Noname manuscript No.
(will be inserted by the editor)

Optimal Design of Measurements on Queueing Systems

**Ben M Parker · Steven Gilmour · John Schormans ·
Hugo Maruri-Aguilar**

the date of receipt and acceptance should be inserted later

Abstract We examine the optimal design of measurements on queues with particular reference to the M/M/1 queue. Using the statistical theory of design of experiments, we calculate numerically the Fisher information matrix for an estimator of the arrival rate and the service rate to find optimal times to measure the queue when the number of measurements are limited for both interfering and non-interfering measurements.We prove that in the non-interfering case, the optimal design is equally spaced. For the interfering case, optimal designs are not necessarily equally spaced. We compute optimal designs for a variety of queuing situations and give results obtained under the D- and D_s- optimality criteria.

Keywords Design of experiments; maximum likelihood estimation; M/M/1 Queue; active measurements

1 Introduction

We examine the optimal design of measurements on queues with particular reference to the M/M/1 queue. This queue consists of arrivals according to a Poisson process with rate λ; if a customer is present in the queue, he is served and departs according to a Poisson process with rate μ.

Whilst a lot is known about the M/M/1 queue, there has been limited research about the optimal times to measure the queue in order to make inference about the parameters. In particular, in some applications (e.g. communications networks) measuring queues can require adding customers to the queue to act as survey customers. This has the effect of altering the future behaviour of the queue, and potentially changing the optimal measurement pattern of the queues: observations interfere with the experiment. We also look in some detail at this interesting interfering case.

Ben M Parker · Steven Gilmour
Southampton Statistical Sciences Research Institute, University of Southampton, UK

John Schormans · Hugo Maruri
Queen Mary University of London

To demonstrate our approach, we set out in this article to measure the number of customers in the queue. As the M/M/1 queue is fully determined by two parameters, λ and μ, we try to optimally estimate these parameters given a fixed number of measurements by determining the optimal time points at which to measure the queue. Our optimality criterion is to minimize the variance of the maximum likelihood estimator of the parameters λ and μ by calculating some function of the Fisher information matrix.

The difficulty in finding the Fisher information of this M/M/1 process (and other queueing systems) is that the conditional probability density function of the queue (i.e. how it evolves after a measurement) is complicated and difficult to evaluate numerically; we provide some methods to overcome this problem within this article, based on previous work by [26], though focusing on the design approach.

1.1 Terminology of queues in communication networks

Although queueing theory is used within a number of fields, electronic engineering and computer science make substantial use of queues for modelling packet communication networks; in these networks communications data is encapsulated in bundles called packets and passed through a communications network. Often, communications networks are modelled as a series of queues, with packets being customers within these queues.

In packet networking samples are usually taken by inserting probes into the network; while these probes are not user data packets it is assumed by the network engineers that they will still provide a good enough estimate of the queueing experienced by the user packets. Probing has the unusual statistical property that measuring a network increases the number of packets within that network, the measurement alters the system being measured. In this paper we use the number of packets present in the buffer as a convenient measure of delay, rather than use delay directly: almost all buffering in packet-based communications networks uses FIFO scheduling, so the sampled instantaneous queue length (termed $y(x)$, and always a non-negative integer) is also the sampled virtual waiting time at the measurement instant. When the measurement process is correctly designed, the measurements provide optimal estimates for either λ, or μ, or both, in the target M/M/1 queue.

1.2 Structure of this article

In section 2 we review previous literature on measuring queues, which to a large extent considers which estimator should be used for particular data. We show in section 3 how to extend previous results to evaluate the Fisher information for any particular measurement times of the queue for unknown parameters λ and μ. We prove a general result about Fisher information in section 4. By considering particular functions of the Fisher information, we find numerically in section 5 the optimal times at which to measure the queue in order to maximise our information about the parameters. This has the effect of minimizing the variance in the estimator of the parameters. In the case where observations do not interfere with the queue, we find that the optimal designs are relatively straightforward to calculate due to symmetry in the problem, but the case where observations interfere with the queue (i.e the probing case) produces more complicated results. We conclude in section 6.

2 Previous Work

2.1 Measuring queues

Earlier research is predominantly concerned with devising probabilistic methods and sometimes using them in model-based prediction, rather than measurement-based inference. This work has covered a very wide range of applications and areas, including queues with self-similar input [28] (prompted by recent discoveries concerning the nature of internet traffic, [24]), priority queueing systems [1], congestion control [13], network scheduling [39], fluid-flow modelling of queues [38] improving datacentre performance [33], and wireless-network specific modelling [18].

In this paper we use an idea first used in our earlier work: we treat all measurement as a numerical experiment, and optimise this process by applying the statistical theory of the design of experiments (DOE). In this way we can address the measurement of queues formally by designing the measurements such that we evaluate when it is best to sample the queue to infer information about queue performance.

2.1.1 Inference and Estimators

Clarke [17] first investigates statistical inference in queues, deriving formulae for maximum likelihood estimators (MLEs) for the M/M/1 queue. He observes the queue until the busy time reaches some fixed value, and notes the number of packets arrived and departed, and the last departure time. The key point is that the busy-time process of the queue can be regarded as separate from the arrival process, and MLEs calculated for λ and μ.

Jenkins [20] compares the relative efficiencies of the direct estimate for the mean waiting time with that suggested by Clarke, and concludes that the MLE has a lower asymptotic variance, particularly for high values of load ρ, defined as $\frac{\lambda}{\mu}$.

Aigner [3] summarises known work at the time (1974) and compares various estimators of different queue parameters for the M/M/1 queue, when the number of packets sampled is fixed. Even for this simple setup, there are a vast number of estimators such as MLEs, least squares estimators, and other more adhoc estimators regarding the ratio of two statistics ("ratio estimators"). Aigner uses asymptotic variance of different estimators as the criterion to decide which is best; he notes that this is a somewhat arbitrary optimality criterion, which does not apply to inference from small samples, and does not take any account of the time needed to gather data. However, Aigner does clearly indicate the difficulty in determining, even for a fixed sampling method and a simple queue, which estimator is the best. Reynolds ([34],section 5) assesses variances of different estimators for a wider class of queues in the fixed time sampling frame.

Basawa and Prahu [10] use probability theory from Billingsley [12] to show how asymptotic normality results can be used for estimators in an M/M/1 queue. They show later [11] derivations for MLEs and information matrices for queues whose arrival and departure distributions come from exponential families with two parameters to be estimated, under various stopping rules.

Achaya [2] extends the work by Basawa and Prahu, and shows how quickly MLEs converge; in other words, how big a sample is needed for the asymptotic theory developed to apply.

In a later paper, Basawa at al.[8] attempt to establish a general framework to find the Fisher information matrix. A possible limitation in applying this work (and most work on MLEs) to real queues, is that the method applies only asymptotically, as the number of samples tends to infinity. They show, at least asymptotically, that the MLE is not affected by the choice of sampling frame, although it does not follow that what is best for large samples is also best for small experiments. Indeed, in most engineering research, there is an implicit stationarity assumption (as described by Roughan [37]) that the traffic rate λ does not change, or in other words that we measure over a short enough period of time that this assumption is valid. As estimators are used over a small number of data gathered in a short period of time, the asymptotic results cannot be relied on; the implication of this result is that in practice the best estimators for a given problem are not necessarily being used by practitioners.

The relevance to this current paper is that previous work concerns large sample inference; here we consider the (sometimes more realistic) case where we have a small number of observations on a queue.

2.2 Partial Information

A real experimenter may be limited in the knowledge he is able to gather. He might not have access to the underlying (user) packets in a system, or measuring all packets may be impractical or impossible. The experimenter may be limited to a number of survey or probe packets from which all inference on the queue must be performed. In the problem studied in this paper, for example, we consider that we may only observe a queue at a fixed number of times, so our information can be thought of as partial.

Basawa et al. [8] consider finding MLEs given only partial information of waiting time data. They show asymptotic consistency and normality of the estimators, and present forms for the MLE and Fisher information for partial information in the special cases of M/M/1 and M/E_k/1, where the service times have the Erlangian distribution. The analytic results show that these MLEs turn out to have rather large variance and are biased. The derivation of these results is relatively complex for any given queue, however, due to the waiting time distribution being non-continuous at time zero.

Basawa [9] extends this work to general (G/G/1) queues with both service and inter-arrival times drawn from exponential families, when only the sample packet's waiting time or system time, together with queue idle times, are known. He derives likelihood equations, and demonstrates numerical methods to solve these.

Chen [15] takes an M/D/1 queue with partial knowledge of waiting and service times for some packets, and tries to find the MLE for the arrival rate λ for k observed packets. Based on the partial data available, a complex form for the MLE is derived, and he proves that the distribution of the MLE is asymptotically normal. Simulation shows the log likelihood to be unimodal, and thus that an MLE can be found numerically (though non-trivially); Chen concludes that the method is more generally applicable, although the exact method to be used will vary depending on what data are available, and what queue is being measured.

None of this research [8, 9, 15] explicitly considers active probing, where extra packets are put into the system and measurements on those packets become our data; instead, data from a random sample of packets is known. The authors therefore do not consider that introducing probe packets into the network may cause interference with the data packets; the case where probes are used and do interfere was previously studied in a Markov chain model of a network router in [31].

2.2.1 Performance measurement in packet-based communications networks

As noted, recent research into packet networks has considered the injection of probe packets to measure the packet level performance, such as packet loss and delay. For example, there has been some consideration of whether it is best to probe at a uniform rate, or to send probes according to some renewal process, such as a Poisson process. Much of this activity has been motivated by the need to guarantee SLAs on packet-level performance (loss and delays) in commercially operated networks, see for example `http://www.verizonbusiness.com/terms/us/products/satellite_services/smb/`. There are patents in existence that relate to, or use, packet level measurements (e.g. [21],[25]), and there has also been considerable research on the potential use of packet-trains (sequences of quasi-contiguous packets) to determine available bandwidth, loss probability etc, e.g. [7],[27]. A packet-train is essentially a performance sampling process that has not been designed in the sense that we consider in this paper.

The PASTA (Poisson Arrivals See Time Averages) theorem, first formalised by Wolff [42] has been a widely used principle in packet probing; it tells us that if we introduce probe packets such that their inter-arrival time is governed by a Poisson process, then the mean of their waiting time is an unbiased estimator of the waiting time of all packets (combining both probe and underlying data packets) in the queue.

Roughan [37] and Bacelli et. al ([5],[6]) show that whilst PASTA is desirable in finding unbiased estimators, there are non-Poisson patterns of arrivals that give estimators with lower variance. Here we extend this work to find the optimal pattern of arrivals where minimising variance of estimators is our optimality criterion.

