Markov-modulated arrival processes in queueing theory

Alain Jean-Marie
INRIA et LIRMM, University of Montpellier 2
161 Rue Ada, 34392 Montpellier Cedex 5, France
ajm@lirmm.fr

Lunteren Conference
January 2005
Plan of the talk

Introduction

- Modeling the traffic of networks
- Markov chains and Markov calculus

Markov-modulated arrival processes

- discrete: MMPP, MAP, BMAP
- continuous: MMRP
- generalization: a Semi-Markovian accumulation process
Decomposition of Markov-Modulated sources

- Markov chains with Markov-modulated speeds
- The MMPP/GI/1 queue
- Equivalent Bandwidth
The mathematical modeling of computer & communication systems necessitates an accurate representation of the arrival process of information/workload.

Depending on the level of the model, this may be:

- the quantity of packets arrived in some network element before some time $t$,
- a quantity of frames (video), requests (transactions), or any other network Application Data Unit, tasks (computing), orders (production),
- a quantity of bytes or bits, or CPU seconds.
Mathematical models of arrivals

The appropriate mathematical object is a counting process:

\[ N(t) = \text{quantity arrived in the interval } [0, t) . \]

Several cases:

- **discrete time**: \( t \in \mathbb{N} \)
- **continuous time**: \( t \in \mathbb{R} \)
- **discrete space**: \( N(t) \in \mathbb{N} \)
- **continuous space**: \( N(t) \in \mathbb{R} \)
Counting process: illustration

Process of arrivals of \textit{events} (arrivals, departures, changes, starts, stops, etc).

\[ N(t) \]

\[ t \]
Modeling constraints

The variety of situations makes the following features necessary:

- relatively complex processes (bursts, temporal correlations, ...)
- possibly large number of sources
- ease of use, for simulation and stochastic calculus: distributions, queueing networks, asymptotics...

... with a mastered algorithmic complexity.

→ Markov-modulated processes have these features
Markov chains

A discrete-time Markov chain is a process \( \{X(n), n \in \mathbb{N}\} \) such that:

- if \( X(n) = i \), then \( X(n + 1) = j \) with probability \( p_{i,j} \),
- jumps are independent.

A Markov chain is fully described by its

transition probabilities: \( p_{i,j}, (i, j) \in \mathcal{E} \times \mathcal{E} \), or its

transition matrix \( \mathbf{P} \).
Example of Markov chain

Transition diagram

Transition matrix

\[ P = \begin{pmatrix} 0.2 & 0.2 & 0.6 \\ 0 & 0.5 & 0.5 \\ 1 & 0 & 0 \end{pmatrix}. \]
Continuous time Markov chains

Let \( \{X(t), t \in \mathbb{R}^+\} \), having the following properties. When \( X \) enters state \( i \):

- \( X \) stays in state \( i \) a random time, exponentially distributed with parameter \( \tau_i \), independent of the past; then

- \( X \) jumps instantly in state \( j \) with probability \( p_{ij} \). We have \( p_{ij} \in [0, 1] \), \( p_{ii} = 0 \) and

\[
\sum_j p_{ij} = 1.
\]

This process is a continuous-time Markov chain with transition rates

\[
q_{ij} = \tau_i p_{ij}.
\]
\[ \tau = \begin{pmatrix} 0.3 \\ 1 \\ 0.6 \end{pmatrix} \quad \mathbf{P} = \begin{pmatrix} 0 & 1 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{2}{3} & 0 \end{pmatrix} \quad \mathbf{Q} = \begin{pmatrix} -0.3 & 0.3 & 0 \\ 0.5 & -1.0 & 0.5 \\ 0.2 & 0.4 & -0.6 \end{pmatrix}. \]
Properties and Analysis

From the computational point of view, the most useful properties of Markov processes are:

- they are described by matrices,
- computing distributions involves the solution of linear problems
- their superposition and composition leads to simple matrix computations.
Superposition of sources

If one superposes several Markov-modulated sources, the resulting process is still Markov-modulated.

The matrices (generators and rates) are obtained using Kronecker sums.

Kronecker product: consider two matrices $A \ (n \times n)$ and $B \ (m \times m)$. Their Kronecker product is a matrix $nm \times nm$ with

$$A \otimes B = \begin{pmatrix}
A_{11}B & \cdots & A_{1n}B \\
\vdots & \ddots & \vdots \\
A_{n1}B & \cdots & A_{nn}B
\end{pmatrix}. $$
**Kronecker sum:** a matrix $nm \times nm$ defined as

