Some asymptotic results for the transient distribution of the Halfin–Whitt diffusion process

QIANG ZHEN\(^1\) and CHARLES KNESSL\(^2\)

\(^1\)Department of Mathematics and Statistics, University of North Florida, 1 UNF DR, Jacksonville, FL 32224, USA
email: q.zhen@unf.edu

\(^2\)Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago, 851 S. Morgan St.(M/C 249), Chicago, IL 60607, USA
e-mail: knessl@uic.edu

(Received 1 February 2014; revised 20 January 2015; accepted 22 January 2015; first published online 20 February 2015)

We consider the Halfin–Whitt diffusion process \(X_d(t)\), which is used, for example, as an approximation to the \(m\)-server \(M/M/m\) queue. We use recently obtained integral representations for the transient density \(p(x, t)\) of this diffusion process, and obtain various asymptotic results for the density. The asymptotic limit assumes that a drift parameter \(\beta\) in the model is large, and the state variable \(x\) and the initial condition \(x_0\) (with \(X_d(0) = x_0 > 0\)) are also large. We obtain some alternate representations for the density, which involve sums and/or contour integrals, and expand these using a combination of the saddle point method, Laplace method and singularity analysis. The results give some insight into how steady state is achieved, and how if \(x_0 > 0\) the probability mass migrates from \(X_d(t) > 0\) to the range \(X_d(t) < 0\), which is where it concentrates as \(t \to \infty\), in the limit we consider. We also discuss an alternate approach to the asymptotics, based on geometrical optics and singular perturbation techniques.

Key words: diffusion processes; asymptotic expansions; singular perturbation

1 Introduction

Halfin and Whitt [13] introduced in their 1981 paper a heavy-traffic limit theorem for the standard \(GI/M/m\) queue. They demonstrated how under certain conditions a sequence of normalized queue length processes converges to a process that behaves like a Brownian motion with drift above zero and like an Ornstein–Uhlenbeck process below zero. We refer to this hybrid diffusion process as the \textit{Halfin–Whitt diffusion}.

We denote the diffusion process as \(X_d(t)\) whose state space is \((-\infty, \infty)\), the process has constant drift \(-\beta\) in the range \(X_d > 0\), and a linear drift \(-X_d - \beta\) in the range \(X_d < 0\), which acts as a “restoring force”. Thus the drift function is continuous at \(X_d = 0\), but not smooth there. When the number of servers \(m\) is large, Halfin and Whitt [13] showed that the number of customers \(N(t)\) in a \(GI/M/m\) system may be approximated by \(N(t) \approx m + \sqrt{m} X_d(t)\). Here we must also scale \(\lambda/(\mu m) = 1 - \beta/\sqrt{m} = 1 - O(m^{-1/2})\), where \(\lambda\) and \(\mu\) are the arrival and service rates in the model. Thus the drift parameter \(\beta\) is defined as the limit of \(\sqrt{m}[1 - \lambda/(\mu m)]\) as \(m \to \infty\). This scale is nowadays known as the Halfin–Whitt regime. The scaling \(\lambda/(\mu m) = 1 + O(m^{-1/2})\) was also used earlier by
Erlang [5] for the $M/M/m/m$ system and Pollaczek [21] for the $M/D/m/m$ system. Note that time $t$ is not scaled in this diffusion limit, and the model is stable for $\beta > 0$.

The Halfin–Whitt diffusion limit and the Halfin–Whitt regime currently get immensely popular due to applications in the modelling of call centres (see [4,10,14]). Related limit theorems for other more general and more complicated models appear in [11,15,19,22], and the mean hitting times for the Halfin–Whitt diffusion were analysed in [17].

Assuming that $X_d(0) = x_0$ with probability one, we let $p(x,t) = p(x,t;x_0,\beta)$ be the density of the diffusion process, which satisfies $p(x,0) = \delta(x-x_0)$, and the Kolmogorov forward equation in equation (2.1). This diffusion arises naturally in the multi-server $M/M/m/m$ queue under the scaling mentioned above. Our concern in this paper is the asymptotics of the transient behaviour of this diffusion.

For $\beta > 0$, it is relatively easy to obtain the steady state limit $p(x,\infty)$ (see equation (2.2)), but the transient density is much more complicated. In [16] integral representations are given for $p(x,t)$, where it proves necessary to consider separately the four cases $x \leq 0$ and $x_0 \leq 0$. The integrals involve inverse Laplace transforms and parabolic cylinder functions. In view of the complexity of the transient density, it is useful to evaluate it in certain asymptotic limits, as the asymptotic results will be simpler and yield more insight. One asymptotic limit worthy of consideration is that of large time, where we estimate the approach to equilibrium. In [9] and [16], estimates are given of the form

\[
\int_{x_0}^{\infty} e^{-r(\beta,t)} dt \approx \frac{
}{
}
\]

in certain asymptotic limits, as the asymptotic results will be simpler and yield more insight. One asymptotic limit worthy of consideration is that of large time, where we estimate the approach to equilibrium. In [9] and [16], estimates are given of the form $p(x,t) \approx e^{-r(\beta,t)}$ for $t \to \infty$ where $r(\beta)$ is the “relaxation rate”, which depends on the drift parameter. In particular, it was shown in [9] that $r(\beta) = \beta^2/4$ if $\beta \leq \beta_*$ and $r(\beta) = r_0(\beta)$ if $\beta \geq \beta_*$, where $\beta_* \approx 1.85722$ is the minimal positive root of $D_{\beta^2/4}(\beta) = 0$. Here $D$ is the parabolic cylinder function $D_\rho(z)$ and $D_\rho'(z) = \frac{d}{dz} D_\rho(z)$ is its derivative. Then $r_0(\beta)$ is the minimal positive solution to

\[
\sqrt{\frac{\beta^2}{4} - r_0(\beta)} D_{r_0(\beta)}(-\beta) = D_{r_0(\beta)}'(-\beta), \quad 0 < r_0(\beta) < \frac{\beta^2}{4},
\]

which can be shown to exist for $\beta \geq \beta_*$. In [9], $r_0(\beta)$ was characterized in a somewhat different form. If we let $\hat{p}(x;\theta) = \int_0^\infty e^{-\theta t} p(x,t) dt$ be the Laplace transform of the density, then $\theta = -\beta^2/4$ corresponds to a branch point of $\hat{p}$ and $\theta = -r_0(\beta)$ is a pole of $\hat{p}$. In [16], we examined further the spectral properties of the Halfin–Whitt diffusion, showing that there is always a continuous spectrum, corresponding to $\Re(\theta) \leq -\beta^2/4$, and that the point spectrum is empty for $\beta < 0$, and may contain any number of points for $\beta > 0$. For $0 < \beta < \beta_*$, $\theta = 0$ is the only pole of $\hat{p}$, but as $\beta$ increases, other poles appear in the range where $\theta$ is real with $\theta \in (-\beta^2/4,0)$. In fact an additional pole appears precisely when $\beta$ increases past a root of $D_{\beta^2/4}'(-\beta) = 0$, and there are infinitely many of these. In the limit $\beta \to +\infty$ the point spectrum approaches $-\theta = \{0,1,2,\ldots\}$, which is that of the standard free space Ornstein–Uhlenbeck process, where the linear drift $-X_d - \beta$ would apply for all $X_d \in (-\infty,\infty)$.

The relaxation rate asymptotics in [9] and [16] assume that $x$, $x_0$ and $\beta$ are fixed, and that only $t \to \infty$. Here we shall consider another asymptotic limit where $\beta$ is assumed large and positive, and $(x,x_0)$ are also scaled to be $O(\beta)$

\[
\beta \to +\infty, \quad x = \beta X, \quad x_0 = \beta X_0, \quad X_0 > 0.
\]

(1.1)
Here $x_0 = O(\beta)$ and $X_0 > 0$, means that the process $X_d(t)$ starts at a large positive value, but as $t$ increases it will most likely be in the range $X_d(t) < 0$, due to the strong negative drift. For very large times we expect that $X_d(t) \approx -\beta$, since this is the equilibrium point in the range $X_d < 0$. As for the $M/M/m$ model, this means initially the number of customers $N(t)$ well exceeds the number of servers. But for $t \to \infty$ the density in the steady state becomes sharply concentrated near $X = -1$ (as the second formula in equation (2.3) becomes $p(x, \infty) \sim e^{-\beta^2(x+1)^2}/\sqrt{2\pi}$, for $X < 0$). Then $N(t)$ tends to be less than the number of servers $m$, and there is no need for customers to queue. Thus our analysis will provide some insight into how the $M/M/m$ model tends to process large queues, at least in this diffusion limit.

We note that our analysis not only indicates where most of the probability mass concentrates, but also predicts how some small amount of probability mass would remain in the range $X > 0$. Our goal is to obtain a “global” description of the density $p(x, t) = p(x, t; x_0, \beta)$ and to show how it behaves in various space-time ranges. We also note that the large $(\beta, x_0)$ scale represents an intermediate case where the discrete model has the probability going from $O(1)$ to being exponentially small. The asymptotic results presented by this paper are very different from the relaxation rate results, as for large $\beta$ the continuous spectrum will not play much of a role, and there will be many points in the point spectrum, which will correspond to poles of $\hat{p}$ that are close to $-\theta = 0, 1, 2, \ldots$. Getting an accurate estimate of these poles for $\beta \to +\infty$, will be crucial to the asymptotic analysis, and in fact our analysis will precisely estimate them.

Our analysis is based on complex variable techniques, the asymptotic evaluation of integrals and sums, and properties of the parabolic cylinder functions $D_p(z)$. Good references on asymptotic methods for evaluating integrals/sums, such as the Laplace method, saddle point method, Euler–Maclaurin formula, singularity analysis, etc., may be found in the books of Bender and Orszag [2], Bleistein and Handelsman [3], Flajolet and Sedgewick [7], Szpankowski [24] and Wong [27]. Basic properties of parabolic cylinder functions appear in the books of Abramowitz and Stegun [1], Gradsteyn and Ryzhik [12], Magnus, Oberhettinger and Soni [18] and Temme [25].

We comment that in the probability literature, the type of asymptotics considered here is often approached using the theory of large deviations (see [6,8,26] for general references and [23] for queueing applications). However, that theory seems to only give a rough approximation, and would estimate, say, only the logarithm log$[p(x, t)]$, leading perhaps to the exponential part of e.g., equation (2.37). Our analysis not only gives the full leading term, it can be used to get the full asymptotic series for $p(x, t)$, and we estimate $p(x, t)$ both where there is significant probability mass and in the tails, where the density may be exponentially small for large $\beta$.

The paper is organized as follows. In Section 2, we summarize our main results (see Theorems 2.2 and 2.3), and interpret them probabilistically. In Section 3, we derive the results for the range $x > 0$, and in Section 4 for $x < 0$. We re-derive the results using an alternate approach in Section 5, which is based on a singular perturbation analysis of the PDE satisfied by $p(x, t)$. This more geometric approach is then compared and contrasted to the asymptotic evaluation of sums and integrals. Note, however, that the two asymptotic analyses are completely independent of one another. Finally, in Section 6, we give some concluding remarks. As the analysis is fairly
2 Summary of results

We let \( p(x, t) \) be the transient probability density of the Halfin–Whitt diffusion \( X_d(t) \), subject to the initial condition \( X_d(0) = x_0 \). The density satisfies the Kolmogorov forward equation

\[
p_t = \begin{cases} 
\beta p_x + p_{xx}, & x > 0 \\
\beta p_x + (xp)_x + p_{xx}, & x < 0 
\end{cases}
\]

(2.1)

with the initial condition \( p(x, 0) = \delta(x-x_0) \) and the “interface” condition \( p(0^-, t) = p(0^+, t) \) and \( p_x(0^-, t) = p_x(0^+, t) \). This diffusion arises naturally in the multi-server M/M/m queue, in the limit where \( m \to \infty \) and traffic intensity \( \rho = \lambda/(\mu m) \to 1 \), with the scaling \( \rho = 1 - \beta/\sqrt{m} = 1 + O(m^{-1/2}) \). Here \( \lambda \) and \( \mu \) are the arrival and service rates in the discrete model and we set \( \mu = 1 \) without loss of generality. Note that the discrete model is stable for \( \rho < 1 \) which corresponds to \( \beta > 0 \). The queue length \( N(t) \) in the M/M/m queue can then be approximated by \( N(t) \approx m + \sqrt{m}X_d(t) \). (For detailed discussion, see Halfin and Whitt [13], Section 2.)

When \( \beta > 0 \) the steady state distribution of the diffusion can be easily obtained as

\[
p(x, \infty) = C(\beta) \begin{cases} 
\beta e^{-\beta x}, & x > 0 \\
\beta e^{-\beta x}e^{-x^2/2}, & x < 0 
\end{cases}
\]

(2.2)

where the normalization constant is \( 1/C(\beta) = 1 + \beta e^{\beta^2/2} \int_{-\infty}^{\beta} e^{-u^2/2} du \). In the limit of \( \beta \to +\infty \) we have \( 1/C(\beta) \sim \sqrt{2\pi} \beta e^{\beta^2/2} \) and then

\[
p(x, \infty) \sim \frac{1}{\sqrt{2\pi}} \begin{cases} 
e^{-\beta x}e^{-\beta^2/2}, & x > 0 \\
e^{-(x+\beta)^2/2}, & x < 0 
\end{cases}
\]

(2.3)

which is concentrated near \( x = -\beta < 0 \).

Now assume that the initial condition \( x_0 > 0 \). In terms of the discrete model this means that the initial number of customers \( N(0) \) exceeds the number of servers \( m \), so that some customers are in the queue. In [16] we obtained an explicit expression for the Laplace transform \( \hat{p}(x, \theta) = \int_0^\infty e^{-\theta t} p(x, t) dt \) of the transient density, which we summarize below.

**Theorem 2.1** If \( x_0 > 0 \) we have the following (see [16]).

(1) For \( x > 0 \)

\[
\hat{p}(x, \theta) = \frac{e^{\theta(x_0-x)/2}}{\sqrt{\theta^2 + 4\theta}} \left[ e^{-|x-x_0|\sqrt{\theta + \beta^2/4}} - e^{-(x+x_0)\sqrt{\theta + \beta^2/4}} \right] + \frac{e^{\theta(x_0-x)/2} e^{-(x+x_0)\sqrt{\theta + \beta^2/4}}}{\sqrt{\theta + \beta^2/4 - R_\theta(\theta)}}.
\]

(2.4)
with
\[ R_\beta(\theta) = \frac{D'_{-\theta}(-\beta)}{D_{-\theta}(-\beta)} = -\frac{d}{d\beta} \log[D_{-\theta}(-\beta)], \]
where \( D_{-\theta}(-\beta) \) is the parabolic cylinder function.

(2) For \( x < 0 \)
\[ \hat{p}(x, \theta) = e^{-x^2/4} e^{\beta(2x_0-x)/2} \frac{e^{-x_0\sqrt{\theta+\beta^2/4}}}{\sqrt{\theta + \beta^2/4 - R_\beta(\theta)}} \frac{D_{-\theta}(-\beta - x)}{D_{-\theta}(-\beta)}. \quad (2.5) \]

When \( x = 0 \), either equation (2.4) or (2.5) may be used to compute \( \hat{p}(0, \theta) \), as \( \hat{p}(0^+, \theta) = \hat{p}(0^-, \theta) \). By inverting the Laplace transform we obtain integral representations for \( p(x, t) \) in the form
\[ p(x, t) = \frac{1}{2\pi i} \int_{Br} e^{\theta t} \hat{p}(x, \theta) d\theta, \quad (2.6) \]
where \( Br \) is a vertical Bromwich contour in the complex \( \theta \)-plane, on which \( \Re(\theta) > 0 \).

In [16] we studied some spectral properties of this problem. The expressions in equations (2.4) and (2.5) have a branch point at \( \theta = -\beta^2/4 \), and there is always a continuous spectrum in the range \( \Re(\theta) \in (-\infty, -\beta^2/4) \). For \( \beta < 0 \) the spectrum is purely continuous, but for \( \beta > 0 \) there may be any number of discrete eigenvalues. For any \( \beta > 0, \theta = 0 \) is a simple pole of equations (2.4) and (2.5) and the residue at this pole corresponds to the steady state solution in equation (2.2). As \( \beta \) increases past each root of \( D'_{\beta^2/4}(-\beta) = 0 \) (there are infinitely many such roots in the range \( \beta > 0 \)), a new discrete eigenvalue enters the spectrum.

The first root occurs when \( \beta = \beta_* \approx 1.85722 \ldots \) and thus for \( 0 < \beta < \beta_* \), \( \theta = 0 \) is the only pole of equations (2.4) and (2.5), but for \( \beta > \beta_* \) there exists at least one pole on the real axis with \( \Re(\theta) \in (-\beta^2/4, 0) \). In [16] we also studied how as \( \beta \to +\infty \) the spectrum approaches that of the standard free space Ornstein–Uhlenbeck process, whose Laplace transformed density would have poles at \( \theta = 0, -1, -2, \ldots \). Understanding the spectrum is useful in estimating the approach to equilibrium, i.e., to obtain \( p(x, t) - p(x, \infty) \) for \( t \to \infty \) and fixed \( x \) and \( x_0 \).

The expressions in equations (2.4)–(2.6) do not yield much immediate insight into the qualitative structure of the density \( p(x, t) \). Thus it is useful to evaluate these expressions in various asymptotic limits, to gain more insight, for example, as to how equilibrium is achieved when \( \beta > 0 \). Here we shall consider the limit defined in equation (1.1).

To interpret some of our results it is useful to consider a fluid approximation to the diffusion process \( X_d(t) \), where we neglect completely the “diffusive” or variability effects. For \( \beta \to +\infty \) we have \( X_d(t) \approx \beta Y_{fl}(t) \) where, since \( x_0 = \beta X_0 \) with \( X_0 > 0 \),
\[ Y_{fl}(t) = \begin{cases} X_0 - t, & 0 \leq t \leq X_0 \\ e^{x_0 - t} - 1, & t \geq X_0. \end{cases} \quad (2.7) \]

Thus the fluid approximation starts at \( X_0 \), leaves the range \( Y_{fl} > 0 \) when \( t \) increases past \( X_0 \), and then remains in the range \( Y_{fl} < 0 \) for \( t > X_0 \), ultimately approaching \( Y_{fl}(\infty) = -1 \). Then equation (2.3) gives the Gaussian spread about this limiting value.
We introduce the small parameter $\varepsilon$ with $\varepsilon = \beta^{-2}$, so that $\beta = 1/\sqrt{\varepsilon}$ and $\varepsilon \to 0^+$, and summarize our main asymptotic results in the $(X,t)$ plane, first for $X > 0$ and then $X < 0$.

**Theorem 2.2**  For $X = x/\beta = x/\sqrt{\varepsilon} > 0$ we have the following asymptotic results:

(i) $0 < t < t_1(X;X_0)$

\[
p(x,t) = e^{-(t+X-X_0)^2/(4\varepsilon t)} \sim \frac{\varepsilon}{2\sqrt{\pi t}} e^{-(X-X_0)^2/(4\varepsilon t)}. \tag{2.8}
\]

(ii) $t_1(X;X_0) < t < t_2(X;X_0)$

\[
p(x,t) = e^{-(t+X-X_0)^2/(4\varepsilon t)} \sim \frac{(1-4\varepsilon)^{1/4} [1 + \sqrt{1 - 4\varepsilon}] e^{f(z_*)/\varepsilon}}{2\varepsilon t \sqrt{1 - 4\varepsilon} - 2(X + X_0)z_*}, \tag{2.9}
\]

where

\[
f(z_*) = -\frac{1}{2} (X + X_0 + 1) \sqrt{1 - 4\varepsilon} + \frac{X - X_0}{2} \frac{X + X_0}{\sqrt{1 - 4\varepsilon}}, \tag{2.10}
\]

and $z_* = z_*(X,t)$ is defined implicitly, as the solution to

\[
t = \frac{X + X_0}{\sqrt{1 - 4\varepsilon}} + 2 \log \left( \frac{1 + \sqrt{1 - 4\varepsilon}}{2\sqrt{\varepsilon}} \right). \tag{2.11}
\]

Furthermore, $t_2$ is implicitly defined as the solution to

\[
t = \frac{(X - X_0)^2}{t} - \sqrt{\mathcal{A}} + 2 \log \left[ \frac{2(X + X_0 + 2) + t + (X - X_0)^2/t + \sqrt{\mathcal{A}}}{2(X + X_0 + 2) - t - (X - X_0)^2/t - \sqrt{\mathcal{A}}} \right], \tag{2.12}
\]

\[
\mathcal{A} = \left[ t + \frac{(X - X_0)^2}{t} \right]^2 - 4(X + X_0)(X + X_0 + 2),
\]

and $t_1$ is implicitly defined as the solution to

\[
t = \frac{(X + X_0)^2}{t} - \sqrt{\mathcal{B}} + 2 \log \left[ \frac{2(X + X_0 + 2) + t + (X + X_0)^2/t + \sqrt{\mathcal{B}}}{2(X + X_0 + 2) - t - (X + X_0)^2/t - \sqrt{\mathcal{B}}} \right], \tag{2.13}
\]

\[
\mathcal{B} = \left[ t + \frac{(X + X_0)^2}{t} \right]^2 - 4(X + X_0)(X + X_0 + 2).
\]

(iii) $t > t_2(X;X_0)$

\[
p(x,t) \sim \frac{1}{2\sqrt{2\pi}} \frac{(1-4\varepsilon)^{1/4} [1 + \sqrt{1 - 4\varepsilon}] e^{f(z_*)/\varepsilon}}{\sqrt{1 - 4\varepsilon} - 2(X + X_0)z_*}, \tag{2.14}
\]

where $z_*$ and $f$ are as in equations (2.11) and (2.10).

On the time scale $t = \log(1/\varepsilon) + O(1)$, with

\[
t = \log(1/\varepsilon) + T = 2 \log \beta + T, \quad T = O(1), \tag{2.15}
\]
we have the explicit expression
\[ p(x, t) \sim \frac{1}{\sqrt{2\pi}} e^{-(X+1/2)/\epsilon} \exp \left[ e^{-(T-X-X_0)} \right]. \] (2.16)

To interpret these results we note that the transient density for a free space Brownian motion with unit negative drift is given by
\[ p_{BM}(X, t; X_0) = \frac{1}{2\sqrt{\pi t}} \exp \left[ -\frac{(t + X - X_0)^2}{4\epsilon t} \right]. \] (2.17)
Here the process starts at \( X_0 \) when \( t = 0 \) and has \( \epsilon \) as its diffusion coefficient.

