BRAVO FOR MANY-SERVER QED SYSTEMS WITH FINITE BUFFERS

D. J. DALEY,* *The University of Melbourne*

JOHAN S. H. VAN LEEUWAARDEN,** *Eindhoven University of Technology*

YONI NAZARATHY,*** *The University of Queensland*

Abstract

This paper demonstrates the occurrence of the feature called BRAVO (balancing reduces asymptotic variance of output) for the departure process of a finite-buffer Markovian many-server system in the QED (quality and efficiency-driven) heavy-traffic regime. The results are based on evaluating the limit of an equation for the asymptotic variance of death counts in finite birth–death processes.

Keywords: Queueing system critical point; stochastic process limit; Poisson distribution asymptotics; stationary Poisson queue asymptotics; Halfin–Whitt regime; queueing output variance asymptotics; birth–death asymptotic variance

2010 Mathematics Subject Classification: Primary 60J27
Secondary 60K25

1. Introduction

The QED (quality and efficiency-driven) regime for many-server systems combines large capacity with high utilization while maintaining satisfactory system performance. In the QED regime the arrival rate λ and the number of servers s are scaled in such a way that while they both increase towards ∞, the traffic intensity $\rho = \rho_s = \lambda/\mu$ (assuming service rate μ/s per server) approaches 1 and

$$(1 - \rho_s)\sqrt{s} \to \beta, \quad \beta \in (-\infty, \infty). \tag{1.1}$$

Halfin and Whitt [11] introduced the QED regime for the GI/M/s system. Under the scaling (1.1), assuming $\beta > 0$, the stationary probability of delay was shown to converge to a non-degenerate limit, bounded away from both zero and one. Limit theorems for other, more general systems were obtained in [10], [13]–[21], and for all those cases, the limiting probability of delay remains in the interval (0, 1). In fact, not only the probability of delay, but several other performance characteristics or objective functions are shown to behave (near) optimally in the QED regime (see [3], [9]).

Associated with the near optimal behaviour is the fact that the process $Q_s(\cdot)$ that counts the total number of customers in the system at any time exhibits relatively small fluctuations. Halfin and Whitt [11] showed that for the GI/M/s system under (1.1) a sequence of standardised

processes $X_s(\cdot)$, with $X_s(t) := (Q_s(t) - s)/\sqrt{s}$, converges as $s \to \infty$ to a diffusion process $X(\cdot)$. This diffusion process behaves like a Brownian motion with drift above zero and like an Ornstein–Uhlenbeck process below zero. The interpretation of this scaled process is as follows: in the case, $X_s(t) > 0$, it represents the scaled number of customers waiting for service, whereas in the case $X_s(t) < 0$, it represents the scaled number of idle servers. This result shows that the natural scale that emerges is of the order $\sqrt{s}$: specifically, both the queue length and the number of idle servers in the system are of the order $\sqrt{s}$.

Due to its favourable behaviour, the many-server QED regime has been a major focal point of applied probability and stochastic operations research for the past 30 years. The many extensions of the Halfin–Whitt exposition in [11] have led to theoretical advances in the areas of stochastic-process limits and asymptotic dimensioning. From an operational point of view, the QED regime has found many applications in the planning, analysis and optimisation of queueing, inventory and service systems (see, for example, [2]).

In this paper we explore the presence of a BRAVO effect in the QED regime. BRAVO is short for balancing reduces asymptotic variance of outputs. This again would be a favourable property of QED, this time for the departure process $N_{\text{dep}}(0, \cdot]$, where $N_{\text{dep}}(0, t]$ counts the number of serviced customers during the time interval $(0, t]$. The study of departure or output processes of queues has a long tradition, see, for example, the classic surveys [5] and [7], yet it is only recently that the surprising phenomenon of the BRAVO effect has been reported. The BRAVO effect is captured in terms of the asymptotic ratio of the variance and the mean of the departure process

$$\mathcal{D} := \lim_{t \to \infty} \frac{\text{var}(N_{\text{dep}}(0, t])}{\text{E}(N_{\text{dep}}(0, t])}.$$

For Poisson processes $\mathcal{D} = 1$, and more generally for renewal processes, $\mathcal{D}$ equals the ratio of the variance of the renewal lifetime and the square of the mean lifetime. Thus, it is initially surprising that for M/M/1/K systems, $\mathcal{D}$ is minimised when the arrival rate λ is equal to the service rate μ with a minimum equal to $\frac{2}{3} + o_K(1)$, where $o_K(1)$ is a term that vanishes as $K \to \infty$. This was shown in [19]. Further, when $K = \infty$, it is well known that $\mathcal{D} = 1$ whenever $\lambda \neq \mu$, yet it was shown in [1] that in the critical case that $\lambda = \mu$, $\mathcal{D} = 2(1 - 2/\pi) \neq \frac{2}{3}$. The work in [1] goes further, generalising this M/M/1 result to GI/GI/1 systems and even multi-server GI/GI/s systems with a finite bounded number of servers s. Hence, by BRAVO we mean that $\mathcal{D} < 1$ when $\rho = 1$. For overviews of BRAVO results we refer the reader to [6] and [18].

In this paper we study the BRAVO effect in the QED regime by equipping the many-server M/M/s system with a finite waiting capacity K. In order to create a finite-capacity effect in the QED regime that is neither dominant nor negligible, it is plausible to assume that K is approximately equal to $\eta \sqrt{s}$, because the natural scale of the queue length is $\sqrt{s}$. More precisely, we study a sequence of systems in which both K and s grow in such a way that

$$\frac{K}{\sqrt{s}} \to \eta \qquad\qquad (1.2)$$

for some positive η. A similar threshold where K is approximately equal to $\eta \sqrt{s}$ in the context of many-server systems in the QED regime has been considered in [2], [12], [17], [23], [24]. Hence, in addition to the parameter β in (1.1) describing the scaled shortfall from one in the system capacity, our system includes a parameter η describing the relative buffer size.

Our result on the BRAVO effect in the QED regime is in terms of

$$\mathcal{D}_{\beta,\eta} := \lim_{s,K\to\infty} \lim_{t\to\infty} \frac{\mathrm{var}(N_{\mathrm{dep}}(0,t])}{\mathrm{E}(N_{\mathrm{dep}}(0,t])}, \tag{1.3}$$

where the outer limit is taken under the constraints at (1.1) and (1.2).

Our analysis leads to explicit representations for $\mathcal{D}_{\beta,\eta}$ and is based on the following general result. Consider a finite, irreducible, birth–death process $Q(\cdot)$ on $\{0,1,\dots,J\}$ with birth rates $\lambda_0, \lambda_1, \dots, \lambda_{J-1}$, and death rates $\mu_1, \mu_2, \dots, \mu_J$. Let $\{\pi_i\}$ denote the stationary distribution for $Q(\cdot)$ with cumulative distribution $P_i = \sum_{j=0}^{i} \pi_j$, and let $N_{\mathrm{dep}}(0,t]$ denote the number of deaths in $(0,t]$. Denote the departure rate by

$$\lambda^* := \lim_{t\to\infty} \frac{\mathrm{E}(N_{\mathrm{dep}}(0,t])}{t} = \sum_{j=1}^{J} \mu_j \pi_j,$$

and write

$$\Lambda_i^* = \frac{\sum_{j=1}^{i} \mu_j \pi_j}{\lambda^*}$$

for what are cumulative probabilities. Then, we know from [19, Theorem 1] that

$$\mathcal{D}_\pi := \lim_{t\to\infty} \frac{\mathrm{var}(N_{\mathrm{dep}}(0,t])}{\mathrm{E}(N_{\mathrm{dep}}(0,t])} = 1 - 2 \sum_{i=0}^{J} (P_i - \Lambda_i^*)\left(1 - \frac{\lambda^*}{\pi_i \lambda_i}(P_i - \Lambda_i^*)\right). \tag{1.4}$$

Note that (1.4) is in a slightly different form from that appearing in [19]; the translation between the two forms is immediate.

In this paper we use (1.4) so as to obtain explicit expressions for $\mathcal{D}_{\beta,\eta}$ in the case of M/M/s/K systems. To do so we use the fact that in birth–death processes with $\lambda_i \equiv \lambda$, we have

$$\lambda^* = \lambda(1 - \pi_J), \qquad \Lambda_i^* = \frac{P_{i-1}}{(1 - \pi_J)}, \qquad P_i - \Lambda_i^* = \frac{\pi_i - \pi_J P_i}{1 - \pi_J},$$

and after basic manipulation, (1.4) can be represented as

$$\mathcal{D}_\pi = 1 - 2 \frac{\pi_J}{1 - \pi_J} \sum_{i=0}^{J} P_i\left(1 - \pi_J \frac{P_i}{\pi_i}\right). \tag{1.5}$$

In the case of the M/M/s/K system, this elegant form proves amenable to manipulation and asymptotics under the QED regime, yielding our desired explicit equations for $\mathcal{D}_{\beta,\eta}$ defined in (1.3). Carrying out these asymptotics is the main contribution of the current paper. A further virtue of the form (1.5), which is of independent interest, is that it demonstrates that (for the case $\lambda_i = \lambda$),

$$\mathcal{D}_\pi \geq \frac{1/2 - \pi_J}{1 - \pi_J}. \tag{1.6}$$

To see this, observe that the function $x \mapsto x(1-x)$ has $\frac{1}{4}$ as its maximum value so that

$$\frac{1}{2}(1 - \mathcal{D}_\pi)(1 - \pi_J) = \sum_{i=0}^{J} \pi_i \frac{\pi_J P_i}{\pi_i}\left(1 - \frac{\pi_J P_i}{\pi_i}\right) \leq \frac{1}{4}.$$

The lower bound (1.6) implies that as long as $\pi_J \to 0$, as in the M/M/s/K system, $\mathcal{D}_{\beta,\eta} \geq \frac{1}{2}$. Our explicit expressions for $\mathcal{D}_{\beta,\eta}$ for $\rho \equiv 1$ in fact establish that $\mathcal{D}_{0,\eta}$ is in the range $(0.6, \frac{2}{3})$ with the exact value depending on η. Hence, for QED systems the magnitude of the BRAVO effect is not exactly the same as for single server systems but it is in a similar range. Similar results are found for QED systems with nonzero β for which $|\beta|$ is not too large.

