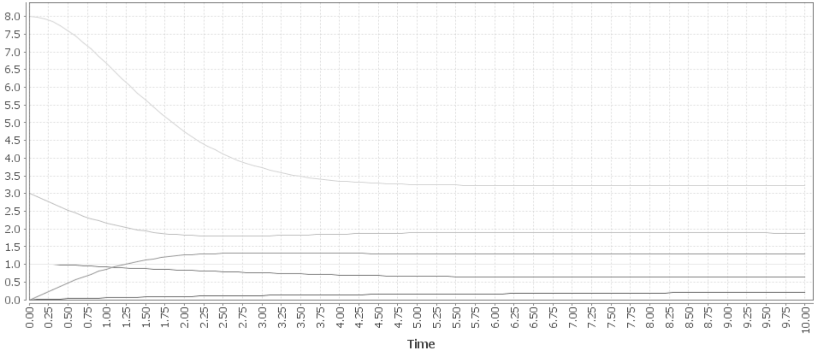
- control policy for switching on and off a cluster of servers
  - `Idle` and `Running` account for the level of load
    `arrival` rate proportional to the number of tokens in `idle`: parallel arrivals
    `served` rate proportional to the minimum between the number of active servers on the number of running jobs at service: jobs run in parallel with a degree of concurrence equal to the number of active servers
    total number of tokens in places `idle` and `running` sets the precision in the discretization of a real level of load
  - `On` and `Off` are the number of servers on and off
    switch-off is decided by the IMM transitions `t0` and `t2`, guarded by enabling functions (not shown) `if(Running<(On*2-2))` and `if(Running>(On*2+2))`
    switch actuation delay modeled by the `t1` and `t3`. (preselection)
  - `Failed` is the number of failed servers
    a server can fail over time while it is in use (`fail`),
    or it can have a concentrated failure at switch on (`ko`)
    random switch between `ko` and `ok` determined by transition weights (not shown) 5 and 95: a concentrated failure occurs with probability $\frac{5}{5+95}$
  - `repair` has a maximum degree of concurrence 2

a GSPN Model for on/off control in a server cluster - 3/4

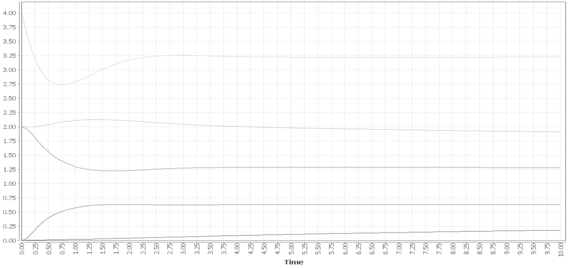
- (from top to bottom at time 10) the average number at time t of:
  - number of pending jobs (`Running`);
  - number of switched off servers (`Off`);
  - number of active servers (`On`);
  - number of switching servers (`ToSwitchOff+ToSwitchOn`);
  - number of failed servers (`Failed`).

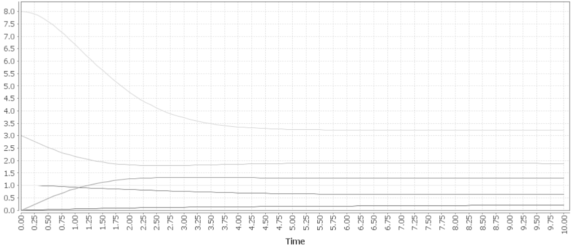
a GSPN Model for on/off control in a server cluster - 4/4

- the same transient rewards from a different initial worst case condition
  - the offered load is maximum (i.e. `idle=0` and `running=8`)
  - and all servers are off (i.e. `on=0` and `off=4`)
- the measure of interest is the settling time to recover the steady state
  - a measure of recoverability



a GSPN Model for on/off control in a server cluster - 5/4





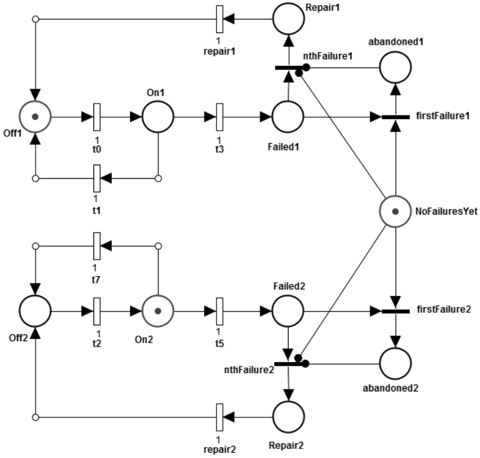
- steady state probabilities invariant to initial conditions (ergodic model)

Exercise... (later discovered to be about irreducibility)

- two servers initially active, but subject to failure
- the first to fail will be abandoned
- the other will be maintained and repaired after each failure

Exercise... (later discovered to be about irreducibility)

- two servers initially active, but subject to failure
- the first to fail will be abandoned
- the other will be maintained and repaired after each failure



qualitative, quantitative, and mixed models

- non-deterministic model: multiple behaviors from the same state
  - qualitative: about existential|universal quantification over the set of behaviors
  - e.g.: Petri Nets (Time Petri Nets, Timed Automata)
- stochastic model: behaviors associated with a measure of probability
  - some say probabilistic when it in includes discrete choices (e.g. the roll of a die) and it is stochastic when it involves the sample of a continuous variable (e.g. the duration of a computation).
  - in Greek, "stochastic" means "able to make reasonable predictions"
  - quantitative: about the probability of sets of behaviors
  - e.g.: Generalized Stochastic Petri Nets (Stochastic Time Petri Nets)
- mixed non-deterministic and stochastic models: some choices have stochastic characterization and others are left non-deterministic
  - ground for stochastic optimization: find the determination of non-deterministic choices that results in the best/worst probabilistic behavior
  - e.g.: Markov Decision Processes (Markov Automata)

