Adv. Appl. Prob. **53,** 687–715 (2021)
*doi:*10.1017/apr.2020.72

VARIATIONAL INFERENCE FOR MARKOVIAN QUEUEING NETWORKS

IKER PEREZ,[*] *University of Nottingham*
GIULIANO CASALE,[**] *Imperial College London*

Abstract

Queueing networks are stochastic systems formed by interconnected resources routing and serving jobs. They induce jump processes with distinctive properties, and find widespread use in inferential tasks. Here, service rates for jobs and potential bottlenecks in the routing mechanism must be estimated from a reduced set of observations. However, this calls for the derivation of complex conditional density representations, over both the stochastic network trajectories and the rates, which is considered an intractable problem. Numerical simulation procedures designed for this purpose do not scale, because of high computational costs; furthermore, variational approaches relying on approximating measures and full independence assumptions are unsuitable. In this paper, we offer a probabilistic interpretation of variational methods applied to inference tasks with queueing networks, and show that approximating measure choices routinely used with jump processes yield ill-defined optimization problems. Yet we demonstrate that it is still possible to enable a variational inferential task, by considering a novel space expansion treatment over an analogous counting process for job transitions. We present and compare exemplary use cases with practical queueing networks, showing that our framework offers an efficient and improved alternative where existing variational or numerically intensive solutions fail.

Keywords: Queueing networks; variational methods; Markov jump process

2010 Mathematics Subject Classification: Primary 60J27
Secondary 60J28

1. Introduction

Queueing networks (QNs) are systems of theoretical and practical interest in the design of computing systems [14], as well as in the optimization of business processes arising in factories, shops, offices, and hospitals [6, 15, 21]. They are formed by interconnected resources routing and serving jobs, and their behaviour often gives rise to complex families of stochastic (jump) processes. In applications, they provide the means to assess modifications, diagnose performance, and evaluate robustness in multiple service infrastructures.

Formally, a QN is associated with a (Markov) jump process: the piecewise-deterministic model for job counts stationed at each resource. Here, a resource consists of a *service station* along with a *queue*, because the processing capabilities of a station are usually limited.

Received 27 June 2019; revision received 3 November 2020.

[*] Postal address: School of Mathematical Sciences, University of Nottingham, Nottingham NG7 2RD, United Kingdom.

[**] Postal address: Department of Computing, Imperial College London, London SW7 2RH, United Kingdom.

Whenever a job is serviced, it is routed from its current resource into a new queue, and we say that the process has *jumped*. Hence, a jump adds a count unit to one resource by subtracting it from another, and we say that jumps in QNs are *coupled*. Finally, the process is parametrized by the set of *intensity* rates, which determine the pace of service across job types and the multiple stations.

Performance evaluation. In applications, the service rates in a QN may be unknown, and the various bottlenecks in the routing mechanism for jobs (across queues) are hard to anticipate. Consequently, network deployments are periodically monitored to collect *performance* measurements, which often include queue lengths, visit counts, or response times. These measurements are a consequence of the stochastic evolution of the underlying Markov model; thus, they form the quantitative basis for drawing inferences on multiple aspects of the network. This inferential process is referred to as *performance evaluation* [8]; it is not limited to estimating the service rates, but also includes (i) quantifying any uncertainty around estimations and (ii) identifying queue length distributions under varying networks conditions.

From a probabilistic viewpoint, the primary inferential objectives are to parametrize the *regular conditional probability* for stochastic network *trajectories* or *paths*, conditioned on the available performance measurements; and to derive closed-form expressions for the induced distributions over (i) intensity rates and (ii) queue lengths over time. In this paper, we will observe that this goal calls for the evaluation of complex intractable integrals, a problem commonly encountered within Bayesian statistical settings [1]. In that context, the approach is generally to deviate from analytical methodology and instead construct numerical representations of the induced distributions (such as density plots and histograms), using intensive *Markov chain Monte Carlo* (MCMC) procedures [22, 27]. However, numerical methods must sample the trajectories of the underlying stochastic model, conditioned on measurements. These work well in analogous problems with jump processes observed in mathematical biology [12] or genetics [11], but scale poorly to complex multivariate settings common in QNs [5], which account for coupled jumps, job priorities, service types, and feedback loops. Also, numerical methods are reliant on *sufficient* readily available measurements to describe the likely behaviour of the system. In real-world applications, performance measurements are scarce [27]; for instance, in large externally managed implementations the monitoring may only be executed *end-to-end* [17], i.e., exclusively for the input and output resources where jobs enter and depart the QN.

Consequently, pragmatic statistical solutions are restricted in scope and often ignore the temporal component of the Markov process. Instead, they construct varied inferential methodologies by relying on the mass representation for the network's stationary distribution [16], and the objective is usually only to offer point estimates for the unknown service rates (see [26] for a review).

1.1. Approximate inferential methods

In order to derive closed-form expressions for the induced conditional distributions over service rates or network trajectories, we must resort to approximate inferential settings of significant probabilistic complexity. Here, the idea is to operate under some (suitably parametrized) *approximating* measure in the probability space. This approximating measure should define alternative sample trajectories for the underlying Markov jump model (for job counts across resources) and should ensure analytical tractability of the various

https://doi.org/10.1017/apr.2020.72 Published online by Cambridge University Press

integrals that arise in the inferential problem. However, the reader will note that closed-form formulae derived under some (conveniently defined) approximating measure will not be necessarily representative of the real induced distributions. An important step under an approximate inferential set-up is to minimize the dissimilarity (or *divergence*) between (i) the proposed approximating measure and (ii) the true *regular conditional probability* over events, conditioned on performance metrics (which also defines a measure).

This approximate procedure is a popular technique in Bayesian statistics, and the resulting closed-form representations are commonly termed *variational* approximations [4], because their derivation relies heavily on functional optimization methods from the field of calculus of variations. The choices made in order to parametrize suitable approximating measures are key to success; however, the entire process is convoluted, and there exist no guidelines suited to every kind of stochastic model. The standard choice in the statistical literature relies on the strongest of all independence assumptions (cf. [7, 18]), describing each marginal resource (within the multivariate Markov jump model) as a completely independent stochastic process. The underlying parametrization is commonly referred to as the *mean-field* approximating measure, because it draws from the physical *mean-field theory*, where high-dimensional stochastic models are studied through approximations of multiple independent models in combination. Overall, the method works well in practice and has been shown to retrieve accurate conditional density representations in multiple application domains, including predator–prey models, epidemics, networks, and similar processes [7, 18, 19, 28, 30, 31].

However, an important commonality across all Markov models addressed in prior work is that jumps in the marginal components take place one at a time. This ensures that the multivariate stochastic process can indeed be approximated by a collection of independent univariate models. Our intend is to approximate the stochastic dynamics for jobs counts across resources in a QN, which are subject to coupled jumps; here, the reader may note that it is probabilistically infeasible to assume full independence, for two independent jump models may not change state at the same time. Currently, there exists no viable alternative or solution that can facilitate an approximate inferential set-up tailored to multivariate systems such as QNs.

Key contributions. In this paper, we will formally describe why coupled jumps hinder the construction of approximating measures as commonly defined in the literature. We will then prove that it is still possible to efficiently enable the approximate inferential task, by considering a novel space expansion treatment of the underlying jump process; in a nutshell, we will do the following:

1. Shift the scope. Traditionally, a QN is formalized as a counting process of jobs *queueing* in the resources. We instead address a process of job *transition* counts across resources.

2. Augment the support space, so that job counts across queues can become negative.

Finally, we present use cases of our proposed procedure for performance evaluation tasks, applied to example inferential problems where (i) numerical methods do not scale and (ii) approximate variational procedures are unusable. The results within this paper are relevant for single- or multi-class Markovian systems (with exponential or phase-type service times), with either finite or infinite processors, and multiple types of service disciplines and probabilistic routings.

Structure. The rest of the paper is organized as follows. In Section 2 we offer a hierarchical formulation of a queueing system, along with the problem statement. Section 3 introduces an

https://doi.org/10.1017/apr.2020.72 Published online by Cambridge University Press

approximating network model and offers a summary of the main results to be presented later in the paper. Sections 4 and 5 include the main contributions of our work; they discuss the treatment of the network system by means of interactions in network resources, and they further present the results, proofs, and technical details that contribute to later algorithmic constructions. In Section 6, we guide the reader through applications of our results within inferential and network evaluation tasks, and in Section 7 we conclude the paper with a discussion.

2. Queueing systems and jump processes

In the following, we employ shorthand notation for densities, base measures, and distributions whenever these are clear from the context. From here on, let $(\Omega, \mathcal{F})$ denote a measurable space with the regular conditional probability property, supporting the various rates, trajectories, and observations. A general-form QN comprises some $M \in \mathbb{N}$ service stations along with a *set of job classes* $\mathcal{C}$. The stations are connected by a *network topology* that governs the underlying routing mechanisms; when a job has been serviced in one station, it can either queue for service at a different station or depart the network. Such a topology is often defined as a set of *routing probability matrices* $\{P^c\}_{c\in\mathcal{C}}$, with elements $p^c_{i,j}$, for all $0 \leq i, j \leq M$, that denote the probability that a job in class $c \in \mathcal{C}$ will immediately transit to queueing station j after service completion in station i. In open queueing systems, the index 0 is used as a virtual external source (and destination) of job arrivals to (and departures from) the network. In closed systems, this index may either not exist, or instead refer to a delay server that routes departing jobs back into the network. Also, it holds that $\sum_{j=0}^{M} p^c_{i,j} = 1$, for all $0 \leq i \leq M$, $c \in \mathcal{C}$.

We address time-homogeneous Markovian systems that are parametrized by exponential inter-arrival and service times, with non-negative rates $\boldsymbol{\mu} = \{\mu^c_i \in \mathbb{R}_+ : 0 \leq i \leq M,\, c \in \mathcal{C}\}$, which may vary across service stations and job classes. The servers in the network stations may have finite or infinite processors, and service disciplines can vary across a range of *processor sharing* (PS) policies, including *first-come first-served* (FCFS) and variations such as *last-come first-served* (LCFS) or *service in random order* (SIRO). In some cases, FCFS processors may require shared processing times across the various job classes (cf. [3]). For simplicity and ease of notation, class *switching*, service *priorities*, and queue-length-dependent service rates are not discussed in detail; however, these follow naturally, and we later present some examples involving them. Under standard exponential service assumptions, the underlying system behaviour is described by a *Markov jump process* $X = (X_t)_{t\geq 0}$ with values defined in a measurable space $(\mathcal{S}, \mathcal{P}(\mathcal{S}))$. Here, $\mathcal{S}$ denotes a countable set of feasible states in the network, usually infinite in open or mixed systems and finite in closed ones; $\mathcal{P}(\mathcal{S})$ denotes the power set of $\mathcal{S}$. We allow for $\mathcal{S}$ to support vectors of integers that represent job counts across the various class types and service stations, and denote by $X^{i,c}_t$ the number of class-c jobs in station $i > 0$ at time $t \geq 0$. Note that here we ignore the queue lengths in the external delay ($i = 0$) within closed systems, since these are uniquely determined given the number of jobs in the remaining stations. The *infinitesimal generator matrix* Q of X is such that

$$\mathbb{P}(X_{t+\mathrm{d}t} = x' | X_t = x) = \mathbb{I}(x = x') + Q_{x,x'}\mathrm{d}t + o(\mathrm{d}t)$$

for all $x, x' \in \mathcal{S}$. This can be an infinite matrix; it is generally sparse, and its entries describe rates for transitions across states in $\mathcal{S}$. Rows in Q must sum to 0 so that $Q_{x,x'} \geq 0$ for all $x \neq x'$, and $Q_x := Q_{x,x} = -\sum_{x'\in\mathcal{S}:x\neq x'} Q_{x,x'}$.

Hence, jumps in the process X are caused by jobs being routed through stations in the underlying network model. We often say that a state $x' \in \mathcal{S}$ is *accessible* from $x \in \mathcal{S}$, and write

https://doi.org/10.1017/apr.2020.72 Published online by Cambridge University Press

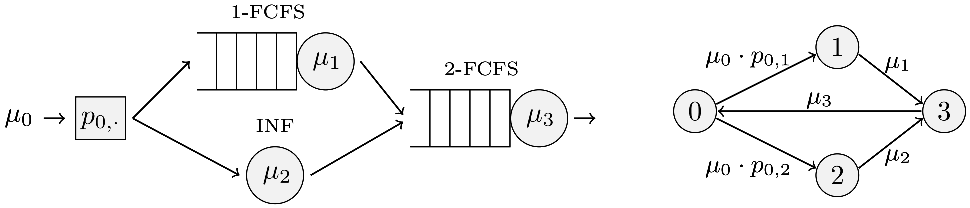

FIGURE 1: Left: open bottleneck network with three service stations. Shaded circles indicate servers; empty rectangles indicate queueing areas. The box is a probabilistic junction for the routing of arrivals. Right: job transition intensities across network stations.

$x \xrightarrow{i,j,c} x'$ for its corresponding jump, if x' may be reached from x by means of a class-c job transition between the stations i and j, in the direction $i \to j$. We further denote by

$$\mathcal{T} = \{(i, j, c) \in \{0, \ldots, M\}^2 \times \mathcal{C} \colon p^c_{i,j} > 0\} \tag{1}$$

the finite set of all *feasible* job transitions in the system, and we remark that the generator Q of X is populated by some positive real-valued rates $\boldsymbol{\lambda} = \{\lambda_\eta \in \mathbb{R}_+ : \boldsymbol{\eta} \in \mathcal{T}\}$ that define the intensities for these job routings, with $\lambda_{i,j,c} = \mu^c_i \cdot p^c_{i,j}$ for all $(i, j, c) \in \mathcal{T}$.

Example 1. In Figure 1 we observe diagrams that illustrate this notation in an open single-class network. On the left, we see three stations with different rates, disciplines, and server counts. The topology P is such that $|\mathcal{T}| = 5$ and $p_{0,1} = 1 - p_{0,2} \in (0, 1)$, $p_{1,3} = p_{2,3} = p_{3,0} = 1$ ($p_{i,j} = 0$ otherwise). On the right, we find the corresponding job transition rates across the four pairs of connected stations. Now, let $\vee$ and $\wedge$ constitute short-hand notation for maxima and minima, respectively. In this single-class example, X monitors counts across the stations, so that $X_t = (X^1_t, X^2_t, X^3_t) \in \mathcal{S}$ for all $t \geq 0$; also, the generator Q is an infinite matrix with $Q_{x,x'} = \lambda_{i,j} \cdot (K_i \wedge x_i)$ for all pairs $x, x' \in \mathcal{S}$ with associated transition $x \xrightarrow{i,j} x'$, where $K_i, x_i \in \mathbb{N}_0$ denote the number of processors and the queue length within station $i \geq 0$. We finally have $K_1 = 1$, $K_2 = \infty$, and $K_3 = 2$; at the virtual source, we always have $K_0 \wedge x_0 = 1$. Thus, note that transition rates in X further depend on the queue lengths and resemble kinetic laws within chemical reaction models [11].