The remainder of this paper is structured as follows. In Section 2 we give our main theorem on the BRAVO effect for the M/M/s/K system, which presents an expression for the asymptotic ratio $\mathcal{D}_{\beta,\eta}$ for both the case $\rho \equiv 1$ (i.e. $\beta = 0$) and the case $\beta \neq 0$ (i.e. $\rho \approx 1 - \beta/\sqrt{s}$). The proofs for these two cases are presented in Sections 3 and 4, respectively. We present some conclusions and ideas for future work in Section 5. Appendix A contains some asymptotic properties of Poisson probabilities that we employ.

2. Main result

We now state the main theorem of this paper, which identifies the BRAVO effect in many-server QED systems in terms of the asymptotic output ratio. Let Φ and ϕ denote the distribution and density of a standard normal random variable.

Theorem 2.1. *Let* $s, K \to \infty$ *in such a way that* $K/\sqrt{s} \to \eta$ *for some finite positive* η.

(a) *Let* $\rho = 1$. *Then*

$$\mathcal{D}_{0,\eta} = \tfrac{2}{3} - L(\eta),$$

where

$$L(\eta) = \frac{(2 - \pi/2)\eta + \sqrt{2\pi}(1 - \log 2 - \pi/12)}{(\eta + \sqrt{\pi/2})^3}.$$

(b) *Let* $\rho = 1 - \beta/\sqrt{s}$ *for finite* $\beta \neq 0$. *Then*

$$\mathcal{D}_{\beta,\eta} = 1 - \frac{2\beta^2 \mathrm{e}^{-\beta\eta} h^2(\eta, \beta)}{\phi(\beta)} f(\eta, \beta) + g(\eta, \beta), \tag{2.1}$$

where

$$h(\eta, \beta) = \frac{1}{1 - \mathrm{e}^{-\beta\eta} + \beta\Phi(\beta)/\phi(\beta)},$$

$$f(\eta, \beta) = \int_{-\beta}^{\infty} \left(1 - \beta\mathrm{e}^{-\beta\eta} h(\eta, \beta)\frac{\Phi(-u)}{\phi(u)}\right)\Phi(-u)\,\mathrm{d}u,$$

and, with $h = h(\eta, \beta)$,

$$g(\eta, \beta) = 2\mathrm{e}^{-\beta\eta} h(1 + \mathrm{e}^{-\beta\eta} h)(1 - \beta\eta - \mathrm{e}^{-\beta\eta} + (1 - 2\beta\eta\mathrm{e}^{-\beta\eta} - \mathrm{e}^{-2\beta\eta})h).$$

We leave it to the reader to check that $\mathcal{D}_{\beta,\eta}$ is a continuous function of β for all finite β including 0.

We start by discussing Theorem 2.1(a). In Figure 1 $\mathcal{D}_{0,\eta}$ is shown as a function of η. Observe that as $\eta \to \infty$ (and this includes the case of a fixed finite number of servers with a large, but finite buffer), $\mathcal{D}_{0,\eta} \to \frac{2}{3}$. Furthermore,

$$\lim_{\eta \to 0} \mathcal{D}_{0,\eta} = 1 - \frac{4(1 - \log 2)}{\pi} \approx 0.6093. \tag{2.2}$$
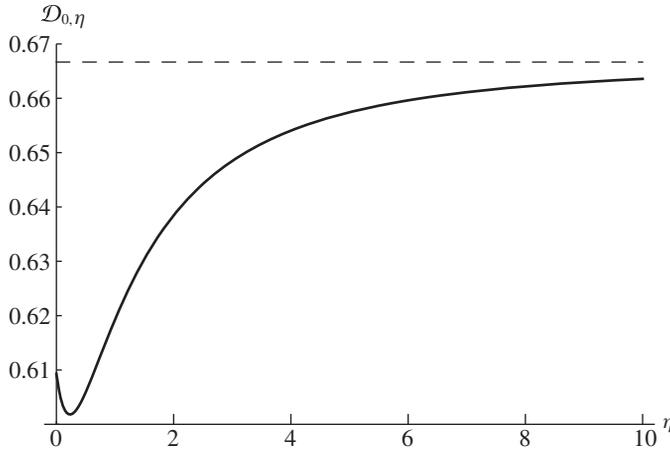
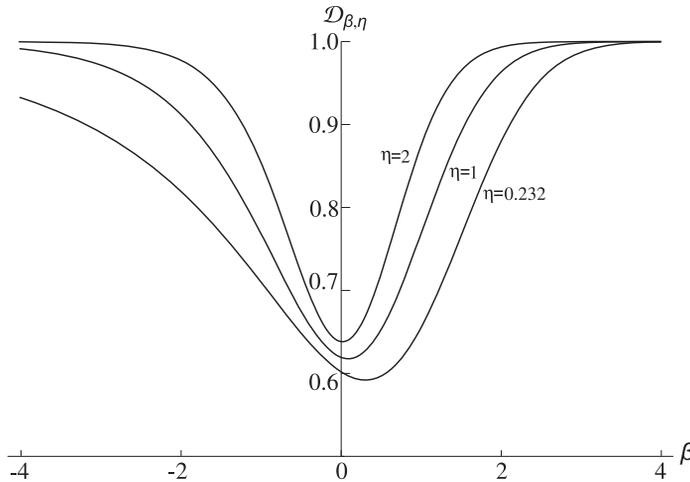FIGURE 1: The asymptotic value of $\mathcal{D}_{0,\eta}$ as a function of η when $\rho \equiv 1$.

FIGURE 2: The asymptotic value of $\mathcal{D}_{\beta,\eta}$ for various values of η.

It is also easy to verify that as a function of η, $\mathcal{D}_{0,\eta}$ has a unique global minimum at

$$\eta = \frac{\sqrt{2\pi}\,(\log 8 - 2)}{4 - \pi} \approx 0.232,$$

yielding $\inf_\eta \mathcal{D}_{0,\eta} \approx 0.6018$.

We now turn to Theorem 2.1(b). In Figure 2 $\mathcal{D}_{\beta,\eta}$ is shown as a function of β and η. The figure suggests that, for fixed η,

$$\lim_{\beta \to \infty} \mathcal{D}_{\beta,\eta} = \lim_{\beta \to -\infty} \mathcal{D}_{\beta,\eta} = 1,$$

and indeed this follows from the exact expression in (2.1). In fact, this can also be explained using the following heuristic reasoning. As $\beta \to \infty$, the system becomes lightly loaded, and

the process $Q_s(\cdot)$ behaves as an infinite-server system, which is reversible and therefore has an asymptotic output ratio equal to one. Also, as $\beta \to -\infty$, the system becomes increasingly overloaded, so that the process $Q_s(\cdot)$ behaves like a single-server system in which all servers work all the time. Such a single-server process is, again, a reversible birth–death process, which has an asymptotic output ratio equal to one. For any finite β, the behaviour of the process $Q_s(\cdot)$ resembles a mixture of an infinite-server system and a single-server system, and it is this alternation between two different stable systems that may give rise to the BRAVO effect, although we cannot give an intuitive argument for it. We do observe that the BRAVO effect is most pronounced for values of β close to zero.

We now present some further arguments about why the scaling relations in (1.1) and (1.2) are the precise scalings needed to obtain non-degenerate limiting behaviour, and as it turns out, also to establish the limiting BRAVO effect. We first show that the scaling relation in fact leads to a well-defined stochastic-process limit of the entire queue-length process. Then we establish, using Gaussian approximations for the Poisson distribution, non-degenerate limits for the stationary distribution.

Let '$\Rightarrow$' denote weak convergence in the space $D[0, \infty)$ endowed with the usual Skorokhod J_1 topology. The following result is proved in [20, Theorem 1.2].

Proposition 2.1. (Weak convergence to a diffusion process.) *Assume (1.1) and (1.2). If $X_s(0) \Rightarrow X(0) \in \mathbb{R}$, then as $s \to \infty$, $X_s(\cdot) \Rightarrow X(\cdot) \wedge \eta$, where the limit $X(\cdot)$ is the diffusion process with infinitesimal drift $-\beta$ when $x > 0$ and $-\beta - x$ when $x < 0$, and constant infinitesimal variance 2.*

With minor abuse of notation, let X, X_s and Q_s without time arguments denote stationary random variables. One of the signature features of the QED regime is that, due to economies of scale, the stationary probability of delay $\mathbb{P}(Q_s \geq s) = \sum_{i=s}^{K+s} \pi_i$ converges to a limit that is neither zero nor one. This feature continues to exist for our model with the finite-capacity scaling. For convenience we state it formally below, showing the relation of $h(\eta, \beta)$ to the probability of delay (the result appears at Equation (4.7) of the unpublished manuscript [17]).