$$A \oplus B = A \otimes I(m) + I(n) \otimes B$$

$$= \begin{pmatrix} A_{11}B & \cdots & \cdots & A_{nn}B \\ \vdots & \ddots & \vdots & \vdots \\ \cdots & \cdots & \ddots & \cdots \\ \cdots & \cdots & \cdots & A_{nn}B \end{pmatrix} + \begin{pmatrix} B_{11}I & \cdots & \cdots & B_{1m}I \\ \vdots & \ddots & \vdots & \vdots \\ \cdots & \cdots & \ddots & \cdots \\ \cdots & \cdots & \cdots & B_{nn}I \end{pmatrix}.$$
Example: for two Markov chains \( \{X_1(t)\} \) and \( \{X_2(t)\} \), we have:

\[
\begin{pmatrix}
-\lambda & \lambda & 0 \\
0 & -\mu & \mu \\
\nu & 0 & -\nu
\end{pmatrix}
\oplus
\begin{pmatrix}
-\alpha & \alpha \\
\beta & \beta
\end{pmatrix}

= \begin{pmatrix}
\alpha & 0 & 0 \\
0 & \alpha & 0 \\
\nu & 0 & -
\end{pmatrix}
\begin{pmatrix}
- \lambda & 0 \\
0 & - \mu \\
\beta & 0
\end{pmatrix}
\begin{pmatrix}
\alpha & 0 & 0 \\
0 & \alpha & 0 \\
\nu & 0 & -
\end{pmatrix}
\begin{pmatrix}
\beta & \beta
\end{pmatrix}
\begin{pmatrix}
- \lambda & 0 \\
0 & - \mu \\
\nu & 0 & -
\end{pmatrix}
\end{pmatrix}
\]
Markov modulated speeds

Consider a Markov chain $Z$ which evolves in some state space with a generator $M = (m_{ab})$.

There is an "environment" $X$ which is a CTMC with generator $G = (g_{ij})$.

When $X$ is in state $i$, the speed of $Z(t)$ (transition rates) is multiplied by $v_i$:

$$\text{rate } a \rightarrow b = m_{ab} \times v_i.$$ 

The generator of the process $(Z(t), X(t))$ has transition rates:

$$
\begin{align*}
(i, a) & \rightarrow (i, b) \quad \text{with rate } m_{ab}v_i \\
(i, a) & \rightarrow (j, a) \quad \text{with rate } g_{ij}
\end{align*}
$$
In block-matrix form:

\[
Q = \begin{pmatrix}
  v_1M + g_{11} & g_{12} & \ldots & g_{1K} \\
g_{21} & v_2M + g_{22} & \ldots & g_{2K} \\
\vdots & \vdots & \ddots & \vdots \\
g_{K1} & g_{K2} & \ldots & v_KM + g_{KK}
\end{pmatrix}
\]

Or, with the Kronecker notation:

\[
Q = G \otimes I + V \otimes M.
\]

where

\[
V = \text{diag}(v_1, \ldots, v_K).
\]
Plan of the talk

Introduction ........................................................................................................3

Markov-modulated arrival processes .................................................................17
  • discrete: MMPP, MAP, BMAP
  • continuous: MMRP
  • generalization: a Semi-Markovian accumulation process

Decomposition of Markov-Modulated sources .................................................38
Markov modulated arrivals

General idea:

• A Markov chain \( \{X(t); t \in \mathbb{R} \text{ or } \mathbb{N}\} \in \mathcal{E} \), the phase

• A counting process \( N(t) \) such that \( \{(X(t), N(t))\} \in \mathcal{E} \times \mathbb{N} \) is a Markov chain.
Let \( \{X(t); t \in \mathbb{R}\} \) be a continuous-time Markov chain.

\( \{N(t); t \in \mathbb{R}\} \) counts the number of jumps of \( X \) in \([0, t)\).
Let \{X(t); t \in \mathbb{R}\} be a continuous-time Markov chain in \mathcal{E}.

Let \(\lambda_i \geq 0\) be an arrival rate, for each \(i \in \mathcal{E}\).

Arrivals occur according to a Poisson process of time-varying rate \(\lambda_{X(t)}\): that is, \(\lambda_i\) as long as \(X(t) = i\).
BMAP: Batch Markov Arrival Process

Also known as “N-process” (N = Neuts), or the “versatile” process.

\{(X(t), N(t)); t \in \mathbb{R}\} is a continuous-time Markov chain with a generator structured as:

\[
Q = \begin{pmatrix}
D_0 & D_1 & D_2 & \ldots \\
D_0 & D_1 & D_2 & \ldots \\
D_0 & D_1 & \ldots & \ldots \\
\end{pmatrix}
\]

A process in the family of Markov additive process.
MMRP: Markov Modulated Rate Process

Let \( \{X(t); t \in \mathbb{R}\} \) be a continuous-time Markov chain over a finite state space \( \mathcal{E} \).

Let \( r_i \) be arrival rates (or accumulation rates), for each \( i \in \mathcal{E} \).

Arrivals occur according to a \textit{fluid} process with rate \( r_{X(t)} \), that is: with rate \( r_i \) as long as \( X(t) = i \).

