Survival Probability for the Stadium Billiard

Carl P. Dettmann ∗1 and Orestis Georgiou †1
∗1School of Mathematics, University of Bristol, United Kingdom

We consider the open stadium billiard, consisting of two semicircles joined by parallel straight sides with one hole situated somewhere on one of the sides. Due to the hyperbolic nature of the stadium billiard, the initial decay of trajectories, due to loss through the hole, appears exponential. However, some trajectories (bouncing ball orbits) persist and survive for long times and therefore form the main contribution to the survival probability function at long times. Using both numerical and analytical methods, we concur with previous studies that the long-time survival probability for a reasonably small hole drops like $\text{Constant} \times (\text{time})^{-\gamma}$; here we obtain an explicit expression for the Constant.

I. INTRODUCTION

Chaotic billiard theory was introduces by Yakov Sinai in 1970 [1]. Since then it has developed to become a well established theory for dynamical systems. A billiard is a dynamical system in which a particle alternates between motion in a straight line and specular reflections from the domain’s boundary. The sequence of reflections is described by the boundary map which completely characterizes the motion of the particle, hence billiards have their boundaries as a natural Poincaré section. Billiard systems are convenient models for many physical phenomena, for example where one or more particles move inside a container and collide with its walls. An excellent and comprehensive mathematical introduction to chaotic billiard theory can be found in the book of Chernov and Markarian [2].

In the early 80’s, mathematicians suggested investigating open systems, systems with holes or leakages, as a means of generating transient chaos [3], retrieving information from distributions [4], and deducing facts about the equivalent closed systems. The key distributions of interest classically are the escape probability density $p_e(t)$, which is given by the trajectories that leave the billiard at time $t$, where $t \geq 0$, $t \in \mathbb{R}$, and also the survival probability of orbits $P(t)$ up to time $t$, given some initial probability measure, typically the equilibrium measure defined in section II below.

These two are related by $P(t) = \int_0^\infty p_e(t')dt'$. Such investigations have naturally been extended to billiard systems as well. Links between billiards and geometrical acoustics [5–9], quantum chaos [10, 11], controlling chaos [12–14], atom optics [15], hydrodynamical flows [16–21], astronomy [22, 23] and cosmology [24], have been established in the context of open dynamical billiards. Furthermore, it has become apparent over the past few years, that the subject of open billiards and their distributions provide a pathway towards understanding chaos and may even open doors to old, but not so forgotten problems such as the Riemann hypothesis [25, 26].

The stadium billiard (see Figure 1 below) is a seemingly simple dynamical system, introduced by Leonid Bunimovich in 1974 [27]. The billiard’s boundary consists of two parallel straight lines and two semicircular arcs. It was later proven by him to be ergodic, to be mixing, to have the Kolmogorov property [28], and in 1996 by Chernov and Haskell to have the Bernoulli property [29]. It has been described as a system with “fully developed chaos” [30]. Its entropy has been numerically estimated in [31] and theoretically in [32]. The stadium billiard is a special case of a chaotic billiard. Being constructed from two fully integrable billiard segments, the circle and the rectangle, it is remarkable that the system remains completely chaotic no matter how short its parallel segments are. It is a limiting case of the larger set of Hamiltonian systems, which Bunimovich refers to as “mushrooms” [33] with cleanly divided phase-space areas, regular and chaotic. The stadium is the fully chaotic limit of the natural mushroom billiard, while the circle is the fully regular limit. If the parallel segments are of length $2a$ say, where $a > 0$, and $a \in \mathbb{R}$, then the Lyapunov exponent $\lambda(a) \to 0$ in both limiting cases of $a \to 0$ and $a \to \infty$. Also, it is well known that the defocusing mechanism, which is one of the two sources of chaos in billiards [34] (the other being the dispersing mechanism), characteristic of all Bunimovich type billiards, requires $a > 0$ in order for any wave-front to defocus and therefore exhibit hyperbolicity.

Wojtkowski in 1986 [35] clarified much of the mechanism behind this hyperbolic behavior. The very existence of the parallel segments of the boundary is also the source of the intermittent behavior found in the stadium billiard. They allow for the existence of a set of marginally stable periodic orbits of zero measure but indeed of great importance. They are the main reason why the stadium is not uniformly hyperbolic. Also, though it is classically and quantum mechanically ergodic, it does not have the property of unique ergodicity [36, 37]. This means that not all eigenfunctions are uniformly distributed and therefore this causes scarring [38]. This is due to the
existence of the so called “bouncing ball” orbits, sometimes called “sticky orbits”. Semi-classically, they have caused much trouble in the treatment of the system as explained in great detail by Tanner [39] since they affect the stability of periodic orbits close to them but do not contribute to individual eigenvalues in the spectrum of the stadium.

Lai-Sang Young’s infinite Markov extension construction called a Young tower in 1998 [41, 42] triggered a series of rigorous mathematical proofs concerning the long time statistical properties of the stadium billiard. In 2004 Markarian [40] proved that asymptotically the billiard map in the stadium has polynomial decay of correlations of order \((\log n)^2 n^{-1}\) (here \(n\) is the number of iterations of the billiard map). This method was then simplified and generalized by Chernov and Zhang in 2005 [43] to include for example the drive-belt stadium where the straight segments are no longer parallel. Bálint and Gouëzel [44] in 2006, used this method to prove that the Birkhoff sums of a sufficiently smooth generic observable with zero mean in the stadium, satisfy a non-standard limit theorem where its convergence to a Gaussian distribution requires a \(\sqrt{n \log n}\) normalization. In 2008, Chernov and Zhang sharpen their previous estimate by removing the \(\log n\) factor [45] and Bálint and Melbourne show that these relations hold for observables smooth in the flow direction as well (this excludes position and velocity)[46]. These results for the rate of decay of correlations can, at least heuristically, be transferred into the context of the open stadium to address problems such as escape rates and survival probabilities.

