Recent Papers on Bounds for Queues

Ward Whitt

Department of Industrial Engineering and Operations Research, Columbia University, New York, NY, 10027; ww2040@columbia.edu

January 25, 2020

Abstract

This is an overview of my papers on bounds for queues, emphasizing recent work with Yan Chen.

1 Overview

In these notes we give a brief overview of bounds for the mean waiting time in the single-server queue. Recent papers Chen and Whitt [2018a,b, 2019] have revisited earlier papers Whitt [1984a,b] and Klincewicz and Whitt [1984].

1.1 The $GI/GI/1$ Model

The $GI/GI/1$ single-server queue has unlimited waiting space and the first-come first-served service discipline. There is a sequence of independent and identically distributed (i.i.d.) service times $\{V_n : n \geq 0\}$, each distributed as V with cumulative distribution function (cdf) G, which is independent of a sequence of i.i.d. interarrival times $\{U_n : n \geq 0\}$ each distributed as U with cdf F. With the understanding that a 0^{th} customer arrives at time 0, V_n is the service time of customer n, while U_n is the interarrival time between customers n and $n+1$.

Let U have mean $E[U] \equiv 1$ and squared coefficient of variation (scv, variance divided by the square of the mean) c_a^2; let a service time V have mean $E[V] \equiv \tau \equiv \rho$ and scv c_s^2, where $\rho < 1$, so that the model is stable. (Let $\equiv$ denote equality by definition.)

Let W_n be the waiting time of customer n, i.e., the time from arrival until starting service, assuming that the system starts with an initial workload W_0 having cdf H_0 with a finite mean. The sequence $\{W_n : n \geq 0\}$ is well known to satisfy the Lindley recursion

$$W_n = [W_{n-1} + V_{n-1} - U_{n-1}]^+, \quad n \geq 1, \tag{1.1}$$

where $x^+ \equiv \max\{x, 0\}$. Let H_n be the cdf of W_n, which is determined by (1.1). Let $W \equiv W_\infty$ (both used) be the steady-state waiting time, satisfying $W_n \Rightarrow W_\infty$ as $n \to \infty$, where $\Rightarrow$ denotes convergence in distribution; see §§X.1-X.2 of Asmussen [2003]. The cdf H_∞ of W_∞ is the unique cdf satisfying the stochastic fixed point equation

$$W_\infty \stackrel{\mathrm{d}}{=} (W_\infty + V - U)^+, \tag{1.2}$$

where $\stackrel{\mathrm{d}}{=}$ denotes equality in distribution. If $P(W_0 = 0) = 1$, then $W_n \stackrel{\mathrm{d}}{=} \max\{S_k : 0 \le k \le n\}$ for $n \le \infty$, $S_0 \equiv 0$, $S_k \equiv X_0 + \cdots + X_{k-1}$ and $X_k \equiv V_k - U_k$, $k \ge 1$. Under the specified finite moment conditions, for $1 \le n \le \infty$, W_n is a proper random variable with finite mean, given by

$$E[W_n|W_0 = 0] \;=\; \sum_{k=1}^{n} \frac{E[S_k^+]}{k} < \infty, \quad 1 \le n < \infty, \quad \text{and} \quad E[W_\infty] = \sum_{k=1}^{\infty} \frac{E[S_k^+]}{k} < \infty. \tag{1.3}$$

1.2 Classical Steady-State Results: Exact, Approximate and Bounds

For the $M/GI/1$ special case, when the interarrival time has an exponential distribution, we have the classical Pollaczek-Khintchine formula

$$E[W] = \frac{\tau\rho(1 + c_s^2)}{2(1-\rho)} = \frac{\rho^2(1 + c_s^2)}{2(1-\rho)}. \tag{1.4}$$

A natural commonly used approximation for the $GI/GI/1$ model, inspired by (1.4), which we call the heavy-traffic approximation, because it is motivated by the early heavy-traffic limit in Kingman [1961], is

$$E[W] \equiv E[W(\rho, c_a^2, c_s^2)] \approx \frac{\rho^2(c_a^2 + c_s^2)}{2(1-\rho)}. \tag{1.5}$$

The heavy traffic limit for the mean states that $(1-\rho)E[W(\rho, c_a^2, c_s^2)] \to (c_a^2 + c_s^2)/2$ as $\rho \uparrow 1$.