2.1. A hierarchical formulation of queueing systems

Within a hierarchical multilevel formulation, rates in $\boldsymbol{\lambda}$ have a distribution (or *image*) $\mathbb{P}_{\boldsymbol{\lambda}} \equiv \boldsymbol{\lambda}_* \mathbb{P}$ under a reference measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$. We assume this to admit a density $f_{\boldsymbol{\lambda}}$ with respect to a base measure that will further induce (by properties of exponential transitions) distributions over the service rates $\boldsymbol{\mu}$ and routing topology. Next, note that a network trajectory over a finite interval is a piecewise deterministic jump process, such that $X \equiv (\boldsymbol{t}, \boldsymbol{x})$ is represented by a sequence of transition times $\boldsymbol{t}$ along with states $\boldsymbol{x}$. Each pair $(\boldsymbol{t}, \boldsymbol{x})$ is furthermore a random variable on a measurable space $(\mathcal{X}, \Sigma_{\mathcal{X}})$ supporting finite $\mathcal{S}$-valued trajectories, and a conditional density $f_{X|\boldsymbol{\lambda}}$ may be defined with respect to a dominating base measure $\mu_{\mathcal{X}}$, such that the regular conditional probability $\mathbb{P}(A|\boldsymbol{\lambda})$, $A \in \mathcal{F}$, satisfies

$$\mathbb{P}(X^{-1}(B)|\boldsymbol{\lambda}) = \int_B f_{X|\boldsymbol{\lambda}}(\boldsymbol{t}, \boldsymbol{x})\, \mu_{\mathcal{X}}(d\boldsymbol{t}, d\boldsymbol{x})$$

https://doi.org/10.1017/apr.2020.72 Published online by Cambridge University Press

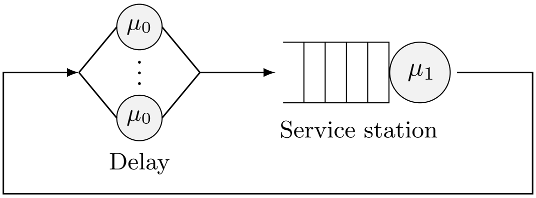

FIGURE 2: Closed QN with a single FCFS service station and a delay.

for all $B \in \Sigma_{\mathcal{X}}$ (see Appendix A for details). In this case,

$$f_{X|\boldsymbol{\lambda}}(\boldsymbol{t}, \boldsymbol{x}) = \pi(x_0)\, e^{Q_{x_I}(T-t_I)} \prod_{i=1}^{I} Q_{x_{i-1},x_i}\, e^{Q_{x_{i-1}}(t_i-t_{i-1})} \tag{2}$$

for every pair of ordered times $\boldsymbol{t} = \{0, t_1, \ldots, t_I\}$ in $[0, T]$ and states $\boldsymbol{x} = \{x_0, \ldots, x_I\}$. Here, $\pi(\,\cdot\,)$ denotes an arbitrary distribution over initial states, and $Q \equiv Q(\boldsymbol{\lambda})$ is the matrix of infinitesimal rates associated with fixed values in $\boldsymbol{\lambda}$. The QN model is thus fully parametrized by a collection of hyperparameters. Analogous modelling choices for continuous-time Markov chains or Markov jump processes can be found in [2, 13] or [32], to name a few.

2.2. Network evaluation and problem statement

Let $T > 0$ denote some arbitrary terminal time and $\boldsymbol{x}_0 \in \mathcal{S}$ an initial state in X. For simplicity, this is assumed to be a 0-valued vector, where no jobs populate the system. Now, let $0 \leq t_1 < \cdots < t_K \leq T$ denote some fixed network monitoring times along with *observation* variables $\{O_k \in \mathcal{O}, k = 1, \ldots, K\}$, for some arbitrary support set $\mathcal{O}$, such that

$$\mathbb{P}\left(\bigcap_{k=1}^{K} O_k^{-1}(\boldsymbol{o}_k) \big| X \right) = \prod_{k=1}^{K} \mathbb{P}(O_k^{-1}(\boldsymbol{o}_k)|X) = \prod_{k=1}^{K} f_{O|X_{t_k}}(\boldsymbol{o}_k) \tag{3}$$

for any sequence of elements $\boldsymbol{o}_1, \ldots, \boldsymbol{o}_K$, where $\boldsymbol{o}_k$ denotes the time-t_k network observation across all stations. Hence, observations are conditionally mutually independent given the network states. The term $f_{O|X_{t_k}}$ stands for a conditional mass function assigned to measurements across the M stations, defined with respect to a counting measure $\mu_{\mathcal{O}}$. In this paper it is assumed that $f_{O|x} > 0$ (everywhere) for all $x \in \mathcal{S}$; however, deterministic observations such as queue lengths can be easily approximated by means of *regularized* indicator functions. Extensions to continuous settings are straightforward.

Example 2. Consider a vanilla closed network as pictured in Figure 2. It includes a single FCFS service station, $K = 1$ processing unit, and a delay routing a population of N jobs in a closed loop. The evolution of $X = (X_t)_{t \geq 0}$ monitors the total number of jobs within the service station, with $X_0 = 0$. Variables $O_k = (O_{0,k}, O_{1,k})$ at times t_k approximate a system with deterministic observations of queue lengths (in both station and delay) and are supported on $\mathcal{O} = \{0, \ldots, N\}^2$. The observation model may thus factor across the network components, so that, for example, $f_{O|x}(\boldsymbol{o}) = \tilde{f}_{O|N-x}(o_0) \cdot \tilde{f}_{O|x}(o_1)$ with $\tilde{f}_{O|x}(o) = \frac{\epsilon}{N} + \mathbb{I}(o = x) \cdot (1 - \frac{N+1}{N}\epsilon)$, and

$$\mathbb{P}(O_k^{-1}(\boldsymbol{o})|X) = \begin{cases} (1-\epsilon)^2, & o_0 = N - X_{t_k},\ o_1 = X_{t_k}, \\ (\epsilon/N)^2, & o_0 \neq N - X_{t_k},\ o_1 \neq X_{t_k}, \\ (1-\epsilon)\cdot \epsilon/N & \text{otherwise}, \end{cases} \tag{4}$$

for a *slack* regularizing parameter $\epsilon > 0$ and all $k = 1, \ldots, K$. This will account for some $\epsilon \cdot 100\%$ faulty measurements. Finally, note that the framework can be generalized to accommodate different observation types. For simplicity of presentation, in this paper we focus on queue lengths.

Now, let $\mathbb{P}(A|\boldsymbol{o}_1, \ldots, \boldsymbol{o}_K)$, $A \in \mathcal{F}$, $(\boldsymbol{o}_1, \ldots, \boldsymbol{o}_K) \in \mathcal{O}^K$, denote the regular conditional probability across global events, conditioned on observations. Our primary interest lies in its induced distribution over the intensity rates (which we denote by $\mathbb{P}_{\boldsymbol{\lambda}|\boldsymbol{o}_1,\ldots,\boldsymbol{o}_K}$) and trajectories. Within (Bayesian) inferential settings, these induced distributions are referred to as *posteriors*. For simplicity, we will restrict the problem statement to intensity rate posteriors; however, the reader will note that an analogous presentation is easily deduced for network trajectories. The posterior distribution exists and admits a density carried by its corresponding *prior* $\mathbb{P}_{\boldsymbol{\lambda}}$ (see Appendix A); moreover, the transformation is proportional to a weighted product of network paths, and is defined by the Radon–Nikodym derivative

$$\frac{d\mathbb{P}_{\boldsymbol{\lambda}|\boldsymbol{o}_1,\ldots,\boldsymbol{o}_K}}{d\mathbb{P}_{\boldsymbol{\lambda}}} = \frac{\int_{\mathcal{X}} \left[\prod_{k=1}^{K} \mathbb{P}(O_k^{-1}(\boldsymbol{o}_k)|\boldsymbol{t}, \boldsymbol{x})\right] f_{X|\boldsymbol{\lambda}}(\boldsymbol{t}, \boldsymbol{x})\, \mu_{\mathcal{X}}(d\boldsymbol{t}, d\boldsymbol{x})}{\mathbb{P}(O_1^{-1}(\boldsymbol{o}_1) \cap \cdots \cap O_K^{-1}(\boldsymbol{o}_K))}, \tag{5}$$

which corresponds to Bayes' equation. There, the denominator denotes a normalizing constant that integrates over trajectories and rates. This transformation will often induce a density representation $f_{\boldsymbol{\lambda}|\boldsymbol{o}_1,\ldots,\boldsymbol{o}_K}$ for the posterior distribution with respect to a suitable (Lebesgue) base. In these cases, we may think of the above derivative as a likelihood ratio. However, this ratio poses a tractability problem; that is, the integral over trajectories cannot be computed analytically and must be approximated. This is a common problem in inferential tasks with jump processes (cf. [12, 24, 23]), and proposed solutions often rely on intensive MCMC procedures that iterate between trajectories and parameters, including direct sampling, rejection sampling, and uniformization-based methods. Yet algorithms may be hard to implement, computationally demanding, or applicable only to reduced classes of problems. In the case of QNs, strong temporal dependencies in the stochastic trajectories X impose hard coupling properties amongst rates and paths [27], which limits the applicability of state-of-the-art numerical solutions to the simplest types of network evaluation problems [22].

Alternatively, the complex integrals in (5) may be approximated through a variational approach, where we suitably parametrize an alternative approximating measure to $\mathbb{P}$ that will decompose the integrand into multiple independent and analytically tractable parts. However, as yet there exists no approximating measure design to achieve this goal, in light of the complexity of jump processes induced by networks of queues. In the following, we present theoretical results leading to a new variational inferential design for use with jump processes, which significantly deviates from and overcomes multiple limitations found in standard methods that resort to full-independence assumptions [18, 19, 7, 28, 30, 31].

3. Overview of results

Note that under the natural measure $\mathbb{P}$ tied to the infinitesimal generator Q, an underlying Markov jump process X as introduced in Section 2 is supported in a set $\mathcal{S}$ of feasible vectors

https://doi.org/10.1017/apr.2020.72 Published online by Cambridge University Press

of integers, which is often just $\mathcal{S} = \mathbb{N}_0^{|\mathcal{C}|\times M}$. Further recall that $\mathcal{T}$ in (1) corresponds to possible transition *routes* in the queueing system, that is, triplets $(i, j, c) \in \{0, \ldots, M\}$ such that a class-c job currently in station i has a non-negative probability $p^c_{i,j} > 0$ of transitioning to station j, under the measure $\mathbb{P}$.

Space expansion. Assume the existence of an approximating measure $\tilde{\mathbb{P}}$ on an augmented space of network paths $\tilde{\mathcal{X}}$, such that we further assign a mass to states with *negative* queue lengths. Under $\tilde{\mathbb{P}}$, rates for transitions across states are induced by a generator $\tilde{Q} = \tilde{Q}(\delta, \boldsymbol{\lambda})$ with

$$\tilde{Q}_{x,x'} = \delta + Q_{x,x'}, \quad \delta > 0, \tag{6}$$

whenever $x \xrightarrow{i,j,c} x'$ and $(i, j, c) \in \mathcal{T}$; that is, whenever a transition from state x to state x' is such that

- a class-c job is serviced at station i and routed to j, and
- $p^c_{i,j} > 0$ within the original probabilistic routing topology under $\mathbb{P}$.

In any other case, it holds that $\tilde{Q}_{x,x'} = Q_{x,x'} = 0$. Note that the original connectivity structure of the QN is preserved under $\tilde{\mathbb{P}}$; however, as described below, jobs may be serviced within queueing stations *even if they do not exist*:

- Following (6), intensities for job transitions between stations i and j are strictly positive whenever $p^c_{i,j} > 0$. If no class-c job exists within station i, then this intensity corresponds to $\delta > 0$.
- In the event of a time-t class-c job departure from station i when $\lim_{s\to t} X^{i,c}_s \le 0$ (the job does not exist), we assume this job to be *virtually generated*, and

  $$X^{i,c}_t = \lim_{s\to t} X^{i,c}_s - 1,$$

  i.e., a unit is subtracted from the state vector at the corresponding index.
- This preserves the global job population count and forces network states to accommodate negative queue lengths.

For values of δ small enough, the $\tilde{\mathbb{P}}$-density assigned to trajectories outside of $\mathcal{X}$ is negligible. Note that a density $\tilde{f}$ in (2) with generator $\tilde{Q}$ in (6) is such that, for any network path $(\boldsymbol{t}, \boldsymbol{x}) \in \tilde{\mathcal{X}} \backslash \mathcal{X}$,

$$\tilde{f}_{X|\boldsymbol{\lambda}}(\boldsymbol{t}, \boldsymbol{x}) \le \prod_{i=1}^{I} \tilde{Q}_{x_{i-1},x_i} = O(\delta^r) \quad \text{as} \quad \delta \to 0$$

for some $r \in \{1, \ldots, I\}$. Thus, $X_* \tilde{\mathbb{P}}(\mathcal{X}) = 1 - \int_{\tilde{\mathcal{X}}\backslash\mathcal{X}} \tilde{f}_{X|\boldsymbol{\lambda}} d\tilde{\mu}_{\mathcal{X}} \xrightarrow{\delta\to 0} 1$, where $\tilde{\mu}$ denotes an appropriately augmented base measure, and the limiting system dynamics under $\tilde{\mathbb{P}}$ will offer a perfect approximation to the original network model. The rest of the paper proceeds as follows:

- In Section 4, we present a counting process Y over job transitions in the augmented network with generator $\tilde{Q}$ in (6), and parametrize an alternative absolutely continuous approximating measure $\mathbb{Q}$. In Lemma 1, we derive a lower bound to the equivalent log-likelihood for the network measurements.

https://doi.org/10.1017/apr.2020.72 Published online by Cambridge University Press

- Propositions 1 and 2 within Section 5 inspect the structure of $\mathbb{Q}$ that best approximates the target regular conditional probability. Corollaries 1 and 2 focus on the rate density $d\mathbb{P}_{\boldsymbol{\lambda}|\boldsymbol{o}_1,\ldots,\boldsymbol{o}_K}$ by looking at conjugacy properties and limiting behaviour as $\delta \to 0$.
- Finally, Section 6 describes applications of our results within inferential tasks, allowing the approximation of (image) measures across the various service rates μ, routing probabilities in $\mathcal{P}$, and trajectories X, conditioned on network measurements. This includes comparisons with existing alternative methods.

4. A counting process over job transitions

A network system as introduced in Section 2 gives rise to a multivariate Markov counting process $Y = (Y_t)_{t\geq 0}$ on $(\Omega, \mathcal{F})$, where each indexed $Y_t = (Y_t^{\boldsymbol{\eta}})_{\boldsymbol{\eta}\in\mathcal{T}} \in \mathcal{S}^Y$ accounts for job transitions across all classes in $\mathcal{T}$, up to a time $t \geq 0$. That is, each $Y_t^{\boldsymbol{\eta}}$ denotes the cumulative count in Y of transitions $x \xrightarrow{\boldsymbol{\eta}} x'$ in X, with $x, x' \in \mathcal{S}$, and $Y_0^{\boldsymbol{\eta}} = 0$ for all $\boldsymbol{\eta} \in \mathcal{T}$. At a basic level, these are simply non-decreasing counting processes for job transitions in the directions defined within $\mathcal{T}$. We further note that $|\mathcal{T}|$ is often small, as underlying network topologies impose strict routing mechanisms. The support set $\mathcal{S}^Y$ for the counting process is determined by the connectivity structure amongst the stations. Under the approximating measure $\tilde{\mathbb{P}}$, it holds that $\mathcal{S}^Y = \mathbb{N}_0^{|\mathcal{T}|}$, since job transitions may occur even if origin stations have no jobs queueing. Now, let

$$\mathcal{T}_{i,c}^{\leftarrow} = \{\boldsymbol{\eta} \in \mathcal{T} : \eta_2 = i,\, \eta_3 = c\} \quad \text{and} \quad \mathcal{T}_{i,c}^{\rightarrow} = \{\boldsymbol{\eta} \in \mathcal{T} : \eta_1 = i,\, \eta_3 = c\}$$

respectively denote the subsets of $\mathcal{T}$ that include transitions of jobs in class $c \in \mathcal{C}$ to and from the network station $i \in \{0, \ldots, M\}$. Also, recall that $X_t^{i,c}$ denotes the number of class-c jobs in station $i > 0$ at time $t \geq 0$; then

$$X_t^{i,c} = \sum_{\boldsymbol{\eta}\in\mathcal{T}_{i,c}^{\leftarrow}} Y_t^{\boldsymbol{\eta}} - \sum_{\boldsymbol{\eta}\in\mathcal{T}_{i,c}^{\rightarrow}} Y_t^{\boldsymbol{\eta}} \tag{7}$$

for all $t \geq 0$, assuming initially empty networked systems. We note that for all $\boldsymbol{\eta} = (i, j, c) \in \mathcal{T}$, it holds that $\boldsymbol{\eta} \in \mathcal{T}_j^{\leftarrow}$ and $\boldsymbol{\eta} \in \mathcal{T}_i^{\rightarrow}$. Thus, paths in X and Y differ in that the former is *coupled*, i.e., a job transition in the direction $\boldsymbol{\eta} = (i, j, c)$ is relevant to (and thus is synchronized across) a pair of marginal processes $(X_t^{i,c})_{t\geq 0}$, $(X_t^{j,c})_{t\geq 0}$; in the latter, this is only relevant to the indexed process $(Y_t^{\boldsymbol{\eta}})_{t\geq 0}$.