Therefore, even though the stadium billiard has been observed to exhibits strong chaotic properties for short times such as approximate exponential decays of the escape times distribution and decays of correlations of initial conditions both numerically and experimentally, it has also been shown to experience a cross-over at longer times, towards an asymptotic power-law behavior [47–49]. Hence, the stadium billiard is an example of a transient chaotic system which exhibits intermittency. Intermittency (described in more detail in section II), in open systems is a relatively new subject and is increasingly being discussed and researched in the context of non-linear Hamiltonian systems. These investigations tend to focus on the asymptotic behavior of distributions and often use the stadium as one of their main examples [50, 51]. In fact, it has recently been suspected that the very long regular flights present in the expanded stadium are the reason why numerically, at least, the moments of displacement diverge from the Gaussian [52]. It is arguably an ideal model for studying the influence of almost regular dynamics near marginally stable boundaries both theoretically [39] and numerically [53]. Armstead Hunt and Ott [54] have carried out a detailed investigation into the asymptotic stadium dynamics and have shown that \(P(t) \sim \text{Const}/t\) for long times but do not calculate the Constant.

In this paper we study the escape from the stadium billiard; we explicitly calculate the measure of the set of orbits causing the asymptotic power-law decay, and obtain an analytic expression for the longer times survival probability function. In contrast with Ref.[25], we do not assume or require that the hole is vanishingly small and in contrast to Ref.[54], we do not use a probabilistic description of the dynamics. The paper is organized as follows. In section II we set the stage and state the reasoning as well as the main ideas of this paper and we also set up our problem and define all the variables and sets required. In section III and IV we consider the two main sets of initial conditions which contribute to the survival probability at long times while in section V we introduce and explain the details of the approximation method used. Finally, in section VI we present our numerical results from computer simulations and compare with the analytical ones. Conclusions and discussions appear at the end, in section VII, where future work is also discussed.

II. THE MAIN IDEAS AND SET-UP

Consider an open stadium as shown in Figure 1. A classical non-interacting particle of unit mass and unit speed experiences elastic collisions on \(\Gamma\), the boundary of the billiard table. The length of the parallel sides is \(2a\), while the radius of the circular segment is \(r\). A hole of size \(\epsilon\) is punched onto one of the straight segments of \(\Gamma\) with \(x\) coordinates \(x \in (h_1, h_2), h_2 = h_1 + \epsilon\). \(x\) is the position coordinate which we only need to take values along the straight segments, \(x \in [-a, a]\). We define the inward pointing normal vector \(\hat{n}\) which defines the angle \(\theta\) made by the reflected particle and \(\hat{n}\). The angle \(\theta \in (-\pi/2, \pi/2)\) and is positive in the clockwise sense from \(\hat{n}\). The billiard flow conserves the phase volume and the corresponding invariant equilibrium measure along the straight segments is \(d\mu = C^{-1} \cos \theta d\theta dx\), where \(C = \int_{-\pi/2}^{\pi/2} \cos \theta d\theta \int_{\epsilon}^{2a} dr = 2|\Gamma| = 2(4a + 2\pi r)\) is the canonical probability measure preserved by the billiard map on the billiard boundary. \(d\mu\) is also the distribution of initial conditions. For the purpose of this paper, no parametrization of the position coordinate is needed along the circular segments of the billiard. Finally, if the particle hits the hole, it will escape; as previously noted we are interested in the long time behavior of the survival probability.

As discussed in the Introduction, the stadium is hyperbolic as a result of the defocusing mechanism. However it is not uniformly hyperbolic, as there exist a small set of parabolic, non-isolated periodic orbits [55] called bouncing ball orbits which is of zero measure. It has been observed that orbits in the chaotic region of the phase space which are close to these marginally stable periodic orbits show almost regular behavior for finite times. As a result, the decay through the hole appears exponential for short times, followed by an algebraic tail [56]. A theory for explaining the intermediate time transition from exponential to power-law decay is that of intermittency and was first introduced.
in 1979 by Manneville and Pomeau in their study of the Lorenz system [57, 58]. Usually, intermittency signifies a small, finite time Lyapunov exponent for unstable periodic orbits approaching regular regions in phase space [59]. Trajectories almost tangent to the circular arcs (“whispering gallery orbits, or rolling orbits”) are intermittent with respect to the collision map, but not with respect to the flow. This is because they are of bounded total path length and sufficiently unstable [39, 46]. Near bouncing ball orbits on the contrary are not of bounded total path length and are generally believed to be the main cause of the intermittency of the flow, therefore exclusively determining the asymptotic tail [49, 54] of the survival probability $P(t)$ and therefore play a significant role in the stadium’s dynamics. These orbits are characterized by small angles $\theta$ (near vertical) that remain small for relatively long periods of time. Chernov and Markarian describe them in their book [2] as orbits with a large number of ‘nonessential collisions’. Semi-classically, it has been suggested that an ‘island of stability’ surrounds this marginally stable family. Its boundary depends explicitly on $\hbar$ and the measure of this island shrinks to zero (compared with the total volume) in the semiclassical limit $\hbar \to 0$ [39].

