Dual Billiards, Twist Maps and Impact Oscillators

Philip Boyland
Institute for Mathematical Sciences
SUNY at Stony Brook
Stony Brook, NY 11794
Internet: boyland@math.sunysb.edu

Abstract. In this paper techniques of twist map theory are applied to the annulus maps arising from dual billiards on a strictly convex closed curve Γ in the plane. It is shown that there do not exist invariant circles near Γ when there is a point on Γ where the radius of curvature vanishes or is discontinuous. In addition, when the radius of curvature is not C^1 there are examples with orbits that converge to a point of Γ. If the derivative of the radius of curvature is bounded, such orbits cannot exist. The final section of the paper concerns an impact oscillator whose dynamics are the same as a dual billiards map. The appendix is a remark on the connection of the inverse problems for invariant circles in billiards and dual billiards.

§0 Introduction

Dual billiards is a dynamical system defined on the exterior of an oriented convex closed curve Γ in the plane. If z is a point in the unbounded component of $\mathbb{R}^2 - \Gamma$, then its image under the dual billiards map Φ is the reflection about the point of tangency in the oriented supporting line to Γ (see Figure 0.1). It is clear that Φ is an area-preserving, and if Γ is strictly convex, it is homeomorphism. Further, Φ is affine invariant, *i.e.* transforming the convex curve by an affine map simply transforms the entire homeomorphism. If the initial curve is an ellipse, then its exterior is foliated by concentric homothetic ellipses that are Φ-invariant.

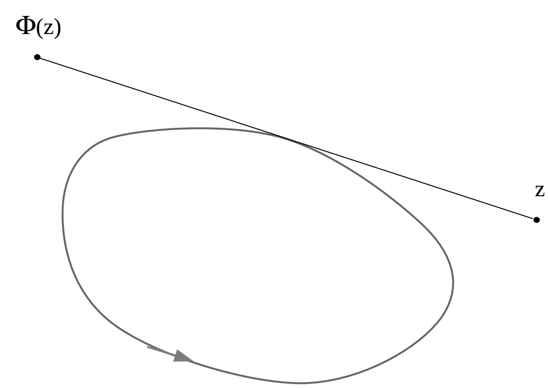

Figure 0.1: The dual billiards map.

The invention of dual billiards is credited to B. H. Newman in [Ms2]. The nomenclature outer billiards or cobilliards is also used. The name dual billiards is apt because this dynamical system is in certain ways the dual of the usual billiards. Indeed, many of the theorems given here are analogs of theorems for billiards. There are, however, fundamental

The author was partially supported by NSF grant # 431-4591-A

differences between the systems. Perhaps the most important is that the phase space for the dual billiards map is non-compact and has infinite area. An obvious obstruction to any explicit duality between the two systems is that the billiards map defined using an ellipse has period-two points of elliptic and saddle type.

One of the goals of this paper is to show how dual billiards fits into recent developments in the theory of area-preserving twist maps of the annulus due mainly to Mather. Once it is known that a dynamical system falls into this class there are a wealth of results available which yield a great deal of information about the dynamics of the map. There are technical differences, however, between the twist maps that arise in dual billiards and the usual classes of twist maps that are studied in the literature.

Section 1 begins by specifying a class of twist maps called half-cylinder twist maps that includes those from dual billiards. The rest of §1 is devoted to describing some standard tools and results of twist map theory and remarking on how these can be adapted to the new class. We only describe that part of twist map theory that is relevant. For a broader perspective the reader is referred to the various papers of Mather (summarized in [M5]), as well as [Bg], [Ms3], [Ms4], chapter 1 of [H] and chapter 10 of [MH]. There are surveys with a more physical point of view in [Mc] and [Me].

Section 2 give necessary background material on convex curves and in §3 it is confirmed that, in the coordinates introduced in §2, the dual billiards map is indeed a half-cylinder twist map. The next section deals with topological circles that are invariant under the dual billiards map. The question of the existence of invariant circles is central in twist map theory. An invariant circle provides a barrier through which orbits cannot pass, and thus the existence of invariant circles is intimately connected with stability questions. Section 4 begins with two results of R. Douady that give existence of invariant circles near Γ and near infinity. These results require a certain smoothness as they ultimately depend on KAM theory.

The inverse problem for invariant circles is considered next: given a convex curve Γ_1 is there a Γ whose dual billiards map has Γ_1 as an invariant circle. The solution to this problem is given by a geometric construction called the area envelope. The analog of this construction for billiards is discussed in the appendix where it is pointed out that both the billiard and dual billiard constructions are a consequence of the constancy of an area function (Mather's big "H") that can be associated with an invariant circle of a twist map (*cf.* Proposition 1.3).