In view of (7), we further define $x_{i,c} = \sum_{\boldsymbol{\eta}\in\mathcal{T}_{i,c}^{\leftarrow}} y_{\boldsymbol{\eta}} - \sum_{\boldsymbol{\eta}\in\mathcal{T}_{i,c}^{\rightarrow}} y_{\boldsymbol{\eta}}$ as the class-c queue length in station $i > 0$ for any $y \in \mathcal{S}^Y$. Then the $\tilde{\mathbb{P}}$-associated infinitesimal generator matrix Ξ of Y is such that $\Xi_{y,y'} \equiv \Xi_{y,\boldsymbol{\eta}} = \delta + \lambda_{\boldsymbol{\eta}} \cdot \left[\Upsilon(y, \eta_1, \eta_3) \vee 0\right]$ with a station *load*

$$\Upsilon(y, i, c) = x_{i,c} \cdot \left(\frac{K_i}{\sum_{c'\in\mathcal{C}} x_{i,c'}} \wedge 1\right) \tag{8}$$

for all jumps $y \xrightarrow{\boldsymbol{\eta}} y'$, $\boldsymbol{\eta} = (i, j, c)$, where the origin station $i > 0$ has PS discipline (here we have set $0/0 = 0$), and

$$\Upsilon(y, i, c) = K_i \wedge x_{i,c} \tag{9}$$

in stations $i > 0$ with an FCFS policy within single-class networks. The station load $\Upsilon(y, i, c)$ differs from the queue length $x_{i,c}$ in that it accounts for the weighting derived from the service

https://doi.org/10.1017/apr.2020.72 Published online by Cambridge University Press

discipline. We further have $\Xi_{y,y'} = \delta + \lambda_{0,j,c}$ for arrivals from virtual stations (in open networks) and $\Xi_{y,y'} = \delta + \lambda_{0,j,c} \cdot (N + \sum_{\eta \in \mathcal{T}^{\leftarrow}_{0,c}} y_{\eta} - \sum_{\eta \in \mathcal{T}^{\rightarrow}_{0,c}} y_{\eta})$ for arrivals from delays, where N denotes the job population in a closed system. Finally, $\Xi_y := \Xi_{y,y} = -\sum_{y' \in \mathcal{S}^Y : y \neq y'} \Xi_{y,y'}$.

4.1. A variational decomposition

The likelihood for observation events in (3) readily transfers to counts Y by means of (7); we thus may write $f_{O|Y_{t_k}}(\boldsymbol{o}_k) \equiv f_{O|X_{t_k}}(\boldsymbol{o}_k)$. Under the measure $\tilde{\mathbb{P}}$, network states can have negative values; the likelihood is undefined in such instances. Now, note that piecewise $\mathcal{S}^Y$-valued trajectories also represent elements $(\boldsymbol{t}, \boldsymbol{y})$ in a space $(\mathcal{Y}, \Sigma_{\mathcal{Y}})$, similar to network paths in X. Let $f_{Y|\boldsymbol{\lambda},\boldsymbol{o}_1,\ldots,\boldsymbol{o}_K}$ be a density function, with respect to some base measure $\mu_{\mathcal{Y}}$, where for all $B \in \Sigma_{\mathcal{Y}}$,

$$\mathbb{P}(Y^{-1}(B)|\boldsymbol{\lambda}, \boldsymbol{o}_1, \ldots, \boldsymbol{o}_K) = \int_B f_{Y|\boldsymbol{\lambda},\boldsymbol{o}_1,\ldots,\boldsymbol{o}_K} \, d\mu_{\mathcal{Y}}.$$

It may be shown by properties of conditional distributions that, conditioned on observations, Y is a non-homogeneous semi-Markov process with *hazard* functions

$$\Lambda_{y,y'}(t) = \Xi_{y,y'} \cdot \frac{\mathbb{P}(\bigcap_{k:t_k>t} O_k^{-1}(\boldsymbol{o_k})|Y_t = y')}{\mathbb{P}(\bigcap_{k:t_k>t} O_k^{-1}(\boldsymbol{o_k})|Y_t = y)} \tag{10}$$

for $y' \neq y$, and $\Lambda_y(t) = -\sum_{y' \neq y} \Lambda_{y,y'}(t)$, so that

$$f_{Y|\boldsymbol{\lambda},\boldsymbol{o}_1,\ldots,\boldsymbol{o}_K}(\boldsymbol{t}, \boldsymbol{y}) = \pi(y_0) \, e^{\int_{t_I}^{T} \Lambda_{y_I}(u)du} \prod_{i=1}^{I} \Lambda_{y_{i-1},y_i}(t_i) \, e^{\int_{t_{i-1}}^{t_i} \Lambda_{y_{i-1}}(u)du}.$$

Here, $\Xi \equiv \Xi(\delta, \boldsymbol{\lambda})$ denotes the generator matrix associated with fixed values in $\boldsymbol{\lambda}$. For a deeper look at conditional jump processes we refer the reader to [25, 9]. This conditional counting process is of key importance; however, the structure of rates in (10) poses a trivial analytical impediment. In our approximating effort, we assume the existence of an alternative measure $\mathbb{Q}$ on $(\Omega, \mathcal{F})$. Under this measure, count trajectories in Y are subject to a *full* decomposition; that is, the $\mathbb{Q}$-law of Y is that of a family of $|\mathcal{T}|$ independent non-homogeneous Poisson counting processes with state-dependent intensity functions $\nu^{\boldsymbol{\eta}} = (\nu_t^{\boldsymbol{\eta}}(\,\cdot\,))_{t \geq 0}$, for all $\boldsymbol{\eta} \in \mathcal{T}$. Here, the following hold:

- Intensity rates for jumps $y \xrightarrow{t,\boldsymbol{\eta}} y'$, $t \geq 0$, are independent of $\boldsymbol{\lambda}$, change over time, and are given by $\nu_t^{\boldsymbol{\eta}}(y_{\boldsymbol{\eta}})$.
- Holding rates in Y evolve according to $|\nu_t(Y_t)|$, with $\nu_t(Y_t) = -\sum_{\boldsymbol{\eta} \in \mathcal{T}} \nu_t^{\boldsymbol{\eta}}(Y_t^{\boldsymbol{\eta}})$.
- The state probability of the multivariate process Y factors across the job transition directions, so that
$$\mathbb{Q}(Y_t = y) = \prod_{\boldsymbol{\eta} \in \mathcal{T}} \mathbb{Q}(Y_t^{\boldsymbol{\eta}} = y_{\boldsymbol{\eta}})$$
for every $y \in \mathcal{S}^Y$.

Similar full-independence decompositions are often referred to as a variational *mean-field* approximations within Bayesian inferential settings [7]. In order to ensure computational

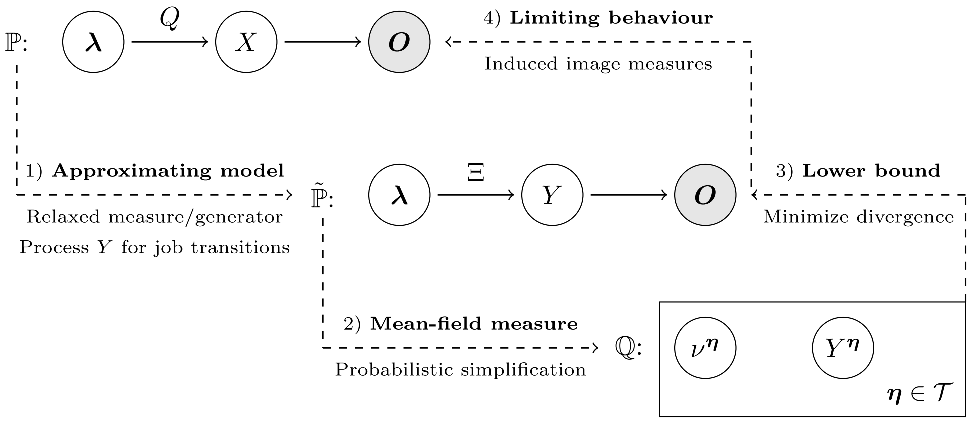

FIGURE 3: Overview of the approximating variational framework presented in this paper, summarizing the various measures used and the primary goals that each step will accomplish.

tractability within forthcoming procedures, the intensity functions ν must be bounded from above by some arbitrary constant, so that $\nu_t^{\boldsymbol{\eta}}(y_{\boldsymbol{\eta}}) \leq \bar{\nu}$ for all $t > 0$, $\boldsymbol{\eta} \in \mathcal{T}$, and $y \in \mathcal{S}^Y$. We finally note that $\mathbb{Q}$ and $\tilde{\mathbb{P}}$ are equivalent on $\mathcal{F}$, as both assign a positive measure to every marginally-increasing sequence of $\mathbb{N}_0^{|\mathcal{T}|}$-valued counts.

4.2. Recapitulation and the variational lower bound

It is important to observe that the intensity functions $\nu^{\boldsymbol{\eta}}$ completely define our alternative measure. In the rest of the paper, we inspect how to fine-tune these functions to ensure that $\mathbb{Q}$ offers a good approximation to the measure $\mathbb{P}$, conditioned on observations in $\boldsymbol{O}$. First, within the diagram in Figure 3 we offer an outline of the various measures, stochastic processes, probabilistic dependencies, and infinitesimal generators introduced up to this point. There, we also summarize the primary steps that help construct a variational framework to enable the inferential task. In short, these are as follows:

- The reference measure $\mathbb{P}$, along with the generator $Q = Q(\boldsymbol{\lambda})$, describes the real *unknown* stochastic network trajectories in X.
- $\tilde{\mathbb{P}}$ is a minor departure from $\mathbb{P}$ and expands the support of X so that queue lengths can become negative. It is associated with the generators $\tilde{Q} = \tilde{Q}(\delta, \boldsymbol{\lambda})$ and $\Xi = \Xi(\delta, \boldsymbol{\lambda})$, for trajectories in X and job transition counts in Y, respectively.
- $\mathbb{Q}$ is an alternative measure under which Y is a set of independent non-homogeneous counting processes. It is associated with intensity functions $\nu^{\boldsymbol{\eta}} = (\nu_t^{\boldsymbol{\eta}}(\,\cdot\,))_{t\geq 0}$ for each count process in the directions $\boldsymbol{\eta} \in \mathcal{T}$.

In our next result, a *variational* lower bound will interrelate the transition rates $\boldsymbol{\lambda}$ for jumps under $\mathbb{P}$ and $\tilde{\mathbb{P}}$ with the intensity functions $\nu^{\boldsymbol{\eta}}$, in light of the available observations/measurements in $\boldsymbol{O}$. From now on, transition rates $\boldsymbol{\lambda}$ are assumed mutually independent under $\mathbb{Q}$ and admit undetermined densities $d\mathbb{Q}_{\boldsymbol{\lambda}} = \{d\mathbb{Q}_{\lambda,\boldsymbol{\eta}},\ \boldsymbol{\eta} \in \mathcal{T}\}$ that must integrate to 1 on $(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+))$.

Lemma 1. (Variational lower bound.) *Define* $\boldsymbol{O} = \bigcap_{k=1}^{K} O_k^{-1}(\boldsymbol{o}_k)$, *and let* $\tilde{\mathbb{P}}$ *and* $\mathbb{Q}$ *be the probability measures on* $(\Omega, \mathcal{F})$ *as defined above. Recall the notation* $\Xi_{y,y'} \equiv \Xi_{y,\boldsymbol{\eta}}$ *for jumps* $y \xrightarrow{\boldsymbol{\eta}} y'$ *with direction* $\boldsymbol{\eta}$. *Then*

$$\log \tilde{\mathbb{P}}(\boldsymbol{O}) \geq \sum_{k=1}^{K} \mathbb{E}^{\mathbb{Q}}_{Y_{t_k}} \big[\log f_{O|Y_{t_k}}(\boldsymbol{o}_k) \big] - \mathbb{E}^{\mathbb{Q}}_{\boldsymbol{\lambda}} \bigg[\log \frac{d\mathbb{Q}_{\boldsymbol{\lambda}}}{d\mathbb{P}_{\boldsymbol{\lambda}}} \bigg]$$
$$- \int_0^T \mathbb{E}^{\mathbb{Q}}_{Y_t,\boldsymbol{\lambda}} \bigg[\sum_{\boldsymbol{\eta} \in \mathcal{T}} \nu_t^{\boldsymbol{\eta}}(Y_t^{\boldsymbol{\eta}}) \log \frac{\nu_t^{\boldsymbol{\eta}}(Y_t^{\boldsymbol{\eta}})}{\Xi_{Y_t,\boldsymbol{\eta}}} - \Xi_{Y_t} + \nu_t(Y_t) \bigg] dt \tag{11}$$

offers a lower bound on the $\tilde{\mathbb{P}}$*-probability of retrieved observation events.*

Proof. Note that

$$\log \tilde{\mathbb{P}}(\boldsymbol{O}) = \log \int_{\mathcal{Y} \times \mathbb{R}_+^{|\mathcal{T}|}} \mathbb{P}(\boldsymbol{O}|Y)\, d(Y, \boldsymbol{\lambda})_* \tilde{\mathbb{P}} = \log \mathbb{E}^{\mathbb{Q}}_{Y,\boldsymbol{\lambda}} \bigg[\mathbb{P}(\boldsymbol{O}|Y) \frac{d(Y, \boldsymbol{\lambda})_* \tilde{\mathbb{P}}}{d(Y, \boldsymbol{\lambda})_* \mathbb{Q}} \bigg]$$
$$\geq \mathbb{E}^{\mathbb{Q}}_{Y} \big[\log \mathbb{P}(\boldsymbol{O}|Y) \big] - \mathbb{E}^{\mathbb{Q}}_{Y,\boldsymbol{\lambda}} \bigg[\log \frac{d(Y, \boldsymbol{\lambda})_* \mathbb{Q}}{d(Y, \boldsymbol{\lambda})_* \tilde{\mathbb{P}}} \bigg], \tag{12}$$

where we use Jensen's inequality for finite measures. This is known as a variational lower bound on the log-likelihood, and

$$\mathbb{E}^{\mathbb{Q}}_{Y} \big[\log \mathbb{P}(\boldsymbol{O}|Y) \big] = \mathbb{E}^{\mathbb{Q}}_{Y} \big[\log \prod_{k=1}^{K} f_{O|Y_{t_k}}(\boldsymbol{o}_k) \big] = \sum_{k=1}^{K} \mathbb{E}^{\mathbb{Q}}_{Y_{t_k}} \big[\log f_{O|Y_{t_k}}(\boldsymbol{o}_k) \big]$$

follows directly from (3). The negative part in (12) is the *Kullback–Leibler* (KL) divergence between image measures of $\mathbb{Q}$ and $\tilde{\mathbb{P}}$. By noting that these share base measures, and Y, $\boldsymbol{\lambda}$ are independent under $\mathbb{Q}$, we have

$$\mathbb{E}^{\mathbb{Q}}_{Y,\boldsymbol{\lambda}} \bigg[\log \frac{d(Y, \boldsymbol{\lambda})_* \mathbb{Q}}{d(Y, \boldsymbol{\lambda})_* \tilde{\mathbb{P}}} \bigg] = \mathbb{E}^{\mathbb{Q}}_{\boldsymbol{\lambda}} \bigg[\log \frac{d\mathbb{Q}_{\boldsymbol{\lambda}}}{d\mathbb{P}_{\boldsymbol{\lambda}}} \bigg] + \mathbb{E}^{\mathbb{Q}}_{\boldsymbol{\lambda}} \bigg[\mathbb{E}^{\mathbb{Q}}_{Y} \bigg[\log \frac{g_Y}{f_{Y|\boldsymbol{\lambda}}} \bigg] \bigg], \tag{13}$$

where g_Y and $f_{Y|\boldsymbol{\lambda}}$ denote the Y-trajectory densities associated with rates $\nu^{\boldsymbol{\eta}}$ and $\Xi(\boldsymbol{\lambda})$, respectively. The last term in (13) is a $\mathbb{Q}$-average of the KL divergence on Y, where the mean is taken across the infinitesimal transition rates. For a fixed starting $Y_0 \in \mathcal{S}^Y$, the inner expectation is shown in [18] to take the equivalent form

$$\mathbb{E}^{\mathbb{Q}}_{Y} \bigg[\log \frac{g_Y}{f_{Y|\boldsymbol{\lambda}}} \bigg] = \lim_{R \to \infty} \sum_{r=0}^{R-1} \mathbb{E}^{\mathbb{Q}}_{Y_{\frac{Tr}{R}}} \Bigg[\sum_{y \in \mathcal{S}^Y} \mathbb{Q}(Y_{\frac{T(r+1)}{R}} = y | Y_{\frac{Tr}{R}}) \log \frac{\mathbb{Q}(Y_{\frac{T(r+1)}{R}} = y | Y_{\frac{Tr}{R}})}{\tilde{\mathbb{P}}(Y_{\frac{T(r+1)}{R}} = y | Y_{\frac{Tr}{R}}, \boldsymbol{\lambda})} \Bigg].$$

