Queueing Systems manuscript No.
(will be inserted by the editor)

How to measure the accuracy of the subexponential approximation for the stationary single server queue

Dmitry Korshunov

Received: date / Accepted: date

Abstract We discuss the problem of establishing an upper bound for the distribution tail of the stationary waiting time D in the $GI/GI/1$ FCFS queue.

Keywords FCFS single server queue · Stationary waiting time · Heavy tails · Large deviations · Long tailed distribution · Subexponential distribution · Integrated tail distribution · Accuracy of approximation · Lower and Upper Bounds

Mathematics Subject Classification (2000) MSC 60K25 · MSC 60F10

1 Introduction

We consider the *first-come-first-served* single server queue system. Let τ be a typical interarrival time and σ be a typical service time. Sequences of independent identically distributed interarrival times $\{\tau_n\}$ with mean $a = \mathbb{E}\tau$ and that of service times $\{\sigma_n\}$ with mean $b = \mathbb{E}\sigma$ are assumed to be mutually independent. We also assume throughout these notes that the distribution of σ has unbounded support, i.e. $B(x) := \mathbb{P}\{\sigma \le x\} < 1$ for all x.

The customers form a single queue in front of the server, and the first customer in the queue moves immediately to a server when it becomes idle.

For $n = 1, 2, \ldots$, let D_n be the waiting time, or the delay which the nth customer experiences upon its arrival into the system. The waiting times D_n satisfy the Lindley recursion [13]: $D_1 = 0$ and

$$D_{n+1} = (D_n + \sigma_n - \tau_{n+1})^+.$$

Recall that D_{n+1} coincides in distribution with $\max(S_k, k \le n)$ where $S_0 = 0$, $S_{n+1} = S_n + \xi_{n+1}$, $\xi_k := \sigma_k - \tau_{k+1}$.

The system is assumed to be *stable*, i.e. $\rho \equiv b/a < 1$. Under this assumption, there exists a unique distribution of the stationary waiting time (delay) D, and the distribution of D_n converges to that of D in the total variation norm as $n \to \infty$.

The stationary delay D in the single server queue is investigated in great details. Let B_r be the *residual distribution*, that is, the distribution on $[0,\infty)$ with density $b^{-1}\overline{B}(x)$. In

D. Korshunov
Sobolev Institute of Mathematics, Novosibirsk 630090, Russia
E-mail: korshunov@math.nsc.ru

particular, it is well known (see, for example, [15,17,2,6]) that if the *residual distribution* B_r is subexponential then D is related to the service time distribution tail $\overline{B}(x) = \mathbb{P}\{\sigma > x\}$ via the equivalence

$$\mathbb{P}\{D > x\} \sim \frac{\rho}{1-\rho}\overline{B}_r(x) \quad \text{as } x \to \infty. \tag{1}$$

Recall that a distribution G on $\mathbb{R}^+$ is *subexponential* if $\overline{G*G}(x) \sim 2\overline{G}(x)$ as $x \to \infty$.

Also, it is known from [12] that the asymptotics (1) implies subexponentiality of B_r; earlier it was proved for the case of exponentially distributed τ in [15] and in [5, Corollary 6.1] in the context of risk theory.

Both from theoretical and practical point of view, it is a question of real interest to know how accurate the asymptotics (1) are. Essentially there are two ways to approach this problem. The fist one is to find the next term (or several terms) of the asymptotical expansion like

$$\mathbb{P}\{D > x\} = \frac{\rho}{1-\rho}\overline{B}_r(x) + c\overline{B}(x) + o(\overline{B}(x)) \quad \text{as } x \to \infty,$$

with an appropriate constant $c > 0$, under some additional assumptions on the distribution B (see, e.g. [3,18,4]). Such asymptotical expansions give an idea how accurate the approximation in (1) may be for very large x. Nevertheless some authors (see, e.g. [1,16]) show that the relative error in (1) can be big even for rather large values of x.