The next results in §4 concern the non-existence of invariant circles near Γ when there is a point on Γ where the radius of curvature vanishes or is discontinuous. In addition, when the radius of curvature is not C^1 there may also exist orbits that converge to a point of Γ. These so-called crash solutions are shown to exist for certain examples in §5. It is also shown that if the derivative of the radius of curvature is bounded, no such orbits exist.

There are no known examples of strictly convex curves for which the dual billiards map lacks invariant circles near infinity. This problem (raised by Moser) remains one of the most important outstanding problems in the theory of dual billiards.

The final section of the paper concerns an impact oscillator whose dynamics are the same as a dual billiards map. Consider a wall that is periodically moving with position given by $p(t)$ where $p(t)$ satisfies $\ddot{p} + p = \rho(t)$ for some 2π-periodic forcing function ρ with $\int \exp(it)\rho(t) = 0$. A simple harmonic oscillator with position specified by $x(t)$ satisfying

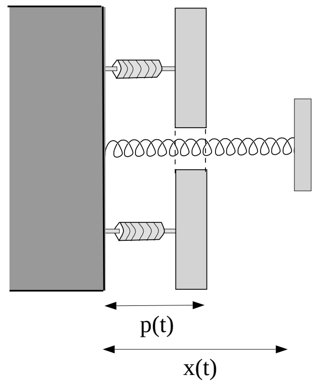

Figure 0.2: The impact oscillator.

$\ddot{x} + x = 0$ is assumed to have perfectly elastic collisions with the wall. Figure 0.2 shows one possible (idealized) realization of the system. On the left is a fixed wall whose position is set at the origin. A pair of pistons periodically drive a second wall that is to the right of the fixed wall. A flat plate is connected to the fixed wall by a spring that passes through a hole in the moving wall. We monitor the motions of the plate as it moves away, collides with the moving wall, moves away, etc. The return map to collisions with the moving wall in appropriate coordinates turns out to be identical to dual billiards on a convex curve Γ whose radius of curvature function ρ is the same as the forcing function for the wall.

As a consequence of the equivalence of the two dynamical systems one can use results about dual billiards to provide information about the oscillator. For example, as a consequence of the Birkhoff-Mather stability theorem (see Remark 4.3.2) when there do not exist Φ-invariant circles near the curve Γ, there must exist orbits that converge to Γ in forward and backward time. Thus as a consequence of Theorem 4.3, if there is an instant at which the forcing function is zero or discontinuous, then there are solutions for the impact oscillator that converge to the wall. In addition, it follows from Theorem 5.1 that there are impact oscillators that have crash solutions that converge to the wall in finite time. These solutions do not exist if the forcing function is C^1 (Theorem 5.2).

The situation near infinity is perhaps of more physical interest. As a consequence of Theorem 4.1, for sufficiently smooth convex curves the dual billiards map has invariant circles near infinity. Thus for sufficiently smooth forcing functions the impact oscillator has no unbounded solutions. Because of the lack of unbounded solutions these oscillators are called *stable*. There have been numerous other impact oscillators studied in the literature. Their stability depends on the details of the oscillator, *eg.* on the restoring force of the particle to the wall, the relation of the frequency of the forcing to that of the free oscillator, etc. Again using the Birkhoff-Mather stability theorem, the non-existence of invariant circles near infinity for dual billiards on some convex curve would imply unbounded solutions for the corresponding impact oscillator. This gives a physical motivation for the question of Moser noted above.

This paper has several purposes, and it contains material of different types. A first goal is to survey known results from the point of view adapted here. For this reason there is an overlap in basic material with other papers that have appeared on dual billiards, for example, [GK], [T1], [T2], and [T3]. A second goal is to give a mathematical framework for future work on dual billiards. Thus there are somewhat technical sections (*eg.* §1.1) that could be skipped by the reader with more physical interests. Despite these rigorous goals we also want to preserve the essentially simple and intuitive geometric character of the problem. Thus there also is descriptive material. Much of this material also serves to illustrate geometric heuristics associated with twist map theory.

Acknowledgments: Most of the results in this paper were obtained while the author was a Postdoctoral Research Fellow at the Mathematical Sciences Research Institute in Berkeley, 1988-89. He would like to thank them (belatedly) for there support. Thanks also to Danny Goroff, Dick Hall, and Robert MacKay for useful and stimulating conversations.

§1 Half-cylinder twist maps

In this section we introduce a class of area-preserving monotone twist maps called half-cylinder twist maps, and point out that the standard results from twist map theory apply to this class. We will see in §3 that a dual billiards map in the appropriate coordinates is a half-cylinder twist map.