... just a mention of a mixed stochastic/non-deterministic Formalism

- Markov Decision Processes
  - discrete time
  - at each step, non-deterministic choice of the discrete probability distribution that determines the next state
  - a strategy resolves the choices, possibly with randomization, possibly depending on the state or the history
- a Formalism for stochastic optimization
  - an MDP Model identifies a (possibly infinite) set of probability spaces
  - an MDP Model with a strategy is fully stochastic, and identifies a unique probability space
  - optimization problem: find the strategy that identifies the probability space where some reward is maximized (or minimized)

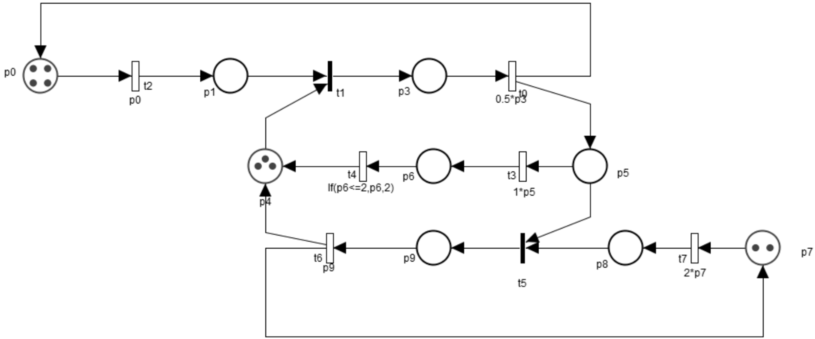
... about Models, Formalisms, Random Variables, and Prob. Spaces

- a Model captures a Case using some Formalism
- a stochastic Model includes Random Variables
- a fully stochastic Model identifies a unique Probability Space

... what are Models intended for?

- philosophically, models serve to gain model-driven insight
- verification of functional requirements
  - usually qualitative: about what is possible or necessary in the set of behaviors of a model
  - e.g., in the RTCA-178B perspective: formalization, disciplined reasoning, automated reasoning
- quantitative evaluation of non-functional requirements
  - early assessment of design choices
  - (main focus of this presentation)

Models are often for quantitative evaluation - 1/2

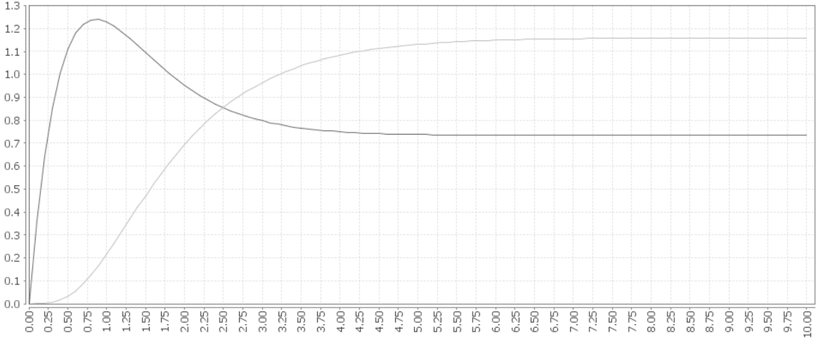


- some possible questions
  - is the "performance" sufficient, and, to what extent ?
  - is the contribution of producers, consumers, and of the buffer balanced ?
  - should be better enhance the rate of production or service ?
  - or should we increase the number of producers or consumers?
  - or should we enhance failure and repair processes, or the number of slots ?

Exercise: use the Oris tool to answer some of the above questions

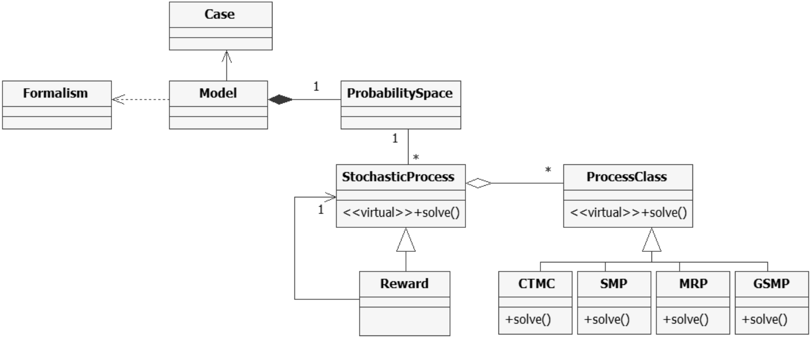
Models are often for quantitative evaluation - 2/2

- probability that producer threads are blocked waiting for free slots on the buffer, evaluated as the mean number of tokens in p1 when p4 is empty (If(p4==0,p1,0)) (the curve tending to 1.159);
- mean number of consumer threads blocked waiting for a busy slot, evaluated as the mean number of tokens in p8 when p5 is empty (If(p5==0,p8,0)) (the curve tending to 0.735)



what are we talking when we talk about evaluating a model

- a Model uses some Formalism to capture a Case in some reality
- the Model identifies one single Probability Space
- ... on which we can define multiple Stochastic Processes and Rewards
- ... amenable to solution techniques depending on the process class



Models, Probability Spaces, and Random Variables

recall: a Model identifies a Probability Space $(\Omega, \mathcal{F}, \mathbb{P})$

- an outcome $\omega \in \Omega$ is a run of the model
- an event in the σ-algebra $\mathcal{F}$ is a set of runs
- the measure of probability $\mathbb{P} : \mathcal{F} \to [0, 1]$
  is induced by stochastic parameters and by the initial condition

recall: a Random Variable is a non-negative function on the set of outcomes Ω

- e.g. the marking of a GSPN, the number of tokens in some specific place, ...