The most familiar upper bound (UB) on $E[W]$ is the Kingman [1962] bound,

$$E[W] \le \frac{\rho^2([c_a^2/\rho^2] + c_s^2)}{2(1-\rho)}, \tag{1.6}$$

which also satisfies the same heavy traffic limit.

A better UB depending on these same parameters was obtained by Daley [1977]. in particular, the Daley [1977] UB replaces the term c_a^2/ρ^2 by $(2-\rho)c_a^2/\rho$, i.e.,

$$E[W] \le \frac{\rho^2([(2-\rho)c_a^2/\rho] + c_s^2)}{2(1-\rho)}. \tag{1.7}$$

Note that $(2-\rho)/\rho < 1/\rho^2$ because $\rho(2-\rho) < 1$ for all ρ, $0 < \rho < 1$.

In contrast to the tight UB that we study, the tight lower bound (LB) for the steady-state mean has been known for a long time; see Stoyan and Stoyan [1974], §5.4 of Stoyan [1983], §V of Whitt [1984a], Theorem 3.1 of Daley et al. [1992] and references there. The LB is

$$E[W] \geq \frac{\rho((1+c_s^2)\rho - 1)^+}{2(1-\rho)}. \tag{1.8}$$

The LB is attained asymptotically at a deterministic interarrival time with the specified mean and at any three-point service-time distribution that has all mass on nonnegative-integer multiples of the deterministic interarrival time. The service part follows from Ott [1987]. (All service-time distributions satisfying these requirements yield the same mean.)

1.3 Motivation: Approximations for Non-Markovian Open Queueing Networks

One source of motivation for the bounds is provided by parametric-decomposition approximations for non-Markovian open networks of single-server queues, as in Whitt [1983], where each queue is approximated by a $GI/GI/1$ queue partially characterized by the parameter vector $(\lambda, c_a^2, \tau, c_s^2)$, obtained by solving traffic rate equations for the arrival rate λ at each queue and after solving associated traffic variability equations to generate an approximating scv c_a^2 of the arrival process. Because the internal arrival processes are usually not renewal and the interarrival distribution is not known, there is no concrete $GI/GI/1$ model to analyze. To gain some insight into these approximations (not yet addressing the dependence among interarrival times), It is natural to regard such approximations for the $GI/GI/1$ model as set-valued functions, applying to all models with the same parameter vector $(\lambda, c_a^2, \tau, c_s^2)$.

For the special case of the $GI/M/1$ model with bounded support for the interarrival-time cdf F, the extremal $GI/M/1$ models were studied in Whitt [1984a], where intervals of bounded support were also used together with the theory of Tchebychev systems, as in Karlin and Studden [1966], drawing on Rolski [1972], Holtzman [1973] and Eckberg [1977].(The focus in Whitt [1984a] was on the mean steady state number of customers in the system, but it is easily seen that the extremal interarrival-time distributions are the same for the mean number of customers in the system and the mean steady-state waiting time, because they both depend on the root of the same equation.) For the $GI/M/1$ model, the extremal distributions are two-point distributions.

Let $\mathcal{P}_{2,2}(M) \equiv \mathcal{P}_{2,2}(m_1, c^2, M)$ be the set of all two-point distributions with mean m_1 and second moment $m_2 = m_1^2(c^2+1)$ with support in $[0, m_1 M]$. The set $\mathcal{P}_{2,2}(M)$ is a one-dimensional parametric family. Any element is determined by specifying one mass point. Let $F_b^{(2)}$ have probability mass $c^2/(c^2+(b-1)^2)$ on $m_1 b$, and mass $(b-1)^2/(c^2+(b-1)^2)$ on $m_1(1-c^2/(b-1))$ for $1+c^2 \leq b \leq M$. The cases $b = 1+c^2$ and $b = M$ constitute the two extremal distributions.