§1.1 Definition of half-cylinder monotone twist maps. Let the one point compactification of the ray $[0, \infty)$ be denoted $[0, \infty]$ with the added point being labeled ∞. Let C be the open half-cylinder $S^1 \times (0, \infty)$ and C^* be $S^1 \times [0, \infty]$, with the circle $S^1 = \mathbb{R}/(2\pi\mathbb{Z})$. The space C^* is homeomorphic to the compact annulus, but we maintain the infinite measure. The universal covers of these spaces are denoted by $\tilde{C}$ and $\tilde{C}^*$, respectively. A tilde will always indicate the lift of a point, function, set, etc. to the universal cover. The map π_1 in various contexts means the projection of a product space onto its first component. We will use the coordinates (θ, γ) for C and C^* and (x, γ) for $\tilde{C}$ and $\tilde{C}^*$. If g is a real valued function, g_i denotes the derivative with respect to the i^{th} variable.

The definition of a half-cylinder monotone twist map requires two pieces of data. The first is a pair of functions $b_0, b_\infty : \mathbb{R} \to \mathbb{R}$ called the *boundary maps*. It is required that b_0 be nondecreasing and continuous from the right, b_∞ be increasing and continuous, $b_0 < b_\infty$, and $b_i(x+1) = b_i(x) + 1$ for $i = 0, \infty$. Let $\mathcal{D} = \{(x, x') \in \mathbb{R}^2 : b_0(x) < x' < b_\infty(x)\}$. For $i = 0, \infty$, $\Lambda_i \subset \mathbb{R}^2$ denotes the graph of b_i as a function of x. Note that $\Lambda_0 \cup \Lambda_\infty$ is contained in the frontier of $\mathcal{D}$.

The second piece of data needed to define a half-cylinder twist map is a C^1-function $h : \mathcal{D} \to \mathbb{R}$ called a *generating function* that satisfies:

(G1) (periodicity) $h(x + 2\pi, x' + 2\pi) = h(x, x')$

(G2) (conditions on the derivatives) $h_1 < 0$ and $h_2 > 0$, and for each $\epsilon > 0$, the derivative Dh is Lipschitz on each $\mathcal{D}_\epsilon = \{(x, x') \in \mathbb{R}^2 : b_0(x) < x' < b_\infty(x) - \epsilon\}$. Further, the mixed partial derivative $h_{12} = h_{21}$ exists, is continuous and satisfies $h_{12} < 0$.

(G3) (limit behavior near Λ_0) For $i = 1, 2$,

$$\lim_{(x,x') \to \Lambda_0} h_i(x, x') = 0 = \lim_{(x,x') \to \Lambda_0} h(x, x').$$

(G4) (limit behavior near Λ_∞) The first derivatives must satisfy

$$\lim_{x' \to b_\infty(x)} h_2(x, x') = \infty = \lim_{x \to b_\infty^{-1}(x)} h_1(x, x')$$

and for each x,

$$h_2(u,v) + h_1(v,w) \to 0 \text{ as } (u,v,w) \to (b_\infty^{-1}(x), x, b_\infty(x)).$$

Definition: *A map $f : C^* \to C^*$ is called a **half-cylinder monotone twist map** if it has a lift $\tilde{f} : \tilde{C}^* \to \tilde{C}^*$ that satisfies*

$$\begin{aligned} \gamma &= -h_1(x, x') \\ \gamma' &= h_2(x, x') \end{aligned} \tag{1.1}$$

where $(x', \gamma') = \tilde{f}(x, \gamma)$ for $(x, \gamma) \in \tilde{C}$, and h satisfies properties (G1) – (G4) above with b_i equal to $\tilde{f}$ restricted to $\mathbb{R} \times \{i\}$.

As a consequence of the properties (G1) – (G4), when restricted to C, a half-cylinder monotone twist map f will always be an area-preserving homeomorphism, and both f and f^{-1} will be locally Lipschitz. Conversely, given b_i and h satisfying (G1) – (G4), (1.1) implicitly defines a half-cylinder monotone twist map f. Using (1.1), $D\tilde{f}(x, \gamma)$ is given by

$$\frac{1}{h_{12}} \begin{pmatrix} -h_{11} & -1 \\ h_{12}^2 - h_{11}h_{22} & -h_{22} \end{pmatrix} \tag{1.2}$$

where each of the derivatives h_{ij} is evaluated at (x, x') with $x' = \pi_1(\tilde{f}(x, \gamma))$.

Remarks

1.1.1 The name "twist map" is given because the image of a vertical arc is always the graph of a function defined on a subset of the circle. From (1.2) one sees that this is a consequence of $h_{12} < 0$. The generating function h can be given a geometric interpretation as in Figure 1.1 ([K]). The area of the region bounded by the x axis, the vertical arc above x', and the image under $\tilde{f}$ of the vertical arc above x is equal to $h(x, x')$.