Stochastic Process

- a Stochastic Process $\mathbb{M}$
  is "a" collection of random variables on a Probability Space $(\Omega, \mathcal{F}, \mathbb{P})$,
  taking values in some measurable state space M,
  indexed by some parameter t taking values in a totally ordered set T:

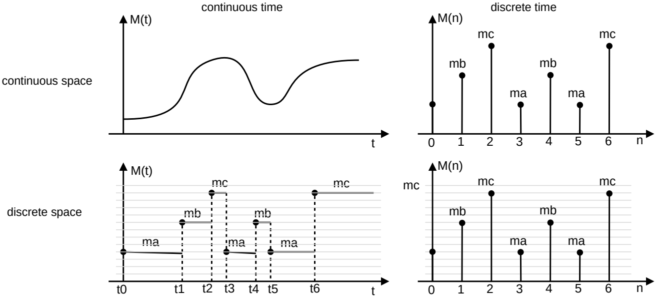
$$\mathbb{M} = \{M(t), t \in T\}$$

Remark: a Model identifies *one* Probability Space,
on which *many* Stochastic Processes can be identified

- with reference to a GSPN Model: the marking, the marking restricted to some subset of places, ... at time t or after n firings, ...

Stochastic Process

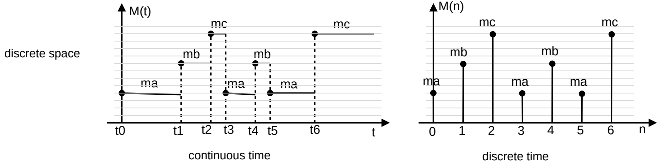
- a Stochastic Process $\mathbb{M} := \{M(t), t \in T\}$ is
  - continuous/discrete-space whether M is continuous/discrete
  - continuous/discrete-time whether T is continuous/discrete



- continuous time processes are assumed to be Right Continuous:

$$M(t_0) = \lim_{t \to t_0^+} M(t)$$

marking process

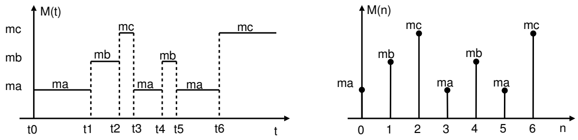


- we mainly focus on discrete-space processes
  - states are reachable markings
  - parameter t may stand for (continuous) time
  - ... or for the (discrete) number of firings
- in particular, we focus on the Marking Process underlying a GSPN

$$\text{marking process} := \quad \mathbb{M} = \{m(t), t \in \mathbb{R}_{\geq 0}\}$$
$$m(t) := \text{marking at time } t$$

embedded chains

- a discrete time process obtained by sampling of a continuous time process is named embedded
- the most common case: sample the state at each transition
- other sampling strategies may be useful as well:
  at equidistant time points, at the transitions of some other process, ...



(a) (b)

example: evolution of the marking of a GSPN through markings *ma*, *mb* e *mc*, observed as a CTC $\mathbb{M} = \{M(t), t \in \mathbb{R}_o^+\}$ (Fig.(a)), or as its embedded DTC$\mathbb{M}^e = \{M(n), n \in \mathbb{N}\}$ (Fig.(b))

Markov condition and time homogeneity

- *Markov condition:*
  the most recent state observation subsumes any previous one

  $$\forall t, \forall n, \forall t \geq t_n \geq ... \geq t_1 \geq t_0 \in T :$$
  $$Prob\{X(t) \leq x | X(t_n) = x_n, X(t_{n-1}) = x_{n-1}, ..., X(t_0) = x_0\}$$
  $$= Prob\{X(t) \leq x | X(t_n) = x_n\}$$

  (the same formulation for continuous and discrete time processes)
- *time-homogeneity:*
  behavior does not depend on the absolute time

  $$Prob\{X(t) \leq x | X(t_n) = x_n\} = Prob\{X(t - t_n) \leq x | X(0) = x_n\}$$

Discrete Time Markov Chain - DTMC

- Discrete Time Markov Chain (DTMC):
  - a discrete time stochastic process $\mathbb{X} = \{X(n), n \in \mathbb{N}\}$
  - that always satisfies the Markov condition:

$$\forall i, j, n, m, \forall M, i_M, i_{M-1}, ..., i_M, m_M \geq m_{M-1} \geq ... \geq m_1 \geq 0,$$
$$Prob\{X(n+m) = j | X(n) = i\} =$$
$$Prob\{X(n+m) = j | X(n) = i, X(n-m_1) = i_1, ...X(n-m_M) = i_M\}$$

- we restrain the treatment to finite state and time homogeneous DTMCs:

$$Prob\{X(n+m) = j | X(n) = i\} = Prob\{X(m) = j | X(0) = i\}$$

DTMC - quantities of interest

- we are interested in evaluating
  - transient probabilities

$$\pi_i(n) := Prob\{X(n) = i\}$$

  - steady state probabilities, if they exist

$$\pi_i := \lim_{n \to \infty} \pi_i(n)$$

DTMC - characterizing quantities

- a DTMC is completely characterized by
  - probability distribution of the initial state:

    $$\pi_i^0 := \pi_i(0)$$

  - transition probabilities at time n:

    $$p_{ij}(n) := Prob\{X(n+1) = j | X(n) = i\}$$

    which, for a time-homogeneous DTMC, take the form:

    $$p_{ij} := p_{ij}(0) = p_{ij}(n)$$

    note that the matrix p_{ij} is stochastic, i.e.:
    $p_{ij} \in [0,1]$ and $\sum_{j=1,l} p_{ij} = 1$, with l:=number of states.
- a finite-state time-homogeneous DTMC
  is thus conveniently figured out as a graph
  - each and every state is a vertex
  - the edge from state i to state j is labelled by the value p_{ij}

on the relation between DTMCs and geometric random variables

- the geometric distribution implements the concept of memoryless behavior for a discrete time random variable; in a similar manner, the DTMC realizes the concept of Markovian behavior, where the future evolution depends on the current state but not on the past history through which the state was reached
- in a DTMC, the sojourn time in state k is a random variable with geometric distribution with rate p_{kk}:

$$Prob\{SJ_k = i\} = p_{kk}^{i-1}(1 - p_{kk}), i \geq 1$$

proof: assuming $H_i(n) := Prob\{SJ_i > n\}$, by the Markov condition:

$$H_i(n+1) := Prob\{SJ_i > n+1\} = Prob\{SJ_i > n\} * p_{ii} = H_i(n) \cdot H_i(n)$$

which yields $H_i(n) = p_{ii}^n$ and thus:

$$Prob\{SJ_i = n\} = p_{ii}^n \cdot (1 - p_{ii})$$

DTMC - Transient analysis

- Chapman-Kolmogorov equations

$$\pi_i(n+1) = \begin{cases} \sum_{j=1}^{I} \pi_j(n) \cdot p_{ji} & \textit{if } n > 0 \\ \pi_i^0 & \textit{if } n = 0 \end{cases}$$

proof: by the law of total probability and the Markov condition

- in vectorial form

$$\underline{\pi}(n) = \begin{cases} \underline{\pi}(n-1) \cdot P & \textit{if } n > 0 \\ \underline{\pi}^0 & \textit{if } n = 0 \end{cases}$$

- computational perspective
  - transient probabilities are derived from initial probabilities $\underline{\pi}^0$ through repeated vector-matrix products
  - largely facilitated by the fact that P is a stochastic matrix, i.e. its elements are non negative values, lower than 1, summing up to 1 on each row.

DTMC - Existence of steady state probabilities

- if a DTMC is finite, time-homogeneous, irreducible and aperiodic, the limit of transient probabilities exists unique and is the solution of

$$\begin{cases} \underline{\pi} = \underline{\pi} \cdot P \\ |\underline{\pi}| = 1 \end{cases}$$

proof: the complex part is existence and unicity (omissis);
the form is just the limit on Chapman-Kologorov equations

- computational perspective
  - steady state probabilities are determined as the solution of a linear system
  - they do not depend on initial probabilities (ergodicicty)

DTMC - irreducibility

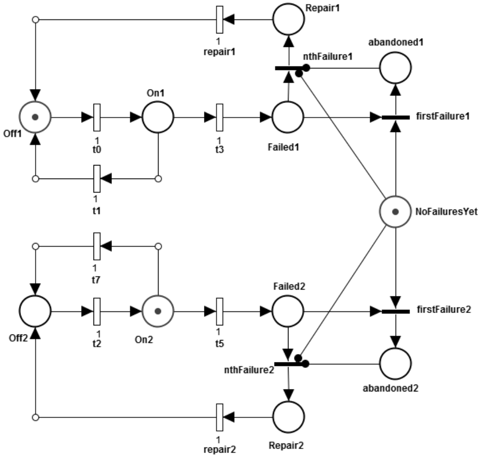
- irreducibility relates to the ability of a system to maintain all its states always reachable
  - a DTMC is irreducible
    if it is strongly connected and all its states are recurrent
  - a state is recurrent if the probability to eventually return back to it is 1
- a Bottom Strongly Connected Component (BSCC)
  is an irreducible subset of states

... more details about irreducibility - TBD

- Irreducibility relates to the ability of a system to maintain all its states always reachable, and can be characterized in terms of transient and recurrent states.
  A state r is recurrent if when reached once, With Probability 1 it will be reached again:
  $$\sum_{n=1}^{\infty} Prob\{first_r = n | X(0) = r\} = 1$$
  Note that this implies that, With Probability 1, r will be visited infinitely often.
  Conversely, a state t is transient if there is a non null probability that when that state is left it will not be reached anymore:
  $$\sum_{n=1}^{\infty} Prob\{X(n) = t | X(0) = t\} < 1$$
  Note that this implies that, With Probability 1, the state t will eventually become un-reachable.
  A DTMC is strongly connected when for any two states i and j there is a non null probability that starting from state i, state j will be eventually reached:

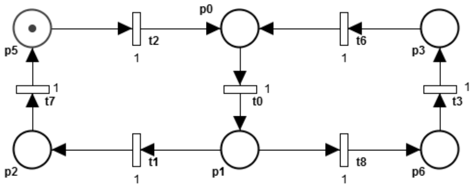
... an example about irreducibility

- a GSPN model with a reducible underlying DTMC (two BSCC)
  - two servers are initially active;
  - the first to fail will be abandoned
  - the other will be repaired after each failure



... more about aperiodicity

- Aperioditicy relates to the ability of a system to exhibit cyclic behaviors that are not constrained to have a length multiple of some period.
  - a DTMC is aperiodic if all its states are aperiodic
  - a state is periodic with period $T > 1$ if all its return times are multiple T
- a GSPN with a periodic underlying DTMC (period 4)