Note that within an infinitesimal time interval a jump in Y may only happen in one direction within $\mathcal{T}$. With this in mind, we retrieve the limit of a Riemann sum in the interval $[0, T]$, i.e.

$$\mathbb{E}^{\mathbb{Q}}_{Y} \bigg[\log \frac{g_Y}{f_{Y|\boldsymbol{\lambda}}} \bigg] = \lim_{R \to \infty} \frac{T}{R} \sum_{r=0}^{R-1} \mathbb{E}^{\mathbb{Q}}_{Y_{\frac{Tr}{R}}} \Bigg[\sum_{\boldsymbol{\eta} \in \mathcal{T}} \nu^{\boldsymbol{\eta}}_{\frac{Tr}{R}}(Y^{\boldsymbol{\eta}}_{\frac{Tr}{R}}) \log \frac{\nu^{\boldsymbol{\eta}}_{\frac{Tr}{R}}(Y^{\boldsymbol{\eta}}_{\frac{Tr}{R}})}{\Xi_{Y_{\frac{Tr}{R}},\boldsymbol{\eta}}} + \frac{R}{T} \log \frac{1 + \frac{T}{R} \nu_{\frac{Tr}{R}}(Y_{\frac{Tr}{R}})}{1 + \frac{T}{R} \Xi_{Y_{\frac{Tr}{R}}}} \Bigg]$$
$$= \int_0^T \mathbb{E}^{\mathbb{Q}}_{Y_t} \bigg[\sum_{\boldsymbol{\eta} \in \mathcal{T}} \nu_t^{\boldsymbol{\eta}}(Y_t^{\boldsymbol{\eta}}) \log \frac{\nu_t^{\boldsymbol{\eta}}(Y_t^{\boldsymbol{\eta}})}{\Xi_{Y_t,\boldsymbol{\eta}}} - \Xi_{Y_t} + \nu_t(Y_t) \bigg] dt,$$

https://doi.org/10.1017/apr.2020.72 Published online by Cambridge University Press

and

$$\mathbb{E}^{\mathbb{Q}}_{Y,\boldsymbol{\lambda}}\left[\log \frac{d(Y,\boldsymbol{\lambda})_*\mathbb{Q}}{d(Y,\boldsymbol{\lambda})_*\tilde{\mathbb{P}}}\right] = \mathbb{E}^{\mathbb{Q}}_{\boldsymbol{\lambda}}\left[\log \frac{d\mathbb{Q}_{\boldsymbol{\lambda}}}{d\mathbb{P}_{\boldsymbol{\lambda}}}\right] \\ + \int_0^T \mathbb{E}^{\mathbb{Q}}_{Y_t,\boldsymbol{\lambda}}\left[\sum_{\boldsymbol{\eta}\in\mathcal{T}} \nu_t^{\boldsymbol{\eta}}(Y_t^{\boldsymbol{\eta}}) \log \frac{\nu_t^{\boldsymbol{\eta}}(Y_t^{\boldsymbol{\eta}})}{\tilde{\Xi}_{Y_t,\boldsymbol{\eta}}} - \tilde{\Xi}_{Y_t} + \nu_t(Y_t)\right]dt$$

completes the proof. □

Thus, the lower bound in (11) depends on both the *latent* variables Y and $\boldsymbol{\lambda}$, accounting for the various counts and rates. On a basic level, this is built from three distinct components: the expected log-observations, the KL divergence across service rate densities, and a $\mathbb{Q}$-weighted divergence across hazard functions and rates, further integrated along the entire network trajectory.

5. A functional representation

The above bound includes the prior rates density $d\mathbb{P}_{\boldsymbol{\lambda}}$ along with the $\tilde{\mathbb{P}}$-generator Ξ for the network with negative queues. In addition, we find the unknown $\mathbb{Q}$-distribution for the count random variables Y_t, along with the hazard functions ν. Hence, by maximizing this bound with respect to ν, we may derive properties on $\mathbb{Q}$ that allow for the construction of approximating distributions to $d\mathbb{P}_{X|\boldsymbol{o}_1,\ldots,\boldsymbol{o}_K}$ and $d\mathbb{P}_{\boldsymbol{\lambda}|\boldsymbol{o}_1,\ldots,\boldsymbol{o}_K}$. In this section, we begin by generalizing work in [18] and present results that (i) accommodate parameter uncertainty in transition rates and (ii) impose computational restrictions in the resulting iterative system of equations. Later, we move on to inspect posterior rate densities and conjugacy properties as $\delta \to 0$.

Proposition 1. *Let $d\mathbb{Q}_{\boldsymbol{\lambda}}$ be some valid joint density assigned to the instantaneous rates $\boldsymbol{\lambda}$ under the approximating variational measure $\mathbb{Q}$. Also, define $Y_t^{\backslash\boldsymbol{\eta}} = \{Y_t^{\boldsymbol{\eta}'} : \boldsymbol{\eta}' \in \mathcal{T}\backslash\{\boldsymbol{\eta}\}\}$. Then the $\mathbb{Q}$-dynamics of Y that optimize the lower bound (11) may be parametrized by a system of equations, so that the intensity functions $\nu_t^{\boldsymbol{\eta}}(y) \le \bar{\nu}$ are given by*

$$\nu_t^{\boldsymbol{\eta}}(y) = \frac{r_t^{\boldsymbol{\eta}}(y+1)}{r_t^{\boldsymbol{\eta}}(y)} e^{\mathbb{E}^{\mathbb{Q}}_{Y_t^{\backslash\boldsymbol{\eta}},\boldsymbol{\lambda}}\left[\log \Xi_{Y_t,\boldsymbol{\eta}} | Y_t^{\boldsymbol{\eta}}=y\right] - k_t^{\boldsymbol{\eta}}(y)/\mathbb{Q}(Y_t^{\boldsymbol{\eta}}=y)} \tag{14}$$

for all $t \in [0, T]$, $\boldsymbol{\eta} \in \mathcal{T}$, and $y \in \mathbb{N}_0$, with $\kappa_t^{\boldsymbol{\eta}}(y) \ge 0$ and

$$\frac{dr_t^{\boldsymbol{\eta}}(y)}{dt} = r_t^{\boldsymbol{\eta}}(y)\mathbb{E}^{\mathbb{Q}}_{Y_t^{\backslash\boldsymbol{\eta}},\boldsymbol{\lambda}}\left[\Xi_{Y_t,\boldsymbol{\eta}} | Y_t^{\boldsymbol{\eta}} = y\right] \\ - \left(1 + \frac{k_t^{\boldsymbol{\eta}}(y)}{\mathbb{Q}(Y_t^{\boldsymbol{\eta}} = y)}\right) r_t^{\boldsymbol{\eta}}(y+1) \frac{e^{\mathbb{E}^{\mathbb{Q}}_{Y_t^{\backslash\boldsymbol{\eta}},\boldsymbol{\lambda}}\left[\log \Xi_{Y_t,\boldsymbol{\eta}} | Y_t^{\boldsymbol{\eta}}=y\right]}}{e^{k_t^{\boldsymbol{\eta}}(y)/\mathbb{Q}(Y_t^{\boldsymbol{\eta}}=y)}} \tag{15}$$

whenever $t \ne t_k$, $k = 1, \ldots, K$, and

$$\lim_{t\to t_k^-} r_t^{\boldsymbol{\eta}}(y) = r_{t_k}^{\boldsymbol{\eta}}(y) \exp\left(\mathbb{E}^{\mathbb{Q}}_{Y_{t_k}^{\backslash\boldsymbol{\eta}}}\left[\log f_{O|Y_{t_k}}(\boldsymbol{o}_k) | Y_{t_k}^{\boldsymbol{\eta}} = y\right]\right) \tag{16}$$

at network observation times. In addition, $\kappa_t^{\boldsymbol{\eta}}(y)(\nu_t^{\boldsymbol{\eta}}(y) - \bar{\nu}) = 0$.

https://doi.org/10.1017/apr.2020.72 Published online by Cambridge University Press

Proof. We identify a stationary point to the Lagrangian associated with this constrained optimization problem, where optimization is with respect to the jump rates and the finite-dimensional distributions of Y, subject to $\nu_t^\eta(y) \le \bar{\nu}$ and the master equation

$$\frac{d\mathbb{Q}(Y_t^\eta = y)}{dt} = \nu_t^\eta(y-1) \cdot \mathbb{Q}(Y_t^\eta = y-1) - \nu_t^\eta(y) \cdot \mathbb{Q}(Y_t^\eta = y) \tag{17}$$

for $y \ge 1$, with $d\mathbb{Q}(Y_t^\eta = 0) = -\nu_t^\eta(0)\mathbb{Q}(Y_t^\eta = 0)dt$. Denote by $\phi_t^\eta(y) = \mathbb{Q}(Y_t^\eta = y)$ the functional representing the marginal $\mathbb{Q}$-probability of the state Y_t in the direction of $\boldsymbol{\eta}$, for all $y \in \mathbb{N}_0$. In view of (11), the object function may be expressed as the functional

$$\begin{aligned}\Phi[\phi, \nu, l] = C + \sum_{k=1}^{K} \mathbb{E}^{\mathbb{Q}}_{Y_{t_k}}\big[\log f_{O|Y_{t_k}}(\boldsymbol{o}_k)\big] \\ - \int_0^T \sum_{\eta \in \mathcal{T}} \mathbb{E}^{\mathbb{Q}}_{Y_t^\eta}\big[\Psi[Y_t^\eta, \phi_t^\eta(Y_t^\eta), \nu_t^\eta(Y_t^\eta), l_t^\eta(Y_t^\eta)]\big]dt\end{aligned}$$

with

$$\begin{aligned}\Psi[Y_t^\eta, \phi_t^\eta(Y_t^\eta), \nu_t^\eta(Y_t^\eta), l_t^\eta(Y_t^\eta)] = \nu_t^\eta(Y_t^\eta)\Big(\log \nu_t^\eta(Y_t^\eta) - \mathbb{E}^{\mathbb{Q}}_{Y_t^{\setminus\eta}, \boldsymbol{\lambda}}\big[\log \Xi_{Y_t,\boldsymbol{\eta}}|Y_t^\eta\big] - 1\Big) \\ + \mathbb{E}^{\mathbb{Q}}_{Y_t^{\setminus\eta}, \boldsymbol{\lambda}}\big[\Xi_{Y_t,\boldsymbol{\eta}}|Y_t^\eta\big] - l_t^\eta(Y_t^\eta)\Big(\nu_t^\eta(Y_t^\eta) + \frac{d\log \phi_t^\eta(Y_t^\eta)}{dt}\Big) \\ + l_t^\eta(Y_t^\eta + 1)\nu_t^\eta(Y_t^\eta) - \frac{k_t^\eta(Y_t^\eta)}{\phi_t^\eta(Y_t^\eta)}(\bar{\nu} - \nu_t^\eta(Y_t^\eta)),\end{aligned}$$

where $l^\eta = (l_t^\eta(\,\cdot\,))_{t \ge 0}$ and $k^\eta = (k_t^\eta(\,\cdot\,))_{t \ge 0}$ are multiplier functions that ensure that (17) and the complementary inequality on rates are satisfied. Above, the term C includes the remainder terms in the lower bound in (11) that are independent of the finite-dimensional distributions of Y under $\mathbb{Q}$. Hence, we obtain the functional derivatives

$$\begin{aligned}\frac{\delta\Phi}{\delta\phi_t^\eta(y)} = \sum_{k=1}^{K} \delta(t - t_k)\, \mathbb{E}^{\mathbb{Q}}_{Y_t^{\setminus\eta}}\big[\log f_{O|Y_t}(\boldsymbol{o}_k)|Y_t^\eta = y\big] - \mathbb{E}^{\mathbb{Q}}_{Y_t^{\setminus\eta}, \boldsymbol{\lambda}}\big[\Xi_{Y_t,\boldsymbol{\eta}}|Y_t^\eta = y\big] - \frac{dl_t^\eta(y)}{dt} \\ - \nu_t^\eta(y)\Big(\log \nu_t^\eta(y) - \mathbb{E}^{\mathbb{Q}}_{Y_t^{\setminus\eta}, \boldsymbol{\lambda}}\big[\log \Xi_{Y_t,\boldsymbol{\eta}}|Y_t^\eta = y\big] - 1 - l_t^\eta(y) + l_t^\eta(y+1)\Big)\end{aligned}$$

and

$$\frac{\delta\Phi}{\delta\nu_t^\eta(y)} = -\phi_t^\eta(y)\big(\log \nu_t^\eta(y) - \mathbb{E}^{\mathbb{Q}}_{Y_t^{\setminus\eta}, \boldsymbol{\lambda}}\big[\log \Xi_{Y_t,\boldsymbol{\eta}}|Y_t^\eta = y\big] + l_t^\eta(y+1) - l_t^\eta(y)\big) - k_t^\eta(y)\,,$$

for all $t \in [0, T]$, $\boldsymbol{\eta} \in \mathcal{T}$, and $y \in \mathbb{N}_0$, to be complemented by the slackness conditions $\kappa_t^\eta(y)(\nu_t^\eta(y) - \bar{\nu}) = 0$, $\kappa_t^\eta(y) \ge 0$, and $\nu_t^\eta(y) \le \bar{\nu}$. By letting $l_t^\eta(y) = -\log r_t^\eta(y)$ and setting the above expressions to 0, we obtain

$$\begin{aligned}\frac{dr_t^\eta(y)}{dt} = r_t^\eta(y) \cdot \Big(\mathbb{E}^{\mathbb{Q}}_{Y_t^{\setminus\eta}, \boldsymbol{\lambda}}\big[\Xi_{Y_t,\boldsymbol{\eta}}|Y_t^\eta = y\big] - \sum_{k=1}^{K} \delta(t - t_k)\, \mathbb{E}^{\mathbb{Q}}_{Y_t^{\setminus\eta}}\big[\log f_{O|Y_t}(\boldsymbol{o}_k)|Y_t^\eta = y\big]\Big) \\ - \Big(1 + \frac{k_t^\eta(y)}{\phi_t^\eta(y)}\Big) \cdot r_t^\eta(y+1) \cdot e^{\mathbb{E}^{\mathbb{Q}}_{Y_t^{\setminus\eta}, \boldsymbol{\lambda}}\big[\log \Xi_{Y_t,\boldsymbol{\eta}}|Y_t^\eta = y\big] - k_t^\eta(y)/\phi_t^\eta(y)}\end{aligned}$$

https://doi.org/10.1017/apr.2020.72 Published online by Cambridge University Press

and

$$\nu_t^{\boldsymbol{\eta}}(y) = \frac{r_t^{\boldsymbol{\eta}}(y+1)}{r_t^{\boldsymbol{\eta}}(y)} \cdot e^{\mathbb{E}^{\mathbb{Q}}_{Y_t^{\setminus\boldsymbol{\eta}},\boldsymbol{\lambda}}\left[\log \Xi_{Y_t,\boldsymbol{\eta}} | Y_t^{\boldsymbol{\eta}}=y\right] - k_t^{\boldsymbol{\eta}}(y)/\phi_t^{\boldsymbol{\eta}}(y)}.$$

Observe above that, for fixed values of ϕ and r, if

$$\frac{r_t^{\boldsymbol{\eta}}(y+1)}{r_t^{\boldsymbol{\eta}}(y)} \cdot e^{\mathbb{E}^{\mathbb{Q}}_{Y_t^{\setminus\boldsymbol{\eta}},\boldsymbol{\lambda}}\left[\log \Xi_{Y_t,\boldsymbol{\eta}} | Y_t^{\boldsymbol{\eta}}=y\right]} < \bar{\nu},$$

then the complementary slackness conditions imply $k_t^{\boldsymbol{\eta}}(y) = 0$; otherwise, $\nu_t^{\boldsymbol{\eta}}(y) = \bar{\nu}$ and

$$k_t^{\boldsymbol{\eta}}(y) = \phi_t^{\boldsymbol{\eta}}(y) \cdot \left[\mathbb{E}^{\mathbb{Q}}_{Y_t^{\setminus\boldsymbol{\eta}},\boldsymbol{\lambda}}\left[\log \Xi_{Y_t,\boldsymbol{\eta}} | Y_t^{\boldsymbol{\eta}} = y\right] - \log \frac{\bar{\nu} \cdot r_t^{\boldsymbol{\eta}}(y)}{r_t^{\boldsymbol{\eta}}(y+1)}\right] \geq 0,$$

yielding a valid system of equations, which leads to (14)–(16) and concludes the proof. □

