ENTROPY MAXIMISATION

AND

QUEUES WITH OR WITHOUT BALKING

An investigation into the impact of generalised maximum entropy solutions on the study of queues with or without arrival balking and their applications to congestion management in communication networks

Neelkamal Paresh SHAH

Submitted for the Degree of

Doctor of Philosophy

School of Electrical Engineering and Computer Science

Faculty of Engineering and Informatics

University of Bradford

2014

Abstract

Neelkamal Paresh Shah

Entropy Maximisation and Queues With or Without Balking

An investigation into the impact of generalised maximum entropy solutions on the study of queues with or without arrival balking and their applications to congestion management in communication networks

Keywords: Queues, Balking, Maximum Entropy (ME) Principle, Global Balance (GB), Queue Length Distribution (QLD), Generalised Geometric (GGeo), Generalised Exponential (GE), Generalised Discrete Half Normal (GdHN), Congestion Management, Packet Dropping Policy (PDP)

Generalisations to links between discrete least biased (i.e. maximum entropy (ME)) distribution inferences and Markov chains are conjectured towards the performance modelling, analysis and prediction of general, single server queues with or without arrival balking. New ME solutions, namely the generalised discrete Half Normal (GdHN) and truncated GdHN ($GdHN_T$) distributions are characterised, subject to appropriate mean value constraints, for inferences of stationary discrete state probability distributions. Moreover, a closed form global balance (GB) solution is derived for the queue length distribution (QLD) of the M/GE/1/K queue subject to extended Morse balking, characterised by a Poisson prospective arrival process, i.i.d. generalised exponential (GE) service times and finite capacity, K. In this context, based on comprehensive numerical experimentation, the latter GB solution is conjectured to be a special case of the $GdHN_T$ ME distribution.

Owing to the appropriate operational properties of the M/GE/1/K queue subject to extended Morse balking, this queueing system is applied as an ME performance model of Internet Protocol (IP)-based communication network nodes featuring static or dynamic packet dropping congestion management schemes. A performance evaluation study in terms of the model's delay is carried out. Subsequently, the QLD's of the GE/GE/1/K censored queue subject to extended Morse balking under three different composite batch balking and batch blocking policies are solved via the technique of GB. Following comprehensive numerical experimentation, the latter QLD's are also conjectured to be special cases of the $GdHN_T$. Limitations of this work and open problems which have arisen are included after the conclusions.

Acknowledgements

First and foremost I salute my supervisor, Professor Demetres Kouvatsos, whose genuine encouragement, guidance and support throughout the duration of my PhD has been invaluable. He has been an inspiration through his care for and assistance to students, passion for high quality work and tireless efforts in disseminating knowledge to and fostering collaboration between the student and professional communities.

I extend my thanks to my supervisor Dr Rod Fretwell for his time and valuable contributions to our discussions.

The selfless love, commitment and support of my parents throughout this journey have been incredible.

I would also like to acknowledge my brothers and sisters in Christ, colleagues, friends, neighbours and others I know who have cheered me on, encouraged me and/or prayed for me during the various phases of my PhD.

Thanks go to Mr Esmaeil Habibzadeh, PhD student of the Department of Computing and Mathematics, for his great help in creating and running simulation programs to validate the analysis of this research work.

And special recognition is due to the School of Electrical Engineering and Computer Science (formerly the School of Computing, Informatics and Media) and the University of Bradford for graciously facilitating and supporting me through this long and arduous journey and for this I am very grateful.

By the grace of the Lord Jesus Christ

Relevant Publications

1. Shah, N., Kouvatsos, D.: Entropy Maximisation, Queueing and Congestion Management Applications, *in preparation*.
2. Shah, N., Kouvatsos, D.: The GE/GE/1/N Queue Subject to State-Dependent Arrival Balking. In the Seventh International Working Conference on the Performance and Security Modelling and Evaluation of Cooperative, Heterogeneous Networks (HETNET's), Ilkley, UK (2013).
3. Shah, N., Kouvatsos, D.: A Queue Conjectured to Bear the Generalised Discrete Half Normal Maximum Entropy QLD. in 27th Annual UK Performance Engineering Workshop, University of Bradford, UK (2011)
4. Shah, N., Kouvatsos, D.D., Fretwell, R.J.: An Analytic Generalisation of a Maximum Entropy Customer Impatience Queueing Solution and its Nonbalking G/M/1/N Equivalence. In The 11th Annual Postgraduate Symposium on the Convergence of Telecommunications, Networking and Broadcasting. Liverpool, UK: The School of Computing and Mathematical Sciences, Liverpool John Moore's University.(2010)

Contents

Acronyms

CDF	Cumulative distribution function
dHN	Discrete Half Normal
dHN_T	Truncated discrete Half Normal
DT	Droptail
EMC	Embedded Markov chain
ERD	Early random drop
G	General
GB	Global balance
GdHN	Generalised discrete Half Normal
$GdHN_T$	Truncated generalised discrete Half Normal
GE	Generalised exponential
GGeo	Generalised geometric
$GGeo_T$	Truncated generalised geometric
GRED-I	Gentle random early detection - Instantaneous
i.i.d.	Independent and identically distributed
IP	Internet protocol
ME	Maximum entropy
MQL	Mean of queue length
PDP	Packet dropping policy
PGF	Probability generating function
pmf	Probability mass function
QLD	Queue Length Distribution
QNM	Queueing network model

RED	Random early detection
RV	Random variable
SCOV	Squared coefficient of variation
VQL	Variance of queue length

1. Introduction

Queueing models with or without arrival balking constitute fundamental investigative tools for the performance modelling, analysis and prediction of discrete flow systems such as business and production systems and transport and communication networks. However, derivation of exact performance distributions for complex queueing systems, such as infinite or finite capacity G/G/1 queues with or without balking, is often an infeasible task. To this end, in this thesis, generalised least biased (i.e. maximum entropy (ME)) queue length distribution (QLD) inferences are devised based on knowledge of limited prior queue length moment information. These least biased QLD inferences are subsequently conjectured to characterise new queueing models with balking, which are applied as ME performance models of network congestion management.

Balking, a customer impatience phenomenon, is the immediate rejection of service by customers arriving to a queue and anticipating an unacceptably long queueing time and/or their service is not required urgently (Haight 1957). Customers which decide to balk do not join the queue and in this thesis are considered to not return afterwards. As such, customers which balk are deemed to be lost from the point of view of the queue. Queueing models with balking have been successfully applied to evaluate multiple real life systems such as:

- Telephone call centres and business service systems where customers gain information about anticipated delay and traffic-flow

control in transport networks (Mendelson et al. 1999; Whitt 1999; Whitt 2005; Guo and Zipkin 2007; Liu 2007; Jouini et al. 2011) and

- Customer acceptance/rejection policies dependent on workload thresholds as administered for example in admission control mechanisms in communication networks (Baccelli et al. 1984; Liu 2007; Boxma and Prabhu 2009).

The ME principle, attributable to Jaynes (Jaynes 1957; Jaynes 1978), is a method of statistical inference by which a least biased probability distribution of a random quantity of any general system can be characterised based on limited prior information. In this context of ME distribution inferences, 'bias' can be defined as commitment to unknown information (Jaynes 1957). The least biased distribution inference is obtained by maximising the uncertainty with respect to what is not known about the random quantity, while satisfying the known partial prior information. The measure of uncertainty maximised in the principle of ME is that proposed by Shannon, who also referred to it as the entropy of the distribution (Shannon 1948).

The ME principle has been employed in the modelling of systems in a vast range of areas from economics and finance to operational research, queueing, statistical mechanics and thermodynamics (Shore and Johnson 1980; Fang et al. 1997). Within the queueing context, ME solutions have been derived for various performance distributions of, among others, single server queues at equilibrium (Shore 1982; El-Affendi and Kouvatsos 1983; Kouvatsos 1986a; Kouvatsos 1986b; Kouvatsos 1988), multiserver queues (Kouvatsos and Almond 1988; Wu and Chan 1989; Arizono et al. 1991),

multiple class queues under priority scheduling disciplines (Kouvatsos and Tabet-Aouel 1989; Kouvatsos and Tabet-Aouel 1994), queues with vacation (Skianis 1997; Ke and Lin 2008), queueing network models (QNM's) (Walstra 1985; Cantor et al. 1986; Kouvatsos and Almond 1988; Kouvatsos 1994; Kouvatsos and Awan 2003) and queues subject to balking (Kemp 2008; Shah et al. 2010).

The technical contributions of this thesis are founded largely on the following works (Morse 1958; El-Affendi and Kouvatsos 1983; Kouvatsos 1986a; Kouvatsos 1986b; Kouvatsos 1988; Kemp 2005; Kemp 2008). They analyse single server infinite and finite - capacity queues at equilibrium, with or without balking under the first come first served (FCFS) scheduling discipline. The approaches used comprise Markov chain models, the global balance (GB) Markov chain steady state probability solution technique and/or the ME principle.

To model customer impatience at queues, Morse (Morse 1958) studied a stable M/M/1 queue with Poisson prospective arrivals subject to balking characterised by an exponential function and independent and identically distributed (i.i.d.) exponential service times. This was solved via the Markov chain model of its equivalent stable M(n)/M/1 queue with state-dependent Poisson arrival rates which accounted for the balking. Kemp, on the other hand, employing the ME principle, subject to the prior information constraints of first moment, variance and the normalisation condition derived the discrete Half Normal (dHN) ME stationary state probability distribution. By means of a Markov chain model, it was shown that the GB solution derived earlier by

Morse was a special case of the dHN (Morse 1958; Kemp 2005; Kemp 2008).

The generalised geometric (GGeo) ME solution for the QLD of a stable, FCFS, ordinary (i.e. without balking) M/G/1 queue, characterised by Poisson arrivals and i.i.d. general (G) service times, was devised in (El-Affendi and Kouvatsos 1983). The GGeo explicitly incorporated the queue stability (or conservation of flow) condition in addition to the queue length mean (MQL) and normalisation prior information constraints. The service-time distribution of the M/G/1 queue possessing the GGeo ME QLD was discovered to be satisfied exactly by the bursty (i.e. variable) generalised exponential (GE) service-time distribution (El-Affendi and Kouvatsos 1983). Following this, Kouvatsos (Kouvatsos 1986a; Kouvatsos 1986b; Kouvatsos 1988) found that the QLD's of both the infinite and finite capacity GE/GE/1 queues are ME distributions each derived from prior knowledge of the MQL and appropriate queue stability and normalisation conditions. The GE/GE/1 queue is characterised by a bursty compound Poisson arrival process with geometrically distributed batch sizes (and thus underlying i.i.d. GE inter-arrival times) and i.i.d. GE service times.

In this work, motivated by the performance modelling of complex queueing systems, the ME principle is employed in the discrete domain to derive generalised least biased distribution inferences. These generalisations arise from the inclusion of additional prior information to that assumed known in their corresponding subclass distributions. The generalised discrete half

normal[1] (GdHN) and truncated GdHN ($GdHN_T$) stationary ME distributions are devised, based on knowledge of the first moment, variance, boundary state probabilities of the infinite or finite support cases, p_0^{∞} (henceforth symbolised simply by p_0) or p_0^K and p_K^K respectively, and the normalising condition.

Subsequently, a closed-form QLD of the FCFS M/GE/1/K queue subject to extended Morse balking is derived via the technique of GB. Based on comprehensive numerical experimentation, a conjecture is proposed that the latter GB solution is a special case of the $GdHN_T$ ME distribution.

Owing to the appropriate operational properties of the M/GE/1/K queue subject to extended Morse balking, the queueing system is applied as an ME performance model of Internet Protocol (IP)-based communication network nodes featuring static or dynamic packet dropping congestion management schemes of the instantaneous, random, early-drop type. A performance evaluation study in terms of the model's delay is conducted.

Furthermore, GB analysis is carried out to derive the QLD's of the FCFS GE/GE/1/K censored queue subject to extended Morse balking under three different composite batch balking and batch blocking (batch balk-block) policies. The term 'censored' implies that the prospective arrival process continues irrespective of whether or not the queue has reached its full capacity. A customer which finds all waiting positions occupied upon its

[1] In the interest of maintaining consistency with existing results in the literature in the context of naming the new ME solutions, the convention used earlier in the case of the GGeo and GE distributions has been adopted here. Therefore, the new ME solution based on the prior information constraints of p_0, in addition to those used in the derivation of the dHN ME distribution, is referred to as the generalised dHN (GdHN).

arrival is either forced to leave the system (i.e. it is lost) or it experiences extra delay at an upstream network queueing station (i.e. it is blocked). Supported by extensive numerical experimentation, the latter GB solutions are conjectured to be special cases of the $GdHN_T$ ME distribution.

The performance evaluation of the GE/GE/1/K queue subject to the above three batch balk-block policies is left as future work. In relation to the characterisation of the performance of the GE/GE/1/K queue subject to balking in comparison to that of equivalent two-phase exponential queues with balking, a fourth batch balk-block policy is introduced and analysed in the Appendix. Moreover, in the interest of the ME approximate analysis of arbitrary, non-exponential QNM's with balking or packet dropping congestion management schemes, preliminary analysis is presented in the Appendix by which an approximation of the departure process from the M/GE/1/K queue subject to extended Morse balking can be computed. The accuracy of this approximation has yet to be assessed.

In this thesis, unless stated otherwise, queueing systems are considered under the following conditions:

- steady state (i.e. statistical equilibrium),
- FCFS scheduling discipline,
- Independent inter-arrival times of prospective customers,
- Independent service durations,
- Independence between inter-arrival times of prospective customers and service times,

- Censored rather than restricted[2] prospective arrival process and
- Prospective customers which balk or those which are blocked are considered to not return to the queue at a later time and are therefore deemed to be lost from the point of view of the queue.

1.1. Motivation

The derivation of exact performance distributions of complex queueing systems, such as infinite or finite capacity G/G/1 queues with or without balking, via classical queueing theoretic analysis is often an infeasible task. However, prior queue length moment information can be obtained either analytically in terms of basic system parameters (which are assumed to be known) or numerically via measurement, without knowledge of the distributions themselves (Kouvatsos 1986a; Kouvatsos and Tabet-Aouel 1994).

Hence one of the main aims of this work was to provide justified estimates of complex queueing system performance distributions when only partial prior queue length moment information (including appropriate boundary queue length state probabilities[3]) is known. The novelty lies in generalising existing least biased QLD inferences in the literature by assuming knowledge of additional prior queue length moment information, resulting in the characterisations of the GdHN and $GdHN_T$ ME solutions. These generalised

[2] In the 'restricted' prospective arrival process, prospective customers cease to arrive while the queue is full. The prospective arrival process resumes on availability of queue capacity.
[3] Boundary queue length state probabilities can be interpreted as moments and in Section 2.2 it is shown how they can be represented as prior information constraints.

solutions were expected to provide more accurate inferences for a larger set of actual distributions than their corresponding subclass distributions (Guiasu 1986).

Furthermore, the motivation for analysing the queueing systems with balking was the mathematical intractability posed by the problem of determining the inter-arrival and service time distributions of the infinite and finite capacity ordinary M/G/1 or G/G/1 queues bearing the GdHN and $GdHN_T$ ME QLD's respectively.

1.2. Contributions

The contributions of this thesis to the body of knowledge are regarded to be the following:

- Characterisation of generalised ME solutions, namely the generalised discrete Half Normal (GdHN) and truncated GdHN ($GdHN_T$) for inferences of stationary probability distributions of discrete random variables (RV's) (Chapter 3).
- Extended Morse Balking:
    - Population-dependent Morse balking has been considered in the more general contexts of a general batch prospective arrival process (comprising a general inter-batch time distribution and general batch size distribution) and/or i.i.d. general service times. Analysis has been carried out for the specific cases of the compound Poisson prospective arrival process with geometrically

distributed batch sizes[4] (characterised by underlying i.i.d. GE prospective inter-arrival times) and/or i.i.d. GE service times (Sections 4.1 and 5.1).

  - New mathematical definitions of the Morse average measure of impatience, α, have been derived in the contexts of a general batch prospective arrival process and/or i.i.d. general service times (Appendix B).

- The aforementioned generalised ME solutions are conjectured to have as special cases the QLD's, respectively, of the novel infinite and finite-capacity M/GE/1 queues subject to extended Morse balking (Chapter 4).
  - A performance evaluation study in terms of the delay of the M/GE/1/K queue subject to extended Morse balking (Section 4.3).
- The aforementioned generalised ME solutions are conjectured to have as special cases the QLD's, respectively, of the new infinite and finite-capacity GE/GE/1 queues subject to extended Morse balking under three different batch balk-block policies[5] (Chapter 5).

Minor contributions of this thesis to the body of knowledge are viewed as the following:

- A new characterisation of the GGeo discrete distribution (Section 2.2.1 and Appendix A).

[4] In Section 5.1 it is described how the analysis can easily be extended to model the case of a compound Poisson prospective arrival process with generally distributed batch sizes subject to population dependent balking under the three policies.

[5] To the best of the author's knowledge, independent batch balking with population-dependent balking probabilities has not been analysed previously in the literature except by the author of this thesis in (Shah and Kouvatsos 2011; Shah and Kouvatsos 2013).

- An approximation of the departure process from the M/G/1/K queue subject to population-dependent balking including an exact analysis of the queueing system's inter-departure time distribution (Section 6.2 and Appendix D).

1.3. Thesis Structure

The ME principle, ME solutions of queueing system performance distributions and queues subject to balking are reviewed in Chapter 2. The GdHN and $GdHN_T$ ME distributions are characterised in Chapter 3. In Chapter 4, the QLD of the M/GE/1/K queue subject to extended Morse balking is derived and its equivalence to the $GdHN_T$ is conjectured. Subsequently, the latter queueing model is applied as an ME performance model of IP-based communication nodes featuring congestion management and a performance evaluation study of the model in terms of its delay is carried out. In Chapter 5, the QLD's of the GE/GE/1/K queue subject to extended Morse balking under the three different batch balk-block policies are solved and these are conjectured to be special cases of the $GdHN_T$. Conclusions are drawn, limitations are identified and open problems arising from this research work are presented in Chapter 6.

An alternative characterisation of the GGeo discrete distribution is proven in Appendix A. Novel definitions of the Morse average measure of impatience, α are derived in Appendix B. The GE/GE/1/K queue subject to balking under Policy IV is studied in Appendix C. Moreover, an approximation of the

departure process from the M/G/1/K queue subject to population-dependent balking is derived in Appendix D.

2. Literature Review

This chapter presents the material which is foundational to the contributions of this thesis commencing with an introduction to the ME principle including the derivation of discrete ME distributions. This is followed by an overview of ME solutions of queueing system performance distributions in the second section with a particular focus on the GGeo and truncated GGeo ($GGeo_T$) ME solutions. The third section introduces different types of balking models and reviews in depth the M/M/1 queue subject to Morse balking and its ME re-interpretation.

2.1. The Maximum Entropy Principle

The ME principle, devised by Jaynes (Jaynes 1957; Jaynes 1978), is a method of statistical inference, addressing the problem of assignment of probabilities to system states or a probability distribution to a random quantity of a system or process, when only limited prior information is available.

The two main approaches to statistical inference occur via the 'frequentist' or 'Bayesian' methods, with the latter usually involving a more general notion of probability (Cox 2006). The Bayesian approach can be further classified into subjective (personalistic) or objective Bayesianism. Under subjective Bayesianism, probabilities are viewed as representations of degrees of personal belief of (rational) individuals. Whereas in objective Bayesianism, probabilities are regarded as encapsulations of a state of knowledge,

independent of the personality or feelings of the individual. Under the frequentist view, probabilities are regarded as measurable and verifiable frequencies in a random experiment (Jaynes 2003).

In the ME principle, since the prior information is incorporated and probability assignments are proposed independently of the personality and feelings of individuals and independently of experiment, this inference technique can be classed as objective Bayesian.

The frequentist approach to statistical inference is based on sampling theory and is concerned with concepts such as parameter estimation (including the method of moments and maximum likelihood estimate), estimators, goodness of fit and hypothesis testing among others.

The roots of inference and the ME principle can be traced at least as far back as Bernoulli's 1713 principle of insufficient reason. That principle advocates assigning equal probabilities to events when the available evidence provides no reason to consider otherwise.

The latter principle however fails when an event has an infinite number of equally possible outcomes. To overcome this limitation, Bernoulli devised a method, in the Binomial distribution, to infer the frequency of successes of an event (repeated independently) from its theoretical probability.

Following this, Bayes, in 1763, submitted an inverse solution to the Binomial distribution by which the theoretical probability of an event could be inferred from its observed frequency.

Subsequently in 1774, Laplace (independently) generalised Bayes' solution (which inferred a population parameter value from an observed frequency) in the following way: the conditional probability of a cause given the observed

event (or a population parameter value) can be inferred from the conditional probabilities of the observed event given each (equally likely) cause (or multiple sample frequencies) according to a simple mathematical relationship. This method proposes a posterior probability from uniform prior probabilities and it can be repeated in turn for each conditional causal probability. Laplace later generalised this latter result by incorporating unequal causal probabilities which were conditional on some prior information and it is this result that is customarily referred to as 'Bayes' theorem' in the literature.

The ME principle on the other hand proposes a probability distribution inference of a random quantity based on prior information, which is often taken to be in the form of moments of the random quantity. Usually a large number of distributions satisfy prior moment and normalisation information. So which one of these distributions is to be chosen to provide the most justified or best inference? The distribution which satisfies the known information while avoiding bias. In this context, the 'unbiased' (or 'least biased') distribution has been defined as the one which is 'maximally non-committal to unknown information' (Jaynes 1957). In other words bias assigns certainty to unknown information. This use of the term 'bias' is to be distinguished from its meaning in the sampling context, where it refers to a persistent (or average) difference between the experimental estimate and corresponding population parameter value.

Any distribution which satisfies the prior moment and normalisation information other than the one with maximum entropy is biased. Selecting a biased distribution estimate amounts to making arbitrary assumptions about

the system information. The least biased inference is therefore derived by maximising the uncertainty regarding what is unknown about the random quantity while satisfying the known partial prior information (Jaynes 1957). Shannon proposed a measure of the uncertainty in predicting the outcomes of a random event or values realised by a random quantity (Shannon 1948), defined by

$$H(\dots p_{-2}, p_{-1}, p_0, p_1, p_2 \dots) = -\sum_{n=-\infty}^{\infty} p_n ln p_n \,, n = \dots, -2, -1, 0, 1, 2 \dots \quad (2.1)$$

where the p_n's are the stationary event or state probabilities. Shannon referred to H as the entropy of the distribution, $p_n, n \in \mathbb{Z}$.

The credibility of Shannon's entropy as an information measure is supported by its satisfaction of numerous postulates which are listed for example in (Csiszár 2008) with references therein giving the corresponding proofs.

In order to ascertain whether the least biased inference satisfies the prior information, it is necessary that the latter is precise enough to enable the explicit verification of its agreement with the inferred distribution. Examples of such prior information are moments and bounds. An example of inadmissible information would be 'the first moment of the RV is probably less than 0.6'. Mathematically, the prior information is incorporated into the ME formalism as an optimisation constraint(s) on the distribution.

According to the ME principle, the least biased estimate of a probability distribution which is unknown but known to exist, is determined by

maximising Shannon's entropy functional, subject to verifiable prior information constraints and the normalisation condition.

Based on the satisfaction of four consistency axioms of inference[6], maximising Shannon's entropy[7] subject to prior moment information has been shown to be a 'uniquely[8] consistent' method of inference of probability distributions. 'Uniquely consistent' in this context means that solely the principle of ME gives congruous inferences when the same prior moment information is incorporated in the formalism in different ways. Furthermore, given prior moment information, the ME principle yields only one distribution inference (Shore and Johnson 1980).

By definition, inferences cannot predict the outcome of an experiment exactly. Nonetheless, irrespective of whether or not the ME inference agrees with the observed results, it still provides the 'best' model in the following sense:

- It is the least biased model given the prior information and
- The maximum entropy frequency distribution derived from certain constraints can be shown to be 'overwhelmingly the most likely one to be observed in a real experiment, provided that the physical

[6] In (Shore and Johnson 1980), the term 'inductive inference' is used. This latter term is not used there in the Fisher sense i.e. a logical process of making sense about population statistics from sample ones (Fisher 1935). Rather, the term refers to the estimation of an underlying distribution, characterising for example the probabilities of states of an arbitrary system, based on available prior information.

[7] Any function that produces identical maxima to Shannon's entropy can be used instead, for example any monotonic function of Shannon's entropy (Shore and Johnson 1980).

[8] Both the ME principle and Kullback's more general principle of minimum cross-entropy (or minimum directed divergence or minimum relative entropy) have been shown to be uniquely consistent methods of inference when the prior information given is in the form of moments (Shore and Johnson 1980).

constraints operative in the experiment are exactly those incorporated in the formalism' (Jaynes 1978).

Persistent discrepancies between the inferred distribution and observed probabilities indicate an ill-constrained ME problem formulation i.e. all the relevant information has not been accounted for or false information has been incorporated in the constraints (Jaynes 1968).

The ME principle has been employed to provide solutions to problems in a vast range of areas not limited to business, economics and finance, decision making, group behaviour, linear and nonlinear programming, nonlinear spectral analysis, parameter estimation, pattern recognition, queueing systems, reliability estimation, statistical mechanics and thermodynamics, system modularity, transportation and urban and regional planning (Shore and Johnson 1980; Fang et al. 1997).

2.1.1. Discrete ME Distributions

Discrete ME distributions are derived below, as that is the context of utility of the ME principle in this thesis. Following the derivation in (Jaynes 1957), consider a RV, N, modelling a discrete quantity of a system or process. N takes integer values over some sample space, S where for example $S = \mathbb{Z}$ but the probabilities $p_n = P(N = n), n = \ldots, -2, -1, 0, 1, 2 \ldots$ are unknown and to be determined. Assume that the limited prior information known about N is in the form of its moments, $E[f_i(n)], i = 1, 2, 3 \ldots m$ and the normalising condition.

Moments of the distribution inferences, $E[f_i(n)]$, may either be probability averages or sample moments, each set producing valid least biased distribution inferences of the RV or sample respectively. The validity of using sample moments has been demonstrated using hypothesis testing and the Bayesian and likelihood criteria in (Jaynes 1978). Maximising Shannon's entropy functional (2.1) subject to the following prior moment information constraints

$$\sum_{n=-\infty}^{\infty} f_i(n)p_n = E[f_i(n)], n = \ldots, -2, -1, 0, 1, 2 \ldots, i = 1, 2, 3 \ldots m \tag{2.2}$$

and normalising condition expressed as

$$\sum_{n=-\infty}^{\infty} p_n = 1.0 \tag{2.3}$$

is a constrained nonlinear optimisation problem soluble by the Lagrange multiplier technique. In the latter method, the extrema of a constrained objective function are obtained via an unconstrained Lagrange function, $L(\ldots p_{-2}, p_{-1}, p_0, p_1, p_2 \ldots)$, defined in terms of the original objective function, the constraints and scalar variables (Lagrangian multipliers). In the case of entropy maximisation subject to prior moment information, the Lagrange function can be defined as

$$L(\ldots p_{-2}, p_{-1}, p_0, p_1, p_2 \ldots) \tag{2.4}$$

$$= -\sum_{n=-\infty}^{\infty} p_n ln p_n + \sum_{i=0}^{m} \beta_i \left(\sum_{n=-\infty}^{\infty} f_i(n) p_n - E[f_i(n)] \right),$$

$$n = \ldots, -2, -1, 0, 1, 2, \ldots; i = 0, 1, 2 \ldots m$$

where $\beta_i, i = 1,2 \ldots m$ are the Lagrangian multipliers corresponding to each of the m prior moment constraints and β_0 is the Lagrangian multiplier associated with the normalising condition. Hence, $f_0(n) = 1, \forall n$. The state probabilities, p_n's at which entropy is maximised are obtained by setting $\partial L/(\partial p_n) = 0, n = \ldots, -2, -1, 0, 1, 2, \ldots$, resulting in the general discrete ME probability distribution in standard form given by

$$p_n = e^{-\sum_{i=0}^{m} \beta_i f_i(n)}, n = \ldots, -2, -1, 0, 1, 2 \ldots, i = 0, 1, 2 \ldots m\,. \tag{2.5}$$

Equation (2.5) can then be represented in terms of its normalising constant, $(1/Z)$, as

$$p_n = \frac{1}{Z} e^{-\sum_{i=1}^{m} \beta_i f_i(n)}, n = \ldots, -2, -1, 0, 1, 2 \ldots, i = 1, 2, 3 \ldots m \tag{2.6}$$

where the inverse of the normalising constant, Z is given by

$$Z = \sum_{n=-\infty}^{\infty} e^{-\sum_{i=1}^{m} \beta_i f_i(n)}, n = \ldots, -2, -1, 0, 1, 2 \ldots, i = 1, 2, 3 \ldots m. \tag{2.7}$$

The Lagrangian multipliers, $\beta_i, i = 1,2,3 \ldots m$ can be determined in terms of the moments via the following partial derivative

$$-\frac{\partial \beta_0}{\partial \beta_i} = E[f_i(n)], i = 1,2,3 \dots m \tag{2.8}$$

where $\beta_0 = \ln Z$.

Without loss of generality, the Lagrangian coefficients, $x_i = e^{-\beta_i}, i = 1,2,3 \dots m$, can be introduced. Making the latter substitutions in (2.6) yields the following general product-form discrete ME distribution

$$p_n = \frac{1}{Z}\prod_{i=1}^{m} x_i^{f_i(n)}, n = \dots, -2, -1,0,1,2 \dots, i = 1,2,3 \dots m \tag{2.9}$$

where $Z = \sum_{n=-\infty}^{\infty}\left(\prod_{i=1}^{m} x_i^{f_i(n)}\right)$.