If the first PDE in equation (2.1) were to apply for all \( x \), rather than only \( x > 0 \), then \( p(x, t) \) would be given exactly by equation (2.17). Theorem 2.2 shows that there are three ranges in the \((X, t)\) plane where the behaviour of \( p(x, t) \) is different. For \( t < t_2(X; X_0) \) items (i) and (ii) (with equations (2.8) and (2.9)) show that \( p \sim p_{BM} \) with an exponentially small error. The error term undergoes a “phase transition” along the curve \( t = t_1(X; X_0) \), and we gave the results in Theorem 2.2 so as to estimate the difference \( p - p_{BM} \) for those space-time ranges where \( p \sim p_{BM} \). We also note that along the first transition curve \( t_1 \) we have
\[ -\frac{1}{4t} (t + X - X_0)^2 = f(z_*) \] (2.18)
Thus along the curve \( t = t_1 \) the right sides of equations (2.8) and (2.9) become roughly comparable, while along \( t = t_2 \) the right side of equation (2.9) becomes roughly comparable to \( p_{BM} \). In the range \( t > t_2 \) \( p_{BM} \) can no longer be used to approximate the density and we obtain the complicated expression in equation (2.14). From equation (2.11) we can show that
\[ z_*(X, t) \sim e^{-t} e^{X+X_0}, \quad t \to \infty, \] (2.19)
and then
\[ f(z_*) \sim -\frac{1}{2} - X + e^{-t} e^{X+X_0}, \quad t \to \infty. \] (2.20)
Thus if we scale time as in equation (2.15) then on the \( T = O(1) \) time scale \( f(z_*) + 1/2 + X \) becomes \( O(\epsilon) \) and equation (2.14) reduces to equation (2.16) as a special case. The latter expression shows how the steady state in equation (2.3) for \( x > 0 \) is slowly achieved on the \( T \)-scale, as \( p(x, t) \) decreases toward \( p(x, \infty) \) as a double exponential function in \( T \).

Roughly speaking, the approximation \( p \approx p_{BM} \) corresponds to sample paths that go directly from \((X_0, 0)\) to \((X, t)\), the right side of equation (2.8) corresponds to sample paths that go from \((X_0, 0)\) to \( X = 0 \) and then “reflect” at the interface before returning to \((X, t)\), and the right side of equation (2.14) corresponds to paths that go from \((X_0, 0)\), hit \( X = 0 \), spend some time in the range \( X < 0 \), and then return to \( X = 0 \) and re-enter the range \( X > 0 \). Note that the exponential part of the right side of equation (2.8) would be consistent with a reflection (in \( X = 0 \)) of the Brownian motion in equation (2.17), but the algebraic factors indicate a more complicated reflection/transmission law, through the interface \( X = 0 \).
When \( t \approx t_1 \) the right sides of equations (2.8) and (2.9) become comparable, and then \( p - p_{BM} \) can be estimated by adding the two expressions. Similarly, when \( t \approx t_2 \) we can approximate \( p \) by \( p_{BM} \) plus the right side of equation (2.14). Finally, we note that Theorem 2.2 applies not only for \( \beta \to +\infty \) with \( X = x/\beta > 0 \), but also on the original \( x \) scale, for \( x > 0 \) (then \( X = O(1/\beta) \)). In fact it can be used to compute \( p(0,t) \) asymptotically by letting \( X = 0 \).

In Figure 1, we plot the transition curves \( t_1 \) and \( t_2 \) for the three values \( X_0 = 0.2, 1, 5 \) to illustrate their dependence on the initial condition. Note that when \( X = 0 \), \( t_1(0;X_0) = t_2(0;X_0) \), as then equations (2.12) and (2.13) coincide.

We next consider the range \( X < 0 \), where the asymptotics of \( p(x,t) \) become much more complicated. It proves useful to divide \( \{ X < 0 \} \) as \( \{ X \leq -1 \} \cup \{ -1 < X < 0 \} \), and the strip \( -1 < X < 0 \) in the \((X,t)\) plane must be further subdivided into seven regions. We first define

\[
t_* = t_*(X;X_0) = \frac{X_0}{\sqrt{-2X - X^2}} + \log \left[ \frac{1 + X}{1 - \sqrt{-2X - X^2}} \right], \quad -1 < X < 0,
\]

and note that \(-2X - X^2 = 1 - (1 + X)^2 \in (0,1)\) when \( X \in (-1,0) \). We also have \( t_* \to +\infty \) as either \( X \to -1^+ \) or \( X \to 0^- \). More precisely, \( t_* \sim X_0/\sqrt{-2X} \) as \( X \to 0^- \) and \( t_* = -\log(1+X) + \log 2 + X_0 + o(1) \) as \( X \to (-1)^+ \). We furthermore define the curves
Transient distribution of the Halfin-Whitt diffusion process

\[ t_c = t_c(X; X_0) = \alpha_c + \log \left[ \frac{X + 1 + \sqrt{(X + 1)^2 + X_0^2/\alpha_c^2} - 1}{1 - X_0/\alpha_c} \right], \]  

(2.22)

where

\[ \alpha_c = \left[ \frac{X_0(X_0 + 2)}{2} - \frac{X_0^2}{2X(X + 2)} - \frac{X_0^{3/2}(X + 1)}{2X(X + 2)} \sqrt{(X^2 + 2X)(X_0 + 4) + X_0} \right]^{1/2}, \]  

(2.23)

and

\[ t_d = t_d(X; X_0) = \alpha_d + \log \left[ \frac{X + 1 + \sqrt{(X + 1)^2 + X_0^2/\alpha_d^2} - 1}{1 - X_0/\alpha_d} \right], \]  

(2.24)

where

\[ \alpha_d = \left[ \frac{X_0(X_0 + 2)}{2} - \frac{X_0^2}{2X(X + 2)} + \frac{X_0^{3/2}(X + 1)}{2X(X + 2)} \sqrt{(X^2 + 2X)(X_0 + 4) + X_0} \right]^{1/2}. \]  

(2.25)

The curves \( t_c \) and \( t_d \) exist only in the range \( X \in (X_{\text{cusp}}, 0) \) where

\[ X_{\text{cusp}} = -1 + \frac{2}{\sqrt{X_0 + 4}}. \]  

(2.26)

Note that \( ((X_{\text{cusp}} + 1)^2 - 1)(X_0 + 4) + X_0 = 0 \). The curves \( t_c \) and \( t_d \) thus coincide at the point \( (X_{\text{cusp}}, t_{\text{cusp}}) \), where

\[ t_{\text{cusp}} = \sqrt{X_0(X_0 + 3)} + \log \left[ \frac{X_0 + 2 + \sqrt{X_0(X_0 + 3)}}{\sqrt{X_0 + 4}} \right]. \]  

(2.27)

When \( X = X_{\text{cusp}} \) we have, in view of (2.27) and (2.21), \( t_{\text{cusp}} > t_* \). From equations (2.22) and (2.23) we see that \( t_c \to \infty \) as \( X \to 0^- \), and comparing the small \( X \) behaviours of \( t_c \) and \( t_* \) we can show that \( t_c > t_* \) for \( X \to 0^- \). On the other hand, \( t_d \) tends to the finite limit \( t_d(0; X_0) \) as \( X \to 0^- \), with

\[ t_d(0; X_0) = \sqrt{X_0(X_0 + 2)} + \log \left[ \frac{\sqrt{X_0 + 2} + \sqrt{X_0}}{\sqrt{X_0 + 2} - \sqrt{X_0}} \right], \]

and also \( \alpha_d \to \alpha_d(0; X_0) = \sqrt{X_0(X_0 + 2)} \). We use the term “cusp” to denote the intersection of \( t_c \) and \( t_d \) since these curves have the common slope \( dt/dX = \sqrt{(X_0 + 3)(X_0 + 4)}/X_0 \) at this point. We can also view the union of \( t_c \) and \( t_d \) as representing a single curve with a cusp point at \( (X_{\text{cusp}}, t_{\text{cusp}}) \). In Section 5, we give more geometric interpretation to \( t_c \) and \( t_d \), since they correspond to the envelope(s) of a certain family of curves, and thus \( t_c \) and \( t_d \) are called “caustic” curves.

In Figure 2, we sketch the curves \( t_*, t_c \) and \( t_d \) in the range \( X \in (-1, 0) \) for \( X_0 = 1 \). This shows that \( t_c \) (which exists only for \( X > X_{\text{cusp}} \)) lies above \( t_* \), while \( t_d \) intersects \( t_* \) at a point, say \( (X_*, t_{\text{cusp}}) \), which may be obtained numerically by simultaneously solving equations (2.21) and (2.24) with equation (2.25). We also sketch (the dashed line) the fluid approximation in equation (2.7), which is the curve \( X = e^{X_0-t} - 1 \in (-1, 0) \), or
Figure 2. The curves \( t_*, t_c, t_d \) and the fluid approximation (the dashed curve) for \( X_0 = 1 \).

\[ t = X_0 - \log(1 + X), \] and this lies below each of the curves \( t_*, t_c, t_d \) (when the latter two exist). While the figure fixes \( X_0 = 1 \), the pattern of intersection is the same for any \( X_0 > 0 \). Hence the three curves \( t_*, t_c, t_d \) split the strip \(-1 < X < 0\) into the following seven regions:

Region I = \( \{ X \leq -1 \} \cup \{ 0 < t < t_*, -1 < X < X_{\text{cusp}} \} \),
Region II = \( \{ t > t_*, -1 < X < X_{\text{cusp}} \} \),
Region III = \( \{ 0 < t < \min\{t_*, t_d\}, X_{\text{cusp}} < X < 0 \} \),
Region IV = \( \{ t_d < t < t_*, X_{\text{cusp}} < X < 0 \} \),
Region V = \( \{ \max\{t_*, t_d\} < t < t_*, X_{\text{cusp}} < X < 0 \} \),
Region VI = \( \{ t > t_c, X_{\text{cusp}} < X < 0 \} \),
Region VII = \( \{ t_* < t < t_d, X_{\text{cusp}} < X < X_* \} \).

It will prove convenient to include \( \{ X \leq -1 \} \) within Region I. Note that \( \min\{t_*, t_d\} = t_* \) if \( X < X_* \) and \( \min\{t_*, t_d\} = t_d \) if \( X > X_* \). The different regions are indicated in Figure 3. The fluid approximation lies within Region I if \( X < X_{\text{cusp}} \) and within Region III if \( X \in (X_{\text{cusp}}, 0) \). The different regions will correspond to different asymptotic formulas for \( p(x, t) \).

Our results will involve the solutions of certain transcendental equations, similar to the one in equation (2.11), so we state two lemmas about these equations.
Figure 3. The curves $t_*, t_c$ and $t_d$ split the strip $-1 < X < 0$ into seven regions (the dashed curve is the fluid approximation).

**Lemma 1** For $t > 0$ and $X < 0$ the equation

$$
t - \frac{X_0}{\sqrt{1 + 4\phi_s}} + \log \left[ \frac{\sqrt{1 + 4\phi_s} - 1}{\sqrt{(1 + X)^2 + 4\phi_s} - 1 - X} \right] = 0, \quad \phi_s > -\frac{1}{4}(1 + X)^2
$$

has a unique solution $\phi_s = \phi_s(X,t)$ for $(X,t)$ in Regions I, III and IV. Also, $\phi_s > 0$ if $X \leq -1$ or $X \in (-1,0)$ and $t < X_0 - \log(1 + X)$ (i.e., below the fluid approximation). When $t = X_0 - \log(1 + X)$ we have $\phi_s = 0$. When $X_0 - \log(1 + X) < t < t_*(X;X_0)$ we have $\phi_s < 0$, and $\phi_s \to -(1 + X)^2/4$ as $t \uparrow t_*$. In Regions II and V–VII, equation (2.35) has no solutions in the range $\phi_s > -(1 + X)^2/4$.

**Lemma 2** For $t > 0$ and $-1 < X < 0$ the equation

$$
t - \frac{X_0}{\sqrt{1 - 4z}} + \log \left[ \frac{1 - \sqrt{1 - 4z}}{1 + X + \sqrt{(1 + X)^2 - 4z}} \right] = 0, \quad z > 0
$$

has a unique solution for $(X,t)$ in Regions II and VI which we denote by $z_1 = z_1(X,t)$. In Region V, equation (2.36) has three solutions that we order as $z_1 < z_2 < z_3$. In Region VII, there is again a unique solution that we now denote by $z_3 = z_3(X,t)$. In Region IV there are two solutions that we order as $z_1 < z_2$, and they lie in the range $z \in (0, -\phi_s)$, where $\phi_s$ is the solution to equation (2.35). For Regions II and V–VII, these conclusions consider
equation (2.36) over the “full” range $z \in (0, (1 + X)^2/4)$. For Regions I and III, equation (2.36) has no solutions in the range $z \in (0, -\phi_s)$ if $\phi_s < 0$.

Note that if we consider the totality of the solutions to equations (2.35) and (2.36), this number is 1 for Regions I–III, VI and VII, and three for Regions IV and V. Thus there are a total of 3 solutions precisely when $t_d < t < t_c$ (for $X \in (X_{\text{cusp}}, 0)$). We can now state our main results for $X < 0$.

Theorem 2.3 For $X = x/\beta = x/\sqrt{\varepsilon} < 0$ we have the following asymptotic results:

(i) $(X, t)$ in Regions I and III

$$p(x, t) \sim G(X, t) \exp \left[ \frac{1}{\varepsilon} F(X, t) \right],$$

where

$$F(X, t) = -X^2/4 - X/2 + X_0/2 + \frac{X_0 \phi_s}{\sqrt{1 + 4\phi_s}} - \left( \frac{X_0}{2} + \frac{1}{4} \right) \sqrt{1 + 4\phi_s}$$

$$+ \frac{1 + X}{4} \sqrt{(1 + X)^2 + 4\phi_s},$$

and

$$G(X, t) = \frac{1}{\sqrt{2\pi}} \left( \frac{1 + 4\phi_s}{1 + 4\phi_s} \right)^{-1/4} \left[ \frac{\sqrt{1 + 4\phi_s} + 1}{\sqrt{(1 + X)^2 + 4\phi_s} + 1 + X} \right]^{1/2}$$

$$\times \left[ \frac{2X_0}{(1 + 4\phi_s)^{3/2}} + \frac{1}{2\phi_s} \left( \frac{1}{\sqrt{1 + 4\phi_s}} - \frac{1 + X}{\sqrt{(1 + X)^2 + 4\phi_s}} \right) \right]^{-1/2},$$

(2.39)

and $\phi_s = \phi_s(X, t)$ is the unique solution to equation (2.35). For $(X, t)$ close to the fluid approximation, we let $X = -1 + e^{-x_0-t} + \sqrt{\varepsilon} \Delta$. Then for $\varepsilon \to 0$ and $\Delta$ fixed, equation (2.37) simplifies to the Gaussian approximation

$$p(x, t) \sim \frac{1}{\sqrt{2\pi} \sqrt{2X_0 + e^{2(t-x_0)} - 1}} \exp \left[ - \frac{e^{2(t-x_0)}}{e^{2(t-x_0)} - 1 + 2X_0} \Delta^2 \right].$$

(2.40)

For $t$ large, with $t = \log(1/\varepsilon) + T$ as in equation (2.15), and $X < -1$, equation (2.37) simplifies to

$$p(x, t) \sim \frac{1}{\sqrt{2\pi}} \exp \left[ - \frac{(1 + X)^2}{2\varepsilon} \right] \exp \left[ e^{x_0 - T(1 + X)} \right],$$

(2.41)

which gives the approach to equilibrium as $T \to \infty$. For $t \approx t_*$ (the upper limit for Region I if $X \in (-1, X_{\text{cusp}})$ and Region III if $X \in (X_{\text{cusp}}, X_*)$), equation (2.37) simplifies to

$$p(x, t) \sim \frac{\sqrt{1 + \sqrt{1 - (1 + X)^2}}}{2\sqrt{\pi} [1 - (1 + X)^2]^{1/4}} \exp \left[ \frac{F_e(X)}{\varepsilon} - \frac{(1 + X)^2}{4\varepsilon^{2/3}} A - \frac{(1 + X)^2}{12} A^3 \right],$$

(2.42)
where
\[ t = t_* + \varepsilon^{1/3} A, \quad A = O(1), \tag{2.43} \]

and
\[ F_e(X) = -\frac{X^2}{4} - \frac{X}{2} + \frac{X_0}{2} - \frac{X_0}{4\sqrt{-X^2 - 2X}} - \left(\frac{X_0}{4} + \frac{1}{4}\right)\sqrt{-X^2 - 2X}. \tag{2.44} \]

(ii) \((X,t)\) in Regions II and VI

\[ p(x,t) \sim \tilde{G}(z_1) \exp\left[\frac{1}{\varepsilon} \tilde{f}(z_1)\right], \tag{2.45} \]

\[ \tilde{f}(z) = \tilde{f}(z;X,t) = -\frac{X^2}{4} - \frac{X}{2} + \frac{X_0}{2} - \frac{zX_0}{\sqrt{1 - 4z}} - \left(\frac{X_0}{2} + \frac{1}{4}\right)\sqrt{1 - 4z} \]
\[ - \frac{1 + X}{4\sqrt{(1 + X)^2 - 4z}}, \]

\[ \tilde{G}(z) = \tilde{G}(z;X,t) = \frac{1}{\sqrt{2\pi}} \frac{(1 - 4z)^{-1/4}}{[(1 + X)^2 - 4z]^{1/4}} \left[\frac{1 + \sqrt{1 - 4z}}{1 + X - \sqrt{(1 + X)^2 - 4z}}\right]^{1/2} \]
\[ \times \left[-\frac{2X_0}{(1 - 4z)^{3/2}} + \frac{1}{2z} \left(\frac{1}{\sqrt{1 - 4z}} + \frac{1 + X}{\sqrt{(1 + X)^2 - 4z}}\right)\right]^{-1/2}, \]

and \(z_1 = z_1(X,t)\) is the unique solution to equation (2.36) in the range \(z \in (0, (1 + X)^2/4)\). For \(t \to \infty\), with \(t = \log(1/\varepsilon) + T\), equation (2.45) simplifies to equation (2.41), so that equation (2.41) applies both to \(X < -1\) and \(X \in (-1,0)\).

(iii) \((X,t)\) in Region VII

\[ p(x,t) \sim \tilde{G}(z_3) \exp\left[\frac{1}{\varepsilon} \tilde{f}(z_3)\right], \tag{2.46} \]

where now \(z_3\) is the unique solution to equation (2.36).

(iv) \((X,t)\) in Region V

\[ p(x,t) \sim \tilde{G}(z_1) \exp\left[\frac{1}{\varepsilon} \tilde{f}(z_1)\right] + \tilde{G}(z_3) \exp\left[\frac{1}{\varepsilon} \tilde{f}(z_3)\right], \tag{2.47} \]

where \(z_1\) and \(z_3\) are the minimal and maximal of the three solutions to equation (2.36).

The equation \(\tilde{f}(z_1(X,t);X,t) = \tilde{f}(z_3(X,t);X,t)\) defines a curve \(t_F\) in the \((X,t)\)-plane within Region V, that satisfies \(t_d < t_F < t_c\). For \(t_d < t < t_F\), the term involving \(z_1\) in equation (2.47) dominates, while for \(t_F < t < t_c\) the term involving \(z_3\) dominates. The three curves \(t_c, t_d\) and \(t_F\) all go through the cusp point \((X_{cusp},t_{cusp})\).

(v) \((X,t)\) in Region IV

\[ p(x,t) \sim G(X,t) \exp\left[\frac{1}{\varepsilon} F(X,t)\right] + \tilde{G}(z_1) \exp\left[\frac{1}{\varepsilon} \tilde{f}(z_1)\right], \tag{2.48} \]

where \(z_1\) is the unique solution of equation (2.36).
The equation $F(X,t) = \tilde{f}(z_1(X,t);X,t)$ defines a curve $t_\Gamma$ within Region IV such that the term in equation (2.48) involving $F$ is the dominate one for $t < t_\Gamma$, while the term involving $\tilde{f}(z_1)$ dominates for $t > t_\Gamma$. The curves $t_\Gamma$ in Regions V and IV are smooth continuations of one another, and as $X \to 0^-$ in Region IV, $t_\Gamma$ coincides with $t_1(0;X_0) = t_2(0;X_0)$ in Theorem 2.2.

(vi) $(X,t) \approx (X_{\text{cusp}},t_{\text{cusp}})$

Introducing the scaled variables $(\xi,\eta)$ with

$$X = X_{\text{cusp}} + \sqrt{\epsilon} \xi = \left( -1 + \frac{2}{\sqrt{X_0 + 4}} \right) + \sqrt{\epsilon} \xi,$$

$$t = t_{\text{cusp}} + \sqrt{\frac{(X_0 + 4)(X_0 + 3)}{X_0}} \sqrt{\epsilon} \xi + \epsilon^{3/4} \eta,$$ (2.49)

we have

$$p(x,t) \sim \frac{\epsilon^{-1/4}}{2\pi} \sqrt{\frac{X_0 + 3}{X_0}} \sqrt{\frac{\sqrt{X_0} \sqrt{X_0 + 3} + X_0 + 2}{\epsilon^2 \frac{\sqrt{X_0}}{4(X_0 + 4)}} \exp \left( -\frac{\xi^2 \sqrt{X_0 + 2\sqrt{X_0 + 3}}}{4\sqrt{X_0}} \right)} \times \exp \left( -\frac{\xi^2 \sqrt{X_0 + 2\sqrt{X_0 + 3}}}{4\sqrt{X_0} \sqrt{X_0 + 3} + X_0} \right) \exp \left( \frac{2(X_0 + 3)^{3/2} + \sqrt{X_0}(2X_0 + 9)}{\epsilon^2 \frac{(X_0 + 3)}{4(X_0 + 4)}} \right) \times J(\xi,\eta),$$ (2.50)

where $J$ has the integral representation

$$J(\xi,\eta) = \int_{-\infty}^{\infty} \exp \left( -\frac{\eta w + \left( \frac{(X_0 + 4)(X_0 + 3)}{X_0} \right)^{3/2} \xi w^2 - \frac{2(X_0 + 3)^{11/2}}{3 X_0^{5/2}} w^4}{\epsilon^{1/4}} \right) dw.$$ (2.51)

The asymptotic structure of $p(x,t)$ is not quite as complicated as Theorem 2.3 suggests, for the following reasons. In equations (2.28)–(2.34) we defined the various regions as open sets, which exclude the curves that separate pairs of regions. For example, the curve $t = t_*$ for $X < X_{\text{cusp}}$, which separates Regions I and II, is not contained in either region. This may suggest that another expansion is needed for $t \approx t_*$ but this is in fact not the case. If we expand equation (2.37) for $t \uparrow t_*$ (and $X < X_{\text{cusp}}$) we obtain the expression in equation (2.42) for $t_* - t = O(\epsilon^{1/3})$. But expanding the Region II result in equation (2.45) for $t \downarrow t_*$ (and $X < X_{\text{cusp}}$) again leads to the expression in equation (2.42), which is a smooth function of the variables $A$ and $X$. Hence equation (2.45) is smooth through the curve $t = t_*$ and the asymptotics of $p(x,t)$ do not undergo a phase transition here. Similarly, going from Region III to Region VII through the curve $t = t_*$ (with $X_{\text{cusp}} < X < X_\ast$) does not lead to a phase transition, as equation (2.46) is the smooth continuation of equation (2.37). Also, going from Regions IV to V through $t = t_*$ (with $X_\ast < X < 0$), does not lead to a phase transition as $\tilde{G}(z_3)e^{T(z_3)/\epsilon}$ is the natural continuation of $Ge^{F/\epsilon}$. Thus there are no transitions near $t = t_*$ for any $-1 < X < 0$. 

