Learning traffic correlations in multi-class queueing systems by sampling queue lengths, with routing applications

Martin Zubeldia
Georgia Institute of Technology
Atlanta, USA

Michel Mandjes
University of Amsterdam
Amsterdam, Netherlands

ABSTRACT

We consider a system of parallel single-server queues. Work of different classes arrives as correlated Gaussian processes with known drifts but unknown covariance matrix Σ, and it is deterministically routed to the different queues according to some routing matrix.

We first provide a procedure to estimate Σ based on the empirical large deviations behavior of the individual queues, for a finite set of routing matrices, and prove that the resulting estimate is asymptotically consistent under minimal technical conditions. We also introduce a more efficient procedure to estimate Σ based on the empirical large deviation behavior of linear combinations of queues, for a single routing matrix, and prove that the resulting estimate is asymptotically consistent under some technical conditions. We establish, however, that in specific cases the latter approach cannot be used due to an inherent loss of information produced by the dynamics of the queues.

Finally, given a well-behaved cost function on the steady-state marginal queue lengths, we show how our procedures can be used to obtain an asymptotically consistent estimator for the cost under *any* routing matrix, and identify an optimal one.

CCS CONCEPTS

• **Mathematics of computing** → **Probabilistic inference problems**; • **Networks** → *Traffic engineering algorithms.*

KEYWORDS

Indirect learning; Large deviations; Gaussian processes; Queueing systems

ACM Reference Format:
Martin Zubeldia and Michel Mandjes. 2021. Learning traffic correlations in multi-class queueing systems by sampling queue lengths, with routing applications. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In networks that handle huge volumes of traffic, such as the backbone of the Internet, routing is done in a mostly static way, by fixing the fractions of traffic that follow each available route, and go to each major data center. This is implemented via a hash value calculated from the flow five-tuple in the packet header, which can be done at a sufficiently high rate. With these high traffic networks in mind, this paper proposes a (self-learning) traffic engineering approach where an optimal static routing is learned based on observations of the network congestion level. In particular, we develop a pair of procedures that learn the characteristics of the incoming traffic by observing the network elements' congestion levels, thus enabling us to identify an optimal static routing. Moreover, this routing can be adapted at a longer timescale, as soon as it is detected that the average congestion has undergone some fundamental change. This design principle for the control of networks with learning algorithms, which focuses on exploiting the underlying structure to more efficiently design the learning schemes, has also been exploited in [2, 8, 9].

First, it is important to realize that control based on traffic intensities only is inadequate. Indeed, the variability on top of the mean rate has a crucial impact on the performance achieved. The incorporation of the traffic streams' variance directly relates to the concept of *effective bandwidth* [4–6, 16]. Informally, an effective bandwidth summarizes the bandwidth consumption of a traffic stream as a function of the performance target imposed (and is in particular increasing in the traffic stream's variability). As long as the sum of the effective bandwidths using a specific link remain below the capacity of this link, the performance target will be achieved.

Effective bandwidths are an intensively used approach in resource allocation. It is noted, though, that in the literature on this topic one typically ignores possible correlations between the different traffic streams, a factor that potentially significantly affects performance. A perhaps more serious conceptual issue in the work on effective bandwidth is that it is assumed that one *knows* the characteristics (in terms of a probabilistic description) of the processes involved. In practice, however, this is usually not the case. This means that one needs an algorithm that can *learn* the optimal static routing. This process can be complicated if the required measurements cannot be performed at the same time scale as the fluctuations of the incoming traffic.

Due to the issues mentioned above, we do not rely on *traffic measurements*, which are often hard to be obtained reliably, in particular when it concerns fine timescales; in line with the approach developed in [13, 14] we rely on *buffer occupancy measurements* performed at a relatively low frequency. We throughout assume that the traffic streams' *mean* transmission rates are known, as these are typically easy to estimate (also from coarse time-scale traffic measurements); see the discussion in [13, 14].

On a more technical note, the network considered in this paper is a set of parallel single server queues. Work of different classes arrives to the system as (possibly correlated) stationary stochastic processes, which are deterministically routed to the different

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

queues. Justified by central limit arguments and backed by extensive experimental studies [7, 17], we assume that these processes are Gaussian (for a textbook treatment of such Gaussian queues we refer to [10]). In our analysis we rely on representations of the logarithmic asymptotics of the queue length distribution. These were developed for single queues in e.g. [1, 3]; in the special, practically relevant case that the Gaussian input process is fractional Brownian motion (fBm), this decay rate has a clean explicit expression. Later these results were extended to more involved network structures in e.g. [15, 18]. Related work on the logarithmic asymptotics of long busy periods in fBm-driven queues can be found in [11], which can be used to describe the speed of convergence of fractional Brownian storage to its stationary limit [12].

Finally, it is important to mention that [13, 14] consider the learning problem in the single-queue context, whereas in the present paper we consider multiple queues in parallel, with the explicit goal of finding an optimal static routing strategy. An interesting new dimension to our approach is due to the multi-dimensional nature of our setup: as it turns out, it entails that in the multi-dimensional context there is a natural limit on how much can be learned about the variance and covariance functions of the arrival processes by observing the (joint, in this case) queue length process. Moreover, while in [13, 14] their claims were mostly justified by numerical simulations, the present paper formally establishes the asymptotic consistency of the proposed estimators.