As an example, consider the case when the first moment of N, $E[N]$ is known, represented as an information constraint by

$$\sum_{n=0}^{\infty} np_n = E[N], n = 0,1,2 \dots \tag{2.10}$$

and the domain of N is known to be the set of non-negative integers, represented as an information constraint by the normalising condition

$$\sum_{n=0}^{\infty} p_n = 1.0\,. \tag{2.11}$$

Then, the least biased inference of the probability distribution of N can be derived by incorporating the prior information constraint of the first moment (2.10) in the general discrete product form ME solution (2.9) resulting in the following discrete ME distribution

$$p_n = \frac{1}{Z} x_1^n, n = 0,1,2\ \ldots. \tag{2.12}$$

Applying the normalising condition (2.11), (2.12) can be re-formulated to

$$p_n = (1 - x_1)x_1^n, n = 0,1,2\ldots \tag{2.13}$$

which is the familiar modified geometric. Moreover, applying the first moment formula, (2.13) becomes

$$p_n = \left(\frac{1}{1 + E[N]}\right)\left(\frac{E[N]}{1 + E[N]}\right)^n, n = 0,1,2\ldots. \tag{2.14}$$

It is to be noted that in general, the Lagrangian multipliers and consequently the ME distribution parameters $x_i, i = 1,2,3\ldots m$ and the inverse of the normalising constant, Z, cannot be expressed explicitly in terms of the moments, $E[f_i(n)]$, however they can be approximated numerically from the latter.

2.2. ME Solutions of Queueing System Performance Distributions

ME solutions have been devised for inferring distributions of various queueing system performance measures such as the queue length, number served in a busy period, busy period length and residence and waiting times. Examples of sets of prior information constraints (excluding the normalisation condition) used in the derivation of ME solutions of queueing system performance distributions are given below.

Equation (2.13) above is noticeable as the QLD of the M/M/1 queue, where $x_1 = \lambda/\mu$ and λ and μ are the average arrival and service rates of the queue. Therefore, given prior information of solely the MQL, $E[N]$ (either as a numerical value or as an analytic expression in terms of the basic queueing parameters λ and μ) and normalisation condition over the set of non-negative integers, the ME principle prescribes the least biased QLD as being that of the M/M/1 queue (Beneš 1965; Cantor et al. 1986).

In (Shore 1982), the MQL, p_0, successive ordinary queue length moments, mean residence time and/or other exact moments of performance measures (as appropriate) were used in ME solutions approximating all the afore-mentioned performance distributions of the M/G/1 queue. Some of these latter ME solutions turn out to become exact in the case of the M/M/1 queue. First moment and first and second ordinary moments were incorporated in two ME solutions both approximating the distribution of number of customers served during the busy period in an M/G/1 retrial queue in (Lopez-Herrero 2002).

Approximate marginal MQL's and marginal server utilisations were included in an ME solution approximating the joint population distribution of a closed QNM of single server queues with non-exponential service times in (Walstra 1985). On the other hand, marginal server utilisations, marginal MQL's and joint first moment (to capture correlation between the individual queueing subsystems) were employed in an ME solution approximating the joint population distribution of two queues in tandem in (Cantor et al. 1986). In (Kouvatsos and Awan 2003), approximate marginal MQL's, marginal server utilisations, marginal queue occupation probabilities and marginal full buffer state probabilities were utilised in ME solutions of G/G/1/K priority queues under the pre-emptive resume (PR) or head-of-the-line (HOL) scheduling disciplines combined with buffer sharing schemes. These latter queueing systems were utilised as building block queues, in the ME approximate analysis by decomposition, of non-exponential open QNM's with space and service priorities under the repetitive service with random destination blocking mechanism.

The MQL and set of state probabilities $\{p_0, p_1, \dots p_{c-1}\}$ were incorporated in an ME solution of the QLD's of M/G/c, G/M/c and G/G/c queues in (Wu and Chan 1989). The ME solution was found to provide an exact inference of the QLD of the G/M/c queue. On the other hand, in (Arizono et al. 1991), the MQL, mean buffer length and $P(all\ c\ servers\ busy)$ were used in an ME solution modelling exactly the QLD of the M/M/c queue. Approximate marginal MQL's and the probabilities of having a minimum of j jobs of class i in service were utilised in ME solutions approximating the QLD's of multiple

class G/G/c queues under the PR scheduling discipline in (Kouvatsos and Tabet-Aouel 1994).

The MQL and server utilisation were used in ME solutions approximating the marginal QLD's of the N-policy M/G/1 queue with removable server in (Wang et al. 2002). Moreover, the MQL, server utilisation, $P(server\ on\ vacation)$ and $P(server\ broken\ down)$ comprised the prior information constraints in ME solutions approximating the marginal QLD's of the N-policy M^X/G/1 queue with server vacation and breakdown in (Ke and Lin 2008). In (Yang et al. 2011), the MQL or second moment of queue length and marginal server utilisation prior information constraints yielded ME solutions approximating the QLD of M/G/1 queues with second optional service and server breakdowns.

The derivations of ME solutions of the QLD's of infinite and finite capacity ordinary queues and properties of these queues pertinent to the contributions of this thesis are detailed below.

2.2.1. ME Solutions of the QLD's of Ordinary Infinite-Capacity Queues

ME solutions have been proposed for the QLD of the ordinary M/G/1 queue in (Shore 1982; El-Affendi and Kouvatsos 1983; Guiasu 1986). The ordinary M/G/1 queue is characterised by a homogeneous Poisson arrival process (with mean rate λ) and i.i.d. general service times (with mean rate μ and

squared coefficient of variation (SCOV), C_s^2). In these ME solutions λ, μ and C_s^2 comprise a basic set of known queueing parameters.

In (El-Affendi and Kouvatsos 1983), an ME solution for the QLD of a stable M/G/1 queue, the generalised geometric (GGeo) ME QLD, was devised. The GGeo explicitly incorporates the queue stability (or conservation of flow) condition (satisfied implicitly by definition) in addition to the MQL and normalisation prior information constraints. Queue stability exists when the average effective arrival rate to a queue coincides with that departing from it, expressed by

$$\lambda = \mu(1 - p_0) \tag{2.15}$$

where p_0 is the probability of the queue being empty. Instability occurs when the average input rate exceeds the average rate of output from the queue. The queue stability condition can be included as a prior moment information constraint as follows[9]

[9] The 'p_0' prior information constraint can be replaced by the server utilisation constraint, $(1 - p_0)$, as they are equivalent prior information constraints both leading to the same ME inference. This assertion is justified by the ME principle's implicit satisfaction of the uniqueness axiom of inference. The uniqueness axiom implies that maximising Shannon's entropy subject to a particular set of prior moment information constraints results in a unique ME distribution inference. Therefore, the same (unique) ME inference obtained twice by entropy maximisation subject to either one of two sets of prior information constraints can only be achieved if the prior information is the same (i.e. if both sets of prior information constraints are equivalent) (Shore and Johnson 1980).

$$\sum_{n=0}^{\infty} u'(n)p_n = p_0 = 1 - \frac{\lambda}{\mu}, \quad u'(n) = \begin{cases} 1, n = 0 \\ 0, n = 1,2,3 \dots \end{cases}. \tag{2.16}$$

Inclusion of the queue stability condition as a prior information constraint assumes knowledge of the values of λ, μ and p_0 (which inherently satisfy the queue stability condition). The 'p_0' prior information supplies additional, non-redundant prior information above the MQL and normalising condition thus yielding a more general ME solution than that derived from solely the latter two.

Within the context of general system modelling, the discrete ME distribution constrained by the prior information of first moment, $E[N]$ (2.10), lower boundary state probability, p_0 (2.16), and normalisation condition over the set of non-negative integers, (2.11), is the GGeo (El-Affendi and Kouvatsos 1983) and it is derived below[10].

Incorporating the first moment and p_0 prior information constraints in the general product form discrete ME solution (2.9) yields the GGeo with form

$$p_n = \begin{cases} \frac{1}{Z}\left(\frac{1}{y}\right), n = 0 \\ \\ \frac{1}{Z}x^n, n = 1,2,3 \dots \end{cases} \tag{2.17}$$

[10] An alternative characterisation of the GGeo is presented in Appendix A.

where x and $(1/y)$ are the Lagrangian coefficients associated with the prior moment constraints, $E[N]$ and p_0, respectively. The Lagrangian coefficient, $(1/y)$, is defined in this inverted manner purely for the convenience of avoiding parameters of this form in the more familiar definition of the GGeo given by

$$p_n = \begin{cases} p_0, n = 0 \\ p_0 y x^n, n = 1,2,3 \ldots \end{cases} \tag{2.18}$$

where $p_0 = 1 \,/\, (Zy)$. By applying the first moment formula (2.10) and normalising condition (2.11), both parameters x and y can be determined in terms of the constraints $E[N]$ and p_0, yielding the following re-parameterised GGeo (Kouvatsos 1988)

$$p_n = \begin{cases} p_0, n = 0 \\ \left(\frac{(1-p_0)^2}{E[N]-(1-p_0)}\right)\left(\frac{E[N]-(1-p_0)}{E[N]}\right)^n, n = 1,2,3 \ldots \end{cases}. \tag{2.19}$$

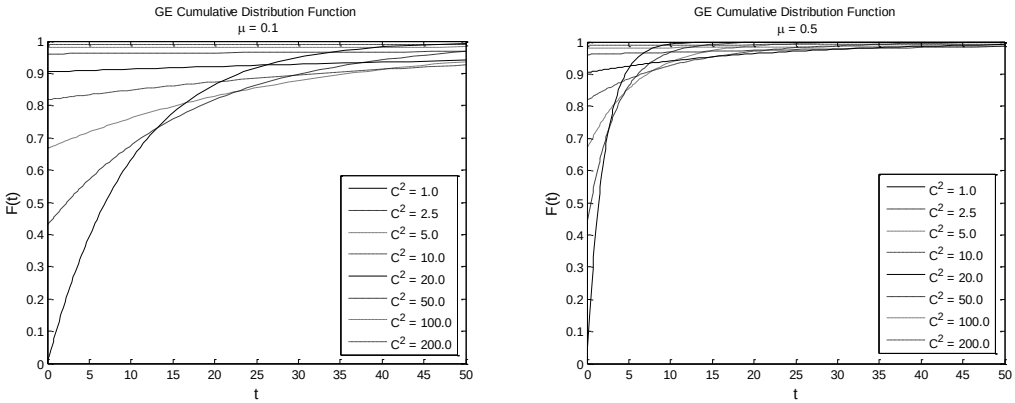
When the first moment prior information constraint is set to the MQL of the M/G/1 queue under consideration and p_0 is specified as $\left(1-(\lambda/\mu)\right)$, then the resulting ME inference, referred to here as the GGeo ME QLD, comprises a subset of all GGeo ME solutions.

The service-time distribution of the M/G/1 queue possessing the GGeo ME QLD was discovered to be satisfied exactly by the GE cumulative distribution function (CDF) (El-Affendi and Kouvatsos 1983). The GE CDF is a two parameter mixed distribution comprising a continuous exponential component and a discrete component at the origin and it can be completely defined in terms of its first two moments as follows

$$F_t = 1 - \frac{2}{C^2+1} e^{-\frac{2}{C^2+1}\mu t}, t \geq 0, \mu > 0 \qquad (2.20)$$

where $(1/\mu)$ is its mean and C^2 is its SCOV. Example profiles of the GE CDF are illustrated in Fig. 1 below over the parameter ranges $\mu = \{0.1, 0.5\}$ and $C^2 = [1, 200]$.

Fig. 1. Example profiles of the generalised exponential (GE) cumulative distribution function with mean rate, $\mu = 0.1$ (left) and $\mu = 0.5$ (right) and squared coefficient of variation, $C^2 = [1, 200]$.



The GGeo was also submitted as an ME solution for the QLD's of G/G/1 and G/M/1 queues, characterised by i.i.d. general inter-arrival times (with known mean rate, λ and SCOV, C_a^2) and, respectively, i.i.d. general or exponential service times (with known mean rate, μ and SCOV, C_s^2 or solely known mean rate, μ). In both these latter two cases, the GGeo ME QLD was again found to be satisfied exactly when the general inter-event time distribution, 'G' was specified as the GE (Kouvatsos 1988).

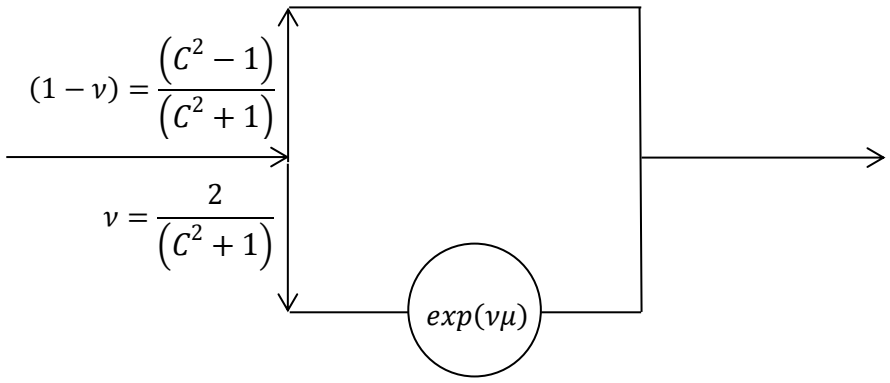
For $C^2 > 1.0$, the GE CDF has also been interpreted as either (Kouvatsos 1988; Kouvatsos 1994)[11]:

- the inter-event time distribution of a compound Poisson process with geometrically distributed batch sizes with mean, $(1/\nu)$, or
- an extremal case of a family of two-phase exponential distributions (e.g. the Hyper-exponential-2 (H_2) distribution) with matching first and second moments, where one of the two phases has zero duration (illustrated in Fig. 2 below).

It has been established experimentally that for $C_a^2, C_s^2 > 1.0$, the GE/GE/1 queue gives pessimistic performance bounds over a large class of equivalent queues characterised by two-phase exponential inter-arrival and service time distributions, such as the H_2 or Coxian-2, with matching first two moments (Kouvatsos 1988; Kouvatsos and Tabet-Aouel 1994).

[11] Interpretations and comparative performance bounds involving the GE distribution when $C^2 < 1$ are given in (Kouvatsos 1988; Kouvatsos 1994)

Fig. 2. An illustration of the two-phase GE distribution interpretation (for $C^2 > 1.0$).



Thus the GE inter-event time distribution provides a justifiable, useful and cost-effective means of analytically modelling burstiness in arrival and/or service processes of queues via the C^2 term.

2.2.2. ME Solutions for the QLD's of Ordinary Finite-Capacity Queues

Analysis based on the ME principle has also been carried out for the QLD's of stable, finite capacity G/M/1/K, M/G/1/K and G/G/1/K queues (Kouvatsos 1986b; Kouvatsos 1986a). For the G/G/1/K queue, the stability condition is expressed as

$$\lambda(1 - p_K^K) = \mu(1 - p_0^K) \tag{2.21}$$

where p_K^K and p_0^K are the fractions of time that the finite-capacity queue is full and empty respectively, λ is the prospective arrival rate (a fraction of which is lost when the queue is full) and μ maintains its former interpretation. Therefore explicit inclusion of the finite-capacity queue stability condition as prior information requires knowledge of both boundary state probabilities, p_0^K and p_K^K, in addition to λ and μ. The boundary state probabilities, p_0^K and p_K^K, can be represented as prior moment information constraints[12] as follows:

$$\sum_{n=0}^{K} u'_K(n)p_n^K = p_0^K = \frac{\mu - \lambda + \lambda p_K^K}{\mu}, \qquad u'_K(n) = \begin{cases} 1, n = 0 \\ 0, n = 1,2 \dots K \end{cases} \tag{2.22}$$

and

$$\sum_{n=0}^{K} f_K(n)p_n^K = p_K^K = \frac{\mu p_0^K + \lambda - \mu}{\lambda}, \qquad f_K(n) = \begin{cases} 0, n = 0,1 \dots K-1 \\ 1, n = K \end{cases} \tag{2.23}$$

respectively.

The discrete distribution having maximum entropy subject to the prior information constraints of the first moment, $E[N_K]$, boundary state probabilities p_0^K and p_K^K and normalisation condition over finite non-negative integer support, $[0,\ K]$ is the GGeo_T (2.24) i.e. the GGeo right-truncated

[12] Analogous to the infinite capacity case, the p_0^K prior information constraint is equivalent to the server utilisation constraint, $(1 - p_0^K)$.

above K and constrained additionally by the upper boundary state probability, p_K^K (Kouvatsos 1986a; Kouvatsos 1986b). The GGeo$_T$ can be defined by

$$p_n^K = \begin{cases} p_0^K, & n = 0 \\ p_0^K y_K (x_K)^n, & n = 1,2,3 \dots K-1 \\ p_0^K z_K y_K (x_K)^K, & n = K \end{cases} \quad (2.24)$$

where the parameters x_K, y_K and z_K correspond to the Lagrangian coefficients associated with the prior moment constraints $E[N_K]$, p_0^K and p_K^K respectively. Note that unlike the GGeo, the parameters of the GGeo$_T$ cannot be expressed explicitly in terms of the prior moment information; however they can be approximated numerically from the latter as follows. By equating the analytic expressions of appropriate moments, $\sum_{n=0}^K f_{Ki}(n) p_n^K$, of the GGeo$_T$ distribution (2.24) to their values represented by, $E[N_K]$, p_0^K and p_K^K, the following system of nonlinear equations arises

$$\frac{(1-p_0^K-p_K^K)(1-x_K)}{(1-(x_K)^{K-1})}\left(\frac{1-(x_K)^{K+1}}{(1-x_K)^2}-\frac{(K+1)(x_K)^K}{(1-x_K)}-K(x_K)^{K-1}\right)+Kp_K^K-E[N_K]=0, \quad (2.25)$$

$$y_K = \frac{\left(1-p_0^K-p_K^K\right)(1-x_K)}{p_0^K x_K (1-(x_K)^{K-1})} \quad (2.26)$$

and

$$z_K = \frac{p_K^K}{p_0^K y_K (x_K)^K} \tag{2.27}$$

which can be solved simultaneously to give values of the parameters x_K, y_K and z_K. In (Kouvatsos 1986b; Kouvatsos 1986a), it was proposed that the parameters, x_K and y_K, be approximated by the analytic expressions for the corresponding parameters of the infinite capacity case, x and y (cf., (2.18) and (2.19)). An asymptotic approximation for the parameter z_K can in turn be determined from the expression for p_0^K and the stability constraint (2.21) (Kouvatsos 1986b). These asymptotic approximations are employed in subsequent works where complex queueing systems are analysed approximately via the ME principle, cf., (Kouvatsos and Almond 1988; Kouvatsos and Denazis 1993; Kouvatsos and Awan 2003; Kouvatsos et al. 2003).

Analogous to the corresponding infinite capacity queues, the G/M/1/K, M/G/1/K and G/G/1/K queues bear the $GGeo_T$ ME QLD when the general inter-event time distribution, 'G' is specified as the GE (Kouvatsos 1986a; Kouvatsos 1986b).

2.3. Queues with Balking

In queueing systems, the balking operation as well as reneging and retrials comprise models of customer impatience (Wang et al. 2010). Queues with balking have been successfully applied to evaluate multiple real life systems (such as those listed in the Introduction).

Following the categorisation of impatience models by (Wang et al. 2010), balking models are classed as wait-based, individual equilibrium and socially optimal or others.

Wait-based balking

In wait-based balking, a customer becomes aware of the queue's delay and either joins or balks based on his/her delay tolerance. Delay information is available at different degrees of precision such as the exact instantaneous workload, average instantaneous workload or long-run average delay (Liu 2007). The effects of the availability of two different approximations of delay information have been compared in (Guo and Zipkin 2009). Wait-based balking may be subdivided into three types: pure threshold, delay function and conditional probability function. Examples of works analysing each of these types are given below.

In (Haight 1957), various discrete customer threshold distributions (called balking distributions) are analysed in the context of the M/M/1 queue subject to balking. Whereas in (Liu 2007), the M/G/1 queue subject to deterministic pure threshold balking was analysed assuming that the exact instantaneous

workload was known. The analysis employed the level crossing theory of stochastic processes. Lu and Mark (Lu and Mark 2004) model an optical burst switch employing fibre delay lines using an M/M/c queue subject to deterministic pure threshold balking. This is analysed via the Markov chain model of queue's equivalent probabilistic balking model. In (Ward and Glynn 2005), customers balk from the GI/GI/1 queue if the conditional average waiting time exceeds the customer's (generally-distributed) threshold. This latter queueing system is analysed via the diffusion approximation method.

The M/M/1 queue subject to balking governed by a reward-cost structure which is dependent on the delay information is analysed in (Guo and Zipkin 2007). In that analysis balking under the above three different precision levels of delay information is studied.

The third type of wait-based balking, namely the conditional probability function, comprises the constant balking probability model since the latter can be interpreted as a balking probability function dependent on the long-run average delay. In (Blackburn 1972), the M/G/1 queue subject to balking and reneging is analysed under a reward-cost structure. The balking probability is fixed and independent of the reward-cost structure. On the other hand, the constant balking probability model is used in the steady state analysis of Markovian queues subject to balking in (Al-Seedy 1995; Al-Seedy 1996). In (Whitt 1999), analysis is carried out for the M/M/c/K queue subject to balking characterised by a probability functional dependent on the probability that the prospective arrival's delay tolerance is greater than the virtual queueing time. Despite being based on approximations of the instantaneous delay, population-dependent balking probability functions as

analysed in (Ancker and Gafarian 1963; Gupta 1995; Zhen et al. 2010) are included under this type of wait-based balking.

Individual equilibrium and socially optimal solutions

This avenue of modelling is widely regarded to go back at least to the pioneering works of Naor (Naor 1969; Economou et al. 2011). Naor studied the M/M/1 queue subject to balking where the decision of prospective arrivals to join (or balk) is guided by the net gains from a reward-cost structure dependent on the observed instantaneous queue size. In that work, individual equilibrium (where payoff for individual customers is maximum) and socially optimal (where overall welfare of the customer population is maximum) solutions were determined (Naor 1969). This framework lends itself to game theoretic analysis since it involves customers (indistinguishable in this case) potentially gaining payoffs through making strategic, informed decisions and taking associated actions.

The unobservable case, where customers are unaware of the queue size, has also been studied. For example in (Hassin and Haviv 1995), the M/M/1 queue subject to balking and deadline-based reneging under a reward-cost structure was considered in the unobservable case. In this latter work, both individual equilibrium and socially optimal solutions were defined. Furthermore, general service time distributions have also been considered within this framework. In (Economou et al. 2011) for example, individual equilibrium and socially optimal strategies were identified for the M/G/1

queue subject to balking and general server vacation times in the unobservable or partially observable (server status information) cases.

A comprehensive review of balking models of this nature is presented in (Hassin and Haviv 2003).

Other balking models

Balking probability sequences as utilised in (Rao and Jaiswal 1969; Zhang et al. 2005; Yue et al. 2006) are included in this category. On the other hand, fuzzy set theory is used to analyse the M/M/1/K queue subject to three balking models characterised by uncertain distributions in (de La Fuente and Pardo 2009).

2.3.1. The Morse Balking Paradigm

Morse (Morse 1958) considered the M/M/1 queue subject to wait-based balking characterised by a conditional probability function. This model is characterised by a single server, infinite capacity queue with a homogenous Poisson prospective arrival process with mean intensity, λ and i.i.d. exponential service times with mean rate, μ. Customers join the queue with conditional probability governed by the Morse joining function defined by

$$q(t) = e^{-\alpha t}, \alpha, t \geq 0 \qquad (2.28)$$

where t is the instantaneous workload (or instantaneous virtual queueing time i.e. the queueing delay from the point of view of a prospective arrival that joins the queue) and α was interpreted by Morse as a measure of the average customer impatience (unwillingness to wait in line) (Morse 1958). The Morse joining function is seen to satisfy two fundamental properties of impatience models namely the non-increasing property (as t increases) and the bounding condition $q(0) = 1.0$.

The 'balking function' refers to the conditional probability with which customers refrain from entering the queue. The Morse balking function defined by

$$1 - q(t) = (1 - e^{-\alpha t}) \tag{2.29}$$

is strictly increasing (for $\alpha > 0$). Owing to Haight's balking paradigm (Haight 1957), the parameter $(1/\alpha)$ can be interpreted qualitatively as an average measure of customers' need for or importance assigned to receiving service.

The qualitative parameter, α, is associated with the overall stance of a customer population towards queueing. In contrast to the pure threshold-based balking models, α enables more varied balking behaviour to be modelled. Whereas in pure threshold-based balking models all customers would balk when the workload exceeds the pre-set threshold, under the Morse balking paradigm, customers from patient populations (or those

having a strong urgency for service) may join, even with relatively large probability at that same or higher workload levels.

The converse would be true for low workloads. In the pure threshold models, all prospective arrivals would join the queue when the instantaneous workload falls below the pre-set threshold. However, the Morse balking model permits customers from a very impatient population (or those indifferent to service) to balk with large probability at the same (low) instantaneous workload.

In most cases in reality t is not known, however a useful estimate of t is t_{av}, the average instantaneous workload conditional on the instantaneous queue length, n. This is because n can usually be obtained as advocated by, among others, the observable queue model reviewed in depth in (Hassin and Haviv 2003). Owing to the memoryless property of the exponential service-time distribution, the average instantaneous workload conditional on n is given by

$$t_{av} = nE[s] = \frac{n}{\mu}, n = 0,1,2 \ldots \quad (2.30)$$

where $E[s]$ is the average service time (per customer).

Substituting t in the Morse joining function (2.28) by t_{av} (2.30) results in the population - dependent Morse joining function given by

$$q(n) = e^{-\alpha\frac{n}{\mu}}, n = 0,1,2 \ldots . \quad (2.31)$$

An equivalent re-parameterised variant of this latter population-dependent joining function (2.31) has been described as a 'very natural' balking model 'that should be appropriate in many applications' (Mendelson et al. 1999).

The QLD of the M/M/1 queue subject to Morse balking defined as

$$p_n = p_0 \left(\frac{\lambda}{\mu}\right)^n e^{-\frac{\alpha n(n-1)}{\mu \quad 2}}, n = 0,1,2 \ldots \quad (2.32)$$

was derived in (Morse 1958). This was achieved by solving the GB equations of the queueing system's birth-death (B-D) chain model, where the effective arrival rates to each state accounted for the fraction of customers which balked.

In (Morse 1958), a graphical method was submitted to obtain an approximate value of the parameter, α. Alternatively, an approximate value of α, for a particular customer population can be obtained numerically by equating the analytic expression of a suitable metric (which is in terms of α), such as the average balking rate (BR), MQL or average waiting time, to its value. For example, given the values of λ, μ and BR, solving

$$\lambda \sum_{n=0}^{\infty} q(n)p_n - (\lambda - BR) = 0 \tag{2.33}$$

for α provides the desired estimate, where $q(n)$ and p_n are given by (2.31) and (2.32) respectively. Other suitable metrics would result in analogous nonlinear equations. A mathematical definition of the parameter, α, owing to equivalence between the Morse and Haight balking paradigms, is derived in Appendix B.

2.3.2. The QLD of the Stable M/M/1 Queue Subject to Balking

Consider a stable M/M/1 queue characterised by a homogeneous Poisson prospective arrival process with rate, λ, subject to balking with population-dependent function $\big(1 - q(n)\big), n = 0,1,2, \ldots$ and i.i.d. exponential service times with mean rate, μ.

It is well known that a homogeneous Poisson arrival process with rate, λ subject to decomposition into S sub-processes by routing arrivals to path j with fixed and independent probabilities, $q_j, j = 1,2 \ldots S$, results in j Poisson sub-processes each with diminished rates, $\lambda q_j, j = 1,2 \ldots S$ (Kleinrock 1975; Kuehn 1979).

In the above queueing system, since customers behave independently of each other with respect to balking, at each state, $n = 0,1,2, \ldots$, the joining probability, $q(n)$, and balking probability, $\big(1 - q(n)\big)$, are fixed and

independent of each other. Therefore, the effective arrival process to the queue is Poisson with state/population - dependent rates given by

$$\lambda_n = \lambda q(n), n = 0,1,2 \ldots . \tag{2.34}$$

As a consequence, the queueing system with balking has an equivalent transformation in the M(n)/M/1 queue without balking but with a population-dependent Poisson arrival process with rates given by (2.34) (Haight 1957). The QLD of the stable M/M/1 queue subject to population-dependent balking can be derived by solving the steady state probability distribution of the B-D chain of its equivalent M(n)/M/1 queue illustrated by

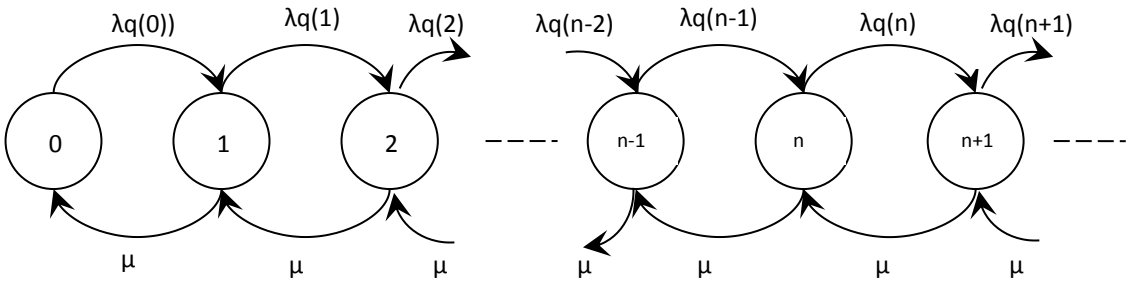


Fig. 3. Birth-Death chain of the M(n)/M/1 with state-dependent arrival rates, $\lambda_n = \lambda q(n)$.

and is given by (Haight 1957)

$$p_n = \left(\frac{\lambda}{\mu}\right)^n \left(\prod_{i=0}^{n-1} q(i)\right) p_0, n = 0,1,2 \ldots . \quad (2.35)$$

The variable rate Poisson arrival process, as well as modelling the effective arrival process to queues with balking, models the arrival process to queueing systems with finite customer populations such as the machine interference system and queues with finite capacity (Courtois and Georges 1971; Gupta and Rao 1996).