For $GI/M/1$, the interarrival-time cdf achieving the UB with mean m_1 and second moment $m_2 = m_1^2(c_a^2+1)$ with support in $[0, m_1 M_a]$, referred to here as $F^{(2)}_{1+c_a^2}$, arises for $b = 1 + c_a^2$. In particular, $F^{(2)}_{1+c_a^2}$ has probability mass $c_a^2/(1+c_a^2)$ on 0 and probability mass $1/(c_a^2+1)$ on $(m_2/m_1) = m_1(c_a^2+1)$.

The corresponding LB interarrival-time cdf, referred to here as $F^{(2)}_{M_a}$, arises for $b = M_a$. In particular, $F^{(2)}_{M_a}$ has probability mass $c_a^2/(c_a^2 + (M_a-1)^2)$ on the upper bound of the support, $m_1 M_a$, and mass $(M_a-1)^2/(c_a^2 + (M_a-1)^2)$ on $m_1(1 - c_a^2/(M_a-1))$. (For the interarrival time, we scale, i.e., choose measuring units for time, so that $m_1 = 1$.) We use the notation $G^{(2)}_{1+c_s^2}$ and $G^{(2)}_{M_s}$ for the corresponding service-time cdf's G with mean ρ and support $[0, \rho M_s]$.

Since the range of possible values is quite large, while the distributions that attain the bounds are unusual (two-point distributions), the papers Klincewicz and Whitt [1984], Whitt [1984b] and Johnson and Taaffe [1990a] focused on reducing the range by imposing shape constraints. In this paper we do not consider shape constraints.

1.4 Related Literature

The literature on bounds for the $GI/GI/1$ queue is well reviewed in Daley et al. [1992] and Wolff and Wang [2003], so we will be brief. The use of optimization to study the bounding problem for queues seems to have begun with Klincewicz and Whitt [1984] and Johnson and Taaffe [1990b]. Bertsimas and Natarajan [2007] provides a tractable semi-definite program as a relaxation model for solving steady-state waiting time of $GI/GI/c$ to derive bounds, while Osogami and Raymond [2013] bounds the transient tail probability of $GI/GI/1$ by a semi-definite program.

Several researchers have studied bounds for the more complex many-server queue. In addition to Bertsimas and Natarajan [2007], Gupta et al. [2010] and Gupta and Osogami [2011] investigate the bounds and approximations of the $M/GI/c$ queue. Gupta et al. [2010] explains why two moment information is insufficient for good accuracy of steady-state approximations of $M/GI/c$. Gupta and Osogami [2011] establishes a tight bound for the $M/GI/K$ in light traffic. Finally, Li and Goldberg [2017] establishes bounds for $GI/GI/c$ intended for the many-server heavy-traffic regime.

2 Three Papers from the 1990's

Here are three papers from the 1990's: Browne and Whitt [1996], Glynn and Whitt [1991] Massey and Whitt [1997].

References

d

S. Asmussen. *Applied Probability and Queues.* Springer, New York, second edition, 2003.

D. Bertsimas and K. Natarajan. A semidefinite optimization approach to the steady-state analysis of queueing systems. *Queueing Systems*, 56:27–39, 2007.

S. Browne and W. Whitt. Portfolio choice and the bayesian kelly criterion. *Journal of Applied Probability*, 28: 1145–1176, 1996.

Y. Chen and W. Whitt. Extremal $GI/GI/1$ queues given two moments. submitted for publication, Columbia University, http://www.columbia.edu/∼ww2040/allpapers.html, 2018a.

Y. Chen and W. Whitt. Algorithms for the upper bound mean waiting time in the $GI/GI/1$ extremal queue. submitted for publication, Columbia University, http://www.columbia.edu/∼ww2040/allpapers.html, 2018b.

Y. Chen and W. Whitt. Set-valued queueing approximations given partial information. submitted for publication, Columbia University, http://www.columbia.edu/∼ww2040/allpapers.html, 2019.