... more details about aperiodicity - TBD

- Aperioditicy relates to the ability of a system to exhibit cyclic behaviors that are not constrained to have a length multiple of some period. The concept is characterized in terms of return time.
  A state i accepts the return time m if there is a non-null probability that starting from i the process will come back to i after m steps:
  $Prob\{first_i = m|X(0) = i\} > 0$
  In general, a state may accept multiple different return times.
  The period of state i is the Maximum Common Divisor among all return times accepted by i.
  A state i is said to be periodic if its period is different than 1. It is said to be aperiodic if it is not periodic.
  TBD: una figura di una DTMC periodica. Poi una con la separazioen delle masse. e poi una con la riduzione di una DTMC peiodica alal forma aperiodica.
  To grasp the real meaning of the concept of periodicity, it can be useful considering what happens if a state i has period $K > 1$ and all the probability mass is initially concentrated in i (i.e. with certainty, the initial state is i). It is not said that after K steps the probability of i will be 1 again, as in general K is not the unique accepted return time.
  Conversely, it is always guaranteed that the probability of i will be 0 in all

steady state probabilities - TBD

- If steady state probabilities exist, they can be determined in straightforward manner by taking the limit on the left-hand-side and the right-hand-side of Eq.(**??**):

$$\lim_{n\to\infty} \underline{\pi}(n+1) = \lim_{n\to\infty} \underline{\pi}(n) \cdot P$$

which yields,

$$\underline{\pi} = \underline{\pi} \cdot P$$

Since P is stochastic, each of its rows sums up to 1 and P cannot have maximum rank so that this equation has (at least) ∞^1 "parallel" solutions. The following condition selects in the multiplicity the solution that satisfies the property of unit-measure:

$$\begin{cases} \underline{\pi} = \underline{\pi} \cdot P \\ |\underline{\pi}| = 1 \end{cases} \tag{1}$$

Note that, according to Eq.(1), if they exist, steady state probabilities are determined as the solution of a linear system.
Also note that Eq.(1) does not depend on the vector of initial probabilities. Which means that, if they exist, steady state probabilities do not depend on the initial distribution. In general, a system whose steady state behavior does not depend on the initial condition is said to be ergodic, which in its greek root means that energy will find its way

derived measures - TBD

- Il tempo di soggiorno SJ_i in uno stato i (inteso come numero di passi di permanenza nello stato) e' una variabile geometrica:

$$Prob\{SJ_i = k\} = p_{ii}^{k-1}(1 - p_{ii}) \tag{2}$$

e il tempo medio speso in i dall'arrivo all'uscita e' quindi:

$$E[SJ_i] = \frac{1}{1 - p_{ii}} \tag{3}$$

- Il tempo medio di ritorno dello stato j (ovvero il numero medio di passi tra due successivi ingressi nello stato j) e' l'inverso della sua probabilita' limite:

$$M_j = \frac{1}{\pi_j} \tag{4}$$

Sia $N_j(n)$ il numero di visite dello stato j entro i primi n passi di esecuzione della catena. $\frac{N_j(n)}{n}$ e' la frequenza di ingresso nello stato j e quindi:

$$\lim_{n \to \infty} \frac{N_j(n)}{n} = \pi_j$$

D'altra parte, se $M_j(n)$ denota il numero medio di passi tra due ingressi in j misurato sui primi n passi, il prodotto di $M_j(n)$ per $N_j(n)$ tende ad essere uguale al numero dei passi eseguiti n:

DTMC - more quantities of interest

- distribution of the first passage time in some state i,
  i.e. probability that state i has been visited within n steps

$$\eta_i(n) := Prob\{first_i \leq n\}$$
$$\text{with } first_i := \min_{k \in \mathbb{N}}\{k | X(k) = i\}$$

- absorption probabilities:
  probability that some (absorbing) state i is eventually visited

$$\eta_i := \lim_{n \to \infty} \eta_i(n)$$

DTMC - Absorption probabilities

- given a list of disjoint sets of states S_1, S_2, ... S_J, evaluate the probability that some state in S_j is reached before any state in any other set.
- as a particular occurrence, this is the problem encountered in the evaluation of the probability that a reducible DTMC reaches some of its Bottom Strongly Connected Components.
- The evaluation of absorption probabilities η_j can be reduced to the solution of a system of linear equations.

... more details on absorption probabilities - TBD

- The problem of first passage time (also called hitting time) is about evaluating the probability that at some time some state has already been visited.
- For a discrete time process, this amounts to evaluating $\eta_j(n) := Prob\{first_j \leq n\}$, where $first_j$ denotes the time of the first visit of state j.
- $\eta_j(n)$ can be evaluated through transient analysis on a modified model where state j is made absorbing by removing all its outgoing edges and adding a self-loop. In so doing, the probability that the modified model is in j at time n is equal to the probability that in the original model at time n the state j has already been visited.

... more details on absorption probabilities - TBD

- A related problem is the evaluation of the probability that some state j_1 is reached before some other state j_2. Or, more generally, given a list of disjoint sets of states S_1, S_2, ... S_J, evaluate the probability that some state in S_j is reached before any state in any other set. As a particular and notable occurrence, this is the problem encountered in the evaluation of the probability that a reducible DTMC reaches some of its Bottom Strongly Connected Components. This will also become relevant in the analysis of the stochastic process underlying a Generalized Stochastic Petri Nets with Immediate transitions.
  The probability that set S_j has been reached before any other set within time n can be evaluated on a modified DTMC where each set S_1, S_2, ... S_J with an absorbing state σ_1, ... σ_J: in this setting the transient probability $\pi_j(n)$ on the modified DTMC will represent; by taking n to ∞, the limit probability $\eta_j := \lim_{n\to\infty} \eta_j(n)$ will represent the probability that S_j is reached before any other set. Probabilities η_j are also called absorption probabilities.
  The evaluation of absorption probabilities η_j can be reduced to the solution of a system of linear equations. To this end, let s_i with $i \in [1, I]$ be the transient states, and σ_j with $j \in [1, J]$ the absorbing states, and let $\pi_i(n)$ and $\eta_j(n)$ be their probabilities. Finally, let p_{ih} be the transition

... more about the evaluation of steady state - TBD

- The characterization of the concepts of irreducibility and aperiodicity indicate that the conditions for the existence of a steady state solution given by Theorem **??** are in fact extremely fair to apply to any finite state and time-homogeneous DTMC. On the one hand, the problem of periodicity can be overcome by separating the analysis of each different periodic sub-DTMC. On the other hand, if the DTMC is reducible, the steady state problem must be formulated separately on each single Bottom Strongly Connected Component; each such component will have its own steady state limit (apart dealing with possible periodicity as above mentioned), and the limits of each BSCC will be combined according to their probability to be eventually reached from the initial state; this probability depends on the initial state probability distribution, so that the model will be not ergodic, and can be determined as a the solution of a problem fo first passage time.