Proposition 2.2. (Probability of delay in QED regime.) *Assume (1.1) and (1.2). Then*

$$\lim_{s \to \infty} \mathbb{P}(Q_s \geq s) = \begin{cases} \dfrac{\eta}{\sqrt{2/\pi} + \eta}, & \beta = 0, \\[3mm] \dfrac{1 - \exp(-\beta\eta)}{1 - \exp(-\beta\eta) + \beta\Phi(\beta)/\phi(\beta)} = (1 - e^{-\beta\eta})h(\eta, \beta), & \beta \neq 0. \end{cases} \quad (2.3)$$

It can be shown that the right-hand side of (2.3) corresponds to the probability $\mathbb{P}(X > 0)$ that the stationary diffusion process is positive (see [4]), as suggested by Proposition 2.1. We choose, however, to give a direct derivation, starting from the exact expression for $\mathbb{P}(Q_s \geq s)$. We do so in order to give insight into the crucial role played by Gaussian approximations of the Poisson distribution; these approximations constitute an important ingredient for the proof of the main theorem (see Appendix A).

Recall the representation for the stationary distribution $\{\pi_i\}$ of the number in the many-server queueing system M/M/s/K, arrival rate λ and service rate μ/s per server,

$$\pi_i = \begin{cases} \dfrac{(s\lambda/\mu)^i}{i!}\pi_0 & \text{for } i = 0, 1, \ldots, s, \\[3mm] (\lambda/\mu)^{i-s}\pi_s & \text{for } i = s, s+1, \ldots, s+K, \end{cases} := \begin{cases} \dfrac{(s\rho)^i}{i!}\pi_0, \\[3mm] \rho^{i-s}\pi_s. \end{cases}$$

For finite $\kappa > 0$, let $\{\varpi_i(\kappa) = e^{-\kappa}\kappa^i/i!, i = 0, 1, \ldots\}$ denote the Poisson distribution with mean κ. Observe that for $i = 0, \ldots, s$,

$$\pi_i = \pi_0 e^{s\rho}\varpi_i(s\rho) := b_s(\rho)\varpi_i(s\rho),$$

where $b_s(\rho) := \pi_0 e^{s\rho}$.

The stationary distribution $\{\pi_i\}$ has three parameters: ρ, K and s, but they must satisfy the constraint

$$1 = \sum_{i=0}^{s+K} \pi_i = \left(\sum_{i=0}^{s} + \sum_{i=s+1}^{s+K}\right)\pi_i.$$

Thus, when $\lambda = \mu$, i.e. $\rho = 1$, the latter sum equals $K\pi_s$, and the former sum equals

$$b_s(1)\sum_{i=0}^{s}\varpi_i(s) = b_s(1)\left(\frac{1}{2} + \frac{\psi_{ss}}{\sqrt{s}}\right)$$

by the central limit property for the Poisson distribution (see Lemma A.2) for constants ψ_{ss} satisfying $\sup_s |\psi_{ss}| < \infty$. Using Stirling's formula,

$$\frac{\pi_s}{\pi_0} = \frac{\varpi_s(s)}{\varpi_0(s)} = \frac{s^s}{s!} = \frac{s^s}{(s/e)^s\sqrt{2\pi s}}\left(1 + \frac{\vartheta_s}{12s}\right),$$

for some ϑ_s for which $|\vartheta_s| < 1$. So $b_s(1) = \pi_0 e^s = \pi_s\sqrt{2\pi s}/[1 + \vartheta_s/(12s)]$, and

$$1 = \pi_s\left[\left(\frac{1}{2} + \frac{\psi_{ss}}{\sqrt{s}}\right)\frac{\sqrt{2\pi s}}{1 + \vartheta_s/(12s)} + K\right].$$

Then

$$1 = \pi_s\sqrt{s}\left(\sqrt{\pi/2} + \frac{K}{\sqrt{s}} + o(1)\right),$$

hence, using (1.2),

$$\pi_s\sqrt{s} \to \frac{1}{\sqrt{\pi/2} + \eta}. \tag{2.4}$$

Also, $b_s(1) = \pi_s\sqrt{s}\sqrt{2\pi}/[1 + \vartheta_s/(12s)] \to \sqrt{2\pi}/(\sqrt{\pi/2} + \eta) =: b_\infty$ as $s \to \infty$.

Moving onto the case of $\beta \neq 0$ we have, $\rho_s = \lambda/\mu = 1 - \beta/\sqrt{s}$ and $s, K \to \infty$ as at (1.2). The sums of terms π_i over $i \leq s$ and $i > s$ now equal

$$b_s(\rho)\sum_{i=0}^{s}\varpi_i(s\rho) = b_s(\rho)\left[\Phi\left(\frac{s - s\rho}{\sqrt{s\rho}}\right) + O(s^{-1/2})\right]$$

and

$$\pi_s\sum_{i=1}^{K}\rho^i = \pi_s\frac{\rho(\rho^K - 1)}{\rho - 1} \tag{2.5}$$

respectively. Here we have used the central limit property of a Poisson distribution with mean $s\rho$ as in Lemma A.2. Substitution for ρ gives, correct to terms that are $O(1/\sqrt{s})$, $\pi_0 e^{s\rho}\Phi(\beta)$

and $\sqrt{s}\pi_s(1-\mathrm{e}^{-\beta\eta})/\beta$ when $s,\ K \to \infty$ as in (1.2). The local asymptotic normality of Poisson probabilities (cf. Lemma A.1) implies

$$
\begin{aligned}
b_s(\rho) &= \pi_s \frac{\pi_0}{\pi_s} \mathrm{e}^{s\rho}\\
&= \frac{\pi_s}{\mathrm{e}^{-s\rho}(s\rho)^s/s!}\\
&= \frac{\sqrt{s\rho}\pi_s}{\phi((s-s\rho)/\sqrt{s\rho})}(1 + O(s^{-1/2}))\\
&= \frac{\sqrt{s}\pi_s}{\phi(\beta)}(1 + O(s^{-1/2})).
\end{aligned}
$$

The terms on the left-hand sides of the relations in (2.5) add to 1, so that

$$
1 + o(1) = \sqrt{s}\pi_s \left(\frac{\Phi(\beta)}{\phi(\beta)} + \frac{1-\mathrm{e}^{-\beta\eta}}{\beta} \right) = \frac{\sqrt{s}\pi_s}{\beta h(\eta,\beta)} \tag{2.6}
$$

and $\sqrt{s}\pi_s \to \beta h(\eta,\beta)$. Also, $\lim_{s\to\infty} b_s(\rho) = \lim_{s\to\infty} \pi_0 \mathrm{e}^{s\rho} = \beta h(\eta,\beta)/\phi(\beta)$. Proposition 2.2 follows from combining (2.5) and (2.6).

3. Proof of Theorem 2.1: the case $\rho = 1$

Since $\pi_i = \pi_s$ for $i \geq s$, $P_i = 1 - (K+s-i)\pi_s$. Write (1.5), with $J = K+s$, as

$$
(1-\pi_J)(1-\mathcal{D}_\pi) = 2\pi_s \sum_{i=0}^{J} \frac{\pi_J}{\pi_s} P_i \left(1 - \frac{\pi_J}{\pi_i} P_i \right) = 2\pi_s \left(\sum_{i=0}^{s-1} + \sum_{i=s}^{s+K} \right)(\cdots), \tag{3.1}
$$

in which the last sum, for which $\pi_s = \pi_i = \pi_J$, equals

$$
2\pi_s \sum_{i'=0}^{K}(1 - [K-i']\pi_s)(K-i')\pi_s = K(K+1)\pi_s^2 - \frac{K(K+1)(2K+1)\pi_s^3}{3},
$$

$$
\begin{aligned}
&\to \frac{\eta^2}{(\sqrt{\pi/2}+\eta)^2} - \frac{2\eta^3/3}{(\sqrt{\pi/2}+\eta)^3}\\
&= \frac{\eta^2\sqrt{\pi/2}+\eta^3/3}{(\sqrt{\pi/2}+\eta)^3}
\end{aligned}
$$

when $s,\ K \to \infty$ as at (1.2), using also (2.4).

It thus remains to consider the sum

$$
S_s := 2 \sum_{i=0}^{s-1} \left(1 - \frac{\pi_s}{\pi_i} P_i \right) \pi_s P_i,
$$

which in terms of Poisson probabilities $\varpi_i := \varpi_i(s)$, their partial sums $\Pi_i(s) := \sum_{j=0}^{i} \varpi_j(s)$ and the multiplier $b_s := b_s(1)$ become

$$
S_s = 2b_s^2 \sum_{i=1}^{s} \left(1 - b_s \varpi_s \frac{\Pi_{s-i}(s)}{\varpi_{s-i}} \right) \varpi_s \Pi_{s-i}(s). \tag{3.2}
$$

Now both components of this sum for S_s have finite limits (see below), and $b_s \to b_\infty$ (see also (2.4)), finite and positive; we use this limit in examining the expression at (3.2).