Having noted, following [49, 54], that the set of orbits surviving for long times is contained in the near bouncing ball orbits, with small angles $\theta$ and position on the straight segments, we now categorize these orbits into two simple families: orbits initially moving towards the hole, and orbits initially moving away from the hole. We would like to identify the set of orbits from these two families which do not escape until a given time $t$. An important result by Lee, that dates back to 1988 [60], states that the angle of a near bouncing ball orbit in the stadium remains small even after a reflection with a semicircular segment. In fact, as will be explained in section IV, small angles can change by at most a factor of 3 after being reflected off the curved billiard boundary. Therefore, given a sufficiently large time constraint $t$, surviving orbits are restricted to angles insufficient to ‘jump over’ the hole, even after a reflection on the circular part of the billiard boundary. In this way we identify the surviving orbits as members of time dependent, monotonically shrinking subsets of the two families of orbits defined above. The measure of these subsets tends to zero as $t \to \infty$. These two subsets are considered in detail in sections III and IV below and are used to calculate, to leading order, the stadium’s survival probability function for long times $P(t)$ (see equation (32) below).

III. CASE I: MOVING TOWARDS THE HOLE

We start by considering trajectories initially on the right of the hole with $x \in (h_2, a)$ moving towards it. These trajectories will prove to be only a part of the survival probability function for long times. However, they are essential in order to construct a complete and accurate expression for the asymptotic limit of the full survival probability function. To ensure that such trajectories will escape when they reach the hole’s vicinity, they must satisfy the following condition:

$$|\theta| < \arctan \left( \frac{\epsilon}{4r} \right),$$  \hspace{1cm} (1)
hence they will definitely not jump over the hole. The set of initial conditions \((x, \theta)\) for trajectories which will escape in exactly time \(t\) satisfies:

\[
t \sin |\theta| - \delta 4r \tan |\theta| \leq x - h_2 \leq a - h_2,
\]

where \(0 < \delta < 1\). For long times \(\delta 4r \tan |\theta|\), which is the horizontal distance from the edge of the hole to where the particle exits the billiard, will shrink to zero \((\sim 1/t)\) as the set of surviving trajectories is limited to near vertical angles. Hence we drop this nonsignificant term in what follows. Notice that we will be using physical time \(t\) for our calculations but the equations are set up as if considering a map between \(p \in \mathbb{N}\) collisions, with \(t = \frac{2\pi p}{\cos \theta}\). This way, we do not need to define equations for the billiard map, which in any case would just be described by the usual reflection map. From equations (1) and (2) we can deduce that the angles must satisfy:

\[
|\theta| < \min\left\{ \arctan \left( \frac{\epsilon}{4r} \right), \arcsin \left( \frac{a - h_2}{t} \right) + \mathcal{O}(1/t^2) \right\}.
\]

The second term in (3) is the dominant one for long times. This leads to the following integral for the conserved measure of the billiard map:

\[
I_r = \frac{2}{C} \int_0^{\arcsin \left( \frac{a - h_2}{t} \right) + \mathcal{O}(1/t^2)} \left( \int_{h_2 + t \sin \theta + \mathcal{O}(1/t)}^a \cos \theta \, dx \right) \, d\theta,
\]

where the subscript \(r\) stands for right and \(C\) was defined in section II. We are integrating over the set of initial conditions, on the right of the hole, which will not escape until time \(t\). Hence we are considering escape times greater than or equal to \(t\). We have also dropped the modulus sign from \(\theta\) and multiplied the whole expression by a factor of 2, due to the vertical symmetry of the problem. This simplifies to

\[
I_r = \frac{(a - h_2)^2}{Ct} + \mathcal{O}(1/t^2).
\]

This result is valid for trajectories satisfying:

\[
\arcsin \left( \frac{a - h_2}{t} \right) + \mathcal{O}(1/t^2) < \arctan \left( \frac{\epsilon}{4r} \right),
\]

that is

\[
t \gtrsim \frac{8ar}{\epsilon},
\]

since the supremum of \(a - h_2\) is \(2a\). We continue with this calculation by adding the analogous contribution \(I_l\) from the small angle trajectories starting on the left of the hole with \(x \in (-a, h_1)\) moving towards it. This operation can easily be calculated from equation (5) by simply sending \(h_1 \mapsto -h_2\) and \(h_2 \mapsto -h_1\).

\[
I_l = \frac{(a + h_1)^2}{Ct} + \mathcal{O}(1/t^2).
\]

Adding the two integrals gives the measure of all initial conditions moving towards the hole that survive until time \(t\):

\[
I_{r+l} = \frac{(a + h_1)^2}{Ct} + \frac{(a - h_2)^2}{Ct} + \mathcal{O}(1/t^2).
\]

Hence, part of the canonical Survival Probability function due to the nonessential orbits initially approaching the hole from either side, for long times satisfying condition (6) is:

\[
P_1(t) = \frac{(a + h_1)^2 + (a - h_2)^2}{2(4a + 2\pi r)t} + \mathcal{O}(1/t^2).
\]

This expression is essentially a sum of contributions from two families of bouncing ball orbits, each proportional to the square of the available length.
IV. CASE II: MOVING AWAY FROM THE HOLE

Numerical simulations confirm that \( P_1(t) \) in equation (8) is indeed not the full expression for the long time survival probability function of the open stadium billiard. The set of orbits accounted for in the previous section is only a fraction of all the marginally stable periodic orbits discussed in the introduction of this paper. In this section we will be considering orbits initially moving away from the hole, so that they experience a reflection process when they collide with the right semicircular end. We only consider the right semicircular end, as we shall later use the symmetry of the stadium to see what happens at the other one. If the initial angles are small, then the final angles (after being reflected at the wings of the stadium) will remain small and therefore survive for long times and account for the remaining set of orbits and built up the long time survival probability. In this section we will investigate and identify exactly the initial conditions which survive for long times \( t \). As in the previous section, this investigation is based on the assumption that equation (1) provides an upper bound on the magnitude of angles considered, therefore ensuring that the orbits considered are indeed bouncing ball orbits that survive for long times.

