Passage-Time and Transient Distributions in Semi-Markov Processes

J.T.Bradley W.J.Knottenbelt P.G.Harrison
Department of Computing, Imperial College, UK*

C.J.Vowden†
Degree 2 Innovations Ltd., UK

Abstract

We discuss passage-time and transient analysis for a particular class of non-Markovian model—the semi-Markov process. We present definitions of first passage-time distributions between clusters of states in a semi-Markov process. The paper then investigates an algorithm for approximating the passage-time distribution which is potentially $O(n^2)$. Finally we touch on how this might be applied to reducing the complexity of transient analysis of SMPs.

1 Motivation

Transient and passage-time distributions are two of the more detailed and potentially useful metrics that can be derived from a stochastic system. Unlike in purely Markov processes, their calculation in the semi-Markov domain is a computationally complex task and therefore rarely carried out for systems of any size.

1.1 Transient distributions

Transient distributions represent the probability that a system will be in a specific state at a time t. Steady-state distributions, on the other hand, only represent the long-run average probability that a system will be in a given state. The steady-state distribution can often (but not always, see section 2.1 and appendix A) be obtained by taking the limit, $t \longrightarrow \infty$, of the transient distribution.

Combining a reward vector with a transient distribution, in much the same way as is done with steady-state distributions, will provide a time-varying metric which conveys much more information than the equivalent steady-state metric.

*180 Queen's Gate, London SW7 2BZ, UK. Email: {jb, wjk, pgh}@doc.ic.ac.uk
†Regus House, 1 Friary, Bristol, BS1 6EA, UK. Email: Chris.Vowden@degree2.com

1.2 Passage-time distributions

Passage-time distributions[1] represent the distribution of how long it takes for a system to go from one state to another. This can be generalised to include distributions from one state to a cluster of states and further from one cluster to another cluster of states.

This can be most useful when analysing, for example, response-time in a queueing node or recovery-time in a network protocol. According to the system definition, the passage-time can then be used to give so-called quantile metrics, the probability that the network bandwidth will drop below 6Mbits^{-1} for example.

Passage-time distributions can be obtained using transient techniques. Likewise transient distributions can be obtained from passage-time distributions, so a good approximate method for finding either should be extensible to the other.

In this paper, we are going to present an approximate iterative method for finding the first passage-time distribution of an SMP and demonstrate how transient distributions be obtained, having got the passage-time information. We will briefly look at two examples of the approximate passage-time technique in action.

2 Transient Analysis

2.1 The steady-state limit

Formally the transient distribution is defined as:

$$T_{ij}(t) = \mathbb{P}(\text{system in state } j \text{ at time } t \text{ having started in state } i) \tag{1}$$

In systems with *pleasant* distributions, we can express the steady-state probability as the limit of the transient probability (using property (a) of [2, p.171]):

$$\lim_{t \longrightarrow \infty} T_{ij}(t) = \pi_j \tag{2}$$

However a broader convergence definition can be obtained by looking at the proportion of time a system spends in a given state (by property (c), [2, p.172]):

$$\lim_{t \longrightarrow \infty} \frac{1}{t} \int_0^t T_{ij}(\tau)\,\mathrm{d}\tau = \pi_j \tag{3}$$

which includes the convergence of equation (2). A proof of this can be found in appendix A. In both cases the steady-state vector is independent of the start state.

[1] Throughout the paper we are referring to first passage-time distributions.

2.2 Traditional technique

This section briefly outlines the traditional [3] analytic method for calculating the transient distribution of an SMP. An n-state semi-Markov process consists of a discrete-time Markov chain, P, and a probability distribution matrix F of sojourn times between states. The matrix of transient probabilities is given analytically in Laplace form by:

$$\hat{T}(s) = \frac{1}{s}(I - \hat{Q}(s)/s)^{-1}(I - \hat{K}(s)/s) \tag{4}$$

where $\hat{A}(s)$ represents the Laplace transform of all the functions in a matrix $A(t)$ and:

$$\begin{aligned}
T_{ij}(t) &= \text{transient probability of system in state } j \text{ having started in state } i \\
Q_{ij}(t) &= \text{weighted pdf from state } i \text{ to state } j \\
&= P_{ij}F_{ij}(t) && (5)\\
K_{ij}(t) &= \delta_{ij}h_i(t) && (6)\\
h_i(t) &= \text{distribution from state } i \\
&= \sum_{j=1}^{n} Q_{ij}(t) && (7)
\end{aligned}$$

The inversion of the matrix $(I - \hat{Q}(s)/s)$ is the sticking point here. The complexity of the inversion is n^3 anyway, but having to do this analytically makes the practical limit about 10–20 states with current symbolic algebra packages.