The second way is to derive explicit bounds for the difference of the left and right sides of (1). In this way it is desirable to establish bound similar to the second term in the corresponding expansion. And this is the main problem in the area under discussion. The perspective solution is likely to be in terms, in particular, of the rate of convergence in the equivalence $\overline{B_r * B_r}(x) \sim 2\overline{B}_r(x)$ as $x \to \infty$.

There are two special cases where the distribution of D may be calculated as an explicit geometric sum generated by B_r. If τ is exponentially distributed with parameter a^{-1} then (see [2, Section VIII.5])

$$\mathbb{P}\{D > x\} = (1-\rho)\sum_{n=0}^{\infty} \rho^n \overline{B_r^{*n}}(x). \tag{2}$$

Another case is just a lattice analog of the first one. If σ and τ take values on $\mathbb{Z}$ and τ is geometrically distributed, that is, $\mathbb{P}\{\tau = n\} = q^{n-1}(1-q)$, $n = 1, 2, \ldots$, then

$$\mathbb{P}\{D \geq N\} = (1-\rho_0)\sum_{n=0}^{\infty} \rho_0^n B_0^{*n}[N,\infty),$$

where the lattice distribution B_0 is defined by $B_0\{N\} := (b-1)^{-1}B[N+1,\infty)$, $N = 1, 2, \ldots$, $\rho_0 = \frac{b-1}{a-1}$. In both exponential and geometric cases some technique developed for geometric sums may be applied. Such an approach was developed for regular varying and Weibull distributions B in [7–10,16]. For general τ, this approach is even less effective because of necessity for explicit estimates for renewal measure generated by the descending ladder hight.

Note that the main difficulty in estimating accuracy in (1) consists in establishing of upper bound, which gives us the worst case scenario, while the lower bound is quite elementary. Indeed, if τ is exponential then the representation (2) together with inequality

$B_r^{*n}(x) \le B_r^n(x)$ implies that (see [9, Theorem 7])

$$\begin{aligned}\mathbb{P}\{D > x\} &\ge (1-\rho)\sum_{n=0}^{\infty}\rho^n(1-B_r^n(x))\\ &= 1-\frac{1-\rho}{1-\rho+\rho\overline{B}_r(x)} = \frac{\overline{B}_r(x)}{\frac{1-\rho}{\rho}+\overline{B}_r(x)},\end{aligned}$$

for a different proof see [11, Theorem 8.7.2].

In the general case the lower bound is slightly different and the proof is more complicated but still quite elementary via an equilibrium identity.

Proposition 1 *If* $\rho < 1$ *then, for any* $x \ge 0$*,*

$$\mathbb{P}\{D > x\} \ge \frac{\mathbb{E}\overline{B}_r(x+\tau)}{\frac{1-\rho}{\rho}+\mathbb{E}\overline{B}_r(x+\tau)}.$$

Proof Let $\xi := \sigma - \tau$ be independent of D. Since D has the stationary distribution, D is equal in distribution to $(D+\xi)^+$. Now fix $x \ge 0$. For $z > 0$ consider the function

$$L_z(y) = \begin{cases} x & \text{if } y \le x,\\ y & \text{if } y \in (x, x+z],\\ x+z & \text{if } y > x+z.\end{cases}$$

Since this function is bounded, $\mathbb{E}L_z(D)$ is finite and $\mathbb{E}L_z(D) = \mathbb{E}L_z(D+\xi)$. Therefore,

$$\mathbb{E}(L_z(D+\xi)-L_z(D)) = 0.$$

We have $|L_z(D+\xi)-L_z(D)| \le |\xi|$ for all z and $L_z(D+\xi)-L_z(D) \to L(D+\xi)-L(D)$ as $z \to \infty$ where

$$L(y) = \begin{cases} x & \text{if } y \le x,\\ y & \text{if } y > x.\end{cases}$$

Hence, by the dominated convergence we obtain the equality

$$\mathbb{E}(L(D+\xi)-L(D)) = 0. \tag{3}$$

We make use of the following bounds. For $y \in [0,x]$,

$$L(y+\xi)-L(y) = (y+\xi-x)\mathbb{I}\{y+\xi > x\} \ge (\xi-x)\mathbb{I}\{\xi > x\},$$

and so

$$\mathbb{E}\{L(D+\xi)-L(D); D \le x\} \ge \mathbb{E}\{\xi-x; \xi > x\}\mathbb{P}\{D \le x\}. \tag{4}$$