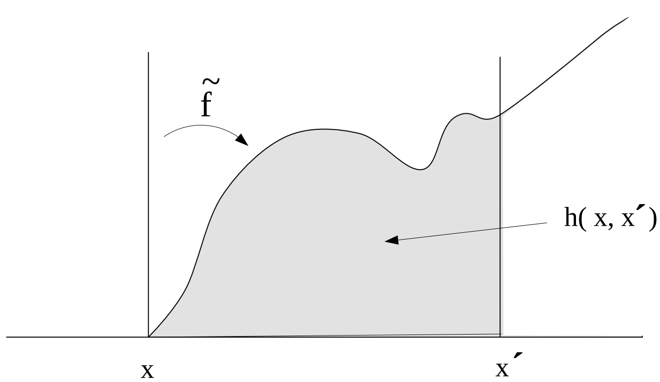

Figure 1.1: The geometric meaning of the generating function.

1.1.2 There are a few differences between half-cylinder twist maps and the classes of twist maps that occur in the literature. These differences are mainly "technical" in the sense that standard arguments work for the class with straightforward alterations. The consideration of these technical differences are forced on us by the twist maps that come from dual billiards.

A twist map is usually defined either on the compact annulus or on the infinite cylinder. In the later case one requires infinite wrapping, *i.e.* the image of a vertical line wraps infinitely around the cylinder. In contrast, half-cylinder twist map are defined on the half-infinite cylinder. They extend to a perhaps non-continuous map on the lower boundary, and can be extended to a homeomorphism on an upper circular boundary that is added at infinity. Thus they do not have infinite wrapping; they can be thought of as being defined on a compact annulus with infinite area.

If a twist map f is differentiable at a point z_0 the ***positive twist*** at z_0, denoted $twist^+(z_0)$, is the angle (measured clockwise) from a vertical vector based at $f(z_0)$ to the image under Df of a vertical vector based at z_0. The lift of a half-cylinder twist map $\tilde{f}$ may not be differentiable, but the image of the vertical line through $\tilde{z}_0 = (x_0, \gamma_0)$ is always the graph of the Lipschitz function of x', $h_2(x_0, x')$. In this case we let $twist^+(z_0) = \text{arcot}(K)$ where K is the best Lipschitz constant for this function at $x'_0 = \pi_1(\tilde{f}(z_0))$. The ***negative twist*** at z_0, denoted $twist^-(z_0)$, is the twist at z_0 of $S \circ \tilde{f}$ where $S(x, \gamma) = (-x, \gamma)$. The ***twist*** at z_0 is $twist(z_0) = \min\{twist^+(z_0), twist_-(z_0)\}$. Note that if h is twice differentiable, $twist(z_0) = \min\{\text{arcot}(h_{11}(\hat{x}_0, x_0), \text{arcot}(h_{22}(x_0, x'_0)\}$, where $\hat{x}_0 = \pi_1(\tilde{f}^{-1}(z_0))$.

In twist map theory it is usually assumed that the twist is bounded away from zero. For half-cylinder twist maps the condition given in (G2) requiring that Dh is Lipschitz in $\mathcal{D}_\epsilon$ insures that the twist is bounded away from zero in neighborhoods of the lower boundary of C. However, f itself may only be only locally Lipschitz as its Lipschitz constant can go to infinity as one approaches the lower boundary. In a half-cylinder twist map the twist will not be bounded away from zero as one approaches infinity as the map extends to the circle at infinity.

§1.2 The variational formulation and dynamics of twist maps. Note that (1.1) says that h is a generating function of a canonical transformation in the sense of classical mechanics (*eg.* [MH], page 47). The function h plays another role and is sometimes called the ***action*** as it can be used in a discrete variational formulation. What follows is the rudiments of this theory adapted for half-cylinder twist maps. For a full treatment the reader is referred to [Bg].

A ***configuration*** is an element $\underline{x} = (x_i)_{i\in\mathbb{Z}}$ of the set $\mathbb{R}^{\mathbb{Z}}$ of bi-infinite sequences of real numbers (the nomenclature "configuration" comes from solid state physics). Given a generating function h and boundary maps b_i as above, let $X = \{\underline{x} \in \mathbb{R}^{\mathbb{Z}} : (x_i, x_{i+1}) \in \mathcal{D} \text{ for all } i\}$. Elements of X are called ***allowable configurations***. The generating function h can be extended to finite segments of allowable configurations via

$$h(x_j, \ldots, x_k) = \sum_{i=j}^{k-1} h(x_i, x_{i+1}).$$

An allowable configuration $\underline{x}$ is called ***stationary*** if for all pairs (j, k) with $j < k$, $h(x_j, \ldots, x_k)$ is stationary with respect to variations with fixed endpoints. The configuration is called ***minimizing*** if for all pairs (j, k) with $j < k$, $h(x_j, \ldots, x_k)$ is minimal with respect to variations with fixed endpoints. Minimizing configurations are always stationary. A configuration is stationary if and only if

$$h_2(x_{i-1}, x_i) + h_1(x_i, x_{i+1}) = 0 \tag{1.3}$$

for all i, or as a consequence of (1.1), precisely when $\underline{x}$ is a sequence of x coordinates of an orbit of the map $\tilde{f}$, *i.e.* for all i, $x_i = \pi_1(\tilde{f}(x_0, \gamma_0))$ for some initial $(x_0, \gamma_0) \in \tilde{C}$.

