Probabilistic Modelling Requires a Certain Imagination

Marcel F Neuts - Tucson, Arizona, USA

1. Introduction

A year ago, a doctoral student in communications engineering came to see me and asked me for the solution to the following problem in the excellent book by Kleinrock (1975, Vol 1, p.80) on the theory and applications of queues:

"Consider a Markov chain with states E_0, E_1, E_2, ... and with transition probabilities

$$p_{ij} = e^{-\lambda} \sum_{n=0}^{j} {}^{i}C_n \, p^n \, q^{i-n} \, \frac{\lambda^{j-n}}{(j-n)!}; \quad p + q = 1 \quad (0 < p < 1).$$

(a) Is this chain irreducible? Periodic? Explain.

(b) We wish to find π_i = equilibrium probability of E_i. Write π_i in terms of p_{ij} and π_j for $j = 0,1,2,...$

(c) From (b) find an expression relating $P(z)$ to $P[1 + p(z-1)]$, where

$$P(z) = \sum_{i=0}^{\infty} \pi_i z^i.$$

(d) Recursively (i.e. repeatedly) apply the result in (c) to itself and show that the n^{th} recursion gives

$$P(z) = e^{\lambda(z-1)(1+p+p^2+...+p^{n-1})} P[1 + p^n(z-1)].$$

(e) From (d) find $P(z)$ and then recognize π_i."

The student showed me a sheaf of papers covered with lengthy calculations in which he had started the solution by calculating the generating functions

$$P(i;z) = \sum_{j=0}^{\infty} p_{ij} z^j ,$$

for $i \geq 0$. An error in a summation index apparently accounted for the fact that he could not evaluate the sum in a closed form.

As is often my lot, I was pressed for time and told the student that I preferably avoided calculations. We would therefore guess at a practical application which could have led to consideration of that Markov chain to gain insight into the needed mathematical analysis. By entering the factor $\exp(-\lambda)$ into the summation, we saw that for each i, the sequence $\{p_{ij}\}$ is the *convolution* of a binomial probability density with parameters i and p and a Poisson density with parameter λ. That observation immediately yields that

$$P(i;z) = (pz + q)^i \exp[-\lambda(1 - z)],$$

so that the steady-state equations

$$\pi_j = \sum_{r=0}^{\infty} \pi_r p_{rj}$$

lead to

$$(1) \qquad P(z) = \sum_{j=0}^{\infty} z^j \sum_{r=0}^{\infty} \pi_r p_{rj} = P[pz + 1 - p] \exp[-\lambda(1 - z)].$$

With these results in place, the remaining parts of the problem lead the student through the essential steps of the mathematical argument. One ultimately recognises the π_i as the terms of a Poisson density with parameter $(1 - p)^{-1}\lambda$.

After doing that problem, I felt that it would be preferable to introduce those mathematical questions as part of a modelling problem, rather than as an exercise in formal manipulations. In my own teaching, I frequently use that approach with success, particularly in postgraduate instruction. In reading research papers, even on formal and abstract results, I am always grateful when the author feels sufficiently free to share some of the intuitive process of discovery, prior to or after rigorous, technical proofs. Such a sharing of imagination has not been fashionable in mathematical presentations during much of this century, but particularly since the 1950s. I believe that this accounts in part for the decline in the general ability of mathematics students to handle complexity and to solve problems that draw on methodologies from several specialities.

My participation in ICOTS 3 provided the final incentive to write an article on the development of imagination, so essential to successful mathematical modelling. I have selected the stated problem which may also be found in Takács (1960, p.21) with a brief reference to statistical mechanics, for several reasons; its source is an important text used by many engineering students; the problem lends itself well to illustrate the ideas on imagination presented in this paper and, with imagination unfettered, it can serve to generate quite exciting *purely mathematical* questions.