Our aim is to use approximation to reduce the complexity if possible and ultimately avoid maintaining a symbolic representation at all.

2.2.1 Transient Distributions from Passage-time Distributions

Also from Pyke's classic semi-Markov paper [3], we have the following relationship between transient functions, $T_{ij}(t)$, and passage-time functions, $L_{ij}(t)$.

$$\hat{T}_{ij}(s) = \begin{cases} \frac{1-\hat{h}_i(s)/s}{1-\hat{L}_{ii}(s)} & : \text{if } i = j \\ \hat{L}_{ij}(s)\hat{T}_{jj}(s) & : \text{if } i \neq j \end{cases} \tag{8}$$

Now, we develop our approximation for the passage-time distribution in the section below. Using this approximation, we will be able to apply it to this formula to find transient functions with the same complexity of solution.

3 Passage-time Analysis

3.1 Some Definitions

We consider a passage-time distribution between two states i and j, $\hat{L}_{ij}(s)$. Take the convolutions of all the distributions out of i to k (successor states of i) with the respective passage-time distributions from k to j. To get our required passage-time, sum the convolutions to cover all possibilities and get a set of linear algebraic equations. The terminating case occurs when j is a successor state of i

$$\hat{L}_{ij}(s) = \sum_{k \neq j} \hat{Q}_{ik}(s) \hat{L}_{kj}(s) + \hat{Q}_{ij}(s) \tag{9}$$

Where there are multiple destination states as represented by the destination vector, $\vec{j}$:

$$\hat{L}_{i\vec{j}}(s) = \sum_{k \notin \vec{j}} \hat{Q}_{ik}(s) \hat{L}_{k\vec{j}}(s) + \sum_{k \in \vec{j}} \hat{Q}_{ik}(s) \tag{10}$$

When there are multiple source states, denoted by the vector $\vec{i}$, the Laplace transform of the passage-time distribution at steady-state is:

$$\hat{L}_{\vec{i}\vec{j}}(s) = \sum_{k \in \vec{i}} \alpha_k \hat{L}_{k\vec{j}}(s) \tag{11}$$

where the weight α_k is the steady-state probability that the system is in state $k \in \vec{i}$ at the starting instant of the passage. As the starting vector of states does not necessarily include all the states α_k is a renormalised version of the standard steady-state vector, $\tilde{\pi}$, as defined for a semi-Markov chain.

$$\alpha_k = \frac{\tilde{\pi}_k}{\sum_{j \in \vec{i}} \tilde{\pi}_j} \tag{12}$$

This means that if a semi-Markov process is not irreducible (i.e. a steady-state vector does not exist) and multiple source states are required then a similar normalised weighting vector must also be specified to produce a meaningful weighted sum of the otherwise distinct distributions.

3.2 Approximating an SMP Passage-time Distribution

If we are trying to find a (first) passage-time distribution from i to the set of states $\vec{v}$, we take the matrix Q_{ij} and construct a second matrix R_{ij} by:

$$R_{ij} = \begin{cases} Q_{ij} & : \text{if } i \notin \vec{v} \\ 0 & : \text{if } i \in \vec{v} \end{cases} \tag{13}$$

Now consider the sequence:

$$\hat{R}(s), \hat{R}^2(s), \hat{R}^3(s), \hat{R}^4(s) \ldots \tag{14}$$

The rth element of the sequence represents the rth-step transition matrix of a system which is absorbing in the vector of states, $\vec{v}$. So $\hat{R}^r_{ij}(s)$ is the distribution of going from state i to j in r steps after one passage of the system. An alternative definition of the first passage time from i to j might therefore be:

$$X(s) = \sum_{i=1}^{\infty} R^i(s) \tag{15}$$

The convergence of this equation will depend on the size (in number of transitions) of the largest cycle within the system as well as the mean time to make that cycle (similar to uniformization for Markov processes). By and large, we can approximate $X(s)$ with $X^{(r)}(s) = \sum_{i=1}^{r} R^i(\mathrm{s})$ with relative confidence, as we will see in the examples in the next section.