Let \( N(t) \) the quantity arrived at time \( t \):

\[
\frac{dN}{dt}(t) = r_{X(t)}.
\]

Note: also known as “Markov drift process”.
Example. $\mathcal{E}$ with three states, $0 < r_1 < r_2$, $r_3 = 0$:
On/Off processes:

- alternating periods On and Off, with IID durations
- while in period On, arrivals according to a fluid process (constant rate) or a discrete process (Poisson or periodic).
Elaborate multiscale processes

Process with arrivals of sessions, requests, packets:

can be modeled as well with hierarchical Markov-modulated arrival processes.
Synthesis

Markov modulated sources of arrivals are described by matrices.

- For a MAP:
  
  the generator $Q$

- For a MMPP/MMRP:
  
  the generator $Q$, and the rate matrix $\Lambda$

- For a BMAP:
  
  the collection of transition rate matrices $D_0, D_1, \ldots$

Most distributions and performance measures are computed using these matrices.
Examples of computations

Average arrival rate

For a MMPP/MMRP, with $\pi$ the stationary probability of $X$,

$$\bar{\lambda} = \pi \Lambda 1 = \sum_{i \in \mathcal{E}} \pi_i \lambda_i.$$

Distribution of arrivals

For a MMPP, if $A_{ij}(k, T) = \mathbb{P}\{k \text{ arrivals and } X(T) = j \mid X(0) = i\}$, then

$$\sum_{k} z^k A_{ij}(k, T) = \left(e^{(Q-(1-z)\Lambda)T}\right)_{ij}.$$
A generalization:

- Start with a **semi-Markov process**: arbitrarily distributed but state-dependent sojourn times, probabilistic jumps.

- Let the quantity accumulate at a “rate” depending on the state,

- plus random increments at jump times
\[ A(t) \]
The process of accumulation is an **independent-increments process**:

- **constant-rate**
- **Poisson**
- **diffusion**

or a mixture of them.
For independent-increment processes, it is known (e.g. Doob (1952)) that:

$$\mathbb{E}(e^{-\nu(x(t)-x(s))}) = e^{-(t-s)\phi(\nu)}.$$ 

For instance:

$$\phi(\nu) = r\nu \quad \text{for a constant-rate accumulation } r$$

$$\phi(\nu) = r(1 - e^{-\nu}) \quad \text{for a Poisson process with rate } r$$

$$\phi(\nu) = r\nu + \frac{1}{2}\sigma^2\nu^2 \quad \text{for a diffusion process with drift } r \text{ and variance } \sigma^2.$$
Distribution of the accumulated quantity

$Q(T)$ being the quantity accumulated at time $T$, consider the Laplace transform:

$$K_{i,j}(\mu, \nu) = \int_0^\infty e^{-\mu T} \int_0^\infty e^{-\nu x} \mathbb{P}\{Q(T) \leq x, X(T) = j | X(0) = i\} \, dx \, dT$$

$$K = (K_{i,j}(\mu, \nu))_{(i,j) \in \mathcal{E} \times \mathcal{E}}$$

$$S = \text{diag} \left( S_i^*(\mu + \phi_i(\nu)) \right)_{i \in \mathcal{E}}$$

$$L = \text{diag} \left( \frac{1}{\mu + \phi_i(\nu)} \right)_{i \in \mathcal{E}}$$

Then (standard arguments, e.g. Cox & Miller (1965) for $K = 2$):

$$K = L \, (I - S) + SPK$$

$$K = (I - SP)^{-1} \, L \, (I - S)$$
Plan of the talk

Introduction

Markov-modulated arrival processes

Decomposition of Markov-Modulated sources

- Markov chains with Markov-modulated speeds
- The MMPP/GI/1 queue
- Equivalent Bandwidth
Decomposition of sources

Principle:

- some source of information is composed of several simpler Markov-modulated sources,
- some computation is required (transients, autocorrelations, distribution of a queue, asymptotics, ...)
- Q: is it possible to reduce the computation to that with the smaller sources?
- A: yes: sometimes, a complexity gain is obtained, sometimes even a full decomposition.

Markov modulated speeds

Consider again the Markov chain $Z$ with generator $M$, modulated by a speed process with generator $G$, and speeds $V$. We have seen that:

$$Q = G \otimes I + V \otimes M.$$ 

**Problem:** compute the transition probabilities, which are the elements of the matrix $e^{Qt}$. A standard method is to diagonalize $Q$: find its eigenvalues and eigenvectors.
\[ Q = G \otimes I + V \otimes M. \]

If one chooses \( x \) and \( y \) such that:

\[
\begin{align*}
  x \ M &= \lambda x \\
  y &= (a_1 x, \ldots, a_N x) = a \otimes x.
\end{align*}
\]

Then

\[
\begin{align*}
  y \ Q &= (a \otimes x) \ (G \otimes I + V \otimes M) \\
        &= aG \otimes xI + aV \otimes xM \\
        &= a \ (G + \lambda V) \otimes x.
\end{align*}
\]

It is enough to choose \( a \) such that \( a(G + \lambda V) = \mu a \) for \( yQ = \mu y \) to hold.
Diagonalization Algorithm

- Find the spectral elements of $M$:

  $\rightarrow (\lambda_i; x_i, y_i) \quad i = 1..K$.