Substitute from Lemmas A.1 and A.2 for the Poisson probabilities and assume for the moment that conditions for uniform convergence are met. Then

$$S_s \approx 2b_\infty^2 \sum_{i=1}^{s} (1 - b_\infty e^{(1/2)x_{si}^2} \Phi(-x_{si})) \frac{1}{\sqrt{2\pi}\sqrt{s}} \Phi(-x_{si}),$$

where $x_{si} = i/\sqrt{s}$. Recognise that this sum is an approximation to the Riemann integral

$$\frac{2b_\infty^2}{\sqrt{2\pi}} \int_0^{\sqrt{s}} [1 - b_\infty e^{(1/2)u^2} \Phi(-u)] \Phi(-u) \, du \qquad (3.3)$$

based on the s intervals determined by the $s + 1$ points $\{i/\sqrt{s}\}$, $i = 0, \ldots, s$. These integrals certainly converge as $s \to \infty$, because the (improper) Riemann integral is finite (its integrand is non-negative and dominated by $\Phi(-u)$ whose integral on $\mathbb{R}_+$ equals $1/\sqrt{2\pi}$).

To prove that the improper integral based on (3.3) is indeed equal to $\lim_{s \to \infty} S_s$, there are essentially two tasks: one relates to the finiteness of the limit of the finite sum, and the other to the convergence of the summands to the function involving the normal distribution function Φ and its density ϕ. These two tasks overlap in the sense that the part of the argument in the latter requiring uniform convergence follows from truncating the infinite sum. It is convenient to rewrite the sum for S_s as an integral, namely

$$S_s = \int_{\mathbb{R}_+} g_s(x) \nu_s(\mathrm{d}x), \qquad (3.4)$$

where for each positive integer s, ν_s is a purely atomic measure on $\mathbb{R}_+$ with support set $\{x_{si}\} := \{i/\sqrt{s} : i = 0, \ldots, s - 1\}$ and mass $2b_s^2 \varpi_s(s) \Pi_{s-i}(s)$ at $i/\sqrt{s}$, and $g_s(x)$ is a right-continuous simple function defined on $\mathbb{R}_+$ equal to $1 - b_s \Pi_s(s)$ at 0 and with upward jumps at each point $i/\sqrt{s}$, $i = 1, \ldots, s$, where $g_s(x_{si} + 0) = 1 - b_s(\varpi_s/\varpi_{s-i}) \Pi_{s-i}(s)$.

We first show that $\Pi_{s-i}(s)/\varpi_{s-i}(s)$ is a monotone sequence (in i), and that its analogue $e^{(1/2)x^2} \Phi(-x)$ is monotonic in x. For the former, recall that the partial sum $\Pi_i(s)$ is a tail integral of a gamma density function, so

$$\frac{\Pi_i(s)}{\varpi_i(s)} = \frac{e^s i!}{s^i} \int_s^\infty e^{-u} \frac{u^i}{i!} \, du = \int_0^\infty e^{-v} \left(1 + \frac{v}{s}\right)^i \mathrm{d}v, \qquad (3.5)$$

which is monotonic increasing in i, hence, $1 - b_s \Pi_s(s) \leq 1 - b_s \varpi_s \Pi_{s-i}(s)/\varpi_{s-i}(s) \leq 1 - b_s \varpi_s$ for $0 \leq i \leq s$. This monotonicity implies that each function g_s has all jumps upwards as asserted.

For the size of the jumps we evaluate (omitting here the common argument s)

$$\varpi_s \left[\frac{\Pi_{i+1}}{\varpi_{i+1}} - \frac{\Pi_i}{\varpi_i} \right] = \varpi_s \frac{\varpi_i[\Pi_i + \varpi_{i+1}] - \varpi_{i+1}\Pi_i}{\varpi_{i+1}\varpi_i} = \varpi_s \left[1 - \left(1 - \frac{i+1}{s}\right) \frac{\Pi_i}{\varpi_i} \right], \quad (3.6)$$

so the jumps in $g_s(x)$ are bounded uniformly in x by $b_s \varpi_s(s) \leq b_s[1 + 1/12s]/\sqrt{2\pi s} \leq B/\sqrt{s}$, for a constant B, uniformly in s.

For the putative limit function $e^{(1/2)x^2} \Phi(-x)$ in the integrand at (3.3), we have for $x \in \mathbb{R}_+$,

$$\sqrt{2\pi} e^{(1/2)x^2} \Phi(-x) = \int_x^\infty e^{(1/2)x^2} e^{-(1/2)v^2} \, \mathrm{d}v = \int_0^\infty e^{-(1/2)w^2 - wx} \, \mathrm{d}w,$$

which is monotonic in x, so $1 - b_\infty \leq 1 - b_\infty e^{(1/2)x^2} \Phi(-x) \leq 1$ for $0 \leq x \to \infty$.

Thus, omitting the multiplier $b_\infty^2/\sqrt{\pi/2}$, the improper integral based on (3.3) is dominated by $\int_0^\infty \Phi(-u)\,du = 1/\sqrt{2\pi}$, which is finite. Hence, given $\varepsilon > 0$, there exists U_ε such that

$$\int_0^{U_\varepsilon} \Phi(-u)[1 - b_\infty e^{(1/2)u^2}\Phi(-u)]\,du$$

is within ε of the similar integral but over $\mathbb{R}_+$.

We turn to the measures ν_s. Consider the sums $S_{sj} := \varpi_s \sum_{i=0}^j \Pi_i(s)$. Using again the integral representation for $\Pi_i(s)$ as in (3.5),

$$
\begin{aligned}
S_{si} &= \varpi_s \sum_{j=0}^i \int_s^\infty e^{-u} \frac{u^j}{j!}\,du \\
&= \varpi_s \int_s^\infty du \int_u^\infty e^{-v} \frac{v^i}{i!}\,dv \\
&= \varpi_s \int_s^\infty e^{-v} \frac{v^i}{i!}(v - s)\,dv \\
&= \varpi_s[(i+1)\Pi_{i+1}(s) - s\Pi_i(s)] \\
&= \varpi_s[(i+1)\varpi_{i+1}(s) - (s - i - 1)\Pi_i(s)] \\
&= \varpi_s[s\varpi_i(s) - (s - i - 1)\Pi_i(s)] \\
&= s\varpi_s\varpi_i\left[1 - \left(1 - \frac{i+1}{s}\right)\frac{\Pi_i(s)}{\varpi_i}\right] \\
&= s\varpi_s^2 \frac{\varpi_i}{\varpi_s}\left[1 - \left(1 - \frac{i+1}{s}\right)\frac{\Pi_i(s)}{\varpi_i}\right].
\end{aligned}
\tag{3.7}
$$

In particular, for $i = s - 1$, $S_{s,s-1} = s\varpi_s\varpi_{s-1} = s\varpi_s^2$, $= \sum_{i=1}^s \varpi_s\Pi_{s-i}(s) = \nu_s(\mathbb{R}_+)/2b_s^2$.

From Stirling's formula (see, e.g. Appendix A) we know that $s\varpi_s^2 = [1 + \vartheta_s/12s]^2/2\pi$ for some $|\vartheta_s| \le 1$, so $\sup_s \nu_s(\mathbb{R}_+) < \infty$ uniformly in s. Therefore, given $\varepsilon > 0$, for every s there exists j_s such that $S_{s,j_s} > S_{s,s-1} - \varepsilon$.

Proposition 3.1. *The measures $\{\nu_s\}$ introduced at (3.4) converge weakly to the limit measure ν_∞ for which*

$$\nu_\infty([0,x]) = \frac{2}{(\sqrt{\pi/2} + \eta)^2} \int_0^x \Phi(-v)\,dv. \tag{3.8}$$

Remark. Because the limit function at (3.8) is continuous, $\sup_{x\in\mathbb{R}_+} |\nu_s([0,x]) - \nu_\infty([0,x])| \to 0$ for $s \to \infty$, i.e. we have not merely weak convergence but uniformity of that convergence on the domain of definition (e.g. Problem 5 in III.Section 1 in [22]).

Proof of Proposition 3.1. Let $f : \mathbb{R}_+ \mapsto \mathbb{R}$ be a bounded continuous function; without loss of generality suppose $0 \le f(x) \le C_f$ for $x \in \mathbb{R}_+$ and finite C_f. From our definitions,

$$\int_{\mathbb{R}_+} f(x)\nu_s(dx) = 2b_s^2 \sum_{i=0}^s f(x_{si})\varpi_s\Pi_{s-i}(s), \tag{3.9}$$

where $x_{si} = i/\sqrt{s}$ as earlier. Because f is bounded and $S_{si} \le \nu_s(\mathbb{R}_+)/2b_s^2 \to 1/(2\pi)$ as $s \to \infty$, there exists $\{j_s\}$ such that the integral at (3.9), written as $\sum_{i=0}^{j_s-1} + \sum_{i=j_s}^s$, is approximated arbitrarily closely for s large enough by the former of these sums; for example $j_s = s^{5/8}$.