A configuration is ***monotone*** if for each fixed $(m, n) \in \mathbb{Z}^2$ either $x_{i+m} > x_i + n$ for all i, or else, $x_{i+m} < x_i + n$ for all i. A basic feature of the theory is that minimizing configurations are always monotone. An orbit (x_i, γ_i) of $\tilde{f}$ is ***monotone*** if the sequence (x_i) is. An orbit (θ_i, γ_i) of f is ***monotone*** if it has a lift to $\tilde{C}$ that is monotone. Roughly speaking, a monotone orbit is one on which f preserves the radial order. An invariant set is called monotone if each of its orbits is.

Given the lift of a half-cylinder twist map $\tilde{f}$ the ***rotation number*** of a point $\tilde{z} \in \tilde{C}^*$ is

$$rot(\tilde{z}, \tilde{f}) := \lim_{n \to \infty} \frac{\pi_1(\tilde{f}^n(\tilde{z})) - \pi_1(\tilde{z})}{2\pi n}$$

if the limit exists. The rotation number of a point $z \in C^*$ is the rotation number of one (and thus all) of its lifts. Note that this is defined only up to a choice of lift of f, *i.e.* up to integer translation. The ***rotation set*** of $\tilde{f}$ is $rot(\tilde{f}) := \{rot(\tilde{z}, \tilde{f}) : \tilde{z} \in \tilde{C}^*\}$. The rotation number of a point under a circle map is defined similarly, only in this case, if $g : S^1 \to S^1$ is nondecreasing (but not necessarily continuous), then the rotation number exists and is the same for all points in the circle. This common number is called the ***rotation number*** of the map, and is denoted $rot(g)$. Another basic feature of twist map theory is that monotone orbits always have a rotation number. This is basically because they behave like orbits of circle maps. Note that any nondecreasing $G : \mathbb{R} \to \mathbb{R}$ that satisfies $G(x + 2\pi) = G(x) + 2\pi$ may also be assigned a rotation number that will be denoted $rot(G)$.

A Denjoy minimal set for a half-cylinder twist map is a compact, invariant set X so that f restricted to X is conjugate to the exceptional minimal set in a Denjoy counterexample on the circle (see *eg.* [Dv], page 108). The phrase "invariant circle" in this paper always means a *homotopically nontrivial* circle that is invariant under the map. Since f restricted to an invariant circle, Ω, is a circle homeomorphism, it has a rotation number that will be denoted $rot(\Omega)$.

Theorem 1.1: *If f is a half cylinder twist map with boundary functions b_i, then:*

(a) (Aubry-Mather) The rotation set of f is $rot(f) = [rot(b_0), rot(b_\infty)]$. For each $p/q \in rot(f)$ with p and q relatively prime, there is a monotone periodic orbit with rotation number p/q and period q. For each $\omega \in rot(f) - \mathbb{Q}$, there is an invariant circle or Denjoy minimal set X_ω that is monotone and each $x \in X_\omega$ has rotation number ω.

(b) (Birkhoff) If $\Omega \subset C$ is a f-invariant circle, then Ω is the graph of a Lipschitz function $S^1 \to (0, \infty)$. Further, if $\mathcal{C}$ denotes the set of invariant circles (including $S^1 \times \{0\}$) with the Hausdorff topology, then $\mathcal{C}$ is closed and the map $rot : \mathcal{C} \to \mathbb{R}$ is continuous and monotonic in the sense that Ω_2 being contained in the open annulus bounded by Ω_1 and $S^1 \times \{0\}$ implies $rot(\Omega_1) > rot(\Omega_2) > rot(b_0)$.

(c) (Mather) If f has an invariant circle $\Omega \subset C$, then for almost every $(x_0, \gamma_0) \in \Omega$,

$$h_{11}(x_0, x_1) + h_{22}(x_{-1}, x_0) > 0$$

where $(x_i, \gamma_i) = f^i(x_0, \gamma_0)$.