Roughan [36] shows that there are fundamental bounds on how accurately network measurements can be made: that no matter how many samples are used in a time interval, there is a limit to the knowledge we can gather about a queue. He makes an analogy to Heisenberg's uncertainty principle in quantum mechanics, where our certainty on position or momentum of a sub-atomic particle cannot be increased above a certain limit no matter how many times we observe it. Although his analytic results focus on measurement of a system where we have 'perfect measurements', he generalises the work to active probing, although he notes that analytic results would be complex in form and derivation.

In this paper we use a well-studied queueing system, the M/M/1, as a model for buffering in a packet network. As with many earlier papers e.g. [5,6,36,37], our intention is that we derive results that are applicable to packet-based communications networks, both where probing is used, and where it is not needed. We use the number of packets present in the buffer as a convenient measure of delay, rather than use delay directly, despite the fact that both probing and passive monitoring (which does not require probes) may measure delay, rather than queue backlog, as delay is often easier to determine. Almost all buffering in packet-based communications networks uses FIFO scheduling, e.g. see [41], so the instantaneous queue length (queue backlog) is anyway equivalent to packet virtual waiting time at that instant.

The insertion of probes into networks takes place where passive measurements are unavailable. Passive measurements are typically available through a management interface available on recent "high-end" routers. Such measurements are typically samples of e.g. the instantaneous queue depth in any target buffer in a chosen output line-card. The usual purpose of inserting probes is to provide an estimator of the experience, e.g. queue depth (delay) encountered, of the actual user traffic where passive measurements are not available (and this does still include a significant proportion of current routers in use in global networks). Probes come with a further cost: they interfere with the real user traffic, altering the performance as a result of increasing the load on the network. For this reason it is important to keep the number of injected probes to a minimum.

However, there is another reason that the number of samples n used to provide an estimator should always be minimised, and importantly this reason also applies to passive measurements sampled without probing. One purpose of performance measurement is to determine, in real-time, whether the network can support another connection. This is called Connection Admission Control (CAC) [19], and for any CAC algorithm to work requires both accuracy and speed, hence the maximum accuracy with the minimum number of samples is required. As an illustration, consider a 1 Gigabit Ethernet link purchased to support VoIP on commercial premises. Assuming 100 byte VoIP packets and 50% load, such a link would carry roughly 8000 active voice connections in parallel. With an average connection holding time of 3 minutes (a widely used figure in network dimensioning), on average a connection completes (and therefore, in steady state, a new connection arrives, and a CAC decision must be made) every 25 milliseconds. (If $n = 5$ samples then the processing time available between samples would be only 10ms.) Any viable CAC algorithm must therefore be fast enough to cope with this; indeed the faster the better since CAC is often centralised (like routing) and would need to cover many links at the same time. Since speed is inversely proportional to the number of samples taken, it is important to minimise the number of samples, however they are taken.

2.3 Measurement of Queues in other fields

The problem of optimally measuring an M/M/1 queue is discussed, in the context of a simple population model, in Pagendam and Pollett [29]. In that research, a Markovian birth/death process is investigated, which is similar to the M/M/1 queue we investigate below; however for that population, it is assumed that $\lambda > \mu$ such that the population growth rate is positive. The Fisher information matrix for the unknown parameters λ and μ is found, however a more crucial difference between that work and ours is that a Gaussian diffusion approximation is used to approximate the process rather than the true model.

A similar method of a Gaussian diffusion approximation is used in [35] to find the optimal times for observing epidemics which evolve according to particular models (SI, SIS, and SI(k)R) are considered. These are all stochastic models which obey the Markov principle, but are slightly different to the M/M/1 model we study below. Robust estimates for an SIS model are demonstrated in [30] for the spread of the crown of thorns starfish (*Acanthaster planci*) in Japan by evaluating the performance of particular optimality criteria.

In [31] the discrete time two parameter birth death process is studied, and this is generalised in [32] for general Markov chains. For a wide class of Markov chains, with a discrete state space and transitions at discrete times, optimal times for measurements are found. The models in this previous research can be thought of as discrete time approximations to the work we present in this paper.

3 Methodology

In this article we investigate how to measure a single server M/M/1 queue optimally. Henceforth, as a strong motivation in the work is packet communications networks, we refer to customers as packets. We assume that we can measure the number of packets in the queue without error at any time point, and that we have a fixed number of observations.

The number of observations may be limited due to logistical, technical, or other constraints; in general we assume that monitoring is in some sense expensive, and we wish to minimise the number of observations we take, or more simply to maximise the information for a fixed number of observations. For example, in a packet communication network, probe packets may take up network bandwidth which could delay user traffic, so we might wish to measure as infrequently as possible, or more likely to get the most information from the same number of measurements.

We measure the amount of information gained in the experiment by the Fisher information for parameters λ and μ using optimality criteria (D, and D_s) described later in section 3.3. These criteria minimise some function of the variance matrix for estimators of the queue, and hence in this article we consider MLEs.

3.1 Model overview and description of problem

Suppose we have an M/M/1 queue into which packets arrive with inter-arrival times distributed exponentially with parameter λ (i.e. defined by a Poisson process with rate λ), and in which service times are exponentially distributed with rate μ. Thus the stochastic behaviour of the queue is completely determined by λ and μ.

We may measure the response, $y(x)$ at any time x, and here $y(x)$ will represent the number of customers in the queue at time x. We will assume the queue has been running for a sufficiently long time that it is in equilibrium at time 0. Our methodology only requires the number of customers in the queue, so the queue discipline (e.g. first in first out) is ignored.

The design problem is, for small n, to choose times $x_1, x_2, \ldots, x_n$ at which to observe the queue in order to maximise our information about some estimators $\hat{\lambda}$, $\hat{\mu}$ or some combination of the two. Here we consider $n \leq 5$ to make computational feasible, and also in view of the restrictions, particularly in active measurement and applications such as Connection Admission Control (CAC) described in section 2.2.1 above.

Our secondary problem, often occurring in active probing of communications networks, is to assume that the measurements interfere with the operation of the queue and that adding a measurement (probe) of size c at time x increases the number of customers in the queue such that $y(x_i^+) = y(x_i) + c$ where time $x^+ = x + \delta$ for infinitesimally small $\delta > 0$. Throughout the rest of this paper, we will define $c = 1$, such that one extra customer is added to the queue per active measurement. We define the notation $y_i = y(x_i)$ to be the number of customers in the queue at the i-th design point.

Morse [26] describes how to find the conditional probability that the queue is in state m at time T given that is in state N at time 0 to be

$$g(N,m,T|\lambda,\mu) = P[(y(T)=m|y(0)=N,\lambda,\mu] = \delta_{m,N} - \left(\frac{\mu}{\pi}\right)\left(\frac{\lambda}{\mu}\right)^{\frac{m-N}{2}} \int_0^T \int_0^{2\pi} \left[\sin N\omega - \sqrt{\frac{\lambda}{\mu}}\sin(N+1)\omega\right]\left[\sin m\omega - \sqrt{\lambda\mu}\sin(m+1)\omega\right] e^{-(\mu+\lambda-2\sqrt{\lambda\mu}cos\omega)u} d\omega du \tag{1}$$

where δ_{ij} is the Kronecker delta, and ω is a constant of integration.

It is known (see e.g. [22]) that the stationary distribution of the number of customers in the M/M/1 queue is geometric with parameter $1-\lambda/\mu$. We assume that the queue has been running sufficiently long before measurement that the number of customers in the queue is determined by this stationary distribution

$$P(Y_1 = N) = \left(1-\frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^N. \tag{2}$$

Thus the probability distribution function for data $\boldsymbol{y} = (y_1, \ldots, y_n)$, given observation times $\boldsymbol{x}$ and parameters λ, μ, is

$$f(\boldsymbol{y}|\boldsymbol{x},\lambda,\mu) = \left(1-\frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^{y_1} \prod_{i=1}^{n-1} g(y_i, y_{i+1}, x_{i+1}-x_i), \tag{3}$$

from which we can write the likelihood of λ, μ given the data as $L(\lambda,\mu|\boldsymbol{y},\boldsymbol{x}) = f(\boldsymbol{y}|\boldsymbol{x},\lambda,\mu)$.

Note if we consider the case where measuring the network adds a probe of size c to the queue, we can consider this within the same framework by noting that the likelihood function changes slightly to become

$$f(\boldsymbol{y}|\boldsymbol{x},\lambda,\mu) = \left(1-\frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^{y_1} \prod_{i=1}^{n-1} g(y_i + c, y_{i+1}, x_{i+1}-x_i).$$

Given a particular design $\boldsymbol{x}$, we can then find the MLE for $\boldsymbol{\theta} = (\lambda \quad \mu)^T$, in the usual way by maximising $f(\boldsymbol{y}|\boldsymbol{x},\lambda,\mu)$. However, here we are interested primarily in calculating the expected Fisher information matrix with the (i,j)-th element defined as

$$I_{ij}(\boldsymbol{x},\boldsymbol{\theta}) = \mathop{\mathbb{E}}_{\boldsymbol{y}\in\mathcal{Y}}\left[\frac{\partial log f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})}{\partial\theta_i}\frac{\partial log f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})}{\partial\theta_j}\right] = \sum_{\boldsymbol{y}\in\mathcal{Y}} \frac{\partial log f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})}{\partial\theta_i}\frac{\partial log f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})}{\partial\theta_j} f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}). \tag{4}$$

Here $\boldsymbol{\theta}$ is the vector of parameters $\boldsymbol{\theta} = (\lambda \quad \mu)^T$ and $\mathcal{Y}$ is the set of all potential queue outcomes, $\mathcal{Y} = \mathbb{Z}^n_{\geq 0}$.

The Fisher information is a matrix that summarises the variance of estimators of the unknown $\boldsymbol{\theta}$ when the queue is observed at time points $\boldsymbol{x}$. We define the information function in section 3.3 below, which allows us to form criteria to compare estimators of the two dimensional $\boldsymbol{\theta}$ in order to find the optimal time points at which to observe the queue.

3.2 Calculating the PDF

Equation (1) is difficult to evaluate numerically, but it is shown in [26] that we can calculate

$$g(N,m,t|,\lambda,\mu) = P[(y(t)=m|y(0)=N,\lambda,\mu] = P(Y(t)=m)) + \sum_{k=1}^{N} Q_m(k,t) - \sum_{k=1}^{\infty} (\lambda/\mu)^k Q_m(k,t), \quad (5)$$

where

$$Q_m(k,t) = (\lambda/\mu)^{\frac{1}{2}(m-k)}\, e^{-(\lambda+\mu)t} \left\{ B_{k-m}(z) - B_{k+m}(z) - \sqrt{\lambda/\mu}\,[B_{k-m-1}(z) - B_{k+m+1}(z)] \right\}, \quad (6)$$

in which $z = 2t\sqrt{\lambda\mu}$ and $B_k(z)$ is the modified Bessel function of the first kind. Morse calls these Hyperbolic Bessel functions, but this term is no longer as frequently used. These are defined by

$$B_k(z) = \frac{1}{2\pi}\int_0^{2\pi} \cos(k\omega)e^{zcos\omega}d\omega$$

and the M/M/1 queue leads to boundary conditions such that

$$Q_m(k,0) = \begin{cases} -1 & \text{if } m = k-1, \\ 1 & \text{if } m = k, \\ 0 & \text{otherwise.} \end{cases}$$

The advantage of this approach for evaluating $g(\cdot)$ is that the Bessel functions are relatively fast to evaluate computationally, whereas evaluating the integral in equation (1) directly, for instance using quadrature or Monte Carlo, would be very slow.