Then using (A.3),

$$\sum_{i=0}^{j_s-1} f(x_{si}) v_s(\{i/\sqrt{s}\}) = 2b_s^2 \sum_{i=0}^{j_s-1} f(x_{si}) \varpi_s \left[\Phi(-x_{si}) + \frac{\psi_{si}}{\sqrt{s}} \right]$$

$$= 2b_s^2 \sum_{i=0}^{j_s-1} f(x_{si}) \Phi(-x_{si}) \varpi_s + 2b_s^2 \sum_{i=0}^{j_s-1} f(x_{si}) \frac{\psi_{si}}{\sqrt{s}} \varpi_s, \quad (3.10)$$

where the constants ψ_{si} are bounded uniformly in i and s, $|\psi_{si}| \le C_\psi$ say. The last sum at (3.10) is bounded by $j_s C_f C_\psi / s\sqrt{2\pi}$, and when $j_s = s^{5/8}$, this bound is arbitrarily small for s sufficiently large. So the left-hand side of (3.9) is arbitrarily close to the sum at (3.10) involving $\Phi(\cdot)$. Each of these sums is an approximation to the Riemann integral $2b_\infty^2 \int_0^{s^{1/8}} f(v)\Phi(-v)\,dv$, which in turn is arbitrarily close (for large enough s) to the improper integral over $\mathbb{R}_+$.

This weak convergence property of $\{v_s\}$ is not directly applicable to the integral at (3.4) in which the functions g_s are not continuous; but they are monotonic and bounded, and have bounded increments with $g_s(x + 1/\sqrt{s}) - g_s(x) \le B/\sqrt{s}$ (see around (3.6)). We show that $g_\infty(x) := \lim_{s\to\infty} g_s(x)$ exists and equals the bounded continuous function $1 - b_\infty e^{(1/2)x^2} \Phi(-x)$ for $x \in \mathbb{R}_+$. To this end, consider $x \in [0, U_\varepsilon]$, where U_ε is as defined below (3.6), and write

$$g_s(x) = 1 - b_s(1) \frac{\Phi(-i_s(x)/\sqrt{s}) + \psi_{s,i_s(x)}/\sqrt{s}}{\phi(i_s(x)/\sqrt{s})[1 + O(s^{-1/8})]/\phi(0)},$$

where $i_s(x) = \min(\lfloor x\sqrt{s} \rfloor, s)$. For given x, $i_s(x)/\sqrt{s} \to x$ as $s \to \infty$, and thus $g_s(x) \to g_\infty(x)$ as defined. Further, because $\phi(i_s(x)/\sqrt{s})$ is bounded away from 0 on $[0, U_\varepsilon]$, this convergence satisfies $\sup_{x\in[0,U_\varepsilon]} |g_s(x) - g_\infty(x)| < B_g/\sqrt{s}$ for some finite constant B_g. We can now write

$$\int_{\mathbb{R}_+} g_s(x) v_s(dx) = \int_{(U_\varepsilon,\infty)} g_s(x) v_s(dx) + \int_{[0,U_\varepsilon]} [g_s(x) - g_{\infty(x)}] v_s(dx)$$

$$+ \int_{[0,U_\varepsilon]} g_\infty(x)[v_s(dx) - v_\infty(dx)] - \int_{(U_\varepsilon,\infty)} g_\infty(x) v_\infty(dx)$$

$$+ \int_{\mathbb{R}_+} g_\infty(x) v_\infty(dx)$$

$$:= T_1 + T_2 + T_3 - T_4 + \int_{\mathbb{R}_+} g_\infty(x) v_\infty(dx). \quad (3.11)$$

Each of the terms $T_1, \dots, T_4$ here can be made smaller than any given positive ε, first by the choice of U_ε for T_4, then by choice of s for T_3 and T_1 by appealing to weak convergence, and finally by choice of s for T_2 by the uniform convergence of $g_s(x)$ in $[0, U_\varepsilon]$.

It remains to demonstrate that the improper integral in (3.3) has the value as asserted. We do this in the proof of the next lemma.

Lemma 3.1. *It holds that*

$$I_+ := \int_0^\infty \Phi(-u)[1 - e^{(1/2)u^2}\Phi(-u)]\,du = \sqrt{2\pi}(1 - \log\sqrt{2}).$$

Proof. Write

$$\int_0^\infty \Phi(-u)\,du - I_+ = \int_0^\infty e^{(1/2)u^2}\,du\,(2\pi)^{-1}\int_u^\infty e^{-(1/2)v^2}\,dv\int_u^\infty e^{-(1/2)w^2}\,dw,$$

noting that the left-hand side equals $1/\sqrt{2\pi} - I_+ =: J_0$ say. Use polar coordinates $(v, w) = (r\cos\theta, r\sin\theta)$ so that $dv\,dw \mapsto r\,dr\,d\theta$ and multiply by 2π, i.e.

$$\begin{aligned}
2\pi J_0 &:= \int_0^\infty e^{(1/2)u^2}\,du\int_u^\infty e^{-(1/2)v^2}\,dv\int_u^\infty e^{-(1/2)w^2}\,dw\\
&= \int_0^\infty e^{(1/2)u^2}\,du\int_{u\sqrt{2}}^\infty \left(\frac{1}{2}\pi - 2\arcsin\frac{u}{r}\right)e^{-(1/2)r^2}r\,dr\\
&= 0 + 2\int_0^\infty e^{(1/2)u^2}\,du\int_{u\sqrt{2}}^\infty \left(-\frac{u}{r^2}\frac{1}{\sqrt{1-u^2/r^2}}\right)(-e^{-(1/2)r^2})\,dr & \text{(a)}\\
&= 2\int_0^\infty e^{-(1/2)z^2\cos 2\theta}z\,dz\int_0^{\operatorname{artan}(1/\sqrt{2})} \frac{z\sin\theta}{z\cos\theta\sqrt{z^2\cos 2\theta}}\,d\theta & \text{(b)}\\
&= 2\int_0^{\operatorname{artan}(1/\sqrt{2})} \frac{\tan\theta\,d\theta}{\sqrt{\cos 2\theta}}\int_0^\infty e^{-(1/2)z^2\cos 2\theta}\,dz & \text{(Fubini's theorem)}\\
&= 2\int_0^{\operatorname{artan}(1/\sqrt{2})} \frac{\tan\theta\,d\theta}{\cos 2\theta}\int_0^\infty e^{-(1/2)y^2}\,dy & (y = z\sqrt{\cos 2\theta})\\
&= \sqrt{2\pi}\int_0^{\operatorname{artan}(1/\sqrt{2})} \frac{\tan\theta\,d\theta}{\cos 2\theta} & \text{(integration)}\\
&= \sqrt{2\pi}\int_0^{1/\sqrt{2}} \frac{t\,dt}{1-t^2} & (t = \tan\theta)\\
&= \sqrt{2\pi}\int_0^{1/2} \frac{(1/2)\,du}{1-u}\\
&= \tfrac{1}{2}\sqrt{2\pi}\log(1/\tfrac{1}{2})\\
&= \sqrt{2\pi}\log\sqrt{2}.
\end{aligned}$$

(Using (a) integration by parts and (b) $(r, u) = (z\cos\theta, z\sin\theta)$.) Recalling the definition of J_0 yields the expression for I_+.

4. Proof of Theorem 2.1: the case $\rho = 1 - \beta/\sqrt{s}$

We now have $\rho = 1 - \beta/\sqrt{s}$ for finite $\beta \neq 0$. The proof is similar to the approach and methods used in the case of $\rho = 1$ above, except the particular algebraic expressions differ. Start by breaking the sum at (3.1) into the two sums as before. Write the first of these as

$$2\sum_{i=1}^s \left(1 - \frac{\pi_J}{\pi_s}\varpi_s(s\rho)\frac{\Pi_{s-i}(s\rho)}{\varpi_{s-i}(s\rho)}\right)\pi_J\Pi_{s-i}(s\rho) = \int_{\mathbb{R}_+} g_s(x)\nu_s(dx), \tag{4.1}$$

where the measure ν_s is atomic as before but now has mass $2b_s^2(\rho)\varpi_s(s\rho)\rho^K\Pi_{s-i}(s\rho)$ located at $x_{si} = i/\sqrt{s}$ for $i = 0, \ldots, s-1$, and the function g_s is a right-continuous simple function defined on $\mathbb{R}_+$ equal to $1 - b_s(\rho)\rho^K\Pi_s(s\rho)$ at 0 and has jumps at each point x_{si}, $(i = 1, \ldots, s)$, where $g_s(x_{si} + 0) = 1 - b_s(\rho)\rho^K\Pi_{s-i}(s\rho)/[\varpi_{s-i}(s\rho)/\varpi_s(s\rho)]$, so that $g_s(\cdot)$ is monotonic.