Proof: Part (a) can be proven using variational methods in which case the monotone orbits are obtained as minimizing orbits (*eg.* [Bg]). It may also be proven using topological methods (*eg.* [MH]), or using monotone recursion relations ([A]). In part (b), the statement that invariant circles are Lipschitz graphs can be proved as in [M1] and [M2], or chapter 1 of [H]. The statement that the set of invariant circles is closed in the Hausdorff topology requires the observation (explained in Remark 1.2.2 below) that condition (G2) implies a uniform bound on the Lipschitz constant of functions $u : S^1 \to (0, \infty)$ whose graphs are invariant circles in $S^1 \times [0, K]$. The monotonicity of the rotation numbers on the circles is a standard consequence of the twist hypothesis.

To prove (c), let α denote the homeomorphism of the real line that is first component of $\tilde{f}$ restricted to the lift of the invariant circle, *i.e.* $\alpha(x) = \pi_1(\tilde{f}(x, u(x)))$. Since $\tilde{f}$ is locally Lipschitz and u is Lipschitz, α and α^{-1} are Lipschitz, and thus their derivatives exist almost everywhere. Equation (1.2) yields that

$$h_1(x, \alpha(x)) + h_2(\alpha^{-1}(x), x) \equiv 0. \tag{1.4}$$

Part (c) then follows by differentiating and using $h_{12} < 0$. □

Remarks:

1.2.1 A geometric interpretation of the condition in Theorem 1.1(c) can be given in terms of "tumbling tangents" (*cf.* [Mc] and [MP]). If h is twice differentiable, using (1.2) one sees that the slope of the image of the vertical tangent vector at the point (x, γ) is given by $h_{22}(x, x')$, where $x' = \pi_1(\tilde{f}(x, \gamma))$. Similarly, under $\tilde{f}^{-1}$, the slope of the image of the vertical tangent vector at the point (x', γ') is $-h_{11}(x, x')$.

Now assume there is an invariant circle that passes through the three points (x_i, γ_i) for $i = -1, 0, 1$ where $(x_{i+1}, \gamma_{i+1}) = \tilde{f}(x_i, \gamma_i)$. Theorem 1.1(b) says that the circle is the graph of a Lipschitz function, and thus the unit tangent vector in the direction of the circle exists almost everywhere; let us assume that it exists at our triple. Since the circle is invariant, the collection of these circle directions is invariant under the induced action of $D\tilde{f}$ on the unit tangent bundle. Now since $\tilde{f}$ is a twist map, the image of a unit vertical vector always points in the positive x direction. Because the three points are on an invariant circle and the circle directions are invariant, the second iterate of the vertical vector at (x_{-1}, γ_{-1}) cannot rotate past the downward vertical.

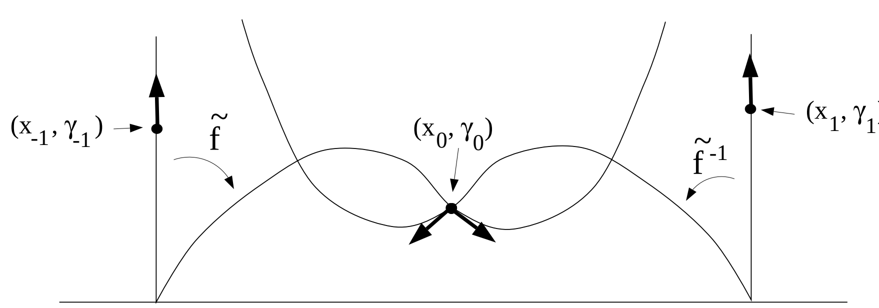

Figure 1.2: The tumbling tangents criterion for the non-existence of invariant circles.

Rather than apply these considerations to the second iterate, it is usually more convenient to iterate forward from (x_{-1}, γ_{-1}) and backwards from (x_1, γ_1). Since the slope of the forward image of a vertical vector is h_{11} and a backward image is $-h_{22}$, the "no tumbling tangents" condition is exactly that given in Theorem 1.1(c). Figure 1.2 shows an example in which there can be no invariant circle containing the three points.

By using more iterates, these considerations can be refined to obtain the converse KAM theory of MacKay and Percival ([MP]). The macroscopic version of tumbling tangents is the existence of non-monotone periodic orbits as in [B+H].

1.2.2 The regularity conditions imposed on half-cylinder twist maps were chosen so that the standard proofs of Theorem 1.1(b) and (c) would go through with little change. In order to differentiate (1.4) almost everywhere, we need that f is locally Lipschitz. For this it is enough that Dh be locally Lipschitz. We also need u to be Lipschitz. For this the proofs ([M1], [M2], [H]) require only f is a locally Lipschitz twist map whose twist is bounded away from zero in a neighborhood of the circle. Again this requires only that Dh be locally Lipschitz.