2.3.3. An ME Re-Interpretation of the Morse Balking Queueing Solution

The discrete Half Normal (dHN) has been characterised as that unique discrete distribution having maximum entropy constrained by the first moment, variance and normalisation condition over non-negative integer support. It has been defined in (Kemp 2008) as

$$p_n = \frac{\theta^n q^{\frac{n(n-1)}{2}}}{\sum_{n=0}^{\infty} \theta^n q^{\frac{n(n-1)}{2}}} = p_0 \theta^n q^{\frac{n(n-1)}{2}}, n = 0,1,2 \ldots, \theta > 0, 0 < q < 1 . \quad (2.36)$$

The justification for this ME distribution being called the 'discrete Half Normal' is that it is the discrete analogue of the unique continuous

distribution on the non-negative half-line having maximum entropy for given mean and variance, the continuous Half-Normal (Kemp 2008).

Notably, ordinary moment (from the second upward) prior information constraints yield the same ME solutions to those obtained when corresponding central moments are used, provided that the first moment information constraint is explicitly incorporated in both ME formulations. This latter requirement is necessary because the first moment information is inherently contained within central moments.

Kemp discovered specific state-transition probabilities of a Markov chain model which result in the dHN stationary state probability distribution (2.36). Furthermore, Kemp also observed that the dHN state probability distribution has as a special case, the QLD of the M/M/1 queue subject to Morse balking (2.32) when $\theta = \lambda/\mu$ and $q = e^{-\alpha/\mu}$ (Morse 1958; Kemp 2005; Kemp 2008). Relating to this observation, Haight had decades earlier discovered the close approximation of the QLD of the M/M/1 queue subject to balking characterised by the modified geometric balking distribution (which is equivalent to the QLD of the M/M/1 queue subject to Morse balking (cf., Appendix B)) to Normal ordinates considered solely at non-negative integer abscissae (Haight 1957).

It is to be recalled that the ME solution of the QLD of a single server queue, derived from the prior information constraints of solely the MQL and normalising condition is the QLD of the ordinary M/M/1 queue. Hence, Kemp's aforementioned observation implies that including the variance of queue length (VQL) prior information constraint in the ME solution of the

QLD of a single server queue is equivalent to maximising the uncertainty arising from the selective behavior (or choice) of prospective arrivals which may either join the queue or balk according to the respective Morse joining or balking functions.

3. Generalised Maximum Entropy Solutions

In this chapter, new ME solutions namely the generalised discrete half normal (GdHN) and truncated GdHN ($GdHN_T$) are characterised. Subsequently the effect of different prior moment information on the profiles of corresponding ME inferences is observed.

The GdHN and $GdHN_T$ comprise least biased stationary probability assignments for discrete values realised by random quantities or for discrete states of any general system or process.

Motivated by the provision of more accurate inferences of queueing system performance distributions, existing ME solutions, namely the GGeo and $GGeo_T$ (cf., Section 2.2) and dHN (cf., Section 2.3.3), are generalised by combining, as appropriate, the prior queue length moment information assumed known in each of the three cases. As a consequence, the GdHN and $GdHN_T$ ME state probability distributions are suited to least biased inferences of the stationary QLD's of, respectively, infinite and finite-capacity ordinary G/G/1 queues, G/G/1 queues subject to extended Morse balking (based on the implication of Kemp's observation) or ordinary G/G/1 queues subject to population-dependent arrivals rates governed by the extended Morse joining function (based on the equivalence between balking and state-dependent arrival rates).

In this context, prior queue length moment constraint information may be known numerically via measurements over a finite observation period or may be derived analytically, in terms of basic system parameters (whose values are assumed to be known), via operational or stochastic properties and/or assumptions (Kouvatsos 1986a).

3.1. The Generalised Discrete Half Normal

The GdHN is characterised as that unique discrete distribution having maximum entropy, given the prior information of the first moment, variance and lower boundary state probability, over non-negative integer support. The respective ME optimisation prior moment constraints, $E[N]$, $E[(N - E[N])^2]$ and p_0 are represented as

$$\sum_{n=0}^{\infty} np_n = E[N], n = 0,1,2 \dots \tag{3.1}$$

$$\sum_{n=0}^{\infty} (n - E[N])^2 p_n = E[(N - E[N])^2], n = 0,1,2 \dots , \tag{3.2}$$

and

$$\sum_{n=0}^{\infty} u'(n)p_n = p_0, \quad u'(n) = \begin{cases} 1, n = 0 \\ 0, n = 1,2,3 \dots \end{cases}. \tag{3.3}$$

Incorporating the prior moment information constraints (3.1) - (3.3) and the following normalising condition

$$\sum_{n=0}^{\infty} p_n = 1.0 \tag{3.4}$$

directly in the general product-form discrete ME solution defined by

$$p_n = \frac{1}{Z}\prod_{i=1}^{m} x_i^{f_i(n)}, n = \dots, -2, -1,0,1,2 \dots, i = 1,2,3 \dots m \tag{3.5}$$

where $Z = \sum_{n=-\infty}^{\infty}\left(\prod_{i=1}^{m} x_i^{f_i(n)}\right)$, yields the GdHN expressed as

$$p_n = \frac{1}{Z}x_1^n x_2^{(n-E[N])^2} x_3^{u'(n)}, n = 0,1,2,3 \dots \tag{3.6}$$

where $Z = \sum_{n=0}^{\infty}\left(x_1^n x_2^{(n-E[N])^2} x_3^{u'(n)}\right)$ and x_1, x_2 and x_3 are the Lagrangian coefficients corresponding to the first moment, variance and p_0 prior moment information constraints respectively. Equation (3.6) can be written equivalently as

$$p_n = \begin{cases} \frac{1}{Z} x_2^{(E[N])^2} x_3, n = 0 \\ \frac{1}{Z} x_1^n x_2^{(n-E[N])^2}, n = 1,2,3 \ldots \end{cases} . \quad (3.7)$$

For the purpose of comparing the GdHN to its subclass distributions, the distributional form of the GdHN in (3.7) is manipulated to resemble as closely as possible, the forms of both the GGeo (2.18) and dHN (2.36) ME distributions, concurrently as follows. Expanding the product $(n - E[N])^2$ and introducing the term $(2E[N] - 1)n$ in the power of x_2 in (3.7) yields

$$p_n = \begin{cases} \frac{1}{Z} x_2^{(E[N])^2} x_3, n = 0 \\ \frac{1}{Z} x_2^{(E[N])^2} \left(x_1 x_2^{-(2E[N]-1)} \right)^n x_2^{(n^2-n)}, n = 1,2,3 \ldots \end{cases} \quad (3.8)$$

where the term $\frac{1}{Z} x_2^{(E[N])^2} x_3$ is clearly the zero state probability, p_0.

Without loss of generality, the parameters γ, ϕ and r are introduced. The substitutions, $(x_3)^{-1} = \gamma$, $\left(x_1 x_2^{-(2E[N]-1)} \right) = \phi$ and $x_2 = r^{\frac{1}{2}}$ in (3.8) yield the re-parameterised GdHN defined as

$$p_n = \begin{cases} p_0, n = 0 \\ p_0 \gamma \phi^n r^{\frac{n(n-1)}{2}}, n = 1,2,3 \ldots \end{cases}, \gamma, \phi, r > 0 \quad (3.9)$$

where despite being known a priori, for the sake of completeness, p_0 can be expressed as

$$p_0 = \left(1 + \sum_{n=1}^{\infty} \gamma \phi^n r^{\frac{n(n-1)}{2}}\right)^{-1}. \quad (3.10)$$

Contrary to the case of the GGeo, the parameters of the GdHN distribution, γ, ϕ and r, cannot be derived explicitly in terms of the constraint information, however, they can be obtained numerically from the latter information.

3.1.1. Properties of the Generalised Discrete Half Normal

The GdHN is observed to reduce to the GGeo (2.18) or the dHN (2.36) discrete ME distributions by setting $r = 1$ or $\gamma = 1$ respectively. Furthermore, setting both $r = 1$ and $\gamma = 1$ yields the modified geometric discrete ME distribution (2.13). Thus, the GdHN is seen to be a generalisation of the GGeo, dHN and modified geometric discrete ME distributions.

When the aforementioned prior moment information constraints (3.1) - (3.3) are specified as infinite capacity queueing system attributes of the MQL, VQL

and empty queue state probability, p_0, then the resulting set of ME solutions, namely the GdHN ME QLD's, comprise a subset of the GdHN ME solutions.

The probability generating function (PGF), $P(z)$, first moment, $E[N]$ and variance, $E[(N - E[N])^2]$ of the GdHN can be defined as

$$P(z) = p_0 + \sum_{n=1}^{\infty} \gamma(\phi z)^n r^{\frac{n(n-1)}{2}} p_0, \tag{3.11}$$

$$E[N] = \sum_{n=1}^{\infty} n\gamma\phi^n \, r^{\frac{n(n-1)}{2}} p_0 \tag{3.12}$$

and

$$E[(N - E[N])^2] = \sum_{n=1}^{\infty} n^2\gamma\phi^n \, r^{\frac{n(n-1)}{2}} p_0 - \left(\sum_{n=1}^{\infty} n\gamma\phi^n \, r^{\frac{n(n-1)}{2}} p_0 \right)^2 \tag{3.13}$$

respectively, where in the latter three cases (3.11) - (3.13), p_0 is given by (3.10).

3.2. The Truncated Generalised Discrete Half Normal

The $GdHN_T$ is that unique discrete ME distribution constrained by the prior information of the first moment, variance and lower and upper boundary state probabilities, over finite, non-negative integer support. The corresponding ME optimisation prior information constraints, $E[N_K]$, $E[(N_K - E[N_K])^2]$, p_0^K, p_K^K and the normalising condition are defined as

$$\sum_{n=0}^{K} np_n^K = E[N_K], n = 0,1,2 \dots K, \tag{3.14}$$

$$\sum_{n=0}^{K} (n - E[N_K])^2 p_n^K = E[(N_K - E[N_K])^2], n = 0,1,2 \dots K, \tag{3.15}$$

$$\sum_{n=0}^{K} u'_K(n) p_n^K = p_0^K, \qquad u'_K(n) = \begin{cases} 1, n = 0 \\ 0, n = 1,2,3 \dots K' \end{cases} \tag{3.16}$$

$$\sum_{n=0}^{K} f_K(n) p_n^K = p_K^K, \qquad f_K(n) = \begin{cases} 0, n = 0,1,2 \dots K-1 \\ 1, n = K \end{cases} \tag{3.17}$$

and

$$\sum_{n=0}^{K} p_n = 1.0\,. \tag{3.18}$$

Incorporating the prior moment information constraints (3.14) - (3.17) and the normalising condition (3.18) directly in the general product form discrete ME solution given by

$$p_n = \frac{1}{Z}\prod_{i=1}^{m} x_{Ki}^{f_{Ki}(n)}, n = \ldots, -2, -1, 0, 1, 2 \ldots, i = 1, 2, 3 \ldots m \qquad (3.19)$$

where $Z = \sum_{n=-\infty}^{\infty}\left(\prod_{i=1}^{m} x_{Ki}^{f_{Ki}(n)}\right)$, yields the GdHN$_T$ expressed as

$$p_n = \frac{1}{Z} x_{K1}^{n} x_{K2}^{(n-E[N_K])^2} x_{K3}^{u'_K(n)} x_{K4}^{f_K(n)}, n = 0, 1, 2, \ldots K \qquad (3.20)$$

where $Z = \sum_{n=0}^{K}\left(x_{K1}^{n} x_{K2}^{(n-E[N_K])^2} x_{K3}^{u'_K(n)} x_{K4}^{f_K(n)}\right)$ and the x_{Ki}, $i = 1,2,3,4$ are the Lagrangian coefficients corresponding to the first moment, variance, p_0^K and p_K^K prior moment information constraints respectively. Equation (3.20) can be written equivalently as

$$p_n^K = \begin{cases} \frac{1}{Z} x_{K2}^{(E[N_K])^2} x_{K3}, & n = 0 \\ \frac{1}{Z} x_{K1}^n x_{K2}^{(n-E[N_K])^2}, & n = 1,2,3 \dots K-1 \\ \frac{1}{Z} x_{K1}^K x_{K2}^{(K-E[N_K])^2} x_{K4}, & n = K \end{cases}. \quad (3.21)$$

For the purpose of comparing the $GdHN_T$ to its subclass distributions, its distributional form expressed in (3.21) is manipulated in an analogous manner to the case of the GdHN. The aim is that the resulting distributional form concurrently resembles, as closely as possible, the forms of both the $GGeo_T$ (2.24) and the truncated dHN (dHN_T)[13] ME distributions. Expanding the product $(n - E[N_K])^2$ and introducing the term $(2E[N_K] - 1)n$ in the power of x_{K2} in (3.21) yields

$$p_n^K = \begin{cases} \frac{1}{Z} x_{K2}^{(E[N_K])^2} x_{K3}, & n = 0 \\ \frac{1}{Z} x_{K2}^{(E[N_K])^2} \left(x_{K1} x_{K2}^{-(2E[N_K]-1)} \right)^n x_{K2}^{(n^2-n)}, & n = 1,2,3 \dots K-1 \\ \frac{1}{Z} x_{K2}^{(E[N_K])^2} x_{K4} \left(x_{K1} x_{K2}^{-(2E[N_K]-1)} \right)^K x_{K2}^{(K^2-K)}, & n = K \end{cases} \quad (3.22)$$

where the term $\frac{1}{Z} x_{K2}^{(E[N_K])^2} x_{K3}$ is clearly the zero state probability, p_0^K. Without loss of generality, the parameters, γ_K, ϕ_K, r_K and ζ_K are introduced. Making

[13] i.e. the dHN right-truncated above K, whose distributional form is identical to that of the dHN defined in (2.36).

the substitutions, $(x_{K3})^{-1} = \gamma_K$, $\left(x_{K1}x_{K2}^{-(2E[N_K]-1)}\right) = \phi_K$, $x_{K2} = r_K^{\frac{1}{2}}$ and $x_{K4} = \zeta_K$ in (3.22), results in the re-parameterised GdHN$_T$ defined by

$$p_n^K = \begin{cases} p_0^K, & n = 0 \\ p_0^K \gamma_K (\phi_K)^n (r_K)^{\frac{n(n-1)}{2}}, & n = 1,2,3 \ldots K-1, \zeta_K, \gamma_K, \phi_K, r_K > 0 \\ p_0^K \zeta_K \gamma_K (\phi_K)^K (r_K)^{\frac{K(K-1)}{2}}, & n = K \end{cases} \quad (3.23)$$

where despite being known a priori, for the sake of completeness p_0^K is defined as

$$p_0^K = \left(1 + \sum_{n=1}^{K-1} \gamma_K (\phi_K)^n (r_K)^{\frac{n(n-1)}{2}} + \zeta_K \gamma_K (\phi_K)^K (r_K)^{\frac{K(K-1)}{2}}\right)^{-1}. \quad (3.24)$$

The parameters of the GdHN$_T$ distribution, γ_K, ϕ_K, r_K and ζ_K, cannot be expressed explicitly in terms of the prior moment constraints, however, they can be obtained numerically from the latter.

3.2.1. Properties of the Truncated Generalised Discrete Half Normal

The GdHN arises as a special case of the $GdHN_T$ when $K \to \infty$. The special case $r_K = 1$ retrieves the $GGeo_T$ (2.24) and setting both $\gamma_K = 1$ and $\zeta_K = 1$ in (3.23) yields the dHN_T. Furthermore setting r_K, γ_K and ζ_K to one in (3.23) results in the truncated modified geometric i.e. the modified geometric (2.13) right-truncated above K. Hence the $GdHN_T$ generalises the $GGeo_T$, dHN_T and the truncated modified geometric discrete ME distributions, as well as their corresponding infinite-support counterparts.

When the aforementioned prior information constraints (3.14) - (3.17) are specified as finite queueing system attributes of MQL, VQL and queue length boundary state probabilities, p_0^K and p_K^K, then the set of ME solutions derived, namely the $GdHN_T$ ME QLD's, comprise a subset of the $GdHN_T$ solutions.

The first moment, $E[N_K]$ and variance, $E[(N_K - E[N_K])^2]$ of the $GdHN_T$ can be defined as

$$E[N_K] = \left(\sum_{n=1}^{K-1} n\gamma_K (\phi_K)^n (r_K)^{\frac{n(n-1)}{2}} p_0^K \right) + K\zeta_K \gamma_K (\phi_K)^K (r_K)^{\frac{K(K-1)}{2}} p_0^K \qquad (3.25)$$

and

$$E[(N_K - E[N_K])^2] = \left(\sum_{n=1}^{K-1} n^2 \gamma_K (\phi_K)^n (r_K)^{\frac{n(n-1)}{2}} p_0^K\right)$$

$$+K^2 \zeta_K \gamma_K (\phi_K)^K (r_K)^{\frac{K(K-1)}{2}} p_0^K \qquad (3.26)$$

$$-\left(\left(\sum_{n=1}^{K-1} n\gamma_K (\phi_K)^n (r_K)^{\frac{n(n-1)}{2}} p_0^K\right) + K\zeta_K \gamma_K (\phi_K)^K (r_K)^{\frac{K(K-1)}{2}} p_0^K\right)^2$$

respectively, where in the above two cases (3.25) - (3.26), p_0^K is given by (3.24).

3.3. The Effect of Different Prior Moment Information

The effect of knowledge of additional or different prior moment information on the profiles of resulting ME distribution inferences is observed in Fig. 4 and Fig 5 below. Specifically, the $GdHN_T$, dHN_T, $GGeo_T$ and truncated modified geometric (Trunc Mod Geom) ME distribution inferences are generated from appropriate prior moment information, which in turn is obtained from an arbitrary distribution, $xi, i = 1,2$.

Fig. 4. Effect of different prior moment information on the profiles of ME distribution estimates of an arbitrary distribution, $x1$.

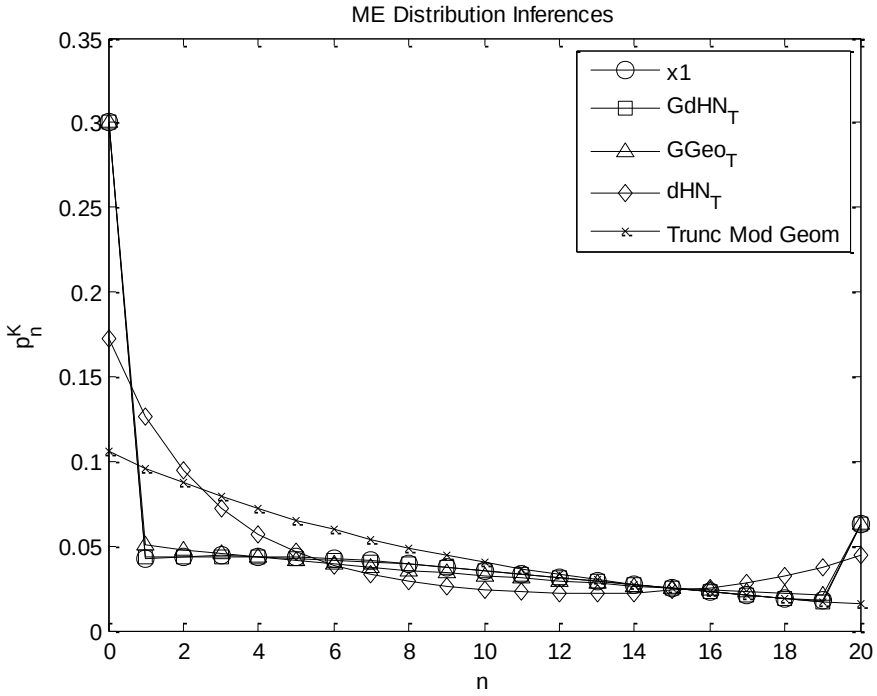
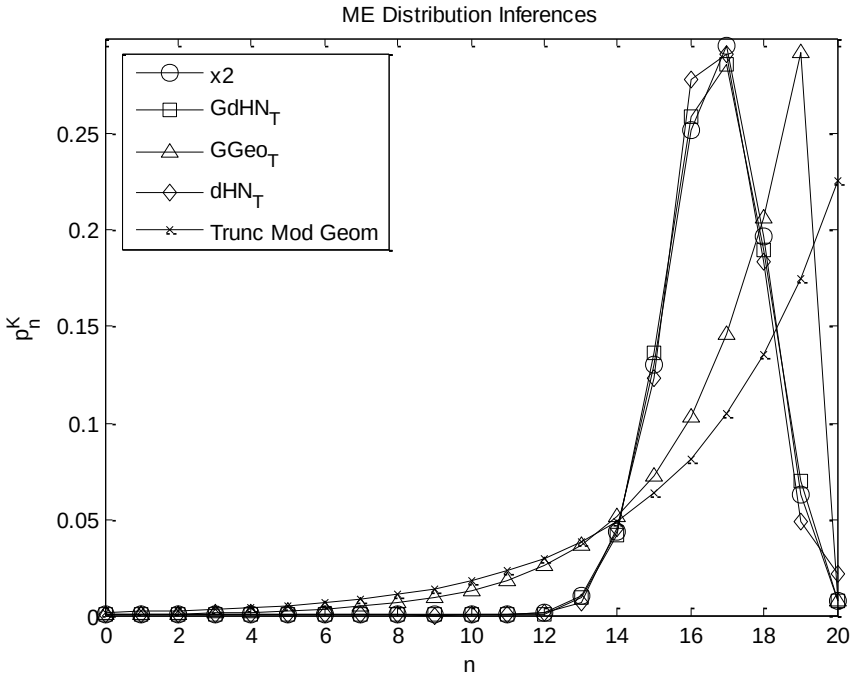


Fig 5. Effect of different prior moment information on the profiles of ME distribution estimates of an arbitrary distribution, $x2$.



The prior moments assumed known, in addition to the normalising condition, in the generation of the different ME distribution inferences in Fig. 4 and Fig 5 above are listed in Table 1 below.

Table 1. Prior moment information associated with corresponding ME distribution inferences.

	$E[N_K]$	$E[(N_K - E[N_K])^2]$	p_0^K	p_K^K
$GdHN_T$	✓	✓	✓	✓
dHN_T	✓	✓	-	-
$GGeo_T$	✓	-	✓	✓
Trunc Mod Geom	✓	-	-	-

4. The M/GE/1/K Queue Subject to Balking

In this chapter, the QLD of the M/GE/1/K queue subject to extended Morse balking is derived and subsequently conjectured to be a special case of the $GdHN_T$ ME distribution. Following this, the aforementioned queueing system is applied as an ME performance model of IP network nodes featuring static or dynamic packet dropping congestion management schemes. In the latter context, a performance evaluation study in terms of the model's delay is carried out.

Employing the celebrated Pollaczek-Khinchin (P-K) transform formula, the service-time distribution of the M/G/1 queue bearing the GGeo ME QLD was discovered to be satisfied exactly by the GE CDF (El-Affendi and Kouvatsos 1983). The P-K transform formula encapsulates a relationship between the QLD and service-time distribution of an M/G/1 queue. The irrational PGF of the GdHN (3.11) renders the latter approach intractable to determine the unique service time distribution of the ordinary M/G/1 queue with GdHN ME QLD and thus solve its QLD in terms of basic queueing system parameters.

By analogy with existing results, it is expected that the GdHN or $GdHN_T$ ME QLD's, which are generalisations of the GGeo and $GGeo_T$ ME QLD's, would characterise, respectively, infinite or finite capacity ordinary M/G/1 and G/G/1 queues, where 'G' is satisfied exactly by a generalisation of the GE. Due to

the dependence of the VQL of an M/G/1 queue upon the first three moments of the service time distribution, it is proposed that the latter generalisation of the GE would be completely defined by its first three moments. Nonetheless, despite the analytic intractability posed by this latter problem, least biased QLD's can be inferred numerically by maximising Shannon's entropy subject to prior queue length moment values and these in turn can be utilised for performance prediction.

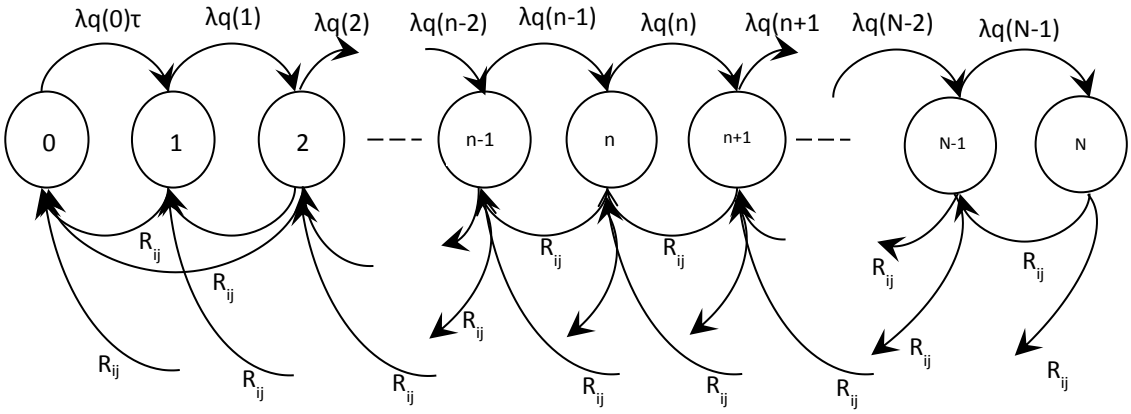
As stated in Chapter 3, two subclass distributions of the $GdHN_T$ are the $GGeo_T$ and dHN_T (and hence the dHN). These have been characterised as QLD's of queueing systems marked respectively by GE service times (Kouvatsos 1986b) and Morse balking (Morse 1958; Kemp 2005; Kemp 2008). In light of this, the queueing system bearing the $GdHN_T$ ME QLD was sought by employing these two operational characteristics in a single queueing system. Hence the M/GE/1/K queue subject to extended Morse balking is analysed below.

4.1. The QLD of the M/GE/1/K Queue Subject to Extended Morse Balking

In this section, the stationary QLD (from a random observer's point of view) of the M/GE/1/K queue subject to extended Morse balking, characterised by a Poisson prospective arrival process (with mean rate, λ), i.i.d. GE service times (with mean rate, μ and SCOV, C_s^2) and finite capacity, K, is derived.

This is carried out by applying the GB principle at each individual state of the Markov chain model of the equivalent M(n)/GE/1/K queue illustrated below.

Fig. 6. Markov Chain Model of the M(n)/GE/1/K Queue.



In **Fig. 6**, the R_{ij}'s are the state transition rates and the effective arrival rates at each state account for the fraction of customers which balk according to the relationship (2.34). The upward state-transition rates are derived as

$$R_{01} = \binom{prospective}{arrival\ rate} \times P\begin{pmatrix} arrival\ joins\ the\ queue\ of\ length\ 0\ and \\ takes\ the\ exponential\ branch\ of\ the \\ two-phase\ GE\ service\ model \end{pmatrix} \qquad (4.1)$$

$$= \lambda\tau q(0)$$

where $q(n)$ is the joining probability of an arriving customer, conditional on the instantaneous queue length, n and τ, the probability of taking the exponential branch in the two-phase GE interpretation (cf., Fig. 2), is given by (Kouvatsos 1994)

$$\tau = \frac{2}{1 + C_s^2} \tag{4.2}$$

and

$$R_{i,i+1} = \binom{prospective}{arrival\ rate} \times P\binom{arrival\ joins\ the\ queue}{of\ length\ i}, i = 1,2,3 \dots K-1, \\ = \lambda q(i) \tag{4.3}$$

and downward state transition rates are derived as

$$R_{ij} = \binom{batch\ service}{completion\ rate} \times P\begin{pmatrix} batch\ which\ completes\ service \\ is\ of\ size\ (i-j)\ and \\ queue\ does\ not\ empty \end{pmatrix}, \\ i = 2,3,4 \dots K, j = 1,2,3 \dots i-1 \\ = (\tau\mu)\tau(1-\tau)^{i-j-1} \tag{4.4}$$

and

$$R_{i0} = \binom{batch\ service}{completion\ rate} \times P\begin{pmatrix} batch\ which\ completes\ service \\ is\ of\ size\ i\ and \\ queue\ empties \end{pmatrix}, \\ i = 1,2,3 \dots K \tag{4.5}$$

$$= (\tau\mu) \times \left(1 - \sum_{k=1}^{i-1} \tau(1-\tau)^{i-k-1} \right)$$

$$= (\tau\mu)(1-\tau)^{i-1}$$

respectively. $R_{ij} = 0$ otherwise. Applying the GB principle to each individual state in the queue's Markov chain yields the following system of $(K+1)$ linear GB equations

$$\sum_{\substack{i=0, \\ i \neq n}}^{K} R_{in}\, p_i^K = p_n^K \sum_{\substack{j=0, \\ j \neq n}}^{K} R_{nj}\,, n = 0,1,2 \ldots K, \tag{4.6}$$

one of which is redundant. Replacing any one of these equations with the normalising condition and solving the system yields the following QLD of the M(n)/GE/1/K queue marked by population-dependent arrival rates, $\lambda q(n), n = 0,1,2 \ldots K-1$,

$$p_n^K = \begin{cases} p_0^K, n = 0 \\ p_0^K \tau \prod_{i=0}^{n-1} \left(\dfrac{\lambda q(i)}{\lambda q(i+1)(1-\tau) + \tau\mu} \right), n = 1,2,3 \ldots K-1 \\ p_0^K \tau \left(\dfrac{\lambda q(K-1)}{\tau\mu} \right) \prod_{i=0}^{K-2} \left(\dfrac{\lambda q(i)}{\lambda q(i+1)(1-\tau) + \tau\mu} \right), n = K \end{cases} \tag{4.7}$$

where λ, $q(n)$, μ and τ are as interpreted above and p_0^K can be obtained by applying the normalising condition.