Throughout this and the following section we will be using \((x_i, \theta_i)\), where \( x_i = x_1 - 2rn\theta_i \), as the coordinates of our initial conditions which lie on the right of the hole, \( x_i \in (h_2, a) \), and move away from it. Due to the stadium’s symmetry, we only need consider the case \( \theta_i > 0 \). We will use \((x_1, \theta_1)\) to indicate the position and angle of a trajectory right after its final collision on a flat segment, while still moving away from the hole. Therefore, the next collision of such trajectories will be on the right semi-circular segment of the billiard. This helps to distinguish between the initial conditions and their transformed final values. Notice that \( \theta_i = |\theta_1| \).

\[ T(x_i, \theta_i) = \frac{2rn}{\cos \theta_i} + \frac{2rm}{\cos \theta_f} + D_f, \]  

where \( n \) and \( m \) are the respective numbers of non essential collisions before and after the reflection process on the right semicircular end, and are defined as:

\[ n = \left\lfloor \frac{a - x_i}{2r \tan |\theta_i|} - \delta_i \right\rfloor, \quad m = \left\lfloor \frac{a - h_2}{2r \tan |\theta_f|} - \delta_f \right\rfloor, \]

with \( 0 < \delta_i, \delta_f < 1 \), and \( f = 3, 4 \). We chose these indices for \( f \) (3 and 4) to indicate the number of collisions comprising the reflection process as the angles change two or three times respectively. \( D_f \) is the time taken for the reflection process at the semicircular end, and is bounded by:

\[ 4r < D_3 < 2r \left( \frac{1}{\cos \theta_i} + \frac{1}{\cos \theta_3} \right), \]

\[ 6r < D_4 < 2r \left( 1 + \frac{1}{\cos \theta_i} + \frac{1}{\cos \theta_4} \right). \]
The number of reflections on the curved boundary alternates between two scenarios, as shown in Figure 2, the case with one collision on the semicircle and the case with two collisions depending on the initial conditions of the trajectory $x_i$ and $\theta_i$. Specifically, $\theta_f$ can be found and defined by the use of small angle approximations as:

$$\theta_3 = \frac{2d_1}{r} - 3\theta_i < 0,$$  \hfill (10)  

$$\theta_4 = \frac{4d_1}{r} - 5\theta_i > 0$$  \hfill (11)

(see Ref.[54, 60]), where $d_1 = a - x_1 = a - x_1 - 2r\cos\theta_i > 0$, is the horizontal distance between $x = x_1$ and $x = a$ as indicated in Figure 2. Taking only the leading order terms of $\theta_1$ and $\theta_f$ from our expressions is justified by the fact that in long time scenarios, unstable periodic orbits have very small angles (before and after being reflected), tracing out near vertical trajectories. Using the above information, equation (9) can be expressed in the following way:

$$T(x_i, \theta_i) = \frac{a - x_i}{|\theta_i|} + \frac{a - h_2}{|\theta_f|} + \Delta_f,$$  \hfill (12)

where $\Delta_f = D_f - 2r(\frac{\delta}{\cos\theta_i} + \frac{\delta_t}{\cos\theta_f})$.

Before we continue, it is essential to find the boundaries of validity for the functions of $\theta_f$. These will define the geometry of the two scenarios. We ask the question: When do we see one and when two collisions at the semicircular ends? This is answered by considering a number of inequalities. The first and most obvious one is $\theta_1 \geq \frac{4a}{\pi r}$ which requires the next collision to be on the circular segment. We also note that $\theta_1 = \frac{3d_1}{r}$ is the transition line between $\theta_3$ and $\theta_4$ and is the case where the reflected particle will hit exactly the point $x = a$, where the circular segment meets the straight segment. Finally, if we are also to satisfy condition (1) we must form two more inequalities: equation (10) gives

$$\theta_i < \frac{2d_1}{3r} + \frac{\arctan\left(\frac{z}{r}\right)}{3},$$  \hfill (13)

and equation (11) gives

$$\theta_i < \frac{4d_1}{5r} - \frac{\arctan\left(\frac{z}{r}\right)}{5}. $$  \hfill (14)

These inequalities enclose a small area in the $x_i, \theta_1$ plane, the plane of initial conditions. Notice that $d_1$ is a function of $x_i$ but also depends on $n$, the number of collisions before the reflection process, which if not equal to zero introduces an extra $\theta_1$ term. This means that inequalities (13) and (14) have to be solved for $\theta_1$ for every $n = 0, 1, 2, 3 \ldots$ This is done and shown in Figure 3, up to and including $n = 2$, where $z$ is taken to be equal to $\arctan(\epsilon/4r)$. These boundaries of validity define the set of orbits which escape after being reflected at the right semicircular segment of the billiard. It is not immediately clear from Figure 3, but the peaks of these spikes are of the same height $\theta_1 = 3\zeta$. However, we note that the set of orbits that survive up to time $T$, where $T$ is large, is not identical with the former set.