However, to get a uniform bound in neighborhoods of the lower boundary for the Lipschitz constants of the functions $u : S^1 \to (0, \infty)$ that define invariant circles, we need a uniform bound on the twist (as defined in Remark 1.1.2). This requires that h_{11} and h_{22} be bounded and hence the requirement in (G1) that Dh is Lipschitz in $\mathcal{D}_\epsilon$.

One can geometrically see the dependence of the Lipschitz constants of the functions u by thinking of the twist as defining a cone about the vertical in the tangent bundle at each point x. By definition, the image under $D\tilde{f}$ of a vertical vector based at $\tilde{f}^{-1}(x)$ misses this cone as does the image of the vertical vector based at $\tilde{f}(x)$ under $D(\tilde{f}^{-1})$. This implies that vectors in this cone are rotated beyond the vertical in either forward or backward iteration. Thus applying the considerations of the previous remark, no tangent to an invariant circle can lie in this cone.

§1.3 Monotone recursion relations. The point of view of Angenent will be used to construct examples in §5.1. The basic idea in [A] is to use (1.3) directly to define a recursion relation that must be satisfied by sequences of x coordinates of orbits.

We will state these ideas just in the limited context needed here. Assume that h and b_i satisfy the properties (G1) – (G4) above and let

$$\Delta(x_{-1}, x_0, x_1) = -(h_1(x_0, x_1) + h_2(x_{-1}, x_0)). \tag{1.5}$$

The properties of Δ that cause it to behave like the monotone recursion relations in [A] are: (1) (monotonicity) Δ is a nondecreasing function of x_{-1} and x_1, (2) (periodicity) $\Delta(x_{-1}, x_0, x_1) = \Delta(x_{-1} + 2\pi, x_0 + 2\pi, x_1 + 2\pi)$, and (3) (coerciveness) which is given by the condition in the first part of (G4). The coerciveness condition differs somewhat from that adopted in [A], but the proof of Theorem 1.2 below can be constructed with little difficulty.

An allowable configuration $\underline{x}$ is called a ***solution*** to the monotone recursion relation given by Δ if $\Delta(x_{i-1}, x_i, x_{i+1}) = 0$ for all i. Note that a solution always gives the sequence of x coordinates of an orbit of the lift $\tilde{f}$ of the half-cylinder twist map defined by the generating function h. The configuration $\underline{x}$ is a ***subsolution*** if $\Delta(x_{i-1}, x_i, x_{i+1}) \geq 0$ for all i and a *supersolution* if $\Delta(x_{i-1}, x_i, x_{i+1}) \le 0$ for all i. The standard partial order on configurations is given by $\underline{x} \le \underline{y}$ if and only if $x_i \le y_i$ for all i. The main result we need is:

Theorem 1.2: (Angenent) *Let Δ be defined by (1.5) for a generating function h associated with a half-cylinder twist map. If $\underline{x}$ and $\underline{y}$ are a subsolution and a supersolution, respectively, of the monotone recursion relation generated by Δ, and further, $\underline{x} \le \underline{y}$, then there exists a solution $\underline{z}$ with $\underline{x} \le \underline{z} \le \underline{y}$.*

§**1.4 The area function of an invariant circle.** The next result concerns the constancy of a certain area that can be associated with an invariant circle. This quantity, sometimes called "Mather's big H", has a simple geometric interpretation in billiards and dual billiards (see the appendix and Theorem 4.2).

Let Ω be a (not necessarily invariant) circle in C that is the graph of $u : S^1 \to (0, \infty)$. We will also use u for the function $\mathbb{R} \to (0, \infty)$ that has $\tilde{\Omega} \subset \tilde{C}$ as its graph. Given the lift of a half-cylinder twist map f, we can define two functions associated with the displacement of $\tilde{\Omega}$ under $\tilde{f}$. Let $\alpha : \mathbb{R} \to \mathbb{R}$ be defined as $\alpha(x) = \pi_1(\tilde{f}(x, u(x)))$, or equivalently, $\alpha(x)$ is the unique solution to $-h_1(x, \alpha(x)) = u(x)$. And let $\beta : \mathbb{R} \to \mathbb{R}$ be defined by $\beta(x) = \pi_1(\tilde{f}(x, \sigma(x)))$ where $\sigma(x) = \sup\{\gamma : (x, \gamma) \in \tilde{f}^{-1}(\tilde{\Omega})\}$. Equivalently, $\beta(x) = \sup\{x' : h_2(x, x') = u(x')\}$ (see Figure 1.3).