For $y > x$,

$$L(y+\xi)-L(y) \ge \xi,$$

and so

$$\mathbb{E}\{L(D+\xi)-L(D); D > x\} \ge \mathbb{E}\xi\mathbb{P}\{D > x\}. \tag{5}$$

Substituting (4) and (5) into (3) we get the inequality

$$\mathbb{E}\{\xi-x; \xi > x\}\mathbb{P}\{D \le x\} \le -\mathbb{E}\xi\mathbb{P}\{D > x\}.$$

Therefore,

$$\mathbb{P}\{D>x\} \geq \frac{\mathbb{E}\{\xi-x;\xi>x\}}{a-b+\mathbb{E}\{\xi-x;\xi>x\}}.$$

Since

$$\begin{aligned}\mathbb{E}\{\xi-x;\xi>x\} &= \mathbb{E}(\xi-x)^+ \\ &= \int_0^\infty \mathbb{P}\{\xi-x>y\}dy = \int_0^\infty \mathbb{P}\{\sigma>x+\tau+y\}dy,\end{aligned}$$

conditioning on τ leads to the equality $\mathbb{E}\{\xi-x;\xi>x\} = b\mathbb{E}B_r(x+\tau)$, so that the required lower bound for $\mathbb{P}\{D>x\}$ follows.

To conclude the discussion on this open problem we mention the possibility of obtaining upper bounds via Lyapunov (test) functions. Such types of bounds are never tight in the case of heavy tails but they are easily computable in explicit way. Let, for example, σ have βth moment finite, $m_\beta := \mathbb{E}\sigma^\beta < \infty$, $\beta \geq 2$. Consider the test function $V(x) = x^\beta \mathbb{I}\{x>0\}$ and calculate the drift of V at point x. Take x_1 such that $\mathbb{E}\max(\sigma-\tau,-x_1) \leq 2(b-a)/3 < 0$. Then, for $x \geq x_1$, we have

$$\begin{aligned}\mathbb{E}V(x+\sigma-\tau)-V(x) &= L'(x)\mathbb{E}\max(\sigma-\tau,-x)+\mathbb{E}V''(x+\theta(\sigma-\tau)(\sigma-\tau)^2/2 \\ &\leq -2V'(x)(a-b)/3+\mathbb{E}V''(x+\sigma)(\sigma-\tau)^2/2 \\ &\leq -2V'(x)(a-b)/3+2^{\beta-2}(V''(x)\mathbb{E}(\sigma-\tau)^2+\mathbb{E}V''(\sigma)(\sigma-\tau)^2)/2,\end{aligned}$$

here we take use of the inequality $V''(x+y) \leq 2^{\beta-2}(V''(x)+V''(y))$. Take $x_2 := 2^{\beta-1}(\beta-1)\mathbb{E}(\sigma-\tau)^2/(a-b)$. Then, for all $x \geq x_2$,

$$\mathbb{E}V(x+\sigma-\tau)-V(x) \leq -V'(x)(a-b)/3+c_1,$$

where $c_1 := 2^{\beta-2}\beta(\beta-1)\mathbb{E}\sigma^{\beta-2}(\sigma-\tau)^2/2$. Put $x_3 := \max(x_1,x_2)$. For all $x \in [0,x_3]$,

$$\begin{aligned}\mathbb{E}V(x+\sigma-\tau)-V(x) &\leq \mathbb{E}V(x+\sigma)-V(x) \\ &\leq \mathbb{E}V'(x_3+\sigma)\sigma =: c_2.\end{aligned}$$