For the purpose of estimating the instantaneous workload, t in the Morse joining function (2.28), it is assumed that prospective arrivals become aware of the instantaneous queue length, n, μ and C_s^2 either before or on arrival. When the service durations are i.i.d., satisfying any general distribution (including the GE CDF), the conditional average instantaneous workload is given by

$$t_{av} = \begin{cases} 0.0, n = 0 \\ \dfrac{(n-1)}{\mu} + \dfrac{1 + C_s^2}{2\mu}, n = 1,2,3 \ldots \end{cases}. \tag{4.8}$$

Substituting t in the Morse joining function (2.28) by t_{av} (4.8) yields the extended Morse population-dependent joining function given by

$$q(n) = \begin{cases} 1.0, n = 0 \\ e^{-\frac{\alpha n}{\mu}} e^{-\frac{\alpha}{\mu}\left(\frac{C_s^2 - 1}{2}\right)}, n = 1,2,3 \ldots \end{cases} \tag{4.9}$$

where α is the Morse average measure of impatience in the context of i.i.d. general service times. Making the substitution $q = e^{-\alpha/\mu}$ in (4.9) and restricting the support to below K, the extended Morse joining function becomes

$$q(n) = \begin{cases} 1.0, n = 0 \\ q^{\left(\frac{C_s^2-1}{2}\right)} q^n, n = 1,2,3 \dots K-1 \\ 0.0, n = K \end{cases} \cdot \quad (4.10)$$

Substitution of $q(n)$ in the QLD (4.7) by the extended Morse joining function defined in (4.10) above yields the following closed form QLD of the M/GE/1/K queue subject to extended Morse balking

$$p_n^K = \begin{cases} p_0^K, n = 0 \\ p_0^K \dfrac{\left(\tau q^{-\left(\frac{C_s^2-1}{2}\right)}\right)\left(\lambda q^{\left(\frac{C_s^2-1}{2}\right)}\right)^n q^{\frac{n(n-1)}{2}}}{\prod_{i=0}^{n-1}\left(\lambda q^{\left(\frac{C_s^2-1}{2}\right)} q^{i+1}(1-\tau)+\tau\mu\right)}, n = 1,2,3 \dots K-1 \\ p_0^K \dfrac{\left(\tau q^{-\left(\frac{C_s^2-1}{2}\right)}\right)\left(\lambda q^{\left(\frac{C_s^2-1}{2}\right)}\right)^K q^{\frac{K(K-1)}{2}}}{\tau\mu\prod_{i=0}^{K-2}\left(\lambda q^{\left(\frac{C_s^2-1}{2}\right)} q^{i+1}(1-\tau)+\tau\mu\right)}, n = K \end{cases} \quad (4.11)$$

where $q = exp(-\alpha/\mu)$, λ, α, μ and τ are as interpreted above and p_0^K can be obtained via the normalising condition.

In this case the parameter α can be estimated in a similar manner to that for the exponential service case described in Section 2.3.1. Assuming that the

overall loss (i.e. total balking and blocking) rate, LR, λ, μ and C_s^2 are known, solving

$$\lambda \sum_{n=0}^{K-1} q(n) p_n^K - (\lambda - LR) = 0 \tag{4.12}$$

for α yields the desired estimate, where $q(n)$ is now given by (4.10) and p_n^K is defined by (4.11). Analogously, other appropriate queueing metrics such as those listed previously in Section 2.3.1 may be used to estimate the value of α. A mathematical definition of the parameter α in the context of i.i.d. general service times is derived in Appendix B.

4.2. Discussion of the Results

The aim of the latter analysis was to determine a queueing system(s) bearing the $GdHN_T$ ME QLD, for the purpose of ME performance modelling and prediction.

The expression for the QLD of the M/GE/1/K queue subject to extended Morse balking (4.11) though bearing some resemblance in form to that of the $GdHN_T$ ME distribution (3.23) cannot be judged solely by observation to be its special case (in contrast to the special case of the QLD of the M/M/1 queue subject to Morse balking and the dHN ME distribution (cf., Section 2.3.3)). Furthermore, limitations were encountered in attempts at an analytic proof of equivalence in distribution between the latter two solutions, i.e. the

GB and ME solutions. Hence validation of this equivalence was sought through other means.

Equivalence in distribution between the QLD of the M/GE/1/K queue subject to extended Morse balking (4.11) and the corresponding $GdHN_T$ ME distribution inference (3.23) was therefore investigated numerically. The QLD's were first generated using input parameter values and from these the corresponding ME solutions were inferred. Subsequently, absolute differences between corresponding state probabilities of the two solutions were obtained in order to investigate the statistical closeness between them. Absolute differences (and/or related measures) have been used in a similar vein in works such as (Chandy et al. 1975; Kouvatsos and Awan 2003). Details of the investigation are as follows.

As part of the experimentation, numerical QLD probabilities of the M/GE/1/K queue subject to extended Morse balking, $p_n^{K\star}, n = 0,1,2,\dots,K$, were generated from equation (4.11) and different combinations of queue input parameter values from **Table 2** below.

Table 2. Parameter values used in the investigation of ME characteristics of the QLD of the M/GE/1/K queue subject to extended Morse balking.

Parameter	Value(s)
λ	[10 20 30 40]
μ	20
C_s^2	[1 5 10 20 50 100 200 500]
K	[5 10 15 20]
q	[0.3 0.6 0.9]

An 'experiment' here refers to the generation of a single set of numerical QLD probabilities, $p_n^{K\star}, n = 0,1,2,\dots,K$, of the M/GE/1/K queue subject to extended Morse balking, from (4.11) and a particular combination of queue input parameter values. It also includes the subsequent computation of the corresponding numerical GdHN$_T$ ME distribution inference, $p_n^{K\dagger}, n = 0,1,2,\dots,K$. By the fundamental principle of counting, there are a total of 384 combinations of queue input parameter values and hence 384 experiments were conducted.

From the numerical QLD's, $p_n^{K\star}, n = 0,1,2,\dots,K$, values of the MQL, VQL, $p_0^{K\star}$ and $p_K^{K\star}$ were either calculated or obtained directly. The latter values comprised the optimisation constraints, in addition to the normalisation condition, in the numerical constrained maximisation of Shannon's entropy functional (2.1), yielding corresponding numerical GdHN$_T$ distribution inference probabilities, $p_n^{K\dagger}, n = 0,1,2,\dots,K$. Finally, for each experiment, the maximum absolute difference between corresponding state probabilities of

the two distributions, referred to here as the 'error', was computed. Error is defined as

$$error = max\left(\left|p_0^{K\star} - p_0^{K\dagger}\right|, \left|p_1^{K\star} - p_1^{K\dagger}\right|, \ldots, \left|p_K^{K\star} - p_K^{K\dagger}\right|\right). \quad (4.13)$$

The experiments were carried out in MATLAB version 7.10.0.499 (R2010a). Both the constraints and change in objective function were satisfied to within the default tolerance of 10^{-6}. Owing to the limitations of the software to produce the GdHN$_T$ inferences when the prior moment constraint, $p_K^{K\star} < 10^{-10}$, a special case of the GdHN$_T$ ME inference was used in those instances. This special case excluded the p_K^K prior information constraint (but not the existence of the probability p_K^K itself) or equivalently set the parameter ζ_K in (3.23) to one.

Over all the 384 experiments, the maximum error encountered was 0.014 with the overwhelming majority of errors being less than 0.01.

Exact equivalence in distribution between the GB and ME solutions is conjectured based on the following two arguments:

1. Errors of comparable magnitude were encountered between corresponding special cases which are known to be equivalent.

   Errors were computed for the following two cases whose exact equivalence (i.e. equivalence between the QLD's and (special cases of) the GdHN$_T$) has been proven in the literature: the M/M/1/K queue subject

to Morse balking (Shah et al. 2010) and the ordinary M/GE/1/K queue (Kouvatsos 1986b; Kouvatsos 1986a). Over all the experiments conducted for the two queueing systems, maximum errors of 0.0114 and 0.000688, respectively, were observed implying that these latter errors and therefore the former errors too can be attributed to numerical limitations of the software package. Larger errors in the case of the M/GE/1/K queue subject to extended Morse balking and the $GdHN_T$ ME distribution inference may be attributed to the effect of MATLAB's computational approximations on a larger set of optimisation constraints.

2. Errors did not increase with increasing C_s^2.

   Earlier ME approximate analysis of ordinary queueing systems featuring i.i.d. GE inter-arrival and/or service times exhibited growing absolute differences between the ME approximations and simulation results as C_a^2 and/or C_s^2 increased (cf., (Kouvatsos and Awan 2003)). Such behaviour was not encountered in this research work but instead errors remained low over all experiments.

Supported by the above experimental evidence and subsequent reasoning, the following conjecture is proposed.

Conjecture I: *The QLD of the M/GE/1/K queue subject to extended Morse balking (4.11) is a special case of the $GdHN_T$ ME distribution (3.23) constrained by the prior information of the queue's MQL, VQL (or second moment of queue length), empty state probability, p_0^K (or equivalently server utilisation), full buffer state probability, p_K^K and the normalisation condition over finite, non-negative integer support* $[0, K]$.

4.2.1. The Infinite-Capacity Special Case

Setting $K \to \infty$ retrieves the QLD of the infinite – capacity M/GE/1 queue subject to extended Morse balking defined by

$$p_n = \begin{cases} p_0, & n = 0 \\ p_0 \dfrac{\left(\tau q^{-\left(\frac{C_s^2-1}{2}\right)}\right)\left(\lambda q^{\left(\frac{C_s^2-1}{2}\right)}\right)^n q^{\frac{n(n-1)}{2}}}{\prod_{i=0}^{n-1}\left(\lambda q^{\left(\frac{C_s^2-1}{2}\right)} q^{i+1}(1-\tau)+\tau\mu\right)}, & n = 1,2,3 \ldots \end{cases}, \tag{4.14}$$

which is analogously conjectured to be a special case of the GdHN (3.9).

4.2.2. The Exponential Service Special Case

For exponential service, $C_s^2 = 1.0$, which implies $\tau = 1.0$ from (4.2). Applying this condition to the QLD (4.11) reduces it to that of the M/M/1/K queue subject to Morse balking, i.e. the dHN$_T$ ME QLD (Shah et al. 2010). Setting $C_s^2 = 1.0$ in (4.14) results in the QLD of the M/M/1 queue subject to Morse balking, i.e. the dHN ME QLD (2.32).

4.2.3. The Non-Balking Special Case

Setting $q(n) = 1, \forall n$ implies no balking and the QLD (4.11) reduces to that of the ordinary M/GE/1/K queue, the $GGeo_T$ (2.24). Similarly, applying the condition, $q(n) = 1, \forall n$ in (4.14) yields the GGeo ME QLD (2.19).

4.3. Case Study: The Evaluation of an ME Performance Model of Congestion Management in Communication Networks

In this section, the M/GE/1/K queue subject to extended Morse balking is exploited as an ME performance model of IP-based network nodes featuring static or dynamic packet dropping congestion management mechanisms.

Queues subject to arrival balking are naturally seen to model admission/dropping policies at service centres including nodes (i.e. routers) in communication networks (Liu 2007; Boxma and Prabhu 2009). In those models, the balking operation is re-interpreted as the node's packet dropping mechanism whereby the scheduler/gateway decides, depending on the level of congestion, whether or not to drop a packet arrival. On the other hand, the customer joining policy is re-interpreted as the node admission scheme.

Congestion is the state of a network whose input traffic is greater than its capacity. It is a network scenario characterised by heavily populated queues and long delays.

Packet dropping (also referred to as active queue management (AQM)) is the most common method by which network congestion management is conducted. The two approaches of congestion management are congestion control and recovery (reactive) and congestion avoidance (pro-active). Rather than control congestion after its onset, congestion avoidance aims to circumvent this eventuality altogether by taking proactive steps to detect and combat congestion early. It aims to keep packet-transfer delay low by maintaining queue lengths at suitable levels while enabling sufficient throughput to traverse the network i.e. it aims to achieve a desirable delay – throughput trade-off (Labrador and Banerjee 1999).

In this thesis, packet dropping policies (PDP's) in the IP networking context are studied.

Based on the re-interpretation of the balking operation as packet dropping in communication network nodes, the extended Morse balking function is re-interpreted as a PDP, namely the extended Morse PDP. The extended Morse PDP can be derived from the extended Morse joining function (4.10) and is defined as

$$1-q(n)=\begin{cases}0.0, n=0\\ 1-q^{\left(\frac{C_s^2-1}{2}\right)}q^n, n=1,2,3\ldots K-1\\ 1.0, n=K\end{cases} \quad (4.15)$$

where n is the instantaneous node queue length, C_s^2 is the SCOV of node service durations, K is the node capacity and $q, 0 < q < 1$ is considered to be a performance tuning parameter set by the node scheduler to achieve desired delay-throughput trade-offs.

As in the case of the joining probabilities defined by (4.10), packet dropping in the extended Morse PDP model (4.15) depends on the conditional average instantaneous workload of the queue in the following way: The node scheduler drops packets, on arrival, with probability, $\big(1 - q(n)\big)$, defined by the extended Morse packet dropping function (4.15) dependent on n, C_s^2 and q. Assuming that values of the average prospective arrival rate to the node, λ, the mean node service rate, μ and C_s^2 can be obtained, desired delay-throughput trade-offs can be achieved by appropriately specifying q.

The extended Morse PDP is proposed as a model of the class of instantaneous, random early drop PDP's. The properties, strengths and limitations of this class of PDP's are discussed below.

The dropping of packets is used as a congestion avoidance technique where (pure) packet marking is unsupported such as in the TCP transmission protocol. In TCP, transmission sources infer existence of incipient congestion via the round-trip timeout mechanism and consequently reduce their transmission window size or transmission rate (as appropriate). In pure packet marking on the other hand, the source is notified to reduce the window for that connection by setting a bit in a packet header. Clearly, network resources are wasted in the packet dropping approach as opposed to packet marking however early dropping (a feature of the extended Morse

PDP), in the first instance, avoids the potentially greater packet loss that would otherwise be experienced under the standard droptail (DT) policy. In the DT policy, all packets are dropped after the queue size reaches a set threshold value. This causes synchronised reduction in sending rates among concurrent flows which in turn results in periods of inefficient link utilisation, a second effect avoided by early dropping. Thus early dropping addresses network congestion preventatively rather than reactively (Floyd and Jacobson 1993).

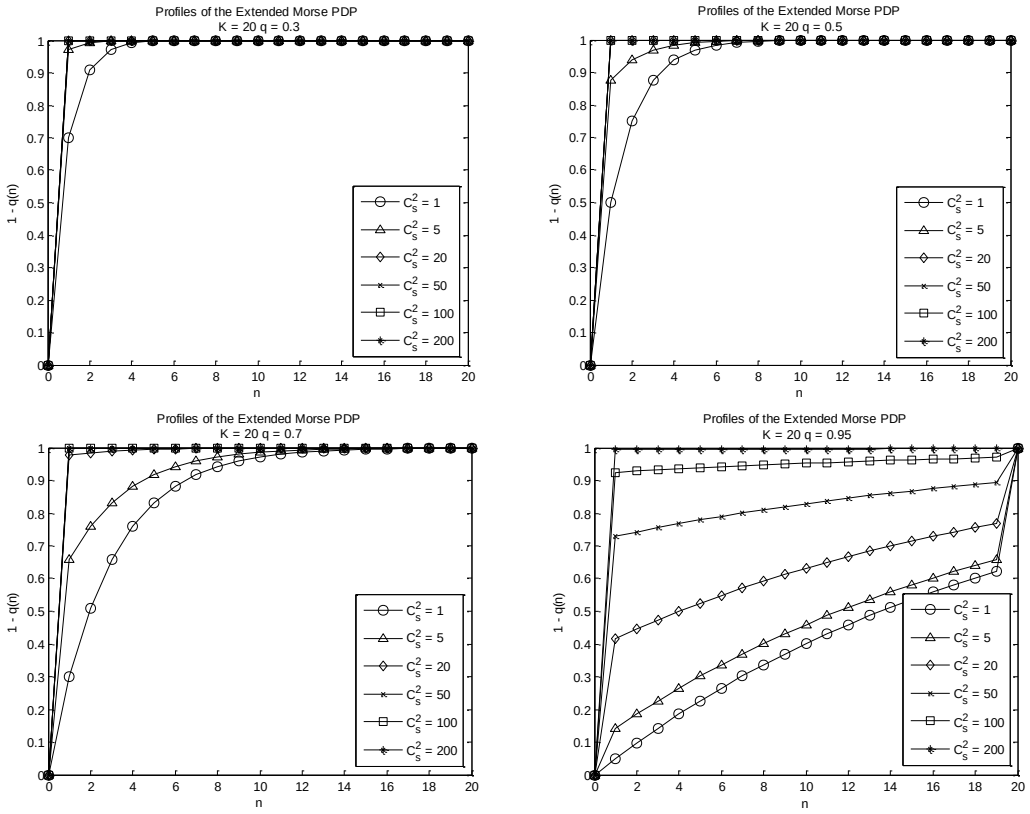
Random dropping, as occurs in the extended Morse PDP, facilitates a fairer admission policy by dropping more packets the heavier the user, thus inadvertently targeting senders proportionally to their bandwidth usage. A positive side-effect of such inadvertent targeted dropping is the minimisation of node under-utilisation.

Moreover, instantaneous queue length policies, such as the extended Morse PDP, avoid the overhead due to processing the average queue lengths for example the exponential weighted moving average (EWMA) used in some congestion avoidance mechanisms. The EWMA is required to be updated after each arrival in the Random Early Detection (RED) congestion avoidance mechanism (Floyd and Jacobson 1993). It is for this reason also that the instantaneous DT mechanism has been favoured as a PDP (Iannaccone et al. 2001). On the other hand, instantaneous policies exhibit bias against bursty or temporal heavy traffic which motivated proposals of RED and related algorithms (Floyd and Jacobson 1993).

PDP's with gentle slopes such as Gentle Random Early Detection - Instantaneous (GRED-I) (Iannaccone et al. 2001) and the extended Morse PDP (cf., Fig. 7 and Fig. 8) have advantages in terms of consecutive losses when contrasted with profiles that have large jumps such as DT and RED. It is highly likely that the jump in the dropping profile results in packets being dropped from many different connections consecutively, causing the undesirable effect of network resource under-utilisation due to synchronisation. Spreading the losses yields better performance as the latter effect is avoided.

The following graphs (Fig. 7) illustrate profiles of the extended Morse packet dropping function for increasing n, C_s^2 and q in the context of the M/G/1/K queue subject to packet dropping under the extended Morse PDP.

Fig. 7. Profiles of the extended Morse packet dropping function in the context of the M/G/1/K queue subject to packet dropping under the extended Morse PDP.



The above profiles immediately illustrate that under the extended Morse PDP, the probability with which prospective arrivals are dropped increases with the instantaneous queue length and/or values of C_s^2 and decreases with increasing q. The variation of dropping probability with C_s^2 corresponds to the change in average delay (i.e. residence time) in the ordinary M/G/1/K queue, which in general increases with increasing C_s^2, as observed in the case of GE service times. Therefore appropriately increasing packet drop probabilities with increasing C_s^2 in the M/G/1/K queue subject to packet dropping would have the desired effect of restricting the average delays to required levels.

The graphs also show that under the extended Morse PDP, arrivals to an empty queue are always admitted.

In addition to the properties described above, the extended Morse PDP (4.15) satisfies other characteristics of practical PDP's including those identified in (Labrador and Banerjee 1999), thus further establishing its usefulness as a PDP model. The properties are summarised below collectively:

- Monotonically increasing (i.e. different combinations of strictly increasing and non-decreasing).

- Fairness – the extended Morse PDP is a fair policy which is unbiased towards particular connections due to its probabilistic nature however it is less accommodating of traffic bursts or temporal heavy traffic due to its use of instantaneous instead of moving average queue length.

- Simplicity – the dropping probability in the extended Morse PDP requires less computation than the moving average queue length PDP's and hence processing overheads are lower and its operational speed greater at the expense of bias against temporal heavy traffic.

- Flexibility – the extended Morse PDP can be set up to behave in either a static or dynamic manner, the former being simpler but providing poorer overall performance results. In the static context, a fixed criterion, for example anticipated hourly traffic loads from previous network measurements can be used to set the performance tuning parameter, q in (4.15) to achieve an anticipated delay – throughput trade-off. On the

other hand q can be set dynamically depending on changes in λ, μ and/or C_s^2 to achieve a desired trade-off in real-time.

- Global Synchronisation – the extended Morse PDP counters global synchronisation through connection-independent probabilistic dropping. In addition, due to its gentle dropping profile (cf., (4.15), Fig. 7 and Fig. 8), the global synchronisation associated with jumps present in some PDP's is avoided.

- Scalability – the absence of per-connection state information in the extended Morse PDP renders it superior, with respect to scalability, to PDP's which store such information. It is not limited by the consequential overheads or performance degradation when implemented across expanding networks.

Thus correspondences have been identified between the extended Morse PDP and instantaneous, random early drop congestion avoidance mechanisms. And other connections have been drawn between properties of the extended Morse PDP and those of practical PDP's. However, a limitation of the extended Morse PDP as a practical PDP is the absence of a lower threshold and consequently packet dropping from queue occupancy of one. This would raise false alarms of incipient congestion resulting in reduced throughput through nodes than would otherwise be obtained. Nevertheless, adjusting q enables average dropping rates and consequently average queue performance levels to be obtained that are comparable to the case when PDP's comprising lower thresholds are used (as demonstrated below).

Hence the M/GE/1/K queue subject to extended Morse balking is submitted as a suitable ME performance model of network nodes featuring packet dropping of the instantaneous, random early drop type.

Profiles resulting from the M/GE/1/K queue subject to different PDP's are compared in Fig. 8 below. Three instantaneous, random, early drop – type PDP's and the DT policy are compared under common mean loss (i.e. total dropping and blocking) rate: GRED-I, Early Random Drop (ERD) (Floyd and Jacobson 1993), extended Morse and DT PDP's. Common values of λ, μ, C_s^2 and K were used while each dropping function's(') parameter(s) was(were) sought by minimising the absolute difference between the loss rate experienced under GRED-I and that under each of the remaining PDP's. The queue capacity, $K = 20$ was chosen based on proposals for optimal buffer requirements in high speed routers (Wischik 2005).

The GRED-I PDP parameter values given by

$$Lower\ Threshold, T_L = \|0.15K\|, \tag{4.16}$$

$$Upper\ Threshold, T_U = \|0.65K\|, \tag{4.17}$$

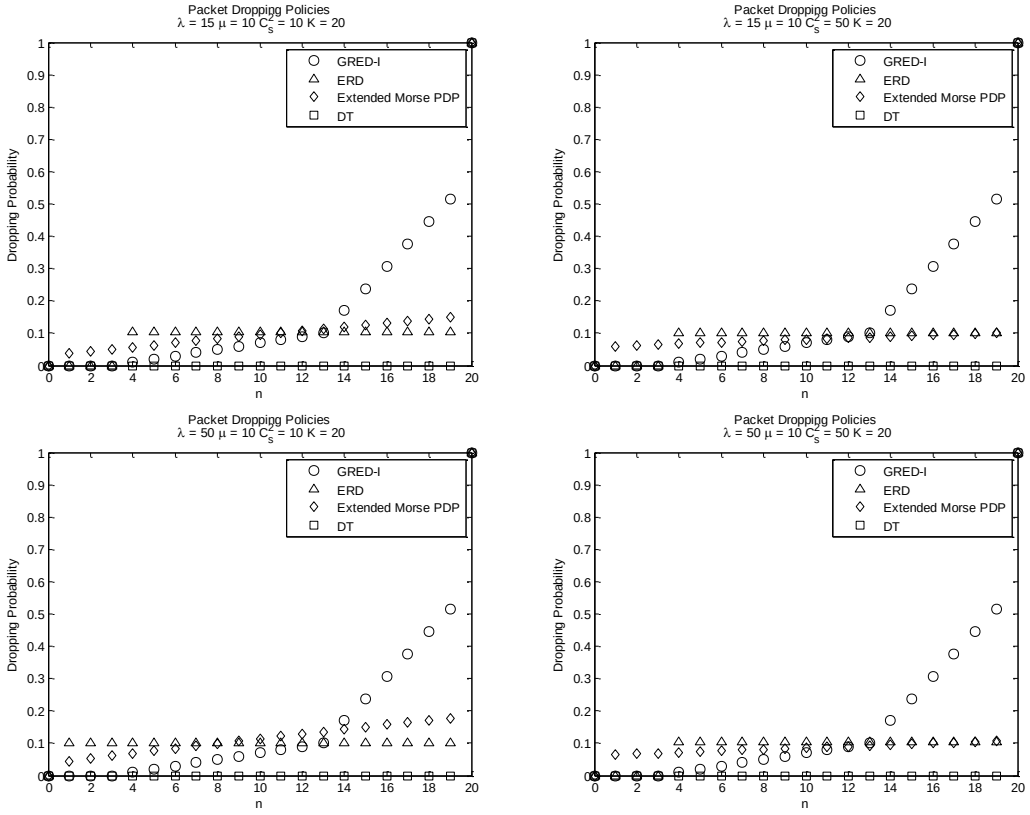
$$Gradient\ 1 = \frac{0.1}{T_U - T_L} \tag{4.18}$$

and

$$Gradient\ 2 = \frac{0.9}{T_U} \tag{4.19}$$

were derived from (Iannaccone et al. 2001).

Fig. 8. Comparison of dropping sequence and function profiles resulting from the M/GE/1/K queue subject to four different PDP's under common loss (i.e. total dropping and blocking) rate.

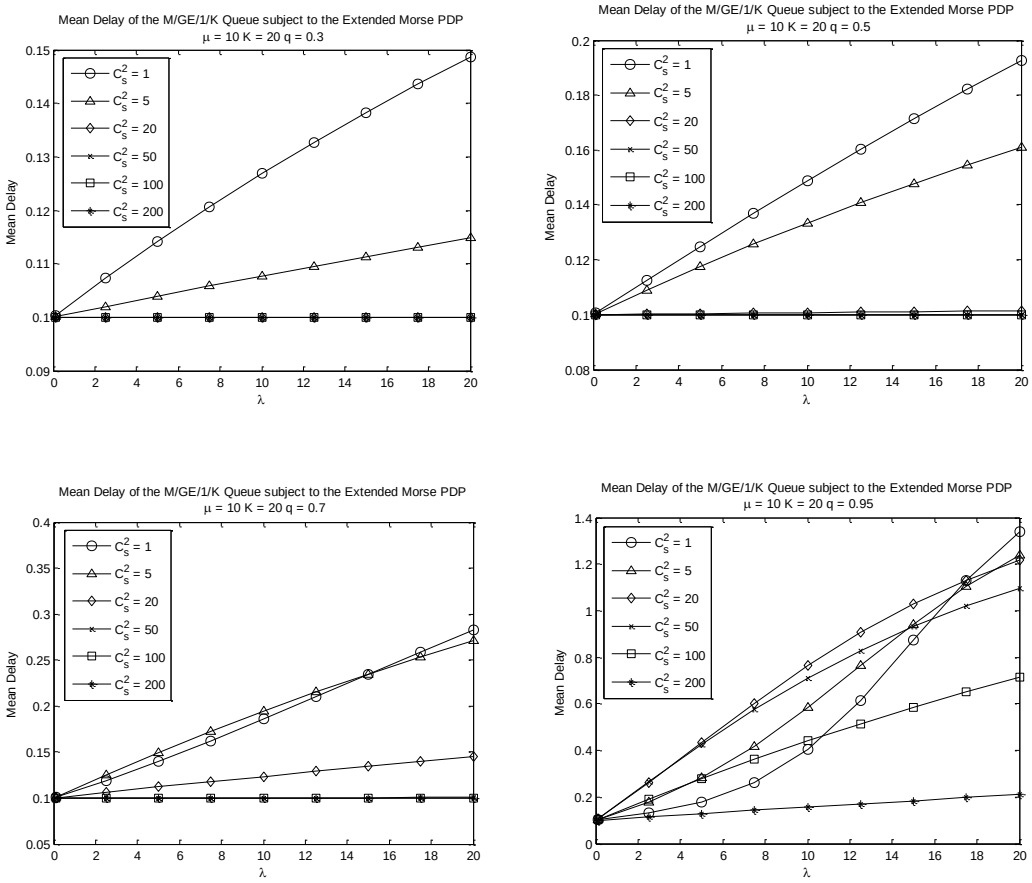


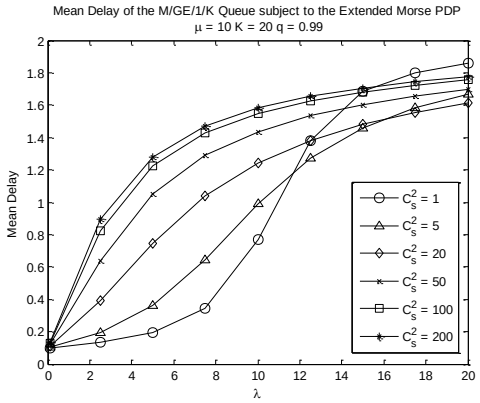
Below, a performance evaluation study is carried out on the M/GE/1/K queue subject to the extended Morse PDP. To this end, three sets of experiments, namely Experiment Set I – III, (illustrated in Fig. 9 - Fig. 11) are conducted to study the variation of mean delay with λ and throughput, $\lambda E[q(n)]$ under common μ and K and increasing values of λ, $\lambda E[q(n)]$, C_s^2 and/or q.