Having found a fast way to calculate $g(m,N,t)$, we can calculate the information in equation (4) directly. We approximate the differentials in that equation as the usual numerical approximation

$$\frac{\partial logf(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})}{\partial\theta_i} = \frac{logf(\boldsymbol{y}|\boldsymbol{x},(\theta_i+\Delta,\theta_j)^T) - logf(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}))}{\Delta}$$

and similarly for θ_j, where in practice $\Delta = 10^{-8}$ gives an accurate approximation.

We truncate the infinite sum in equation (4) to remove regions which do not contribute to the amount of information. The Expected Fisher information matrix is approximated by the truncated sum $I^k(\boldsymbol{x},\boldsymbol{\theta})$ where

$$I^k_{ij}(\boldsymbol{x},\boldsymbol{\theta}) = \sum_{\boldsymbol{y}\in\mathcal{Y},\max\boldsymbol{y}\leq k} \frac{\partial log f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})}{\partial\theta_i}\frac{\partial log f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})}{\partial\theta_j} f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}) \tag{7}$$

and choose k to be the smallest integer such that $\det(I^{k+1}) < \det(I^k) + \epsilon$ and $\det(I^{k+1}) < \det(I^k)$. For the examples here, we pick $\epsilon = 0.01$. In practice we find that the Fisher information rapidly converges for moderate k, although this convergence becomes slower as $\lambda \to \mu$; this is because the first geometric term in equation (3) for the initial response y_1 of the queue decays exponentially with increasing y, and contributes very little to the information; a queue is vastly unlikely to contain a very high number of customers so this will not contribute much towards the information function, although as $\lambda \to \mu$ the backlogs that the queue can reach are higher, so contribute more, so we must choose a higher limit k at which to truncate our information.

3.3 Fisher Information and the Information function

In this article we focus on three objectives

- We wish to minimize the area of the joint confidence ellipsoid of λ and μ; this is the D-optimality criterion that maximises $\det(I(\boldsymbol{x},\boldsymbol{\theta}))$.
- We wish to minimize the length of the confidence interval for μ considering λ as a nuisance parameter. This is D_s optimality, and found by maximising $\det(I(\boldsymbol{x},\theta))/I_{11}(\boldsymbol{x},\boldsymbol{\theta})$.
- Similarly, we disregard the service rate μ, but we wish to minimize the length of the confidence interval for λ. This is D_s optimality, and found by maximising $\det(I(\boldsymbol{x},\theta))/I_{22}(\boldsymbol{x},\boldsymbol{\theta})$.

For a given problem, we will typically wish to find the set of inputs at which observing the queue gives us most information.

Without loss of generality, we assume that $x_1 = 0$ (so we start the experimental clock with our first observation). We let $d_i = x_{i+1} - x_i$ for $i = 1,\ldots,n-1$ be the differences between observation times of the queue, so we can specify a design with n observations by a vector of $n-1$ differences. This search is carried out over $\mathbb{R}^{n-1}_{0+}$ and again, the optimal design is given by a vector of differences d^*.

4 Proof that optimal design is equally spaced (non-interfering case)

Lemma 1 *For the M/M/1 queue with a maximum capacity where the observations are non-interfering (probe size $c = 0$), optimal designs are equally spaced.*

Proof Let $\mathcal{Y}^r = \{\boldsymbol{y} \in \mathcal{Y} | y_i \leq r \quad \forall i\}$, representing the queue where the maximum number of packets in a queue is r. This queue remains Markovian. Let $f(\cdot)$ and $g(\cdot)$ in this proof represent the initial and transition probabilities in this queue, which will be adjusted versions of $f(\cdot)$ and $g(\cdot)$ for the unbounded queue.

Consider the design with spacings $\epsilon_1, \ldots, \epsilon_{n-1}$, i.e the design $(0, \epsilon_1, \epsilon_1 + \epsilon_2, \ldots, \sum_{k=1}^{n-1} \epsilon_k)$. The (i, j)-th element of the information matrix I is given in Equation (4).

The D-optimal design maximises $\det I = I_{11} I_{22} - (I_{12})^2$. This is achieved with respect to spacings ϵ_k when $\frac{\partial}{\partial \epsilon_k} \det I = 0$, i.e

$$\left[\frac{\partial}{\partial \epsilon_k} I_{11}\right] I_{22} + \left[\frac{\partial}{\partial \epsilon_k} I_{22}\right] I_{11} - 2\left[\frac{\partial}{\partial \epsilon_k} I_{12}\right] I_{12} = 0. \tag{8}$$

As $\mathcal{Y}^r$ is a finite set, the interchange between expectation and derivative is guaranteed under the regularity of the likelihood so that

$$\begin{aligned}\frac{\partial}{\partial \epsilon_k} I_{ij} &= \mathop{\mathbb{E}}_{\boldsymbol{y} \in \mathcal{Y}^m} \left\{ \frac{\partial}{\partial \epsilon_k} \left[\frac{\partial log f(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})}{\partial \theta_i} \frac{\partial log f(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})}{\partial \theta_j} \right] \right\} \\ &= \mathop{\mathbb{E}}_{\boldsymbol{y} \in \mathcal{Y}^m} \left\{ \frac{\partial}{\partial \epsilon_k} \left[\frac{\partial log f(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})}{\partial \theta_i} \right] \frac{\partial log f(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})}{\partial \theta_j} + \frac{\partial}{\partial \epsilon_k} \left[\frac{\partial log f(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})}{\partial \theta_j} \right] \frac{\partial log f(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})}{\partial \theta_i} \right\}. \end{aligned} \tag{9}$$

Now, by the Markov principle, the log-likelihood can be simplified so that

$$\begin{aligned} \frac{\partial}{\partial \epsilon_k} \left[\frac{\partial log f(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})}{\partial \theta_i} \right] &= \frac{\partial}{\partial \epsilon_k} \left[\frac{\partial}{\partial \theta_i} log(f(y_1|\boldsymbol{x}, \boldsymbol{\theta})) + log(f(y_2|y_1, \boldsymbol{x}, \boldsymbol{\theta})) + \ldots + log(f(y_n|y_{n-1}, \boldsymbol{x}, \boldsymbol{\theta})) \right] \\ &= \frac{\partial}{\partial \epsilon_k} \left[\frac{\partial}{\partial \theta_i} log(f(y_1|x_1 = 0, \boldsymbol{\theta})) + log(g(y_1, y_2, \epsilon_1, |\boldsymbol{\theta})) + \ldots + log(f(y_{n-1}, y_n, \epsilon_{n-1}, \boldsymbol{\theta})) \right] \\ &= \frac{\partial}{\partial \theta_i} \left[\frac{\partial}{\partial \epsilon_k} log(f(y_1|x_1 = 0, \boldsymbol{\theta})) + log(g(y_1, y_2, \epsilon_1, |\boldsymbol{\theta})) + \ldots + log(g(y_{n-1}, y_n, \epsilon_{n-1}, \boldsymbol{\theta})) \right] \\ &= \frac{\partial}{\partial \theta_i} \left[\frac{\partial}{\partial \epsilon_k} log(g(y_k, y_{k+1}, \epsilon_k|\boldsymbol{\theta})) \right], \end{aligned} \tag{10}$$

which depends only on ϵ_k. Substituting Equation (10) into Equation (9), and this into Equation (8) allows us to find a value of ϵ_k^* that maximises $\det I$, although we may need to do this numerically. However, by symmetry this value of ϵ_k^* is the same for all k: call it ϵ^*. Thus the D-optimal design is given by $(0, \epsilon^*, 2\epsilon^*, \ldots, (n-1)\epsilon^*)$ so the optimal design is equally spaced.□

Corollary 1 *For the M/M/1 queue, where observations are non-interfering (probe size* $c = 0$*), (i.e.* $r \to \infty$*) in the case above) optimal designs are equally spaced.*

Proof For all practical cases, taking large r is appropriate as real queues, particularly in packet communication networks, have finite capacity. In the numerical results that follow, we only consider finite queues for computational reasons. However, the result is also true as $r \to \infty$, although strictly the in exchange of the order of expectation and differentiation must be justified.

In overview, we need to show that the information contributed by busy queues (as $\boldsymbol{y} \to \infty$) is negligible. (It may be that we see highly loaded queues infrequently, but seeing this gives us a lot of information.) In other words, we consider the summand on the right hand side of Equation (4),

$$s(y) = \frac{\partial log f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})}{\partial \theta_i} \frac{\partial log f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})}{\partial \theta_j} f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})$$

We need show only that $s(\boldsymbol{y}) \to 0$ for large r. Consider the two observation queue $\boldsymbol{y} = (y_1, y_2)$ (the result for larger numbers of observations is similar). It is clear from the definition of the density function $f(\cdot)$ in Equation (3) that, recalling $\lambda < \mu$, as $y_1 \to \infty$, $f(y) \to 0$. Similarly, note that the integrand in $g(\cdot)$ defined in Equation (1) is bounded as it is the scaled sum of sine functions. For small y_1, as $y_2 \to \infty$, $m \to \infty$ such that $g(y) \to 0$. Thus $f(y)$ is zero for either large y_1 or large y_2 (or both).

We now need only show that $\frac{\partial log f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})}{\partial \theta_i}$ is not infinite, so that $s(y)$ is finite. By differentiating Equation (3), it is clear that the first two terms representing the stationary distribution of the queue do not given an unbounded differential for $\lambda < \mu$, so we need only consider whether differentiating $g(\cdot)$ produces an unbounded function. It is easiest to consider Equations (5) and (6), and we see that trivially the differential of $g(\cdot)$ is only unbounded if the differential of $Q_m(k,t)$ is unbounded. By the properties of the Bessel functions of the first kind, these are only unbounded as $t \to \infty$, i.e if ϵ_1, the time difference between observations 1 and 2, is infinite, a case we do not consider here. Thus the differential of the density function $f(\cdot)$ is finite, $f(\cdot)$ itself approaches zero for large m, so the result is proven.□

Note that in the case where probes interfere with the measurement ($c > 0$), the system is no longer Markovian; the optimal ϵ_k may depend on previous spacings ϵ_{k-1}. Thus we consider the entire likelihood, and take a numerical approach to optimisation.