$X^{(r)}(s)$ can be used to approximate the passage-time functions in equation (8) and thus the transient functions themselves, $T_{ij}(t)$.

4 Examples

We present two examples of our passage-time approximation in action. The first is a purely Markovian system, for compatibility with previous Markovian techniques. The second is a fully semi-Markovian example. In both cases we present an analytically calculated result by which to compare the approximations. Where we have needed to use a Laplace inversion algorithm, we have used the Euler method from Abate and Whitt [1].

4.1 Markovian Example

We set out a Markov system with DTMC, P and rate matrix, S. We aim to find the passage times $L_{06}(t)$ and $L_{02}(t)$.

$$P = \begin{pmatrix} 0 & p_1 & 0 & 1-p_1 & 0 & 0 & 0 \\ 1-p_2 & 0 & 0 & 0 & p_2 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1-p_3 & 0 & 0 & p_3 \\ p_4 & 0 & 1-p_4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \tag{16}$$

$$S = \begin{pmatrix} 0 & s_{01} & 0 & s_{03} & 0 & 0 & 0 \\ s_{10} & 0 & 0 & 0 & s_{14} & 0 & 0 \\ s_{20} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & s_{36} \\ 0 & 0 & 0 & s_{43} & 0 & 0 & s_{46} \\ s_{50} & 0 & s_{52} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & s_{65} & 0 \end{pmatrix} \tag{17}$$

Using values $p_1 = 2/3, p_2 = 1/2, p_3 = 1/4, p_4 = 1/3, s_{01} = 1/4, s_{03} = 1, s_{10} = 1/3, s_{14} = 2/3, s_{20} = 1, s_{36} = 1/2, s_{43} = 1, s_{46} = 4/5, s_{50} = 1/5, s_{52} = 2/5, s_{65} = 1$. E is a matrix of exponential rates, so the distribution matrix $F_{ij}(t) = s_{ij}\exp(-s_{ij}t)$

Figures 1 and 2 show the passage-time distributions $L_{06}(t)$, $L_{02}(t)$ as calculated using equation (9). It takes longer for the algorithm to converge in the second case as more of the system is exposed to the passage-time.

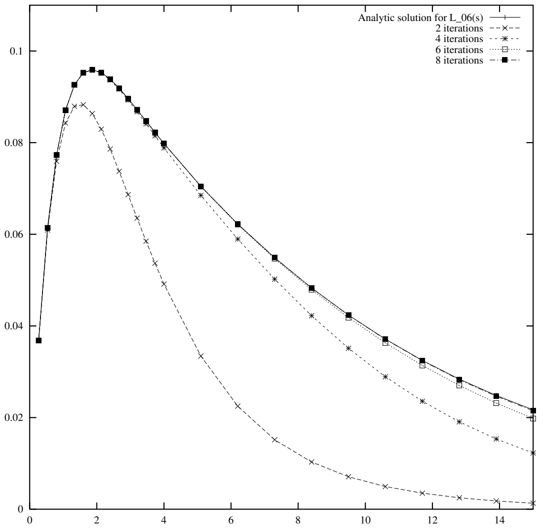

Fig. 1. Four iterative approximations of $L_{06}(t)$

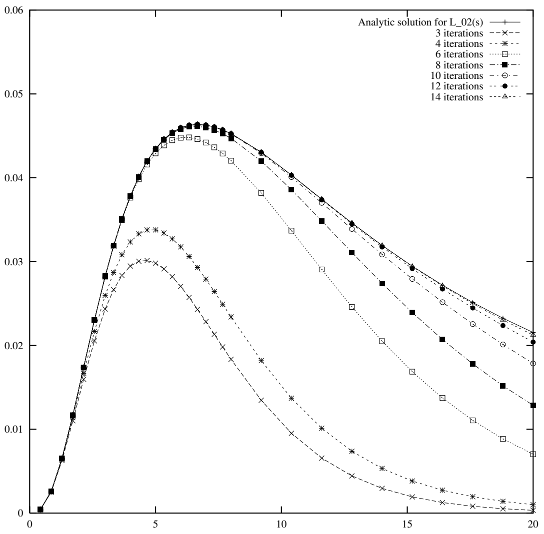

Fig. 2. Seven iterative approximations of $L_{02}(t)$

4.2 Semi-Markov Example

Finally, we look at a truly semi-Markov example of a system which employs different types of distribution (figure 3).