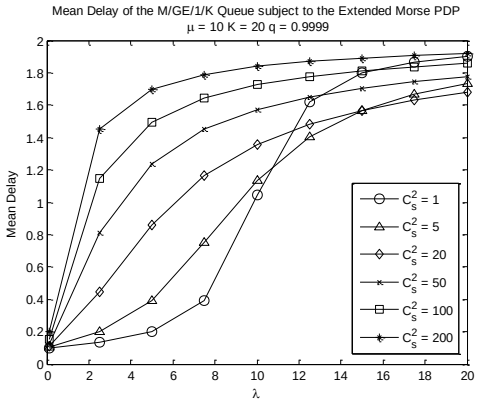
In Experiment Set I (**Fig. 9**), mean delay is observed as λ is increased for different values of C_s^2 and fixed q. Each experiment is repeated for increasing values of q. In the second set, Experiment Set II (**Fig. 10**), mean

delay is observed for different values of C_s^2, as $\lambda E[q(n)]$ is increased. $\lambda E[q(n)]$ is increased by increasing λ while q is fixed. As in the first set, each experiment is repeated for increasing values of q. Finally, Experiment Set III (**Fig. 11**) investigates the variation in mean delay, for different values of C_s^2 and increasing $\lambda E[q(n)]$ by increasing q while maintaining consistent λ to the queue. Each experiment in the last set is repeated for increasing values of λ.

Fig. 9. Mean delay of the M/GE/1/K queue subject to the extended Morse PDP plotted against increasing $\boldsymbol{\lambda}$ for fixed $\boldsymbol{q}$ (Experiment Set I).







As is expected, for fixed C_s^2 and q, the mean delay is larger for increasing λ ($\lambda \geq 0.1$) as the mean throughput increases. Secondly, the relatively larger delay for each successive experiment in the set is attributable to the larger q which corresponds to higher packet admission probabilities and consequently greater throughput.

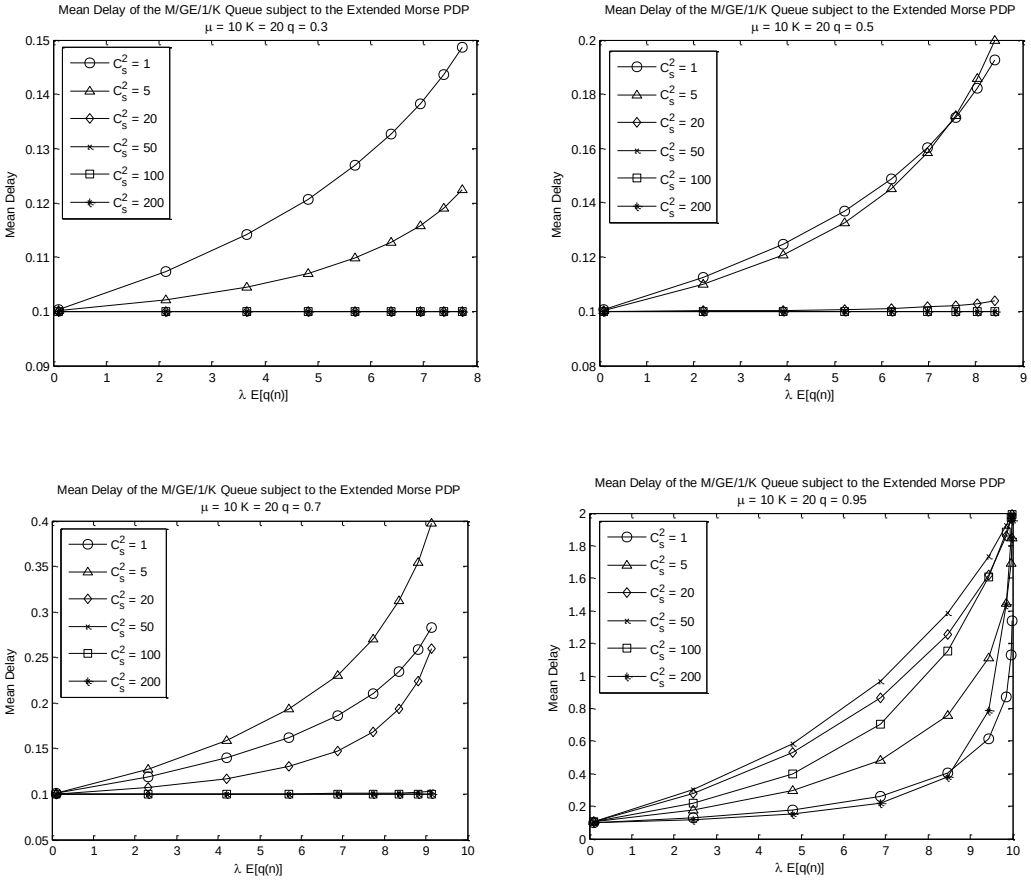
A third observation especially evident in the first and second experiments of Set I, namely the reduction in mean delay for increasing C_s^2, is in stark contrast to that for ordinary queues characterised by i.i.d. GE inter-arrival and/or service times. In the ordinary GE queues, it was observed for the most part that increases in C^2 resulted in degraded performance (Kouvatsos and Awan 2003). This different behaviour exhibited in the current model is attributed to greater dropping probability for increasing C_s^2 under the extended Morse PDP (4.15), resulting in lower throughput and therefore lower delay. This different behaviour is observed to occur, at least in part, over a large range of values of $q, q \leq 0.95$. As expected, the variation in mean delay illustrated by the last two graphs, where $q \to 1.0$, conforms to that of the ordinary M/GE/1/K queue.

The performance of the ordinary M/GE/1/K queue can be explained as follows: A positive linear relationship exists between C_s^2 of the GE service-time distribution and the corresponding mean size of batches completing service (Kouvatsos 1994). Therefore, for the same overall departure rate, the larger the C_s^2, the larger the mean size of batches served. Larger batches in turn require longer to form than smaller ones for the same arrival rate and thus, on the whole, customers will remain in the queue for a comparatively greater length of time.

In the fourth experiment, for lower values of C_s^2, the different behaviour resumes when λ is greater than around 18. This is because at these values of q (i.e. $q \approx 0.95$) and λ (i.e. $\lambda \approx 18$), throughput decreases as C_s^2 increases at a much greater rate than at lower values of λ. Hence, the significantly greater throughput associated with smaller C_s^2's results in greater delay. Whereas at lower values of λ, the longer times required to form larger batch sizes (marked by larger C_s^2's) outweighs the delay-impact of slightly greater throughput for smaller C_s^2's.

However, further explanation for the different behaviour is required as illustrated by the results of Experiment Set II (**Fig. 10**). These show that the behaviour persists even for the same throughput. That is to say that the lower throughput as a result of greater packet dropping is an insufficient explanation in itself for reduced delay as C_s^2 increases.

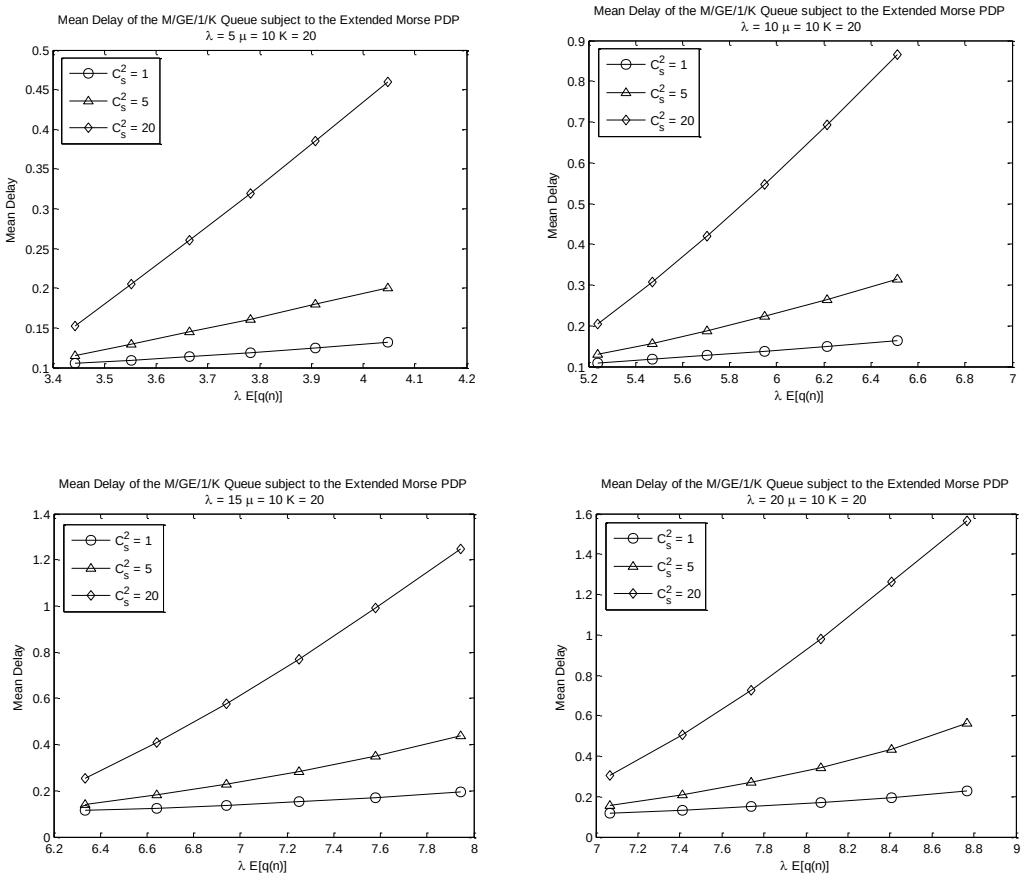
Fig. 10. Mean delay of the M/GE/1/K queue subject to the extended Morse PDP plotted against increasing throughput for fixed $\boldsymbol{q}$ (Experiment Set II).



In Experiment Set II the abscissae with common throughput are achieved by varying λ. The persistence of reduced mean delay for increasing C_s^2 is further explained by the relatively greater probabilities of lower queue occupancy levels, the higher the value of C_s^2 (for fixed q) owing to the profile of the extended Morse PDP. This condition corresponds to the formation of relatively smaller departing batch sizes on the whole and therefore lower average delays experienced by individual customers. This effect gradually decreases as the value of q increases and the model's performance approaches that of the ordinary M/GE/1/K queue as $q \rightarrow 1.0$.

Nevertheless the latter two sets of experiments illustrate the adverse impact on the mean delay of the congestion management model due to increases in λ, $\lambda E[q(n)]$ and/or q. For most values of q however, delay improvements are experienced with increasing C_s^2 (while q is fixed) at the expense of greater packet loss. This implies that if the variability of packet lengths (and thus C_s^2) increased, then in order to maintain a constant throughput, the node scheduler would need to increase q.

Fig. 11. Mean delay of the M/GE/1/K queue subject to the extended Morse PDP plotted against increasing throughput for fixed $\boldsymbol{\lambda}$ (Experiment Set III).



In the third set of experiments (Fig. 11), abscissae with common throughput are achieved by increasing the value of q as C_s^2 is increased. Values of C_s^2 are restricted to the range $[1,20]$ due to the laborious process of obtaining

numerical solutions for the cases $C_s^2 > 20.0$ within the strict tolerance levels detailed in subsection 4.2. Nonetheless, the pattern observed for this limited range continues for the cases $C_s^2 > 20.0$.

The outcome bears similarities to that of the ordinary case i.e. increasing mean delays for increasing C_s^2. This behaviour is attributed to the greater likelihood of larger queue occupancies due to the effect of increasing values of q (cf., (4.15)). The consequence of this is the formation of relatively larger departing batch sizes on average thus imposing longer average delays per customer. Therefore from Experiment Sets II and III it can be deduced that the larger departing batches and implied greater range of queue occupancy levels are achieved more readily by increasing q than by increasing λ.

In this latter set of experiments it is observed again that the effect of increasing λ, $\lambda E[q(n)]$ and/or q is performance degradation in terms of the mean delay.

Overall, the three sets of experiments demonstrate how varying the performance tuning parameter, q enables different delay – throughput trade-offs to be achieved for different values of C_s^2, λ and $\lambda E[q(n)]$.

5. The GE/GE/1/K Queue Subject to Balking

In this chapter, the QLD's of the GE/GE/1/K queue subject to population-dependent balking under three different composite batch balking and batch blocking (batch balk-block) policies are solved. Subsequently, the QLD's of the latter systems under extended Morse balking are conjectured to be special cases of the $GdHN_T$ ME distribution.

Following the derivation in Chapter 4 of the QLD of the M/GE/1/K queue subject to extended Morse balking and its conjectured equivalence to the $GdHN_T$ discrete ME distribution, it was decided to investigate the ME characteristics of a generalisation of this latter queueing system. The generalisation is achieved via an enhancement from the Poisson prospective arrival process to the bursty compound Poisson prospective arrival process with geometrically distributed batch sizes (characterised by i.i.d. GE prospective inter-arrival times). The Poisson process as a model for communication network traffic has been found to result in optimistic performance predictions. In response to this, the bursty compound Poisson process with geometrically distributed batch sizes has been proposed as a more realistic traffic model through its representation of the variability of inter-event times in addition to their mean. The latter traffic model has been found to yield more realistic queueing system performance predictions than the Poisson process (Kouvatsos 1986a; Kouvatsos 1988; Kouvatsos 1994).

5.1. The QLD of the GE/GE/1/K Queue Subject to Population-Dependent Balking

In this section, the stationary QLD's (from a random observer's point of view) of the GE/GE/1/K queue subject to population-dependent balking under three different batch balk-block policies are derived via GB analysis of the queues' Markov chain models. The GE/GE/1/K queue subject to balking is characterised by a compound Poisson prospective arrival process with geometrically distributed batch sizes (and thus i.i.d. GE prospective inter-arrival times with mean rate, λ and SCOV, C_a^2) subject to balking, i.i.d. GE service times (with mean rate, μ and SCOV, C_s^2) and finite capacity, K.

With respect to batch prospective arrivals subject to balking, members of an arriving batch can behave either uniformly (Ke 2007) or autonomously (Artalejo et al. 2005). When members of a batch behave uniformly, the entire batch joins or balks as a single entity (referred to in this thesis as complete batch balking). Whereas in the second case, each member of a batch decides autonomously whether to join or balk from the queue (referred to in this thesis as independent batch balking). In (Artalejo et al. 2005), independent batch balking with a constant balking probability for customers of all batches is used, irrespective of instantaneous queue size. To the best of the author's knowledge, independent batch balking with population-dependent balking probabilities (studied below) has not been analysed previously in the literature except by the author of this thesis in (Shah and Kouvatsos 2011; Shah and Kouvatsos 2013).

Two batch blocking policies have been widely adopted in the literature, namely complete and partial batch blocking (Manfield and Tran-Gia 1982; Kaufman and Rege 1996). In complete batch blocking, the whole arriving batch is rejected if the available queue capacity is insufficient to accommodate it entirely. On the other hand in partial batch blocking, as many customers from a batch prospective arrival as can be accommodated fill the available queue capacity with the remaining customers being rejected.

Combinations of the above batch balking and batch blocking policies have been studied in the past for example in (Choudhury et al. 1994). They analysed the multiple class $M^G(n)/M/1/K$ resource-sharing model characterised by a compound Poisson arrival process with generally distributed batch sizes and with state-dependent arrival rates, i.i.d. exponential service times and finite capacity, K subject to either the complete or partial batch blocking policies. State-dependent arrival rates were considered to be uniform between all the customers in a batch thus rendering the model equivalent to the $M^G/M/1/K$ queue subject to complete batch balking and either complete or partial batch blocking.

The three batch balk-block policies analysed in this thesis are described below:

1. Complete batch balking and complete batch blocking (Policy I)

The joining/balking and blocking behaviour of each of the members of a batch is identical resulting in the batch behaving as a single entity. The

entire batch either proceeds to join the queue or balks following a decision to join or balk respectively. If there is insufficient capacity for the entire batch to join then the whole batch is rejected. This policy can be interpreted as unity between members of a batch being preserved throughout the system operation regardless of the level of resource availability.

2. Complete batch balking and partial batch blocking (Policy II)

When there is sufficient capacity for the entire batch, the joining/balking behaviour of each of the members of a batch is identical. However, when there is insufficient capacity for the entire batch, the uniform behaviour continues solely in the case of the decision to balk. Following a join decision, as many customers as can be accommodated enter the queue successively from the head of the batch and the remaining customers are blocked.

3. Independent batch balking and partial batch blocking (Policy III)

Members of a batch behave independently with respect to joining/balking such that each successive member of a batch makes an autonomous decision to join or balk. Individual batch members which decide (autonomously) to join the queue proceed to occupy successive positions in the queue until it becomes full. After this, subsequent members are blocked. Under this policy, customer autonomy is upheld irrespective of the level of resource availability.

The state transition rates, R_{ij}'s of the Markov chain models of the GE/GE/1/K queue subject to population-dependent balking under the above three policies are presented below. It is assumed that all the members of a batch

'see' the same instantaneous queue size, n. Hence all the members of a batch which elect to join the queue, do so with the same conditional probability defined by the population-dependent joining function, $q(n), n = 0,1,2 \ldots K-1$. Customers which balk and those which are blocked are deemed to be lost from the viewpoint of the queue.

The downward state transition rates of the Markov chain models associated with the GE/GE/1/K queue subject to population-dependent balking under each of the three batch balk-block policies above are identical to each other and to those associated with the M/GE/1/K queue subject to population-dependent balking, and thus are given by (4.4) and (4.5).

The corresponding upward state transition rates are derived below and notably, these can easily be extended to characterise the Markov chain models of the M^G/GE/1/K queue subject to population-dependent balking under each of the above three batch balk-block policies. The latter queueing system is marked by a compound Poisson prospective arrival process with generally distributed batch sizes, i.i.d. GE service times and finite capacity, K. This can be achieved by simply replacing the geometric batch size distribution, $\sigma(1-\sigma)^{i-1}, i = 1,2,3 \ldots$ with the desired one.

1. *Complete batch balking and complete batch blocking (Policy I)*

$$R_{ij} = \binom{prospective\ batch}{arrival\ rate} \times P\begin{pmatrix} batch\ arrival \\ joins\ the\ queue \\ of\ length\ i \end{pmatrix} \times P\binom{batch\ arrival\ is}{of\ size\ (j-i)}, \quad (5.1)$$

$$i = 1,2,3 \ldots K-1, j = i+1, i+2, i+3 \ldots K$$

$$= (\sigma\lambda)q(i)\sigma(1-\sigma)^{j-i-1}$$

where σ is the inverse of the mean of arriving batch sizes and it can be defined in terms of C_a^2 as (Kouvatsos 1994)

$$\sigma = \frac{2}{1+C_a^2}. \tag{5.2}$$

$$\begin{aligned} R_{0j} &= \begin{pmatrix} prospective\ batch \\ arrival\ rate \end{pmatrix} \times P\begin{pmatrix} batch\ arrival \\ joins\ the\ empty \\ queue \end{pmatrix} \\ &\times P\begin{pmatrix} batch\ arrival\ is\ of\ size\ k, k = j, j+1, j+2 \dots K, \\ with\ the\ first\ \ (k-j)\ customers\ exiting\ the\ queue\ via \\ the\ zero\ branch\ of\ the\ service\ process\ and\ the \\ remaining\ j\ customers\ residing\ in\ the\ queue \end{pmatrix}, \\ & j = 1,2,3 \dots K-1 \\ &= (\sigma\lambda)q(0)\sum_{k=j}^{K}\left((1-\tau)^{k-j}\tau\sigma(1-\sigma)^{k-1}\right) \\ &= \frac{\sigma^2\lambda\tau q(0)(1-\sigma)^{j-1}}{1-(1-\tau)(1-\sigma)}\left(1-\left((1-\tau)(1-\sigma)\right)^{(K-j+1)}\right) \end{aligned} \tag{5.3}$$

$$\begin{aligned} R_{0K} &= \begin{pmatrix} prospective\ batch \\ arrival\ rate \end{pmatrix} \times P\begin{pmatrix} batch\ arrival \\ joins\ the\ empty \\ queue \end{pmatrix} \times P\begin{pmatrix} batch\ arrival\ is \\ of\ size\ K \end{pmatrix} \\ &= (\sigma\lambda)q(0)\tau\sigma(1-\sigma)^{K-1} \end{aligned} \tag{5.4}$$

2. *Complete batch balking and partial batch blocking (Policy II)*

$$R_{ij} = \binom{prospective\ batch}{arrival\ rate} \times P\begin{pmatrix} batch\ arrival \\ joins\ the\ queue \\ of\ length\ i \end{pmatrix} \times P\binom{batch\ arrival\ is}{of\ size\ (j-i)},$$

$$i = 1,2,3 \dots K-2, j = i+1, i+2, i+3 \dots K-1 \tag{5.5}$$

$$= (\sigma\lambda)q(i)\sigma(1-\sigma)^{j-i-1}$$

$$R_{iK} = \binom{prospective\ batch}{arrival\ rate} \times P\begin{pmatrix} batch\ arrival \\ joins\ the\ queue \\ of\ length\ i \end{pmatrix}$$

$$\times P\binom{batch\ arrival\ is\ of\ size\ k}{k = K-i, K-i+1, K-i+2, \dots \infty}, i = 1,2,3 \dots K-1 \tag{5.6}$$

$$= (\sigma\lambda)q(i)\sum_{k=K-i}^{\infty}\left(\sigma(1-\sigma)^{k-1}\right)$$

$$= (\sigma\lambda)q(i)(1-\sigma)^{K-i-1}$$

$$R_{0j} = \binom{prospective\ batch}{arrival\ rate} \times P\begin{pmatrix} batch\ arrival \\ joins\ the\ empty \\ queue \end{pmatrix}$$

$$\times P\begin{pmatrix} batch\ arrival\ is\ of\ size\ k, k = j, j+1, j+2 \dots \infty, \\ with\ the\ first\ \ (k-j)\ customers\ exiting\ the\ queue\ via \\ the\ zero\ branch\ of\ the\ service\ process\ and\ the \\ remaining\ j\ customers\ residing\ in\ the\ queue \end{pmatrix},$$

$$j = 1,2,3 \dots K-1 \tag{5.7}$$

$$= (\sigma\lambda)q(0)\sum_{k=j}^{\infty}\left((1-\tau)^{k-j}\tau\sigma(1-\sigma)^{k-1}\right)$$

$$= \frac{(\sigma\lambda)q(0)\tau\sigma(1-\sigma)^{j-1}}{1-(1-\tau)(1-\sigma)}$$

$$R_{0K} = \binom{prospective\ batch}{arrival\ rate} \times P\begin{pmatrix} batch\ arrival \\ joins\ the\ empty \\ queue \end{pmatrix}$$

$$\times P\begin{pmatrix} batch\ arrival\ is\ of\ size\ k, k = K, K+1, K+2 \dots \infty, with \\ the\ first\ j, j = 0,1,2 \dots (k-K)\ customers\ exiting\ the\ queue \\ via\ the\ zero\ branch\ of\ the\ service\ process, \\ K\ customers\ residing\ in\ the\ queue\ and\ the\ remaining \\ members\ of\ the\ batch\ being\ blocked \end{pmatrix} \quad (5.8)$$

$$= (\sigma\lambda)q(0)\sum_{k=K}^{\infty}\left(\sum_{j=0}^{k-K}\left((1-\tau)^{j}\tau\sigma(1-\sigma)^{k-1}\right)\right)$$

$$= \frac{(\sigma\lambda)q(0)\tau(1-\sigma)^{K-1}}{1-(1-\tau)(1-\sigma)}$$

3. *Independent batch balking and partial batch blocking (Policy III)*

$$R_{ij} = \binom{prospective\ batch}{arrival\ rate} \times P\begin{pmatrix} batch\ arrival\ is\ of\ size\ k, \\ k = j-i, j-i+1, j-i+2, \dots \infty \\ and\ (j-i)\ from\ the\ k\ customers \\ join\ the\ queue\ of\ length\ i \end{pmatrix},$$

$$i = 1,2,3 \dots K-1, j = i+1, i+2, i+3 \dots K-1 \quad (5.9)$$

$$= (\sigma\lambda)\sum_{k=j-i}^{\infty}\sigma(1-\sigma)^{k-1}\left(\binom{k}{j-i}q(i)^{j-i}\left(1-q(i)\right)^{k-(j-i)}\right),$$

$$0.0 < q(i) < 1.0$$

$$R_{iK} = \binom{prospective\ batch}{arrival\ rate} \times P\begin{pmatrix} batch\ arrival\ is\ of\ size\ k, \\ k = K-i, K-i+1, K-i+2, \dots \infty \\ and\ (K-i)\ from\ the\ k\ customers \\ join\ the\ queue\ of\ length\ i, with\ the \\ rest\ balking\ or\ being\ blocked \end{pmatrix},$$

$$i = 1,2,3 \dots K-1 \quad (5.10)$$

$$= (\sigma\lambda)\left(\left(q(i)\right)^{K-i}\right)\sum_{k=K-i}^{\infty}\sigma(1-\sigma)^{k-1}\sum_{\substack{m_i, m_{i+1} \dots m_K: \\ m_i+m_{i+1}+\cdots+m_K=k-(K-i)}}\left(\left(1-q(i)\right)^{m_i+m_{i+1}+\cdots+m_{K-1}}\right),$$

$$0.0 < q(i) < 1.0$$

where the $m_i, m_{i+1}, \dots, m_K$'s are the summands comprising compositions resulting from $\big(k-(K-i)\big)$ into $(K-i+1)$ parts. In (Nijenhuis and Wilf 1978), an algorithm, NEXCOM, is provided to efficiently generate compositions.

$$R_{0j} = \binom{prospective\ batch}{arrival\ rate} \times P\left(\begin{array}{c} batch\ arrival\ is\ of\ size\ k, k = j, j+1, j+2 \dots \infty, of\ which \\ (j+l), l = 0,1,2 \dots (k-j)\ customers\ join\ and\ (k-j-l)\ balk. \\ Of\ those\ (j+l)\ which\ join, the\ first\ l\ exit\ the\ queue\ via\ the \\ zero\ branch\ of\ the\ service\ process\ and\ the\ remaining\ j \\ customers\ opt\ to\ reside\ in\ the\ queue. \end{array}\right),$$

$$j = 1,2,3 \dots K-1$$

$$= (\sigma\lambda) \sum_{k=j}^{\infty} \sum_{l=0}^{k-j} (1-\tau)^l \tau\sigma(1-\sigma)^{k-1} \left(\binom{k}{j+l} q(0)^{j+l} \big(1-q(0)\big)^{k-j-l} \right), \tag{5.11}$$

$$0.0 < q(0) < 1.0$$

$$R_{0K} = \binom{prospective\ batch}{arrival\ rate} \times P\left(\begin{array}{c} batch\ arrival\ is\ of\ size\ k, k = K, K+1, K+2 \dots \infty, of\ which \\ (K+l), l = 0,1,2 \dots (k-K)\ customers\ join\ and\ (k-K-l)\ balk\ or \\ are\ blocked.\ Of\ those\ (K+l)\ which\ join, the\ first\ l\ exit\ the \\ queue\ via\ the\ zero\ branch\ of\ the\ service\ process\ and \\ the\ remaining\ K\ customers\ opt\ to\ reside\ in\ the\ queue. \end{array}\right),$$

$$= (\sigma\lambda) \Big(\big(q(0)\big)^{K+l} \Big) \sum_{k=K}^{\infty} \sum_{l=0}^{k-K} (1-\tau)^l \tau\sigma(1-\sigma)^{k-1} \sum_{\substack{m_0, m_1 \dots m_K: \\ m_0+m_1+\cdots+m_K = k-K-l}} \Big(\big(1-q(0)\big)^{m_0+m_1+\cdots+m_{K-1}} \Big), \tag{5.12}$$

$$0.0 < q(0) < 1.0$$

where the $m_0, m_1, \dots, m_K$'s are the summands of compositions of $(k-K-l)$ into $(K+1)$ parts.

The system of GB equations (4.6) holds here and its solution in the context of the GE/GE/1/K queue subject to population-dependent balking under Policies I – III results in the following recursive (Policies I and III) and closed-form (Policy II) QLD's:

1. QLD of the GE/GE/1/K queue subject to population-dependent balking under Policies I and III

$$p_n^K = \begin{cases} p_0^K, n = 0 \\ p_0^K \dfrac{(1-\tau)R_{01} + \tau\sum_{j=1}^{K} R_{0j}}{\tau\mu + (1-\tau)\sum_{j=2}^{K} R_{1j}}, n = 1 \\ \dfrac{(1-\tau)\sum_{i=0}^{n-1} R_{in}p_i^K + p_{n-1}^K\left(\tau\mu + \sum_{j=n}^{K} R_{n-1,j}\right) - \sum_{i=0}^{n-2} R_{i,n-1}p_i^K}{\tau\mu + (1-\tau)\sum_{j=n+1}^{K} R_{nj}}, n = 2,3,4 \dots K-1 \\ \dfrac{\sum_{i=0}^{K-1} R_{iK}p_i^K}{\tau\mu}, n = K \end{cases} \quad (5.13)$$

where the R_{ij}'s are given by (5.1) - (5.4) for Policy I and (5.9) - (5.12) for Policy III.