Combining this altogether we deduce, for all x,

$$\mathbb{E}V(x+\sigma-\tau) \leq V(x)-V'(x)(a-b)/3+c_3,$$

where $c_3 := \max(c_1,c_2)+V'(x_3)(a-b)/3$. Therefore, as follows from [14, Theorem 14.3.7], $\mathbb{E}V'(D) \leq 3c_3/(a-b)$, that is,

$$\mathbb{E}D^{\beta-1} \leq 3c_3/(a-b)\beta.$$

Now if we consider service time with Pareto type distribution, for example $\mathbb{P}\{\sigma>x\} = x^{-\alpha}$, $x \geq 1$, $\alpha > 1$, then we may take $\beta = \alpha-\varepsilon$, $\varepsilon > 0$. In this case $\mathbb{E}\sigma^\beta = O(1/\varepsilon)$ as $\varepsilon \downarrow 0$ and $c_3 = O(1/\varepsilon)$, so that $\mathbb{E}D^{\beta-1} \leq c/\varepsilon$ with explicit constant c. Hence, the Chebyshev inequality implies the following upper bound,

$$\mathbb{P}\{D>x\} \leq \frac{c}{\varepsilon x^{\beta-1}} = \frac{c}{\varepsilon x^{\alpha-1-\varepsilon}},$$

which is heavier than the actual asymptotics

$$\mathbb{P}\{D>x\} \sim \frac{\alpha-1}{(a-b)x^{\alpha-1}},$$

but may be still useful.

Similar calculations can be carried out for the moments of order e^{x^β}, $\beta \in (0,1)$, with further application to Weibull type distributions.

References

1. Abate, J., Choudhury, G. L., Whitt, W.: Waiting-time tail probabilities in queues with long-tail service-time distributions. Queueing Syst. **16**, 311–338 (1994)
2. Asmussen, S.: Applied Probability and Queues, 2nd edn. Springer, New York (2003)
3. Baltrunas, A.: Second-order asymptotics for the ruin probability in the case of very large claims. Sib. Math. J. **40**, 1034–1043 (1999)
4. Borovkov, A. A., Borovkov, K. A.: Asymptotic Analysis of Random Walks. Heavy-Tailed Distributions. Cambridge University Press, Cambridge (2008)
5. Embrechts, P., Veraverbeke, N.: Estimates for the probability of ruin with special emphasis on the possibility of large claims. Insur. Math. Econ. **1**, 55–72 (1982)
6. Foss, S., Korshunov, D., and Zachary, S.: An Introduction to Heavy-Tailed and Subexponential Distributions. Springer, Berlin (2011)
7. Kalashnikov, V.: Geometric Sums: Bounds for Rare Events with Applications. Kluwer Academic, Dordrecht (1997)
8. Kalashnikov, V.: Bounds for ruin probabilities in the presence of large claims and their comparison. Insur. Math. Econ. **20**, 146–147 (1997)
9. Kalashnikov, V., Tsitsiashvili, G.: Tails of waiting times and their bounds. Queueing Syst. **32**, 257–283 (1999)
10. Kalashnikov, V., Tsitsiashvili, G.: Asymptotically correct bounds of geometric convolutions with subexponential components. J. Math. Sci. **106**, 2806–2819 (2001)
11. Korolev, V. Yu., Bening, V. E., Shorgin, S. Ya.: Mathematical Foundations of Risk Theory. Fizmatlit, Moscow (2007) (in Russian)
12. Korshunov, D. A.: On distribution tail of the maximum of a random walk. Stoch. Process. Appl. **72**, 97–103 (1997)
13. Lindley, D. V.: The theory of queues with a single server. Proc. Camb. Philos. Soc. **48**, 277–289 (1952)
14. Meyn, S. P., Tweedie, R. L.: Markov chains and stochastic stability. Springer, Berlin (1993)
15. Pakes, A. G.: On the tails of waiting-time distribution. J. Appl. Probab. **12**, 555–564 (1975)
16. Richards, A.: On upper bounds for the tail distribution of geometric sums of subexponential random variables. Queueing Syst. **62**, 229–242 (2009)
17. Veraverbeke, N.: Asymptotic behavior of Wiener-Hopf factors of a random walk. Stoch. Process. Appl. **5**, 27–37 (1977)
18. Willekens, E., Teugels, J. L.: Asymptotic expansions for waiting time probabilities in an $M/G/1$ queue with long-tailed service time. Queueing Syst. **10**, 295–313 (1992)