To motivate what is to follow we have a look momentarily to Figure 4 below, which on its left panel shows a numerical simulation which identifies the set of initial conditions $(x_i, \theta_i)$ which survive until time $t = 50$. Notice that the peaks of the spikes grow in height as we move from right to left, moving away from the end of the flat segment at $x = a$, and therefore in a sense increasing the count of pre-reflection collisions $n$.

To find the survival probability function, $P_2(t)$, of these trajectories we must now consider $T(x_i, \theta_i)$ as the parameter $t$. Hence we can rearrange equation (12) according to (10) and (11), to form two expressions which depend not only on $x_i$, $\theta_i$ and $n$, but also on the time $t$:

$$f_3(x_i, \theta_i, t) \equiv (a - x_i)\left(\frac{2d_1}{r} - 3\theta_i\right) - (a - h_2)\theta_i + (\Delta_3 - t)\left(\frac{2d_1}{r} - 3\theta_i\right)\theta_i = 0$$ \hfill (15)  

$$f_4(x_i, \theta_i, t) \equiv (a - x_i)\left(\frac{4d_1}{r} - 5\theta_i\right) + (a - h_2)\theta_i + (\Delta_4 - t)\left(\frac{4d_1}{r} - 5\theta_i\right)\theta_i = 0$$ \hfill (16)

The above expressions are conic sections as they are quadratic in both $x_i$ and $\theta_i$ and describe hyperbolas in the plane of initial conditions. This is not immediately obvious because of the factors $x_i$ and $\theta_i$ which are hidden in the
The set of initial conditions initially on the right of the hole, which collide and reflect on the right semicircular segment of the stadium and do not cross over the hole are defined by the boundaries of validity. These are shown here for \( n = 0, 1, 2 \). The dotted diagonal lines are given by \( \theta_i = \frac{(a - x_i)}{2r(n + 1)} \) and separate the plane into the relative areas of \( n \). The top triangle of each spike (dark blue) is for \( f = 3 \) (one collision on semicircle) while the bottom (light blue) is for \( f = 4 \) (two collisions on semicircle). They are separated by the straight lines given by \( \theta_i = \frac{3(a - x_i)}{2r(3n + 2)} \). The remaining two sets of lines which define the spikes are given by the solutions of equations (13) and (14). Notice that all spikes have the same maximum height of \( 3z = 3 \arctan(c/4r) \).

The two hyperbolas, (15) and (16), approach each other as an effect of increasing the time \( t \) and tilt and shift discontinuously when increasing \( n \). We notice that if we impose these hyperbolas onto the boundaries of validity we found earlier (see Figure 3), we are essentially imposing a time constraint on the set of initial conditions which will survive up to time \( t \). Their effect will be for smaller \( n \) to erode the area enclosed by the inequalities, therefore sharpening them, and for larger \( n \) to thicken them from either side, effectively shaping them into a series of spikes, allowing for larger and larger values of \( \theta_i \) as we increase \( n \) from left to right.

This effect can be seen on the right of Figure 4, where the boundaries of validity for \( n = 0, 1, 2 \) and 3 are eroded by the hyperbolas which are time dependent, causing each spike to grow taller as we move away from the edge of the straight segment of the billiard. It can be easily seen that the pictures in Figure 4 are almost identical to a very small error, confirming that we are measuring the correct set of initial conditions and therefore the orbits they describe. The thickening effect is a consequence of the survival probability function’s set up. Trajectories which fall just outside of the area enclosed by the boundaries of validity but for large enough \( n \), are trapped between the two hyperbolas, will not eventually escape through the hole, i.e. they will jump over it, but they will still survive until the given time \( t \).
We notice that for finite \( t, n \) is also finite. The key of the relation between \( t \) and \( n \) lies in the conic section equations (15)-(16). Solving them for \( n \) we discover that for \( n = \frac{-3r+2t-2\Delta_3}{4r} \) and \( n = \frac{-5r+4t-4\Delta_4}{8r} \), respectively, the conic sections are no longer hyperbolas but turn into negative parabolas. Hence, we can define the maximum number of pre-reflection collisions for finite time as:

\[
N_{\text{max}}(t) = \left\lfloor \frac{-3r+2t-2\Delta_3}{4r}, \frac{-5r+4t-4\Delta_4}{8r} \right\rfloor
\]

for large \( t \), where the lower square brackets \( \lfloor h \rfloor \) are defined as the integer part of \( h \) (also known as the floor function). Actually, as we shall find out in the next section, this term \((N_{\text{max}})\) is never reached in practice (see \( N_3 \) in equation (22) below).

It seems that we have the area of interest well defined and bounded. We thus need to integrate over the area of each spike and then sum them all up to \( N_{\text{max}} \) for any given \( t \). Multiplying the result by a factor of 2 (vertical symmetry), would eventually give the measure of the orbits initially on the right of the hole moving away from it.

Integrating hyperbolas and then summing their enclosed areas is a lengthy and unpleasant process. This calculation has been done numerically and an accurate result has been obtained successfully and is presented in section VII. However, this calculation can only be carried out numerically, as an analytical result is in our opinion impossible to obtain. Therefore we shall present a simpler approximation method for \( P_2(t) \), which is analytically tractable, but still accurate to leading order in \( 1/t \).