Continuous Time Markov Chain - CTMC

- Continuous Time Markov Chain (CTMC):
  - a continuous time stochastic process $\mathbb{X} = \{X(t), t \in \mathbb{R}_{\geq 0}\}$
  - that is right-continuous

$$\forall t_0 \in \mathbb{R}_{\geq 0} \ \exists \lim_{t \to t_0^+} X(t) = X(t_0)$$

  - and always satisfies the Markov condition

$$\begin{aligned}
&\forall i, j, t, \tau, \\
&\forall M,\ i_M, i_{M-1}, ..., i_M, \forall \tau_M \geq \tau_{M-1} \geq ... \geq m_1 \geq 0, \\
&Prob\{X(t+\tau) = j | X(t) = i\} = \\
&Prob\{X(t+\tau) = j | X(t) = i, X(t-\tau_1) = i_1, ...X(t-\tau_M) = i_M\}
\end{aligned}$$

- the assumption of right-continuity *conceals* vanishing states (i.e. states with null sojourn time)
- we restrain the treatment to finite state and time-homogeneous CTMCs:

$$Prob\{X(t+\tau) = j | X(t) = i\} = Prob\{X(\tau) = j | X(0) = i\} \qquad (10)$$

Embedded DTMC of a CTMC - TBD

- In general, given a continuous time chain $\mathbb{X} = \{X(t), t \in \mathbb{R}\}$, a sequence of time points $\{t_n\}_{n=0}^{\infty}$ identifies an an embedded chain $\mathbb{X}^e = \{X(t_n), n \in \mathbb{N}\}$.
- The sequence $\{t_n\}_{n=0}^{\infty}$ can be defined in various ways fitting different objectives. For instance, we might assume that the time points are equidistant, i.e. $t_n = n * \delta$ where δ is a time tick. The most notable way to embed a discrete time chain $\mathbb{X}^e$ into a continuous time chain $\mathbb{X}$ is to assume that t_n is the time of the $n-$th transition in the execution of the process; this is often called "the" embedded chain of the process.
- Given a CTMC $\mathbb{M} = \{M(t), t \in \mathbb{R}\}$, any of its embedded chains is a DTMC, as it can be easily verified that the embedding preserves the Markov condition.

CTMC - quantities of interest

- we are interested in evaluating
  - transient probabilities
$$\pi_i(t) := Prob\{X(t) = i\}$$
  - steady state probabilities, if they exist
$$\pi_i := \lim_{t\to\infty} \pi_i(t)$$

CTMC - Characterizing quantities

- a CTMC is completely characterized by
  - the distribution of probability of the initial state

$$\pi_i^0 := \pi_i(0)$$

  - the infinitesimal generator matrix Q:

$$q_{ij} := \frac{d}{dt} p_{ij}(t)_{|_0} = \lim_{dt \to 0^+} \frac{Prob\{M(dt) = j | M(0) = i\}}{dt}$$

    where $p_{ij}(t)$ are continuous transition probabilities

$$p_{ij}(t) := Prob\{M(t) = j | M(0) = i\}$$

- a CTMC is conveniently figured out as a graph
  - each and every state is a vertex
  - the edge from state i to state j is labelled by the value q_{ij}
  - for some good reasons, self-loops are not significant and the diagonal elements of Q are not shown
  - TBD: concept of intensity in the flow of the probability mass
  - TBD: let some picture appear here, even the questionable example of the light Off/On/Failed
  - TBD: A finite and time-homogeneous CTMC can be conveniently represented as graph, where each state i is a vertex and each arc from vertex i to j is labeled with the element q_{ij} of the infinitesimal generator. This arc will represnet the flow of the mass of probability from state i to state j

CTMC - properties of the infinitesimal generator

- since $\mathbb{M}$ is right-continuous:

$$\forall i \neq j, P_{ij}(0) = 0 \text{ and } P_{ii}(0) = 1$$

- for any fixed i and t, $P_{ij}(t)$ is a discrete distribution:

$$\sum_{i=1}^{I} q_{ij}(t) = \frac{d}{dt} \sum_{i=1} P_{ij}(t) = \frac{d}{dt} 1 = 0$$

and thus:

$$\forall i, q_{ii} = - \sum_{j=1, j \neq i}^{I} q_{ij}$$

- since $P_{ij}(t)$ is non negative and $Pij(0) = 0$:

$$\forall i \neq j, q_{ij} \geq 0$$

on the relation between CTMCs and exponential random variables

- The Exponential distribution realizes the concept of memoryless behavior for a continuous time random variable. In a similar manner, the CTMC realizes the concept of Markovian behavior, where the future evolution depends on the current state but not on the past history through which the state was reached.
- in a CTMC, the sojourn time in state k is a random variable with negative exponential distribution with rate q_{kk}:

$$F_{SJ_k}(t) = 1 - e^{q_{kk}t}$$

proof: a consequence of the Markov condition, similar to the proof that a memoryless continuous time variable must be exponentially distributed