Now suppose we fix $X \in (X_{cusp}, X_*)$ and increase $t$, thus going from Region III to VII to V and finally VI. As mentioned above going from III to VII is just a smooth continuation. Going from VII to V leads to the appearance of the term involving $z_1$ in equation (2.47). However, along $t = t_d$ the term involving $z_1$ is the dominant one asymptotically, as $\tilde{f}(z_3) > \tilde{f}(z_1)$, so the transition involves only the exponentially small error term. Going from V to VI leads to the disappearance of the term involving $z_3$, but by then the term involving $z_1$ dominates, and this occurs precisely as $t$ increases past $t_F$, by definition. Thus along $t \approx t_c$ the phase transition again involves only the exponentially small correction term, which now arises from $z_3$.

If we fix $X \in (X_*, 0)$ and increase $t$ we go from Region III to IV to V and finally VI. Going from III to IV leads to the birth of the term involving $z_1$ in equation (2.48), but for $t \approx t_d$ this is exponentially smaller than the first term in equation (2.48). Going from IV to V involves crossing $t_c$ and this is a smooth continuation that we already discussed. Finally, going from V to VI again leads to the disappearance of the $z_3$ term, which is by then dominated by $z_1$, as was the case for $X \in (X_{cusp}, X_*)$.

This discussion shows that the only true transition in the leading term for $p(x, t)$ occurs along the curve $t_F$, as when $t$ increases past this curve $\tilde{G}(z_1)e^{f(z_1)/\epsilon}$ begins to dominate $\tilde{G}(z_3)e^{f(z_3)/\epsilon}$ or $Ge^{F/\epsilon}$ (as the latter two are smooth continuations of one another). Then the only other non-uniformity or transition in the asymptotics of $p(x, t)$ occurs in a small neighbourhood of the cusp point $(X_{cusp}, t_{cusp})$, defined precisely via the scaling in equation (2.49), where the approximation for $p(x, t)$ involves the complicated integral in equation (2.51), in terms of the local $(\xi, \eta)$ coordinate system.

In Section 4 we shall establish Theorem 2.3 (and also Lemmas 1 and 2) by using the Laplace transform in equation (2.5) and the integral in equation (2.6), which we shall expand using the saddle point method and singularity analysis. Then the curves $t_r$, $t_c$ and $t_d$ will naturally come out of that analysis. However, in Section 5 we shall briefly discuss an alternate, more geometrical approach. This will regain all of the results in Theorem 2.3, albeit in a slightly different form. The alternate approach will make it clear that only the exterior and interior (defined by $t_d < t < t_c$ for $X_{cusp} < X < 0$) of the cusped region need to be considered separately. Thus Regions I, II, III, VI and VII will lead to one result for $p(x, t)$, while regions IV and V will need a different analysis.

In Figure 4 we indicate the curves $t_c, t_d$ and $t_F$, for the three initial conditions $X_0 = 1, 2$ and 5. Finally we comment that to obtain results for the scale $X = O(\sqrt{\epsilon})$ (thus $x = O(1)$) we can simply expand the Regions III and IV results for small $X$. When $X = 0$ the Region III expression agrees with the approximation $p_{BM}(0, t; X_0)$ in equation (2.17), and this remains true in Region IV as long as $t < t_F|_{X=0}$. For $t > t_F|_{X=0}$ setting $X = 0^-$ in equation (2.48), where now $p(0, t) \sim \tilde{G}(z_1)e^{f(z_1)/\epsilon}|_{X=0}$, we obtain equation (2.14) with $X = 0^+$ (and $z_* = z_*(0^+, t)$). Thus all the asymptotic approximations are continuous along $X = 0$.

Theorems 2.2 and 2.3 apply only for a fixed $X_0 > 0$, and break down if $X_0 \to 0^+$. Note that the cusp point $(X_{cusp}, t_{cusp}) \to (0, 0)$ as $X_0 \to 0$ and then a completely separate asymptotic analysis is needed, which we do not attempt here. Theorem 2.1 holds for any $x_0 \geq 0$ but the asymptotics of $p(x, t)$ will be very different if $X_0 = x_0/\beta$ is small. When $x_0 < 0$ a different expression for the Laplace transform must be used (see [16]) and again the asymptotics of $p(x, t)$ will be completely different.
Here we establish the asymptotic results in Theorem 2.2, which apply for \( X > 0 \) (thus \( x > 0 \)) and \( X = 0 \). We shall use equation (2.4) in equation (2.6) and expand the resulting contour integral by the saddle point method and singularity analysis. General reference on methods for asymptotically evaluating integrals and sums can be found in [2, 3, 27]. We shall also need to use various properties of the parabolic cylinder functions \( D_{-\theta}(-) \), and they are summarized in [1, 12, 18, 25].

Since we are taking \( \beta \to +\infty \), the function in equation (2.4) will have many poles in the range \( \theta \) real and \( \theta \in (-\beta^2/4, 0) \), in addition to the branch cut along \( \Im(\theta) = 0, \Re(\theta) \in (-\infty, -\beta^2/4) \). Let us denote by \( \theta = \theta_N \) the \( N \)th pole of equation (2.4), with \( \theta_0 = 0 \) corresponding to the steady state limit \( p(x, \infty) \), and \( \theta_N < 0 \) with \( 0 < |\theta_1| < |\theta_2| < \cdots \). The number of poles in the range \((-\beta^2/4, 0)\) is certainly finite for finite \( \beta \), and in view of
equation (2.4) and the definition of $R_\beta(\theta)$, the poles are the solutions of

$$\sqrt{\theta + \frac{\beta^2}{4}} = R_\beta(\theta) = \frac{D'_{-\theta}(-\beta)}{D_{-\theta}(-\beta)}, \quad \frac{\beta^2}{4} < \theta \leq 0. \quad (3.1)$$

Note that $D_{-\theta}(-\beta)$ is well known to be an entire function of $\theta$ (and also $\beta$, see [18]) and that zeros of $D_{-\theta}(-\beta) = 0$ are not poles of the $\hat{p}(x, \theta)$, again by equation (2.4). We have $D_\beta(-\beta) = e^{-\beta^2/4}$ so that $\theta = 0$ is a root of equation (3.1) for any $\beta > 0$. Note also that $D'_{-\theta}(z)$ is the derivative of the function with respect to its argument $z$, hence $D'_{-\theta}(-\beta) = -\frac{\partial}{\partial \beta} D_{-\theta}(-\beta)$.

In [16] we estimated the poles $\theta_N$ for $\beta \to +\infty$ and $N$ fixed, showing that $\theta_N \approx -N$ with an exponentially small error

$$\theta_N = -N + \frac{e^{-\beta^2/2}}{\sqrt{2\pi(N-1)!}} \beta^{2N-3} \left[1 + o(1)\right], \quad \beta \to \infty,$$

$$= -N + \frac{e^{3/2-N}}{(N-1)!} \exp \left(-\frac{1}{2e}\right) \left[1 + o(1)\right], \quad \epsilon \to 0. \quad (3.2)$$

This applies also to $N = 0$ since $\theta_0 = 0$ exactly. The present analysis, however, will require computing the contribution(s) from poles $\theta_N$ for large values of $N$, and we shall see that a natural scaling will have $N = O(\epsilon^{-1}) = O(\beta^2)$.

Using equations (2.4) and (2.6) we can decompose $p(x, t)$ into the two parts $p(x, t) = p_1(x, t) + p_2(x, t)$, with

$$p_1(x, t) = \frac{1}{2\pi i} \int_{Br} e^{\theta t} e^{\theta(x_0-x)/2} e^{-\sqrt{\theta^2 + \beta^2}/4} \sqrt{\beta^2 + 4\theta} \, d\theta = \frac{1}{2\sqrt{\pi t}} e^{-(t+X-X_0)^2/(4\epsilon t)}, \quad (3.3)$$

where we set $(x, x_0) = (\beta(X, X_0))$ and $\beta^2 = \epsilon^{-1}$, and

$$p_2(x, t) = \frac{1}{2\pi i} \int_{Br} e^{\theta t} e^{\theta(x_0-x)/2} e^{-(x+x_0)\sqrt{\theta^2 + \beta^2}/4} H(\theta) \, d\theta,$$

$$H(\theta) = -\frac{1}{\sqrt{\beta^2 + 4\theta} + \sqrt{\theta^2 + \beta^2}/4 - R_\beta(\theta)}. \quad (3.4)$$

We again have $\Re(\theta) > 0$ on the $Br$ contours, so these lie to the right of all poles. The contour integral in equation (3.3) may be explicitly evaluated and the result is precisely the density of the free space Brownian motion in equation (2.17). The term $p_2$ gives the derivation of $p$ from $p_{BM}$, and represents the effects of the interface $X = 0$. Note that the first term in the expression for $H(\theta)$ would correspond to absorption at $X = 0$, but the second part of $H$ indicates a much more complicated reflection/transmission law, that involves the parabolic cylinder functions.

We can obtain an alternate representation of $p_2(x, t)$ by shifting the contour $Br$, on which $\Re(\theta) > 0$, into another vertical contour $Br(\theta_{sa})$ on which $-\beta^2/4 < \Re(\theta_{sa}) < 0$. But in shifting we must take into account the residues from the poles. We thus have the following:
Lemma 3 The difference \( p(x, t) - p_{BM}(x, t) = p_2(x, t) \) between the density of the present diffusion and a free space Brownian motion with drift has the alternate representation

\[
p_2(x, t) = \sum_{N=0}^{[-\theta_{sa}]} e^{\theta_{sa} t} e^{\beta(x_0-x)/2} e^{-(x+x_0)\sqrt{\theta_{sa} + \beta^2/4}} h_N \\
+ \frac{1}{2\pi i} \int_{Br(\theta_{sa})} e^{\theta t} e^{\beta(x_0-x)/2} e^{-(x+x_0)\sqrt{\theta + \beta^2/4}} H(\theta) d\theta \\
\equiv \text{SUM} + \text{INT},
\]

where \( Br(\theta_{sa}) \) is a vertical contour on which \( \Re(\theta_{sa}) \in (-\beta^2/4, 0) \), \([\cdot]\) is the greatest integer function, and

\[
h_N = \lim_{\theta \to \theta_N} \left\{ \left( \frac{\theta - \theta_N}{\sqrt{\beta^2 + 4\theta}} \right)^{-1} \left[ -\frac{1}{\sqrt{\beta^2 + 4\theta}} + \frac{1}{\sqrt{\beta^2/4 - R_\beta(\theta)}} \right] \right\} \\
= \left[ \frac{1}{\sqrt{\beta^2 + 4\theta}} - \frac{d}{d\theta} R_\beta(\theta) \right]^{-1} \bigg|_{\theta = \theta_N}
\]

is the residue of \( H(\theta) \) at the pole \( \theta_N \), where \( 0 = \theta_0 > \theta_1 > \theta_2 > \cdots \).

By choosing \( \theta_{sa} \) to be the saddle point of the integrand, the integral in equation (3.5) may be easily evaluated for \( \beta \to +\infty \), and then we need to estimate \( h_N \) for \( \beta \) large and evaluate the residue sum by a Laplace type method. Getting the asymptotics of \( p \) then involves determining whether the integral in equation (3.5) dominates the sum, and then whether \( p_1 \) dominates \( p_2 \). Note that for short times the approximation \( p \sim p_{BM} \) certainly holds, so that \( p_1 \gg p_2 \), but for \( t \to \infty \), \( p_1 = p_{BM} \to 0 \) and then \( p_2 \gg p_1 \), as \( p_2 \) must approach the steady state limit in equation (2.2) or (2.3). In fact the steady state limit comes from the pole \( \theta_0 = 0 \), so that for sufficiently large time, the residue sum in equation (3.5) must dominate the integral, and the summand with \( N = 0 \) must ultimately dominate the other terms in the sum. The details of these asymptotic transitions will ultimately lead to Theorem 2.2.

We first summarize the asymptotics of the integral and the sum in equation (3.5) below as two lemmas. The proofs are deferred to the Appendix.

Lemma 4 The integral in equation (3.5) has the following asymptotic expansion:

\[
\text{INT} \sim \frac{\varepsilon e^{(X_0-X)/2t}}{4\sqrt{\pi} \sqrt{X + X_0}} \frac{\sqrt{1 + 4\phi^{**}} + 1}{\left[ 1 + 4\phi^{**} \right]^{3/4}} \exp \left\{ \frac{1}{\varepsilon} \left[ \phi^{**} t - \frac{X + X_0}{2} \sqrt{1 + 4\phi^{**}} \right] \right\},
\]

where

\[
\phi^{**} = \phi^{**}(X, t) = \frac{(X + X_0)^2}{4t^2} - \frac{1}{4}.
\]

We note that the expression in equation (3.7) is the same as the right side of equation (2.8), since \( \sqrt{1 + 4\phi^{**}} = (X + X_0)/t \). The saddle \( \phi^{**} \leq 0 \) for \( X + X_0 \leq t \). When \( t < X + X_0 \), \( \theta_{sa} > 0 \) and the sum in equation (3.5) is absent, then equation (3.7) gives the leading term
for \(p_2(x, t)\). For \(t > X + X_0\) we must estimate the residue sum in equation (3.5). Note also that the right side of equation (2.8) (or (3.7)) is always exponentially smaller than \(p_{BM}\). The two become comparable only along \(X = 0\) (more precisely for \(X = O(\varepsilon)\)), but then \(p_{BM}\) is still larger since equation (3.7) contains an additional \(O(\varepsilon)\) algebraic factor that is absent from \(p_{BM}\).

**Lemma 5** For \(t > X + X_0\) the sum in equation (3.5) has the following asymptotic expansions:

(i) \(X + X_0 < t < t_+(X; X_0)\)

\[
\text{SUM} \sim \frac{\sqrt{\varepsilon}}{4\pi} \frac{t (t + X + X_0)^{3/2}}{\sqrt{t - X - X_0}} \frac{1}{(X + X_0)^2} e^{f(\varepsilon\tilde{N})/\varepsilon} \equiv \text{ENDPOINT},
\]

where

\[
f(\varepsilon\tilde{N}) = -\varepsilon t\tilde{N} - \frac{X + X_0 + 1}{2} \sqrt{1 - 4\varepsilon\tilde{N}} + \frac{X_0 - X}{2} + 2\varepsilon\tilde{N} \log \left( \frac{1 + \sqrt{1 - 4\varepsilon\tilde{N}}}{2\sqrt{\varepsilon\tilde{N}}} \right),
\]

and

\[
t_+(X; X_0) = \sqrt{(X + X_0)(X + X_0 + 2)} + 2 \log \left( \frac{\sqrt{X + X_0 + 2} + \sqrt{X + X_0}}{\sqrt{2}} \right).
\]

(ii) \(t > t_+(X; X_0)\)

\[
\text{SUM} \sim \text{ENDPOINT} + \frac{\sqrt{2\pi} g(z_*)}{\sqrt{-f''(z_*)}} e^{f(z_*)/\varepsilon},
\]

where \(z_*\) is the unique solution to

\[
t = \frac{X + X_0}{\sqrt{1 - 4z_*}} + 2 \log \left( \frac{1 + \sqrt{1 - 4z_*}}{2\sqrt{z_*}} \right).
\]

(iii) \(t \approx X + X_0\) with \(t = X + X_0 + \varepsilon\tilde{t}, \tilde{t} = O(1)\), and \(M = M(X, \tilde{t}) = [\tilde{t}/(2(X + X_0))]\),

\[
\text{SUM} \sim \frac{e^{-M}}{\sqrt{2\pi} M!} \exp \left[ -\frac{1}{\varepsilon} \left( X + \frac{1}{2} \right) \right].
\]

To obtain the leading term for \(p_2 = (p - p_{BM})\) we must add equation (3.7) to the results in Lemma 5. For \(t < X + X_0\), only the integral is present. For \(X + X_0 < t < t_+\), \(p_2\) will be asymptotic to the sum of equations (3.7) and (3.8), but the former dominates since \(- (t - X - X_0)^2/(4t) > f(-\phi_*)\) is equivalent to

\[
X + \frac{X + X_0}{2t} > \frac{1}{2} \left[ 1 - \frac{(X + X_0)^2}{t^2} \right] \log \left[ \frac{t + X + X_0}{\sqrt{t^2 - (X + X_0)^2}} \right],
\]

(3.12)
and equation (3.12) is obviously true in view of the inequality \( \xi^{2m+1} > \xi^{2m+1}/(2m+1) \) 
\((m = 1, 2, 3, \ldots)\), which when summed over \( m \) yields

\[
\frac{2\xi}{1 - \xi^2} > \sum_{m=0}^{\infty} \frac{2\xi^{2m+1}}{2m+1} = \log \left( \frac{1 + \xi}{1 - \xi} \right), \quad 0 < \xi < 1.
\]

Setting \( \zeta = (X + X_0)/t \) we see that the left side of equation (3.12) dominates the right side, even without the \( X \) term in the former. Hence \( p_2 \) can be asymptotic to either equation (3.7), which must be true for \( t < X + X_0 \), or to the \( z_\ast \) contribution in equation (3.10), which must occur for very large times \( t \). The two contributions are comparable when

\[
-\frac{X - (t - X - X_0)^2}{4t} = f(z_\ast),
\]

which can only occur for \( t > t_+ \). In view of equation (2.10) this equation is equivalent to

\[
\frac{t}{4} + \frac{(X + X_0)^2}{4t} = \frac{X + X_0 + 1}{2} \sqrt{1 - 4z_\ast} + \frac{X + X_0}{\sqrt{1 - 4z_\ast}} z_\ast.
\]  

Setting \( U = \sqrt{1 - 4z_\ast} \) so that \( z_\ast = (1 - U^2)/4 \), equation (3.13) becomes the quadratic equation

\[
(X + X_0 + 2)U + \frac{X + X_0}{U} = t + \frac{(X + X_0)^2}{t},
\]

whose solution is

\[
U = \frac{1}{2(X + X_0 + 2)} \left[ t + \frac{(X + X_0)^2}{t} + \sqrt{\mathcal{B}} \right],
\]  

where \( \mathcal{B} \) is defined below equation (2.13). Then using equation (3.14) in (2.11) yields the equation in (2.13) for the space time curve \( t_t(X; X_0) \), which is the lower curve in Figure 1. We have thus established items (i) and (ii) in Theorem 2.2. To establish item (iii) we define a second transition curve \( t_2(X; X_0) \), along which

\[
-\frac{(t + X - X_0)^2}{4t} = f(z_\ast),
\]

and this leads to equation (2.12). Then for \( t > t_2 \), \( p(x, t) \) becomes asymptotic to the term in equation (3.10) arising from \( z_\ast \). In Section 5, we give more geometric interpretation to the various terms in the expansion of \( p(x, t) \), and of the curve \( t_\ast \). This completes the analysis for fixed \( t = \mathcal{O}(1) \).

Now suppose we consider large time scales \( t \approx \log(1/\epsilon) \), setting \( T = t - \log(1/\epsilon) = \mathcal{O}(1) \), as in equation (2.15). Now \( [-\phi_{\ast\ast}/\epsilon] \sim 1/4 \) and we have \( e^{-N} e^{-N t} = e^{-N T} = \mathcal{O}(1) \). Then the asymptotics of the sum in equation (3.5) may be obtained by summing the \( N = \mathcal{O}(1) \) result in equation (2.14), thus obtaining

\[
p(x, t) \sim \frac{1}{\sqrt{2\pi}} \exp \left[ -\frac{1}{\epsilon} \left( X + \frac{1}{2} \right) \right] \sum_{N=0}^{\infty} \frac{e^{-N(T - X - X_0)}}{N!}.
\]  

Evaluating the series in equation (3.15) gives the double exponential (in \( T \)) approximation in equation (2.16), and for \( T \to \infty \) \((t - \log(1/\epsilon) \to \infty)\) this reduces to the steady state formula for \( X \) (or \( x \)) \( > 0 \) in equation (2.3). On the \( T \)-scale the sum in equation (3.5) remains a discrete sum and every pole contributes roughly equally to its asymptotic behaviour. Though the argument leading to the evaluation of equation (3.5) is different for the \( t \) and \( T \) time scales, the result in equation (2.14) actually reduces to equation
(2.16), for large times \( t \). This is easily seen by expanding \( z^* \) and \( f(z^*) \) for \( t \to \infty \), which leads to equations (2.19) and (2.20). This completes the derivation of Theorem 2.2.

4 Analysis for \( x < 0 \)

We shall obtain Theorem 2.3 by using equation (2.5) in equation (2.6) and again expanding the integral for \( \beta \to +\infty \) with \( x_0 = \beta X_0 = O(\beta) \) (and \( X_0 > 0 \)). The analysis for \( x < 0 \) is somewhat more complicated than for \( x > 0 \), and equation (2.5) involves the special functions \( D_{-\theta} (\cdot) \), which we have to approximate in various ranges of \( \theta, \beta \) and \( x \).

First we let, as in Section 3, \( \theta_{sa} \) have any real value in the range \((-\beta^2/4, \infty)\), and obtain from equation (2.6) the following alternate representation.