V. APPROXIMATING HYPERBOLAS

In this section we will take equations (15) and (16) and argue that for large enough times \( t \), the sections of the hyperbolas that are of interest can be simply and accurately described by straight lines. We can visually confirm this from Figure 4; however, further investigations have shown that the distance between the foci of each hyperbola converges to zero faster than the lengths of the integration limits (see Figure 4; however, further investigations have shown that the distance between the foci of each hyperbola converges to zero faster than the lengths of the integration limits (on \( \theta_i = (a-x_i)/2rn \) and \( \theta_i = (a-x_i)/2rn(n+1) \) as \( t \to \infty \). In fact, for any \( n \) we find that the corresponding rates are \( \sim t^{-1.5} \) and \( \sim t^{-1} \). This in turn shows that the error made by approximating hyperbolas by straight lines, after integrating and summing over \( n \), is still negligible with respect to the leading term \((t^{-2} \text{ rather than } t^{-1})\). This approximation method will later be verified by further numerical simulations where we calculate the error between the approximate solution and the numerical integration result.

In this section we will use \( \tau = t - \Delta \) without any subscript to avoid unnecessary confusion given that for long times, \( \Delta \) will completely vanish from our results. We consider the same initial conditions as in the previous section. For the first step of the approximation method, we must find the coordinates where equations (15) and (16) meet with \( \theta_i = \frac{3(a-x_i)}{2(1+n)r} \). We also need to find the coordinates where equation (16) meets with \( \theta_i = \frac{a-x_i}{2rn} \) and finally equation (15) with \( \theta_i = \frac{a-x_i}{2rn} \). These three points along with \((a,0)\) define the four corners of a quadrilateral in phase space shown in Figure 5. Here are their coordinates:

\[
\begin{align*}
A & = (a + \frac{6(a-h_2)(2+3n)r}{(4+6n)r - 3r - 3(a-h_2) - 3(a-h_2)(4+6n)r^2}, \\
B & = (a + \frac{2(a-h_2)(n+1)r}{3(2nr - \tau)}, \\
C & = (a + \frac{2(a-h_2)(1+n)r}{3(2(1+n)r - \tau)}, \\
D & = (a - \frac{(a-h_2)^2r}{2n(r - \tau)(2(1+n)r - \tau)}).
\end{align*}
\]

The next step is to form four equations, one for each side of the quadrilateral by using these coordinates for \( n \ge 0 \). Afterwards, elementary integration methods in \( x_i \) are used to produce explicit functions for the area of each spike with \( \tau \) and \( n \) as the only parameters:

\[
\text{Area}_1 = \frac{(a-h_2)^2r}{2nr - \tau) (2(1+n)r - \tau)}.
\]

Before we insert equation (18) into a sum, we must figure out the upper limit of \( n \) for which this expression is valid. We do this by finding the smallest integer value of \( n \) for which the \( x_i \) coordinate of point \( A \) is smaller than \( h_2 \) and call it \( N_1 \). This is because, for increasing values of \( n \), all the corners \((A, B, C)\) shift to the left causing the spike to tilt and stretch but when \( n \ge N_1 \) point \( A \) is no longer a valid coordinate. A closer look at the situation reveals that point \( A \) will split in to two points, \( A_1 \) and \( A_2 \) say, both situated on the line \( x_i = h_2 \). For \( n > N_1 \), the quadrilateral...
FIG. 5: (Color online) The corners which make up the polygons which approximate the hyperbolas, for any \(n\), are defined by the coordinates of points \(A, B, C\) and the corresponding coordinate of the endpoint of the straight segment (in this case \((a,0)\)). The \(x_i\) and \(\theta_i\) coordinates of these corners are clearly marked by the dotted lines. The color coding is similar to that of Figure 3. The green/dashed vertical line indicates the hypothetic position of the closest edge of the hole (in this case \(x_i = h_2\)) which acts and changes the shape of the spike by defining two new corners \(A_1\) and \(A_2\) instead of \(A\). This process is replaced by a pentagon as the spike’s peak (point \(A\)) overshoot the vicinity of the hole’s location. This process is best described diagrammatically in Figure 5 where the four corners \(A, B, C\) and \((a,0)\) are replaced by \(A_1, A_2, B, C\) and \((a,0)\) which are then fed back to the integration method to produce a new expression describing the area of the truncated spike. The same will happen to all corners, and different combinations of them will be necessary to produce the corresponding area functions. In other words, as the \(x_i\) coordinates of corners \(A, C\) and \(B\), in this order, overshoots the hole’s location at \(x_i = h_2\), as we increase \(n\), a new area function \((\text{Area}_j, j = 1 - 4)\) via integration, with a new expiration number (summation limit) via equation solving in \(n\), will be required. This process produces three more area functions, and therefore four summations, each with different limits:

\[
\text{Area}_2 = 
\frac{(a - h_2)^2 \left[ 1152 n^4 r^4 + \left( -224 r^3 + 313 r^2 \tau - 114 r \tau^2 + 9 \tau^3 \right) \tau + 192 n^3 \tau^3 (14 r - 9 \tau) \right] + 4 n^2 r^2 \left( 512 r^2 + 225 \tau^2 - 762 r \tau \right) + 4 n r \left( 128 r^3 + 270 r^2 \tau^2 - 45 \tau^3 - 100 r^2 \tau \right) \right]}{\tau \left( 16 (1 + n) (2 + 3 n) r - 3 (5 + 8 n) \tau \right) (2 n r - \tau) \left( 2 (1 + n) \tau - r \right) \left[ -9 \tau + 4 n \left( 4 + 6 n \right) r - 3 \tau \right]},
\]  