2. QLD of the GE/GE/1/K queue subject to population-dependent balking under Policy II

$$p_n^K = \begin{cases} p_0^K, n = 0 \\ p_0^K \dfrac{\sigma\lambda q(0)\tau}{(\sigma\lambda q(1)(1-\tau) + \tau\mu)} \prod_{i=2}^{n} \left(\dfrac{(1-\sigma)\tau\mu + \sigma\lambda q(i-1)}{\sigma\lambda q(i)(1-\tau) + \tau\mu} \right), n = 1,2,3 \dots K-1 \\ p_0^K \dfrac{1}{\tau\mu} \left(\dfrac{(1-\sigma)\tau\mu + \sigma\lambda q(K-1)}{\sigma + \tau(1-\sigma)} \right) \left(\dfrac{\sigma\lambda q(0)\tau}{\sigma\lambda q(1)(1-\tau) + \tau\mu} \right) \prod_{i=2}^{K-1} \left(\dfrac{(1-\sigma)\tau\mu + \sigma\lambda q(i-1)}{\sigma\lambda q(i)(1-\tau) + \tau\mu} \right), n = K \end{cases} \quad (5.14)$$

5.1.1. Batch prospective arrivals subject to extended Morse balking

In the Morse balking paradigm, the joining (and balking) probabilities are dependent on the (exact or average) instantaneous workload and average impatience of the customer population (cf., (2.28), (2.29) and (2.31)). These probabilities are independent of the prospective arrival process. Therefore, in the case of any prospective arrival process of single customers and i.i.d. general service times, the extended Morse joining probabilities are identical to those of the Poisson prospective arrival case defined by (4.9).

Since the extended Morse joining probabilities are independent of the prospective arrival process, consider a general batch prospective arrival process, characterised by a general batch inter-arrival time distribution and general batch size distribution subject to extended Morse balking. In the latter case under complete batch balking with complete or partial batch blocking (i.e. Policies I or II), based on the assumption that all members of a batch balk with identical probability (cf., Section 5.1), the parameter α is re-interpreted as the Morse average measure of impatience of customer batches, α_B. The extended Morse joining function now becomes

$$q(n) = \begin{cases} 1.0, n = 0 \\ e^{-\frac{\alpha_B n}{\mu}} e^{-\frac{\alpha_B}{\mu}\left(\frac{C_S^2-1}{2}\right)}, n = 1,2,3 \ldots \end{cases} . \qquad (5.15)$$

In the case of batch prospective arrivals subject to extended Morse balking under independent batch balking with partial batch blocking (i.e. Policy III), since members of a batch elect to join or balk autonomously albeit with identical probabilities, the corresponding Morse average measure of impatience, α, applies to individual customers. Hence in this case, the extended Morse joining function defined by (4.9) holds.

Analogous to the estimation of α in the context of single prospective arrivals (cf., Sections 2.3.1 and 4.1), the parameters α_B and α in the batch prospective arrival context may also be estimated numerically as shown below. In the specific context of compound Poisson prospective arrivals with geometrically distributed batch sizes (characterised by i.i.d. GE prospective inter-arrival times) subject to extended Morse balking under Policies I, II or III, assuming knowledge of the loss (i.e. overall balking and blocking) rate of individual customers, LR, α_B (Policies I or II) or α (Policy III) can be estimated by solving

$$\sum_{n=0}^{K-1}\left(\sum_{i=1}^{K-n} R_{n,n+i}\right) p_n^K - (\lambda - LR) = 0 \qquad (5.16)$$

for α_B or α, where the upward state-transition rates, R_{ij}'s are given by (5.1) - (5.12) (as appropriate for each of the policies) and the corresponding $q(n)$'s are given by (5.15) (Policies I or II) or (4.9) (Policy III) respectively. Alternatively, α_B or α can be estimated by setting up analogous equations using other appropriate metrics.

Mathematical definitions of the parameters α_B and α in the context of batch prospective arrivals and i.i.d. general service times are presented in Appendix B.

5.2. Discussion of the Results

In this section, for the purpose of ME performance modelling and prediction of queueing systems, investigations are carried out to determine equivalence in distribution between the QLD's of the GE/GE/1/K queue subject to extended Morse balking under the three different batch balk-block policies and their corresponding $GdHN_T$ ME inferences.

Expressions for the QLD's of the GE/GE/1/K queue subject to extended Morse balking under the three batch balk-block policies are obtained by substituting $q(n)$ with the extended Morse joining function (4.10), in (the appropriate upward state-transition rates of) (5.13), for Policies I and III, and (5.14), for Policy II, respectively. As in the case of the M/GE/1/K queue subject to extended Morse balking (cf., Section 4.2), in this case too, equivalence in distribution between the QLD's and corresponding ME distribution inferences cannot be concluded by observation alone. Hence, equivalence in distribution was investigated numerically by an analogous procedure to that used for the M/GE/1/K queue subject to extended Morse balking (cf. Section 4.2). Details of the investigation are as follows.

For the experimentation, numerical QLD probabilities, $p_n^{K\star}, n = 0,1,2 \dots K$, of the GE/GE/1/K queue subject to extended Morse balking under each of the

three batch balk-block policies were generated from the respective equations (5.13) and (5.14) with $q(n)$ specified as the extended Morse joining function (4.10) and different combinations of queue input parameter values from **Table 3** below.

Table 3. Parameter values used in the validation of ME characteristics of the QLD's of the GE/GE/1/K queue subject to extended Morse balking under Policies I – III.

Parameter	Value(s)
λ	[10 20 30 40]
μ	20
C_a^2, C_s^2	[1 5 10 20 50 100 200 500]
K	[5 10 15 20] (Policies I & II)
	[5 10] (Policy III)
q	[0.3 0.6 0.9]

Analogous to the earlier experimentation, an 'experiment' here refers to the generation of a single set of numerical QLD probabilities, $p_n^{K\star}, n = 0,1,2 \ldots K$, from a particular combination of queue input parameter values for the GE/GE/1/K queue subject to extended Morse balking under one of the batch balk-block policies. It also includes the generation of the corresponding numerical GdHN$_T$ ME distribution inference, $p_n^{K\dagger}, n = 0,1,2, \ldots, K$. By the

fundamental principle of counting, there are a total of 7680 combinations[14] of queue input parameter values and hence a total of 7680 experiments were conducted. From each of these numerical QLD's, $p_n^{K\star}, n = 0,1,2 \ldots K$, values of the MQL, VQL, $p_0^{K\star}$ and $p_K^{K\star}$ were either calculated or obtained directly. The latter values comprised the optimisation constraints, in addition to the normalisation condition, in the numerical constrained maximisation of Shannon's entropy functional (2.1), yielding the corresponding numerical GdHN$_T$ distribution inference probabilities, $p_n^{K\dagger}, n = 0,1,2, \ldots, K$. Finally, for each experiment the error, given by (4.13), was computed.

The experiments were carried out in MATLAB version 7.10.0.499 (R2010a). The range of queue capacities used in the case of Policy III was smaller due to the memory and speed limitations of the PC used for the experiments. And therein lies the main weakness of the above solution of the QLD of the GE/GE/1/K queue subject to balking under Policy III, namely the high computational demands of its implementation. Both the constraints and change in objective function were satisfied to within the default tolerance of 10^{-6}. Owing to the limitations of the software to produce the GdHN$_T$ distribution inferences when the prior moment constraint information, $p_K^{K\star} < 10^{-10}$, the special case of the GdHN$_T$ ME inference used in the earlier experimentation (cf. Section 4.2) was adopted in those instances.

The largest error observed for the case of the GE/GE/1/K queue subject to extended Morse balking under Policy I was 0.018 with the overwhelming majority of errors less than 0.005. For Policy II, the largest error observed

[14] The numbers of combinations of queue input parameter per policy are 3072, 3072 and 1536 respectively, resulting in 7680 in total.

was 0.019 with the overwhelming majority of errors below 0.007. Under the third policy, the maximum error encountered was 0.0277 with the overwhelming majority lying below 0.008.

Exact equivalence in distribution between the above GB and ME solutions is conjectured based on the following arguments, which are analogous to those used in the earlier experimentation of the M/GE/1/K queue subject to extended Morse balking:

1. Errors of comparable magnitude were encountered in analogous experiments conducted for special cases which are known to be equivalent.

   Errors were computed for the case of the ordinary GE/GE/1/K queue for which exact equivalence has been proven in (Kouvatsos 1986b). Over all the experiments conducted for the latter queueing system, the largest error encountered was 0.000848 providing confirmation that these errors and by extension those of the GE/GE/1/K queue subject to extended Morse balking under the three policies can be attributed to numerical limitations of the software package. The comparatively larger errors encountered for the GE/GE/1/K queue subject to extended Morse balking under the three policies may again be attributed to the effect of MATLAB's computational approximations on a larger set of optimisation constraints.

2. Errors did not increase with increasing C_a^2 and C_s^2.

   Errors were not found to increase and remained low with increasing C_a^2 and C_s^2 in contrast to certain existing ME approximations of ordinary queueing systems marked by i.i.d. GE inter-arrival and/or service times. In the latter case, absolute differences between the ME approximations and simulation results were found to grow with increasing C_a^2 and/or C_s^2 (cf., (Kouvatsos and Awan 2003)).

Supported by the above experimental evidence and subsequent reasoning, the following conjecture is proposed.

Conjecture II. *The QLD's of the GE/GE/1/K queue subject to extended Morse balking under Policies I, II or III* (5.13) *and* (5.14) *are special cases of the GdHN$_T$ ME distribution* (3.23) *constrained by the prior information of the individual queues' MQL, VQL (or second moment of queue length), empty state probability, p_0^K (or equivalently server utilisation), full buffer state probability, p_K^K and the normalisation condition over finite, non-negative integer support* $[0, K]$.

5.2.1. The Infinite-Capacity Special Case

Setting $K \to \infty$ in the above state-transition rates (5.1) – (5.12) and appropriate QLD's (5.13) or (5.14) yields the QLD's of the infinite-capacity GE/GE/1 queue subject to either complete batch balking (as a special case of both Policies I and II) or independent batch balking (as a special case of

Policy III). When the joining function, $q(n)$ is specified as the extended Morse joining function, then the QLD's of the latter infinite-capacity queues are conjectured to be special cases of the GdHN.

5.2.2. The Poisson Prospective Arrival Special Case

Putting $C_a^2 = 1.0$ implies $\sigma = 1.0$ (from (5.2)). From the illustration of the GE two-phase interpretation (Fig. 2) it can be seen how the latter condition results in single prospective arrivals with exponential inter-arrival times i.e. Poisson prospective arrivals. Applying this condition to the solutions of the GE/GE/1/K queue subject to balking under any of the three policies yields the QLD of the M/GE/1/K queue subject to population-dependent balking (4.7).

5.2.3. The Exponential Service Special Case

Setting $C_s^2 = 1.0$ implies $\tau = 1.0$ from (4.2). From the illustration of the GE two-phase interpretation (Fig. 2) it can be seen how the latter condition results in exponential service of single customers only. Under this condition, the solutions of the GE/GE/1/K queue subject to balking under each of the three policies reduce to the corresponding QLD's of the GE/M/1/K queue subject to balking under each of the three policies. When $q(n)$ is specified as the Morse joining function, then the latter QLD's are conjectured to be special cases of the $GdHN_T$.

5.2.4. The Non-Balking Special Case

Specifying $q(n) = 1.0, n = 0,1,2 \ldots K-1$ in the state-transition rates of the GE/GE/1/K queue subject to balking under Policy I ((5.1) – (5.4)) and substituting these in (5.13) yields the QLD of the ordinary GE/GE/1/K queue subject to complete batch blocking.

Moreover, the ordinary GE/GE/1/K queue under partial batch blocking (solved in (Kouvatsos et al. 1989)) arises as a special case of the GE/GE/1/K queue subject to balking under Policies II or III when the condition $q(n) = 1.0, n = 0,1,2 \ldots K-1$ is specified.

These special cases are useful in the case of extended Morse balking where $q(0) = 1.0$.

6. Conclusions

Novel generalised least biased inferences, namely the generalised discrete Half Normal (GdHN) and truncated GdHN ($GdHN_T$) ME solutions have been characterised, subject to prior knowledge of the first moment, variance, boundary state probabilities, p_0 (infinite support case) or p_0^K and p_K^K (finite support case) and the normalising condition. These latter ME solutions have been devised particularly for least biased inferences of the stationary QLD's of infinite and finite-capacity, respectively, ordinary G/G/1 queues, G/G/1 queues subject to extended Morse balking or ordinary G/G/1 queues subject to population-dependent arrival rates governed by the extended Morse joining function.

Subsequently the closed-form QLD of the M/GE/1/K queue subject to population-dependent balking was derived via the technique of GB. Specifically under the population-dependent extended Morse balking regime, the latter QLD was conjectured, based on extensive numerical experimentation, to be a special case of the $GdHN_T$ ME distribution. Furthermore, owing to its appropriate operational properties, the extended Morse balking function (4.15) was applied as a model of the class of instantaneous, early random drop congestion management mechanisms. As a consequence, the novel M/GE/1/K queue subject to extended Morse balking was submitted as a suitable ME performance model of IP-based communication network nodes featuring such congestion management mechanisms set up to run either statically or dynamically.

A performance evaluation study of the ME congestion management model was carried out by assessing the impact on its mean delay due to increasing the values of the squared coefficient of variation (SCOV) of node service durations, C_s^2, overall traffic arrival rate to the node, λ, node throughput, $\lambda E[q(n)]$ and extended Morse packet dropping policy (PDP) performance tuning parameter, q. On the whole, the consequence of these conditions was performance degradation with the exception of the variation in mean delay for increasing C_s^2 when q was fixed. Under the latter conditions, improvements in mean delay were experienced for increasing C_s^2 (while q was fixed). This observation implies that if the variability of packet lengths (and thus C_s^2) were to increase, then in order to maintain a constant throughput, the node scheduler would need to increase the value of q. The performance study demonstrated how varying the performance tuning parameter, q, enables different delay – throughput trade-offs to be achieved for different values of C_s^2, λ and/or $\lambda E[q(n)]$.

Following this development, a generalisation from the Poisson to the compound Poisson prospective arrival process with geometrically distributed batch sizes (characterised by i.i.d. GE prospective inter-arrival times) was made. This necessitated modelling combinations of batch balking and batch blocking (batch balk-block) policies. Consequently, the QLD's of the novel GE/GE/1/K queue subject to population-dependent balking under three batch balk-block policies were solved via GB analysis of the queues' Markov chain models. Based on extensive numerical experimentation, the QLD's of the GE/GE/1/K queue, subject to extended Morse balking under the three batch

balk-block policies were conjectured to be special cases of the $GdHN_T$ ME distribution.

In addition to the analysis of extended Morse balking in the context of the above GE queues, more general queues characterised by a general batch prospective arrival process, subject to extended Morse balking and/or i.i.d. general service times were considered. Furthermore, in this latter context, new definitions of the Morse average measure of impatience, α, have been derived based on equivalence between the Morse and Haight balking paradigms.

6.1. Limitations

A future aim related to this research work is to eventually use the novel queueing models devised in this thesis as building blocks in the ME approximate analysis of non-exponential queueing network models (QNM's) with arbitrary topology, network blocking mechanisms and balking (or packet dropping congestion management mechanisms). In the latter context, the main limitation of this research work is the independence assumed between prospective arrivals to the queues whether Poisson or compound Poisson prospective arrivals with geometrically distributed batch sizes. On the contrary, long range dependence (LRD) has been observed in real network and World Wide Web traffic (Garrett and Willinger 1994; Leland et al. 1994; Beran et al. 1995; Crovella and Bestavros 1997). The independence assumption results in relatively optimistic queue performance predictions

compared to the case where incoming traffic is correlated (Fretwell and Kouvatsos 2002).

Nonetheless, the new queueing systems solved in this thesis can be useful when modelling network queues with small buffers since under this condition, traffic correlation is limited and consequently the queue behaviour conforms more closely to one fed by a renewal process (Fretwell and Kouvatsos 2002). Furthermore, based on the results obtained for the discrete time queues in (Kouvatsos et al. 2000), it may be that under certain parameter values, in performance terms, a (continuous time) queue fed by short range dependent traffic can be approximated with tolerable accuracy by a corresponding queue with the compound Poisson prospective arrival process with geometrically distributed batch sizes (employed in this thesis).

Notably, LRD input traffic results in heavy-tailed QLD's (Norros 1994; Erramilli et al. 1996). Whereas the ME principle, as employed in this thesis, generates inferences suited to modelling 'extensive' systems marked by subsystem independence. Maximising more general, non-extensive entropy functions such as the Havrda-Charvát-Tsallis entropy function (Havrda and Charvát 1967; Tsallis 1988), subject to the commonly used queueing system prior information constraints, yields heavy-tailed ME solutions, which are applicable to modelling queueing systems fed by LRD input traffic. Some initial results in this vein have been published in the literature, for example in (Assi 2000; Kouvatsos and Assi 2002; Karmeshu and Sharma 2005; Karmeshu and Sharma 2006c; Karmeshu and Sharma 2006a; Karmeshu and Sharma 2006b; Kouvatsos and Assi 2011a; Kouvatsos and Assi 2011b). An axiomatic characterisation of non-extensive entropy maximisation using

the Havrda-Charvát-Tsallis entropy function has been carried out in (Kouvatsos and Assi 2011a). Since Shannon's entropy arises as a special case of numerous non-extensive entropy functions, the new ME solutions and QLD's devised in this thesis are special cases of the heavy-tailed solutions derived using the non-extensive entropies, subject to the same prior information constraints. As such, the results of this thesis provide special cases which would be useful in testing and establishing new heavy-tailed ME queueing solutions.

6.2. Future Work

Some open problems stemming from this research work are listed below:

1. Proof of conjectures I (cf., Section 4.2) and II (cf., Section 5.2).
2. It is to be recalled that experimentally, for $C_a^2, C_s^2 > 1.0$, the ordinary GE/GE/1 queue has been found to give pessimistic performance bounds over a large class of equivalent queues characterised by two-phase exponential inter-arrival and service time distributions, such as the H_2 or Coxian-2, with matching first two moments (Kouvatsos 1988; Kouvatsos and Tabet-Aouel 1994).

   It is therefore proposed that an analogous investigation is carried out into the performance of the GE/GE/1/K queue, subject to balking, relative to that of equivalent two-phase queues, subject to balking. Owing to such a performance comparison, a fourth batch balk-block policy, Policy IV, has been devised and it is introduced, described and analysed in Appendix C.

It still remains to be determined whether the QLD of the GE/GE/1/K queue, subject to extended Morse balking under Policy IV is a special case of the $GdHN_T$ ME distribution.

3. A performance comparison of the GE/GE/1/K queue, subject to extended Morse balking under the four batch balk-block policies which lend themselves to modelling PDP's of the instantaneous, random, early drop type at IP-based network nodes subject to bursty arrivals.
4. Analyse how the variance of queue length (VQL) prior information constraint captures the uncertainty arising from the selective behaviour of prospective arrivals which join or balk according to the Morse and extended Morse balking policies.
5. Assess the accuracy of the approximation of the departure process from the M/GE/1/K queue, subject to extended Morse balking presented in Appendix D. Extend this approximation to include the compound Poisson prospective arrival process with geometrically distributed batch sizes (and thus i.i.d. GE prospective inter-arrival times), Whitt's asymptotic approximation for the departure process and network blocking mechanisms.

Following the approximate analysis of the departure process from the GE/GE/1/K queue, subject to population-dependent balking and utilising existing flow formulae for splitting and merging of GE-type traffic streams (cf., (Kouvatsos 1994; Kouvatsos et al. 2011)), the GE/GE/1/K queue, subject to extended Morse balking is envisaged to play the role of a building block model in the ME approximate analysis of non-exponential QNM's with arbitrary topology, network blocking mechanisms and balking

(or packet dropping congestion management schemes) (Whitt 1982; Whitt 1984; Kouvatsos 1986b; Kouvatsos 1986a; Kouvatsos 1994; Kouvatsos and Awan 2003; Kouvatsos et al. 2011).

6. Use non-extensive entropy functions, such as the Havrda-Charvát-Tsallis entropy function (Havrda and Charvát 1967; Tsallis 1988) to devise new heavy tail inferences of queueing system performance distributions.
7. Determine the general inter-event time distribution characterising the infinite and finite capacity ordinary M/G/1 and G/G/1 queues bearing the GdHN and $GdHN_T$ ME QLD's respectively. At the outset, this inter-event time distribution, which is a generalisation of the GE distribution, is named the '*Kouvatsos distribution*' in honour of my research supervisor Professor Demetres D. Kouvatsos.

Appendices

Appendix A: An alternative characterisation of the GGeo discrete ME distribution

In this section, the GGeo discrete ME distribution is characterised as a random sum and this result is applied to the solution of the QLD of the GE/M/1 queue.

A1. The GGeo as a random sum

The GGeo is characterised as the random sum of i.i.d. geometric RV's bounded by the modified geometric distribution. This is demonstrated below.

Let G represent a geometric RV with probability mass function (pmf) given by

$$P(G = i) = (1 - \sigma)^{i-1}\sigma, i = 1,2,3 \dots, \quad \text{(A1)}$$

let H model a modified geometric RV with pmf defined by

$$P(H = j) = (1 - \nu)\nu^{j}, j = 0,1,2 \dots \quad \text{(A2)}$$

and let the PGF, $G_X(z)$, of the RV, X, be given by $G_X(z) = \sum_{n=0}^{\infty} z^n P(X = n)$.

Then, the RV, N, with pmf, $P(N = n) = p_n, n = 0,1,2 \ldots$, resulting from the random sum of i.i.d. copies of the RV, G bounded by the pmf of H can be derived by applying the theorem of total generating functions as follows.

The PGF of N can be given by

$$
\begin{aligned}
G_N(z) &= \sum_{n=0}^{\infty} \big(G_G(z)\big)^n P(H = n) \\
&= \sum_{n=0}^{\infty} \left(\frac{\sigma z}{1-(1-\sigma)z}\right)^n (1-v)v^n \\
&= \frac{(1-v)(1-(1-\sigma)z)}{1-\big(\sigma v + (1-\sigma)\big)z} \\
&= \frac{(1-v)}{1-\big(\sigma v + (1-\sigma)\big)z} \\
&\quad - \frac{(1-v)(1-\sigma)z}{1-\big(\sigma v + (1-\sigma)\big)z}.
\end{aligned}
\tag{A3}
$$

Inverting $G_N(z)$ (A3) yields the GGeo pmf defined as

$$
p_n = \begin{cases} (1-v), n = 0 \\ (1-v)\left(\dfrac{\sigma v}{\sigma v + (1-\sigma)}\right)\big(\sigma v + (1-\sigma)\big)^n, n = 1,2,3 \ldots \end{cases}. \tag{A4}
$$

Furthermore, following the derivation of the first two moments of p_n (A4), the GGeo pmf can be defined in terms of its mean $E[N]$ and SCOV, C_N^2 as

$$p_n = \begin{cases} 1 - \dfrac{2E[N]}{(C_N^2+1)E[N]+1}, n = 0 \\ \dfrac{2E[N]}{(C_N^2+1)E[N]+1}\left(1 - \dfrac{(C_N^2+1)E[N]-1}{(C_N^2+1)E[N]+1}\right)\left(\dfrac{(C_N^2+1)E[N]-1}{(C_N^2+1)E[N]+1}\right)^{n-1}, n = 1,2,3 \ldots \end{cases} . \quad \text{(A5)}$$

A2. The stationary QLD of the GE/M/1 queue

Consider a single server, infinite-capacity, ordinary queue with (single) Poisson arrivals and i.i.d. exponential service times (i.e. the M/M/1 queue). Its QLD is the modified geometric with pmf defined by

$$p_n = (1-\rho)\rho^n, n = 0,1,2 \ldots \quad \text{(A6)}$$

where $\rho = \lambda/\mu$ and λ and μ are the mean arrival and service rates of the queue respectively.

Now consider that the arrivals come in independent batches of geometrically-distributed sizes with pmf defined by (A1) (i.e. the GE/M/1 queue). If λ_B is the mean batch arrival rate, then the mean arrival rate of (individual) customers to the queue, $\lambda_{ind} = \lambda_B(1/\sigma)$.

In this case, the effective service-time distribution per batch is the distribution characterised by the random sum of i.i.d. exponential RV's, each with mean rate, μ_{ind}, bounded by the geometric RV, G. This is known to be the exponential distribution with revised rate $\sigma\mu_{ind}$ (Feller 1966). Therefore the effective batch service rate, $\mu_B = \sigma\mu_{ind}$.

This latter queue has i.i.d. exponential inter-batch arrival times and i.i.d. exponential batch service times. Therefore, the distribution of number of batches in the queue is equivalent to the QLD of the M/M/1 queue with pmf given by

$$P(number\ of\ batches\ in\ queue = n) = (1 - \upsilon)\upsilon^n, n = 0,1,2 \ldots \quad \text{(A7)}$$

where $\upsilon = \lambda_B/\mu_B = \lambda_{ind}/\mu_{ind}$.

At any given instant, the total number of individual customers in the latter queue is simply the sum of the total numbers of customers in each of the batches in the buffer and the residual number of customers of the batch in service. Owing to the memoryless property of the geometric distribution, the distribution of the number of residual customers of the batch in service is also geometric with the same mean, $(1/\sigma)$. Therefore, the distribution of total number of individual customers in the GE/M/1 queue is the random sum of i.i.d. geometric RV's, each with mean $(1/\sigma)$, bounded by the modified geometric defined by (A7) and the result is the GGeo QLD defined by (A4). This corresponds to QLD of the GE/M/1 queue derived in (El-Affendi and Kouvatsos 1983) via the ME approach.

Appendix B: Definitions of the Morse average measure of impatience, α

This section presents an alternative approach to determine the value of the Morse average measure of impatience, α based on equivalence between the Morse and Haight balking paradigms.

B1. The Haight balking paradigm

An aspect which affects the decision of a prospective arrival to join a queue or not is the importance of receiving service. The level of importance falls within the range from minimal need for service (where a non-empty queue will not be joined) to extreme urgency for service (where queues of all occupancy levels are joined). This concept was introduced in (Haight 1957) and was modelled as follows: Considering his/her desire for service prior to arrival, each customer selects a queue size, B, above which he/she will balk from the queue according to the balking distribution, $P(B = n), n = 0,1,2$ ....

Within the context of the M/M/1 queue, Haight devised a method to derive the queue length-dependent joining function, $q(n), n = 0,1,2$ ... from the balking distribution as follows: By definition, a customer only joins the queue when his/her balking threshold is above the instantaneous queue length. This can be stated probabilistically as

$$P(a\ customer\ joins\ when\ instantaneous\ queue\ length = n), n = 0,1,2 \ldots$$
$$= P(the\ customer's\ balking\ threshold\ \geq n+1)\,. \tag{B1}$$

Relationship (B1) can in turn be represented symbolically as

$$\begin{aligned} q(n) &= P(B \geq n+1), n = 0,1,2 \ldots \\ &= 1 - P(B \leq n) \\ &= 1 - \sum_{k=0}^{n} P(B = k)\,. \end{aligned} \tag{B2}$$

Specifically, it can be shown that when the balking distribution is modified geometric with pmf given by

$$P(B = n) = \left(\frac{E[B]}{1+E[B]}\right)^{n}\left(1 - \frac{E[B]}{1+E[B]}\right), n = 0,1,2 \ldots \tag{B3}$$

where $E[B]$, the mean of the balking distribution, is the mean instantaneous queue size tolerated by a customer population above which customers will balk from the M/M/1 queue, then the corresponding joining function can be derived as (Haight 1957)

$$q(n) = 1 - \sum_{k=0}^{n} \left(\frac{E[B]}{1+E[B]}\right)^{k} \left(1 - \frac{E[B]}{1+E[B]}\right), n = 0,1,2 \ldots$$

$$= \left(\frac{E[B]}{1+E[B]}\right)^{n}. \quad \text{(B4)}$$

B2. The relationship between the Haight and Morse balking paradigms

The expressions for $q(n)$ in (B4) and (2.31) are observed to be of the same form implying equivalence between balking according to the modified geometric balking distribution and Morse balking. Owing to this latter equivalence, α in the context of the M/M/1 queue subject to Morse balking can be expressed as

$$\alpha = -\mu \ln\left(\frac{E[B]}{1+E[B]}\right) \quad \text{(B5)}$$

where μ is the mean service rate. Hence it is deduced that in general, α can be determined quantitatively in terms of μ and appropriate moments of the balking distribution. Indeed, in the case of Poisson arrivals subject to extended Morse balking at an infinite-capacity queue characterised by i.i.d. general (including GE) service times (i.e. the M/G/1 (including M/GE/1)

queue subject to extended Morse balking), the balking distribution generalises to the GGeo as follows.

Haight (Haight 1957) devised a method to determine the balking distribution from the corresponding joining function as follows:

$$P(B = n) = q(n) - q(n+1)\,. \tag{B6}$$

Applying (B6) to the extended Morse joining function (4.9), the corresponding balking distribution is found to be

$$P(B = n) = \begin{cases} 1 - e^{-\frac{\alpha}{\mu}\left(\frac{C_S^2+1}{2}\right)}, n = 0 \\ e^{-\frac{\alpha}{\mu}\left(\frac{C_S^2+1}{2}\right)}\left(1 - e^{-\frac{\alpha}{\mu}}\right)e^{-\frac{\alpha}{\mu}(n-1)}, n = 1,2,3\ldots \end{cases} \tag{B7}$$

where C_S^2 is the SCOV of the service time distribution. (B7) is seen to be the GGeo pmf, which can equivalently be expressed in terms of its mean and SCOV by (A5). By this latter equivalence, α in the context of the M/G/1 queue subject to extended Morse balking can be defined by

$$\alpha = -\mu \ln\left(\frac{(C_B^2 + 1)E[B] - 1}{(C_B^2 + 1)E[B] + 1}\right) \tag{B8}$$

where C_B^2 is the SCOV of the balking threshold values chosen by individual members of a customer population.