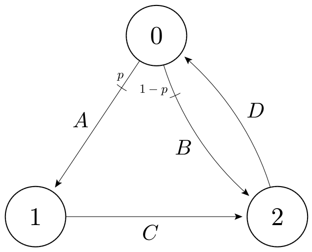

Fig. 3. Lots of distributions.

The distributions used in the transition system of figure 3 are given by:

$$\begin{aligned} A &\sim \Gamma(\alpha, \lambda) && (18)\\ B &\sim \text{Hyper}\left(\frac{2}{3}, \frac{1}{3}; \lambda, \mu\right) && (19)\\ C &\sim \mathrm{U}\left[\frac{1}{\lambda}, \frac{1}{\mu}\right] && (20)\\ D &\sim \delta\left[\frac{1}{\mu}\right] && (21) \end{aligned}$$

where:

$$\begin{aligned} \hat{F}_{01}(s) &= \left(\frac{\lambda}{\lambda + s}\right)^{\alpha} && (22)\\ \hat{F}_{02}(s) &= \frac{2}{3}\frac{\lambda}{\lambda + s} + \frac{1}{3}\frac{\mu}{\mu + s} && (23)\\ \hat{F}_{12}(s) &= \frac{\lambda\mu}{(\lambda - \mu)s}\left[e^{-s/\lambda} - e^{-s/\mu}\right] && (24)\\ \hat{F}_{20}(s) &= e^{-s/\mu} && (25) \end{aligned}$$

The following parameter values are used: $\alpha = 5/2$, $\lambda = 3/2$, $\mu = 1.23$ and $p = 3/10$. Figure 4 is the passage-time distribution between states 2 and 1 as calculated using equation (9). Figure 5 shows the converging iterations of the approach. Note that even though the distribution has a discontinuous derivative ($t = 1/\mu$), the Laplace inversion algorithm (Euler Method [1]) has no numerical difficulty inverting the function.

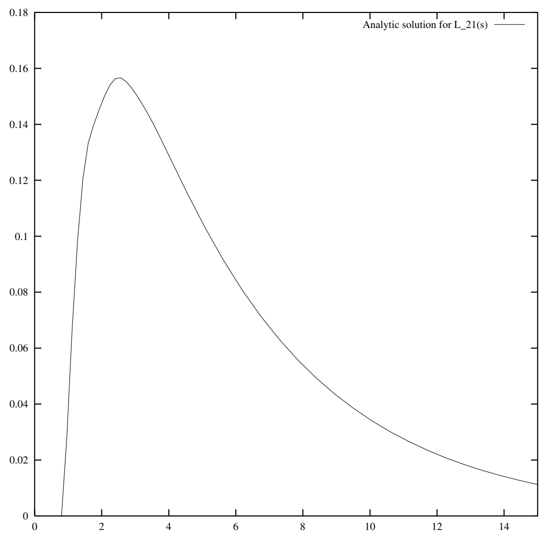

Fig. 4. Analytic solution of $\hat{L}_{21}(s)$

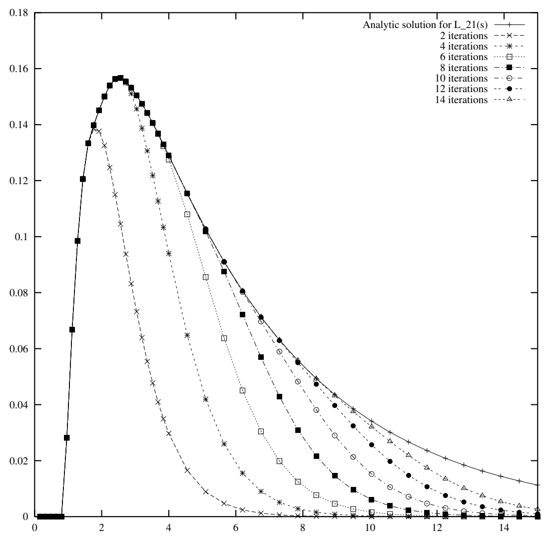

Fig. 5. Seven iterative approximations of $\hat{L}_{21}(s)$

5 Conclusion

We have presented definitions of passage-time distributions for SMPs between clusters of states. We have presented and $O(n^2)$ algorithm for creating successive approximations to the passage-time.