5 Numerical Results

For these results, we henceforth assume that $\mu = 1$. We do this for convenience, but without loss of generality as we can scale time t such that $\mu t = 1$, the service rate is always 1 per unit of time.

5.1 Computational note

Computation was performed using MATLAB on the IRIDIS High Performance Computing Facility at the University of Southampton; computation of the optimal design involves intensive evaluation of equation (7).

For example, in the contour plots that follow, for each plot an information function was evaluated at 10,000 design points $\boldsymbol{x}$.

In order to find optimal designs, we have used the Nelder-Mead algorithm as described by [23] to maximise the information. This is implemented within Matlab as `fminsearch`. In general, the procedure converges quickly; for example with $n = 3$ design points, to find a design such that its information value is within 0.1 of the global maximum takes around 20 evaluations of Equation (7). The smooth nature of the information surface seems well suited to this optimisation, and we have found no great differences between the optimization approach, and a more exhaustive search method.

The evaluation of Equation (4) involves summing the likelihood function for each possible $\boldsymbol{y} \in \mathcal{Y}$, which is the set of all possible responses of length n from observing the queue n times. We truncate this Equation (4) to Equation (7) as described in section 3.2 to only look at responses which contribute to the information function, however as $\lambda \to \mu$ the queue can reach high numbers of customers often, so the sum in Equation (7) must include higher values of k. In other words, the value of k should increase together with the load $\frac{\lambda}{\mu}$.

For each likelihood function, we must evaluate $g(\cdot)$ in Equation (5); the second term in this series is an infinite sum, and we find numerically it converges very quickly, although this convergence becomes slower as as $\lambda \to \mu$. Thus for high λ, we must not only evaluate $g(\cdot)$ more times, but its evaluation becomes harder; each evaluation of $g(\cdot)$ involves evaluating numerous Bessel functions, which is a slow step. However, for each value of λ we only evaluate these once, and store these for quick re-use within our code, which decreases the running time substantially.

We see below that in the non-interfering case, Lemma 1 tells us we need only consider one spacing, so the computational time does not increase for increasing n. For the interfering case, as the size of all potential sets of responses $y \in \mathcal{Y}$ for our observations in equation (7) is multiplied by k for each extra observation, we can see that for fixed λ and μ the calculation is approximately $O(k^n)$.

5.2 Optimal designs for non-interfering observations

For the M/M/1 queue in the non interfering case, we know that the optimal designs for D-optimality, and for D_s-optimality for λ and μ occur for equally spaced observations (Lemma 1), although the spacing of these observations may be different for number of observations $n \geq 2$.

Figure 2 shows the optimal design found for the distance between observations as a function of λ, with the optimal function value shown as Figure 1. These are presented for $n = 2$ and $n = 3$.

We see from Figure 2 that, for all three criteria studied, the optimal spacing between observations is higher for both small and large λ relative to μ (recall μ is fixed at 1); optimal spacing is at a minimum for

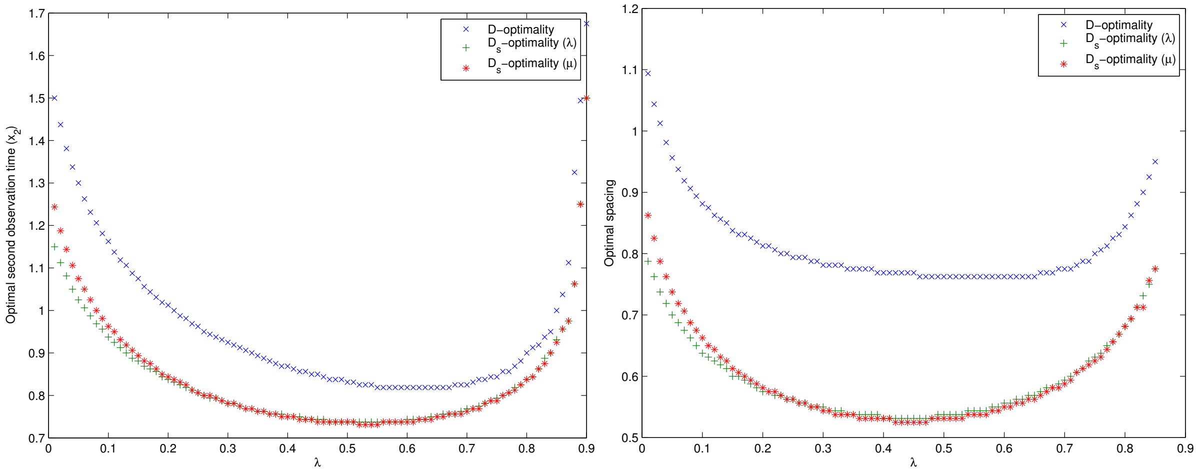

Fig. 1 Optimal spacings for three optimality criteria. Non-interference case, for $n = 2$ (left) and $n = 3$ (right) observations.

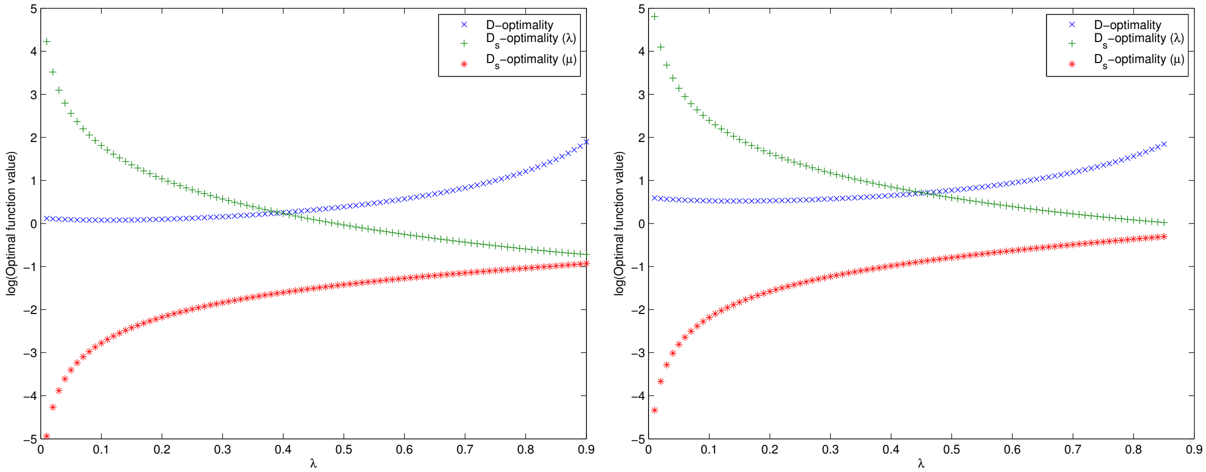

Fig. 2 Optimal values (log scale) for three optimality criteria. Non-interference case, for $n = 2$ (left) and $n = 3$ (right) observations.

medium loads. The optimal spacings are very similar for D_s optimality for both λ and μ, although both these criteria are a little different from the D-optimality criterion. Slightly longer gaps between observations are better when both λ and μ are to be estimated than if just one of these is. We see that the optimal spacing is closer together (note the smaller scale on the y-axis) for $n = 3$ than $n = 2$.

The optimal function value plots (Figure 1) can be interpreted thus; recall that we may only observe the number of customers in a queue, and not individual arrivals or departures. If we wish to estimate λ precisely, we would prefer to do this with a queue with higher loads. We are more likely to see arrivals in a highly loaded queue than with a lightly loaded queue, so we can be more confident of the estimator's precision. Conversely, if we wish to estimate μ we would rather do this with a lightly loaded queue. We can observe when the queue decreases more readily, and we do not have to worry that an arriving customer will mask a departure. The D-optimality criterion can be regarded as a compromise between these two previous criteria, and it is interesting to note that our overall information increases (and thus the precision of the joint confidence region of λ and μ decreases) with increasing load. This last is perhaps not intuitively obvious, but nevertheless interesting.

We now investigate how much a non-optimal observation affects the precision of any estimates. Figure 3 shows the information function value for different values of λ for varying space between 2 observations (left column) and 3 observations (right column). The overall variability of the information functions show how crucial the choice of spacing between measurements can be. Consider for example D_s optimality for λ when $n = 2$ with the true value of $\lambda = 0.1$ (the top line in the middle graph on the left). The optimal design has spacings of 0.9375s, with an optimal information function value of 6.12. If we wrongly used a spacing of, say, 4 seconds, we would get an information function value 1.16 an efficiency of 19%. Thus our estimator $\hat{\lambda}$ might have more than five times the variance of the best we could do. Choosing the right spacing between observations is therefore crucial for estimates with low variance. To show that choosing the spacing correctly is important for more n, for $n = 5$ observations and the same parameter values, the efficiency of spacing at 4 seconds compared to the optimum (spacing at 0.3937s) is 28.5%.

For the values of λ tabulated here, we see that there is not a large difference between locations of the optimum. We see that for each of the criteria there is a quite sharp ascent to the optimal spacing, then a less sharp descent. If we are going to measure at a non-optimal rate, it is thus better to measure at slightly bigger intervals than slightly smaller intervals.

For all these graphs, for spacing between observations approximately equal to 1, the information function is never too far from maximum. For the M/M/1 queue without interference, if we were to suggest that we measure uniformly at an interval approximately equal to the service rate, this would be an effective rule of thumb which would mean that none of our estimators have a variance which is too much higher than the optimum variance.

5.3 Interfering case

We now repeat the above analysis, but assume that observing the queue adds one customer of size $c = 1$ to the queue's load; as explained above, in packet communications networks, this is a common situation where probing a network involves adding probe packets to the network.