1.1 Our contribution

We introduce two procedures to estimate the variance and covariance functions of the arrival processes based on estimates of the steady-state queue length distributions. These can be used when setting up an algorithm to determine an optimal routing matrix. The two learning procedures can be characterized in greater detail as follows:

(i) The first procedure adapts the inversion formula that was introduced in [13, 14] for a single queue to obtain an estimate of the variance functions of the input processes to our multi-queue network. It does so by using the empirical marginal steady-state queue length distributions for appropriately chosen routing matrices.
(ii) The second procedure is based on a new inversion formula that directly obtains an estimate of the covariance functions of the input processes to the queues from the empirical pairwise joint steady-state queue length distributions for a single routing matrix.

Since the variance and covariance functions of the input processes are known linear combinations of the variance and covariance functions of the arrival processes, the latter are obtained by solving a system of linear equations with known coefficients that only depend on the routing matrices. In this respect, we prove that:

(i) Under mild conditions on the arrival processes, the first procedure yields asymptotically consistent estimators for the true variance and covariance functions of the arrival processes.
(ii) When the input processes are short-range dependent and non-negatively correlated, our second procedure yields asymptotically consistent estimators for the true variance and covariance functions of the arrival processes.
(iii) When the input processes are long-range dependent and negatively correlated, it is impossible to accurately estimate the covariance functions of the input processes from the large deviations behavior of the steady-state distribution of linear combinations of queue lengths.

Finally, our learning algorithms can be used to obtain estimates of the steady-state queue length distributions for any routing matrix (i.e., also for routing matrices different from the one used in the estimation), and to efficiently learn an optimal one (i.e., minimizing a given cost function).

The above results are in the full version of this article [19].

ACKNOWLEDGEMENTS

This work was partially supported by ONR grant N00014-17-1-2790.

REFERENCES

[1] R. Addie, P. Mannersalo, and I. Norros. 2002. Most probable paths and performance formulae for buffers with Gaussian input traffic. *Europeans Transactions on Telecommunications* 13 (2002), 183–196.
[2] J. Dai and M. Gluzman. 2020. Queueing network controls via deep reinforcement learning. (2020). Preprint arXiv:2008.01644.
[3] N. Duffield and N. O'Connell. 1995. Large deviations and overflow probabilities for the general single-server queue, with applications. *Mathematical Proceedings of the Cambridge Philosophical Society* 118 (1995), 363–374.
[4] A. Elwalid and D. Mitra. 1993. Effective bandwidth of general Markovian traffic sources and admission control of high speed networks. *IEEE/ACM Transactions on Networking* 1 (1993), 329–343.
[5] R. Guérin, H. Ahmadi, and M. Naghshineh. 1991. Equivalent capacity and its application to bandwidth allocation in high-speed networks. *IEEE Journal on Selected Areas in Communications* 9 (1991), 968–981.
[6] J. Hui. 1988. Resource allocation for broadband networks. *IEEE Journal Selected Areas Communication* 6 (1988), 1598–1608.
[7] J. Kilpi and I. Norros. 2002. Testing the Gaussian approximation of aggregate traffic. In *Internet Measurement Workshop*. 49–61.
[8] S. Krishnasamy, A. Arapostathis, R. Johari, and S. Shakkottai. 2018. On learning the $c\mu$ rule in single and parallel server networks. In *Proceedings of the 56th Annual Allerton Conference*. 153–154.
[9] S. Krishnasamy, R. Sen, R. Johari, and S. Shakkottai. 2021. Learning unknown service rates in queues: a multiarmed bandit approach. *Operations Research, to appear* (2021).
[10] M. Mandjes. 2007. *Large Deviations for Gaussian Queues: Modelling Communication Networks*. John Wiley & Sons.
[11] M. Mandjes, P. Mannersalo, I. Norros, and M. van Uitert. 2006. Large deviations of infinite intersections of events in Gaussian processes. *Stochastic Processes and their Applications* 9 (2006), 1269–1293.
[12] M. Mandjes, I. Norros, and P. Glynn. 2009. On convergence to stationarity of fractional Brownian storage. *The Annals of Applied Probability* 18 (2009), 1385–1403.
[13] M. Mandjes and R. van de Meent. 2005. Inferring traffic burstiness by sampling the buffer occupancy. In *NETWORKING 2005*. 303–315.
[14] M. Mandjes and R. van de Meent. 2009. Resource dimensioning through buffer sampling. *IEEE/ACM Transactions on Networking* 17 (2009), 1631–1644.
[15] M. Mandjes and M. van Uitert. 2005. Sample-path large deviations for tandem and priority queues with Gaussian inputs. *The Annals of Applied Probability* 15 (2005), 1193–1226.
[16] A. Pras, L. Nieuwenhuis, R. van de Meent, and M. Mandjes. 2009. Dimensioning network links: a new look at equivalent bandwidth. *IEEE Network* 23 (2009), 5–10.
[17] R. van de Meent, M. Mandjes, and A. Pras. 2006. Gaussian traffic everywhere?. In *Proceedings of the 2006 IEEE International Conference on Communications*. 573–578.
[18] M. Zubeldia and M. Mandjes. 2021. Large deviations for acyclic networks of queues with correlated Gaussian inputs. *Queueing Systems* (2021).
[19] M. Zubeldia and M. Mandjes. 2021. Learning traffic correlations in multi-class queueing systems by sampling queue lengths, with routing applications. *Performance Evaluation* 152 (2021).