**Lemma 6** For \( x < 0 \), \( p(x, t) \) is given by

\[
p(x, t) = \sum_{N=0}^{[-\theta_{sa}]} e^{\theta_N t} e^{-x^2/4 + \beta(x_0 - x)/2} e^{-x_0 \sqrt{\theta_N + \beta^2/4} \frac{D_{-\theta_N}(-\beta - x)}{D_{-\theta_N}(-\beta)}} h_N + \frac{1}{2\pi i} \int_{B(\theta_{sa})} e^{\theta t} e^{-x^2/4 + \beta(x_0 - x)/2} e^{-x_0 \sqrt{\theta + \beta^2/4} - R_\beta(\theta)} \frac{D_{-\theta}(-\beta - x)}{D_{-\theta}(-\beta)} d\theta, \tag{4.1}
\]

where \( h_N \) is as in (3.6), \( \theta_{sa} \) is real, and \( B(\theta_{sa}) \) is a vertical Bromwich contour on which \( \Re(\theta) = \theta_{sa} \in (-\beta^2/4, \infty) \). If \( \theta_{sa} > 0 \) the sum in (4.1) is absent, and we assume that \( \theta_{sa} \neq \theta_N \) for any \( N \).

Next we give some asymptotic results for the parabolic cylinder functions.

**Lemma 7** For \( \varepsilon \to 0^+ \) and \( \phi > 0 \) we have the asymptotic result

\[
\frac{D_{-\phi/\varepsilon}(-(1 + X)/\sqrt{\varepsilon})}{D_{-\phi/\varepsilon}(-1/\sqrt{\varepsilon})} \sim \left[ \frac{1 + 4\phi}{(1 + X)^2 + 4\phi} \right]^{1/4} \left[ \frac{\sqrt{1 + 4\phi} + 1}{\sqrt{(1 + X)^2 + 4\phi} + 1 + X} \right]^{1/2} \times \exp \left[ \frac{\phi}{\varepsilon} \log \left( \frac{\sqrt{1 + 4\phi} - 1}{\sqrt{(1 + X)^2 + 4\phi} - 1 - X} \right) \right] \times \exp \left\{ \frac{1}{4\varepsilon} \left[ -\sqrt{1 + 4\phi} + (1 + X)\sqrt{(1 + X)^2 + 4\phi} \right] \right\}. \tag{4.2}
\]

For \(-1 < X < 0 \) (thus \( 1 + X > 0 \)) and \(-(1 + X)^2/4 < \phi < 0 \), (4.2) holds as long as \(-\phi/\varepsilon = -\theta\) remains bounded away from the non-negative integers 0, 1, 2, ... .

The derivation of Lemma 7 is deferred to the Appendix.

**Lemma 8** For \( \varepsilon \to 0 \) and \( \phi < 0 \), with \( \phi \approx -(1 + X)^2/4 \) and

\[
\phi = -\frac{1}{4} (1 + X)^2 + \left( \frac{1 + X}{2} \right)^{2/3} \varepsilon^{2/3} \delta, \quad \delta = O(1), \tag{4.3}
\]
we have

\[ D_{-\phi/\varepsilon} \left( -\frac{1 + X}{\sqrt{\varepsilon}} \right) \sim \varepsilon^{\phi/(2\alpha)} \frac{2\pi(1 + X)^{1/3}}{2^{1/3}\varepsilon^{1/6}} \exp \left\{ \frac{1}{\varepsilon} \left[ -\frac{(1 + X)^2}{8} - \phi \log \left( \frac{1 + X}{2} \right) \right] \right\} \]

\[ \times \left[ \sin(\pi \theta) \text{Bi}(\delta) + \cos(\pi \theta) \text{Ai}(\delta) \right], \quad (4.4) \]

where \( \text{Ai}(\cdot) \) and \( \text{Bi}(\cdot) \) are the Airy functions. The expression in (4.4) holds away from zeros of \( \left[ \sin(\pi \theta) \text{Bi}(\delta) + \cos(\pi \theta) \text{Ai}(\delta) \right] \).

This result is obtained by using well-known [20] results for approximating parabolic cylinder functions by Airy functions.

Lemmas 7 and 8 give the approximations to the integrand in equation (4.1) that we shall need. Now consider the sum in equation (4.1), which is needed for \( \theta_{sa} < 0 \). We have already estimated \( h_N \) in equation (B 10) and shown that \( \theta_N \sim -N \) with an exponentially small error. We thus have

\[ D_{-\theta_N}(-\beta - X) \sim \frac{D_N((1 + X)/\sqrt{\varepsilon})}{D_N(-1/\sqrt{\varepsilon})} \]

\[ = \exp \left\{ -\frac{1}{\varepsilon} \left( \frac{X^2}{4} + \frac{X}{2} \right) \right\} \frac{He_N((1 + X)/\sqrt{\varepsilon})}{He_N(1/\sqrt{\varepsilon})}, \quad (4.5) \]

where \( He_N(\cdot) \) is the Nth Hermite polynomial (thus \( He_0(Z) = 1, He_1(Z) = Z \), etc.), which may be obtained, for example, from the integral

\[ He_N(Z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (Z + iw)^N e^{-w^2/2} dw. \]

We shall need equation (4.5) only for \(-1 < X < 0\) so the argument(s) of \( He_N(\cdot) \) will be always positive. Using the saddle point method we can expand equation (4.6) for \( N \to \infty \), \( Z \to \infty \) and a natural scaling has \( N = O(Z^2) \). Then different expansions are needed for \( 4N/Z^2 < 1 \), and for \( 4N - Z^2 = O(N^{1/3}) \). Below we summarize only the main results.

**Lemma 9** For \( \varepsilon \to 0 \), \( N \to \infty \), and \( X \in (-1, 0) \) we have

\[ He_N \left( \frac{1 + X}{\sqrt{\varepsilon}} \right) \sim \sqrt{\frac{N}{2}} \left[ \frac{(1 + X)^2}{4\varepsilon} - N \right]^{-1/4} \left[ \frac{1 + X}{2} - \sqrt{\frac{(1 + X)^2}{4\varepsilon} - N} \right]^{-1/2} \]

\[ \times \exp \left\{ \frac{(1 + X)^2}{4\varepsilon} - \frac{N}{2} - \frac{1 + X}{\sqrt{\varepsilon}} \sqrt{\frac{(1 + X)^2}{4\varepsilon} - N} \right\} \]

\[ \times \exp \left\{ N \log \left[ \frac{1 + X}{2\sqrt{\varepsilon}} + \sqrt{\frac{(1 + X)^2}{4\varepsilon} - N} \right] \right\}, \quad (4.7) \]

where \((1 + X)^2/(4\varepsilon N) > 1\). For \((1 + X)^2/(4\varepsilon N) \approx 1\), with the scaling

\[ \varepsilon N = \frac{1}{4}(1 + X)^2 - \left( \frac{1 + X}{2} \right)^{2/3} \varepsilon^{2/3} W, \quad W = O(1), \quad (4.8) \]
we have

\[ H \varepsilon_N \left( \frac{1 + X}{\sqrt{\varepsilon}} \right) \sim 2^{1/6-N} \sqrt{\pi} \left( \frac{1 + X}{\sqrt{\varepsilon}} \right)^{N+1/3} \exp \left[ \frac{(1 + X)^2}{8 \varepsilon} \right] \text{Ai}(W), \tag{4.9} \]

where \text{Ai}(\cdot) is the Airy function.

We now asymptotically evaluate the expression in Lemma 6, starting with the integral. We scale \( \theta = \phi/\varepsilon \), use equation (4.2) to approximate the ratio of parabolic cylinder functions, and note that as long as we remain bounded away from the poles \( \theta_N \), equation (A1) may be used to approximate \( R^\beta(\theta) \), and then

\[ \frac{d\theta}{\sqrt{\theta + \beta/4 - R^\beta(\theta)}} \sim \frac{d\phi}{\sqrt{\varepsilon} \sqrt{1 + 4 \phi}}. \]

The integrand in equation (4.1) thus becomes

\[
\begin{align*}
\frac{1}{\sqrt{\varepsilon} (1 + 4 \phi)^{1/4}} & \left[ \frac{1}{(1 + X)^2 + 4 \phi} \right]^{1/4} \left[ \frac{\sqrt{1 + 4 \phi} + 1}{\sqrt{(1 + X)^2 + 4 \phi} + 1 + X} \right]^{1/2} \\
& \times \exp \left\{ \frac{1}{\varepsilon} \left[ - \frac{X^2}{4} - \frac{X}{2} + \frac{X_0}{2} + \phi t - \left( \frac{X_0}{2} + \frac{1}{4} \right) \sqrt{1 + 4 \phi} + \frac{1 + X}{4} \sqrt{(1 + X)^2 + 4 \phi} \right] \right\} \\
& \times \exp \left[ \frac{\phi}{\varepsilon} \log \left( \frac{\sqrt{1 + 4 \phi} - 1}{\sqrt{(1 + X)^2 + 4 \phi} - 1 - X} \right) \right] \\
& = \varepsilon^{-1/2} \mathcal{G}(\phi; t, X, X_0) \exp \left[ \mathcal{F}(\phi; t, X, X_0)/\varepsilon \right], \tag{4.10} \end{align*}
\]

and this we now integrate over the \( \phi \) variable. So that there are saddle points where \( \partial \mathcal{F}/\partial \phi = 0 \), or

\[ t - \frac{X_0}{\sqrt{1 + 4 \phi}} + \log \left[ \frac{\sqrt{1 + 4 \phi} - 1}{\sqrt{(1 + X)^2 + 4 \phi} - 1 - X} \right] = 0. \tag{4.11} \]

We write the solution of equation (4.11) as \( \phi_s = \phi_s(X, t; X_0) \). First observe that \( \phi = 0 \) is a solution when \( t - X_0 + \log(1 + X) = 0 \) (when \( -1 < X < 0 \)), and this corresponds precisely to the fluid approximation \( X = -1 + e^{X_0-t} \) in equation (2.7), where \( Y_{fl}(t) < 0 \) for \( t > X_0 \). The equation (4.11) applies only for \( \phi > -1/4 \) and \( \phi > -(1 + X)^2/4 \), as only then equation (4.2) can be used to approximate the integrand in equation (4.1). In Figure 5, we sketch the “surface” \( \phi_s(X, t) \) over the \( (X, t) \) plane for \( X_0 = 0 \). The curve \( X + 1 = e^{X_0-t} \) divides the \( (X, t) \) plane into regions where \( \phi_s > 0 \) and \( \phi_s < 0 \). We have \( \phi_s > 0 \) when \( X \leq -1 \), and also when \( 0 < t < X_0 - \log(1 + X) \) with \( X \in (-1, 0] \). If \( X = 0 \) we may solve equation (4.11) explicitly to give \( \phi_s(0, t; X_0) = (X_0^2 - t^2)/(4t^2) \), and then

\[ \frac{1}{\varepsilon} \left[ \frac{X_0}{2} + t \phi_s(0, t; X_0) - \frac{X_0}{2} \sqrt{1 + 4 \phi_s(0, t; X_0)} \right] = -\frac{1}{4\varepsilon t}(t - X_0)^2, \]

which is the exponential part in \( p_{bsm} \) in equation (2.17), and this we previously remarked holds also at \( X = 0 \). For any \( X \leq 0 \) we can obtain the behaviour of \( \phi_s \) at \( t \to 0 \), by expanding equation (4.11) for \( \phi \to \infty \), which leads to \( \phi_s \sim (X_0 - X)^2/(4t^2) \) as \( t \to 0 \). For
Figure 5. The curve $X + 1 = e^{X_0 - t}$ divides the surface $\phi_s(X,t)$ into regions where $\phi_s > 0$ and $\phi_s < 0$.

$t \to \infty$ and $X \leq -1$ we have $\phi_s \to 0$ and from equation (4.11) we can obtain the more precise estimate

$$\phi_s \sim \begin{cases} -(X + 1) e^{X_0 - t}, & X < -1 \\ e^{2(X_0 - t)}, & X = -1 \end{cases}, \quad t \to \infty. \tag{4.12}$$

Thus for large times and $X \leq -1$ the saddle point becomes exponentially close to the origin in the $\phi$-plane. For $X \leq -1$ we have $\partial \phi_s / \partial t < 0$ for all times, and $\phi_s$ decreases from $+\infty$ to 0 as $t$ increases from 0 to $+\infty$. In contrast, for $-1 < X < 0$ we again have $\partial \phi_s / \partial t < 0$ but now $\phi_s$ decreases from $+\infty$ to 0 as $t$ increases from 0 to $X_0 - \log(1 + X)$ (which is $> 0$), and $\phi_s$ further decreases from 0 to $-(1 + X)^2/4$ as $t$ increases from $X_0 - \log(1 + X)$ to $t_*$, where $t_*$ is obtained by setting $\phi = -(1 + X)^2/4$ in equation (4.11) and this leads to the expression in equation (2.21). For $\phi < -(1 + X)^2/4$ the analysis leading to equation (4.11) ceases to be valid, as then Lemma 7 no longer applies.

Using the standard saddle point estimate, we observe that $\partial^2 F / \partial \phi^2 > 0$ at $\phi = \phi_s$ and the steepest descent directions are $\text{arg}(\phi - \phi_s) = \pm \pi/2$, so we approximate the integral by

$$\frac{1}{\sqrt{2\pi}} \mathcal{G}(\phi_s; t, X; X_0) \left[ \frac{\partial^2 F}{\partial \phi^2} (\phi_s; t, X; X_0) \right]^{-1/2} \exp \left[ \frac{1}{\varepsilon} F(\phi_s; t, X; X_0) \right]. \tag{4.13}$$
But, in view of equation (4.10) and our definitions of $G$ and $F$ above equation (4.11), equation (4.13) is the same as equation (2.37), with equations (2.38) and (2.39).

When $\phi_s > 0$ the sum in equation (4.1) is absent, so we have established item (i) in Theorem 2.3, at least for the range $\{ X \leq -1 \}$ and for $t < X_0 - \log(1+X)$ with $X \in (-1, 0)$, which correspond to portions of Regions I and III. For the portions of these regions where $\phi_s < 0$ we must also estimate the sum in equation (4.1), which we shall do shortly.

For $X < -1$ and large times $t$ we can use the estimate in equation (4.12) to simplify equation (2.37), and this leads to equation (2.41), on the $T$ time scale. Along the fluid approximation $\phi_s = 0$, and by expanding equation (4.11) for small $\phi$ we obtain the estimate

$$
\phi_s \sim \frac{t - X_0 + \log(1 + X)}{1 - 2X_0 - (1 + X)^{-2}}, \quad t \to X_0 - \log(1 + X).
$$

(4.14)

Using equation (4.14) in equations (2.37)–(2.39) gives the Gaussian approximation in equation (2.40), which applies only on the $\Delta$-scale, and thus for $X - (-1 + e^{X_0 - t}) = O(\sqrt{\epsilon})$.

We now take $-1 < X < 0$ and consider ranges where $-(1 + X)^2/4 < \phi_s < 0$, and also the case(s) $t > t^*$, where $\phi_s$ fails to exist. By the saddle point method, setting $\theta_s = \phi_s/\epsilon$ allows us to obtain the estimate in equation (4.13) for the integral. The sum in equation (4.1) we then estimate by

$$
\sum_{N=0}^{[-\phi_s/\epsilon]} \sqrt{\epsilon} \tilde{g}(\epsilon N) \exp \left[ \frac{1}{\epsilon} \tilde{f}(\epsilon N) \right],
$$

(4.15)

where

$$
\tilde{f}(z) = -\frac{X^2}{4} - \frac{X}{2} + \frac{X_0}{2} - zt - \left( \frac{X_0}{2} + \frac{1}{4} \right) \sqrt{1 - 4z} - \frac{1 + X}{4} \sqrt{(1 + X)^2 - 4z}
$$

$$
+ z \log \left[ \frac{(1 + X + \sqrt{(1 + X)^2 - 4z})(1 + \sqrt{1 - 4z})}{4z} \right],
$$

(4.16)

and

$$
\tilde{g}(z) = \frac{1}{2\pi (1 - 4z)^{1/4} [(1 + X)^2 - 4z]^{1/4}} \left[ \frac{1 + \sqrt{1 - 4z}}{1 + X - \sqrt{(1 + X)^2 - 4z}} \right]^{1/2}.
$$

(4.17)

To obtain equations (4.15)–(4.17) we used equation (B 10) to estimate $h_N$ (also approximating $N!$ by Stirling’s formula), equations (4.5) and (4.7) to estimate the parabolic cylinder functions in equation (4.1), and the facts that $\beta^2 = \epsilon^{-1}$ and $\theta_N \sim -N$. In equations (4.16) and (4.17) we emphasize the dependence of $z$, but of course $\tilde{f}$ depends also on $t$ and $X$, and $\tilde{g}$ depends also on $X$.

Note also that if we identify $z$ with $-\phi$, apart from a factor of $\epsilon$, the summand in equation (4.15) is nearly identical to equation (4.10), which corresponds to the integrand in equation (4.1), except that the term $\sqrt{(1 + X)^2 - 4z}$ in the summand is replaced by $-\sqrt{(1 + X)^2 + 4\phi}$ in equation (4.10). Thus equation (4.10) and the summand in equation (4.15) may be viewed as different branches of the same function, and in Section 5 we explore further this connection. We now analyse Region IV and those portions of Regions...
I and III where \( \phi_s < 0 \). We evaluate equation (4.15) by the Laplace method, which requires that we find the maximum of \( \tilde{f}(z) \) over the range \( z \in [0, -\phi_s] \).

From equation (4.16) we have

\[
\tilde{f}'(z) = -t + \frac{X_0}{\sqrt{1-4z}} + \log \left[ \frac{1 + X + \sqrt{(1 + X)^2 - 4z}}{1 - \sqrt{1-4z}} \right],
\]

(4.18)

and

\[
\tilde{f}''(z) = \frac{2X_0}{(1-4z)^{3/2}} - \frac{1}{2z\sqrt{1-4z}} - \frac{1 + X}{2z\sqrt{(1 + X)^2 - 4z}}.
\]

(4.19)

Also, \( \tilde{f}(0) = - (1 + X)^2/2 \), \( \tilde{f}'(0) = +\infty \) and

\[
\tilde{f}'(-\phi_s) = \log \left[ \frac{1 + X + \sqrt{(1 + X)^2 + 4\phi_s}}{1 + X - \sqrt{(1 + X)^2 + 4\phi_s}} \right] > 0,
\]

(4.20)

where we used the fact that \( \phi_s \) satisfies equation (4.11). After some calculation we can explicitly solve \( \tilde{f}''(z) = 0 \) to find that it has two roots, at \( z = z_{\pm} \), where

\[
z_{\pm} = \frac{(X_0 + 1)^2(X + 1)^2 + 2(X_0 + 1) - 3(X + 1)^2 \pm X_0^{3/2}(X + 1)\sqrt{(X_0 + 4)(X + 1)^2 - 4}}{8(X_0 + X + 2)(X_0 - X)}.
\]

For \( X \in (-1, 0) \) the roots \( z_{\pm} \) are real only when \( X \in (X_{\text{cusp}}, 0) \), so we see the cusp point in equation (2.26) entering the analysis. At the cusp point we have the following:

\[
z_+ = z_- = \frac{3}{4(X_0 + 3)} > 0, \quad X = X_{\text{cusp}} = -1 + \frac{2}{\sqrt{X_0 + 4}}.
\]

(4.21)

At the points \( z = z_{\pm} \) we need to consider the sign(s) of \( \tilde{f}'(z) \), and we note that, using equation (4.11), \( \tilde{f}'(z_-) = 0 \) implies \( t = t_d(X; X_0) \), and \( \tilde{f}'(z_+) = 0 \) implies \( t = t_c(X; X_0) \), where \( t_c \) and \( t_d \) are the curves in equations (2.22)–(2.25).

In Region I, \( X < X_{\text{cusp}} \) and \( \tilde{f}''(z) < 0 \), and thus \( \tilde{f}' \) decreases from \( \infty \) at \( z = 0 \), to the value in equation (4.20) when \( z = -\phi_s \). Hence \( \tilde{f}' > 0 \) and the maximum of \( \tilde{f} \) occurs at the upper limit of the sum in equation (4.15). By expanding in Taylor series and noting that \( \varepsilon \frac{[-\phi_s/\varepsilon]}{\varepsilon} = -\phi_s + O(\varepsilon) \), we estimate equation (4.15) by

\[
\sqrt{\varepsilon} g(-\phi_s) \exp \left[ \frac{1}{\varepsilon} \int \left( \varepsilon \left[ -\frac{\phi_s}{\varepsilon} \right] \right) \sum_{m=0}^{\infty} \exp^{-m \tilde{f}'(-\phi_s)} \right].
\]

(4.22)

Evaluating the geometric sum in equation (4.22), and using equations (4.11), (4.16) and (4.18), we ultimately obtain

\[
\frac{\sqrt{\varepsilon} (1 + \sqrt{1 + 4\phi_s})^{1/2}(1 + X + \sqrt{(1 + X)^2 + 4\phi_s})^{3/2}}{8\pi \sqrt{-\phi_s} (1 + 4\phi_s)^{1/4} (1 + X)^{3/4} + 4\phi_s} \exp \left[ \frac{1}{\varepsilon} \tilde{f}' \left( \varepsilon \left[ -\frac{\phi_s}{\varepsilon} \right] \right) \right].
\]

(4.23)

But comparing the magnitude of equation (4.23), which is roughly of order \( O(e^{[\tilde{f}'(-\phi_s)/\varepsilon]} \) to the saddle point contribution in equation (2.37) (or (4.10)), which is \( O(e^F/\varepsilon) \), we see that
the latter dominates, as \( F(X, t) \geq \tilde{f}(-\phi_s) \). Hence we have established equation (2.37) in all of Region I.

In the part of Region III above the fluid approximation we have \( \phi_s < 0 \). Now \( t < t_* \) if \( X \in (X_{\text{cusp}}, X_*) \) and \( t < t_d \) if \( X \in (X_*, 0) \), where we recall that \((X_*, t_{**})\) is the intersection of the curves \( t_*(X; X_0) \) and \( t_d(X; X_0) \), so the intersection point depends only on \( X_0 \). If we view \( \tilde{f}(z) \) over \( z \in (0, (1 + X)^2/4) \) then \( \tilde{f}''(z) = 0 \) at \( z = z_\pm \). We may have \( -\phi_s < z_- \), \(-\phi_s \in (z_-, z_+) \), or \(-\phi_s > z_+ \), depending on the values of \((X, t)\). However, since \( t < t_d \) we have \( \tilde{f}'(z_+) > 0 \) and \( \tilde{f}'(z_-) > 0 \) and thus again \( \tilde{f}(z) > 0 \) for \( z \in [0, -\phi_s] \) and the estimate in equation (4.23) holds for the sum in equation (4.1), but the saddle contribution in equation (4.10) dominates the sum. We have thus established equation (2.37) for the part of Region III where \( \phi_s < 0 \).