- if $\mathbb{M}$ is a CTMC with infinitesimal generator Q, and $\mathbb{M}^e$ is its embedded chain with transition matrix P:

$$p_{ij} = \frac{q_{ij}}{-q_{ii}}$$

proof: TBD

... more details

Theorem

For a time-homogeneous CTMC with infinitesimal generator Q, the sojourn time in any state i is an EXP variable with rate $-q_{ii}$.

prof: Let $H_i(t)$ be the holding time in state i:

$$H_i(t) := Prob\{t1 \geq t | M(0) = i\}$$

where t_1 denotes the time of the first transition after time 0.

$$\frac{dH_i(t)}{dt} = \lim_{dt \to 0} \frac{Prob\{t_1 > t+dt | M(0)=i\} - Prob\{t_1 > t | M(0)=i\}}{dt}$$
$$= \lim_{dt \to 0} \frac{-Prob\{t_1 \in [t, t+dt] | M(0)=i\}}{dt}$$

- By applying the Markov condition, this can be rewritten as:

$$= \lim_{dt \to 0} \frac{-Prob\{t_1 > t | M(0)=i\} \cdot Prob\{t_1 < dt | M(0)=i\}}{dt}$$
$$= H_i(t) \cdot \lim_{dt \to 0} \frac{\sum_{j=1, j \neq i}^{l} P_{ij}(dt) - P_{ij}(0)}{dt}$$

  where P_{ij} denotes the continuous transition matrix of the CTMC, and where the assumption of continuity implies that $j \neq i \to P_{ij}(0) = 0$.
- Moreover, by definition the definition and properties of the infinitesimal generator q:

... more details

Theorem

Given a time-homogeneous CTMC $\mathbb{M}$ *with infinitesimal generator Q, called* $\mathbb{M}^e := \{M(t_n), n \in \mathbb{N}\}$ *the embedded DTMC with* t_n *denoting the time of the n-th transition of* $\mathbb{M}$*, the transition probability* p_{ij} *of* $\mathbb{M}^e$ *is equal to*

$$p_{ij} = \frac{q_{ij}}{-q_{ii}}$$

prof: Let $G_{ij}(t)$ be the kernel defined as:

$$G_{ij}(t) := Prob\{t1 \leq t \wedge M(t_1) = j | M(0) = i\}$$

where t_1 denotes the time of the first transition after time 0.

$$\frac{dG_{ij}(t)}{dt} = \lim_{dt \to 0} \frac{Prob\{t_1 \leq t+dt \wedge M(t_1)=j | M(0)=i\} - Prob\{t_1 \leq t \wedge M(t_1)=j | M(0)=i\}}{dt}$$
$$= \lim_{dt \to 0} \frac{Prob\{t_1 \in [t, t+dt] \wedge M(t_1)=j | M(0)=i\}}{dt}$$

By applying the Markov condition, and then the definition of infinitesimal generator, this can be rewritten as:

$$= H_i(t) \cdot q_{ij} = e^{q_{ii} t} \cdot q_{ij}$$

We can thus finally write the following differential equation

CTMC - transient analysis - TBD

- Chapman-Kolmogorov equations

$$\frac{d}{dt}\pi_j(t)|_0 = \ldots$$

proof: a consequence of the law of total probability and the Markov condition

- in vectorial form

$$\underline{\pi}'(t) = \underline{\pi}(t) \cdot Q$$

Chapman Kolmogorov equations - TBD

- Given the initial state probability distribution $\underline{\pi}^0$ and the infinitesimal generator Q, transient probabilities $\underline{\pi}(t)$ are determined by Chapman-Kolmogorov equations:

$$\begin{aligned}
&\frac{d}{dt}\pi_j(t)|_0 = \lim_{dt\to 0} \frac{Prob\{M(t+dt)=j\}-Prob\{M(t)=j\}}{dt} = \\
&\lim_{dt\to 0} \frac{\sum_{i=1}^{l} Prob\{M(t+dt)=j \wedge M(t)=i\} - Prob\{M(t)=j \wedge M(t)=i\}}{dt} = \\
&\lim_{dt\to 0} \frac{\sum_{i=1}^{l} (Prob\{M(t+dt)=j|M(t)=i\} - Prob\{M(t)=j|M(t)=j\}) \cdot Prob\{M(t)=i\}}{dt} = \\
&\sum_{i=1}^{l} q_{ij} \cdot \pi_i(t)
\end{aligned} \tag{11}$$

- Note that this comprises an application of the law of total probability under the assumption of the Markov condition and time-homogeneity: the set of events that can lead to $M(t+dt)=j$ is partitioned into (disjoint and total) events $(M(t)=i) \wedge (M(t+dt)=j|M(t)=i)$; the total probability is obtained by summing up the probability of each event; the Markov condition is applied to express the probability of $(M(t+dt)=j|M(t)=i)$ independently from states visited before t; time-homogeneity is applied to express the probability of transition $(M(t+dt)=j|M(t)=i)$ independently of t.
- In vectorial form, Eq.(11), can be rewritten as:

$$\underline{\pi}'(t) = \underline{\pi}(t) \cdot Q \tag{12}$$

CTMC - Steady state analysis - TBD

- for a finite, time-homogeneous, and irreducible CTMC,
  steady state probabilities exist,
  and they do not depend on the initial distribution (ergodicity)
- derive the solution directly from Chapman Kolmogorov equations ...
- system of linear equations ... TBD