- For each $i$, find the spectral elements of $G + \lambda_i V$:

  $\rightarrow (\mu_{ij}; a_{ij}, b_{ij}) \quad i = 1..K, \ j = 1..N$.

- Obtain the spectral elements of $Q$:

  $\rightarrow (\mu_{ij}; a_{ij} \otimes x_i, b_{ij} \otimes y_i) \quad i = 1..K, \ j = 1..N$. 
Complexity:

- so let $N$ be the size of the state space, $K$ the number of speeds

- $Q$ is of size $NK \times NK$

- diagonalizing directly is $O(N^3K^3)$

- this algorithm is $O(K^3 + KN^3)$.

It is not even necessary to store the “big” matrix.
Markov modulated queues

Discrete queues: Markov-modulated arrivals

- exponential/Erlang/Cox service distribution $\rightarrow$ method of phases, QBDs
- general IID services: method of the embedded Markov chain.

Fluid queues:

- partial differential equations (Chapman-Kolmogoroff).
In both cases, the results are:

- Computation through matrix formulas, generating functions, Laplace transforms.

- Spectral expansions of stationary and transient probabilities:

\[
P\{W > x; X = i\} = \sum_p a_{i,p} e^{-z_p x}.
\]

→ asymptotics, or bounds.

\[
P\{W > x; X = i\} \sim a_{i,1} e^{-z_1 x}, \quad x \to \infty.
\]
The MMPP/GI/1 queue

Arrivals: MMPP with $N$ states, generator $Q$ and matrix of rates $\Lambda$;

Services: independent with a general distribution $H(x)$, of Laplace transform $H^*(s)$.

Distribution of the workload $W$:

$$W^*(s) = s(1 - \rho) \ g \ [sI + Q - (1 - H^*(s))\Lambda]^{-1} \ 1,$$

$g$ vector to be determined.
This requires diagonalizing $sl + Q - (1 - H^*(s)) \Lambda$, which can be done more efficiently using the fact that if:

$$A = A^{(1)} \oplus \ldots \oplus A^{(K)},$$

and that for all $k$, $A^{(k)}$ is diagonalizable with

$$A^{(k)} = R^{(k)} D^{(k)} S^{(k)},$$

where $R^{(k)} S^{(k)} = I^{(k)}$ and $D^{(K)} = \text{diag}(\omega_{i}^{(k)})$. Then:

$$A = \left( \bigotimes_{k=1}^{K} R^{(k)} \right) \left( \bigoplus_{k=1}^{K} D^{(k)} \right) \left( \bigotimes_{k=1}^{K} S^{(k)} \right).$$

This work since $Q$ and $\Lambda$ have precisely this structure.

$\implies$ complexities reduced from $(\sum_k N_k)^3$ to $\sum_k N_k^3$. 

Queues
Consider the \textbf{multiplexing problem}: \( K \) sources feed a buffer with \textit{finite buffer space} \( B \) and \textit{service capacity} \( C \) units of work/s.
For each source \( k \), let \( \rho_k \) be the average rate of arrival of information (the “bandwidth”). Then the queue with infinite buffer is stable if and only if

\[
\sum_k \rho_k < C.
\]

But for the overflow probabilities

\[
\Pr\{W^B = B\} \sim \Pr\{W^\infty \geq B\}
\]

is there a similar property?
Yes, for Markov-Modulated sources.

Assume source \( k \) has rate matrix \( \mathbf{L}^{(k)} \) and generator \( \mathbf{Q}^{(k)} \).

Let \( g^{(k)}(z) \) be the largest eigenvalue of \( \mathbf{L}^{(k)} - \frac{1}{z}\mathbf{Q}^{(k)} \).

For \( B \) large and \( \alpha \) small,

\[
\mathbb{P}\{W^\infty \geq B\} \leq \alpha \iff \sum_k g^{(k)} \left( \frac{\log(\alpha)}{B} \right) \leq C.
\]

The quantity \( g^{(k)} \left( \frac{\log(\alpha)}{B} \right) \) is the equivalent bandwidth at level \( \log(\alpha)/B \).

Proved by Elwalid and Mitra, generalized by Kulkarni for general Markov-Renewal sources.
Bibliography

Fluid models


Bibliographie


**MMPP, MAP, BMAP...**


Asymptotics, bounds and equivalent bandwidth