Now consider Region IV, where \( t \in (t_d, t_*) \) and \( X \in (X_*, 0) \). Then \( \tilde{f}'(z_-) < 0 \) and \( \tilde{f}'(z_+) > 0 \), so \( \tilde{f}(z) \) has two zeros, say at \( z_1 \) and \( z_2 \), which satisfy \( 0 < z_1 < z_- < z_2 < z_+ \). Also, \( z_1 \) will be a local maximum of \( \tilde{f} \) while \( z_2 \) will be a local minimum. The global maximum of \( \tilde{f}(z) \) will come from either \( z = z_1 \), or the endpoint \( z = -\phi_s \). But the endpoint contribution will be dominated again by the saddle point contribution to the integral in equation (4.1). The contribution to the sum from \( z_1 \) can be obtained by the Laplace estimate, as

\[
\sqrt{2\pi \tilde{g}(z_1)} \left[ -\tilde{f}''(z_1) \right]^{-1/2} \exp \left[ \frac{1}{\epsilon} \tilde{f}(z_1) \right].
\]  

(4.24)

But \( \sqrt{2\pi \tilde{g}(z_1)} \left[ -\tilde{f}''(z_1) \right]^{-1/2} \) is the same as \( \tilde{G}(z_1) \) in equation (2.48). Then in Region IV the asymptotics are governed by either \( \epsilon N = z_1 \) from the sum or \( \phi = \phi_s \) in the integral, and either of these may dominate. This establishes item (v) in Theorem 2.3.

In Regions II and V–VII we have \( t > t_* \). Now we set \( \epsilon \theta sa = -(1 + X)^2/4 \) in equation (4.1) and the integrand has no saddle points, as equation (4.11) has no real solutions. Later we shall examine the case \( t \approx t_* \) to see what happens to the solution of equation (4.11) as \( t \) increases past \( t_* \). Now we must estimate the sum

\[
\sum_{N=0}^{[(1+X)^2]/(4\epsilon)} \sqrt{\epsilon \tilde{g}(\epsilon N)} \exp \left[ \frac{1}{\epsilon} \tilde{f}(\epsilon N) \right].
\]

(4.25)

Our approximation to the summand fails near the upper limit, where \( z = (1 + X)^2/4 \). But this will be important asymptotically only for \( t \approx t_* \), and this case we shall analyse separately. We note that

\[
\tilde{f}' \left( \frac{(1 + X)^2}{4} \right) = t_* - t < 0
\]

(4.26)

so now \( \tilde{f} \) will no longer have a local maximum at the upper limit in equation (4.25).

In Region V we have \( t_d < t < t_c \) if \( X \in (X_{\text{cusp}}, X_*) \) and \( t_* < t < t_c \) if \( X \in (X_*, 0) \). In either case \( \tilde{f}'(z_-) < 0 \) and \( \tilde{f}'(z_+) > 0 \), so that \( \tilde{f}'(z) = 0 \) has three roots, at \( z_1 < z_2 < z_3 \), with \( z_1 \) and \( z_3 \) corresponding to local maxima of \( \tilde{f} \) and \( z_2 \) a local minimum. Then the contribution from the maximum at \( z_1 \) is again given by equation (4.24). Adding to this the contribution from \( z_3 \) leads to equation (2.47) in Theorem 2.3. Either contribution may dominate, and the two are comparable along the curve \( t_F \), as previously discussed.
In Region VII we have $X \in (X_{\text{cusp}}, X_*)$ and $t_* < t < t_d$. Now $\tilde{f}'(z_-) > 0$, $\tilde{f}'(z_+) > 0$ which together with equation (4.26) shows that $\tilde{f}'(z) = 0$ has a unique root, call it $z_3$ ($> z_+$), and this leads to equation (2.46) in Theorem 2.3.

For Region VI we have $\tilde{f}'(z_-) < 0$ and $\tilde{f}'(z_+) < 0$, and, as always, $\tilde{f}'(0) = +\infty$. Thus there is again a unique root, which we call $z_1$, and we obtain equation (2.45).

For Region II, $\tilde{f}''(z) = 0$ has no solutions in $z \in (0, (1 + X)^2/4)$, since $X < X_{\text{cusp}}$. Then $\tilde{f}'(0) = +\infty$ and $\tilde{f}' < 0$ at $z = (1 + X)^2/4$ so that $\tilde{f}'$ has a unique zero and we again obtain equation (2.45).

It remains to consider the case $t \approx t_*$, where the approximations leading to the summand in equations (4.15) and (4.25) cease to be valid. Also, we must separately analyse when $(X, t) = (X_{\text{cusp}}, t_{\text{cusp}})$, where the local maxima $z_1$ and $z_3$ of $\tilde{f}(z)$ coalesce, and the standard Laplace estimate in equation (4.24) no longer holds (at the cusp point $z_1 = z_3$ and $\tilde{f}''(z_1) = 0$).

We introduce the scaling

$$ t = t_*(X; X_0) + \epsilon^{1/3} A, \quad A = O(1) \quad (4.27) $$

as in equation (2.43), so that $t - t_* = O(\epsilon^{1/3})$. Furthermore we shift the $Br$ contour in (4.1) to $\Re(\phi) = -(1 + X)^2/4$. Then scaling $\phi$ as in equation (4.3), we use equation (4.4) to approximate $D_{-\phi/\epsilon}(-(1 + X)/\sqrt{\epsilon})$ and (C 5) to approximate $D_{-\phi/\epsilon}(-1/\sqrt{\delta})$. Then the integrand in equation (4.1) becomes, in this asymptotic limit,

$$ \frac{\sqrt{\pi}}{2} \left[ \sqrt{1 - (1 + X)^2} + 1 \right]^{1/2} \left[ \text{Bi}(\delta) + \cot(\pi\theta) \text{Ai}(\delta) \right] e^{\mathcal{M}/\epsilon} \quad (4.28) $$

where

$$ \mathcal{M} = -\frac{(X + 1)^2}{8} - \frac{X^2}{4} - \frac{X}{2} + \frac{X_0}{2} + \phi t - \frac{\phi}{2} - \left( \frac{X_0}{2} + \frac{1}{4} \right) \sqrt{4 + 4\phi + \phi \log \left( \frac{1 - \sqrt{1 + 4\phi}}{1 + X} \right)}, $$

and equation (4.28) is to be integrated with respect to $\delta$, since $d\phi = (1 + X)^2/2 \epsilon^{2/3} d\delta$. Using the scaling in equation (4.3) we can rewrite $\mathcal{M}$ in equation (4.29) as

$$ \mathcal{M} = F_\epsilon - \frac{1}{4} (1 + X)^2 \epsilon^{1/3} A + \epsilon A_1 \delta + o(\epsilon), \quad (4.30) $$

as this simply amounts to a Taylor expansion about $\phi = -(1 + X)^2/4$. In equation (4.30), $F_\epsilon$ is obtained by setting $4\phi = -(1 + X)^2$ and $t = t_*$ in equation (4.29), and then $F_\epsilon$ is as in equation (2.44). Also, in equation (4.30) $A_1 = [(1 + X)/2]^{2/3} A$. Using equation (4.30) in equation (4.28) we must evaluate the integral

$$ \frac{1}{2\pi i} \int_{Br} \left[ \text{Bi}(\delta) + \cot(\pi\theta) \text{Ai}(\delta) \right] e^{\mathcal{M}/\epsilon} \, d\delta, \quad (4.31) $$

where $Br$ has been shifted so that $\Re(\delta) = 0$ on this contour. Now

$$ \cot(\pi\theta) = i \frac{e^{i\pi\theta} + e^{-i\pi\theta}}{e^{i\pi\theta} - e^{-i\pi\theta}} \quad \text{and} \quad \theta = \frac{\phi}{\epsilon} = -\frac{1}{4\epsilon} (1 + X)^2 + \epsilon^{-1/3} \left( \frac{1 + X}{2} \right)^{2/3} \delta. $$
For $\delta$ positive imaginary we have $\cot(\pi \theta) \sim -i$, while for $\delta$ negative imaginary we have $\cot(\pi \theta) \sim i$. Hence parametrizing the $Br$ contour in equation (4.31) and separately calculating the contributions from $\Im(\delta) > 0$ and $\Im(\delta) < 0$ we are led to

$$\frac{1}{2\pi i} \int_{-\infty}^{\infty} \left[ \text{Bi}(\delta) - i \text{Ai}(\delta) \right] e^{\delta t} d\delta + \frac{1}{2\pi i} \int_{-\infty}^{0} \left[ \text{Bi}(\delta) + i \text{Ai}(\delta) \right] e^{\delta t} d\delta. \quad (4.32)$$

Using the symmetry relations between Airy functions $\text{Ai}(z) \pm i \text{Bi}(z) = 2e^{\pm \pi i/3} \text{Ai}(z e^{\mp 2\pi i/3})$, performing some contour rotations in equation (4.32), and then multiplying the result by the factors in equation (4.28) that are not contained in equation (4.31), we ultimately obtain

$$\mathcal{P}_+ \sim \frac{\left[ \sqrt{1 - (1 + X)^2} + 1 \right]^{1/2}}{2\sqrt{\pi} \left[ 1 - (1 + X)^2 \right]^{1/4}} \exp \left[ \frac{F_c}{\epsilon} - \frac{(1 + X)^2}{4\epsilon^{2/3}} A \right] \int_{0}^{\infty} \text{Ai}(\delta) \left( e^{A_1 \omega^2 \delta} + e^{A_1 \omega \delta} \right) d\delta,$$

where $\omega = e^{\pm 2\pi i/3}$ is a cube root of unity. Thus, $\mathcal{P}_+$ represents the contribution to $p(x, t)$ from the integral in equation (4.1) (with $\theta_{sa} = -(1 + X)^2/(4\epsilon)$), for $t - t_* = O(\epsilon^{1/3})$.

To compute the contribution from the residue sum in equation (4.1), which we denote by $\mathcal{P}_-$, we use equation (4.5), approximate $He_N((1 + X)/\sqrt{\epsilon})$ using equation (4.9), and approximate $He_N(1/\sqrt{\epsilon})$ using equation (4.7) with $X = 0$. Note that near $t = t_*$ we must carefully consider the contribution from the upper endpoint in the sum in equation (4.1), since now $\phi_s \sim -(1 + X)^2/4$. We thus obtain

$$\frac{\epsilon^{1/3}}{2^{1/3} \sqrt{\pi}} \frac{\left[ 1 + \sqrt{1 - (1 + X)^2} \right]^{1/2}}{(1 + X)^{2/3} \left[ 1 - (1 + X)^2 \right]^{1/4}} \text{Ai}(W) e^{\mathcal{S}/\epsilon}, \quad (4.33)$$

where

$$\mathcal{S} = -\frac{(1 + X)^2}{8} + \frac{X_0}{2} + \frac{z}{2} - z t - \left( \frac{X_0}{2} + \frac{1}{4} \right) \sqrt{1 - 4z + z} \log \left( \frac{1 + X}{2} \right) + z \log \left( \frac{1 + \sqrt{1 - 4z}}{2z} \right).$$

Expanding $\mathcal{S}$ in the above equation about $t = t_*$ and $z = (1 + X)^2/4$ by using the scaling in equations (4.27) and (4.8), we obtain

$$\mathcal{S} = F_c - \frac{(1 + X)^2}{4} \epsilon^{1/3} A + A_1 W + o(\epsilon). \quad (4.34)$$

With equations (4.33) and (4.34), summing equation (4.33) over $N$ corresponds to $\epsilon^{-1/3}$ times an integral over $W > 0$, in view of the Euler–Maclaurin formula and equation (4.8), as $W > 0$ corresponds to $\epsilon N < (1 + X)^2/4$. Then the contribution to equation (4.1) from the sum is, to leading order,

$$\mathcal{P}_- \sim \frac{1}{2\sqrt{\pi}} \frac{\left[ \sqrt{1 - (1 + X)^2} + 1 \right]^{1/2}}{\left[ 1 - (1 + X)^2 \right]^{1/4}} \exp \left[ \frac{F_c}{\epsilon} - \frac{(1 + X)^2}{4\epsilon^{2/3}} A \right] \int_{0}^{\infty} \text{Ai}(W) e^{A_1 W} dW.$$

To obtain the expansion of $p$ for $t \approx t_*$ we simply add $\mathcal{P}_+$ to $\mathcal{P}_-$. While both $\mathcal{P}_+$ and $\mathcal{P}_-$ involve integrals of Airy functions, their sum is much simpler. To see this let
\[ \mathcal{H}(Z) = \int_{0}^{Z} \text{Ai}(W)(e^{ZW} + e^{\omega ZW} + e^{\omega^2 ZW})dW. \]

We have \( \mathcal{H}(0) = 3 \int_{0}^{\infty} \text{Ai}(W)dW = 1. \)

Using the Airy equation \( \text{Ai}''(W) = W \text{Ai}(W), \) some integration by parts, and the facts that \( 1 + \omega + \omega^2 = 0 \) and \( \omega^3 = 1, \) we obtain

\[
\mathcal{H}'(Z) = \int_{0}^{Z} W \text{Ai}(W)(e^{ZW} + \omega e^{\omega ZW} + \omega^2 e^{\omega^2 ZW})dW
\]
\[
= -\int_{0}^{Z} Z^2 \text{Ai}(W)(e^{ZW} + \omega e^{\omega ZW} + \omega^2 e^{\omega^2 ZW})dW
\]
\[
= -Z^2 \mathcal{H}(Z).
\]

It follows that \( \mathcal{H}(Z) = e^{-Z^2/3}, \) and then \( P_+ \) and \( P_- \) sum to give the explicit formula in equation (2.42), since \( A_1^3 = (1 + X)^2 A_3/4. \)

In crossing \( t = t_* \) we go from Region I to II \((-1 < X < X_{cusp}), \) from Region III to VII \((X_{cusp} < X < X_*), \) or from Region IV to V \((X_* < X < 0). \) In the first two crossings equation (2.42) gives the approximation to \( p(x, t), \) but in going from Region IV to V we must also consider the contribution from \( z_1. \) Then equation (2.42) gives the smooth continuation of the term \( Ge^{F/\varepsilon} \) in equation (2.48) and it becomes the \( z_3 \) contribution in equation (2.47).

Also, we can show directly that the approximation to \( p(x, t) \) is smooth as \( t \) increases past \( t_* \). For example, consider the transition from Region I to II. Then for \( t < t_* \), equation (4.11) (with \( \phi = \phi_s \)) holds. For \( \phi_s \) slightly greater than \(-1 + X)^2/4 \) we let \( \phi_s = -(1 + X)^2/4 + \varepsilon_0. \) Then expanding equation (4.11) about \( t = t_* \) and for small \( \varepsilon_0 \) yields

\[
t - t_* + O(\varepsilon_0) = \log \left[ \frac{1 + X - 2\sqrt{\varepsilon_0}}{1 + X} \right] \sim -\frac{2\sqrt{\varepsilon_0}}{1 + X},
\]

and hence

\[
\varepsilon_0 \sim \frac{1}{4}(1 + X)^2 \left[ t - t_*(X; X_0) \right]^2.
\] (4.35)

By setting \( 4\phi_s + (1 + X)^2 = U \) and expanding equation (4.11) for small \( U \) we obtain

\[
t - t_* = -\frac{\sqrt{U}}{1 + X} + \frac{1 - (1 + X_0)(1 + X)^2}{2(1 + X)^2[1 - (1 + X)^2]^{3/2}} U - \frac{U^{3/2}}{3(1 + X)^3} + O(U^2).
\] (4.36)

Also for small \( U, F(X, t) \) in equation (2.37) becomes

\[
F = F_* + \frac{1 + X}{4} - \frac{\sqrt{U}}{1 + X} + \frac{(1 + X_0)(1 + X)^2 - 1}{8[1 - (1 + X)^2]^{3/2}} U + O(U^2)
\]
\[
= F_* + \frac{(1 + X)^2}{4}(t_* - t) - \frac{U^{3/2}}{12(1 + X)} + O(U^2),
\] (4.37)

where we used equation (4.36). But \( \sqrt{U} \sim (1 + X)(t_* - t) \) so that \( U^{3/2} \sim (t_* - t)^3(1 + X)^3 \) and equation (4.37) becomes the exponential part of equation (2.42). The algebraic part of equation (2.42) corresponds simply to \( G(X, t_*(X; X_0)). \) Thus equation (2.42) is the
smooth limiting form of the Region I approximation as $t \uparrow t^*$, and a completely analogous
calculation shows that equation (2.42) is also the limiting form equation (2.45) as $t \downarrow t^*$.
To go from equation (2.37) to equation (2.45) we can simply replace $\sqrt{(1 + X)^2 + 4\phi_s}$ by
$-\sqrt{(1 + X)^2 + 4\phi_s}$, and $(1 + X)^2 + 4\phi_s$ has, in view of equation (4.35), a double zero at
$t = t^*$.

Finally, we consider the vicinity of the cusp point $(X_{cusp}, t_{cusp})$, defined by equations
(2.26) and (2.27). Then certainly $t > t^*$, and at the cusp the Regions II, V, VI and VII
meet. We evaluate the sum in equation (4.25), with equations (4.16) and (4.17). Now the
maxima of $\tilde{f}$ at $z_1$ and $z_3$ are very close, and the estimate in equation (4.24) no longer
holds. Note that at the cusp point $z_+ = z_-$ and both $\tilde{f}'$ and $\tilde{f}''$ vanish.

Again applying the Laplace method to equation (4.25), we expand the summand about
$$z = z_{cusp} = \frac{3}{4(X_0 + 3)},$$
and we note that equation (4.38) agrees with equation (4.21). Let us now write $\tilde{f}(z; X, t)$
and $\tilde{g}(z; X)$, to indicate the dependence on all three variables in equations (4.16) and
(4.17). Then, after some calculation,
$$\tilde{g}(z_{cusp}; X_{cusp}) = \frac{1}{2\pi} \sqrt{\frac{X_0 + 3}{X_0}} \sqrt{\sqrt{X_0(X_0 + 3)} + X_0 + 2}$$
(4.39)
and
$$\tilde{f}(z_{cusp}; X_{cusp}, t_{cusp}) = \frac{\sqrt{X_0}}{4(X_0 + 4)} \left[ -2(X_0 + 3)^{3/2} + \sqrt{X_0}(2X_0 + 9) \right].$$

We shall need some of the higher order terms in the Taylor expansion of $\tilde{f}(z; X, t)$ about
$(z, X, t) = (z_{cusp}, X_{cusp}, t_{cusp})$.

From equation (4.18) we find that
$$\tilde{f}'(z_{cusp}) \sim t_{cusp} - t + \sqrt{(X_0 + 4)(X_0 + 3)/X_0}(X - X_{cusp}),$$
where we first set $z = z_{cusp}$ in equation (4.18) and then expanded the result for $(X, t) \to (X_{cusp}, t_{cusp})$. Also, from equation (4.19) we have, as $X \to X_{cusp}$,
$$\tilde{f}''(z_{cusp}) \sim 2[(X_0 + 4)(X_0 + 3)/X_0]^{3/2}(X - X_{cusp}),$$
$$\tilde{f}'''(z_{cusp}) \sim O(X - X_{cusp}), \quad \tilde{f}^{(iv)}(z_{cusp}) \sim -16(X_0 + 3)^{11/2}/X_0^{5/2}.$$
If we introduce the scaled variables $(\xi, \eta)$ in equation (2.49) and also set $z - z_{cusp} = \varepsilon^{1/4} w = O(\varepsilon^{1/4})$, then
$$\frac{\tilde{f}'(z_{cusp})}{4! \varepsilon}(z - z_{cusp})^4 \sim -\frac{2(X_0 + 3)^{11/2}}{3X_0^{5/2}} w^4 = O(1), \quad \frac{\tilde{f}''(z_{cusp})}{3! \varepsilon}(z - z_{cusp})^3 = O(\varepsilon^{1/4}),$$
and
$$\frac{\tilde{f}'''(z_{cusp})}{2! \varepsilon}(z - z_{cusp})^2 \sim \left[ \frac{(X_0 + 4)(X_0 + 3)}{X_0} \right]^{3/2} \xi^2, \quad \frac{\tilde{f}'(z_{cusp})}{\varepsilon}(z - z_{cusp}) \sim -\eta w.$$
We thus have the approximation
\[
\tilde{f}(z) = \frac{\tilde{f}_0(X,t)}{\varepsilon} = \tilde{f}_0(X_0 + 3, \varepsilon) - \eta w + \left( \frac{X_0 + 4(X_0 + 3)}{X_0} \right)^{3/2} \xi^2 w^2 + \frac{2(X_0 + 3)^{11/2}}{3X_0^{5/2}} w^4 + O(\varepsilon^{1/4}), \tag{4.40}
\]
where \( \tilde{f}_0(X,t) = \tilde{f}(z_{\text{cusp}}; X,t) \) is obtained by setting \( z = z_{\text{cusp}} \) in equation (4.16). Then we can further expand \( \tilde{f}_0(X,t) \) about \( (X,t) = (X_{\text{cusp}}, t_{\text{cusp}}) \) in a double Taylor series, and this ultimately leads to
\[
\tilde{f}_0(X,t) = \tilde{f}(z_{\text{cusp}}; X_{\text{cusp}}, t_{\text{cusp}}) - \eta w + \frac{3}{4(X_0 + 3)} \xi^2 + o(1) \tag{4.41}
\]
where we again used the scaled variables \( (\xi, \eta) \) in equation (2.49). Using equation (4.41) in equations (4.40) and (4.39), we obtain the leading order approximation to the summand in equation (4.25) for \( z - z_{\text{cusp}} = O(\varepsilon^{1/4}) \) and \( (X,t) \) near the cusp. The sum is asymptotically the same as \( \varepsilon^{-1} \) times an integral over \( z \), or \( \varepsilon^{-3/4} \) times an integral over \( w \in (-\infty, \infty) \). We thus obtain the approximation in equations (2.50) and (2.51), that involves the integral \( J \).

This completes the analysis of the range \( X < 0 \).

5 Alternate approach

We discuss a different approach to the asymptotics, based on geometrical optics and singular perturbations, which will lead to results equivalent to those in Theorems 2.2 and 2.3.