Corollary 1. (Distributed network monitoring.) *Assume that network observations are distributed and independent across the stations, so that*

$$f_{O|Y_{t_k}}(\boldsymbol{o}_k) = \prod_{i=1}^{M} f_{O|\{Y_{t_k}^{\boldsymbol{\eta}}\,:\,\boldsymbol{\eta}\in\mathcal{T}_i\}}(\boldsymbol{o}_k^i)$$

for some conditional mass function

$$f_{O|\{Y_{t_k}^{\boldsymbol{\eta}}\,:\,\boldsymbol{\eta}\in\mathcal{T}_i\}},$$

where $\mathcal{T}_i = (\cup_{c\in\mathcal{C}}\, \mathcal{T}_{i,c}^{\leftarrow}) \cup (\cup_{c\in\mathcal{C}}\, \mathcal{T}_{i,c}^{\rightarrow})$ *is the set of job transitions relevant to network activity in station* $i > 0$, *and* $\boldsymbol{o}_k^i$ *denotes the time-*t_k *observations across classes in the station. Further assume that* $\bar{\nu} = \infty$, *so that there exists no bound on intensity rates* $\nu_t^{\boldsymbol{\eta}}(y)$ *under* $\mathbb{Q}$. *Then the system of equations in Proposition 1 reduces to*

$$\nu_t^{\boldsymbol{\eta}}(y) = \frac{r_t^{\boldsymbol{\eta}}(y+1)}{r_t^{\boldsymbol{\eta}}(y)} e^{\mathbb{E}^{\mathbb{Q}}_{Y_t^{\setminus\boldsymbol{\eta}},\boldsymbol{\lambda}}\left[\log \Xi_{Y_t,\boldsymbol{\eta}} | Y_t^{\boldsymbol{\eta}}=y\right]}$$

for all $t \in [0, T]$, $\boldsymbol{\eta} \in \mathcal{T}$, *and* $y \in \mathbb{N}_0$, *with*

$$\frac{dr_t^{\boldsymbol{\eta}}(y)}{dt} = r_t^{\boldsymbol{\eta}}(y)\mathbb{E}^{\mathbb{Q}}_{Y_t^{\setminus\boldsymbol{\eta}},\boldsymbol{\lambda}}\left[\Xi_{Y_t,\boldsymbol{\eta}} | Y_t^{\boldsymbol{\eta}} = y\right] - r_t^{\boldsymbol{\eta}}(y+1) e^{\mathbb{E}^{\mathbb{Q}}_{Y_t^{\setminus\boldsymbol{\eta}},\boldsymbol{\lambda}}\left[\log \Xi_{Y_t,\boldsymbol{\eta}} | Y_t^{\boldsymbol{\eta}}=y\right]}$$

whenever $t \neq t_k$, $k = 1, \ldots, K$, *and*

$$\lim_{t\to t_k^-} r_t^{\boldsymbol{\eta}}(y) = r_{t_k}^{\boldsymbol{\eta}}(y) e^{\mathbb{E}^{\mathbb{Q}}_{Y_t^{\setminus\boldsymbol{\eta}}}\left[\log f_{O|\{Y_t^{\boldsymbol{\eta}'}\,:\,\boldsymbol{\eta}'\in\mathcal{T}_{\eta_1}\}}(\boldsymbol{o}_k^{\eta_1}) | Y_t^{\boldsymbol{\eta}}=y\right]} \cdot e^{\mathbb{E}^{\mathbb{Q}}_{Y_t^{\setminus\boldsymbol{\eta}}}\left[\log f_{O|\{Y_t^{\boldsymbol{\eta}'}\,:\,\boldsymbol{\eta}'\in\mathcal{T}_{\eta_2}\}}(\boldsymbol{o}_k^{\eta_2}) | Y_t^{\boldsymbol{\eta}}=y\right]}$$

accounting for observations at origin and departure stations in $\boldsymbol{\eta} \in \mathcal{T}$.

In Proposition 1, we derived a system of equations with iterated dependencies, given a distribution $\mathbb{Q}_{\boldsymbol{\lambda}}$ for the service rates. These equations can reconstruct the hazard rates $\nu =$

https://doi.org/10.1017/apr.2020.72 Published online by Cambridge University Press

$\nu^{\boldsymbol{\eta}} = (\nu_t^{\boldsymbol{\eta}}(\,\cdot\,))_{t\geq 0}$ (for each counting process) that best approximate the trajectories of Y under $\tilde{\mathbb{P}}$, given the observations $\boldsymbol{O}$. Probabilities over the states of Y may ultimately be derived by means of the master equation (17). In Corollary 1, we further notice that by simplifying the network observation model and easing restrictions on rates under the approximating measure $\mathbb{Q}$, we retrieve a result analogous to that previously presented in [18, 7]. However, this is reportedly problematic and can cause a computational bottleneck when reconstructing the jump rates $\nu_t^{\boldsymbol{\eta}}(y)$, as these may approach infinity at observation times. When presenting our experimental results at the end of the paper, we will rely on the formulae within Proposition 1.

Next, we wish to determine the optimal form for $\mathbb{Q}_{\boldsymbol{\lambda}}$, given a family of hazard rates ν, such that we best approximate the conditional densities $d\mathbb{P}_{\boldsymbol{\lambda}|\boldsymbol{o}_1,\ldots,\boldsymbol{o}_K}$.

Proposition 2. *Let densities for the infinitesimal rates* $\boldsymbol{\lambda}$ *be defined with respect to the (Lebesgue) product base measure* $\mu_{\boldsymbol{\lambda}}$, *so that* $d\mathbb{Q}_{\boldsymbol{\lambda}} = g_{\boldsymbol{\lambda}} d\mu_{\boldsymbol{\lambda}}$ *with* $g_{\boldsymbol{\lambda}} = \prod_{\boldsymbol{\eta}\in\mathcal{T}} g_\lambda^{\boldsymbol{\eta}}$ *and marginal densities* $g_\lambda^{\boldsymbol{\eta}} = d\mathbb{Q}_{\lambda,\boldsymbol{\eta}}/d\mu_\lambda$. *Also, let* $\nu_t^{\boldsymbol{\eta}}(y)$, $\boldsymbol{\eta}\in\mathcal{T}$, *be some (independent) intensity functions assigned to* Y *under the approximating variational measure* $\mathbb{Q}$. *Finally, define* $\boldsymbol{\lambda}_{\backslash\boldsymbol{\eta}} = \{\lambda_{\boldsymbol{\eta}'} : \boldsymbol{\eta}' \in \mathcal{T}\backslash\{\boldsymbol{\eta}\}\}$ *and recall the definitions for network station loads* Υ *in (8) and (9). Then, as* $\delta \to 0$ *in (6), the distribution* $\mathbb{Q}_{\boldsymbol{\lambda}}$ *that optimizes the lower bound (11) is such that*

$$g_\lambda^{\boldsymbol{\eta}}(z) \propto e^{\mathbb{E}^{\mathbb{Q}}_{\boldsymbol{\lambda}\backslash\boldsymbol{\eta}}[\,\log f_{\boldsymbol{\lambda}}(z)] - z\cdot\int_0^T \mathbb{E}^{\mathbb{Q}}_{Y_t}[\Upsilon(Y_t,\eta_1,\eta_3)\vee 0]dt} \cdot z^{\int_0^T \mathbb{E}^{\mathbb{Q}}_{Y_t^{\boldsymbol{\eta}}}[\nu_t^{\boldsymbol{\eta}}(Y_t^{\boldsymbol{\eta}})]dt}$$

up to a normalizing constant, for $z\in\mathbb{R}_+$ *and every* $\boldsymbol{\eta}\in\mathcal{T}$.

Proof. We again identify a stationary point to the Lagrangian associated with a constrained optimization problem, with respect to arbitrary (positive) densities $g_\lambda^{\boldsymbol{\eta}}$ with $\int_{\mathbb{R}_+} g_\lambda^{\boldsymbol{\eta}} d\mu_\lambda = 1$. Since $\mathbb{P}_{\boldsymbol{\lambda}}$ and $\mathbb{Q}_{\boldsymbol{\lambda}}$ share base measures, the object function can be written as

$$\Phi[g] = C - \sum_{\boldsymbol{\eta}\in\mathcal{T}} \mathbb{E}^{\mathbb{Q}}_{\lambda_{\boldsymbol{\eta}}}[\Psi[\lambda_{\boldsymbol{\eta}}, g_\lambda^{\boldsymbol{\eta}}]] - \sum_{\boldsymbol{\eta}\in\mathcal{T}} l^{\boldsymbol{\eta}}\left[\int_{\mathbb{R}_+} g_\lambda^{\boldsymbol{\eta}} d\mu_\lambda - 1\right],$$

where $\{l^{\boldsymbol{\eta}}\}_{\boldsymbol{\eta}\in\mathcal{T}}$ are non-functional Lagrange multipliers, and

$$\Psi[\lambda_{\boldsymbol{\eta}}, g_\lambda^{\boldsymbol{\eta}}] = \log g_\lambda^{\boldsymbol{\eta}} - \frac{1}{|\mathcal{T}|}\mathbb{E}^{\mathbb{Q}}_{\boldsymbol{\lambda}\backslash\boldsymbol{\eta}}[\,\log f_{\boldsymbol{\lambda}}] + \int_0^T \mathbb{E}^{\mathbb{Q}}_{Y_t}\left[\nu_t^{\boldsymbol{\eta}}(Y_t^{\boldsymbol{\eta}})\log\frac{\nu_t^{\boldsymbol{\eta}}(Y_t^{\boldsymbol{\eta}})}{\Xi_{Y_t,\boldsymbol{\eta}}} + \Xi_{Y_t,\boldsymbol{\eta}} - \nu_t^{\boldsymbol{\eta}}(Y_t)\right]dt.$$

The term C includes the remainder terms in the lower bound in (11) that are independent of the rates $\boldsymbol{\lambda}$. It follows that

$$\begin{aligned}\frac{\delta\Phi}{\delta g_\lambda^{\boldsymbol{\eta}}} &= \mathbb{E}^{\mathbb{Q}}_{\boldsymbol{\lambda}\backslash\boldsymbol{\eta}}[\,\log f_{\boldsymbol{\lambda}}] - \log\,(g_\lambda^{\boldsymbol{\eta}}) - 1 - l^{\boldsymbol{\eta}}\\ &\quad - \int_0^T \mathbb{E}^{\mathbb{Q}}_{Y_t}\left[\nu_t^{\boldsymbol{\eta}}(Y_t^{\boldsymbol{\eta}})\log\frac{\nu_t^{\boldsymbol{\eta}}(Y_t^{\boldsymbol{\eta}})}{\Xi_{Y_t,\boldsymbol{\eta}}} + \Xi_{Y_t,\boldsymbol{\eta}} - \nu_t^{\boldsymbol{\eta}}(Y_t)\right]dt\end{aligned}$$

in its support set $\mathbb{R}_+$, for all $\boldsymbol{\eta}\in\mathcal{T}$. By equating the above to 0, considering constraints, and analysing the relevant terms up to proportionality, we find that

$$g_\lambda^{\boldsymbol{\eta}} \propto \exp\left(\mathbb{E}^{\mathbb{Q}}_{\boldsymbol{\lambda}\backslash\boldsymbol{\eta}}[\,\log f_{\boldsymbol{\lambda}}] + \int_0^T \mathbb{E}^{\mathbb{Q}}_{Y_t}\left[\nu_t^{\boldsymbol{\eta}}(Y_t^{\boldsymbol{\eta}})\log\Xi_{Y_t,\boldsymbol{\eta}} - \Xi_{Y_t,\boldsymbol{\eta}}\right]dt\right),$$

so that

$$g_\lambda^{\boldsymbol{\eta}}(z) \propto e^{\mathbb{E}^{\mathbb{Q}}_{\boldsymbol{\lambda}\backslash\boldsymbol{\eta}}[\,\log f_{\boldsymbol{\lambda}}(z)] - \int_0^T z\cdot\mathbb{E}^{\mathbb{Q}}_{Y_t}[\Upsilon(Y_t,\eta_1,\eta_3)\vee 0]dt + \int_0^T \mathbb{E}^{\mathbb{Q}}_{Y_t^{\boldsymbol{\eta}}}[\nu_t^{\boldsymbol{\eta}}(Y_t^{\boldsymbol{\eta}})\log\,(\delta + z\cdot\Upsilon(Y_t,\eta_1,\eta_3))]dt}$$

https://doi.org/10.1017/apr.2020.72 Published online by Cambridge University Press

and

$$g_{\lambda}^{\boldsymbol{\eta}}(z) \propto e^{\mathbb{E}^{\mathbb{Q}}_{\boldsymbol{\lambda}\setminus\boldsymbol{\eta}}[\log f_{\boldsymbol{\lambda}}(z)] - z\cdot\int_0^T \mathbb{E}^{\mathbb{Q}}_{Y_t}[\Upsilon(Y_t,\eta_1,\eta_3)\vee 0]dt} \cdot z^{\int_0^T \mathbb{E}^{\mathbb{Q}}_{Y_t^{\boldsymbol{\eta}}}[\nu_t^{\boldsymbol{\eta}}(Y_t^{\boldsymbol{\eta}})]dt}$$

as $\delta \to 0$, for $z \in \mathbb{R}_+$ and every $\boldsymbol{\eta} \in \mathcal{T}$. □

Corollary 2. (Conjugate prior.) *Assume that the prior density on* $\boldsymbol{\lambda}$ *also factors across the individual rates, so that* $d\mathbb{P}_{\boldsymbol{\lambda}} = \prod_{\boldsymbol{\eta}\in\mathcal{T}} f_{\lambda}^{\boldsymbol{\eta}} d\mu_{\boldsymbol{\lambda}}$, *where* $f_{\lambda}^{\boldsymbol{\eta}}$ *for* $\boldsymbol{\eta} \in \mathcal{T}$ *denote gamma density functions with shape* $\alpha_{\boldsymbol{\eta}}$ *and rate* $\beta_{\boldsymbol{\eta}}$. *Then, as* $\delta \to 0$ *in (6), these are conjugate priors, and*

$$\lambda_{\boldsymbol{\eta}} \overset{\mathbb{Q}}{\sim} \Gamma\Big(\alpha_{\boldsymbol{\eta}} + \int_0^T \mathbb{E}^{\mathbb{Q}}_{Y_t^{\boldsymbol{\eta}}}[\nu_t^{\boldsymbol{\eta}}(Y_t^{\boldsymbol{\eta}})]dt,\, \beta_{\boldsymbol{\eta}} + \int_0^T \mathbb{E}^{\mathbb{Q}}_{Y_t}[\Upsilon(Y_t,\, \eta_1,\, \eta_3) \vee 0]dt\Big)$$

for all $\boldsymbol{\eta} \in \mathcal{T}$.

Hence, as the network model with negative queues under $\tilde{\mathbb{P}}$ offers a better approximation of its original counterpart ($\delta \to 0$), we may numerically approximate posterior distributions across the infinitesimal rates in $\boldsymbol{\lambda}$, under the variational measure $\mathbb{Q}$. In the special case with independent gamma prior densities, this is an easily interpretable posterior where the shape and rate parameters depend, respectively, on the integrated expected jump intensities and the integrated expected station loads in (8) and (9).

6. Applications

The results in this paper suggest an iterative approximation procedure for (5) by means of coordinate ascent. Here, we iteratively update the values of the various rates, functions, and Lagrange multipliers while evaluating and assessing convergence in the lower bound (11) to the log-likelihood. This is a standard approach in variational inference when looking for (local) maxima [4], and the problem is known to be convex.

Maximizing the bound by calibrating the measure $\mathbb{Q}$ will yield an approximation to the regular conditional probability of events in $\mathcal{F}$ under $\tilde{\mathbb{P}}$, conditioned on the observations. This may then be projected over the rates $\boldsymbol{\lambda}$ or the trajectories in X and Y. These projected densities are valid for approximating the conditional distributions of the service rates $\boldsymbol{\mu}$ and routing probabilities $\mathcal{P}$ in the original QN system, given observations. The final iterative procedure is described below.