Not being a psychologist, I find it too difficult to discuss imagination in general. Moreover, the creative processes of the human mind are so much more easily (and pleasurably) demonstrated than analysed, that I shall rather proceed by example; successively donning the hats of a communications engineer, an applied mathematician, and an algorithmic probabilist. However, the hats of these various guilds are only metaphors for different modes of mathematical imagination and not for separate professional communities. One of the important features of a fertile mathematical imagination is the ability to recognise structural and methodological similarities across subjects and disciplines. I firmly believe that mathematical education needs a period of reintegration in which the student learns to think of mathematics as a living language, the tool *par excellence* for systematic and precise discourse on problems of scientific or technological interest. Presentations of calculus as a bag of isolated tricks or of probability as an obsessive study of balls in urns are therefore mostly educational disasters. Even the "successful" student is deprived of the delights of imaginative discovery. Mathematics, which should be anything but that, is often perceived as a sterile and formal activity.

2. Some communications models

We imagine a number of situations that are quite representative of modelling problems encountered in communications engineering. I shall give these situations some suggestive names. These should not limit the scope of application or of our imagination. I would urge the reader to search for other tales by which to introduce mathematical problems of a similar nature. It is likely that different perspectives shall lead to other questions than those discussed here.

My e-mail buffer: Every day, I receive a number of messages by electronic mail. These are stored in a list of files for review at the end of the day. A message may be kept in the file for one more day, or deleted. Assume that the numbers of new messages received each day are independent, identically distributed random variables with (common) probability density $\{a_j\}$ on the nonnegative integers, and that at each review, the retention or removal of messages in the file corresponds to independent Bernoulli trials. A message is retained for one more day with probability p or is removed with probability 1 - p. I wish to study the stochastic process describing the number of messages in my e-mail file at the beginning of each day. In particular, I shall need the "stationary probability density" of the length of my list to assess, for various values of K, the fraction of time that there will be more than K stored messages.

A discrete telephone trunk model: Consider a time-slotted communications link and suppose that the numbers of new calls starting at the beginning of every time slot are independent, identically distributed random variables with probability density $\{a_j\}$, $j \geq 0$. A call remains active for a number of time slots which has a geometric probability density, that is, the probability that a call starting at time t ends at time $t + k$ is $p^{k-1}(1 - p)$ for $k \geq 1$. We are interested in the number of calls active at the beginning of each time slot, in particular the stationary probability density of that number.

A packet stream: We again consider a time-slotted communications link and let the numbers of new calls starting at the beginning of every time slot be independent, identically distributed random variables with probability density $\{a_j\}$. As in the preceding situation, calls have independent, geometrically distributed durations and, for

as many time slots as a call is active, it submits one packet per time slot to the system.

We are interested in *simulating* the stream of packets submitted during successive slots, but wish to avoid "initialisation effects" in our simulation runs. We would like to choose the initial number of active calls according to an appropriate probability law, so as to mimic a packet stream *which has been flowing for a very long time*. That is, of course, accomplished by choosing the initial number of active calls according to the stationary density of an appropriate Markov chain.

All three "applied" questions lead to the same problem in Markov chains, which is a generalisation of the problem selected from Kleinrock (1975). As we shall see, concrete models provide us with mental constructs leading naturally to new mathematical questions. Questions uninteresting for one model often become important in other settings. For example, for the buffer model we may want to consider the time required to reach an *empty buffer*, starting from a given positive initial buffer content. For the telephone trunk model, we may enquire into the probability distribution of the *maximum number of active calls* during a given time period. Such questions, which do not lend themselves well to analytic solution, are nevertheless computationally tractable over a useful range of parameters. They are therefore not only stimulating to the imagination, but lead to the exploration of other methods of solution. For a different probability model, some examples are discussed in Neuts (1980). Before embarking upon discussions of technical matters, we address some modelling issues.

Discrete time: It is often natural and mathematically convenient to model certain processes in discrete time. In doing so, it is important to spell out carefully how the accounting will be done during a unit time slot. In our case, we first count the items remaining after the Bernoulli trials and next add new additions to the count. We can write this formally as

$$X_{n+1} = B[X_n,p] + v_n \,, \tag{2}$$

which says that the count X_{n+1} at time n+1 consists of the successes in X_n Bernoulli trials with probability p of success, augmented by the number v_n of new arrivals during the n-th time slot. The independence assumption on the numbers of new arrivals implies that the sequence $\{X_n\}$ is a Markov chain.