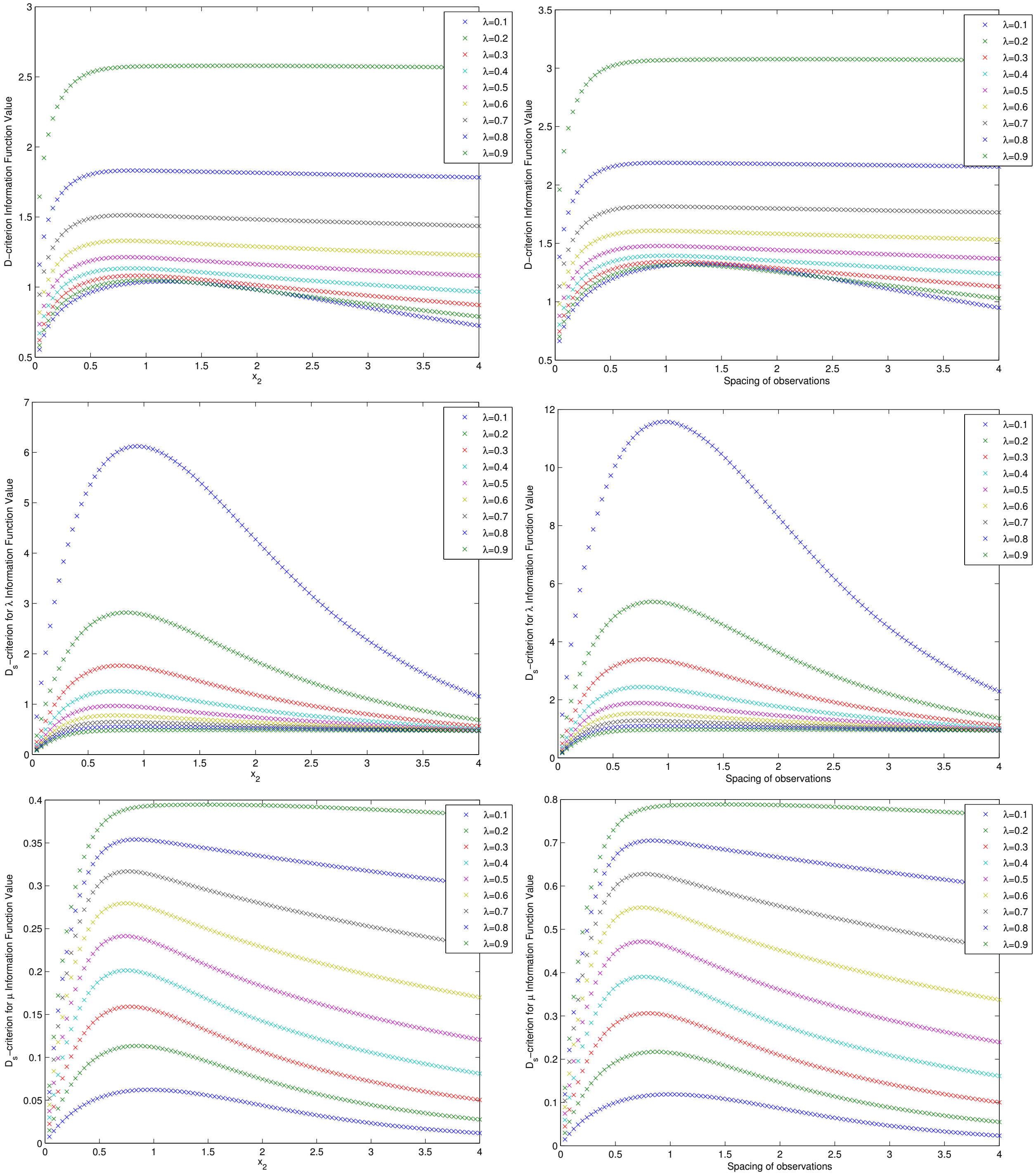

Fig. 3 Non-interference case, $n = 2$ (Left) and $n = 3$ (right): Information values for varying spacings (Top) D-criterion (Middle) D_S for λ (Bottom) D_s for μ.

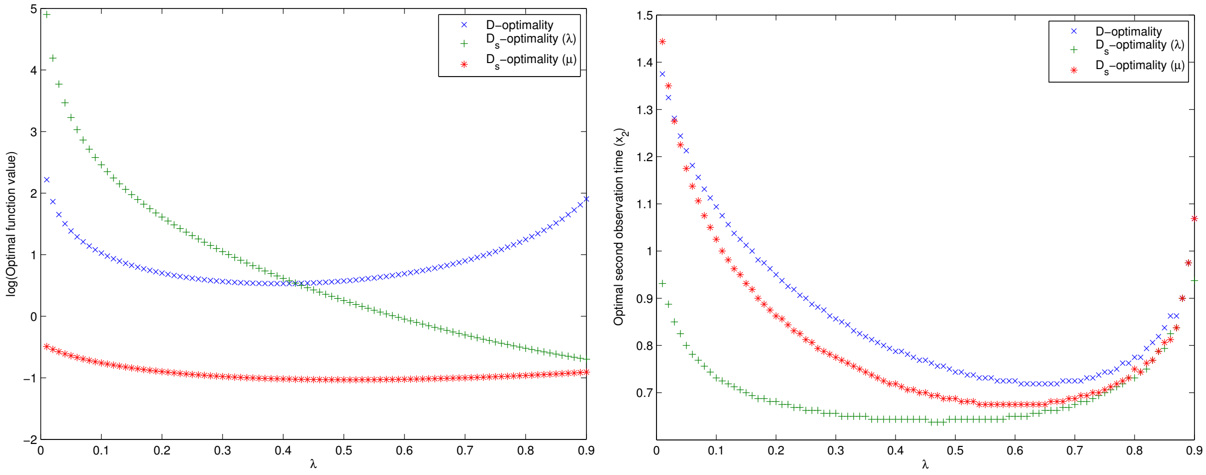

Fig. 4 Information values for varying spacings. Interfering case, $n = 2$: (Left) Optimal spacing value (Right)

5.3.1 Two observations

We repeat the analysis as per the non-interfering case above for the interfering case; we present as Figure 4 the optimal design found for distance between two observations, together with the optimal function value for $n = 2$ found again using Nelder-Mead.

We again present as Figure 4 the information function value for $n = 2$ observations, here for the interfering case. We see a broadly similar pattern as before, with the optimal spacing decreased somewhat for estimating λ alone, and increased somewhat for estimating μ alone, and for both parameters simultaneously. We also, in general, get slightly more information about μ when observations interfere with the queue; this is because we have more customers in the queue to observe departing (we are always guaranteed to have at least one after an observation that interferes), so we can make some inference about it. In the non-interfering case, particularly for low loads, we may observe an empty queue, which will give us little information about service rates as there are no customers to depart.

We present as Figure 5 the information function values for varying spacing between observations in the $n = 2$ case; there does not seem to be much qualitative difference for estimating λ through the D_s criterion; intuitively when estimating λ, having more arrivals caused by measurement should make no difference, as long as we know when these measurements occur. However, for low values of λ, the information function curve has become a lot more pronounced for both the D criterion and the D_s criterion for μ. The information function tails off a lot quicker, leading to low efficiency if we use an incorrect measurement rate. This ties in with intuition; for a very heavily loaded queue, an additional observation will not perturb the queue strongly. However, for lightly loaded queues, the addition of a further customer in measurement may have a larger effect.

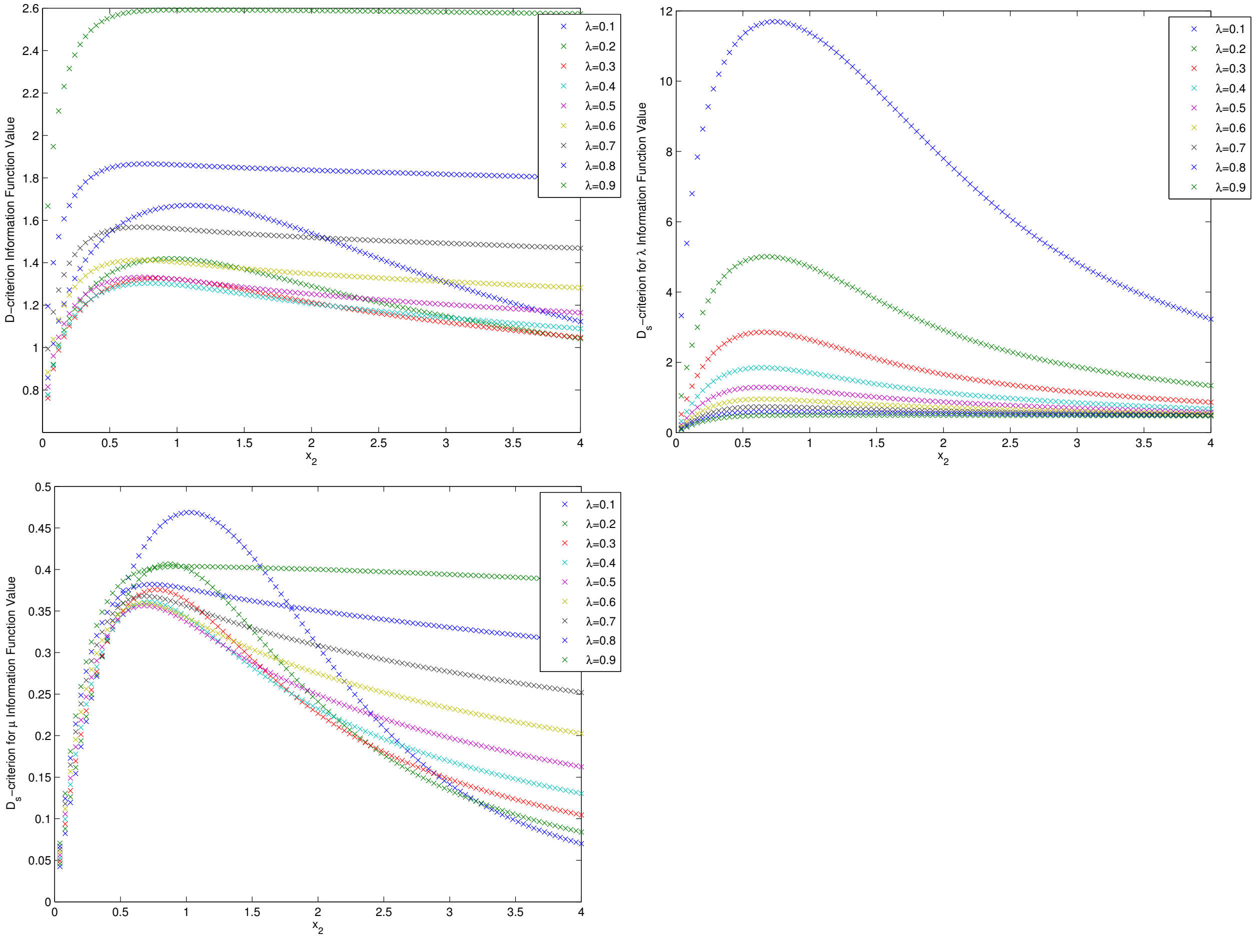

Fig. 5 Information function values for varying spacing between observations. (Left) D-criterion. (Middle) D_s-criterion for λ. (Right) D_s criterion for μ.

We see again that our rule of thumb for measuring the queue at about the same interval as the service rate is not a bad design, at least for this $n = 2$ interfering case.

5.3.2 Exact design for three observations

We can also calculate the information functions for three observations (which yield two time differences $d_1 = x_2 - x_1 = x_2$ and $d_2 = x_3 - x_2$) for ease of display of the information function. The design problem reduces to choosing differences in times d_1 and d_2 to find the maximum value of the information function described in equation (4).