1. Input network observations in (3) and assign a (conjugate) gamma (image) density $d\mathbb{P}_{\boldsymbol{\lambda}}$ across job transition intensities $\boldsymbol{\lambda}$, with shape parameters $\alpha_{\boldsymbol{\eta}}$ and a (shared) rate parameter $\beta_{\boldsymbol{\eta}} = \beta$, $\boldsymbol{\eta} \in \mathcal{T}$.
2. Define a discretization grid of the time interval $[0, T]$, and operate through interpolation within points in the grid.
3. Set an arbitrary density $d\mathbb{Q}_{\boldsymbol{\lambda}}$. Fix $\kappa_t^{\boldsymbol{\eta}}(y) = 0$, $r_t^{\boldsymbol{\eta}}(y) = 1$, and input (valid) arbitrary starting values $\mathbb{Q}(Y_t^{\boldsymbol{\eta}} = y)$, for all $t \in [0, T]$, $\boldsymbol{\eta} \in \mathcal{T}$, and $y \in \mathbb{N}_0$.
4. Iterate until convergence:

   4.1. In each direction $\boldsymbol{\eta} \in \mathcal{T}$, numerically compute $r_t^{\boldsymbol{\eta}}(y)$ for every $t \in [0, T]$ and $y \in \mathbb{N}_0$, by means of (15)–(16). Then update intensity and slack functions $\nu_t^{\boldsymbol{\eta}}(y)$, $\kappa_t^{\boldsymbol{\eta}}(y)$ with

https://doi.org/10.1017/apr.2020.72 Published online by Cambridge University Press

(14), so that $k_t^\eta(y) = 0$ if

$$\frac{r_t^\eta(y+1)}{r_t^\eta(y)} \cdot e^{\mathbb{E}^{\mathbb{Q}}_{Y_t^{\setminus\eta},\lambda}\left[\log \Xi_{Y_t,\eta} | Y_t^\eta = y\right]} < \bar{\nu},$$

and

$$\nu_t^\eta(y) = \bar{\nu}, \qquad k_t^\eta(y) = \mathbb{Q}(Y_t^\eta = y) \cdot \left[\mathbb{E}^{\mathbb{Q}}_{Y_t^{\setminus\eta},\lambda}\left[\log \Xi_{Y_t,\eta} | Y_t^\eta = y\right] - \log \frac{\bar{\nu} \cdot r_t^\eta(y)}{r_t^\eta(y+1)}\right]$$

otherwise. Renew transient state probabilities in Y by means of the master equation (17), for all $t \in [0, T]$.

4.2. Derive expected intensities

$$\mathbb{E}^{\mathbb{Q}}_{Y_t^\eta}[\nu_t^\eta(Y_t^\eta)]$$

and station loads

$$\mathbb{E}^{\mathbb{Q}}_{Y_t}[\Upsilon(Y_t, \eta_1, \eta_3) \vee 0],$$

for all directions $\boldsymbol{\eta} \in \mathcal{T}$ and times $t \in [0, T]$. Update $\mathbb{Q}$-densities so that

$$\lambda_\eta \overset{\mathbb{Q}}{\sim} \Gamma\left(\alpha_\eta + \int_0^T \mathbb{E}^{\mathbb{Q}}_{Y_t^\eta}[\nu_t^\eta(Y_t^\eta)]dt,\ \beta + \int_0^T \mathbb{E}^{\mathbb{Q}}_{Y_t}[\Upsilon(Y_t, \eta_1, \eta_3) \vee 0]dt\right)$$

for all $\boldsymbol{\eta} \in \mathcal{T}$. The likelihood ratio in (5) can be computed at this stage.

4.3. Evaluate the lower bound (11), given the current densities and infinitesimal rates under the approximating measure $\mathbb{Q}$. Assess variation in the bound across iterations and establish convergence.

5. Finally, infer the structure of the various service rates and routing probabilities in the QN system.

5.1. Note that

$$\mathbb{E}^{\mathbb{Q}}_{Y_t}[\Upsilon(Y_t, \eta_1, \eta_3) \vee 0]$$

remains the same across directions $\boldsymbol{\eta} \in \mathcal{T}$ with shared origin station. Since

$$\mu_i^c = \sum_{\eta \in \mathcal{T}_{i,c}^{\rightarrow}} \lambda_\eta,$$

it holds that

$$\mu_i^c \overset{\mathbb{Q}}{\sim} \Gamma\left(|\mathcal{T}_{i,c}^{\rightarrow}| \cdot \alpha_\eta + \sum_{\eta \in \mathcal{T}_{i,c}^{\rightarrow}} \int_0^T \mathbb{E}^{\mathbb{Q}}_{Y_t^\eta}[\nu_t^\eta(Y_t^\eta)]dt,\ \beta \right.$$
$$\left. + \int_0^T \mathbb{E}^{\mathbb{Q}}_{Y_t}[\Upsilon(Y_t, \eta_1, \eta_3) \vee 0]dt\right)$$

for all $0 \leq i \leq M$, $c \in \mathcal{C}$.

5.2. Retrieve distributions for routing probabilities by noting that $p_{i,j}^c = \lambda_{i,j,c}/\mu_i^c$ for all $0 \leq i, j \leq M$, $c \in \mathcal{C}$. This suggests a Dirichlet distribution.

The following examples treat open and closed network models; source code can be found at https://github.com/IkerPerez/variationalQueues.

https://doi.org/10.1017/apr.2020.72 Published online by Cambridge University Press

6.1. Single-class closed network

We begin by analysing the small closed network example previously shown in Figure 2. We recall that this includes one FCFS service station, with $K_1 = 1$ processing unit, along with a delay, together processing a population of N jobs cyclically in a closed loop. All jobs belong to the same class and have equal service rates; we denote by μ_1 the job processing rate within the service station. On completion, a job proceeds to the delay station where it awaits for an exponentially distributed time before being routed back to the queue. We use μ_0 to denote the delay rate, and we note that the arrival rate to the queue is directly proportional to the number of jobs at the delay.

Both stations are independent and μ_0 is fixed in order to ensure model identifiability within the service station. In this instance, the network topology is deterministic and trivial, and the evolution of $X = (X_t)_{t \geq 0}$ monitors the total number of jobs within the service station, with $X_0 = 0$. The generator Q of X is finite and satisfies

$$Q = \left[\begin{array}{cccccccc} & 0 & 1 & 2 & \dots & N-2 & N-1 & N \\ 0 & -N\mu_0 & N\mu_0 & 0 & \dots & 0 & 0 & 0 \\ 1 & \mu_1 & -N\mu_0 + \mu_0 - \mu_1 & (N-1)\mu_0 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ N-1 & 0 & 0 & 0 & \dots & \mu_1 & -(\mu_0 + \mu_1) & \mu_0 \\ N & 0 & 0 & 0 & \dots & 0 & \mu_1 & -\mu_1 \end{array}\right],$$

where row and column labels denote the number of jobs in the queueing station. Since μ_0 is fixed, our interest lies in $\lambda \equiv \lambda_{1,0} = \mu_1 \cdot p_{1,0} = \mu_1$. We assign to this rate a distribution $\mathbb{P}_\lambda \equiv \lambda_\star \mathbb{P}$ with (gamma) density f_λ such that its hyperparameters fix some reasonably uninformative *prior* knowledge on the system. We monitor the delay and FCFS service station at fixed and equally spaced times $t_1 < \cdots < t_K$ in an interval $[0, T]$. Here, variables O_k are supported on $\mathcal{O} = \{0, \dots, N\}^2$, and the observation model factors across the network components so that $f_{O|x}(\boldsymbol{o}) = \tilde{f}_{O|N-x}(o_0) \cdot \tilde{f}_{O|x}(o_1)$ with $\tilde{f}_{O|x}(o) = \frac{\epsilon}{N} + \mathbb{I}(o = x) \cdot (1 - \frac{N+1}{N}\epsilon)$ and

$$\mathbb{P}(O_k^{-1}(\boldsymbol{o})|X) = \begin{cases} (1-\epsilon)^2, & o_0 = N - X_{t_k},\ o_1 = X_{t_k}, \\ (\epsilon/N)^2, & o_0 \neq N - X_{t_k},\ o_1 \neq X_{t_k}, \\ (1-\epsilon) \cdot \epsilon/N & \text{otherwise} \end{cases} \tag{18}$$

for $\epsilon > 0$ and all $k = 1, \dots, K$. This accounts for some $\epsilon \cdot 100\%$ faulty measurements; also, we note that a system with discrete observations is approximated as $\epsilon \to 0$. Now, assume there exist some sample observations $\boldsymbol{o}_1, \dots, \boldsymbol{o}_K$ from a model realization in this closed network. These can easily be produced from (18) given a trajectory $(X_t)_{t \in [0,T]}$. In order to produce the trajectory from (2), given the service rates, we may employ Gillespie's algorithm [10] or faster uniformization-based alternatives [24].

Remark 1. The transformation $d\mathbb{P}_{\lambda|\boldsymbol{o}_1,\dots,\boldsymbol{o}_K}/d\mathbb{P}_\lambda$ is such that, conditioned on $\boldsymbol{o}_1, \dots, \boldsymbol{o}_K$, the distribution over λ admits a density carried by the Lebesgue measure μ_λ, so that $d\mathbb{P}_{\lambda|\boldsymbol{o}_1,\dots,\boldsymbol{o}_K} = f_{\lambda|\boldsymbol{o}_1,\dots,\boldsymbol{o}_K}\, d\mu_\lambda$. In this simple example, numerical MCMC procedures [22] or basic *generator-matrix exponentiations* combined with a forward–backward algorithm [32] can offer such density approximations; however, this is reportedly very inefficient when N is large. Moreover,

https://doi.org/10.1017/apr.2020.72 Published online by Cambridge University Press

in involved networks/processes for complex applications (see next example), such alternatives are simply unusable (i.e. they do not scale).

In the following, we analyse simulated data ($N=50,\ \mu_0=0.1,\ \epsilon=0.2,\ T=100,\ K=50$) by assigning a conjugate gamma density to $\lambda\equiv\mu_1$, so that $\lambda\sim\Gamma(\alpha,\beta)$ with $\alpha=5$ and $\beta=2$ under the reference measure $\mathbb{P}$. Recall that μ_0 is fixed to ensure model identifiability, and $X_t\in\{0,\dots,N\}$ denotes the number of jobs in the service station at any time $t\geq 0$. For later reference, the stationary distribution of the system is given by

$$\pi_{\mathbb{P}}(x|\lambda)=\lim_{t\to\infty}\mathbb{P}(X_t=x|\lambda)=\Big(\frac{\mu_0}{\lambda}\Big)^x\frac{1}{(N-x)!}\Big/\sum_{x=0}^{N}\Big(\frac{\mu_0}{\lambda}\Big)^x\frac{1}{(N-x)!},$$

so that, assuming the observations are sufficiently spaced and the system has reached stationarity, it holds that

$$\mathbb{P}\Big(\bigcap_{k=1}^{K}O_k^{-1}(\boldsymbol{o}_k)\big|\lambda\Big)\approx\prod_{k=1}^{K}\sum_{x=0}^{N}f_{O|x}(\boldsymbol{o}_k)\pi_{\mathbb{P}}(x|\lambda)=\frac{\prod_{k=1}^{K}\sum_{x=0}^{N}f_{O|x}(\boldsymbol{o}_k)\left(\frac{\mu_0}{\lambda}\right)^x\frac{1}{(N-x)!}}{\left(\sum_{x=0}^{N}\left(\frac{\mu_0}{\lambda}\right)^x\frac{1}{(N-x)!}\right)^K},\qquad(19)$$

where $f_{O|x}(\boldsymbol{o}_k)$ is as defined in (18). Note that here $|\mathcal{T}|=2$, and the process $Y=(Y_t^{0,1},Y_t^{1,0})_{t\geq 0}$ monitors transitions between the delay and service station, in both the directions $0\to 1$ and $1\to 0$. The lower bound to the log-likelihood in (11) reduces to

$$\log\tilde{\mathbb{P}}(\boldsymbol{O})\geq\sum_{k=1}^{K}\mathbb{E}^{\mathbb{Q}}_{Y_{t_k}}\big[\log\mathbb{P}(O_k^{-1}(\boldsymbol{o}_k)|Y_{t_k}^{0,1}-Y_{t_k}^{1,0})\big]$$
$$-\mathbb{E}^{\mathbb{Q}}_{\lambda}\bigg[\log\frac{g_\lambda^{1,0}}{d\mathbb{P}_\lambda}\bigg]+\int_0^T\mathbb{E}^{\mathbb{Q}}_{Y_t,\lambda}\big[\Psi[Y_t,\nu_t,\lambda]\big]dt\qquad(20)$$

with

$$\Psi[Y_t,\nu_t,\lambda]=\nu_t^{1,0}(Y_t^{1,0})+\nu_t^{0,1}(Y_t^{0,1})-2\delta-\lambda\cdot\mathbb{I}(Y_{t_k}^{0,1}-Y_{t_k}^{1,0}>0)$$
$$-\mu_0\cdot(N+Y_{t_k}^{1,0}-Y_{t_k}^{0,1})-\nu_t^{1,0}(Y_t^{1,0})\log\frac{\nu_t^{1,0}(Y_t^{1,0})}{\delta+\lambda\cdot\mathbb{I}(Y_{t_k}^{0,1}-Y_{t_k}^{1,0}>0)}$$
$$-\nu_t^{0,1}(Y_t^{0,1})\log\frac{\nu_t^{0,1}(Y_t^{0,1})}{\delta+\mu_0\cdot(N+Y_{t_k}^{1,0}-Y_{t_k}^{0,1})},$$

so that it contains only two hazard functions in the approximating measure $\mathbb{Q}$, namely $\nu^{0,1}$ and $\nu^{1,0}$. In (20), we again notice that the lower bound is dominated by three distinguishable components: (i) the expected log-observations, (ii) the KL divergence across the service rate density, and (iii) a weighted $\mathbb{P}$-to-$\mathbb{Q}$ divergence in the expected path likelihood, further integrated along the entire network trajectory. The differential equations for functionals in (16) reduce to

$$\frac{dr_t^{0,1}(y)}{dt}=r_t^{0,1}(y)\big(\delta+\mu_0\cdot\mathbb{E}^{\mathbb{Q}}_{Y_t^{1,0}}\big[(Y_t^{1,0}-y)\vee 0)\big]\big)$$
$$-\frac{1+k_t^{0,1}(y)/\mathbb{Q}(Y_t^{0,1}=y)}{e^{k_t^{0,1}(y)/\mathbb{Q}(Y_t^{0,1}=y)}}r_t^{0,1}(y+1)e^{\mathbb{E}^{\mathbb{Q}}_{Y_t^{1,0}}\left[\log(\delta+\mu_0\cdot[(Y_t^{1,0}-y)\vee 0])\right]}$$

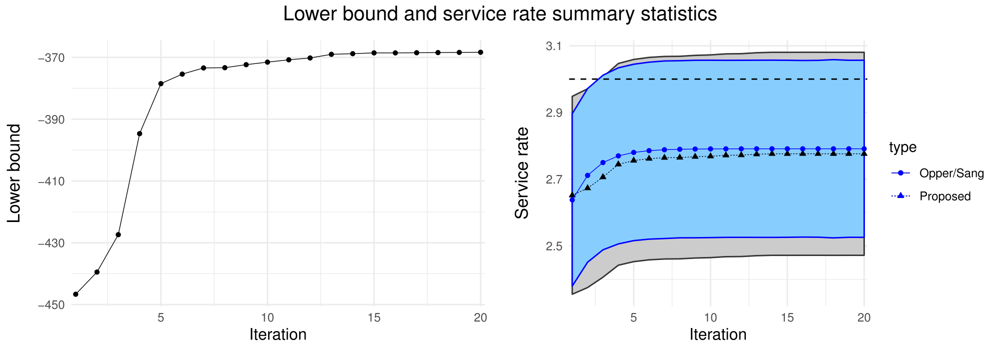

FIGURE 4: Left, evolution of lower bound to the log-likelihood during the inferential procedure. Right, 95% credible intervals and point estimates for the service rate λ under $\mathbb{Q}$; the back (grey) corresponds to the proposed method, while the front (blue) is for an (adapted) traditional variational procedure.

and

$$\begin{aligned}\frac{dr_t^{1,0}(y)}{dt} &= r_t^{1,0}(y)\big(\delta + \mathbb{E}_\lambda^{\mathbb{Q}}[\lambda]\cdot \mathbb{Q}(Y_t^{0,1} > y)\big)\\ &\quad - \frac{1+k_t^{1,0}(y)/\mathbb{Q}(Y_t^{1,0}=y)}{e^{k_t^{1,0}(y)/\mathbb{Q}(Y_t^{1,0}=y)}} r_t^{1,0}(y+1)e^{\mathbb{E}^{\mathbb{Q}}_{Y_t^{0,1},\lambda}\left[\log(\delta+\lambda\cdot\mathbb{I}(Y_t^{0,1}>y))\right]}.\end{aligned}$$

In Figure 4 (left) we observe the evolution of the lower bound (20) during the iterative inferential procedure, for a sufficiently small and negligible value of δ. There, we notice that the procedure has converged to a (local) optimum within approximately 13 iterations. On the right-hand side of the figure, we further observe credible intervals and point estimates across iterations for λ, under the approximating measure $\mathbb{Q}$. Displayed in the back (in grey) are the estimates obtained using the proposed method in this paper; in the front (in blue), we find approximations extracted by adapting variational procedures in [18], to allow for prior knowledge and conjugacy properties.