Scaling \( x = \beta X \) in equation (2.1) and using \( \varepsilon = \beta^{-2} \) we argue that an asymptotic solution to equation (2.1) is in the form
\[
p(x,t) = e^{F(X,t)/\varepsilon} \left[ G(X,t) + \varepsilon G^{(1)}(X,t) + O(\varepsilon^2) \right]. \tag{5.1}
\]
Then we distinguish \( X > 0 \) and \( X < 0 \) by setting \( (F,G) = (F^+, G^+) \) for \( X > 0 \) and \( (F,G) = (F^-, G^-) \) for \( X < 0 \). From equations (5.1) and (2.1) we find that \( F^\pm \) satisfy the “eikonal” equation(s)
\[
F^+_t = \left( F^+_X \right)^2 + F^+_X; \quad X, t > 0, \tag{5.2}
\]
\[
F^-_t = \left( F^-_X \right)^2 + (X + 1)F^-_X; \quad X < 0, t > 0, \tag{5.3}
\]
while \( G^\pm \) satisfy the “transport” equation(s)
\[
G^+_t = (2F^+_X + 1)G^+_X + F^+_{XX} G^+; \quad X, t > 0, \tag{5.4}
\]
\[
G^-_t = (2F^-_X + 1)G^-_X + F^-_{XX} G^-; \quad X < 0, t > 0. \tag{5.5}
\]
The initial condition for \( p \) is \( p(x,0) = \delta(x - x_0) = \sqrt{\varepsilon} \delta(X - X_0) \), and the interface conditions at \( X = 0 \), in view of equation (5.1), imply asymptotically that
\[
F^+(0^+, t) = F^-(0^-, t), \tag{5.6}
\]
Then $F_X$ is continuous at $X = 0$ automatically, which follows from equations (5.2), (5.3) and (5.6).

A singular perturbation analysis would first consider the short time scale $t = \varepsilon \tau_0 = O(\varepsilon)$ and a spatial scale near the initial condition, with $X - X_0 = \varepsilon \xi = O(\varepsilon)$. Then in terms of $(\xi, \tau_0)$, $p$ would satisfy the PDE $p_{\tau_0} = p_{\xi \xi} + p_{\xi}$, for $\tau_0 > 0$ and $-\infty < \xi < \infty$, and the initial condition $p = \varepsilon^{-1/2} \delta(\xi)$ at $\tau_0 = 0$. But then $p \sim \varepsilon^{-1/2} (4\pi \tau_0)^{-1/2} \exp\left[-(\xi + \tau_0)^2/(4\tau_0)\right]$, so on the $(\xi, \tau_0)$ scale $p$ may be approximated by the free space Brownian motion with unit negative drift.

On the $(X,t)$ scale we solve the non-linear PDE equation (5.2) by the method of characteristics. The appropriate solution, which is necessary to asymptotically match to that on the $(\xi, \tau_0)$ scale, is a singular solution where all characteristics, also called “rays”, start from the point $(X_0,0)$. Since equation (5.2) is “constant-coefficient”, the rays are all straight lines, and every point in the quarter plane $\{(X,t) : X > 0, t > 0\}$ is reached by a unique ray from $(X_0,0)$. The solution of equation (5.2) corresponding to this ray family has $F^+ = -(X - X_0 + t)^2/(4t)$. Then we can also solve equation (5.4), which is a first order linear PDE, by the method of characteristics. The solution will involve an arbitrary function of $(X - X_0)/t$, which is a function of which ray we are on. By using asymptotic matching between the $(X,t)$ and $(\xi, \tau_0)$ scales we can uniquely determine this unknown function, and ultimately obtain $G^+ = 1/\sqrt{4\pi t}$. We thus conclude that the free space Brownian motion $p_{BM}$ applies also on the $(X,t)$ scale. But it will not be the leading term for $p$ for all $X > 0$, as we show below. This must be true, since, for example, $p_{BM} \to 0$ as $t \to \infty$ while the density $p(x,t)$ approaches the steady state in equation (2.2) or (2.3).

Next we consider the range $X < 0$ and solve equations (5.3) and (5.5) subject to the continuity conditions equations (5.6) and (5.7), the latter yielding

$$F^-(0,t) = -\frac{1}{4t}(t - X_0)^2, \quad (5.8)$$

$$G^-(0,t) = \frac{1}{2\sqrt{\pi t}}. \quad (5.9)$$

We can study in more detail the interface at $X = 0$, by considering the scale $X = O(\varepsilon)$. Then to leading order this will simply regain the conditions in equations (5.8) and (5.9). But the next term(s) in the expansion will show a small $O(\varepsilon)$ probability that a “direct” ray, starting from $(X_0,0)$, will reflect in the $t$-axis ($X = 0$). This small reflection can be used to construct a new ray family, which we call the “reflected” rays, and this will ultimately yield the right side of equation (2.8) in Theorem 2.2. However, as we discussed in Section 2, the reflected rays can never be the dominant part of $p(x,t)$. Even along $X = 0$ they are smaller than $p_{BM}$ (the direct ray expansion) by a factor $O(\varepsilon)$, which is also evident in equation (2.8). For $X > 0$ the reflected rays lead to a solution exponentially smaller than $p_{BM}$, and thus we do not consider the reflected rays further here.

We solve equation (5.3) subject to equation (5.8), so that $X = 0$ provides the “initial manifold” for the first order PDE, and hence the solution for $X < 0$ is not a singular solution. However, the solution will have certain “singular” features, as we show below.
Omitting the details, the solution of equations (5.3) and (5.8) leads to the rays

\[ t = \alpha + \tau, \quad (5.10) \]

\[ X = \left( \frac{1}{2} - \frac{X_0}{2\alpha} \right) e^\tau + \left( \frac{1}{2} + \frac{X_0}{2\alpha} \right) e^{-\tau} - 1. \quad (5.11) \]

Thus equations (5.10) and (5.11) give the mapping between the original \((X,t)\) variables and the \((\alpha, \tau)\) ray variables. When \(\tau = 0\), \(X = 0\) and thus \(\alpha\) is the value of \(t\) where the given ray hits the initial manifold. The solution \(F^-\) can be expressed in terms of the ray variables as

\[ F^- = -\frac{1}{2} \left( \frac{1}{2} - \frac{X_0}{2\alpha} \right)^2 (e^{2\tau} - 1) - \frac{\alpha}{4} + \frac{X_0}{2} - \frac{X_0^2}{4\alpha}, \quad (5.12) \]

and we note that when \(\tau = 0\), \(\alpha = t\) and thus equation (5.8) is satisfied. We can set \(\tau = t - \alpha\) in equation (5.11) and then this equation defines a one parameter family of curves in the \((X,t)\) plane, for \(\alpha > 0\), with \(\alpha\) indexing the family. In Figure 6, we sketch the rays in the space-time plane for \(X < 0\). The figure suggests that some rays maintain a negative slope \(dt/dX\) for all \(\tau > 0\), while others change their slope and ultimately return to \(X = 0\) at some time \(t > \alpha\). Furthermore, while some regions of the \((X,t)\) plane are reached by a single ray, in other regions the rays intersect and then \((X,t)\) corresponds to more than one value of \((\alpha, \tau)\) (the figure suggests that in the multi-valued range, exactly three rays reach a given \((X,t)\)). Below we establish all these results analytically.
Transient distribution of the Halfin-Whitt diffusion process

Setting $\tau = t - \alpha$ in equation (5.11) with $X = 0$ and solving for $t$ leads to

$$t - \alpha = \log \left( \frac{\alpha + X_0}{\alpha - X_0} \right). \quad (5.13)$$

Certainly $t = \alpha$ corresponds to $X = 0$ for all $\alpha > 0$, which corresponds to the $(-)$ sign in equation (5.13), but the $(+)$ sign gives a second solution if $\alpha > X_0$. Thus for $\alpha < X_0$ a ray will not return to $X = 0$, but if $\alpha > X_0$ the ray will return at time

$$t = \alpha + \log \left( \frac{\alpha + X_0}{\alpha - X_0} \right), \quad \alpha > X_0. \quad (5.14)$$

This return time becomes infinite as $\alpha \downarrow X_0$, but becomes very close to $\alpha$ if $\alpha \to +\infty$, which corresponds to an almost immediate return to $X = 0$.

When $\alpha = X_0$ we have $X = e^{-t} = e^{0 - t} - 1$, which is precisely the fluid approximation in the range $X < 0$ (or $t > X_0$). Thus all rays that begin from $X = 0$ at times before the fluid approximation never return to $X = 0$ and maintain a negative slope, while all rays that begin at times $t > X_0$ return to $X = 0$. Also, from equation (5.11), if $\alpha < X_0$, a ray will have $X \to -\infty$ as $t \to \infty$, while if $\alpha = X_0$, the fluid approximation $\to -1$ as $t \to \infty$. Note also that $F^- = 0$ if $\alpha = X_0$.

It is also instructive to find the minimum value of $t$ where the rays return to $X = 0$. When $t = X_0^+$ the return value is very large. To find the minimum value we minimize the right side of equation (5.14), which yields $\alpha = \sqrt{X_0(X_0 + 2)} > 0$ and thus the returned rays hit $X = 0$ for times $t \geq t_{\text{min}}(X_0)$, with

$$t_{\text{min}}(X_0) = \sqrt{X_0(X_0 + 2)} + 2 \log \left( \frac{\sqrt{X_0 + 2} + \sqrt{X_0}}{\sqrt{2}} \right). \quad (5.15)$$

Thus for $t < t_{\text{min}}$, $X = 0$ has one ray leaving and for $t > t_{\text{min}}$ one ray leaves and at least one returns.

Next we consider the geometric envelope(s) of the rays, which Figure 6 suggests will play a role in the analysis. Writing the rays as

$$X = \mathcal{F}(t, \alpha) = \left( \frac{1}{2} + \frac{X_0}{2\alpha} \right) e^{\alpha t} + \left( \frac{1}{2} - \frac{X_0}{2\alpha} \right) e^{\alpha t} - 1, \quad (5.16)$$

the envelope is obtained by setting $\partial \mathcal{F} / \partial \alpha = 0$, which yields

$$t - \alpha = \frac{1}{2} \log \left( \frac{\alpha^2 + X_0 - X_0}{\alpha^2 + X_0 - X_0} \right). \quad (5.17)$$

Setting equation (5.17) in equation (5.16) and performing some algebraic simplifications leads to the quartic equation

$$(X^2 + 2X)\alpha^4 - X_0[(X_0 + 2)X^2 + 2(X_0 + 2)X - X_0] \alpha^2 + X_0^2 X^2 + 2X^2 - X_0^2 - 2X_0) = 0, \quad (5.18)$$

which is also a quadratic equation for $\alpha^2$. Solving equation (5.18) yields

$$\alpha^2 = \frac{X_0}{2|X|(X + 2)} \left[(X_0 + 2)|X|(X + 2) + X_0 \pm \sqrt{X_0(X + 1)\sqrt{(X_0 + 4)(X + 1)^2 - 4}} \right]. \quad (5.19)$$
The expression in equation (5.19) is real only for $X > -1 + 2 / \sqrt{X_0 + 4} = X_{\text{cusp}}$. Then $z_+$ corresponds to $z_c$ in equation (2.23), and $z_-$ to $z_d$ in equation (2.25). The envelope(s) of the rays can then be obtained from equation (5.17), with $z_c$ corresponding to the “upper caustic” $t = t_c(X; X_0)$ in equation (2.22) and $z_d$ to the “lower caustic” $t = t_d(X; X_0)$ in equation (2.24). An alternate way of obtaining the caustic curves is to examine where the Jacobian between $(X, t)$ and $(z, \tau)$ coordinates vanishes (or is infinite). Using equations (5.10) and (5.11) to compute the Jacobian $\partial (X, t) / \partial (z, \tau)$ again leads to the two caustic curves $t_c$ and $t_d$, which are explicit functions of $X$ and $X_0$. When $X = X_{\text{cusp}}$, $z_c = z_d = \sqrt{X_0(X_0 + 3)} \equiv z_{\text{cusp}}$ (which exceeds both $X_0$ and $\sqrt{X_0(X_0 + 2)}$, the minimum $z$ needed for a ray to return to $X = 0$) and then $t_d = t_c = t_{\text{cusp}}$, as in equation (2.27). At $X = X_{\text{cusp}}$ the two caustic curves have the same slope, $dt/dX = \sqrt{(X_0 + 3)(X_0 + 4)/X_0}$, and thus form a cusp. Also, $t_c \to +\infty$ as $X \to 0^-$, but $t_d$ approaches a finite value, with $t_d(0; X_0) = t_{\min}(X_0)$ as in equation (5.15). When $X \to 0^-$, $z_c \sim X_0 / \sqrt{-2X} \to +\infty$ while $z_d \to \sqrt{X_0(X_0 + 2)}$.

We are now in a position to relate the present ray expansion results to those in Theorem 2.3. First consider $z < X_0$. Then we write $t$ in terms of $X$ and $z$ as

$$t = z + \log \left[ \frac{\sqrt{(X + 1)^2 - 1 + X_0^2}/z^2}{X_0/z - 1} \right]$$

(5.20)

and then $F^-$ in equation (5.12) becomes

$$F^- = -\frac{1}{4} \left[ (X + 1)^2 + \frac{X_0}{z} - 1 - (X + 1)\sqrt{(X + 1)^2 - 1 + \frac{X_0^2}{z^2}} \right] - \frac{X_0^2}{4z} - \frac{z}{4} + \frac{X_0}{2}.$$  

(5.21)

Comparing equation (5.21) with $F(X, t)$ in equation (2.38) we see they agree precisely if $z$ and $\phi_s$ (the saddle location) are related by $z = X_0 / \sqrt{1 + 4\phi_s}$. Then also equation (5.20) is equivalent to equation (2.35). We have thus established the equivalence of the ray and saddle point approaches, at least for the exponential parts of the approximation, and for $z < X_0$, which corresponds to $0 < t < X_0 - \log(1 + X)$ if $X \in (-1, 0)$, and to all $t$ if $X \leq -1$.

Now take $z > X_0$. Then we solve again equation (5.16) for $t$ in terms of $X$ and $z$, but now must consider both branches of the quadratic, hence

$$t = z + \log \left[ \frac{X + 1 - \sqrt{(X + 1)^2 - 1 + X_0^2}/z^2}{1 - X_0/z} \right],$$

(5.22)

or

$$t = z + \log \left[ \frac{X + 1 + \sqrt{(X + 1)^2 - 1 + X_0^2}/z^2}{1 - X_0/z} \right].$$

(5.23)

When equation (5.22) holds, we differentiate along a ray (thus with $z$ fixed) to get

$$\frac{dt}{dX} = - \left[ (X + 1)^2 - 1 + \frac{X_0^2}{z^2} \right]^{-1/2} < 0.$$
If equation (5.23) holds then $dt/dX$ is the negative of equation (5.24) and thus $dt/dX > 0$. Then along a ray $dt/dX$ changes sign (or $dX/dt = 0$) when $\alpha = X_0/\sqrt{1 - (1 + X)^2}$. But then equations (5.22) and (5.23) lead to precisely the curve $t_\ast = t_\ast(X; X_0)$ in equation (2.21). This curve is not itself a ray, but represents the locus of the minimal values of $X$ (or maximal $|X|$) that the rays attain, before changing slope and returning to $X = 0$ at the time in equation (5.14). This is also illustrated in Figure 6 (the dashed curve).

Now assume that $X_0 - \log(1 + X) < t < t_\ast$, with $X \in (-1, 0)$, and that we are in the range where $dt/dX < 0$. We define

$$R_1(\alpha) = t - \alpha - \log \left[ \frac{X + 1 - \sqrt{(X + 1)^2 - 1 + X_0^2/\alpha^2}}{1 - X_0/\alpha} \right]$$

(5.25)

and show that $R_1(\alpha)$ has a unique zero. We have, from equation (5.25),

$$R'_1(\alpha) = -1 + \frac{X_0}{\alpha^2 - X_0^2} - \frac{X_0^2(X + 1)}{\alpha(\alpha^2 - X_0^2)\sqrt{(X + 1)^2 - 1 + X_0^2/\alpha^2}}$$

and we shall show that $R'_1(\alpha) < 0$, or

$$\frac{X_0^2}{\alpha}(X + 1) > (X_0^2 + X_0 - \alpha^2)\sqrt{(X + 1)^2 - 1 + X_0^2/\alpha^2}. \quad (5.26)$$

If $\alpha > \sqrt{X_0(X_0 + 1)}$, then equation (5.26) is obviously true. If $X_0 < \alpha < \sqrt{X_0(X_0 + 1)}$ we square both sides of equation (5.26) and after some algebraic manipulation obtain the equivalent inequality

$$\alpha^3 - (X_0^2 + X_0)\alpha + X_0^2 > 0. \quad (5.27)$$

But equation (5.27) has roots at $\alpha = X_0$ and $\alpha = \left[ -X_0 \pm \sqrt{X_0(X_0 + 4)} \right]/2 < X_0$. Hence equation (5.27) is true for all $\alpha > X_0$ and thus $R'_1(\alpha) < 0$. We also have $R_1(\alpha) \to t - X_0 + \log(X + 1) > 0$ as $\alpha \to X_0^+$ and

$$R_1 \left( \frac{X_0}{\sqrt{1 - (1 + X)^2}} \right) = t - t_\ast < 0, \quad (5.28)$$

and thus $R_1(\alpha)$ has a unique zero. Then we again obtain equation (5.21), which is equivalent to the saddle point approximation in equation (2.37). We have now covered Regions I and III via the ray approach.

We shall next analyse Regions II, VI and VII, where $dt/dX > 0$ and $t > t_\ast$. Then $R_1(\alpha) = 0$ has no solutions, and we need to analyse the root(s) of

$$R_2(\alpha) = t - \alpha - \log \left[ \frac{X + 1 + \sqrt{(X + 1)^2 - 1 + X_0^2/\alpha^2}}{1 - X_0/\alpha} \right],$$

whose derivative is

$$R'_2(\alpha) = -1 + \frac{X_0}{\alpha^2 - X_0^2} + \frac{X_0^2(X + 1)}{\alpha(\alpha^2 - X_0^2)\sqrt{(X + 1)^2 - 1 + X_0^2/\alpha^2}}.$$
For Region II, we shall show that $R_2'(z) > 0$, or
\[
\frac{X_0^2}{z}(X+1) > (z^2 - X_0^2 - X_0)\sqrt{(X+1)^2 - 1 + \frac{X_0^2}{z^2}}. 
\] (5.29)
If $X_0 < z \leq \sqrt{X_0(X_0 + 1)}$, equation (5.29) is obviously true. If $z > \sqrt{X_0(X_0 + 1)}$ we square both sides of equation (5.29) and obtain the equivalent inequality
\[
\left[ \frac{X_0^4}{z^2} - (z^2 - X_0^2 - X_0)^2 \right] (X+1)^2 > \left( \frac{X_0^2}{z^2} - 1 \right) (z^2 - X_0^2 - X_0)^2. 
\] (5.30)
Clearly $X_0^2/z^2 - 1 < 0$ so if $X_0^4/z^2 - (z^2 - X_0^2 - X_0)^2 \geq 0$, which occurs if $X_0 < z \leq [X_0 + \sqrt{X_0(X_0 + 4)}]/2$, then equation (5.30) is true for any $X \in (-1, 0)$. If, on the other hand, $X_0^4/z^2 - (z^2 - X_0^2 - X_0)^2 < 0$, then we multiply both sides of equation (5.30) by $-1$, changing $>$ to $<$, and obtain
\[
X + 1 < \sqrt{\frac{(z^2 - X_0^2)(z^2 - X_0^2 - X_0^0)^2}{z^2(z^2 - X_0^2 - X_0^0)^2 - X_0^0}}. 
\] (5.31)
But the inequality in equation (5.31) implies, in view of equation (5.18), that $X < X_{\text{cusp}}$. Thus $R_2'(z) > 0$ in Region II. Then $R_2(z) \to -\infty$ as $z \to X_0^+$ and
\[
R_2\left( \frac{X_0}{\sqrt{1-(1+X)^2}} \right) = t - t_\ast > 0, 
\] (5.32)
so that $R_2(z)$ has a unique root for $z > X_0$.

We next consider Regions VI and VII. We note that for $X \in (X_{\text{cusp}}, 0)$ the equation $R_2'(z) = 0$ has roots precisely at $z_c$ and $z_d$ (or $z_{\pm}$ in equation (5.19)).

For Region VI, where $t > t_\ast$ and $X \in (X_{\text{cusp}}, 0)$, we have $R_2(z_d) = t - t_d > 0$, $R_2(z_c) = t - t_c > 0$, $R_2(X_0^*) = -\infty$ and the equality in equation (5.32) holds. Thus $R_2(z)$ has a unique root for $z > X_0$.

For Region VII, where $t_\ast < t < t_d$ and $X \in (X_{\text{cusp}}, X_\ast)$, we have $R_2(X_0^*) = -\infty$ and the equality in equation (5.32) holds as usual, but now $R_2(z_d) = t - t_d < 0$ and $R_2(z_c) = t - t_c < 0$. Again, $R_2(z)$ has a unique root for $z > X_0$.

We thus have shown that in Regions VI, VI and VII, $R_2(z)$ has a unique root for $z > X_0$. If we identify $z = X_0/\sqrt{1-4z}$, and note that, using equations (5.23) and (5.10)–(5.12),
\[
F^- = -\frac{1}{4} \left[ (X+1)^2 + \frac{X_0}{z} - 1 + (X+1)\sqrt{(X+1)^2 - 1 + \frac{X_0^2}{z^2}} \right] - \frac{X_0^2}{4z} - \frac{z}{4} + \frac{X_0}{2}, 
\] (5.33)
then $F^-$ agrees with $\tilde{f}(z_1)$ (cf. equation (2.45)) or $\tilde{f}(z_3)$ (cf. equation (2.46)). Here we use the expression below equation (2.45) for $\tilde{f}(z)$ and the fact that $z_1$ or $z_3$ is the unique solution to equation (2.36), which is equivalent to equation (5.23).