To check that ν_s has finite total mass, and that there is a bound on $\nu_s(\mathbb{R}_+)$ that is uniform in s, note that $\lim_{s,K\to\infty} b_s^2(\rho)\rho^K$ is finite for finite β and equal to $[\beta h(\eta,\beta)/\phi(\beta)]^2(1-e^{-\beta\eta})$ (see around (2.7)); for the rest, mimic the calculation leading to (3.7), now with $\rho = 1 - \beta/\sqrt{s}$, in computing

$$S_{si} := \varpi_s(s\rho) \sum_{j=0}^{i} \Pi_j(s\rho) = s\rho\varpi_s(s\rho)\varpi_i(s\rho)\left[1 - \left(1 - \frac{i+1}{s\rho}\right)\frac{\Pi_i(s\rho)}{\varpi_i(s\rho)}\right]$$

so that

$$S_{s,s-1} = s\rho\varpi_s(s\rho)\varpi_{s-1}(s\rho)\left[1 - \left(1 - \frac{1}{\rho}\right)\frac{\Pi_{s-1}(s\rho)}{\varpi_{s-1}(s\rho)}\right]$$

$$= s\varpi_s^2(s\rho)\left[1 + \frac{\beta}{\rho\sqrt{s}}\frac{\Pi_{s-1}(s\rho)}{\varpi_{s-1}(s\rho)}\right],$$

which for given (finite) β is bounded in s. Thus, $\sup_s \nu_s(\mathbb{R}_+) < \infty$. Note that

$$\nu_s(\mathbb{R}_+) = 2b_s^2(\rho)\rho^K S_{s,s-1} \to \frac{\beta^2 h^2(\eta,\beta)(1-e^{-\beta\eta})}{\pi\phi^2(\beta)}\left(1 + \beta\sqrt{\frac{\pi}{2}}\right) \qquad (s \to \infty).$$

From Lemma A.2 (Appendix A), for $\rho = 1 - \beta/\sqrt{s}$, $0 \le i \le s$ and s large,

$$\Pi_{s-i}(s\rho) - \frac{\psi_{si}}{\sqrt{s\rho}} = \Phi\left(\frac{s-i-s\rho}{\sqrt{s\rho}}\right)$$

$$= \Phi(-x_{si} + \beta + O(s^{-1/2}))$$

$$= \Phi(-x_{si} + \beta) + O(s^{-1/2}),$$

where $o(\cdot)$ in the last term is uniform in i and s. Hence,

$$\Pi_{s-i}(s\rho) = \Phi(-x_{si} + \beta) + \frac{\tilde{\psi}_{si}}{\sqrt{s}}, \tag{4.2}$$

where by Lemma A.2 $\tilde{\psi}_{si}$ is uniformly bounded in i and s.

For $x \in \mathbb{R}_+$ define $i_s(x) = \lfloor x\sqrt{s} \rfloor$ so that $i_s(x)/\sqrt{s} \to x$ as $s \to \infty$, and

$$g_s(x) = 1 - \frac{b_s(\rho)\rho^K \Pi_{s-i_s(x)}(s\rho)}{\varpi_{s-i_s(x)}(s\rho)/\varpi_s(s\rho)}$$

$$= 1 - b_s(\rho)\frac{\rho^K[\Phi(-x_{s,i_s(x)} + \beta) + \tilde{\psi}_{s,i_s(x)}/\sqrt{s}]}{\phi(-x_{s,i_s(x)} + \beta)/\phi(\beta)[1 + O(s^{-1/2})]}.$$

Then for given x and $s \to \infty$, the right-hand side $\to g_\infty(x)$ defined by

$$g_\infty(x) := 1 - \beta h(\eta,\beta)\frac{(1-e^{-\beta\eta})\Phi(-x+\beta)}{\phi(-x+\beta)} \tag{4.3}$$

at a uniform rate of convergence that is $O(s^{-1/2})$ for x on compact sets like U_ε where the denominator above is bounded away from 0, $b_\infty(\rho) := \lim_{s\to\infty} b_s(\rho) = \beta h(\eta,\beta)/\phi(\beta)$ as below (2.7), and

$$\lim_{s\to\infty}\left(1 - \frac{\beta}{\sqrt{s}}\right)^{\sqrt{s}\eta} = e^{-\beta\eta}.$$

Proposition 4.1. *The measures $\{v_s\}$ defined in (4.1) converge weakly to the limit measure v_∞ for which*

$$v_\infty([0, x]) = \frac{2[\beta h(\eta, \beta)]^2 (1 - e^{-\beta\eta})}{\phi(\beta)} \int_{(-\beta, x - \beta]} \Phi(-v)\, dv. \qquad (4.4)$$

Proof. The same argument as used in establishing Proposition 3.1 holds, subject to using (4.2) in place of (A.4) from Lemma A.2. The analogue of (3.10) for $\beta \neq 0$, for bounded continuous f, is

$$\sum_{i=0}^{j_s - 1} f(x_{si}) v_s\left(\left\{\frac{i}{\sqrt{s}}\right\}\right) = \sum_{i=0}^{j_s - 1} f(x_{si}) 2b_s^2(\rho) \rho^K \varpi_s(s\rho)\left[\Phi(-x_{si} + \beta) + \frac{\tilde{\psi}_{si}}{\sqrt{s}}\right].$$

Each term involving some $\tilde{\psi}_{si}$ is at most $O(1/s)$, so when $j_s < O(s)$ their sum is negligibly small for large s. Since $\varpi_s(s\rho) = e^{-(1/2)\beta^2}/\sqrt{2\pi s}[1 + O(s^{-\frac{1}{2}})]$ the other term is an approximation to the Riemann integral

$$2\left(\lim_{s\to\infty} b_s^2(\rho)\rho^K\right)\phi(\beta) \int_0^{j_s/\sqrt{s}} f(u)\Phi(-u + \beta)\, du,$$

from which we deduce (4.4).

An argument similar to that leading to (3.11) shows that the integrals at the right of (4.1) converge to the limit $\int_{\mathbb{R}_+} g_\infty(x) v_\infty(dx)$.

It remains to consider the analogue of the last sum at (3.1) for which $i \geq s$. As earlier, calculate P_i via its tail which now reads

$$P_i = 1 - \sum_{j=i+1}^{J} \pi_j$$

$$= 1 - \pi_J \sum_{j=i+1}^{J} \rho^{j-J}$$

$$= 1 - \rho^{i+1-J} \frac{\pi_J(\rho^{J-i} - 1)}{\rho - 1}$$

$$= 1 + \frac{\rho\pi_J(1 - \rho^{-(J-i)})}{1 - \rho},$$

so

$$\left(1 - \frac{\pi_J}{\pi_i} P_i\right) P_i = (1 - \rho^{J-i} P_i) P_i$$

$$= \left(1 + \frac{\rho\pi_J(1 - \rho^{-(J-i)})}{1 - \rho}\right)\left[1 - \rho^{J-i}\left(1 + \frac{\rho\pi_J(1 - \rho^{-(J-i)})}{1 - \rho}\right)\right]$$

$$= \left(1 + \frac{\rho\pi_J(1 - \rho^{-(J-i)})}{1 - \rho}\right)(1 - \rho^{J-i})\left(1 + \frac{\rho\pi_J}{1 - \rho}\right)$$

$$= \left(1 + \frac{\rho\pi_J}{1 - \rho}\right)\left[1 - \rho^{J-i} + \frac{\rho\pi_J(2 - \rho^{J-i} - \rho^{-(J-i)})}{1 - \rho}\right].$$

Hence, for the second part of the sum of (3.1) for the case $\rho \neq 1$ and $i \geq s$ we have

$$2\pi_J \sum_{i=s}^{J} \left(1 - \frac{\pi_J}{\pi_i} P_i\right) P_i$$

$$= 2\pi_J\left(1 + \frac{\rho\pi_J}{1-\rho}\right)\left[(K+1)\left(1 + \frac{2\rho\pi_J}{1-\rho}\right) - \frac{\rho^K(\rho^{-K-1}-1)}{\rho^{-1}-1}\left(1 + \frac{\rho\pi_J}{1-\rho}\right)\right.$$

$$\left. - \frac{\rho^{-K}(\rho^{K+1}-1)}{\rho-1} \cdot \frac{\rho\pi_J}{1-\rho}\right]$$

$$= 2\pi_J\left(1 + \frac{\rho\pi_J}{1-\rho}\right)\left[\left(K+1 - \frac{1-\rho^{K+1}}{1-\rho}\right)\left(1 + \frac{\rho\pi_J}{1-\rho}\right)\right.$$

$$\left. + \left(K+1 - \frac{1-\rho^{K+1}}{\rho^K(1-\rho)}\right)\frac{\rho\pi_J}{1-\rho}\right]. \tag{4.5}$$