Next, in Figure 5 (left) we find the $\mathbb{P}$-prior density for λ (flat density in the back of the figure) with the final-iteration $\mathbb{Q}$-posterior superimposed (grey density in the front); the additional red and blue densities represent

- the posterior density obtained through Metropolis–Hastings MCMC, by means of strong stationarity assumptions leading to the likelihood function shown in (19), and
- the approximate posterior obtained by adapting variational procedures in [18].

On the right-hand side we have the network observations on both the service station and the delay. Delay observations are displayed by subtracting their value from the job population N (which represents a second measurement on the service station). Whenever both observations match, these are displayed with a large dot. Along with it, we find the following:

- in grey, a mean-average network trajectory and 95% credible interval for job counts on the service station $X_t = (Y_t^{0,1} - Y_t^{1,0})_{t\geq 0}$, under the approximating measure $\mathbb{Q}$ and with methods introduced in this paper;
- in blue, a similar confidence interval and mean-average path obtained using benchmark methods in [18].

https://doi.org/10.1017/apr.2020.72 Published online by Cambridge University Press

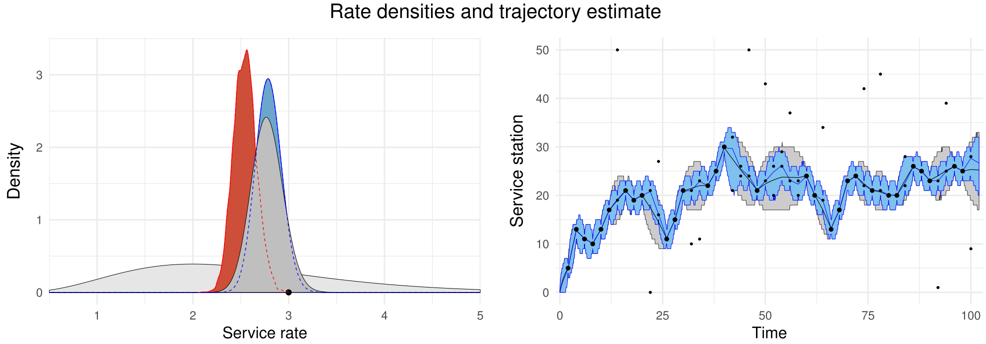

FIGURE 5: Left, prior density (light grey flat density in the back) and posterior density (dark grey density in the front) for λ, along with MCMC (red) and traditional variational (blue) density estimates. The black dot on the horizontal axis represents the original value in the network simulation. Right, network observations along with mean-average network trajectory and 95% credible interval for job counts in the service station; in grey, our proposed method, in blue, existing variational alernative method.

Noticeably, the average trajectory under our proposed variational technique flows through the most informative observations (thick dots), and the credible interval widens up to account for some faulty measurements within either network station. On the other hand, traditional variational approaches quickly converge to a local optimum and restrict mean-average dynamics, further compressing confidence intervals in regions with noisy data. In the next example, we notice how this poses a problem for traditional methods; that is, within complex and *synchronized* stochastic processes, we will fail to obtain sensible estimates for network parameters. Moreover, in our next example, inference by MCMC/forward–backward methods is virtually intractable (cf. [22]).

6.2. Multi-class parallel tandems with bottleneck and service priorities

We analyse an open multi-class QN as pictured in Figure 6. In this network, there exists two classes ($c = 1, 2$) of jobs that simultaneously transit the system. The first class consists of high-priority jobs with low arrival and service intensity rates. The second class consists of low-priority jobs with high arrival and service rates. Once a job enters the system, a probabilistic routing junction (pictured as a square within the figure) sends this job through either a PS or a priority-FCFS tandem; later, it will be serviced within an *infinite* station before leaving the network. In the top processor-sharing tandem, each station has five processing units. These will fraction their working capacity as seen in (8), in order to simultaneously service all jobs regardless of their class and priority level; however, service rates will differ depending on the job class. By contrast, the bottom tandem includes two FCFS stations with a single processing unit and priority scheduling. Here, low-priority jobs are serviced only if each station is entirely empty of any high-priority jobs. Consequently, the station loads in (9) are rewritten so that

$$\Upsilon(y, i, 1) = 1 \wedge x_{i,1} \quad \text{and} \quad \Upsilon(y, i, 2) = (1 \wedge x_{i,2}) \cdot \mathbb{I}(x_{i,1} < 1)$$

at stations $i \in \{2, 4\}$ and for any $y \in \mathcal{S}^Y$, where we recall that

$$x_{i,c} = \sum_{\eta \in \mathcal{T}_{i,c}^{\leftarrow}} y_\eta - \sum_{\eta \in \mathcal{T}_{i,c}^{\rightarrow}} y_\eta$$

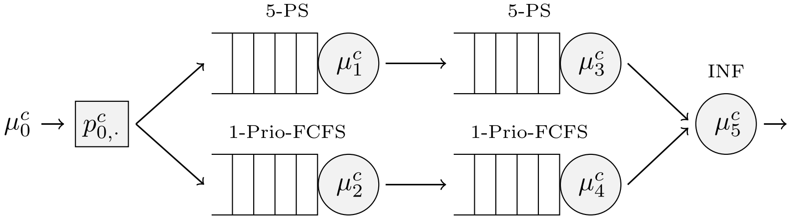

FIGURE 6: Open QN with one routing junction (pictured as a square) and five service stations with varied disciplines and processing rates.

and thus $x_{2,c} = y_{0,2,c} - y_{2,4,c}$, $x_{4,c} = y_{2,4,c} - y_{4,5,c}$, for $c = 1, 2$. Due to the presence of service priorities, the ordering of jobs within the queue is irrelevant (this is also the case with *random order* disciplines); hence, our inferential framework allows for different service rates assigned to jobs in each class. Finally, the last service station includes an infinite number of processing units, and processing rates also differ depending on the job class.

We analyse synthetic data created during a time interval $[0, T]$ ($T = 100$), with arrival intensities $\mu_0^1 = 0.5$, $\mu_0^2 = 3$, routing probabilities $p_{0,i}^c = 0.5$, $i, c \in \{1, 2\}$, and service rates as shown in Table 1. We collect a reduced set of *noiseless* and equally spaced observations with $K = 50$; these are essentially snapshots of the full system state across its service stations and job classes, so that $\mathcal{O} = \mathbb{N}^{10}$, and the observation density in (3) is defined with

$$f_{O|x}(\boldsymbol{o}) = \prod_{i=1}^{5} \prod_{c=1}^{2} \mathbb{I}(x_{i,c} = o_{i,c})$$

for $x \in \mathcal{S}$, where $o_{i,c}$ is an indexed observation in the element $\boldsymbol{o}$ denoting the class-c queue length at station $i > 0$. (Source code for the data simulation process may be found at https://github.com/IkerPerez/variationalQueues.) Within the inferential procedure, this observation likelihood must be approximated with some regularized variant similar to (18), while taking $\epsilon \to 0$. Next, we assign conjugate gamma priors to the various service intensities; in order to ensure identifiability in the problem, arrival rates and routing probabilities are fixed, and we focus this inferential task on the various service stations. Hence $\boldsymbol{\lambda} \equiv \{\mu_i^c : c = 1, 2 \text{ and } i = 1, \cdots, 5\}$, and we set $\lambda_{\boldsymbol{\eta}} \sim \Gamma(1, 0.3)$ under the reference measure $\mathbb{P}$, for all $\boldsymbol{\eta} \in \mathcal{T}$.

In the following, we omit the cumbersome mathematical details related to this complex model formulation, and we focus on discussing prior choices, calibration of the algorithm, results, and method comparisons following the inferential procedure.

Remarks on using MCMC data-augmentation for inference. Transient inference in a stochastic system with priorities is especially challenging, because of the strong dependencies this generates on the queue lengths across the stations and classes. Specifically,

- data-augmentation methods relying on MCMC techniques do not scale (cf. [27; 22]), as dependences yield very autocorrelated output chains;
- there exist no analytic product-form distributions to enable approximate inferential methods under assumptions of system stationarity, as discussed in the previous example; and
- *generator-matrix* exponentiations with a forward–backward algorithm [32] are simply unscalable to such large multivariate systems.

https://doi.org/10.1017/apr.2020.72 Published online by Cambridge University Press

TABLE 1: Summary statistics for posterior service rates in the QN in Figure 6.

	Real	O/S	Summary		Quantiles				
			Mean	StDev	2.5%	25%	50%	75%	97.5%
μ_1^1	**0.25**	0.364	0.307	0.043	0.228	0.276	0.304	0.335	0.397
μ_2^1	**1.5**	1.242	1.387	0.188	1.043	1.256	1.378	1.508	1.778
μ_3^1	**0.25**	0.421	0.339	0.049	0.250	0.305	0.337	0.371	0.442
μ_4^1	**1.5**	1.482	1.635	0.219	1.233	1.482	1.625	1.777	2.093
μ_5^1	**0.5**	0.837	0.761	0.075	0.622	0.709	0.758	0.810	0.915
μ_1^2	**0.5**	0.551	0.501	0.041	0.424	0.473	0.499	0.528	0.584
μ_2^2	**4.0**	3.496	3.740	0.298	3.177	3.534	3.731	3.935	4.346
μ_3^2	**0.5**	0.568	0.504	0.041	0.425	0.475	0.502	0.531	0.588
μ_4^2	**4.0**	3.265	3.670	0.296	3.112	3.465	3.661	3.863	4.270
μ_5^2	**1.0**	0.976	0.984	0.056	0.877	0.946	0.983	1.021	1.097

Remarks on using benchmark variational methods for inference. Note that traditional variational methods (cf. [18, 7]) are centred around populations or *lengths* in the individual queues. In this example, populations may not be factorized under an approximating measure $\mathbb{Q}$, since system jumps are synchronized; i.e., a jump down in one queue corresponds to a jump up in another. As a consequence, pairs of approximating rates under $\mathbb{Q}$ will be interlinked with the same *real* transition rate under $\mathbb{P}$, and derivations such as the lower bound in (11) or Equations (14)–(16) are unattainable. For the sake of completeness and comparisons, we adapt existing variational algorithms to the current task; however, we must make the following changes:

- We allow a factorization $\mathbb{Q}(X_t = \boldsymbol{x}) = \prod_i \mathbb{Q}(X_{i,t} = x_i)$, so that jobs may be virtually created and removed in any queue; i.e., jobs do not transit a network, but rather reach and depart servers individually. The full population of jobs in the network is not preserved.
- We duplicate intensities for transitions in the real model; that is, we have a rate for (i) a job departing a queue and (ii) the job arriving at another. Technically, a job could arrive at a new server before it departs the previous one; synchronization is lost; point estimates for parameters are averaged across pairs and weighted for network load.

As we will see, this leads to drastic performance issues that render the method unusable.

6.2.1 *Algorithm calibration.* Within our method, prior choices in the system state Y must initially accommodate a strictly positive, albeit not necessarily large, likelihood for low-priority jobs to be serviced at any point in time. Here, we achieve this by assigning Poisson process priors to task transition counts in Y; that is, we first run the master equation (17) with some user-specified constant intensity rates. This creates monotone mean-average queue lengths in the service stations, and we ensure that they flow aligned to the network observations in every instance.

Also, the presence of strong temporal dependencies will often trigger the approximating rates ν in (14) to become unreasonably large, ultimately rendering the algorithm computationally infeasible. This is a phenomenon also observed in [7, 18], within the context of simpler

https://doi.org/10.1017/apr.2020.72 Published online by Cambridge University Press

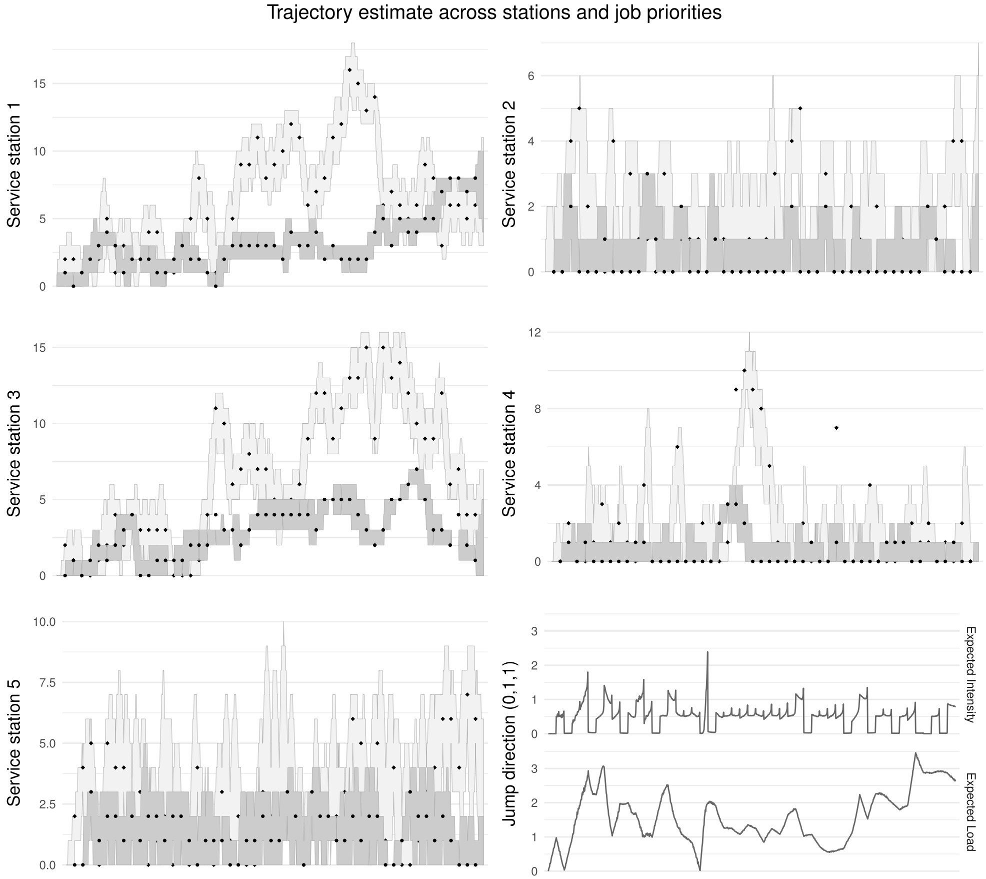

FIGURE 7: 95% credible intervals for queue lengths across the service stations. Dark (light) grey corresponds to high-priority (low-priority) jobs. Bottom right panel shows expected jump intensity and station load in the direction $\eta = (0, 1, 1)$.

stochastic dynamics. To ensure computational tractability, we exploit the capping functionals k as in introduced in Proposition 1 and set a global rate cap of $\bar{\nu} = 50$. Furthermore, we run the differential equations for r in (16) in log-form. Specific details can be found within the aforementioned source code.

6.2.2 *Results.* In the plots in Figure 7, with the exception of the bottom right one, we observe 95% credible intervals for the queue length processes $X_t^{i,c}$ over time, across the various service stations and job classes. In the figure, intervals in dark grey relate to high-priority jobs, and their corresponding queue length observations are represented by black circles. This information is superimposed over its analogue for low-priority jobs, where intervals are coloured in light grey and observations represented by small diamonds. These interval approximations ignore small positive densities that are sometimes assigned to negative queue lengths. Note that this is a consequence of employing counts across job transitions in Y as a basis for inference on X; however, we recall that this is necessary to overcome the coupling challenges described in Sections 2 and 4. Overall, we note that the variational flow captures the collected observations

https://doi.org/10.1017/apr.2020.72 Published online by Cambridge University Press

well, with some few exceptions in the stations with priority scheduling; hence, it offers a good basis for building approximate estimates for parameters and the likelihood ratio (5).

The bottom right plot in Figure 7 shows an overview of the expected jump intensity

$$\mathbb{E}^{\mathbb{Q}}_{Y_t^{\boldsymbol{\eta}}}[\nu_t^{\boldsymbol{\eta}}(Y_t^{\boldsymbol{\eta}})]$$

and station load

$$\mathbb{E}^{\mathbb{Q}}_{Y_t}[\Upsilon(Y_t, \eta_1, \eta_3) \vee 0]$$

in the direction $\boldsymbol{\eta} = (0, 1, 1)$ at times $t \in [0, T]$. The sharp peaks in the intensities come at observation times, ensuring that the process density transits through the observations. Finally, we notice that the expected station load differs from the estimate of the high-priority queue length in station 1, as this process combines and weights the queue length across the two priorities according to (8).