We present contour plots for our three criteria with the x- and y-axis being the differences d_1 and d_2 for arrival rates of $\lambda = 0.3$, 0.5, and 0.7. The results are presented as Figure 6.

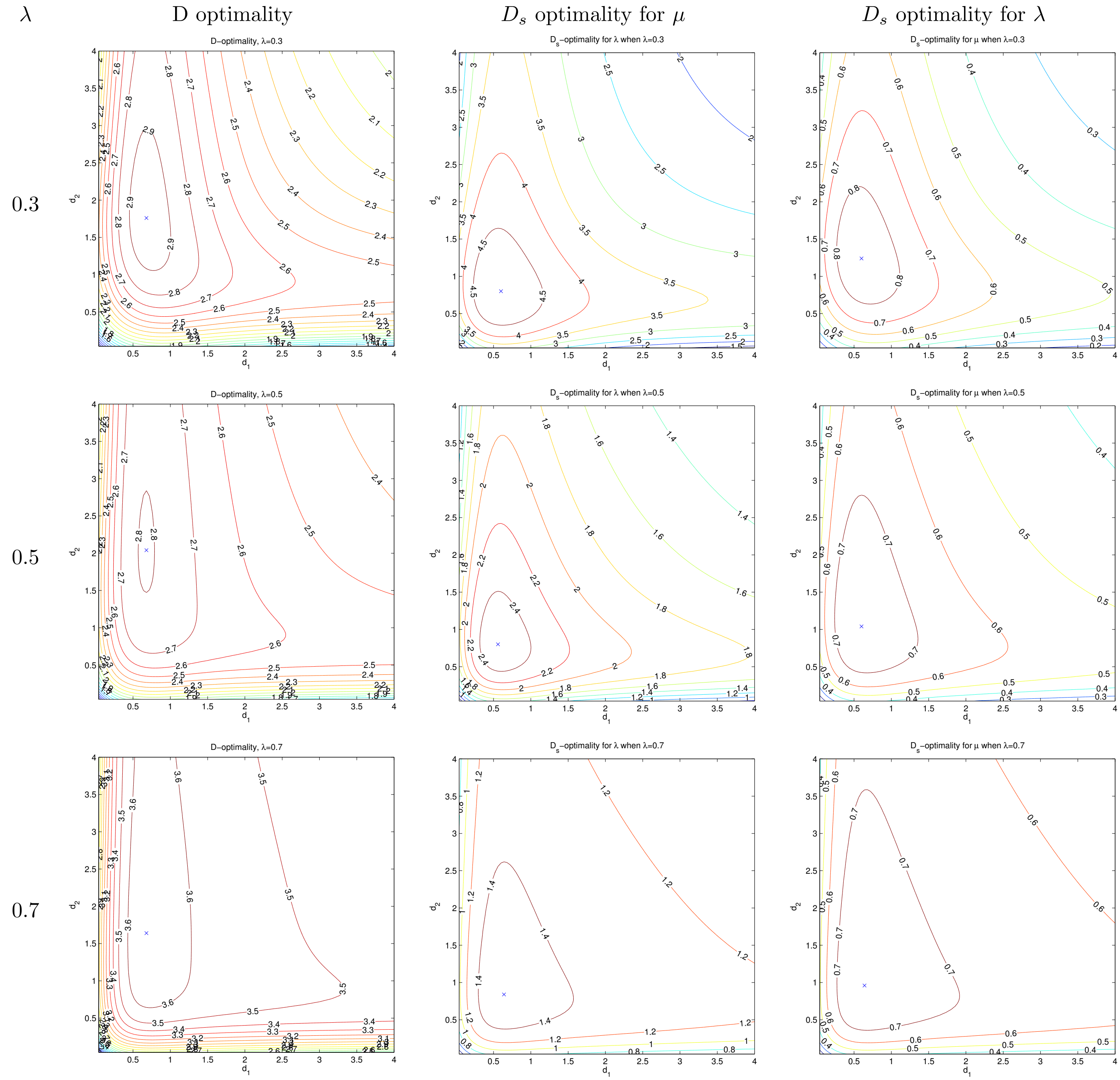

Fig. 6 Optimal designs for D and D_s criteria

A common feature of the information surface is that measuring slightly too quickly produces very bad results for D-optimality, as seen in the sharp gradient in the bottom left of the D-optimality contour plots; however, measuring slightly too slowly does not produce such a rapid tail off in the information function.

Note that, although the queue is Markovian in nature, information and hence the information functions are not in general invariant to changing the order of differences in observations; i.e. $I(0, d_1, d_1 + d_2) \neq I(0, d_2, d_1 + d_2)$. This is demonstrated by the asymmetric nature of the information function contours above.

D	$n=2$		$n=3$		$n=4$		$n=5$	
λ	d_1^*	ψ_1^*	d_i^*	ψ_1^*	d_i^*	ψ_1^*	d_i^*	ψ_1^*
0.1	1.16	1.04	0.88	1.30	0.71	1.51	0.59	1.68
0.2	1.01	1.05	0.81	1.30	0.69	1.50	0.60	1.67
0.3	0.93	1.08	0.78	1.33	0.71	1.53	0.66	1.70
0.4	0.87	1.13	0.77	1.38	0.76	1.58	0.74	1.75
0.5	0.83	1.21	0.76	1.47	0.81	1.67	0.86	1.84
0.6	0.82	1.33	0.76	1.60	0.87	1.81	1.07	1.99
0.7	0.83	1.51	0.78	1.81	0.93	2.03		
0.8	0.90	1.83	0.84	2.19				
0.9	1.73	2.58						

Table 1 Optimal spacings and information function values, non-interfering case, D-optimality

$D_s(\lambda)$	$n=2$		$n=3$		$n=4$		$n=5$	
λ	d_1^*	ψ_1^*	d_i^*	ψ_1^*	d_i^*	ψ_1^*	d_i^*	ψ_1^*
0.1	0.94	2.47	0.64	3.31	0.49	3.93	0.39	4.45
0.2	0.84	1.68	0.58	2.26	0.44	2.69	0.36	3.04
0.3	0.78	1.33	0.55	1.80	0.42	2.14	0.34	2.43
0.4	0.75	1.12	0.53	1.53	0.41	1.83	0.34	2.07
0.5	0.74	0.98	0.54	1.35	0.42	1.62	0.34	1.84
0.6	0.74	0.88	0.56	1.22	0.44	1.47	0.37	1.67
0.7	0.77	0.80	0.59	1.12	0.49	1.35		
0.8	0.84	0.74	0.68	1.04	0.58	1.27		
0.9	1.50	0.70						

Table 2 Optimal spacings and information function values, non-interfering case, D_s-optimality for λ

$D_s(\mu)$	$n=2$		$n=3$		$n=4$		$n=5$	
λ	d_1^*	ψ_1^*	d_i^*	ψ_1^*	d_i^*	ψ_1^*	d_i^*	ψ_1^*
0.1	0.96	0.25	0.66	0.34	0.51	0.40	0.41	0.45
0.2	0.84	0.34	0.58	0.45	0.44	0.54	0.36	0.61
0.3	0.78	0.40	0.54	0.54	0.42	0.64	0.34	0.73
0.4	0.75	0.45	0.53	0.61	0.41	0.73	0.33	0.83
0.5	0.74	0.49	0.53	0.67	0.41	0.81	0.34	0.92
0.6	0.74	0.53	0.55	0.73	0.43	0.88	0.36	1.00
0.7	0.76	0.56	0.59	0.78	0.48	0.95		
0.8	0.84	0.59	0.68	0.83	0.58	1.01		
0.9	1.50	0.63						

Table 3 Optimal spacings and information function values, non-interfering case, D_s-optimality for μ

Also note that there does not exist an optimal design which is equally spaced; the maximum of the contour plot is not on the $d_1 = d_2$ line. The optimal design is not, in general, uniform probing, in this interfering case.

5.4 Optimal design by Nelder-Mead

We present as Tables 1 to 6, the optimal differences between observations for D, and D_s optimality, for $n = 2, 3, 4, 5$ observations, in both the interfering and non-interfering case. These were generated by non-

D	$n = 2$		$n = 3$		$n = 4$		$n = 5$	
λ	d_1^*	ψ_1^*	d_1^*, d_2^*	ψ_1^*	d_1^*, d_2^*, d_3^*	ψ_1^*	$d_1^*, d_2^*, d_3^*, d_4^*$	ψ_1^*
0.1	1.09	2.79	0.81, 1.59	4.78	0.74, 0.99, 2.31	6.72	0.70, 0.87, 1.19, 3.62	8.72
0.2	0.95	2.01	0.73, 1.64	3.44	0.66, 0.97, 3.19	4.90	0.64, 0.83, 1.29, 4.66	6.41
0.3	0.86	1.76	0.69, 1.76	2.95	0.63, 0.98, 4.15	4.23	0.61, 0.81, 1.45, 6.21	5.54
0.4	0.79	1.70	0.67, 1.92	2.77	0.62, 1.00, 5.91	3.97		
0.5	0.74	1.77	0.67, 2.05	2.81	0.63, 1.03, 10.08	4.00		
0.6	0.73	2.00	0.68, 1.97	3.06	0.64, 1.03, 21.03	4.33		
0.7	0.73	2.46	0.70, 1.66	3.66				
0.8	0.78	3.48	0.75, 1.54	5.07				
0.9	1.07	6.72						

Table 4 Optimal spacings and information function values, interfering case, D-optimality

$D_s(\lambda)$	$n = 2$		$n = 3$		$n = 4$		$n = 5$	
λ	d_1^*	ψ_2^*	d_1^*, d_2^*	ψ_2^*	d_1^*, d_2^*, d_3^*	ψ_2^*	$d_1^*, d_2^*, d_3^*, d_4^*$	ψ_2^*
0.1	0.73	3.42	0.71, 0.83	4.05	0.75, 0.82, 0.88	4.51	0.78, 0.83, 0.87, 0.92	4.92
0.2	0.68	2.24	0.62, 0.79	2.79	0.64, 0.74, 0.84	3.16	0.67, 0.75, 0.81, 0.89	3.47
0.3	0.66	1.69	0.58, 0.78	2.21	0.59, 0.71, 0.84	2.56	0.61, 0.70, 0.79, 0.88	2.83
0.4	0.64	1.36	0.57, 0.79	1.85	0.56, 0.70, 0.85	2.18		
0.5	0.64	1.14	0.58, 0.80	1.59	0.56, 0.71, 0.89	1.91		
0.6	0.65	0.98	0.59, 0.82	1.39	0.58, 0.73, 0.95	1.69		
0.7	0.68	0.86	0.63, 0.85	1.23				
0.8	0.73	0.77	0.69, 0.93	1.10				
0.9	0.94	0.71						