D. J. Daley. Inequalities for moments of tails of random variables, with queueing applications. *Zeitschrift fur Wahrscheinlichkeitsetheorie Verw. Gebiete*, 41:139–143, 1977.

D. J. Daley, A. Ya. Kreinin, and C.D. Trengove. Inequalities concerning the waiting-time in single-server queues: a survey. In U. N. Bhat and I. V. Basawa, editors, *Queueing and Related Models*, pages 177–223. Clarendon Press, 1992.

A. E. Eckberg. Sharp bounds on Laplace-Stieltjes transforms, with applications to various queueing problems. *Mathematics of Operations Research*, 2(2):135–142, 1977.

P. W. Glynn and W. Whitt. A new view of the heavy-traffic limit for infinite-server queues. *Advances in Applied Probability*, 23(1):188–209, 1991.

V. Gupta and T. Osogami. On Markov-Krein characterization of the mean waiting time in $M/G/K$ and other queueing systems. *Queueing Systems*, 68:339–352, 2011.

V. Gupta, J. Dai, M. Harchol-Balter, and B. Zwart. On the inapproximability of $M/G/K$: why two moments of job size distribution are not enough. *Queueing Systems*, 64:5–48, 2010.

J. M. Holtzman. The accuracy of the equivalent random method with renewal inouts. *Bell System Techn ical Journal*, 52(9):1673–1679, 1973.

M. A. Johnson and M. R. Taaffe. Matching moments to phase distributions: Density function shapes. *Stochastic Models*, 6(2):283–306, 1990a.

M. A. Johnson and M.R. Taaffe. Matching moments to phase distributions: nonlinear programming approaches. *Stochastic Models*, 6(2):259–281, 1990b.

S. Karlin and W. J. Studden. *Tchebycheff Systems; With Applications in Analysis and Statistics*, volume 137. Wiley, New York, 1966.

J. F. C. Kingman. The single server queue in heavy traffic. *Proc. Camb. Phil. Soc.*, 77:902–904, 1961.

J. F. C. Kingman. Inequalities for the queue $GI/G/1$. *Biometrika*, 49(3/4):315–324, 1962.

J. G. Klincewicz and W. Whitt. On approximations for queues, II: Shape constraints. *AT&T Bell Laboratories Technical Journal*, 63(1):139–161, 1984.

Y. Li and D. A. Goldberg. Simple and explicit bounds for multii-server queues with universal $1/(1-\rho)$ and better scaling. arXiv:1706.04628v1, 2017.

W. A. Massey and W. Whitt. Uniform acceleration expansions for Markov chains with time-varying rates. *Annals of Applied Probability*, 9(4):1130–1155, 1997.

T. Osogami and R. Raymond. Analysis of transient queues with semidefinite optimization. *Queueing Systems*, 73:195–234, 2013.

T. J. Ott. Simple inequalties for the $D/G/1$ queue. *Opperations Research*, 35(4):589–597, 1987.

T. Rolski. Some inequalities for $GI/M/n$ queues. *Zast. Mat.*, 13(1):43–47, 1972.

D. Stoyan. *Comparison Methods for Queues and Other Stochastic Models*. John Wiley and Sons, New York, 1983. Translated and edited from 1977 German Edition by D. J. Daley.

D. Stoyan and H. Stoyan. Inequalities for the mean waiting time in single-line queueing systems. *Engineering Cybernetics*, 12(6):79–81, 1974.

W. Whitt. The queueing network analyzer. *Bell Laboratories Technical Journal*, 62(9):2779–2815, 1983.

W. Whitt. On approximations for queues, I: Extremal distributions. *AT&T Bell Laboratories Technical Journal*, 63(1):115–137, 1984a.

W. Whitt. On approximations for queues, III: Mixtures of exponential distributions. *AT&T Bell Laboratories Technical Journal*, 63(1):163–175, 1984b.

R. W. Wolff and C. Wang. Idle period approximations and bounds for the $GI/G/1$ queue. *Advances in Applied Probability*, 35(3):773–792, 2003.