Next we examine the interior of the caustics, where $t_d < t < t_\ast$, so certainly $X \in (X_{\text{cusp}}, 0)$. This corresponds to Regions IV and V. The equation $R_1(z) = 0$ has a unique zero in Region IV and no solution in Region V. Note that equation (5.28) holds in Region IV. However, $R_2(z)$ will now have multiple roots, and also $R_2(z_c) = t - t_c < 0,$
\( R_2(x_d) = t - t_d > 0, \) \( R_2(X_0^+) = -\infty \) and the equality in equation (5.32) holds. When \( t > t_1, t > t_d \) and \( t < t_e \) (Region V) we have three sign changes in \( R_2(x) \) over the interval \( x \in (X_0, X_0/\sqrt{1 - (1 + X)^2}) \), and we denote the three roots by \( x_1 < x_2 < x_3 \), where \( x_j = x_j(X, t) \). When \( t_d < t < t_e \) (Region IV) there are two sign changes and we call the roots \( x_1 < x_2 \), and then let \( x_3 \) be the unique root of \( R_1(x) = 0 \). For both Regions IV and V, a given \((X, t)\) corresponds to three distinct values of \((x, \tau)\), so that the mapping from ray to \((X, t)\) coordinates in 3-to-1. In the five other regions this mapping is 1-to-1 and only a single ray reaches a point \((X, t)\) that is within Regions I–III, VI or VII.

Omitting the details, given \( F^- \) in the single valued case we can easily solve the linear PDE in equation (5.5) subject to equation (5.9), with the result

\[
G^- = \sqrt{\frac{X_0}{2\pi}} \left| (x^2 + \alpha X_0 - X_0) e^{-2(t-\tau)} - (x^2 - \alpha X_0 - X_0) \right|^{-1/2}. \tag{5.34}
\]

We can show that equation (5.34) agrees with \( G \) in equation (2.37), and also with \( \tilde{G} \) in equations (2.45) and (2.46). Then with equations (5.34) and (5.12) we have the leading order ray approximation \( p \sim G^- e^{F^-/\epsilon} \), and this applies in the union of Regions I–III, VI and VII, including the curves that separate them. We observed already in Section 4 that \( Ge^{F/\epsilon} \) in equation (2.37) may be smoothly continued from Regions I and III into II, VI and VII, and in particular is smooth along \( t = t_e \). It is only upon trying to eliminate \( t = x + \tau \) in equation (5.12) that leads to the two different expressions in equations (5.21) and (5.33). Outside the caustic range, the rays with \( dt/dX < 0 \) correspond to the saddle point \( \phi_s \) in the integral in equation (4.1), while rays with \( dt/dX > 0 \) correspond to the unique interior maximum of \( f \) in equation (4.15) or (4.25), which arises from the residue sum in equation (4.1).

In Regions IV and V, three rays from \( X = 0 \) reach a given \((X, t)\), so we write

\[
p(x, t) \sim G_{1}^- e^{F_{1}^-/\epsilon} + G_{2}^- e^{F_{2}^-/\epsilon} + G_{3}^- e^{F_{3}^-/\epsilon}, \tag{5.35}
\]

where \( F_{j}^- \) corresponds to \((x_j, \tau_j)\) for \( j = 1, 2, 3 \). The dominant contribution to \( p \) will come from the maximal \( F_{j}^- \). In Region IV, \( F_{3}^- \) arises from rays with \( dt/dX < 0 \) and this term corresponds to \( Ge^{F/\epsilon} \) in equation (2.48). But equation (2.48) has only two terms while equation (5.35) has three. We recall that in Region IV our analysis of the residue sum in equation (4.1) led to the equation \( \tilde{f}'(z) = 0 \) which had two roots \( z_1 < z_2 \), with \( z_1 \) (resp. \( z_2 \)) being a maximum (resp. minimum) of \( \tilde{f}(z) \). But the Laplace method would never consider minima, and \( \tilde{f}(z_1) \) in equation (2.48) corresponds to \( F_{1}^- \) in equation (5.35), while \( F_{2}^- \) would correspond to the minimum of \( \tilde{f} \) at \( z = z_2 \), since it is also a root of \( \tilde{f}'(z) = 0 \). Geometrically, for every \((X, t)\) in Region IV, three rays meet, one with \( dt/dX < 0 \) (\( F_{3}^- \)) and two with \( dt/dX > 0 \) (\( F_{1}^- \) and \( F_{2}^- \)). But \( F_{1}^- > F_{2}^- \) so the middle term in equation (5.35) can never be dominant asymptotically. Also, the term with \( F_{2}^- \) corresponds to a ray that has been tangent to one of the caustic curves before reaching \((X, t)\), while the \( F_{1}^- \) ray has not been tangent to a caustic. For Region V comparing equation (5.35) to equation (2.47) we now have the correspondence \((f(z_1), \tilde{f}(z_1)) \leftrightarrow (F_{1}^-, F_{3}^-) \) and the middle term in equation (5.35) is again negligible, corresponding to the local minimum of \( \tilde{f} \) at \( z = z_2 \), and the \( F_{2}^- \) ray has been tangent to a caustic before reaching \((X, t)\). In Region V all three rays that reach \((X, t)\) have \( dt/dX > 0 \).
Thus we may drop the middle term in equation (5.35) as it is exponentially smaller than one of the other two terms. Along the caustic \( t = t_c \) we have \( F_1^- = F_2^- \) but then the \( F_3^- \) term dominates. Along \( t = t_c \) we have \( F_2^- = F_3^- \) but then \( F_1^- \) dominates. We also have \( G_1^- \) and \( G_2^- \), obtained from equation (5.34), becoming singular along \( t = t_d \), while \( G_2^- \) and \( G_3^- \) become singular along the caustic \( t = t_c \). But the \( G_j^- \) corresponding to the dominant term in equation (5.35) will not be singular along the caustic. However, at the cusp point \((X, t) = (X_{\text{cusp}}, t_{\text{cusp}})\), \( F_1^- \), \( F_2^- \) and \( F_3^- \) all agree, and the corresponding \( G_j^- \) all become singular. Thus here a genuine non-uniformity develops, which we analysed using the Laplace method and obtained equation (2.50). It should also be possible to start directly from the PDE and analyse a vicinity of the cusp, say using the local \((\xi, \eta)\) variables in equation (2.49), and thus obtain equation (2.50) more directly. However, this would require significant additional work which we do not attempt here. Deciding whether \( F_1^- \) or \( F_3^- \) in equation (5.35) dominates amounts to solving \( F_1^- = F_3^- \), and this leads to the implicit curve \( t_{\Gamma} \) which we already discussed, and evaluated numerically. Certainly \( t_d < t_{\Gamma} < t_c \) and \( t_{\Gamma} \) hits \( X = 0 \) and a finite point.

Now we return to the range where \( X > 0 \). We must consider the effects of the rays that returned to \( X = 0 \) after traveling in the range \( X < 0 \), on the range \( X > 0 \). We recall that these rays hit \( X = 0 \) at all times \( t \geq t_{\text{min}} \), cf. equation (5.15). The rays that return to \( X = 0 \) continue into the range \( X > 0 \) and maintains the same slope \( dt/dX > 0 \) at \( X = 0^- \). The continuity of ray slopes follows from the continuity of \( F_X \) at \( X = 0 \), and this follows from equations (5.2) and (5.3). Let us denote this new ray family for \( X > 0 \) by \( F^R \), so \( F^R \) will again satisfy the PDE in equation (5.2). But now the initial condition for \( F^R \) at \( X = 0 \) must be obtained from the value of \( F^-(0^-, t) \) for a ray that returned. By differentiating equation (5.16) implicitly along a ray, we find that \( dt/dX = \alpha/X_0 \), at the time in equation (5.14) when the ray returns to \( X = 0 \). Then we find that the “returned” rays correspond to

\[
X = \frac{X_0}{\alpha} \left[ t - \frac{\alpha}{\alpha - X_0} \log \left( \frac{X + X_0}{X - X_0} \right) \right] = \frac{X_0}{\alpha} \tau, \tag{5.36}
\]

which is a family of straight lines. Also, equation (5.36) defines the \( \tau \) variable for this new family, with \( \tau > 0 \) and \( \alpha > \sqrt{X_0^2(X_0 + 2)} \). To obtain \( F^R \) we solve equation (5.3) subject to equation (5.33) at \( X = 0 \), which yields

\[
F^R(0^+, t) = \frac{X_0}{2} - \frac{\alpha}{4} - \frac{X_0^2}{4\alpha} - \frac{X_0}{2\alpha}.
\]

We let \( \alpha = X_0/\sqrt{1 - 4z_*} \) where now \( z_*(X, t) \) satisfies

\[
t = \frac{X + X_0}{\sqrt{1 - 4z_*}} + \log \left[ \frac{1 + \sqrt{1 - 4z_*}}{1 - \sqrt{1 - 4z_*}} \right].
\]

Then the solution \( F^R \) is given by

\[
F^R(X, t) = -\left( \frac{1}{2} + \frac{X_0}{2\alpha} \right)^2 \tau + \frac{X_0}{2} - \frac{\alpha}{4} - \frac{X_0^2}{4\alpha} - \frac{X_0}{2\alpha} = \frac{X_0 - X}{2} - \frac{X + X_0 + 1}{2} \sqrt{1 - 4z_*} - \frac{z_*(X + X_0)}{\sqrt{1 - 4z_*}},
\]
which agrees with \( f(z^*) \) in equation (2.10) (with (2.11)). By solving for \( G^R(X, t) \), using continuity with \( G^- \) at \( X = 0 \), we can similarly obtain the factor in the right side of equation (2.9) that multiplies \( e^{f(z_*)/\varepsilon} \).

The ray family in equation (5.36) also has a caustic curve, which can be obtained by setting the derivative with respect to \( z \) equal to zero, and then eliminating \( z \). This yields the curve

\[
t = t_+(X; X_0) = \sqrt{(X + X_0)(X + X_0 + 2)} + 2 \log \left[ \frac{\sqrt{X + X_0 + 2} + \sqrt{X + X_0}}{\sqrt{2}} \right],
\]

and this is the same as that in equation (3.9). Also, the returned rays fill only the domain \( t > t_+ \) (and \( X > 0 \)) and are absent for \( t < t_+ \). We have \( t_+(0; X_0) = t_d(0; X_0) \) so we may view \( t_+ \) as a caustic in \( X > 0 \) that was induced by the lower caustic \( t_d \) in the range \( X < 0 \). Now consider both the direct ray and returned ray expansions for \( X > 0 \). For \( t < t_+ \) only the direct rays are present and then \( p \sim p_{BM} \). For \( t > t_+ \) every point \((X, t)\) is reached by three rays, the direct path from \((X_0, 0)\), and two “returned” rays from \( X = 0 \). One returned ray reached \((X, t)\) by first being tangent to the caustic \( t_+ \), and the other returned ray reached \((X, t)\) before being tangent to the caustic. The former returned rays correspond to the minimum of \( f(z) \) in equation (B.13) at \( z_{**} \), which we discussed below equation (B.18). But again clearly \( f(z_{**}) < f(z_*) \) so we have, for \( t > t_+ \),

\[
p(x, t) \sim G^R e^{F^R/\varepsilon} + p_{BM}.
\]

Along \( t = t_+ \), \( p_{BM} \) dominates, but for times sufficiently large the first term in equation (5.38) overtakes the direct ray solution. Here \( F^R \) is understood to arise from the returned ray that was not tangent to \( t_+ \) before hitting \((X, t)\). The transition from \( p_{BM} \) to the returned ray result occurs precisely when \( F^R = -(t + X - X_0)^2/(4t) \), and this corresponds to equation (2.18) and the curve \( t = t_d(X; X_0) \).

In Figure 7 we sketch the returned rays in the range \( t > t_+ \) above the caustic for \( X_0 = 0.1 \). Note that for larger values of \( X_0 \), such as \( X_0 \geq 1 \), equation (5.37) numerically resembles a straight line (and in fact \( t_+ \sim X \) as \( X \to \infty \)), and it is difficult to see the curvature of \( t_+ \) or the tangency of the rays to this curve. In Figure 8, we sketch all the rays for both \( X < 0 \) and \( X > 0 \), which illustrates the multi-valuedness. Note that numerically part of the curve \( t = t_+ \) where the rays change slope is hard to distinguish from the upper caustic \( t = t_c \). Our analysis here shows that the ray approach is much different from the saddle/singularity analysis, but ultimately yields equivalent results. Either method may be used to obtain higher order terms in equation (5.1).

### 6 Concluding remarks

We provided a thorough analysis of the asymptotic behaviours of the transient distribution under the scale in equation (1.1). The results provided some insight into how the \( M/M/m \) model tends to process large queues, and how equilibrium is achieved for large \( \beta \). The classical \( M/M/m \) queueing model, also called the Erlang-C model, is the model most often used in call-centre analysis. One measure of service quality is the probability that a caller must wait in the queue before service, i.e., \( \Pr[N(t) > m] \), which is equivalent.
Figure 7. Returned rays in the range \( t > t_+ \) above the caustic with \( X_0 = 0.1 \) (the dashed curve is \( t_+ \)).

to \( \Pr[X_d(t) > 0] \). Thus, Theorem 2.2 is of particular importance. The approximations equations (2.16) and (2.41) are also important. These two expressions show how the steady state in equation (2.3) is slowly achieved on the \( T \)-scale equation (2.15), as \( p(x,t) \) decreases toward \( p(x,\infty) \) as a double exponential function in \( T \).

From a methodological viewpoint, our asymptotic analysis had some novel features. The problem of a saddle point coalescing with a pole (or another algebraic singularity) has been previously studied in detail, see [3] and [27]. Our analysis led to an integrand with a unique saddle and also many closely spaced simple poles, which coalesce with the saddle for certain parameter values. But the asymptotic transition(s) caused by this coalescence is quite unlike the classic results in [3] and [27]. Our analysis also involved two caustic curves that come together at a cusp point in the space-time \((x,t)\) plane, and we were able to precisely analyse the neighbourhood of the cusp point, using a type of discrete version of the Laplace method.

In relation to the discrete \( M/M/m \) model, the present analysis involved a double limit, first letting \( m \to \infty \) with \( \beta = \sqrt{m[1-\lambda/(\mu m)]} \) fixed, and then taking \( \beta \to \infty \). Thus a natural question is how large can \( \beta \) be, in relation to \( m \), so that the present asymptotic results still
have relevance to the discrete queue. For the discrete model we can take $\frac{\lambda}{\mu} = m\nu$ with $\nu < 1$, which means that the system is, for $m \rightarrow \infty$, lightly loaded and in the steady state customers will rarely need to queue. If the initial number of customers is $N(0) = mn_*$ with $n_* > 1$, then as time evolves the number of customers will decrease, on average, from this initial value to the value $N(\infty) = \frac{\lambda}{\mu} = m\nu < m < N(0)$. This situation is completely analogous to the diffusion process starting at $X_d(0) = x_0 > 0$ and migrating to the stable equilibrium at $X_d = -\beta$. It is possible to give an explicit expression for the transient probability distribution, $\Pr[N(t) = n|N(0) = n_0]$, of the discrete queue. But this expression would be even more complicated than those in Theorem 2.1, and the corresponding integrals would involve hypergeometric functions. We believe that ultimately the discrete model could also be analysed asymptotically, say in the limit $m \rightarrow \infty$ with $N(t)$ and $N(0)$ scaled to be $O(m)$. If $\nu < 1$ the asymptotic results would be different from the results here, but we think that the overall structure, as far as the different regions in the $(x, t)$ plane in Theorems 2.2 and 2.3, of the problem would be similar. Letting $\beta \rightarrow \infty$ in the diffusion model and letting $\nu \uparrow 1$ and $n_* \downarrow 1$ in the discrete model should correspond to an asymptotic matching region between the two, and this would imply that $\beta$ could be made as large as $m^{\gamma}$ for any $\gamma < 1/2$. However, the study of such a matching region would involve an in depth asymptotic analysis of the discrete queue, and this we did not attempt here.
Acknowledgement

Charles Knessl was partly supported by NSA grants H 98230-08-1-0102 and H 98230-11-1-0184.

References

[1] Abramowitz, M. & Stegun, I. A. (1972) Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables (Tenth Printing), Dover, New York.

[2] Bender, C. M. & Orszag, S. A. (1978) Advanced Mathematical Methods for Scientists and Engineers, McGraw-Hill, New York.

[3] Bleistein, N. & Handelsman, R. A. (1986) Asymptotic Expansions of Integrals, Dover, New York.

[4] Borst, S. Mandelbaum, A. & Reiman, M. (2004) Dimensioning large call centers. Oper. Res. 52, 17–34.

[5] Brockmeyer, E. Halstrøm, H. L., & Jensen, A. (1948) The life and works of A. K. Erlang. Trans. Danish Acad. Tech. Sci. 2, 277.

[6] Dembo, A. & Zeitouni, O. (1993) Large Deviations Techniques and Applications, Jones and Bartlett, Boston.

[7] Flajolet, P. & Sedgwick, R. (2009) Analytic Combinatorics, Cambridge University Press, Cambridge.

[8] Friedlin, M. A. & Wentzell, A. D. (1984) Random Perturbations of Dynamical Systems, Springer-Verlag, New York.

[9] Gamarnik, D. & Goldberg, D. A. (2013) On the rate of convergence to stationarity of the $M/M/N$ queue in the Halfin–Whitt regime. Ann. Appl. Probab. 23, 1879–1912.

[10] Gans, N. Koole, G. & Mandelbaum, A. (2003) Telephone call centers: tutorial, review and research prospects. Manuf. Serv. Oper. Manag. 5, 79–141.

[11] Garnett, O. Mandelbaum, A. & Reiman, M. (2002) Designing a call center with impatient customers. Manuf. Serv. Oper. Manag. 4, 208–227.

[12] Gradshteyn, I. S. & Ryzhik, I. M. (2007) Table of Integrals, Series and Products, 7th ed., Elsevier/Academic Press, Amsterdam.

[13] Halfin, S. & Whitt, W. (1981) Heavy-traffic limits for queues with many exponential servers. Oper. Res. 29, 567–588.

[14] Janssen, A. J. E. M., van Leeuwaarden, J. S. H. & Zwart, B. (2011) Refining square-root safety staffing by expanding Erlang C. Oper. Res. 59, 1512–1522.

[15] Jelenković, P. Mandelbaum, A. & Momčilović, P. (2004) Heavy traffic limits for queues with many deterministic servers. Queueing Syst. 47, 53–69.

[16] van Leeuwaarden, J. S. H. & Knessl, C. (2011) Transient behavior of the Halfin–Whitt diffusion. Stoch. Process. Appl. 121, 1524–1545.

[17] Maglaras, C. & Zeevi, A. (2004) Diffusion approximations for a multiclass Markovian service system with “guaranteed” and “best-effort” service levels. Math. Oper. Res. 29, 786–813.

[18] Magnus, W. Oberhettinger, F. & Soni, R. P. (2008) Formulas and Theorems for the Special Functions of Mathematical Physics, Springer-Verlag, New York.

[19] Mandelbaum, A. & Momčilović, P. (2008) Queues with many servers: The virtual waiting-time process in the QED regime. Math. Oper. Res. 33, 561–586.

[20] Olver, F. W. J. (1959) Uniform asymptotic expansions for Weber parabolic cylinder functions of large orders. J. Res. Nat. Bur. Standards Sect. 63B, 131–169.

[21] Pollaczek, F. (1931) Über zwei Formeln aus der Theorie des Wartens vor Schaltergruppen. Elektr. Nachr. 8, 256–268.

[22] Reed, J. (2009) The $G/GI/N$ queue in the Halfin–Whitt regime. Ann. Appl. Probab. 19, 2211–2269.
Appendix A Proof of Lemma 4

We consider the integral in equation (3.5), and use the fact that $\beta \to +\infty$ to simplify the function $H(\theta)$. Using asymptotic properties of the parabolic cylinder functions $D_{-\theta}(-\beta)$ and scaling $\theta$ as $\theta = \varepsilon/\phi$ and $\theta_{sa} = \varepsilon/\phi^\ast$, we obtain the following result (see [16])

$$
\frac{D'_{-\theta}(-\beta)}{D_{-\theta}(-\beta)} = -\frac{1}{\sqrt{\varepsilon}} \sqrt{\phi + \frac{1}{4} + \frac{\sqrt{\varepsilon} [1 + \sqrt{1 + 4\phi}]}{2(1 + 4\phi)}} + O(\varepsilon^{3/2}).
$$

(A 1)

Thus we are taking $\beta \to +\infty$ with $\theta = O(\varepsilon^{-1}) = O(\beta^2)$. The result in equation (A 1) applies for $\phi > 0$, and also for $\phi \in (-1/4, 0)$ as long as we remain bounded away from the zeros of $D_{-\theta}(-\beta) = D_{-\phi}(1/\sqrt{\varepsilon})$. For $\varepsilon \to 0^+$ these zeros will be very close to $\phi = -\varepsilon N$ (then $\theta = -N$) where $N$ is an integer. We shall show that the poles of $H(\theta)$ are also very close to negative integer values. Thus as long as $-\theta_{sa}$ in equation (3.5) is not close to a positive integer, we can use equation (A 1) to approximate $H(\theta)$, thus obtaining

$$
H(\theta) = \frac{1 + \sqrt{1 + 4\phi}}{2(1 + 4\phi)^2} \varepsilon^{3/2} + O(\varepsilon^{5/2}).
$$

(A 2)

Note that the second term in the expansion in equation (A 1) is needed to obtain the leading term in equation (A 2), since $\sqrt{\theta + \beta^2/4 - R_\phi(\theta)} \sim 2\sqrt{\theta + \beta^2/4} = \sqrt{4\theta + \beta^2}$ in this limit. With equation (A 2) and the scaling $(x, x_0) = \beta(X, X_0)$ the integral in equation (3.5) becomes

$$
\text{INT} = \frac{\sqrt{\varepsilon}}{2\pi i} e^{(X_0 - X)/(2\varepsilon)} \int_{B(\phi^\ast)} \frac{1 + \sqrt{1 + 4\phi}}{2(1 + 4\phi)^2} \left[ 1 + O(\varepsilon) \right] \times \exp \left\{ \frac{1}{\varepsilon} \left[ \phi t - \frac{X + X_0}{2} \frac{\sqrt{1 + 4\phi}}{1 + 4\phi} \right] \right\} d\phi.
$$

(A 3)

The right side of equation (A 3) is in the standard form for applying the saddle point method for $\varepsilon \to 0^+$. There is a saddle at

$$
\frac{d}{d\phi} \left[ \phi t - \frac{X + X_0}{2} \sqrt{1 + 4\phi} \right] = 0, \quad \text{i.e.,} \quad \phi = \frac{(X + X_0)^2}{4t^2} - \frac{1}{4} = \phi^\ast(X, t).
$$

By integrating through the saddle in the imaginary direction we obtain the standard estimate in equation (3.7).
Appendix B  Proof of Lemma 5