(19)

\[
\text{Area}_3 = \frac{(a - h_2)^2 \left( 32 n^2 r^2 + 16 n^2 \tau^2 - (14 n + 3 n^2) \tau \right)}{4 r \left( 1 + n \right) \left( 2 n r - \tau \right) \left[ -9 \tau + 4 n \left( 4 + 6 n \right) r - 3 \tau \right]},
\]  

(20)

\[
\text{Area}_4 = \frac{(a - h_2)^2}{4 (n + n^2) r}.
\]  

(21)

Having all the puzzle pieces at hand, we form an expression for the invariant measure of all the initial conditions moving away from the hole from the right. The sum over the areas of quadrilaterals \((\text{Area}_1)\) added to the sum of
pentagons ($\text{Area}_2$) and another sum of quadrilaterals ($\text{Area}_3$) and finally the infinite sum of triangles, $\text{Area}_4$, gives:

$$\text{Area}_{\text{Right}} = \sum_{n=0}^{N_1} \text{Area}_1 + \sum_{n=N_1}^{N_2} \text{Area}_2 + \sum_{n=N_2}^{N_3} \text{Area}_3 + \sum_{n=N_3}^{\infty} \text{Area}_4,$$

where

$$N_1 = \left[ \frac{3t - 3\Delta - 16r}{24r} \right], \quad N_2 = \left[ \frac{3t - 3\Delta - 8r}{8r} \right], \quad \text{and} \quad N_3 = \left[ \frac{3t - 3\Delta}{8r} \right].$$

All these sums, except the second one, where simplified as follows by Mathematica v.6, by allowing the summation limits to acquire their non-integer values. This is allowed since the upper limit $N_l \sim t$ ($l = 1, 2, 3$), and therefore losing or gaining a term from the end of each summation effectively makes no difference whatsoever for long times.

$$\sum_{n=0}^{N_1} \text{Area}_1 = \frac{(a - h_2)^2(8r + 3\tau)}{2\tau(9r - 8r)},$$

$$\sum_{n=N_1}^{N_2} \text{Area}_3 = \frac{32(a - h_2)^2r(192r^2 + 56\tau^2 + 3\tau)}{(8r + 3\tau)(-8r - \tau)(64r^2 - 9r^2 - 72r\tau)},$$

$$\sum_{n=N_2}^{\infty} \text{Area}_4 = \frac{(a - h_2)^2}{4rN_3}.$$

The simplification of the second sum (that of $\text{Area}_2$) requires a more lengthy and tricky process, as it can not be simplified explicitly by any conventional means. This is so, not only because $\text{Area}_2$ has the most complicated of the four expressions, but also because its sum covers the largest range over $n$ ($N_2 - N_1 \sim t/4$). Therefore, for large $t$, $n$ is never small. By using a substitution of the form $u = 1/t$, assuming $u$ to be small for large $t$ and then substituting $n = s/u$, where $s$ is of $O(1)$, before expanding $\text{Area}_2$ into a power series effectively incorporates the effect of large $n$ into the leading order term of the series. We get:

$$\text{Area}_2 = \sum_{k=0}^{\infty} \alpha_k u^k = -\frac{(a - h_2)^2(1 - 16sr + 32s^2r^2)}{32(s^2r(-1 + 2sr)^2)} + O(u^3),$$

We reverse the substitution, and simplify the sum to obtain:

$$\sum_{n=N_1}^{N_2} \text{Area}_2 = (a - h_2)^2 \left( \frac{12ru(\Psi^{(0)}(z_1) - \Psi^{(0)}(z_2) + \Psi^{(0)}(z_3) - \Psi^{(0)}(z_4))}{32r} + \frac{-\Psi^{(1)}(z_1) - \Psi^{(1)}(z_2) + \Psi^{(1)}(z_3) + \Psi^{(1)}(z_4)}{32r} \right),$$

where

$$z_1 = \frac{8r - 9/u - 3\Delta}{24r}, \quad z_2 = \frac{8r + 3/u - 3\Delta}{24r}, \quad z_3 = \frac{3(1/u - \Delta)}{8r}, \quad z_4 = -\frac{1/u + 3\Delta}{8r},$$

and $\Psi^{(k)}$’s are polygamma functions. The polygamma function of order $k$ is defined as the $(k + 1)$th derivative of the logarithm of the gamma function:
\[ \Psi^{(k)}(z) = \frac{d^{(k+1)}}{dz^{(k+1)}} \ln \Gamma(z). \]

Fortunately the polygamma functions are of the form \( z = \frac{a}{bu} + c \), where \( a, b \) and \( c \) are constants, and can be expanded as a Taylor series to leading order as follows:

\[
\Psi^{(0)}(\frac{a}{bu} + c) = \ln(a/b) - \ln u + O(u),
\]

\[
\Psi^{(k\geq2)}(\frac{a}{bu} + c) = (-1)^{k-1}(k-1)!\left(\frac{bu}{a}\right)^k + O(u^{k+1}).
\]

Substituting these expressions into equation (27) will simplify the expression dramatically, finally leaving us with the survival probability function:

\[ P(t) = \frac{1}{4(4a + 2\pi r)t} \left( \frac{3ln3 + 4}{4a + 2\pi r} \right) + O(1/t^2). \]

VI. RESULT AND NUMERICAL SIMULATION

It remains to add the probability measure of the two types of trajectories to obtain the asymptotic limit of the survival probability function:

\[ P_s(t) = P_1(t) + P_2(t), \]

where the subscript \( s \) stands for the straight lines we approximated the hyperbolas with. This gives:

\[ P_s(t) = \frac{4(4a + 2\pi r)\Psi^{(0)}(\frac{a + h_1}{a - h_2}) + O(1/t^2)}{4(4a + 2\pi r)t} \]

which is valid only for trajectories satisfying (6), i.e. sufficiently large \( t \).