The assumptions that the balking distribution has infinite support and is independent of finite queue capacity, K, render the formulae for α, (B5) and (B8), valid for finite-capacity queues.

B3. The batch prospective arrival case

In the case of a general batch prospective arrival process subject to population-dependent balking, it is assumed that all the members of a batch join (or balk from) the queue with identical population-dependent probabilities due to each 'seeing' the same instantaneous queue size (cf., Section 5.1).

Under complete batch balking with complete or partial batch blocking (i.e. Policies I or II), members of an arriving batch elect to either join or balk from the queue uniformly as a group. As a consequence of the above assumption, the correspondence of this scenario in the Haight paradigm is that members of an arriving batch subject to population-dependent balking select their common balking threshold, B_B, after forming the batch. B_B is the common queue size chosen collectively by all the members of a batch, above which the batch will balk. The RV, B_B, is distributed according to the batch balking distribution, $P(B_B = n), n = 0,1,2 \ldots$, with mean, $E[B_B]$ and SCOV, $C_{B_B}^2$.

Following the method used above to derive the balking distribution of the M/G/1 (including M/GE/1) queue subject to extended Morse balking (cf., (B6)

- (B8)), it is clear that the batch balking distribution associated with a general batch prospective arrival process subject to extended Morse balking under Policies I or II, is also the GGeo pmf. However, in this latter context, the parameter α is re-interpreted as the Morse average measure of impatience of customer batches, α_B (cf., Section 5.1.1), defined by

$$\alpha_B = -\mu \ln\left(\frac{\left(C_{B_B}^2 + 1\right)E[B_B] - 1}{\left(C_{B_B}^2 + 1\right)E[B_B] + 1}\right). \tag{B9}$$

Under independent batch balking with partial batch blocking (i.e. Policy III), members of a batch prospective arrival elect to join or balk autonomously, albeit with the same joining or balking probabilities. This balking operation can be seen to have correspondence in the Haight paradigm when each member of a batch choses his/her balking threshold, B, individually. Bearing in mind the above assumption that all the members of a batch prospective arrival join (or balk from) the queue with identical probabilities, following from (B1), the following equality holds for a batch prospective arrival of size $k = 1,2,3$ ...:

$$P(j\ from\ k\ customers\ join\ when\ the\ instantaneous\ queue\ length = n), n = 0,1,2 \ldots$$
$$= P\begin{pmatrix} j\ customers\ have\ balking\ thresholds\ \geq (n+1) \\ and\ (k-j)\ customers\ have\ balking\ thresholds\ \leq \mathrm{n} \end{pmatrix}. \quad \text{(B10)}$$

Due to the independence between members of a batch, (B10) can be evaluated as

$$\binom{k}{j} q(n)^j \big(1-q(n)\big)^{k-j} = \binom{k}{j} \Big(P\big(B \geq (n+1)\big)\Big)^j \Big(1-P\big(B \geq (n+1)\big)\Big)^{k-j}. \quad \text{(B11)}$$

Hence, $q(n) = P\big(B \geq (n+1)\big)$ and consequently in the context of a general batch prospective arrival process subject to extended Morse balking under Policy III and i.i.d. general service times, α can be defined by (B8) above.

The assumptions employed above, namely that the balking distribution has infinite support and is independent of finite queue capacity, K, render the formulae for α, in this context of general batch prospective arrival processes subject to extended Morse balking under Policies I, II or III and i.i.d. general service times, valid for finite-capacity queues.

Appendix C: The GE/GE/1/K queue subject to balking under Policy IV

In this section a fourth batch balk-block policy, Policy IV is introduced and described. Subsequently, the upward state transition rates of the Markov chain model of the GE/GE/1/K queue subject to population-dependent balking under Policy IV are presented.

The definition of Policy IV is motivated by the proposed investigation into the performance of the GE/GE/1/K queue subject to balking, relative to that of equivalent queues with balking characterised by two-phase exponential distributions, such as the H_2 or Coxian-2. These two-phase distributions characterise inter-arrival times of bursty single arrivals or bursty service times of single customers.

Consider a single server, finite-capacity queue subject to population-dependent balking where the prospective arrival process is characterised by a two-phase exponential distribution. In such a queueing system, each successive arrival, albeit part of a burst, could 'see' a greater instantaneous queue length due to the potential joining of preceding arrivals. Therefore, in order to compare the performance of such a queueing system with that of an equivalent GE/GE/1/K queue subject to balking, Policy IV is devised.

Coinciding with Policy III, under Policy IV, members of a batch behave independently with respect to joining/balking. However, each successive member of a batch, from the head onwards, is assumed to 'see' the potentially updated instantaneous queue length following the potential joining

of preceding members of its batch. Furthermore, as in Policy III, the blocking is of the partial batch type.

The downward state transition rates of the Markov chain model of the GE/GE/1/K queue subject to population-dependent balking under Policy IV are given by equations (4.4) and (4.5). The upward state transition rates, R_{ij}'s, are given by

$$R_{ij} = \binom{prospective\ batch}{arrival\ rate} \times P\begin{pmatrix} batch\ arrival\ is\ of\ size\ k, \\ k = j-i, j-i+1, j-i+2, \dots \infty \\ and\ (j-i)\ from\ the\ k\ customers \\ join\ the\ queue\ initially\ of\ length\ i \end{pmatrix},$$

$$i = 1,2,3 \dots K-1, j = i+1, i+2, i+3, \dots K$$

$$= (\sigma\lambda)\left(\prod_{r=i}^{j-1} q(r)\right)\sum_{k=j-i}^{\infty}\sigma(1-\sigma)^{k-1}\sum_{\substack{m_i, m_{i+1} \dots m_j: \\ m_i + m_{i+1} + \cdots + m_j = k-(j-i)}}\left(\prod_{p=i}^{j}\left(1-q(p)\right)^{m_p}\right), \quad \text{(C1)}$$

$$0.0 < q(n) < 1.0$$

where the $m_i, m_{i+1}, \dots, m_j$'s are the summands comprising compositions resulting from $\left(k-(j-i)\right)$ into $(j-i+1)$ parts and

$$R_{0j} = \binom{prospective\ batch}{arrival\ rate} \times P\begin{pmatrix} batch\ arrival\ is\ of\ size\ k, k = j, j+1, j+2 \dots \infty, of\ which \\ (j+l), l = 0,1,2 \dots (k-j)\ customers\ join\ and\ (k-j-l)\ balk. \\ Of\ those\ (j+l)\ which\ join, the\ first\ l\ exit\ the\ queue\ via\ the \\ zero\ branch\ of\ the\ service\ process\ and\ the\ remaining\ j \\ customers\ opt\ to\ reside\ in\ the\ queue. \end{pmatrix}, \quad \text{(C2)}$$

$$j = 1,2,3 \dots K$$

$$= (\sigma\lambda)\left((q(0)^{l})\prod_{r=0}^{j-1} q(r)\right)\sum_{k=j}^{\infty}\sum_{l=0}^{k-j}(1-\tau)^{l}\tau\sigma(1-\sigma)^{k-1}\sum_{\substack{m_0,m_1\ldots m_j:\\ m_0+m_1+\cdots+m_j=k-j-l}}\left(\prod_{p=0}^{j}\left(1-q(p)\right)^{m_p}\right),$$

$$0.0 < q(n) < 1.0$$

where the $m_0, m_1, \ldots, m_j$'s are the summands of compositions of $(k-j-l)$ into $(j+1)$ parts.

The above R_{ij}'s of the GE/GE/1/K queue subject to population-dependent balking under Policy IV, (C1) and (C2), reduce to the following special cases (which correspond to those special cases considered for Policies I-III in Section 5.2):

- Setting $K \to \infty$ yields the R_{ij}'s of the infinite-capacity GE/GE/1 queue subject to independent batch balking of the Policy IV-type, where each successive member of a batch 'sees' a potentially updated instantaneous queue length.
- Putting $C_a^2 = 1.0$ yields the R_{ij}'s of the M/GE/1/K queue subject to population-dependent balking,
- Setting $C_s^2 = 1.0$ results in the R_{ij}'s of the GE/M/1/K queue subject to balking under Policy IV and
- Specifying $q(n) = 1.0, n = 0,1,2 \ldots K-1$ yields the R_{ij}'s of the ordinary GE/GE/1/K queue subject to partial batch blocking.

It is expected, in light of results from experiments carried out for the corresponding ordinary queues, cf., (Kouvatsos 1988; Kouvatsos and Tabet-Aouel 1994), that for $C_a^2, C_s^2 > 1.0$, the GE/GE/1/K queue subject to balking

under Policy IV would give pessimistic performance bounds over the equivalent two-phase queueing systems.

Appendix D: An approximation of the departure process from the M/G/1/K queue subject to balking

In this section, an approximation of the departure process from the M/G/1/K queue subject to population-dependent balking is formulated. This involves the exact analysis of the inter-departure time distribution of the queueing system.

In queueing networks, the departures from one queue comprise the prospective arrivals to its downstream queue(s). Hence, for the analysis of QNM's it is necessary to study the departure process from queues. It is well known that in general, departure processes of infinite and finite capacity M/G/1 queues are non-renewal point processes with correlated inter-departure intervals, except in a few special cases such as the M/M/1, M/G/1/1 and M/D/1/2 queues. Furthermore, the consequential dependence between inter-arrival times at downstream queues adversely affects their performance (Whitt 1984; Bertsimas and Nakazato 1990; Hu 1996; Takagi and Nishi 1998).

The complicated analysis of departure processes has spurred the proposals of approximations such as modelling the (non-renewal) queue output processes with renewal ones with matching moments. The choice of the renewal process interval distribution is motivated in part by analytic

tractability (Kuehn 1979; Whitt 1982). In (Whitt 1982), two methods to determine the moments of the renewal interval are identified and applied, namely the stationary interval and asymptotic methods. In the former method, the moments of a single stationary interval of the non-renewal process are used. In the latter case, the moments of the renewal interval are obtained by matching the asymptotic behaviour of the moments of partial sums of successive intervals of both processes.

The asymptotic method was proposed as a means to account for the underlying correlation between inter-departure intervals. This correlation is not captured in the stationary interval method, nonetheless, this latter procedure is employed here as a preliminary step to approximating the departure process from the M/GE/1/K queue with balking. Furthermore, network blocking mechanisms are not modelled at this early stage. However, these can be incorporated from works such as (Kouvatsos et al. 2011) where arbitrary open QNM's with network blocking mechanisms and GE-type flows have been analysed approximately within the ME framework.

In (Takagi and Nishi 1998), the Laplace-Stieltjes transform (LST) of the inter-departure time distribution of the ordinary M/G/1/K queue is presented. It is seen to be comprised of the mixture of the general service time density and the convolution of the general service time density and the exponential inter-arrival time density[15] following the emptying of the system. The inter-departure time distribution of the M/G/1/K queue subject to population-dependent balking can be seen to have the same form as that of the ordinary

[15] Due to the memoryless property of the exponential distribution, the density of residual inter-arrival times (resulting from the emptying of the queue) is equal to the inter-arrival time density.

M/G/1/K queue, since, while the queue is occupied the service process is unaffected by the effective arrival process. By definition, the service process is independent of the prospective arrival process. Furthermore, if customers balk with a fixed probability when the queue is empty, then the effective arrival process during these periods is still Poisson, albeit with a revised rate.

Consider an M/G/1/K queue characterised by Poisson prospective arrivals (with mean rate λ) subject to state-dependent balking with probability $\big(1-q(n)\big), n = 0,1,2 \ldots K-1$, i.i.d. general service times (with mean rate, μ and SCOV, C_s^2) and finite capacity, K. Let the service time density, $b(t)$ have LST, $B^*(s)$, where $B^*(s) = \int_0^\infty e^{-st} b(t)dt$. Moreover, let the stationary inter-departure time density function, $d(t)$ have LST, $D^*(s)$. Following (Takagi and Nishi 1998), the LST of the stationary inter-departure time density can be defined by

$$D^*(s) = \pi_o^{K-1}\left(\frac{\lambda q(0)}{s+\lambda q(0)}\right)B^*(s) + (1-\pi_o^{K-1})B^*(s) \qquad \text{(D1)}$$

where π_o^{K-1} is the steady state probability that a (single) departing customer leaves behind an empty queue and can be computed from its derivation given in subsection D1 below. For GE (batch) service, an equivalent H_2 distribution can be used to obtain π_o^{K-1} approximately. Details of how such an equivalent H_2 distribution can be generated are presented in the Appendix of (Kouvatsos 1988).

The mean, $E[D]$ and SCOV, C_D^2 of $d(t)$ are given by

$$E[D] = \frac{1}{\mu} + \frac{\pi_o^{K-1}}{\lambda q(0)} \tag{D2}$$

and

$$C_D^2 = \left(\frac{\left(\lambda q(0)\right)^2 (C_s^2 + 1) + 2\mu^2 \pi_0^{K-1} \left(1 + (\lambda q(0)/\mu)\right)}{\lambda q(0) + \mu \pi_0^{K-1}} \right) - 1 \,. \tag{D3}$$

The ME approximate solution of arbitrary open QNM's devised in (Kouvatsos 1994; Kouvatsos et al. 2011) and references therein, advocates decomposing the QNM's into individual queueing stations and analysing these in isolation with GE-type (overall and effective) inter-arrival and service times. Furthermore, the stationary compound Poisson process with geometrically-distributed batch sizes (and thus underlying GE inter-event time distribution) is a renewal process (Kouvatsos 1994).

Therefore, it is proposed that the above stationary departure interval moments (D2) and (D3) be used to characterise an appropriate GE prospective inter-arrival time distribution (2.20) to the downstream queue(s) and thus approximate the departure process with an appropriate renewal one towards the ME approximate analysis of QNM's with balking.

The accuracy of the above approximation may be assessed by comparing analytic performance results against simulation of a downstream queue in an appropriate tandem queueing system as carried out in (Whitt 1984) and

references therein. Such a tandem system may comprise the infinite-capacity M/GE/1 queue subject to extended Morse balking connected to a downstream infinite-capacity GE/GE/1 queue subject to extended Morse balking under either the complete or independent batch balking policies. The infinite capacity cases can be modelled by setting K to an appropriately large value. Following the incorporation of network blocking mechanisms into the model, the accuracy of the approximation in the context of analogous finite-capacity queues can be assessed.

D1. The QLD from the viewpoint of customers departing from the M/G/1/K queue subject to balking

The stationary QLD from the point of view of departing customers from the M/G/1/K queue subject to population-dependent balking is derived below.

The stochastic process with state description, $N(t)$, modelling the number of customers present at time t, in an M/G/1 (and more general) queueing system(s) is insufficient to summarise the complete past history of the system. This renders the stochastic process $N(t)$ non-Markovian. However, in such stochastic processes, epochs of the Markovian type (referred to in the literature as regeneration points) may occur. A regeneration point possesses the beneficial characteristic that the knowledge of the past history of the process at that particular epoch has no predictive value. Within the context of the M/G/1 queue, these regeneration points occur at the instants immediately after departures. The instantaneous queue population at future

departure epochs can be derived from the current value and the number of arrivals which subsequently join the queue. Values of the instantaneous queue population at past departure epochs are not required. Thus, the progression of instantaneous queue population at these departure epochs comprises a discrete-parameter Markov chain embedded within the non-Markovian process, $N(t)$. Fortunately, the state probability distribution of this embedded Markov chain (EMC) is also the QLD of the M/G/1 queue from the point of view of departing customers (Kendall 1951).

The stochastic process, $N(t)$, modelling the number of customers over time in the M/G/1/K queue subject to population-dependent balking, despite being non-Markovian, can also be seen to possess regeneration points at the instants immediately following departures. The instantaneous queue population at future departure epochs can be derived from the value at the current departure epoch and the number of subsequent effective arrivals (i.e. the number of prospective arrivals which join the queue noting that some potentially balk or are blocked). Therefore, the evolution of instantaneous queue length at the departure epochs can be modelled by a discrete-parameter Markov chain. The state probability distribution of the Markov chain embedded at the departure epochs, $\pi_n^{K-1}, n = 0,1,2 \dots K-1$ provides the QLD from the point of view of departures from the M/G/1/K queue subject to population-dependent balking. A combinatorial approach is taken to determine the state-transition probabilities of the EMC, as carried out also in (Dick 1970). In the latter work, the M/G$^{[a,b]}$/1 queue with batch service and Poisson prospective arrivals subject to balking characterised by a constant

probability is analysed. The population-dependent balking model analysed below therefore extends that of (Dick 1970).

The stationary QLD from the point of view of departures from the equivalent M(n)/G/1/K queue with a state-dependent Poisson arrival process has been solved via the supplementary variable technique in (Gupta and Rao 1996). Furthermore, the stationary QLD from the point of view of departures from the M(n)/G(n)/1/K queue with state-dependent arrival and service rates and queue length–dependent service times has been analysed in (Courtois and Georges 1971). In the latter work, the EMC approach is used with results from renewal theory. Therefore, the analysis of π_n^{K-1} below presents an alternative solution to a resolved problem.

Consider an M/G/1/K queue characterised by Poisson prospective arrivals (with mean rate λ) subject to population-dependent balking with function $\big(1-q(n)\big), n = 0,1,2 \dots K-1$, i.i.d. general service times (with density $b(t)$, mean rate, μ and SCOV, C_s^2) and finite capacity, K.

Additional notation used in the derivation is defined at the outset:

C_r denotes the r^{th} customer to join the queue

x_r represents the service duration of C_r

d_r is the number of customers left behind by C_r on departure from the queue

v_r symbolises the number of prospective arrivals which join the queue during service of C_r

The state-transition probabilities, p_{ij}'s, of the discrete-parameter Markov chain modelling the number of customers in the queue at departure epochs are one-step transition probabilities of the right stochastic matrix, $\boldsymbol{P}$ given by

$$\boldsymbol{P} = \begin{bmatrix} p_{00} & p_{01} & & p_{0,K-1} \\ p_{10} & p_{11} & & p_{1,K-1} \\ & & \ddots & \\ & 0 & & p_{K-2.K-1} \\ & & & p_{K-1,K-1} \end{bmatrix},$$

where

$$p_{ij} = \begin{cases} P(d_{r+1} = j/d_r = i), i = 0, j = i, i+1, i+2, \dots, K-1 \\ P(d_{r+1} = j/d_r = i), i = 1,2,3 \dots K-1, j = i-1, i, i+1, \dots K-1 . \\ 0, \quad j < i-1, \text{ since only single departures occur} \end{cases} \quad \text{(D4)}$$

Immediately following the departure of C_{r+1}, the number of customers left behind in the queue, d_{r+1}, can be obtained from the number of customers left behind immediately following the departure of C_r and the number of prospective arrivals which join the queue during the service of C_{r+1}, as follows:

$$d_{r+1} = \begin{cases} \nu_{r+1}, i = 0 \\ d_r + \nu_{r+1} - 1, i = 1,2,3 \dots K-1 \end{cases}. \quad \text{(D5)}$$

Notably, when C_r leaves behind an empty system, C_{r+1} joins the queue next and proceeds straight to service. The number of customers left behind by C_{r+1} is simply the number of customers which join the queue during its service i.e. ν_{r+1}. Therefore, it is seen that C_{r+1} leaving behind ν_{r+1} customers on departure from the queue occurs exclusively from either one of two preceding states: $d_r = 0$ or $d_r = 1$.

Heeding the dependence of the ν_{r+1}'s on i due to population-dependent balking, and incorporating the expressions for d_{r+1} in (D5), the p_{ij}'s can be defined as

$$p_{ij} = \begin{cases} P(\nu_{r+1} = j/d_r = 0), i = 0, j = i, i+1, i+2, \dots, K-1 \\ P(\nu_{r+1} = j - i + 1/d_r = i), i = 1,2,3 \dots K-1, j = i-1, i, i+1, \dots K-1 \\ 0, \quad j < i-1, \end{cases}. \quad \text{(D6)}$$

The probability of a transition from either $d_r = 0$ or $d_r = 1$ to $d_{r+1} = \nu_{r+1}$ is simply the probability that ν_{r+1} prospective arrivals join the queue during the service of C_{r+1}. Therefore, the p_{0j}'s and corresponding p_{1j}'s are equal as expressed in (D6) above and consequently the derivations of p_{0j} will be omitted henceforth.

Since all the customers are statistically identical, the v_{r+1}'s are identically distributed. Hence, following the substitution $(j - i + 1) = k$ in (D6), the p_{ij}'s can be represented by the state transition probabilities

$$\alpha_{i,k} = P(v_{r+1} = k/d_r = i), i = 1,2,3 \ldots K-1; k = 0,1,2 \ldots K-i\,. \qquad \text{(D7)}$$

Now $\boldsymbol{P}$ can equivalently be populated with the $\alpha_{i.k}$'s as follows:

$$\boldsymbol{P} = \begin{bmatrix} \alpha_{1,0} & \alpha_{1,1} & & \cdots & & \alpha_{1,K-1} \\ \alpha_{1,0} & \alpha_{1,1} & & & & \alpha_{1,K-1} \\ 0 & \alpha_{2,0} & \alpha_{2,1} & & & \\ 0 & & \alpha_{3,0} & & & \\ & & & \ddots & & \\ & & 0 & \alpha_{K-2,0} & \alpha_{K-2,1} & \alpha_{K-2.2} \\ & & & 0 & \alpha_{K-1,0} & \alpha_{K-1,1} \end{bmatrix}.$$

By definition, the Poisson prospective arrival process is independent of the customer number, r, the queue size (and hence d_r) and the service process. Moreover, the effective arrival process is also independent of the service process (Courtois and Georges 1971). However, the number of prospective arrivals which join the queue, v_{r+1}, during the service of C_{r+1} depends on the length of its service duration, x_{r+1}, and the number of customers left in the queue immediately following the departure of C_r. The probability of k prospective arrivals joining the queue over a service-time duration, t,

conditional on i customers left in the queue immediately following the last departure, can be derived by applying the theorem of total probability yielding

$$\alpha_{i,k} = \int_0^{\infty} P\big(v_{r+1} = k/(x_{r+1} = t, d_r = i)\big) b(t) dt, \tag{D8}$$

$$i = 1,2,3 \dots K-1; k = 0,1,2 \dots K-i\,.$$

The conditional probability of k joining from all the possible numbers of l prospective arrivals during the service duration, $x_{r+1} = t$, given that the first prospective arrival during x_{r+1} finds i customers in the queue can be determined by applying again the theorem of total probability giving

$$P\big(v_{r+1} = k/(x_{r+1} = t, d_r = i)\big) = \sum_{l=k}^{\infty} P(k\ join\ from\ l\ during\ x_{r+1}/d_r = i) \frac{\lambda t^l}{l!} e^{-\lambda t}, \tag{D9}$$

$$i = 1,2,3 \dots K-1; k = 0,1,2 \dots K-i$$

where the conditional probability $P(k\ join\ from\ l\ during\ x_{r+1}/d_r = i), i = 1,2,3 \dots K-1$ models all the different permutations of k joining and $(l-k)$ balking from the l prospective arrivals. For illustrative purposes, the latter conditional probability is derived below for two examples.

$$P(1\ joins\ from\ 3\ during\ x_{r+1}/d_r = i)$$

$$= q(i)\big(1-q(i+1)\big)^2 + \big(1-q(i)\big)q(i)\big(1-q(i+1)\big)$$
$$+ \big(1-q(i)\big)^2 q(i) \qquad \text{(D10)}$$
$$= q(i)\Big(\big(1-q(i+1)\big)^2 + \big(1-q(i)\big)\big(1-q(i+1)\big)$$
$$+ \big(1-q(i)\big)^2\Big)$$

and

$$P(2\ join\ from\ 4\ during\ x_{r+1}/d_r = i)$$

$$= q(i)q(i+1)\big(1-q(i+2)\big)^2$$
$$+ q(i)\big(1-q(i+1)\big)q(i+1)\big(1-q(i+2)\big)$$
$$+ q(i)\big(1-q(i+1)\big)^2 q(i+1)$$
$$+ \big(1-q(i)\big)q(i)q(i+1)\big(1-q(i+2)\big)$$
$$+ \big(1-q(i)\big)q(i)\big(1-q(i+1)\big)q(i+1) \qquad \text{(D11)}$$
$$+ \big(1-q(i)\big)^2 q(i)q(i+1)$$
$$= q(i)q(i+1)$$
$$\times \Big(\big(1-q(i+2)\big)^2 + \big(1-q(i+1)\big)\big(1-q(i+2)\big)$$
$$+ \big(1-q(i+1)\big)^2 + \big(1-q(i)\big)\big(1-q(i+2)\big)$$
$$+ \big(1-q(i)\big)\big(1-q(i+1)\big) + \big(1-q(i)\big)^2\Big).$$

These permutations can be represented collectively by the expression

$$P(k\ join\ from\ l\ during\ x_{r+1}/d_r = i) = \left(\prod_{n=i}^{i+k-1} q(n)\right)\left(\sum_{\substack{m_i, m_{i+1} \ldots m_{i+k}: \\ m_i + m_{i+1} + \cdots + m_{i+k} = (l-k)}} \left(\prod_{p=i}^{i+k} \left(1 - q(p)\right)^{m_p}\right)\right), \quad \text{(D12)}$$

$$i = 1,2,3 \ldots K-1; k = 0,1,2 \ldots K-i, q(K) = 0.0$$

where the $m_i, m_{i+1}, \ldots, m_{i+k}$'s are the summands comprising compositions resulting from $(l-k)$ into $(k+1)$ parts. When $k = 0$, $\prod_{n=i}^{i+k-1} q(n) = 1.0$.

Therefore, the state transition probabilities can be defined by

$$\alpha_{i,k} = \int_0^\infty \sum_{l=k}^{\infty} \left(\prod_{n=i}^{i+k-1} q(n)\right)\left(\sum_{\substack{m_i, m_{i+1} \ldots m_{i+k}: \\ m_i + m_{i+1} + \cdots + m_{i+k} = (l-k)}} \left(\prod_{p=i}^{i+k} \left(1 - q(p)\right)^{m_p}\right)\right) \frac{\lambda t^l}{l!} e^{-\lambda t}\, b(t) dt, \quad \text{(D13)}$$

$$i = 1,2,3 \ldots K-1; k = 0,1,2 \ldots K-i\ .$$

This completes the derivation of the $\alpha_{i,k}$'s and hence the population of the state transition probability matrix, $\boldsymbol{P}$. The QLD, $\pi_n^{K-1}, n = 0,1,2 \ldots K-1$ can now be computed from both the vector equation $\pi_n^{K-1}(\boldsymbol{P} - \boldsymbol{I}) = 0, n = 1,2,3 \ldots K-1$ and the normalisation condition.

D1.1. Validation of Analysis

In this subsection, the analysis of the QLD, $\pi_n^{K-1}, n = 0,1,2 \dots K-1$, is verified by comparing the results of experiments conducted for the M/G/1/K queue subject to population-dependent balking with analogous ones in (Gupta and Rao 1996). In the latter work, the supplementary variable solution technique was applied to determine the stationary QLD, $P_n^+, n = 0,1,2 \dots K-1$, from the point of view of departing customers from the equivalent M(n)/G/1/K queue (Gupta and Rao 1996).

Two sets of experiments were conducted, one for the special case of the ordinary M/G/1/K queue and the second for a machine interference model represented by the M/G/1/K queue subject to balking characterised by the function $q(n) = (N_s - n), n = 0,1,2 \dots K-1$, where N_s is the size of the source.

The different service-time densities, $b(t)$, employed in the experiments and resulting state transition probabilities of the corresponding EMC's, $\alpha_{i,k}$'s, are defined below in (D14) - (D25). For all the $\alpha_{i,k}$'s below, $i = 1,2,3 \dots K-1$ and $k = 0,1,2 \dots K-i$ and when $k = 0$, $\left(\prod_{n=i}^{i+k-1} q(n)\right) = 1.0$. Furthermore, the $m_i, m_{i+1}, \dots, m_{i+k}$'s are the summands comprising compositions resulting from $(l-k)$ into $(k+1)$ parts.