For $\rho = 1 - \beta/\sqrt{s}$ and $s, K \to \infty$ as at (1.2)

$$\frac{\rho\pi_J}{1-\rho} = \frac{\sqrt{s}\pi_s\rho^{K+1}}{\beta} \approx \frac{\sqrt{s}\pi_s(1 - e^{-\beta\eta})}{\beta},$$

so the limit of the right-hand side of (4.5) is the same as the limit of

$$-2\pi_s\rho^K\left(1 + \frac{\sqrt{s}\pi_s(1 - e^{-\beta\eta})}{\beta}\right) \tag{4.6}$$

$$\times \left[\left(K - \frac{\sqrt{s}(1-(1-e^{-\beta\eta}))}{\beta}\right)\left(1 + \frac{\sqrt{s}\pi_s(1 - e^{-\beta\eta})}{\beta}\right)\right.$$

$$\left. + \left(K - \frac{\sqrt{s}(1-(1-e^{-\beta\eta}))}{(1-e^{-\beta\eta})\beta}\right)\frac{\sqrt{s}\pi_s(1 - e^{-\beta\eta})}{\beta}\right]$$

$$\to -2\beta h(\eta, \beta)(1 - e^{-\beta\eta})\left(1 + \frac{\beta h(\eta, \beta)(1 - e^{-\beta\eta})}{\beta}\right)$$

$$\times \left[\left(\eta - \frac{1-(1-e^{-\beta\eta})}{\beta}\right)\left(1 + \frac{\beta h(\eta, \beta)(1 - e^{-\beta\eta})}{\beta}\right)\right.$$

$$\left. + \left(\eta - \frac{1-(1-e^{-\beta\eta})}{\beta(1-e^{-\beta\eta})}\right)\frac{\beta h(\eta, \beta)(1 - e^{-\beta\eta})}{\beta}\right]$$

$$= -2\widetilde{C}\left(1 + \frac{\widetilde{C}}{\beta}\right)\left[\left(\eta - \frac{1-(1-e^{-\beta\eta})}{\beta}\right)\left(1 + \frac{\widetilde{C}}{\beta}\right) + \left(\eta - \frac{1-(1-e^{-\beta\eta})}{\beta(1-e^{-\beta\eta})}\right)\frac{\widetilde{C}}{\beta}\right], \tag{4.7}$$

where $\widetilde{C} = \beta h(\eta, \beta)(1 - e^{-\beta\eta})$. Combining (4.3) and (4.4), and (4.7), as required for the sum at (3.1), now yields

$$1 - \mathcal{D}_{\beta,\eta} = 2 \lim_{s\to\infty} \sqrt{s}\pi_s \int_{-\beta}^{\infty} (1 - e^{-\beta\eta})$$

$$\times \left(1 - \frac{(1 - e^{-\beta\eta})\phi(\beta)\Phi(-u)}{\phi(u)}\right)\Phi(-u)\,\mathrm{d}u + g(\eta, \beta). \tag{4.8}$$

We should like at this point to evaluate the integral at (4.8). For this purpose we have mimicked the steps followed in establishing the expression for I_+ in Lemma 3.1, but have succeeded only in reaching the expressions in (4.9) and (4.10) below in which final quadratures are unresolved. Inspection of (4.8) shows that we must evaluate, for both positive and negative β,

$$J_\beta := \int_\beta^\infty \frac{[\Phi(-u)]^2}{\phi(u)} \, du = \int_\beta^\infty e^{(1/2)u^2} \, du \int_u^\infty \int_u^\infty \frac{e^{-(1/2)(v^2+w^2)}}{2\pi} \, dv \, dw.$$

By following steps similar to those in the proof of Lemma 3.1 we deduced first that for $\beta > 0$,

$$J_\beta = -\frac{\Phi(-|\beta|)}{\sqrt{2\pi}} \log 2 + \int_1^\infty \frac{\beta e^{-(1/2)\beta^2 x^2}}{2\pi} \log\left(1 + \frac{1}{x^2}\right) dx. \tag{4.9}$$

For $\beta < 0$, integration takes place (cf. second line of the evaluation of J_0 in the proof of Lemma 3.1) over $0 < r < \infty$, with angular coordinate having arc-lengths 2π for $r < |\beta|$, $2\pi - 4\arccos(|\beta|/r) = 4\arcsin(|\beta|/r)$ for $|\beta| < r < |\beta|\sqrt{2}$, and $(1/2)\pi + 2\arcsin(|\beta|/r)$ for $r > |\beta|\sqrt{2}$. Ultimately this yields

$$J_{-|\beta|} = J_{|\beta|} - J_0 + \int_0^{|\beta|} e^{(1/2)u^2} \, du + \frac{4\log 2}{\sqrt{2\pi}} \left[\Phi(|\beta|\sqrt{2}) - \frac{1}{2}\right]$$

$$- \int_0^{\sqrt{2}} \frac{4|\beta|e^{-(1/2)\beta^2 x^2}}{2\pi} \log\left(1 + \frac{1}{x^2}\right) dx. \tag{4.10}$$

5. Conclusion

The analysis presented in this paper both answers questions and raises further points that need clarification. Recall that in the setup in [19], Nazarathy and Weiss observed that for M/M/1/K systems with $K \to \infty$, $\mathcal{D}_\pi \to 1$ for $0 < \rho < \infty$ except when $\rho = 1$, for which they found $\mathcal{D}_\pi \to 2/3$. A similar effect occurs in multi-server M/M/s/K systems. This is illustrated in Figure 3.

Our theorem shows that in the multi-server case, this discontinuity is merely a matter of scale: setting $\beta = (1-\rho)\sqrt{s}$ and considering

$$\mathcal{D}_\beta = \lim_{t\to\infty} \left.\frac{\text{var}(N_{\text{dep}}(0,t]))}{\text{E}(N_{\text{dep}}(0,t]))}\right|_{\rho=1-\beta/\sqrt{s}},$$

the limit function is no longer discontinuous but reflects the asymptotic variance over a much finer range of values as in Figure 2 of Section 2.

In turn, this emphasises that, for the systems that we have considered, there exists some subtle interplay between the service facility on the one hand and customers on the other. In terms of customers there may be a 'deficit' (so some servers are idle) or a 'surfeit' (so some customers are waiting), and when the system is 'balanced', meaning that $\rho \approx 1$, the occasions when the system is either empty or full are rare (i.e. both π_0 and π_J are 'small') so that mostly periods with above or below average net arrivals are balanced by periods of above or below average productive service, and 'consequently' the variability of throughput of the system as measured by the output N_{dep} is to some extent 'smoothed'. This intuitive explanation of the BRAVO effect is one we have not been able to translate into a mathematical explanation.

In other work concurrently under preparation, we have examined the many-server M/M/s system with reneging to see whether it too exhibits the BRAVO effect. In studying a system

with reneging, the departure process $N_{\text{dep}}(0, \cdot]$ no longer consists precisely of the deaths in the birth–death process $Q_s(\cdot)$ but only of a subset of those deaths. Consequently the second equality at (1.5) is no longer available as an expression for $\mathcal{D}_\pi$, but an independent computation, with the advantage from an expository viewpoint that it relies on more primitive results, yields a substitute expression

$$1 - 2 \sum_{i=0}^{J} (P_i - \Lambda_i^*) \left(q_{i+1} - \frac{\lambda^*}{\pi_i \lambda_i} (P_i - \Lambda_i^*) \right),$$

where q_i is the probability of having a death at state i of $Q_s(\cdot)$ counts as an increment of $N_{\text{dep}}(0, \cdot]$. A complete derivation of this result has been deferred from this paper, in which context the present work should have been more self-contained.

The same substitute expression can also cope with a sequence of systems with balking, where we expect the BRAVO effect may still be observed, but whether such a system must be finite, with a buffer in place as well as the balking mechanism, remains to be discovered.

Appendix A. Asymptotic normality of Poisson probabilities

In formulating the condition at (1.2) and establishing the double limit property (1.3) we use both local and central limit properties of the Poisson distribution with mean κ, here denoted $\{\varpi_i\} := \{\varpi_i(\kappa)\} := \{e^{-\kappa} \kappa^i / i!\}$ and write $\Pi_i(\kappa) := \sum_{j=0}^{i} \varpi_j(\kappa)$ for the corresponding distribution function. For convenience we state these properties with their respective rates of convergence in two lemmas below. We draw on the exposition in Feller [8] concerning Stirling's formula and limit results for the binomial distribution (p. 54 and p. 183 respectively).

For Stirling's formula Feller shows that for any positive integer, s,

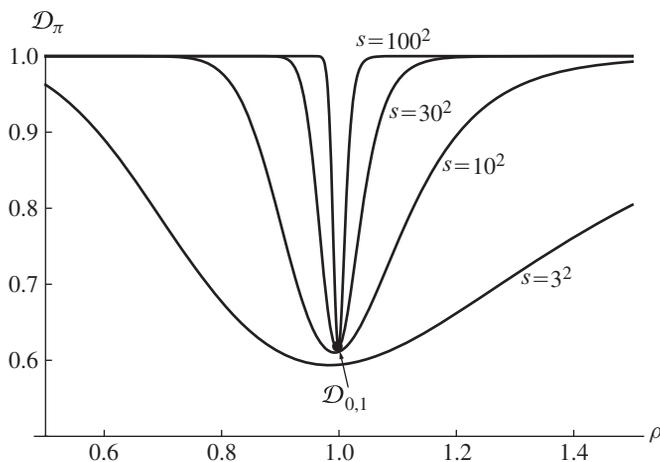$$e^{1/(12s+1)} < \frac{s!}{\sqrt{2\pi s} (s/e)^s} < e^{1/(12s)},$$

FIGURE 3: $\mathcal{D}_\pi$ as a function of ρ in M/M/s/K systems with $K = \sqrt{s}$ for increasing values of s. The black dot is $\mathcal{D}_{0,1}$ corresponding to $s \to \infty$ with $\eta = 1$.

from which it follows that

$$\varpi_s(s) = e^{-s}\frac{s^s}{s!} = \frac{1}{\sqrt{2\pi s}}\left(1 - \frac{\gamma_s}{12s}\right),$$

where $0 < \gamma_s \to 1$ as $s \to \infty$. We use a local central limit theorem for Poisson probabilities, including a rate of convergence, as follows.

Lemma A.1. *For $i, s \to \infty$ such that $i/\sqrt{s} \to \gamma$ for fixed finite $\gamma > 0$, uniformly for $i < \lfloor s^{5/8} \rfloor$,*

$$\frac{\varpi_{s-i}(s)}{\varpi_s(s)} = \frac{\phi(\gamma)}{\phi(0)}(1 + O(1/s^{1/8})),$$

where the rate of convergence holds uniformly for $i \le s^{5/8}$.