In the left panel of Figure 6 below, we compare \( P_s \) (equation (32)) with \( P_d \) which is obtained by a direct numerical simulation using Mathematica v6., consisting of 1.5 million initial conditions distributed according to the invariant measure of the billiard map. We see that \( P_s \) gives a good prediction of the numerical survival probability for long times \( P_d \). We have tested this result with other values of the parameters: \( a, r, h_1, h_2 \) as well. What is even more important is that equation (32) is found to be an asymptotic formula. This is shown in the right panel of Figure 6, where we have plotted the \( (P_s - P_d) \) at regular intervals of time, and fitted it to an inverse time curve \( D/t^2 \), where \( D \) is some constant. Here, \( P_d \) is the result obtained by numerically summing over the areas of each spike (see Figure 4), found by the integrated difference of \( \theta_3(x_i, t, u) \) and \( \theta_4(x_i, t, u) \) which define the hyperbolas in the \( x, \theta \) plane. We find that the \( (P_s - P_d) \) fits perfectly into \( D/t^2 \), where \( D \) needs to be calculated by a numerical fit. Thus we confirm that the approximation chosen in section V was justifiable from the asymptotic convergence to the integral \( P_h \). \( D \) is simply the coefficient of the second order term in:

\[ P(t) = \frac{(3\ln3 + 4)(a + h_1)^2 + (a - h_2)^2}{4(4a + 2\pi r)t} + \frac{D}{t^2} + O(1/t^2). \]
FIG. 6: (Colour online) \textit{Left}: Plot comparing the survival probability $P_s$ (light green/horizontal line) found by equation (32), with the numerical survival probability $P_d$ (dark blue/curve), found by direct numerical simulation, both multiplied by the time $t$. \textit{Right}: The difference between equation $P_s$ and $P_h$ which is found by numerically integrating hyperbolas and summing the relative areas under the spikes created, decays as $D/t^2$. $D$ is the coefficient of the second order term in equation (33).

VII. CONCLUSION AND DISCUSSION

In this paper, we have investigated the open stadium billiard and managed to derive, using phase space methods, an expression consisting of the two main contributions (see equations (8) and (31)) to the long time asymptotic tail of the survival probability function of the stadium billiard. Both expressions are to leading order in $t$. The second one (equation (31)) is an approximate result which converges to the true result as $\sim t^{-2}$ which means the errors are $O(1/t^2)$ and hence do not appear in the simple closed form of equation (32). The expression has been confirmed through numerical simulation (see Figure 6 \textit{Left}). In total, we confirm that the survival probability of the stadium for long times goes as $\text{Constant} t^{-2}$, and we find that the \text{Constant} depends quadratically on the lengths of the parallel segments of the billiard on either side of the hole and hence the size of the hole as well as its position on one of the straight segments of the boundary (see equation (32)).

In the context of stadia, there is a variety of possible shapes for which one can observe similar properties. In this paper we only considered the standard stadium that is a construction of two parallel straight lines and two complete semicircular arcs. It is also possible to construct different ergodic stadia by using circular arcs of lengths less than $\pi r$ or by using elliptical arcs [61]. In both cases we expect ergodicity and an initial strong decay of the survival probability followed by an asymptotic power law decay at longer times, provided that the parallel straight sides are still present, and that the dynamics remain defocusing [62]. Hence, similar methods used in the present paper should be applicable to some variations of the stadium geometry as explained above.

At this point, we would like to comment on the $\ln 3$ term, which first appeared in equation (28). Similar terms where found in the work of both Bálint and Gouëzel [44] and Armstead’s \textit{et al.} [54] as well as several other papers relating to the stadium’s bouncing ball orbits and its long time dynamics. It appears, that the $\ln 3$ term is a direct consequence of the geometry of our stadium billiard. More specifically, the circular curvature of the boundary near the straight segments, leads to a reflected final angle $|\theta_f| \in ([\theta_i]/3, 3|\theta_i|)$, if $\theta_i$ is small enough; this follows from equations (10) and (11) above. Hence we propose that any bounded change in the curvature of the focusing segments of the billiard, such that the boundary remains $C^1$ smooth, would change the dynamics quantitatively but not qualitatively (i.e. $P(t) \sim \text{Constant} t^{-2}$).

Further work on this subject may include researching the open stadium with holes on the circular segments. Such an example is expected to behave very similarly to the case described in the present paper as is numerically shown in [53]. This is because the trajectories which dominate and survive for long times, again will be characterised by small (near vertical) angles. Their collisions will mainly be with the straight segments of the billiard, but also on very short segments of the semicircular arcs. What is obviously different in such a situation is that the number of collisions with the semicircular arcs is not restricted to only one, as was the case here. This fact will complicate the dynamics substantially. Therefore one might prefer to choose a probabilistic approach to such a problem, as suggested by in Armstead’s \textit{et al.} [54], rather than an analytic one. One might also be interested in addressing this from a different perspective such as a semiclassical approach or even a quantum mechanical one and hopefully obtain some sort of correspondence between results.
Acknowledgements

We would like to thank Uzy Smilansky for helpful discussions, the anonymous referees for their helpful suggestions and OG’s EPSRC Doctoral Training Account number SB1715.
14