Next, we find in Table 1 summary statistics for the posterior service rates under the approximating variational measure $\mathbb{Q}$, along with point estimates obtained by adapting benchmark variational techniques in [18]. We observe that the proposed framework allows us to gain a good overview of the system properties and variability in the processing speed across the various stations; by contrast, existing methods are far from offering reasonable approximations to system behaviour (they instead seem to construct an *averaged* estimate of network flow). Noticeably, there exist a few significant deviations from real values, within the posterior estimates for high-priority service rates in PS stations. This is likely due to a combination of sampling variance, high model complexity, and the limitations of such approximate variational procedures for transient analyses of stochastic processes.

7. Discussion

In this paper, we have enabled the variational evaluation of approximating measures for partially-observed *coupled* systems of jump stochastic processes, with a focus on mixed systems of QNs. Furthermore, we have presented a flexible approximate Bayesian framework, capable of overcoming the challenges posed by coupling properties, and applicable in scenarios where existing MCMC or variational solutions are unusable. To achieve this goal, we have built on existing variational theory (see [18, 7]) and discussed an alternate optimization procedure with slack variables and inequality constraints that can address computational limitations within existing techniques. Notably, results within this paper contribute to existing Bayesian statistical literature in [27, 29, 22] and allow for the study of the latent stochastic behaviour across complex mixed network models, by means of an augmented process for interactions in the resources.

Even though the proposed framework relies on an approximated network model as a basis for inference (which ensures the absolute continuity across base measures), and while it further analyses queue lengths by means of augmented job transitions in the resources, we have shown that we can reliably capture the finite-dimensional posterior distributions of the various marginal stochastic processes and offer a good overview of the network structure and likely flow of workload. This is important as it can enable evaluation and uncertainty quantification tasks in several networked systems, found in many application domains, where full data observations may be hard to retrieve. Current state-of-the-art alternatives rely on strong assumptions leading to stationary analyses of such systems, or use alternate MCMC procedures that reportedly encounter limitations due to existing computational constraints [27, 22].

https://doi.org/10.1017/apr.2020.72 Published online by Cambridge University Press

Appendix A. Construction

Let $(\Omega, \mathcal{F})$ be a measurable space with the regular conditional probability property; also, let $0 \le t_1 < \cdots < t_K \le T$ be some fixed observation times, with $T > 0$. In a standard QN with M stations, Ω may denote a product set supporting instantaneous rates, trajectories, and observations, and $\mathcal{F}$ the corresponding product σ-algebra. The space of rates and observations will consist of trivial Borel algebras and power sets, so that $\boldsymbol{\lambda}$ is an $(\mathbb{R}^n_+, \mathcal{B}(\mathbb{R}^n_+))$-valued random variable of rates in the infinitesimal generator matrix Q of X, where $n \in \mathbb{N}$ denotes an arbitrary number determined by the network topology. In addition, $\{O_k{:}k = 1, \dots, K\}$ corresponds to random measurement variables for the network monitoring activity, each defined on $(\mathcal{O}, \mathcal{P}(\mathcal{O}))$, where $\mathcal{O}$ denotes an arbitrary countable support set for observations in every service station. A network trajectory $X = (X_t)_{0\le t\le T}$ is an $(\mathcal{S}, \mathcal{P}(\mathcal{S}))$-valued stochastic process with a countably infinite support set $\mathcal{S}$. Note that this is a piecewise deterministic jump process, so that $X = (\boldsymbol{t}, \boldsymbol{x})$ is formed by a sequence of transition times $\boldsymbol{t}$ along with states $\boldsymbol{x}$. Every pair $(\boldsymbol{t}, \boldsymbol{x})$ can be further defined as a random variable on a measurable space $(\mathcal{X}, \Sigma_{\mathcal{X}})$, with $\mathcal{X} = \cup_{i=0}^{\infty}([0, T] \times \mathcal{S})^i$ and $\Sigma_{\mathcal{X}}$ the corresponding union σ-algebra. This space can support all finite $\mathcal{S}$-valued trajectories and allows the assignment of a dominating base measure $\mu_{\mathcal{X}}$ with respect to which to define a trajectory density. For details, we refer the reader to [9].

Let $\mathbb{P}$ be a reference probability measure on $(\Omega, \mathcal{F})$. For all $A \in \mathcal{B}(\mathbb{R}^n_+)$, we write

$$\mathbb{P}(\boldsymbol{\lambda}^{-1}(A)) = \mathbb{P}_{\boldsymbol{\lambda}}(A) = \int_A f_{\boldsymbol{\lambda}}(\boldsymbol{a})\, \mu_{\mathbb{R}^n_+}(d\boldsymbol{a}),$$

where $f_{\boldsymbol{\lambda}}$ denotes the joint density function of n independent gamma-distributed variables. Hence, we assume that the distribution of instantaneous rates under $\mathbb{P}$ admits a density carried by the (Lebesgue) measure $\mu_{\mathbb{R}^n_+}$. Next, let $\kappa_1{:}\mathcal{F} \times \mathbb{R}^n_+ \to [0, 1]$ be a regular conditional probability, i.e., a Markov kernel that defines a probability measure on $\mathcal{F}$ for all $\boldsymbol{\lambda} \in \mathbb{R}^n_+$, with

$$\mathbb{P}(B \cap \boldsymbol{\lambda}^{-1}(A)) = \int_A \kappa_1(B, \boldsymbol{a}) f_{\boldsymbol{\lambda}}(\boldsymbol{a})\, \mu_{\mathbb{R}^n_+}(d\boldsymbol{a})$$

for $A \in \mathcal{B}(\mathbb{R}^n_+)$ and $B \in \mathcal{F}$. By definition, $\kappa_1(B, \boldsymbol{a}) = \mathbb{P}(B|\boldsymbol{\lambda} = \boldsymbol{a})$, and most importantly,

$$\kappa_1(X^{-1}(C), \boldsymbol{a}) = \int_C f_{X|\boldsymbol{\lambda}=\boldsymbol{a}}(\boldsymbol{t}, \boldsymbol{x})\, \mu_{\mathcal{X}}(dt, d\boldsymbol{x})$$

for all $C \in \Sigma_{\mathcal{X}}$ (note this often constitutes an intractable integral). The conditional density $f_{X|\boldsymbol{\lambda}=\boldsymbol{a}}$ is such that for every $I \in \mathbb{N}$ and pair of ordered times $\boldsymbol{t} = \{0, t_1, \dots, t_I\}$ in $[0, T]$ and states $\boldsymbol{x} = \{x_0, \dots, x_I\}$ in $\mathcal{S}$ we have

$$f_{X|\boldsymbol{\lambda}=\boldsymbol{a}}(\boldsymbol{t}, \boldsymbol{x}) = \pi(x_0)\, e^{Q_{x_I}(T-t_I)} \prod_{i=1}^{I} Q_{x_{i-1},x_i}\, e^{Q_{x_{i-1}}(t_i-t_{i-1})},$$

where $Q \equiv Q(\boldsymbol{a})$ is the matrix of infinitesimal transition rates in X associated to values in $\boldsymbol{a}$. Finally, network observations are assumed to be discrete events, independent of transition rates given a trajectory. Thus, there exists a kernel $\kappa_2{:}\mathcal{F} \times (\mathcal{X} \times \mathbb{R}^n_+) \to [0, 1]$ such that

$$\mathbb{P}(O_k \in D|X = (\boldsymbol{t}, \boldsymbol{x}), \boldsymbol{\lambda} = \boldsymbol{a}) = \kappa_2(O_k^{-1}(D), (\boldsymbol{t}, \boldsymbol{x}), \boldsymbol{a}) = \sum_{d\in D} f_{O_k|(\boldsymbol{t},\boldsymbol{x})}(d)\, \mu_{\mathcal{O}}(d)$$

for all $k = 1, \dots, K$ and $D \in \mathcal{P}(\mathcal{O})$. Here, $f_{O_k|(\boldsymbol{t},\boldsymbol{x})}$ defines an arbitrary probability mass function on $\mathcal{O}$ carried by a counting measure; in our applications, each observation depends only on the state of the system at the observation time, so the above expression could be further simplified.

Under the above model construction, the support over infinitesimal rates is a standard Borel space, and the existence of a posterior distribution is guaranteed (cf. [20]). Also, measures induced by the kernel κ_2 are σ-finite and such that $\kappa_2(\cdot, (\boldsymbol{t}, \boldsymbol{x}), \boldsymbol{a}) << \mu_{\mathcal{O}}$ for every $((\boldsymbol{t}, \boldsymbol{x}), \boldsymbol{a}) \in \mathcal{X} \times \mathbb{R}^n_+$. The posterior is thus carried by its corresponding prior and defined by means of the Radon–Nikodym derivative

$$\frac{d\mathbb{P}_{\boldsymbol{\lambda}|O_1=o_1,\dots,O_K=o_K}}{d\mathbb{P}_{\boldsymbol{\lambda}}}(\boldsymbol{a}) = \frac{\int_{\mathcal{X}} \prod_{k=1}^K f_{O_k|(\boldsymbol{t},\boldsymbol{x})}(o_k) f_{X|\boldsymbol{\lambda}=\boldsymbol{a}}(\boldsymbol{t}, \boldsymbol{x})\, \mu_{\mathcal{X}}(d\boldsymbol{t}, d\boldsymbol{x})}{\int_{\mathbb{R}^n_+} \int_{\mathcal{X}} \prod_{k=1}^K f_{O_k|(\boldsymbol{t},\boldsymbol{x})}(o_k) f_{X|\boldsymbol{\lambda}=\boldsymbol{a}}(\boldsymbol{t}, \boldsymbol{x})\, \mu_{\mathcal{X}}(d\boldsymbol{t}, d\boldsymbol{x}) f_{\boldsymbol{\lambda}}(\boldsymbol{a}) \mu_{\mathbb{R}^n_+}(d\boldsymbol{a})},$$

where we employ the shorthand notation $d\mathbb{P}_{\boldsymbol{\lambda}|\cdot}(\boldsymbol{a}) = \mathbb{P}_{\boldsymbol{\lambda}}(d\boldsymbol{a}|\cdot)$.

Acknowledgements

The authors would like to thank the anonymous reviewers, who greatly improved the quality of this paper. This work was supported by RCUK through the Horizon Digital Economy Research grants EP/G065802/1 and EP/M000877/1.

References

[1] Armero, C. and Bayarri, M. J. (1994). Prior assessments for prediction in queues. *Statistician* **43**, 139–153.

[2] Baele, G., van de Peer, Y. and Vansteelandt, S. (2010). Using non-reversible context-dependent evolutionary models to study substitution patterns in primate non-coding sequences. *J. Molec. Evolution* **71**, 34–50.

[3] Baskett, F., Chandy, K. M., Muntz, R. R. and Palacios, F. G. (1975). Open, closed, and mixed networks of queues with different classes of customers. *J. Assoc. Comput. Mach.* **22**, 248–260.

[4] Blei, D. M., Kucukelbir, A. and McAuliffe, J. D. (2017). Variational inference: a review for statisticians. *J. Amer. Statist. Assoc.* **112**, 859–877.

[5] Bobbio, A., Gribaudo, M. and Telek, M. (2008). Analysis of large scale interacting systems by mean field method. In *2008 Fifth International Conference on Quantitative Evaluation of Systems*, Institute of Electrical and Electronics Engineers, Piscataway, NJ, pp. 215–224.

[6] Buzacott, J. A. and Shanthikumar, J. G. (1993). *Stochastic Models of Manufacturing Systems*, Vol. **4**. Prentice Hall, Englewood Cliffs, NJ.

[7] Cohn, I., El-Hay, T., Friedman, N. and Kupferman, R. (2010). Mean field variational approximation for continuous-time Bayesian networks. *J. Mach. Learning Res.* **11**, 2745–2783.

[8] Cooper, R. B. (1981). Queueing theory. In *Proceedings of the ACM'81 conference*, Association for Computing Machinery, New York, pp. 119–122.

[9] Daley, D. J. and Vere-Jones, D. (2007). *An Introduction to the Theory of Point Processes, Volume II: General Theory and Structure*. Springer, New York.

[10] Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* **81**, 2340–2361.

[11] Golightly, A. and Wilkinson, D. J. (2015). Bayesian inference for Markov jump processes with informative observations. *Statist. Appl. Genet. Molec. Biol.* **14**, 169–188.

[12] Hobolth, A. and Stone, E. A. (2009). Simulation from endpoint-conditioned, continuous-time Markov chains on a finite state space, with applications to molecular evolution. *Ann. Appl. Statist.* **3**, 1204–1231.

[13] Huelsenbeck, J. P., Bollback, J. P. and Levine, A. M. (2002). Inferring the root of a phylogenetic tree. *Systematic Biol.* **51**, 32–43.

[14] Kleinrock, L. (1976). *Queueing Systems, Vol. II: Computer Applications*. John Wiley, New York.

[15] Koole, G. and Mandelbaum, A. (2002). Queueing models of call centers: an introduction. *Ann. Operat. Res.* **113**, 41–59.

[16] Kraft, S., Pacheco-Sanchez, S., Casale, G. and Dawson, S. (2009). Estimating service resource consumption from response time measurements. In *Valuetools '09: Proceedings of the Fourth International ICST Conference on Performance Evaluation Methodologies and Tools*, Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, Brussels, pp. 1–10.

[17] Liu, Z., Wynter, L., Xia, C. H. and Zhang, F. (2006). Parameter inference of queueing models for IT systems using end-to-end measurements. *Performance Evaluation* **63**, 36–60.

https://doi.org/10.1017/apr.2020.72 Published online by Cambridge University Press

[18] Opper, M. and Sanguinetti, G. (2008). Variational inference for Markov jump processes. In *Advances in Neural Information Processing Systems*, MIT Press, Boston, MA, pp. 1105–1112.

[19] Opper, M. and Sanguinetti, G. (2010). Learning combinatorial transcriptional dynamics from gene expression data. *Bioinformatics* **26**, 1623–1629.

[20] Orbanz, P. and Teh, Y. W. (2011). Bayesian nonparametric models. In *Encyclopedia of Machine Learning*, Springer, Boston, MA, pp. 81–89.

[21] Osorio, C. and Bierlaire, M. (2009). An analytic finite capacity queueing network model capturing the propagation of congestion and blocking. *Europ. J. Operat. Res.* **196**, 996–1007.

[22] Perez, I., Hodge, D. and Kypraios, T. (2017). Auxiliary variables for Bayesian inference in multi-class queueing networks. *Statist. Comput.* **28**, 1187–1200.

[23] Perez, I. and Kypraios, T. (2019). *Scalable Bayesian inference for population Markov jump processes*. Preprint. Available at https://arxiv.org/abs/1904.08356.

[24] Rao, V. A. and Teh, Y. W. (2013). Fast MCMC sampling for Markov jump processes and extensions. *J. Mach. Learning Res.* **14**, 3295–3320.

[25] Serfozo, R. F. (1972). Conditional Poisson processes. *J. Appl. Prob.* **9**, 288–302.

[26] Spinner, S., Casale, G., Brosig, F. and Kounev, S. (2015). Evaluating approaches to resource demand estimation. *Performance Evaluation* **92**, 51–71.

[27] Sutton, C. and Jordan, M. I. (2011). Bayesian inference for queueing networks and modeling of internet services. *Ann. Appl. Statist.* **5**, 254–282.

[28] Wang, C., Blei, D. and Heckerman, D. (2012). *Continuous time dynamic topic models*. Preprint. Available at https://arxiv.org/abs/1206.3298.

[29] Wang, W., Casale, G. and Sutton, C. (2016). A Bayesian approach to parameter inference in queueing networks. *ACM Trans. Model. Comput. Simul.* 27, 2:1–2:26.

[30] Zechner, C. *et al.* (2014). Scalable inference of heterogeneous reaction kinetics from pooled single-cell recordings. *Nature Methods* **11**, 197–202.

[31] Zhang, B., Pan, J. and Rao, V. A. (2017). Collapsed variational Bayes for Markov jump processes. In *Advances in Neural Information Processing Systems*, MIT Press, Boston, MA, pp. 3749–3757.

[32] Zhao, T. *et al.* (2016). Bayesian analysis of continuous time Markov chains with application to phylogenetic modelling. *Bayesian Anal.* **11**, 1203–1237.

https://doi.org/10.1017/apr.2020.72 Published online by Cambridge University Press