We have suggested a finite sum approach to creating a transient distribution which relies on the passage-time results—if this works, it will also be an $O(n^2)$ technique. In both instances the original analytic methods are $O(n^3)$.

Our next improvement will be to use a finite-space representation of a general distribution. At the moment the complexity savings are lost in the increased space representation that occurs when the matrix operations act on the analytic representation.

References

[1] ABATE, J., AND WHITT, W. The Fourier-series method for inverting transforms of probability distributions. *Queueing Systems 10*, 1 (1992), 5–88.

[2] COX, D. R., AND MILLER, H. D. *The Theory of Stochastic Processes*. Methuen, 1965.

[3] PYKE, R. Markov renewal processes with finitely many states. *Annals of Mathematical Statistics 32*, 4 (December 1961), 1243–1259.

A Establishing Compatibility of Convergence Definitions

Statement of problem

We have proposed two definitions for the convergence of the transient probabilities to the steady-state probabilities:

i. Well-behaved case

$$\lim_{t \longrightarrow \infty} T_{ij}(t) = \pi_j \tag{26}$$

ii. General case

$$\lim_{t \longrightarrow \infty} \frac{1}{t} \int_0^t T_{ij}(\tau)\, d\tau = \pi_j \tag{27}$$

Proof that (ii) includes (i)

Suppose the transient probability converges in the sense of (i):

$$\lim_{t \longrightarrow \infty} T_{ij}(t) = \pi_j \tag{28}$$

Then given any $\varepsilon > 0$, there exists a time T such that $|T_{ij}(t) - \pi_j| < \varepsilon$ for all $t \geq T$. We split the integral from (ii) into two ranges:

$$\frac{1}{t} \int_0^t T_{ij}(\tau)\, d\tau = \underbrace{\frac{1}{t} \int_0^T T_{ij}(\tau)\, d\tau}_{\longrightarrow 0 \text{ as } t \longrightarrow \infty} + \frac{1}{t} \int_T^t T_{ij}(\tau)\, d\tau \tag{29}$$

$$\lim_{t \longrightarrow \infty} \frac{1}{t} \int_0^t T_{ij}(\tau)\, d\tau = \lim_{t \longrightarrow \infty} \frac{1}{t} \int_T^t T_{ij}(\tau)\, d\tau \tag{30}$$

But when $t \geq T$, we know that $\pi_j - \varepsilon < T_{ij}(t) < \pi_j + \varepsilon$, so:

$$\frac{1}{t} \int_T^t (\pi_j - \varepsilon)\, d\tau < \frac{1}{t} \int_T^t T_{ij}(\tau)\, d\tau < \frac{1}{t} \int_T^t (\pi_j + \varepsilon)\, d\tau \tag{31}$$

$$(\pi_j - \varepsilon)\left(1 - \frac{T}{t}\right) < \frac{1}{t} \int_T^t T_{ij}(\tau)\, d\tau < (\pi_j + \varepsilon)\left(1 - \frac{T}{t}\right) \tag{32}$$

$$\pi_j - \varepsilon < \lim_{t \longrightarrow \infty} \frac{1}{t} \int_T^t T_{ij}(\tau)\, d\tau < \pi_j + \varepsilon \tag{33}$$

Finally, from equation (30) we have that for any $\varepsilon > 0$:

$$\pi_j - \varepsilon < \lim_{t \longrightarrow \infty} \frac{1}{t} \int_0^t T_{ij}(\tau)\, d\tau < \pi_j + \varepsilon \tag{34}$$

and it follows that $\lim_{t \longrightarrow \infty} \frac{1}{t} \int_0^t T_{ij}(\tau)\, d\tau = \pi_j$. ■