The Exponential Distribution

$$b(t) = \mu e^{-\mu t}, t > 0, \mu > 0 \tag{D14}$$

$$\alpha_{i,k} = \mu \sum_{l=k}^{\infty} \left(\prod_{n=i}^{i+k-1} q(n) \right) \left(\sum_{\substack{m_i, m_{i+1} \ldots m_{i+k}: \\ m_i + m_{i+1} + \cdots + m_{i+k} = (l-k)}} \left(\prod_{p=i}^{i+k} \left(1 - q(p)\right)^{m_p} \right) \right) \left(\frac{\lambda^l}{(\lambda + \mu)^{l+1}} \right) \tag{D15}$$

The Erlang-r Distribution

$$b(t) = \frac{\mu^r t^{r-1} e^{-\mu t}}{(r-1)!}, t > 0, \mu > 0, r = 1,2,3 \ldots \tag{D16}$$

$$\alpha_{i,k} = \left(\frac{\mu^r}{(r-1)!} \right) \sum_{l=k}^{\infty} \left(\prod_{n=i}^{i+k-1} q(n) \right) \left(\sum_{\substack{m_i, m_{i+1} \ldots m_{i+k}: \\ m_i + m_{i+1} + \cdots + m_{i+k} = (l-k)}} \left(\prod_{p=i}^{i+k} \left(1 - q(p)\right)^{m_p} \right) \right) \left(\frac{\lambda^l (l + r - 1)!}{(\lambda + \mu)^{l+r}} \right) \tag{D17}$$

The Deterministic Distribution

$$b(t) = \begin{cases} 1.0, t = 1/\mu, \mu > 0 \\ 0, \text{ otherwise} \end{cases} \tag{D18}$$

$$\alpha_{i,k} = \sum_{l=k}^{\infty} \left(\prod_{n=i}^{i+k-1} q(n) \right) \left(\sum_{\substack{m_i, m_{i+1} \dots m_{i+k}: \\ m_i + m_{i+1} + \dots + m_{i+k} = (l-k)}} \left(\prod_{p=i}^{i+k} \left(1 - q(p)\right)^{m_p} \right) \right) \left(\frac{\lambda}{\mu}\right)^{l} \left(\frac{e^{-\frac{\lambda}{\mu}}}{l!} \right) \quad \text{(D19)}$$

The r-Phase Hyper-exponential (H_r) Distribution

$$b(t) = \sum_{m=1}^{r} \alpha_m \mu_m e^{-\mu_m t}, t > 0, \mu_m > 0, 0 < \alpha_m < 1, \sum_{m=1}^{r} \alpha_m = 1 \quad \text{(D20)}$$

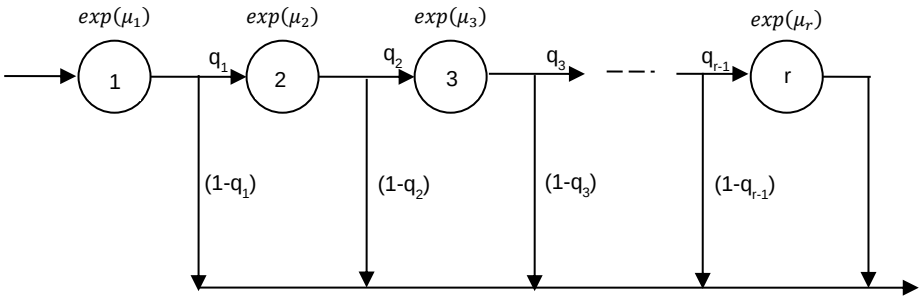
$$\alpha_{i,k} = \sum_{l=k}^{\infty} \left(\prod_{n=i}^{i+k-1} q(n) \right) \left(\sum_{\substack{m_i, m_{i+1} \dots m_{i+k}: \\ m_i + m_{i+1} + \dots + m_{i+k} = (l-k)}} \left(\prod_{p=i}^{i+k} \left(1 - q(p)\right)^{m_p} \right) \right) \lambda^{l} \sum_{m=1}^{r} \left(\frac{\alpha_m \mu_m}{(\lambda + \mu_m)^{l+1}} \right) \quad \text{(D21)}$$

The r-Phase Hypo-exponential (h_r) Distribution

$$b(t) = \sum_{m=1}^{r} \mu_m e^{-\mu_m t} \left(\prod_{n=1, n \neq m}^{r} \frac{\mu_n}{\mu_n - \mu_m} \right), t > 0, \mu_m, \mu_n > 0 \quad \text{(D22)}$$

$$\alpha_{i,k} = \sum_{l=k}^{\infty} \left(\prod_{n=i}^{i+k-1} q(n) \right) \left(\sum_{\substack{m_i, m_{i+1} \dots m_{i+k}: \\ m_i + m_{i+1} + \dots + m_{i+k} = (l-k)}} \left(\prod_{p=i}^{i+k} \left(1 - q(p)\right)^{m_p} \right) \right) \lambda^{l} \sum_{m=1}^{r} \frac{\mu_m}{(\lambda + \mu_m)^{l+1}} \left(\prod_{n=1, n \neq m}^{r} \frac{\mu_n}{\mu_n - \mu_m} \right) \quad \text{(D23)}$$

The r-Phase Coxian (C_r) Distribution



$$b(t) = \sum_{m=1}^{r}\left(\prod_{n=1}^{m-1} q_n\right)(1-q_m)\sum_{s=1}^{m}\mu_s e^{-\mu_s t}\left(\prod_{u=1,u\neq s}^{m}\frac{\mu_u}{\mu_u-\mu_s}\right), \tag{D24}$$

$$t>0, \mu_s, \mu_u > 0, 0 < q_m, q_n < 1, \prod_{n=1}^{0} q_n = 1.0, (1-q_r) = 1.0$$

$$\alpha_{i,k} = \sum_{l=k}^{\infty}\left(\prod_{n=i}^{i+k-1} q(n)\right)\left(\sum_{\substack{m_i, m_{i+1}\ldots m_{i+k}:\\ m_i+m_{i+1}+\cdots+m_{i+k}=(l-k)}}\left(\prod_{p=i}^{i+k}(1-q(p))^{m_p}\right)\right)\lambda^l\sum_{m=1}^{r}\left(\prod_{n=1}^{m-1} q_n\right)(1-q_m)\sum_{s=1}^{m}\frac{\mu_s}{(\lambda+\mu_s)^{l+1}}\left(\prod_{u=1,u\neq s}^{m}\frac{\mu_u}{\mu_u-\mu_s}\right) \tag{D25}$$

The experiments were carried out in MATLAB version 7.10.0.499 (R2010a). Due to the high computational demands, the maximum number of prospective Poisson arrivals permitted per service duration, in all the experiments, was set to 15 i.e. $l = k, k+1, k+2, \ldots 15$. The results of the experiments are presented in **Table 4** and **Table 5** below.

Table 4. Stationary QLD's from the point of view of departing customers from the ordinary M/G/1/11 queue with the exponential, deterministic or H_2 service-time distributions, where $q(n) = 0.999999, n = 0,1,2 \ldots 10$ to model the case of no balking.

n	Exponential		Deterministic		H_2	
	$\lambda = 1.0$ $\mu = 5.0$		$\lambda = 1.0$ $\mu = 2.0$		$\lambda = 1.0$ $\mu_1 = 0.8, \mu_2 = 0.4$ $\alpha_1 = 0.4, \alpha_2 = 0.6$	
	P_n^+	π_n^{10}	P_n^+	π_n^{10}	P_n^+	π_n^{10}
0	0.800000	0.800000	0.500001	0.500002	0.000787	0.000787
1	0.160000	0.160000	0.324361	0.324361	0.001466	0.001466
2	0.032000	0.032000	0.122600	0.122600	0.002772	0.002772
3	0.006400	0.006400	0.037788	0.037788	0.005263	0.005263
4	0.001280	0.001280	0.010909	0.010909	0.010007	0.010007
5	0.000256	0.000256	0.003107	0.003107	0.019038	0.019038
6	0.000051	0.000051	0.000884	0.000884	0.036223	0.036223
7	0.000010	0.000010	0.000252	0.000252	0.068924	0.068924
8	0.000002	0.000002	0.000072	0.000072	0.131147	0.131147
9	0.000000	0.000000	0.000020	0.000020	0.249544	0.249544
10	0.000000	0.000000	0.000006	0.000006	0.474830	0.474829

Table 5. Stationary QLD's from the point of view of departing customers from the M/G/1/5 queue subject to balking according to the function $q(n) = (N_s - n), n = 0,1,2 \ldots 4$, where N_s is the size of the source. Service time distributions used are the exponential, Erlang-10, deterministic, H_2, h_4, C_2 and Erlang-15.

$N_s = 5$

n	Exponential		Erlang-10		Deterministic	
	$\lambda = 1.0$ $\mu = 5.0$		$\lambda = 1.0$ $\mu = 2.0$		$\lambda = 1.0$ $\mu = 2.0$	
	P_n^+	π_n^4	P_n^+	π_n^4	P_n^+	π_n^4
0	0.398343	0.398342	0.026605	0.026605	0.019762	0.019762
1	0.318674	0.318674	0.138124	0.138124	0.126259	0.126259
2	0.191205	0.191204	0.330395	0.330395	0.336035	0.336035

3	0.076482	0.076482	0.366209	0.366209	0.381139	0.381139
4	0.015296	0.015298	0.138668	0.138668	0.136806	0.136806

$N_s = 5$

	H_2		h_4		C_2	
	$\lambda = 1.0$ $\mu_1 = 0.53, \mu_2 = 1.97$ $\alpha_1 = 0.21, \alpha_2 = 0.79$		$\lambda = 4.0$ $\mu_1 = 20, \mu_2 = 40$ $\mu_3 = 60, \mu_4 = 120$		$\lambda = 0.65$ $\mu_1 = 1.5, \mu_2 = 2.0$ $q_1 = 0.5$	
	P_n^+	π_n^4	P_n^+	π_n^4	P_n^+	π_n^4
0	0.050429	0.050512	0.090485	0.090484	0.048291	0.049217
1	0.126482	0.126564	0.236853	0.236852	0.136739	0.138304
2	0.245264	0.245111	0.334693	0.334691	0.272483	0.273456
3	0.331082	0.330345	0.257343	0.257342	0.342539	0.341076
4	0.246742	0.247469	0.080626	0.080631	0.199948	0.197948

$N_s = 40$

	Exponential		Erlang-15		H_2	
	$\lambda = 1.0$ $\mu = 5.0$		$\lambda = 1.0$ $\mu = 10.0$		$\lambda = 1.0$ $\mu_1 = 0.37, \mu_2 = 3.63$ $\alpha_1 = 0.09, \alpha_2 = 0.91$	
	P_n^+	π_n^4	P_n^+	π_n^4	P_n^+	π_n^4
0	0.000273	0.000273	0.000002	0.000002	0.000055	0.000055
1	0.002128	0.002128	0.000067	0.000067	0.000649	0.000652
2	0.016174	0.016174	0.001782	0.001782	0.007459	0.007475
3	0.119686	0.119686	0.043128	0.043128	0.08344	0.083527
4	0.861739	0.861739	0.955021	0.955021	0.908396	0.908291

The results obtained from the stochastic analysis above are seen to closely match those of (Gupta and Rao 1996). The largest error in terms of absolute difference between corresponding probabilities, $|\pi_n^{K-1} - P_n^+|, n = 0,1,2 \ldots K-1$, was 0.002 and around 82% of the errors lay below 8×10^{-5}.

The discrepancies between the two QLD's, π_n^{K-1} and P_n^+, may be attributed firstly to the restriction of a maximum of 15 Poisson prospective arrivals per service duration in the analysis of π_n^{K-1} above. Secondly, errors may have arisen due to limitations of the software package in carrying out the matrix inversion to determine the QLD, π_n^{K-1}.

And therein lies the main drawback of the above solution of π_n^{K-1}, namely the high computational demands of its implementation. This drawback can be overcome by using a different solution approach such as the supplementary variable technique.

References

Al-Seedy, R. O. (1995). "The truncated queue: M/M/2/m/m + Y with balking spares, machine interference and an additional server for longer queues (Krishnamoorthi discipline)." Microelectronics Reliability **35**(11): 1423-1427.

Al-Seedy, R. O. (1996). "Analytical solution of the state-dependent Erlangian queue: M/Ej/1/N with balking." Microelectronics Reliability **36**(2): 203-206.

Ancker, C. J., Jr. and A. V. Gafarian (1963). "Some Queuing Problems with Balking and Reneging. I." Operations Research **11**(1): 88-100.

Arizono, I., Y. Cui, et al. (1991). "An Analysis of M/M/s Queueing Systems Based on the Maximum Entropy Principle." The Journal of the Operational Research Society **42**(1): 69-73.

Artalejo, J. R., I. Atencia, et al. (2005). "A discrete-time Geo[X]/G/1 retrial queue with control of admission." Applied Mathematical Modelling **29**(11): 1100-1120.

Assi, S. A. (2000). An Investigation into Generalised Entropy Optimisation with Queueing Systems Applications. MSc Dissertation, Dept. of Computing, School of Informatics, University of Bradford.

Baccelli, F., P. Boyer, et al. (1984). "Single-Server Queues with Impatient Customers." Advances in Applied Probability **16**(4): 887-905.

Beneš, V. E. (1965). Mathematical Theory of Connecting Networks and Telephone Traffic, Academic Press Inc.

Beran, J., R. Sherman, et al. (1995). "Long-range dependence in variable-bit-rate video traffic." Communications, IEEE Transactions on **43**(2/3/4): 1566-1579.

Bertsimas, D. J. and D. Nakazato (1990) "The Departure Process from a GI/G/1 Queue and its Applications to the Analysis of Tandem Queues", Research Report No. 3275-91, Sloan School of Management, Massachusetts Institute of Technology.

Blackburn, J. D. (1972). "Optimal Control of a Single-Server Queue with Balking and Reneging." Management Science **19**(3): 297-313.

Boxma, O. J. and B. J. Prabhu (2009) "Analysis of an M/G/1 Queue with Customer Impatience and an Adaptive Arrival Process", Research Report No. 2009-028, EURANDOM, Eindhoven University of Technology.

Cantor, J., A. Ephremides, et al. (1986). "Information theoretic analysis for a general queueing system at equilibrium with application to queues in tandem." Acta Informatica **23**(6): 657-678.

Chandy, K. M., U. Herzog, et al. (1975). "Approximate Analysis of General Queuing Networks." IBM Journal of Research and Development **19**(1): 43-49.

Choudhury, G. L., K. K. Leung, et al. (1994). Resource-Sharing Models with State-Dependent Arrivals of Batches. Research Report, AT&T Bell Laboratories, New Jersey, USA.

Courtois, P. J. and J. Georges (1971). "On a Single-Server Finite Queuing Model with State-Dependent Arrival and Service Processes." Operations Research **19**(2): 424-435.

Cox, D. R. (2006). Principles of Statistical Inference, Cambridge University Press.

Crovella, M. E. and A. Bestavros (1997). "Self-similarity in World Wide Web traffic: evidence and possible causes." Networking, IEEE/ACM Transactions on **5**(6): 835-846.

Csiszár, I. (2008). "Axiomatic Characterizations of Information Measures." Entropy **10**(3): 261-273.

de La Fuente, D. and M. J. Pardo (2009). Development of queueing models with balking and uncertain data using Fuzzy Set Theory. IEEE International Conference on Industrial Engineering and Engineering Management, 2009.

Dick, R. S. (1970). "Some Theorems on a Single-Server Queue with Balking." Operations Research **18**(6): 1193-1206.

Economou, A., A. Gómez-Corral, et al. (2011). "Optimal balking strategies in single-server queues with general service and vacation times." Performance Evaluation **68**(10): 967-982.

El-Affendi, M. A. and D. D. Kouvatsos (1983). "A maximum entropy analysis of the M/G/1 and G/M/1 queueing systems at equilibrium." Acta Informatica **19**(4): 339.

Erramilli, A., O. Narayan, et al. (1996). "Experimental queueing analysis with long-range dependent packet traffic." IEEE/ACM Trans. Netw. **4**(2): 209-223.

Fang, S.-C., J. R. Rajasekera, et al. (1997). Entropy optimization and mathematical programming, Kluwer Academic Publishers, Boston.

Feller, W. (1966). An Introduction to Probability Theory and Its Applications, John Wiley & Sons Inc.

Fisher, R. A. (1935). "The Logic of Inductive Inference." Journal of the Royal Statistical Society **98**(1): 39-82.

Floyd, S. and V. Jacobson (1993). "Random early detection gateways for congestion avoidance." Networking, IEEE/ACM Transactions on **1**(4): 397-413.

Fretwell, R. and D. Kouvatsos (2002). "LRD and SRD traffic: review of results and open issues for the batch renewal process." Performance Evaluation **48**(1–4): 267-284.

Garrett, M. W. and W. Willinger (1994). "Analysis, modeling and generation of self-similar VBR video traffic." SIGCOMM Comput. Commun. Rev. **24**(4): 269-280.

Guiasu, S. (1986). "Maximum Entropy Condition in Queueing Theory." The Journal of the Operational Research Society **37**(3): 293-301.

Guo, P. and P. Zipkin (2007). "Analysis and Comparison of Queues with Different Levels of Delay Information." Management Science **53**(6): 962-970.

Guo, P. and P. Zipkin (2009). "The effects of the availability of waiting-time information on a balking queue." European Journal of Operational Research **198**(1): 199-209.

Gupta, S. M. (1995). "Queueing model with state dependent balking and reneging: its complementary and equivalence." SIGMETRICS Perform. Eval. Rev. **22**(2-4): 63-72.

Gupta, U. C. and T. S. S. S. Rao (1996). "Computing steady state probabilities in λ(n)/G/1/K queue." Performance Evaluation **24**(4): 265-275.

Haight, F. A. (1957). "Queueing with Balking." Biometrika **44**(3/4): 360-369.

Hassin, R. and M. Haviv (1995). "Equilibrium strategies for queues with impatient customers." Operations Research Letters **17**(1): 41-45.

Hassin, R. and M. Haviv (2003). To Queue or Not to Queue: Equilibrium Behaviour in Queueing Systems, Kluwer Academic Publishers.

Havrda, J. and F. Charvát (1967). "Quantification method of classification processes. Concept of structural *α*-entropy." Kybernetika **3**(1): 30-35.

Hu, J.-Q. (1996). "The Departure Process of the GI/G/1 Queue and Its MacLaurin Series." Operations Research **44**(5): 810-815.

Iannaccone, G., M. May, et al. (2001). "Aggregate traffic performance with active queue management and drop from tail." SIGCOMM Comput. Commun. Rev. **31**(3): 4-13.

Jaynes, E. T. (1957). "Information Theory and Statistical Mechanics." Physical Review **106**(4): 620 - 630.

Jaynes, E. T. (1968). "Prior Probabilities." Systems Science and Cybernetics, IEEE Transactions on **4**(3): 227-241.

Jaynes, E. T. (1978). Where do we Stand on Maximum Entropy? The Maximum Entropy Formalism Conference, M.I.T., USA, M.I.T. Press, Cambridge, MA, USA.

Jaynes, E. T. (2003). Probability Theory : The Logic of Science. NY, USA, Cambridge University Press.

Jouini, O., Z. Akşin, et al. (2011). "Call Centers with Delay Information: Models and Insights." Manufacturing & Service Operations Management **13**(4): 534-548.

Karmeshu and S. Sharma. (2005). Long Tail Behaviour of Queue Lengths in Broadband Networks: Tsallis Entropy Framework. Technical Report, School of Computing and System Sciences, J. Nehru University, New Delhi, India.

Karmeshu and S. Sharma (2006a). Power Law and Tsallis Entropy: Network Traffic and Applications. Chaos, Nonlinearity, Complexity. A. Sengupta, Ed., Springer Berlin Heidelberg. **206:** 162-178.

Karmeshu and S. Sharma (2006b). "q-Exponential product-form solution of packet distribution in queueing networks: maximization of Tsallis entropy." Communications Letters, IEEE **10**(8): 585-587.

Karmeshu and S. Sharma (2006c). "Queue length distribution of network packet traffic: Tsallis entropy maximization with fractional moments." Communications Letters, IEEE **10**(1): 34-36.

Kaufman, J. S. and K. M. Rege (1996). "Blocking in a shared resource environment with batched Poisson arrival processes." Performance Evaluation **24**(4): 249-263.

Ke, J.-C. (2007). "Operating characteristic analysis on the M[x]/G/1 system with a variant vacation policy and balking." Applied Mathematical Modelling **31**(7): 1321-1337.

Ke, J.-C. and C.-H. Lin (2008). "Maximum entropy approach for batch-arrival queue under N policy with an un-reliable server and single vacation." J. Comput. Appl. Math. **221**(1): 1-15.

Kemp, A. W. (2005). "Steady-state Markov chain models for certain q-confluent hypergeometric distributions." Journal of Statistical Planning and Inference **135**(1): 107.

Kemp, A. W. (2008). The Discrete Half-Normal Distribution. Advances in Mathematical and Statistical Modeling. B. C. Arnold, N. Balakrishnan, J. M. Sarabia and R. Minguez, Eds., Birkhäuser Boston**:** 353-360.

Kendall, D. G. (1951). "Some Problems in the Theory of Queues." Journal of the Royal Statistical Society. Series B (Methodological) **13**(2): 151-185.

Kleinrock, L. (1975). Queueing Systems Vol. 1:Theory, John Wiley & Sons.

Kouvatsos, D. and S. Assi (2011a). Generalised Entropy Maximisation and Queues with Bursty and/or Heavy Tails. Network Performance Engineering. D. Kouvatsos, Ed., Springer Berlin / Heidelberg. **5233:** 357-392.

Kouvatsos, D. and S. Assi (2011b). On the Analysis of Queues with Heavy Tails: A Non-Extensive Maximum Entropy Formalism and a Generalisation of the Zipf-Mandelbrot Distribution. Performance Evaluation of Computer and Communication Systems. Milestones and Future Challenges. K. Hummel, H. Hlavacs and W. Gansterer, Eds., Springer Berlin Heidelberg. **6821:** 99-111.

Kouvatsos, D. and I. Awan (2003). "Entropy maximisation and open queueing networks with priorities and blocking." Performance Evaluation **51**(2-4): 191.

Kouvatsos, D., P. Georgatsos, et al. (1989) "GE-Type Stochastic Algebra for GE-Type Queues", Research Report No. RS-89-013, University of Bradford.

Kouvatsos, D. and N. Tabet-Aouel (1994). "An ME-based approximation for multi-server queues with preemptive priority." European Journal of Operational Research **77**(3): 496-515.

Kouvatsos, D. D. (1986a). "Maximum entropy and the G/G/1/N queue." Acta Inf. **23**(5): 545-565.

Kouvatsos, D. D. (1986b). A maximum entropy queue length distribution for the G/G/1 finite capacity queue. Proceedings of the 1986 ACM SIGMETRICS joint international conference on Computer performance modelling, measurement and evaluation. Raleigh, North Carolina, USA, ACM**:** 224-236.

Kouvatsos, D. D. (1988). "A Maximum Entropy Analysis of the G/G/1 Queue at Equilibrium." The Journal of the Operational Research Society **39**(2): 183-200.

Kouvatsos, D. D. (1994). "Entropy maximisation and queueing network models." Annals of Operations Research **48**(1): 63-126.

Kouvatsos, D. D., J. S. Alanazi, et al. (2011). "A Unified ME Algorithm for Arbitrary Open QNM's with Mixed Blocking Mechanisms." Numerical Algebra, Control and Optimization **1**(4): 781-816.

Kouvatsos, D. D. and J. Almond (1988). "Maximum entropy two-station cyclic queues with multiple general servers." Acta Inf. **26**(3): 241-267.

Kouvatsos, D. D. and S. A. Assi (2002). An Investigation into Generalised Entropy Optimisation with Queueing Systems Applications. The 3rd

Annual Postgraduate Symposium on the Convergence of Telecommunications, Networking and Broadcasting (PGNet 2002), Liverpool, UK, The School of Computing and Mathematical Sciences, Liverpool John Moores University.

Kouvatsos, D. D., I. Awan, et al. (2003). "Performance modelling of GPRS with bursty multiclass traffic." Computers and Digital Techniques, IEE Proceedings - **150**(2): 75-85.

Kouvatsos, D. D., I. U. Awan, et al. (2000). "A cost-effective approximation for SRD traffic in arbitrary multi-buffered networks." Computer Networks **34**(1): 97-113.

Kouvatsos, D. D. and S. G. Denazis (1993). "Entropy maximised queueing networks with blocking and multiple job classes." Performance Evaluation **17**(3): 189-205.

Kouvatsos, D. D. and N. Tabet-Aouel (1989). "A Maximum Entropy Priority Approximation for a Stable G/G/1 Queue." Acta Inf. **27**(3): 247-286.

Kuehn, P. (1979). "Approximate Analysis of General Queuing Networks by Decomposition." Communications, IEEE Transactions on **27**(1): 113-126.

Labrador, M. A. and S. Banerjee (1999). "Packet dropping policies for ATM and IP networks." Communications Surveys & Tutorials, IEEE **2**(3): 2-14.

Leland, W. E., M. S. Taqqu, et al. (1994). "On the self-similar nature of Ethernet traffic (extended version)." Networking, IEEE/ACM Transactions on **2**(1): 1-15.

Liu, L. (2007). Service Systems with Balking Based on Queueing Time. PhD Thesis, Statistics and Operations Research, University of North Carolina at Chapel Hill.

Lopez-Herrero, M. J. (2002). "On the number of customers served in the M/G/1 retrial queue: first moments and maximum entropy approach." Computers & Operations Research **29**(12): 1739-1757.

Lu, X. and B. L. Mark (2004). "Performance modeling of optical-burst switching with fiber delay lines." Communications, IEEE Transactions on **52**(12): 2175-2183.

Manfield, D. and P. Tran-Gia (1982). "Analysis of a Finite Storage System with Batch Input Arising out of Message Packetization." Communications, IEEE Transactions on **30**(3): 456-463.

Mendelson, H., R. R. Pillai, et al. (1999). "Inferring Balking Behavior From Transactional Data." Oper. Res. **47**(5): 778-784.

Morse, P. M. (1958). Queues, Inventories and Maintenance, The Analysis of Operational Systems with Variable Demand and Supply, John Wiley & Sons, Inc.

Naor, P. (1969). "The Regulation of Queue Size by Levying Tolls." Econometrica **37**(1): 15-24.

Nijenhuis, A. and H. S. Wilf (1978). Combinatorial Algorithms for Computers and Calculators. New York, Academic Press, Inc.

Norros, I. (1994). "A storage model with self-similar input." Queueing Systems **16**(3): 387-396.

Rao, S. S. and N. K. Jaiswal (1969). "On a Class of Queuing Problems and Discrete Transforms." Operations Research **17**(6): 1062-1076.

Shah, N. and D. Kouvatsos (2011). A Queue Conjectured to Bear the Generalised Discrete Half Normal Maximum Entropy QLD. 27th Annual UK Performance Engineering Workshop, University of Bradford.

Shah, N. and D. Kouvatsos (2013). The GE/GE/1/N Queue Subject to State-Dependent Arrival Balking. Seventh International Working Conference on Performance & Security Modelling and Evaluation of Cooperative Heterogeneous Networks, Ilkley, UK.

Shah, N., D. D. Kouvatsos, et al. (2010). An Analytic Generalisation of a Maximum Entropy Customer Impatience Queueing Solution and its Nonbalking G/M/1/N Equivalence. The 11th Annual Postgraduate Symposium on the Convergence of Telecommunications, Networking and Broadcasting, Liverpool, UK, The School of Computing and Mathematical Sciences, Liverpool John Moore's University.

Shannon, C. E. (1948). "A Mathematical Theory of Communication." The Bell System Technical Journal **27**: 379-423 and 623-656.

Shore, J. and R. Johnson (1980). "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy." Information Theory, IEEE Transactions on **26**(1): 26-37.

Shore, J. E. (1982). "Information theoretic approximations for M/G/1 and G/G/1 queuing systems." Acta Informatica **17**(1): 43.

Skianis, C. (1997). Arbitrary queueing network models with blocking and server vacations: approximate analysis of queueing network models of manufacturing and computer communication systems with finite capacities, server vacation periods and different types of building

block queues using the principle of minimum relative entropy and the generalised exponential distribution. PhD Thesis, Department of Computing, University of Bradford.

Takagi, H. and T. Nishi (1998). "Correlation of Interdeparture Times in M/G/1 and M/G/1/K Queues." Journal of the Operations Research Society of Japan **41**(1): 142-151.

Tsallis, C. (1988). "Possible generalization of Boltzmann-Gibbs statistics." Journal of Statistical Physics **52**(1-2): 479-487.

Walstra, R. J. (1985). "Nonexponential networks of queues: a maximum entropy analysis." SIGMETRICS Perform. Eval. Rev. **13**(2): 27-37.

Wang, K.-H., S.-L. Chuang, et al. (2002). "Maximum entropy analysis to the N policy M/G/1 queueing system with a removable server." Applied Mathematical Modelling **26**(12): 1151.

Wang, K., N. Li, et al. (2010). Queueing System with Impatient Customers: A Review. 2010 IEEE International Conference on Service Operations and Logistics and Informatics (SOLI).

Ward, A. and P. Glynn (2005). "A Diffusion Approximation for a GI/GI/1 Queue with Balking or Reneging." Queueing Systems **50**(4): 371-400.

Whitt, W. (1982). "Approximating a Point Process by a Renewal Process, I: Two Basic Methods." Operations Research **30**(1): 125-147.

Whitt, W. (1984). "Approximations for departure processes and queues in series." Naval Research Logistics Quarterly **31**(4): 499-521.

Whitt, W. (1999). "Improving Service by Informing Customers About Anticipated Delays." Management Science **45**(2): 192-207.

Whitt, W. (2005). "Engineering Solution of a Basic Call-Center Model." Management Science **51**(2): 221-235.

Wischik, D. (2005). Buffer requirements for high-speed routers. 31st European Conference on Optical Communication (ECOC 2005), Institution of Electrical Engineers.

Wu, J.-S. and W. C. Chan (1989). "Maximum Entropy Analysis of Multiple-server Queueing Systems." Journal of the Operational Research Society **40**(9): 815-825.

Yang, D. Y., K. H. Wang, et al. (2011). "First two moment entropy maximisation approach for M/G/1 queues with second optional service and server breakdowns." International Journal of Services Operations and Informatics **6**(4): 310-331.

Yue, D., W. Yue, et al. (2006). Performance Analysis of an M/M/c/N Queueing System with Balking, Reneging and Synchronous Vacations of Partial Servers. The Sixth International Symposium on Operations Research and its Applications (ISORA'06), Xinjiang, China.

Zhang, Y., D. Yue, et al. (2005). Analysis of an M/M/1/N Queue with Balking, Reneging and Server Vacations. The Fifth International Symposium on OR and Its Applications 2005, Tibet, China.

Zhen, Q., J. H. Leeuwaarden, et al. (2010). "On a processor sharing queue that models balking." Mathematical Methods of Operations Research **72**(3): 453-476.