Table 5 Optimal spacings and information function values, interfering case, D_s-optimality for λ

$D_s(\mu)$	$n = 2$		$n = 3$		$n = 4$		$n = 5$	
λ	d_1^*	ψ_3^*	d_1^*, d_2^*	ψ_3^*	d_1^*, d_2^*, d_3^*	ψ_3^*	$d_1^*, d_2^*, d_3^*, d_4^*$	ψ_3^*
0.1	1.03	0.68	0.68, 1.48	1.05	0.59, 0.90, 1.85	1.34	0.55, 0.77, 1.04, 2.20	1.59
0.2	0.86	0.64	0.63, 1.32	0.98	0.56, 0.83, 1.80	1.25	0.52, 0.71, 0.96, 2.31	1.49
0.3	0.78	0.61	0.61, 1.23	0.93	0.55, 0.79, 1.81	1.19	0.52, 0.69, 0.93, 2.45	1.43
0.4	0.72	0.60	0.60, 1.14	0.90	0.55, 0.78, 1.83	1.15		
0.5	0.69	0.60	0.60, 1.05	0.88	0.56, 0.78, 1.83	1.12		
0.6	0.68	0.60	0.61, 0.99	0.88	0.58, 0.78, 1.77	1.10		
0.7	0.69	0.61	0.64, 0.95	0.88				
0.8	0.74	0.62	0.69, 0.98	0.89				
0.9	0.94	0.64						

Table 6 Optimal spacings and information function values, interfering case, D_s-optimality for μ.

linear optimisation by Nelder-Mead as discussed above. When no result is shown, the Nelder-Mead algorithm did not converge to find an optimal design within 24 hours of CPU time. See Section 5.1 above.

In the non-interfering case (Tables 1 to 3) we see results with a similar pattern to that already seen; the optimal spacing is further apart for both high and low λ (relative to $\mu = 1$). As the number of observations increases, the optimal spacing increases slightly.

The pattern of results for the interfering case (Tables 4 to 6) is not difficult to interpret. For instance, reading from Table 6 for $n = 3$ and $\lambda = 0.4$, optimal spacings are 0.60 and 1.14, thus the D_s optimal design has observation times 0, 0.60, and 1.74. We see that optimal spacing between design points is not regular. For estimating λ alone the patterns are most regular; for estimating μ they are less regular, and for D-optimality they are highly irregular.

In particular, D-optimality seems to be particularly interesting in that observations often occur a long time after previous ones. Note for $n = 4$ and $n = 5$ the high value of the last spacing, indicating that one point is to be chosen a large distance away from the previous points. We believe this result hints at the following intuition: we gain some information about a queue's behaviour either by observing transitions over a short period and seeing the short term behaviour of a queue, or by making observations far enough apart in time for them to be functionally independent, so that we are sampling from the stationary distribution of the queue. As in this case measuring a queue interferes with it, we need to allow time for the queue to return to its stationary distribution before measuring again. We saw a similar result for observing Markov chains in [31] and [32].

5.5 Bayesian designs

The optimal designs found above share a feature common with many other problems in non-linear optimal design: the optimal design for determining the unknown parameter depends on that unknown parameter, i.e. the information about θ depends on the unknown θ which we are trying to measure. To recommend a practical design for a practitioner to use is therefore still not possible without further assumptions.

Spall [40] suggests three practical approaches to this problem:

1. Find a design which is locally optimal based on a nominal value of θ, hoping this design is near-optimal across the parameter space of θ.
2. Use a sequential design.
3. Assign a prior to θ, and use a Bayesian strategy.

We have shown above, e.g in Figure 3, the difference in the function values for varying parameter values. For example, if we were to pick a D_s-optimal design point believing $\lambda = 0.1$, when λ was in reality 0.9, we would pick a suboptimal design. As, in practice, μ is also unknown to us, we might pick a very inefficient design. See [16] for more discussion about an optimal design.

This indicated that, in some cases, a local design might be rather poor ; we resolve it here by the common method of assuming a prior $p(\boldsymbol{\theta})$ for our unknown θ. We present a "pseudo"-Bayesian approach here, as we are only using the prior distribution in order to design the experiment assuming we will perform maximum likelihood estimation.

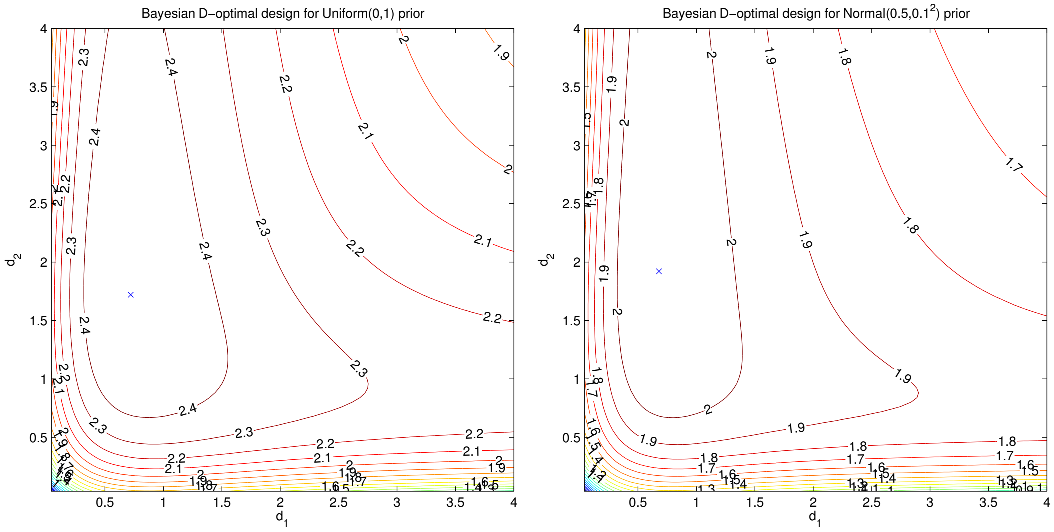

Fig. 7 Contour plots of D function values for $n = 3$ observations for varying Bayesian priors on λ

- Uniformly distributed in the range (0.05,0.85) (a vague prior on this range) (left panel);
- Normally distributed with mean 0.5 and standard deviation 0.1 (right panel);

Chaloner and Verdinelli [14] define the design $\boldsymbol{x}$ which maximizes

$$\psi_{BD}(\boldsymbol{x}) = \int_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \log\det\left(I(\boldsymbol{x}|\boldsymbol{\theta})\right)p(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{11}$$

as the Bayesian D-optimal design. The gain in information between prior and posterior above is equivalent to the Shannon Information. We can thus, for each $\boldsymbol{x}$, evaluate $\psi_{BD}(\boldsymbol{x})$ and find the design which maximises this criterion. For this approach and for most practical problems we require that $\{\lambda \geq \mu\} \cap \boldsymbol{\Theta} = \emptyset$, i.e there is no prior probability that the service rate is less than the arrival rate. This is consistent with the probability density in Equation (3) which is only valid when $\lambda < \mu$. We once again without loss of generality scale $\mu = 1$.

In practice, apart from for trivial priors, this calculation is not tractable analytically, but we can form a numerical approximation to Equation (11) by discretizing over a suitable grid for θ of grid size δ to get an approximate calculation for the optimality criterion as

$$\psi^{\Delta}_{BD}(\boldsymbol{x}) = \sum_{0\leq i<\lfloor 1/\Delta+1\rfloor} \log\det[I(\boldsymbol{x}|\lambda = \Delta i)]p(\Delta(i-0.5) \leq \lambda < \Delta(i+0.5))$$

We present as Figures 7,8 and 9 the D-optimal information contour plot for grid size (of θ) of $\Delta = 0.1$ for several simple priors on λ. The optimal design (marked with a “x”) is fairly consistent across all priors. As we have seen before, distances between observations that are too small lead to a much worse design than

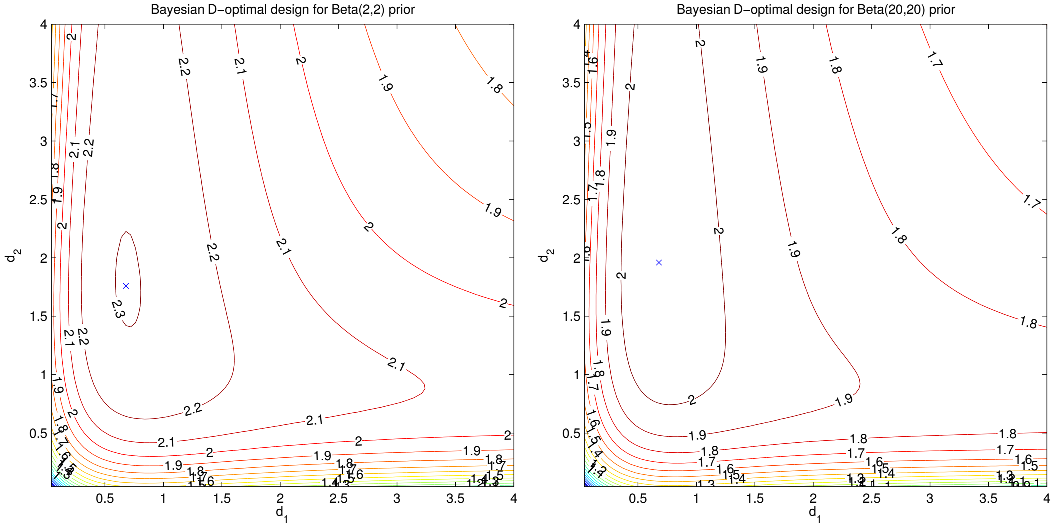

Fig. 8 Contour plots of D function values for $n = 3$ observations for varying Bayesian priors on λ

– Beta(2,2) distributed with endpoints scaled to be (0.05,0.85) (left panel);
– Beta(20,20) distributed with endpoints scaled to be (0.05,0.85) (right panel).