Proof. Analogous to Feller's proof of the local asymptotic normality of binomial probabilities, write

$$\frac{\varpi_s(s)}{\varpi_{s-i}(s)} = \frac{s^i}{s(s-1)\cdots(s-i+1)} = \frac{1}{(1-t_1)\cdots(1-t_{i-1})}, \tag{A.1}$$

where for $j = 1, \dots, i-1$, $t_j = j/s$. For $|t| \le \frac{1}{2}$, $1/(1-t) = \exp[-\log(1-t)] = e^{t+(1/2)t^2+\vartheta_t t^3}$, where $|\vartheta_t| < 1$. Then for positive integers $i \le \frac{1}{2}s$, the ratio at (A.1) equals

$$\exp\left(\frac{(i-1)i}{2s} + \bar{\vartheta}_i\frac{(i-1)i(2i-1)}{6s^2}\right), \tag{A.2}$$

where $0 < \bar{\vartheta}_i \le 1$. Let $j_s \to \infty$ as $s \to \infty$ such that $j_s^3/s^2 \to 0$. Then for (i, s) with $i < j_s$, the quantity at (A.2) differs from $e^{(1/2)\gamma^2}$ by a multiplicative factor that is dominated by $\exp(\bar{\vartheta}_i(j_s^3/s^2))$, so for $j_s = s^{5/8}$, this last term yields a factor $O(1/s^{1/8})$, whose expansion then completes the proof.

Lemma A.2. *There exist finite constants ψ_{sj} that are uniformly bounded in both j and s such that*

$$\sum_{i=0}^{j} \varpi_i(s) = \Phi\left(\frac{j-s}{\sqrt{s}}\right) + \frac{\psi_{sj}}{\sqrt{s}}. \tag{A.3}$$

Proof. Let X_s denote a Poisson random variable with mean s, so that X_s is expressible as the sum of s i.i.d. Poisson random variables X_1 with mean 1, variance 1, and third absolute moment about the mean $\beta_3 = 1 + 2/e$. Then by the Berry–Esseen theorem (e.g. [22] p. 342), for a positive constant $C < 4/5$,

$$\sup_i \left| \Pr\{X_s \le i\} - \Phi\left(\frac{i-s}{\sqrt{s}}\right) \right| \le \frac{C\beta_3}{\sqrt{s}}.$$

Equation (A.3) follows, with an explicit bound for the constant, and we write this in the form, for uniformly bounded quantities ψ_{si},

$$\Pi_i(s) = \Phi\left(\frac{i-s}{\sqrt{s}}\right) + \frac{\psi_{si}}{\sqrt{s}}. \tag{A.4}$$

Corollary A.1. *The constants* $\{\psi_{si}\}$ *in the Berry–Esseen expansion (A.4) for the Poisson distribution satisfy* $\sum_{i=1}^{s} \psi_{si} = o(s)$.

Proof. From around (3.7),

$$S_{s,s-1} = \sum_{i=0}^{s-1} \varpi_s(s)\Pi_i(s) = s\varpi_s^2 = \frac{1}{2\pi}\left(1 + \frac{\vartheta_s}{12s}\right)^2 \to \frac{1}{2\pi} \qquad (s \to \infty),$$

where $|\vartheta_s| \le 1$, while substituting from (A.4) into the expression for $S_{s,s-1}$ gives

$$S_{s,s-1} = \sum_{i=0}^{s-1} \varpi_s(s)\left[\Phi(\frac{i-s}{\sqrt{s}}) + \frac{\psi_{si}}{\sqrt{s}}\right]$$

$$= (1 + \vartheta_s/12s)\left[\sum_{i=1}^{s} \Phi(-x_{si})\frac{1}{\sqrt{2\pi}\sqrt{s}} + \frac{1}{\sqrt{2\pi}s}\sum_{i=0}^{s-1} \psi_{si}\right], \qquad (A.5)$$

where $x_{si} = i/\sqrt{s}$. Examining these last two sums, the former is an approximation, based on the s intervals determined by the $s+1$ points $\{i/\sqrt{s}\}$ $(i = 0, \ldots, s)$, to the integral $\int_0^{\sqrt{s}} \Phi(-u)\,du/\sqrt{2\pi}$. This in turn is a finite portion of the integral $\int_0^{\infty} \Phi(-u)\,du/\sqrt{2\pi}$; this improper (Riemann) integral exists because $\Phi(u)$ is directly Riemann integrable on $(-\infty, 0)$ and equals $1/\sqrt{2\pi}$. This means that the Riemann integral approximations converge to $1/2\pi = \lim_{s\to\infty} S_{s,s-1}$. This last equality implies that the last sum at (A.5) $\to 0$ for $s \to \infty$.

The Berry–Esseen theorem is typically proved using Fourier techniques. We have not explored how such techniques would, presumably, provide another way of establishing the corollary. Analogues of Lemmas A.1 and A.2 for general Poisson distributions $\{\varpi_i(\lambda)\}$ for (large) λ, not necessarily an integer, are readily constructed.

Acknowledgements

References

[1] AL HANBALI, A., MANDJES, M., NAZARATHY, Y. AND WHITT, W. (2011). The asymptotic variance of departures in critically loaded queues. *Adv. Appl. Prob.* **43,** 243–263.

[2] ARMONY, M. AND MAGLARAS, C. (2004). On customer contact centers with a call-back option: customer decisions, routing rules, and system design. *Operat. Res.* **52,** 271–292.

[3] BORST, S., MANDELBAUM, A. AND REIMAN, M. I. (2004). Dimensioning large call centers. *Operat. Res.* **52,** 17–34.

[4] BROWNE, S. AND WHITT, W. (1995). Piecewise-linear diffusion processes. In *Advances in Queueing*, CRC, Boca Raton, FL, pp. 463–480.

[5] DALEY, D. J. (1976). Queueing output processes. *Adv. Appl. Prob.* **8,** 395–415.

[6] DALEY, D. J. (2011). Revisiting queueing output processes: a point process viewpoint. *Queueing Systems* **68,** 395–405.

[7] DISNEY, R. L. AND KÖNIG, D. (1985). Queueing networks: a survey of their random processes. *SIAM Rev.* **27,** 335–403.

[8] FELLER, W. (1968). *An Introduction to Probability Theory and Its Applications*, Vol. I, 3rd edn. John Wiley, New York.

[9] GANS, N., KOOLE, G. AND MANDELBAUM, A. (2003). Telephone call centers: tutorial, review, and research prospects. *Manufacturing Service Operat. Manag.* **5,** 79–141.

[10] GARNETT, O., MANDELBAUM, A. AND REIMAN, M. (2002). Designing a call center with impatient customers. *Manufacturing Service Operat. Manag.* **4,** 208–227.

[11] HALFIN, S. AND WHITT, W. (1981). Heavy-traffic limits for queues with many exponential servers. *Operat. Res.* **29,** 567–588.

[12] JANSSEN, A. J. E. M., VAN LEEUWAARDEN, J. S. H. AND SANDERS, J. (2013). Scaled control in the QED regime. *Performance Evaluation* **70,** 750–769.

[13] JANSSEN, A. J. E. M., VAN LEEUWAARDEN, J. S. H. AND ZWART, B. (2011). Refining square-root safety staffing by expanding Erlang C. *Operat. Res.* **59,** 1512–1522.

[14] JELENKOVIĆ, P., MANDELBAUM, A. AND MOMČILOVIĆ, P. (2004). Heavy traffic limits for queues with many deterministic servers. *Queueing Systems* **47,** 53–69.

[15] MAGLARAS, C. AND ZEEVI, A. (2004). Diffusion approximations for a multiclass Markovian service system with 'guaranteed' and 'best-effort' service levels. *Math. Operat. Res.* **29,** 786–813.

[16] MANDELBAUM, A. AND MOMČILOVIĆ, P. (2008). Queues with many servers: the virtual waiting-time process in the QED regime. *Math. Operat. Res.* **33,** 561–586.

[17] MASSEY, W. A. AND WALLACE, R. (2005). An asymptotically optimal design of the M/M/c/K queue for call centers. Unpublished manuscript.

[18] NAZARATHY, Y. (2011). The variance of departure processes: puzzling behavior and open problems. *Queueing Systems* **68,** 385–394.

[19] NAZARATHY, Y. AND WEISS, G. (2008). The asymptotic variance rate of the output process of finite capacity birth–death queues. *Queueing Systems* **59,** 135–156.

[20] PANG, G., TALREJA, R. AND WHITT, W. (2007). Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Prob. Surveys* **4,** 193–267.

[21] REED, J. (2009). The $G/GI/N$ queue in the Halfin–Whitt regime. *Ann. Appl. Prob.* **19,** 2211–2269.

[22] SHIRYAYEV, A. N. (1984). *Probability*. Springer, New York.

[23] WHITT, W. (2004). A diffusion approximation for the $G/GI/n/m$ queue. *Operat. Res.* **52,** 922–941.

[24] WHITT, W. (2005). Heavy-traffic limits for the $G/H_2^*/n/m$ queue. *Math. Operat. Res.* **30,** 1–27.