General distributions: I have not required the probability density $\{a_r\}$ to be Poisson for the following reasons. (a) The Poisson assumption leads to an explicit solution, but at the cost of a severe restriction on the model and therefore on its practical applicability. As we shall see, the same mathematical analysis yields the solution to the general version. For many practical problems, explicit analytic solutions are rare, while feasibly computable solutions are much more common. I believe that textbooks place too high a premium on analytic tractability. (b) With a general probability density $\{a_r\}$, the answer to the first question in Kleinrock's problem is no longer obvious. Some care is now needed to see that irreducibility of the Markov chain requires that $0 < a_0 < 1$. When that does not hold, there exists some $N > 0$, such that each day there arrive at least N new messages. States below N will then not be attainable. With that mild restriction on a_0, aperiodicity of the chain is assured by the Bernoulli trials.

Geometric holding times: Particularly in the second and third situations, the assumption of geometric holding times is a severe restriction. It should be made clear to the student that, without that restriction, analysis of the problem becomes much more

difficult. It we insisted that calls have a general holding time distribution, the analysis by Markov processes would require keeping track of the remaining holding times of all active calls. Insight into this point will give the student an understanding of the evolution of the literature, say, on the theory of queues. Many models have analytically tractable solutions under the assumption of exponential distributions and a significant portion of the literature is devoted to relaxing that very stringent assumption. The model which we are discussing here, has a continuous-time analogue in the GI/M/∞ queue with group arrivals (see Takács, 1962). Efforts to remove the assumption of exponential holding times in that model have not yet been very successful. Discrete models in which the holding time of calls or the generation of packets during a call is described by Markovian mechanisms, are the subject of recent research relevant to packetised communication flows. See, for example, Ramaswami and Latouche (1988).

Mixed geometric holding times: After perusal of this paper, the reader may wish to consider the following first attempt to remove the assumption of geometric holding times. There are two categories of messages, with geometric holding times with retention probabilities p_1 and p_2 respectively. The arrival of new messages occurs as before, but a new message is now of the first type with probability θ or of the second with probability $1 - \theta$ according to Bernoulli trials. The holding time distribution is now a mixture of two geometric densities. The numbers of customers of both types now form a Markov chain on the set of pairs (i_1,i_2) of nonnegative integers. We suggest exploring how far the analysis in Section 2 can be extended to this model.

3. Analytic issues

The analysis sketched in Section 1 essentially carries over to the general case, but some minor technical issues not present in the Poisson case need to be addressed. In the interest of brevity, only the essential points of proofs are discussed. We denote by $A(z)$, the probability generating function of the sequence $\{a_r\}$. The generating functions $P(i;z)$ are given by $P(i;z) = (pz + q)^i A(z)$, and the steady-state equations lead to

$$P(z) = \sum_{j=0}^{\infty} z^j \sum_{r=0}^{\infty} \pi_r p_{rj} = P[pz + 1 - p]\, A(z). \tag{3}$$

Successively replacing z by $p^n z + 1 - p^n$ in equation (3) and forming the products of the left and right hand sides of the resulting equations, we obtain the formal solution

$$P(z) = \prod_{n=0}^{\infty} A[p^n z + 1 - p^n], \tag{4}$$

whenever the infinite product converges. Since $|P(z)| \le P(|z|)$, it suffices to establish convergence for z in $[0,1]$ and, since each factor is then positive and increases in z to 1 on that interval, convergence at $z = 0$ guarantees convergence for all z with $|z| \le 1$. The necessary and sufficient condition for positive recurrence of the Markov chain is therefore the convergence of the infinite product

(5) $$\prod_{n=0}^{\infty} A[1 - p^n] = \pi_0.$$

By the classical convergence criterion that convergence is assured if and only if the series

(6) $$\sum_{n=0}^{\infty} [1 - A(1 - p^n)],$$

converges. Henceforth we concentrate attention on the cases where condition (6) holds.