Steady state analysis - TBD

Theorem

For a finite, time-homogeneous, and irreducible CTMC, steady state probabilities exist unique, i.e. there exists unique the limit:

$$\pi_i = \lim_{t \to \infty} \pi_i(t)$$

- Note that for a reducible CTMC, absorption probabilities in each Bottom Strongly Connected Component can be computed on the embedded chain, and Theorem3 can then be applied in each BSCC. To this end, it is relevant that the steady state solution oes not depend on the initial probability distribution, which makes it insensitive to the particular state where the absorption in a BSCC occurs.
  TBD ... se la CTMC che stiamo studiando deriva da una GSPN l'ipotesi di finitezza diventa che la GSPN abbia finite marcatire raggiungibili; l'ipotesi time-homogeneous segue direttamente da avere i paraetri costanti sul modello (i.e. la topologia della rete e il parametro R(t,m) sono immutabili).
  TBD .... una volta che il teorema si applica, poi determinare la soluzione e' facile perche' se esiste steady state allora la derivata del vettore delle

CTMC - transient analysis - TBD

- analytic form of transient probabilities as exponential matrix ... TBD
- uniformization ... TBD

Transient analysis

- Given the vector of initial probabilities $\underline{\pi}^0$, Eq.(12) supports derivation of transient probabilities ...
  ...TBD... forma con matrix exponential... usando la decomposizione di Jordan si vede che le componenti sono expolinomiali con tassi corrispondenti alla parte reale degli autovalori di Q e gradi polinomiali dipendenti dal grado di confluenza degli autovettori... si calcola con 21 dubious ways ... in generale ha alcuni difetti che ne rendono instabile la truncation: Q non è stocastica, questo è un sviluppo di Taylor e quindi funziona male quando ci si allontana da 0, e comunque sono polinomi che prima o poi esplodono.
  ... TBD ... esiste un altro approccio che è quello che funziona: uniformization ... scomporre la prob di essere in i al tempo t in essere in i dopo n passi e avere n passi entro t sommando per n=0,inf (total probability law), ... poi la prob di essere in i dopo n passi e' indipendente dal tempo speso per fare gli n passi (poprrieta' di Markov e notably il fatto che su una exp il tempo trascorso è indipendente da come è risolta una race) ... resta il problema che il numero di passi fatti entro il tempo t non si esprime in forma chiusa ... il problema è che ciascuno stato ha un diverso tempo di soggiorno e quindi il numero di passi fatti dipende dal percorso nella storia passata ... se aggiungiamo dei self-loops viene

CTMC - embedded chain

- let $\mathbb{M} := \{M(t), t \in \mathbb{R}_{geq0}\}$ be a CTMC,
  and let $\mathbb{M}^e := \{M(n), n \in \mathbb{N}\}$ be the discrete time chain that samples $\mathbb{M}$ at each state transition
- $\mathbb{M}^e$ always satisfies the Markov condition and is thus a DTMC
- this is usually called "the" embedded chain of $\mathbb{M}^e$,
  but other embedded chains could be built as well with other samplings (e.g. sample at equidistant time points, at each visit of a set of states,...)

underlying stochastic process of a GSPN

- the continuous time marking process $\mathbb{M} := \{M(t), t \in \mathbb{R}_{\geq 0}\}$ of a GSPN always satisfies the Markov condition

proof: according to the semantics of GSPNs, given the current marking, the distribution of the time to the next firing and the probability of each feasible next event are determined, and they are not conditioned by any previous observation

  - $\mathbb{M}$ is thus a CTMC
  - since $\mathbb{M}$ is abstracted so as to be right continuous, vanishing states with enabled IMM transitions are concealed

- also the discrete time marking process $\mathbb{M}^d := \{M(n), n \in \mathbb{N}\}$ of a GSPN always satisfies the Markov condition
  - and thus comprises a DTMC
  - which observes also vanishing states, but misses continuous time durations
  -

- a full characterization of the stochastic process underlying a CTMC requires that both $\mathbb{M}$ and $\mathbb{M}^d$ be characterized
  - derive the infinitesimal generator Q of $\mathbb{M}$
  - the transition matrix P of $\mathbb{M}^d$
  - and the distribution of initial probabilities π^0

The marking process of a GSPN - 1/2

- let $\mathbb{M}^e$ and $\mathbb{M}$ be the discrete time and continuous time marking processes of a GSPN, respectively:

  $\mathbb{M}^e := \{M(n), n \in \mathbb{N}\}$
  with $M(n) :=$ marking after n firings

  $\mathbb{M} := \{M(t), t \in \mathbb{R}\}$
  with $M(t) :=$ marking at time t

- Since in the GSPN all transitions are either IMM or EXP, the time elapsed in the current marking does not condition the future evolution of the net, i.e. the marking completely determines the stochastic characteristics of the future behavior independently of the past history.
- According to this, $\mathbb{M}^e$ is a DTMC and $\mathbb{M}$ is a CTMC.
- The two stochastic processes are completely characterized as soon as we have determined the transition probabilities of the DTMC and the rate of the Exponential distribution of the sojourn time in a state of the CTMC.

The marking process of a GSPN - 2/2

Theorem

The states of the DTMC $\mathbb{M}^e$ are all and only the reachable markings, and the transition matrix P is defined as:

$$p_{ij} = \frac{\sum_{m_i \xrightarrow{t_h} m_j} R(t_h, m_i)}{\sum_{m_i \xrightarrow{t_h}} R(t_h, m_i)}$$

Theorem

The states of the CTMC $\mathbb{M}^e$ are all and only the tangible reachable markings, and the rate of the sojourn time in the state i is the sum of the rates of the EXP transitions enabled in the tangible marking m_i, i.e. the element q_{ii} of the infinitesimal generator is:

$$q_{ii} = \sum_{m_i \xrightarrow{t_h}} R(t_h, m_i)$$