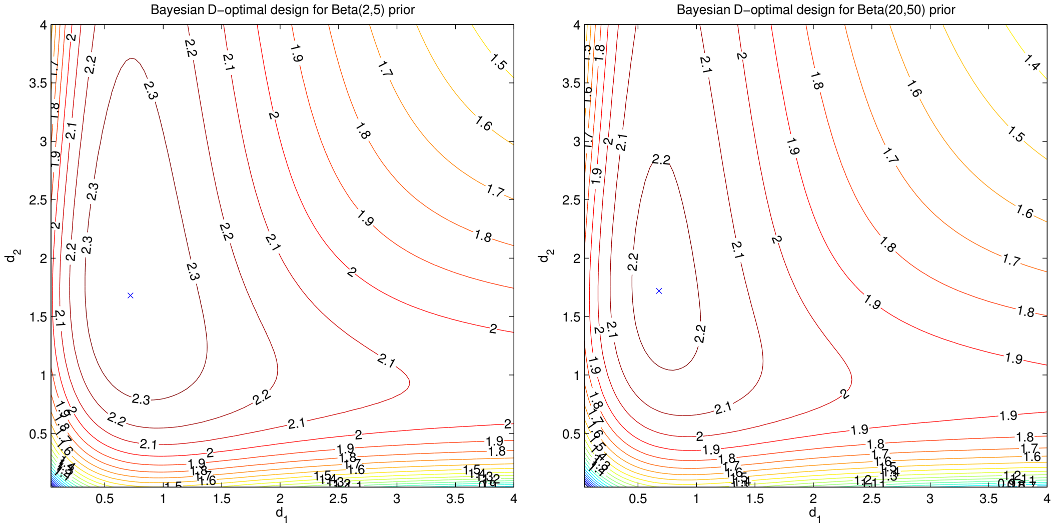

Fig. 9 Contour plots of D function values for $n = 3$ observations for varying Bayesian priors on λ

– Beta(2,5) distributed with endpoints scaled to be (0.05,0.85) (left panel);
– Beta(20,50) distributed with endpoints scaled to be (0.05,0.85) (right panel).

those that are too large. It is interesting to note how similar the plots are, and that the optimal designs are not sensitive to choice of any of these quite different priors.

Recall that a Beta(α, β) distribution has mean $\alpha/(\alpha+\beta)$ and variance $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$. Hence the left and right plots of Figure 8 have the same mean value for the unknown $\lambda = 0.5$, but the variance decreases from 0.05 to 0.005, i.e. we are more certain about the value of unknown λ in the right plot than in the left. This increase in certainty does not change the overall shape of the contour plots, although does seem to suggest that the optimal design moves from approximately $(d_1, d_2) = (0.68, 1.76)$ to $(d_1, d_2) = (0.68, 1.96)$

As we go from Figure 8 to Figure 9, the mean of the distributions changes from 0.5 to 0.29. This decrease in mean seems here seems to increase the density of the contour plot, meaning that the plot is steeper and there is more difference in information between the optimal values and sub-optimal values. We have seen previously that we in general have more certainty about the design for low values of λ (the variation in the queue is simply lower for low load) so this ties in with our intuition as well.

We have presented priors here for a number of particular cases, and each of these is computationally expensive as it involves calculating information contour plots for each point in the discretized grid for θ (in the examples above, 8 function values). We might extend this technique by making a finer discretization, although the increased computation would, except in simple cases, suggest a sampling (Monte Carlo) approximation to equation 11. This technique is discussed in, for example, Atkinson et al.[4] (section 18.5). Although computationally difficult, once contour plots can be calculated for any $\boldsymbol{\theta}$ as discussed in 5, combining these with a prior is a relatively simple numerical calculation.

6 Conclusions

There has been limited previous work on optimal design of measurement times for stochastic systems, and particularly for the M/M/1 queue we have studied here. We have adapted a method to evaluate the autocovariance function of the queue quickly, and used this to numerically calculate the optimal designs for this queue.

We have shown that, whilst uniform designs are appropriate for observing a queue where the observations do not cause interference, we must use a non-standard pattern for observation in the case where observations interfere with the queue. Our numerical results present some insights into the features of a good measurement regime, for the simple structure of the M/M/1 queue.

We have studied the more practical case of a Bayesian design where we have some uncertainty on the unknown parameters we wish to measure; throughout this work we have shown that measuring at the wrong rate will produce estimators with much bigger variance, compared to adopting a more considered measurement regime.

Further research might usefully focus on what properties of a stochastic system or queue determine the optimal design, and how these designs might be calculated more efficiently for systems with complicated autocovariance functions.

Acknowledgements The authors acknowledge the use of the IRIDIS High Performance Computing Facility, and associated support services at the University of Southampton, in the completion of this work.

References

1. J. Abate and W. Whitt. Asymptotics for M/G/1 low-priority waiting-time tail probabilities. *Queueing Systems*, 25(1):173233, 1997.
2. S. Acharya. On normal approximation for maximum likelihood estimation from single server queues. *Queueing Systems*, 313, 1999.
3. D. Aigner. Parameter estimation from cross-sectional observations on an elementary queuing system. *Operations Research*, 22, 1974.
4. A. Atkinson, A. Donev, and R. Tobias. *Optimum Experimental Designs, with SAS.* Oxford University Press, 2007.
5. F. Baccelli, S. Machiraju, D. Veitch, and J. Bolot. The role of PASTA in network measurement. *ACM SIGCOMM Computer Communication Review*, 36(4):231–242, 2006.
6. F. Baccelli, S. Machiraju, D. Veitch, and J. Bolot. On optimal probing for delay and loss measurement. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 291–302. ACM, 2007.
7. F. Baccelli, S. Machiraju, D. Veitch, and J. Bolot. Probing for loss: The case against probe trains. *Communications Letters, IEEE*, 15(5):590592, 2011.
8. I. Basawa, U. Bhat, and R. Lund. Maximum likelihood estimation for single server queues from waiting time data. *Queueing Systems*, 24:155–167, 1997.
9. I. Basawa, U. Bhat, and J. Zhou. Parameter estimation in queueing systems using partial information. Technical report, Ohio State University, June 2006.
10. I. Basawa and N. Prabhu. Estimation in single server queues. *Naval Research Logistics Quarterly*, 28(3), 1981.
11. I. Basawa and N. Prabhu. Large sample inference from single server queues. *Queueing Systems Theory and Applications*, 3(4):289–304, 1988.
12. P. Billingsley. *Statistical Inference for Markov Processes.* The University of Chicago Press, 1962.
13. S. Bodas, D. Shah, and D. Wischik. Congestion control meets medium access: throughput, delay, and complexity. In *Proceedings of the 12th ACM SIGMETRICS/PERFORMANCE joint international conference on Measurement and Modeling of Computer Systems*, page 399400, 2012.
14. K. Chaloner and I. Verdinelli. Bayesian experimental design: a review. *Statistical Science*, 10(3):273–304, 1995.
15. T. Chen. Parameter estimation for partially observed queues. *IEEE Transactions on Communications*, 42 (9):2730–2739, 1994.
16. H. Chernoff. Locally optimal designs for estimating parameters. *The Annals of Mathematical Statistics*, 24(4):586–602, 1953.
17. A. Clarke. Maximum likelihood estimates in a simple queue. *The Annals of Mathematical Statistics*, 28(4):1036–1040, Dec 1957.
18. E. Coffman, P. Robert, F. Simatos, S. Tarumi, and G. Zussman. A performance analysis of channel fragmentation in dynamic spectrum access systems. *Queueing Systems*, page 128, 2012.
19. M. Grossglauser and D. N. Tse. A time-scale decomposition approach to measurement-based admission control. *IEEE/ACM Transactions on Networking (TON)*, 11(4):550–563, 2003.

20. J. Jenkins. The relative efficiency of direct and maximum likelihood estimates of mean waiting time in the simple queue M/M/l. *Journal of Applied Probability*, 9(2):396–403, 1972.
21. A. Khisti, C. Huitema, and A. Dube. Controlling admission of data streams onto a network based on end-to-end measurements, July 2007.
22. L. Kleinrock. *Queueing Systems: Theory Vol 1.* John Wiley & Sons Inc, 1975.
23. J. C. Lagarias, J. A. Reeds, M. H. Wright, and P. E. Wright. Convergence properties of the Nelder-Mead simplex method in low dimensions. *SIAM Journal of Optimization*, 9:112147, 1998.
24. W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson. On the self-similar nature of ethernet traffic (extended version). *Networking, IEEE/ACM Transactions on*, 2(1):115, 1994.
25. S. Love, G. Pollock, P. Goldsack, and E. Kirshenbaum. System and method for monitoring communication networks using data stream characterization, June 2005.
26. P. M. Morse. Stochastic properties of waiting lines. *Journal of the Operations Research Society of America*, 3(3):255–261, 1955. ArticleType: research-article / Full publication date: Aug., 1955 / Copyright 1955 INFORMS.
27. M. Nilsson. Measuring available path capacity using short probe trains. In *Network Operations and Management Symposium (NOMS), 2010 IEEE*, page 910913, 2010.
28. I. Norros. A storage model with self-similar input. *Queueing systems*, 16(3):387396, 1994.
29. D. Pagendam and P. Pollett. Optimal sampling and problematic likelihood functions in a simple population model. *Environmental Modeling and Assessment*, 14(6):759–767, 2009.
30. D. Pagendam and P. Pollett. Robust optimal observation of a metapopulation. *Ecological Modelling*, 221(21):2521–2525, Oct. 2010.
31. B. Parker, S. Gilmour, and J. Schormans. Measurement of packet loss probability by optimal design of packet probing experiments. *IET Communications*, 3(6):979, 2009.
32. B. M. Parker, S. G. Gilmour, and J. A. Schormans. Design of experiments for categorical repeated measurements in packet communication networks. *Technometrics*, 53(4):339–352, November 2011.
33. C. Raiciu, S. Barre, C. Pluntke, A. Greenhalgh, D. Wischik, and M. Handley. Improving datacenter performance and robustness with multipath TCP. In *ACM SIGCOMM Computer Communication Review*, volume 41, page 266277, 2011.
34. J. Reynolds. The covariance structure of queues and related processes: A survey of recent work. *Advances in Applied Probability*, 7(2):383–415, Jun., 1975.
35. J. Ross, D. Pagendam, and P. Pollett. On parameter estimation in population models II: multi-dimensional processes and transient dynamics. *Theoretical Population Biology*, 75(23):123–132, May 2009.
36. M. Roughan. Fundamental bounds on the accuracy of network performance measurements. *Proceedings of the 2005 ACM SIGMETRICS international conference on Measurement and modeling of computer systems*, pages 253–264, 2005.
37. M. Roughan. A Comparison of Poisson and Uniform Sampling for Active Measurements. *IEEE Journal on Selected Areas in Communications*, 24(12):2299–2312, 2006.
38. D. Shah and D. Wischik. Fluid models of congestion collapse in overloaded switched networks. *Queueing Systems*, 69(2):121143, 2011.
39. D. Shah and D. Wischik. Log-weight scheduling in switched networks. *Queueing Systems*, page 140, 2012.
40. J. Spall. *Introduction to Stochastic Search and Optimization: Estimation, Simulation and Control.* John Wiley & Sons Inc, 2003.
41. The ISP Column. The QoS emperor's wardrobe. `http://www.internetsociety.org/publications/isp-column-june-2012-qos-emperors-wardrobe-1`, June 2012.
42. R. Wolff. Poisson Arrivals See Time Averages. *Operations Research*, 30(2):223–231, 1982.