If the mean μ'_1 of the probability density $\{a_r\}$ is finite, it follows from (5) that the mean $\psi'_1 = P'(1-)$ of the steady-state density $\{\pi_i\}$ is given by $\psi'_1 = (1 - p)^{-1}\mu'_1$. If we are *only* concerned with the stationary mean length of our list of messages, that quantity is easily computed and depends only on the *mean* of the density $\{a_r\}$. That is an example of an *insensitivity theorem*, of which a number have been established in queueing theory. Such a result is both appealing (because of its mathematical simplicity) and potentially deceptive, as it says that the mean ψ'_1 is not affected by the *variability* of the numbers of incoming messages. ψ'_1 is therefore, in general, not an informative descriptor of the Markov chain.

The terms of the series in (6) are increasing positive functions of p and the series trivially converges at p = 0, and diverges at p = 1. This suggests that there exists a constant p*, which depends on the generating function A(z), such that the Markov chain is positive recurrent if and only if $p < p^*$. For the Poisson case, it is clear that $p^* = 1$, and that expression for ψ'_1 suggests that this may also be the case for all densities $\{a_r\}$, *with finite mean.* We now prove that this is so.

Theorem 1: Provided the mean μ'_1 *exists,* $p^* = 1$, *or equivalently, the equilibrium condition (6) holds for all* p *with* $0 < p < 1$.

Proof: We recall the formula

$$1 - A(z) = (1 - z) \sum_{n=0}^{\infty} z^n \sum_{k=n+1}^{\infty} a_k ,$$

and the fact that if μ'_1 exists, is also given by the sum

$$\sum_{n=0}^{\infty} \sum_{k=n+1}^{\infty} a_k .$$

Setting $z = 1 - p^n$, we obtain

$$1 - A(1 - p^n) = p^n \sum_{r=0}^{\infty} (1 - p^n)^r \sum_{k=r+1}^{\infty} a_k < p^n \mu'_1 ,$$

and convergence in (6) follows by comparison. ❐

Theorem 1 adequately settles the existence of steady-state probabilities for most applications. It raises, however, a few questions laden with purely mathematical delight,

such as: How can one characterise p* when $\{a_r\}$ has an infinite mean? Is there an example of a sequence $\{a_r\}$ with infinite mean for which p* = 1? Such questions are problems in analysis, yet they are not without probabilistic interest. For $\{a_r\}$ to have an infinite mean, the numbers of new arrivals per time slot must exhibit strong variability. We are therefore asking whether a Bernoulli deletion process can "cope" with such erratic inputs in keeping the buffer size from increasing beyond all bounds.

The special case of Poisson input gives an instance of a *closure theorem.* If $\{a_r\}$ is a Poisson density, so is the density $\{\pi_i\}$. As in other areas of mathematics, we can ask just how "special" that result is. This leads to questions such as: If $\{\pi_i\}$ is a Poisson density, what can we say about $\{a_r\}$? Are there other special classes C of probability densities on the nonnegative integers, such that if $\{a_r\} \in C$ so does $\{\pi_i\}$? There is also the *inverse problem*: For which (positive) densities $\{\pi_i\}$ is it possible to find a probability density $\{a_r\}$ such that their probability generating functions satisfy the equation (4) for some constant p with $0 < p < 1$? The answer to the first question is easy; it is given in Theorem 2. For the other questions, we can only offer suggestions and refer to literature that may be relevant to their investigation.

Theorem 2: If $\{\pi_i\}$ *is a Poisson density, so is the density* $\{a_r\}$.
Proof: This follows immediately from equation (3). If $P(z) = \exp[-a(1 - z)]$, then clearly $A(z) = \exp[-a(1 - p)(1 - z)]$. ❐

The second question has not yet been answered to any degree of generality, but the following result shows that there are sets C, other than the Poisson densities that are *closed*, in the sense that they contain both sequences $\{a_r\}$ and $\{\pi_i\}$.