To estimate the sum in equation (3.5) for $t > X + X_0$, we need to evaluate the residue(s) $h_N$. We first locate the poles $\theta_N$ for $\beta \to +\infty$ with $N = O(\beta^2)$. Note that for $N = O(1)$, equation (3.2) applies. Using the symmetry relation (see [12], p. 1030)

$$D_{-\theta}(-\beta) = e^{-\pi i \theta} D_{-\theta}(\beta) + \frac{\sqrt{2\pi}}{\Gamma(\theta)} e^{\pi i (1-\theta)/2} D_{\theta-1}(i\beta), \quad (B1)$$

and noting that the poles satisfy

$$\frac{d}{d\beta} D_{-\theta}(-\beta) + \sqrt{\theta + \frac{\beta^2}{4}} D_{-\theta}(-\beta) = 0, \quad (B2)$$

we must solve

$$D'_{-\theta}(\beta) + \sqrt{\theta + \frac{\beta^2}{4}} D_{-\theta}(\beta) = -ie^{\pi i \theta/2} \left( \frac{d}{d\beta} D_{\theta-1}(i\beta) + \sqrt{\theta + \frac{\beta^2}{4}} D_{\theta-1}(i\beta) \right). \quad (B3)$$

In the limit $\beta \to +\infty$ with $\theta = O(\beta^2)$ we have the estimates (see [1]), for $\beta^2 + 4\theta > 0$,

$$D_{-\theta}(\beta) \sim \left( 1 + \frac{\theta}{W^2} \right)^{1/2} W^{-\theta} \exp \left[ \frac{\theta}{2} - \frac{\beta^2}{4\sqrt{\beta^2 + 4\theta}} \right], \quad (B4)$$

$$W = \frac{1}{2} \left[ \beta + \sqrt{\beta^2 + 4\theta} \right] = \frac{1}{2\sqrt{\theta}} \left[ 1 + \sqrt{1 + 4\phi} \right],$$

and

$$D_{\theta-1}(i\beta) \sim -ie^{\pi i \theta/2} \left( 1 + \frac{\theta}{W^2} \right)^{1/2} W^{\theta-1} \exp \left[ -\frac{\theta}{2} + \frac{\beta^2}{4\sqrt{\beta^2 + 4\theta}} \right]. \quad (B5)$$

Also, analogous estimates of the derivatives show that

$$\frac{D'_{-\theta}(\beta)}{D_{-\theta}(\beta)} = -\sqrt{\theta + \frac{\beta^2}{4}} - \frac{1}{2(\beta^2 + 4\theta)} \left[ \beta - \sqrt{\beta^2 + 4\theta} \right] + O(\beta^{-3}), \quad (B6)$$

$$\frac{\frac{d}{d\beta} D_{\theta-1}(i\beta)}{D_{\theta-1}(i\beta)} \sim \sqrt{\theta + \frac{\beta^2}{4}}.$$  

Now, we know that the solutions of equations (B2) and (B3) are on the real axis. But in view of equations (B4) and (B5) the ratio of the left side of equation (B3) to the bracketed factor in the right side is roughly of the order of $O \left[ \exp \left( \theta - \beta \sqrt{\theta + \beta^2/4} \right) \right]$, which is exponentially small in this limit. Thus if equation (B3) is to hold then $1/\Gamma(\theta)$ must also be exponentially small, so that $\theta$ must be very close to a zero of $\Gamma^{-1}(\theta)$, or a pole of $\Gamma(\theta)$, and these occur at $\theta = 0, -1, -2, \ldots$. We know that $\theta = 0$ is a pole exactly, and near a pole we can approximate $\Gamma$ by

$$\frac{1}{\Gamma(\theta)} = (-1)^N N!(\theta + N) + O((\theta + N)^2), \quad (B7)$$
which is just a Taylor expansion. Using equations (B 7) and (B 6) we can replace equation (B 3) by the asymptotic relation
\[ -\frac{1}{2} \frac{1}{\beta^2 + 4\theta} \left[ \beta - \sqrt{\beta^2 + 4\theta} \right] \frac{D_{-\theta}(\beta)}{D_{\theta-1}(i\beta)} \sim \sqrt{2\pi}(-1)^{N+1}i\Gamma(1)(\theta + N) e^{\pi i/2} \sqrt{\beta^2 + 4\theta} \quad (B 8) \]
and this can be used to estimate \( \theta_N + N \), which we summarize below.

**Lemma 10** The roots \( \theta_N \) of (3.1) for \( \beta = 1/\sqrt{\theta} \to +\infty \) satisfy \( \theta_N \sim -N \) with the correction term
\[ \theta_N + N \sim \frac{e^{-N}(\beta^2 - 4N)^{-3/2}}{\sqrt{2\pi(N-1)!}} \exp \left[ \frac{-\beta}{2} \frac{\sqrt{\beta^2 - 4N}}{2} \right]. \quad (B 9) \]

Note that equation (B 9) remains valid for \( N = O(1) \), as then it reduces to equation (3.2), since for \( N \ll \beta^2 \) we have \( \beta \sqrt{\beta^2 - 4N} = \beta^2 - 2N + o(1) \). Expression (B 9) thus applies for all \( N = O(1) \) and \( N = O(\beta^2) \), as long as \( N/\beta^2 < 1/4 \). Lemma 10 follows immediately from equation (B 8) by using equations (B 4) and (B 5) to estimate the parabolic cylinder functions, and asymptotically replacing \( \theta \) by \( -N \) in all terms except \( \theta + N \). Also, equation (B 9) is consistent with \( \theta_0 = 0 \), if we define \((-1)! = \Gamma(0) = \infty \). For \( N \gg 1 \), we can approximate \((N-1)! = \Gamma(N) \) by \( N^N e^{-N} \sqrt{2\pi}/N \), due to Stirling’s formula.

Having estimated the location of the poles we now evaluate asymptotically the residuals \( h_N \) in equation (3.5). We use equations (3.4) and (B 1) and note that only the second term in equation (3.4) will contribute to the residues. Thus,
\[ \frac{1}{\sqrt{\theta + \frac{\beta^2}{4} - R_\theta(\theta)}} = \frac{D_{-\theta}(\beta)}{\sqrt{\theta + \frac{\beta^2}{4}}} D_{-\theta}(\beta) + \frac{d}{d\beta} D_{-\theta}(\beta) \]
\[ = \sqrt{\theta + \frac{\beta^2}{4}} D_{-\theta}(\beta) + D_{-\theta}(\beta) + i\sqrt{2\pi} D_{-\theta}(\beta) \exp \left[ \frac{\pi i \theta}{\beta} \right] \frac{D_{\theta-1}(i\beta)}{\sqrt{\theta + \frac{\beta^2}{4}}} D_{\theta-1}(i\beta) \]
and the residue will be the numerator evaluated at \( \theta = \theta_N \), divided by the derivative of the denominator at \( \theta = \theta_N \). Since \( \Gamma^{-1}(\theta) \sim (-1)^N N!/(\theta + N) \) we use equations (B 9), (B 4) and (B 5) to conclude that the numerator is asymptotically the same as \( D_N(\beta) \), while the derivative of the denominator is asymptotically dominate by \( \frac{d}{d\theta} \Gamma^{-1}(\theta) \sim (-1)^N N! \), \( \theta \to \theta_N \). Hence the residue is asymptotically given by
\[ h_N \sim \frac{(-1)^N e^{\pi i N/2}}{i \sqrt{2\pi} N! \sqrt{\beta^2 - 4N}} \frac{D_N(\beta)}{D_{-\theta}(\beta)} \sim \frac{2(\theta_N + N)}{\beta - \sqrt{\beta^2 - 4N}} \beta^2 - 4N \]
\[ \sim \frac{1}{N!} \frac{\beta + \sqrt{\beta^2 - 4N}}{2} \left( \frac{\beta + \sqrt{\beta^2 - 4N}}{2} \right)^{2N} e^{-N} \exp \left[ \frac{-\beta}{2} \frac{\sqrt{\beta^2 - 4N}}{2} \right]. \quad (B 10) \]
Here we again used equations (B 4), (B 5), (B 8) and (B 9). Using equation (B 10) we have thus approximated the summand in equation (3.5) as follows.
Lemma 11 The summand in equation (3.5) has the asymptotic expansion

\[
\sqrt{\varepsilon} g(\varepsilon N) \exp \left[ \frac{1}{\varepsilon} f(\varepsilon N) \right], \quad 0 < \varepsilon N < \frac{1}{4},
\]  

where

\[
f(z) = f(z; X, t) = -z t - \frac{X + X_0}{2} \sqrt{1 - 4z} + \frac{X_0 - X}{2} \sqrt{1 - 4z} + 2z \log \left[ \frac{1 + \sqrt{1 - 4z}}{2\sqrt{z}} \right],
\]

\[
g(z) = \frac{1 + \sqrt{1 - 4z}}{4\pi \sqrt{1 - 4z} \sqrt{z}},
\]

and for \( N = O(1) \) the expansion is

\[
\frac{1}{N!} \frac{1}{\sqrt{2\pi}} \varepsilon^{-N} \exp \left[ -\frac{1}{\varepsilon} \left( X + \frac{1}{2} \right) \right] e^{-N(t - X - X_0)}. \]  

To obtain equation (B 11) we simply multiplied equation (B 10), after expanding \( N! \) by Stirling’s formula, by the factors that multiply \( h_N \) in equation (3.5), using also \( \theta_N \sim -N \) and \( x\sqrt{\theta_N} + \beta^2/4 \sim \frac{1}{2} \beta X \sqrt{\beta^2 - 4N} = \frac{1}{2} e^{-1} X \sqrt{1 - 4\varepsilon N}, \) etc. To obtain equation (B 12) we also used \( \sqrt{\beta^2 - 4N} \sim \beta = e^{-1/2} \) for \( N = O(1) \).

To evaluate asymptotically the sum in equation (3.5) we first note that for \( t = O(1) \) the expression in equation (B 12) becomes asymptotically larger with increasing \( N \), due to the factor \( \varepsilon^{-N} \). Then surely the main asymptotic behaviour cannot come from the range \( N = O(1) \), and we thus consider the sum

\[
\sum_{N=0}^{\lceil -\phi^{**}/\varepsilon \rceil} \sqrt{\varepsilon} g(\varepsilon N) e^{f(\varepsilon N)/\varepsilon}, \]  

for \( t > X + X_0 \), since then \( -\phi^{**} > 0 \). This is a Laplace type sum and the main contribution will come from the global maxima of the function \( f(z) \) over the range \( z \in (0, -\phi^{**}) \). Note that \( f(z) \) is smooth for \( z \in (0, 1/4) \), and \( -\phi^{**} < 1/4 \) for any fixed \( t > X + X_0 \). We have

\[
f'(z) = -t - \frac{X + X_0}{2} \sqrt{1 - 4z} + 2 \log \left( \frac{1 + \sqrt{1 - 4z}}{2\sqrt{z}} \right), \]  

and

\[
f''(z) = \frac{2(X + X_0)}{(1 - 4z)^{3/2}} - \frac{1}{z \sqrt{1 - 4z}}. \]  

Thus \( f'(z) \to +\infty \) as either \( z \to 0^+ \) or \( (1/4)^- \), and also

\[
f'(-\phi^{**}) = 2 \log \left( \frac{t + X + X_0}{\sqrt{t^2 - (X + X_0)^2}} \right) > 0.
\]

Thus \( f(z) \) always has a local maximum at \( z = -\phi^{**} \), which corresponds to the upper limit on the sum in equation (B 13). For \( t \) sufficiently small we have \( f'(z) > 0 \) for \( z \in (0, 1/4) \) in view of equation (B 14). However for \( t \) large \( f'(z) = 0 \) will have solution(s). Solving
$f''(z) = 0$ leads to $z = \bar{z} \equiv 1/[2(X + X_0) + 4]$ and thus the minimum value of $f'(z)$ occurs at $\bar{z}$ and the minimum value is

$$\min_{0 < z < 1/4} f'(z) = -t + \sqrt{(X + X_0)(X + X_0 + 2)} + 2 \log \left( \frac{\sqrt{X + X_0 + 2 + \sqrt{X + X_0}}}{\sqrt{2}} \right). \quad (B\ 16)$$

Now, $\bar{z}$ may or may not lie within the interval $(0, -\phi_*)$, but it will lie in this range for $t$ sufficiently large, as $-\phi_* \rightarrow 1/4$ as $t \rightarrow \infty$. By equating $\bar{z}$ to $-\phi_*$ we see that $\bar{z} \in (0, -\phi_*)$ precisely when

$$t > \sqrt{(X + X_0)(X + X_0 + 2)}. \quad (B\ 17)$$

If equation $(B\ 17)$ holds then certainly $t > X + X_0$, which is needed for the sum in equation $(3.5)$ to come into play. It follows that for $t > \sqrt{(X + X_0)(X + X_0 + 2)}$, $f'$ will have a unique minimum in the range $z \in (0, -\phi_*)$, and if $f'(\bar{z}) < 0$ then $f' = 0$ will have exactly two roots in this range. In view of equation $(B\ 16)$ let us define

$$t_+(X; X_0) = \sqrt{(X + X_0)(X + X_0 + 2)} + 2 \log \left( \frac{\sqrt{X + X_0 + 2 + \sqrt{X + X_0}}}{\sqrt{2}} \right). \quad (B\ 18)$$

Then for $t > t_+$ the function $f'$ has two zeros, call these $0 < z < z_* < |\phi_*|$, and $z_*$ is a local maximum of $f$ (since $f''(z_*) < 0$) while $z_*$ is a local minimum.

We next compute the contributions to equation $(B\ 13)$ from $z = -\phi_*$ and then from $z = z_*$ By the Laplace method, this requires expanding the summand in equation $(B\ 13)$ about a maximum of $f$. Setting

$$\tilde{N} = \tilde{N}(X, t) = \left[ \frac{t^2 - (X + X_0)^2}{4 \epsilon^2 t^2} \right] = \left[ -\frac{\phi_*}{\epsilon} \right],$$

the contribution from $z = -\phi_*$ is

$$\sqrt{\epsilon} g(\epsilon \tilde{N}) e^{f(\epsilon \tilde{N})/\epsilon} \sum_{m=0}^{\infty} e^{-m f'(\epsilon \tilde{N})} [1 + O(\epsilon)] \sim \sqrt{\epsilon} g(\epsilon \tilde{N}) \frac{e^{f(\epsilon \tilde{N})/\epsilon}}{1 - e^{-f'(\epsilon \tilde{N})}}, \quad (B\ 19)$$

where we set $N = \tilde{N} - m$ in the sum over $N$ in equation $(3.5)$. This yields equation $(3.8)$.

The contribution from $z = z_*$ is given by, using the Laplace method,

$$\sqrt{\epsilon} g(z_*) \sum_{N} e^{f(z_*)/\epsilon} \exp \left[ -\frac{1}{2\epsilon} f''(z_*)(z - z_*)^2 \right]$$

$$\sim \frac{1}{\sqrt{\epsilon}} g(z_*) \epsilon^{f(z_*)/\epsilon} \int_{-\infty}^{\infty} \exp \left( -\frac{|f''(z_*)| u^2}{2\epsilon} \right) du = \sqrt{2\pi} \frac{g(z_*)}{\sqrt{-f''(z_*)}} e^{f(z_*)/\epsilon},$$

and this yields equation $(3.10)$ for $(t > t_+)$. When $t - (X + X_0) = \epsilon \tilde{t} = O(\epsilon)$, $[-\phi_*/\epsilon] \sim M = O(1)$ and only a few terms in equation $(3.5)$ (up to $N = M$) are present, and the last term dominates the others, leading to equation $(3.11)$. This case is not important to the asymptotics of $p(x, t)$, or even $p_2(x, t)$. The right side of equation $(3.8)$ is the same as equation $(B\ 19)$, after we evaluate $f'$ and $g$ at $\epsilon \tilde{N} \sim -\phi_*$. In equation $(3.10)$ the last term is the contribution from the local maximum.
of \( f \) at \( z = z_* \), and we note that \( \sqrt{2\pi}g(z_*) \left[ -f''(z_*) \right]^{-1/2} \) is precisely the algebraic factor that multiplies \( e^{f(z_*)/\epsilon} \) in equation (2.9), in view of equation (B15) and the definition of \( g(z) \) below equation (B11). Also, \( f(z_*) \) in equation (2.10) is equivalent to that computed from the expression below equation (B11), when we use the fact that \( z_* \) satisfies equation (2.11). The leading term in equation (3.10) can come from either the endpoint term, if \( f(-\phi_{**}) > f(z_*) \), or from \( z_* \), if \( f(-\phi_{**}) < f(z_*) \). Along \( t = t_+ \) the endpoint contribution will be dominant, but for \( t \) sufficiently large \( f(-\phi_{**}) \rightarrow -\infty \) and \( f(z_*) \rightarrow -X - 1/2 \) so the \( z_* \) contribution will dominate. However, we are only interested in estimating \( p_2 \), and then the full density \( p \).

Appendix C Proof of Lemma 7

We shall scale \( \theta = \phi/\epsilon = O(\epsilon^{-1}) \) and \( x = \beta X = O(\beta) = O(1/\sqrt{\epsilon}) \), and we note that the function \( D_{-\phi/\epsilon}(Z/\sqrt{\epsilon}) \) has the integral representation

\[
D_{-\phi/\epsilon}(\frac{Z}{\sqrt{\epsilon}}) = \frac{\epsilon^{\phi/(2\epsilon)}}{i\sqrt{2\pi\epsilon}} \exp\left(\frac{Z^2}{4\epsilon}\right) \int_{B_r} \exp\left(\frac{1}{\epsilon} \tilde{F}(U; \phi, Z)\right) dU, \tag{C1}
\]

where

\[
\tilde{F} = -\phi \log U - ZU + \frac{1}{2} U^2, \tag{C2}
\]

\(|\Re(U)| > 0 \) on the vertical contour \( Br \), and \( \log U \) is defined to be real for \( U \) real and positive.

From equation (C2) we see that equation (C1) has saddle points where \( \tilde{F}'(U) = 0 \), or

\[
U_{\pm} = \frac{1}{2} \left[ Z \pm \sqrt{Z^2 + 4\phi} \right], \tag{C3}
\]

and also \( \tilde{F}_{UU} = 1 + \phi/U^2 \). Then using the steepest descent method we find that the major contribution comes from the saddle \( U_+ \) in equation (C3) and obtain, for \( \phi > 0 \),

\[
D_{-\phi/\epsilon}(\frac{Z}{\sqrt{\epsilon}}) \sim \epsilon^{\phi/(2\epsilon)} \left(\frac{2\phi}{Z^2 + 4\phi}\right)^{1/4} \epsilon^{-1/2} \exp\left\{ \frac{1}{\epsilon} \left[ \frac{\phi}{2} \frac{Z}{4} \sqrt{Z^2 + 4\phi - Z} \right. \right. \\
\left. \left. - \phi \log \left( \frac{Z + \sqrt{Z^2 + 4\phi}}{2} \right) \right] \right\}. \tag{C4}
\]

We shall also need the expansion of \( D_{-\phi/\epsilon}(\cdot) \) for \( \phi < 0 \), and in this range the function may have zeros. The expression in equation (C4) still holds if \( \phi < 0 \) and \( Z > 0 \), as long as \( Z^2 + 4\phi > 0 \). However, equation (C4) does not apply for \( \phi < 0 \) and \( Z < 0 \), which is what we need to approximate \( D_{-\theta}(-\beta) \) in equation (4.1) if \( \theta_{su} < 0 \). But, we can use the symmetry relation in equation (B1) for the case \( \phi < 0 \) and \( Z < 0 \) (setting \( -\beta = Z \)). Then equation (C4) may be used to estimate both terms in the right side of equation (B1) (to get \( D_{\phi/\epsilon}(iZ/\sqrt{\epsilon}) \) we just replace \( (\phi, Z) \) by \( (\epsilon - \phi, -iZ) \) and simplify the expression). As long as \( \Gamma^{-1}(\theta) \neq 0 \), or \( \phi \) remains bounded away from \( -\epsilon N; N = 0, 1, 2, \ldots \), the second term in equation (B1) is exponentially larger than the first and we thus obtain,
for $\phi = \varepsilon\theta < 0$, $Z < 0$ and $Z^2 + 4\phi > 0$,

$$D_{\phi/\varepsilon} \left( \frac{Z}{\sqrt{\varepsilon}} \right) \sim 2 \sin(\pi\theta) \varepsilon^{\phi/(2\varepsilon)} \frac{1}{(Z^2 + 4\phi)^{1/4}} \left( \frac{-2\phi}{\sqrt{Z^2 + 4\phi} - Z} \right)^{1/2} \times \exp \left\{ \frac{1}{\varepsilon} \left[ \frac{\phi}{2} - \frac{Z}{4\sqrt{Z^2 + 4\phi}} - \phi \log \left( \frac{-Z - \sqrt{Z^2 + 4\phi}}{2} \right) \right] \right\}. \quad (C\ 5)$$

This applies as long as $\theta \neq 0, -1, -2, -3, \ldots$, more precisely as long as we are away from exponentially small neighbourhoods of $\theta = -N$, which is where the zeros of $D_{\phi/\varepsilon}(Z/\sqrt{\varepsilon})$ are located for $\phi, Z < 0$. We can estimate these zeros precisely using equation (B1), similarly as we estimated the roots of equation (3.1), but the present set of zeros will not be required for the analysis. Setting $Z = -1 - X$ and then $Z = -1$ and taking the ratio of the two expressions leads to the result in equation (4.2).

We note that in equation (C4) the most rapidly varying factor in $\varepsilon$ is $\varepsilon^{\phi/(2\varepsilon)} = O[\exp(-1 \log \varepsilon)]$, but this disappears upon taking the ratio. In going from $\phi > 0$ to $\phi < 0$ in equation (4.2) we must only replace

$$\log \left[ \frac{\sqrt{1 + 4\phi} - 1}{\sqrt{(1 + X)^2 + 4\phi} - 1 - X} \right] \text{ by } \log \left[ \frac{1 - \sqrt{1 + 4\phi}}{1 + X - \sqrt{(1 + X)^2 + 4\phi}} \right].$$

Also, the factor $\sin(\pi\theta)$ in equation (C5) disappears upon taking the ratio of equation (C5) with the two different $Z$ values. But, equation (4.2) does become invalid when $\theta \approx -\varepsilon N$ for integer values of $N$. Lemma 7 does not apply for $\phi < 0$ and $X < -1$, but this range will not be needed in our analysis.