Theorem 3: If the density $\{a_r\}$ *has the generating function*

$$A(z) = \exp\{-b[1 - (\phi z + \phi' z^2)]\},$$

where $b > 0$, *and* $0 \le \phi = 1 - \phi' \le 1$, *the probability generating function* $P(z)$ *of* $\{\pi_i\}$ *is given by*

$$P(z) = \exp\{-a[1 - (\theta z + \theta' z^2)]\},$$

where

$$a = b(1 - p^2)^{-1}(1 + 2p - p\phi), \text{ and } \theta = 1 - \theta' = (2p - p\phi + \phi)(2p - p\phi + 1)^{-1}a.$$

Proof: This is shown by substituting the proposed forms of A(z) and P(z) into equation (4) and equating coefficients in the quadratic expressions which are the exponents in both sides of the resulting equation. ❐

The densities corresponding to A(z) and P(z) are each the convolution of an ordinary Poisson density with a Poisson density on the *even* nonnegative integers. Theorem 3 suggests considering densities with probability generating functions A(z) of the form $\exp\{-a[1 - B(z)]\}$, where B(z) is a probability generating function on the positive integers. We leave that potential generalisation to the initiative of the reader.

The third question is related to the decomposability (or divisibility) properties of discrete distributions with respect to the convolution product. There is a developing literature on this subject. We refer to Steutel and van Harn (1979) and van Harn, Steutel and Vervaat (1982) for examples and further references. The analysis draws on the theory of functional equations and is quite arduous. Equation (2) is an example of a (simple) stochastic difference equation. This suggests yet another mathematical perspective from which the models discussed here may be considered.

An entirely different class of generalisations is obtained by enlarging the class of probability models so as to incorporate fluctuations in the input process. By analogy with investigations in the theory of queues, discussed in detail in Neuts (1981, 1989), we may assume that the input process is *modulated* by an irreducible finite Markov chain with m states and the transition probability matrix C. We shall not dwell on that generalisation, except to note that discussion of the convergence of the infinite product (5), which are elementary in the scalar case, draws on mathematical properties which have received much less attention.

We conclude this section by drawing attention to the third motivating example presented in Section 2. The numbers of packets generated at successive time slots could be modelled here as a Markov chain and we discussed the issue of determining the initial conditions which yield the stationary version of the packet stream. It is clear that, for genuine and important problems of communications engineering, there are much more weighty questions to be answered. These deal with issues of statistical analysis and model fitting, with spectral analysis to assess the fluctuations in the packet stream and with the response of queues subject to such arrival processes. Generalisations of the present model lead to the study of stationary integer-valued stochastic processes with a discrete parameter. These subjects are matters of current investigations, which are far from complete. It is remarkable that, starting from a textbook problem, there is a short thread of mathematical imagination leading to issues of current research interest.

4. Algorithmic issues

As I have repeatedly discussed, the consideration of algorithmic aspects of mathematical models is of the greatest educational interest. While traditional methods of analysis emphasise formal manipulations and asymptotic behaviour, the examination of algorithmic procedures invariably requires the student carefully to consider the *structural* features of the problem. Correct algorithmic analysis draws on talents of abstraction that generally are still inadequately developed by standard undergraduate courses. We note with dismay, that few if any textbooks on probability or stochastic models discuss algorithmic issues or list problems of sufficient complexity to offer algorithmic challenge. We shall limit our discussion to the elementary algorithmic aspects of the model in Section 2. The reader can find additional examples or further discussions of education aspects of algorithmic analysis in Carson and Neuts (1975) and Neuts (1973, 1974, 1980, 1981, 1984a,b and 1986).

To evaluate the probability density $\{\pi_i\}$ of equation (4), for a given density $\{a_r\}$, we need two basic modules. The first is an algorithm to evaluate the *convolution* of two discrete (lattice) densities; the second an algorithm to evaluate the density with generating function $B[\alpha z + 1 - \alpha]$ for a given density $\{b_r\}$ with generating function $B(z)$.