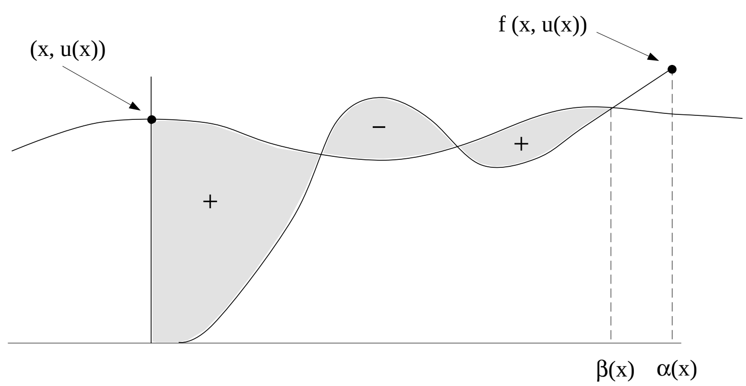

Figure 1.3: The definitions of α, β, and the area function.

From the definitions, α is continuous and $\beta \le \alpha$. Since $h_{12} < 0$, β is strictly increasing. Note that if $\eta > \beta(x)$, then since $\lim_{x' \to b_\infty(x)} = \infty$, we have $h_2(x, \eta) > u(\eta)$. Also note that $\tilde{\Omega}$ is $\tilde{f}$-invariant if and only if $\alpha = \beta$.

If the generating function associated with f is h, define the ***area function*** of Ω as

$$A(x) := \int_x^{\beta(x)} u(\eta)\, d\eta - h(x, \beta(x)).$$

Thus A(x) measures the signed area between the Ω and the image of the vertical arc above x (see Figure 1.3 again). Mather proves in [M3] that the area function of an invariant circle is always constant.

Proposition 1.3: (Mather) *Given a half-cylinder twist map f, the graph Ω of a continuous function $u : S^1 \to (0, \infty)$ is an f-invariant circle if and only if its area function is constant.*

Proof: First assume that Ω is f-invariant. By Theorem 1.1(b), u is a Lipschitz function and since f is locally Lipschitz, $\alpha = \beta$ is Lipschitz and so its derivative exists almost everywhere. Differentiating $A(x)$ using its definition and then applying (1.2) one gets that $dA/dx = 0$ almost everywhere, and since A is continuous, it is constant.

Now conversely, assume that A is constant and proceed by contradiction. If $\alpha \neq \beta$, then since α is continuous and $\beta \leq \alpha$, there exists x and $\bar{x}$ with $x < \bar{x}$ and $\alpha(\eta) > \beta(\bar{x})$ for all $\eta \in (x, \bar{x})$. Thus for these η, $h_1(\eta, \beta(\bar{x})) > h_1(\eta, \alpha(\eta)) = -u(\eta)$. Note also that above we showed that $h_2(x, \eta) > u(\eta)$ for all $\eta > \beta(x)$. Now by assumption, $A(x) - A(\bar{x}) = 0$, and so

$$\begin{aligned}
-\int_x^{\bar{x}} u(\eta)\, d\eta &= -h(x, \beta(x)) + h(\bar{x}, \beta(\bar{x})) - \int_{\beta(x)}^{\beta(\bar{x})} u(\eta)\, d\eta \\
&= -h(x, \beta(x)) + h(x, \beta(\bar{x})) - \int_{\beta(x)}^{\beta(\bar{x})} u(\eta)\, d\eta \\
&\quad - h(x, \beta(\bar{x})) + h(\bar{x}, \beta(\bar{x})) \\
&= \int_{\beta(x)}^{\beta(\bar{x})} (h_2(x, \eta) - u(\eta))\, d\eta + \int_x^{\bar{x}} h_1(x, \beta(\bar{x}))\, d\eta \\
&> -\int_x^{\bar{x}} u(\eta)\, d\eta
\end{aligned}$$

where in the last inequality we have used the fact that β is strictly increasing. □

§2 Convex curves and envelope coordinates.

In this section we review some standard material from the theory of convex curves in the plane. For more details the reader is referred to [E], [G1], and [G2].

§2.1 The height function and the radius of curvature. If Γ is a convex closed curve in the plane, let $U(\Gamma)$ denote the unbounded component of $\mathbb{R}^2 - \Gamma$. It will be convenient to assume that the origin is *not* contained in $U(\Gamma)$ and that Γ is oriented in a counterclockwise direction. Let $\mathbf{u}_\theta = (\cos(\theta), \sin(\theta))$ and $\mathbf{n}_\theta = (-\sin(\theta), \cos(\theta))$. For each $\theta \in S^1 = \mathbb{R}/(2\pi\mathbb{Z})$, $\mathcal{L}_\theta$ is the supporting line to Γ that is parallel to $\mathbf{n}_\theta$ and near its point of intersection Γ is oriented in the same direction as $\mathbf{n}_\theta$. It will also be convenient to assume that $\mathcal{L}_0$ intersects Γ in an extremal point. The ***height function*** of Γ, $p(\theta)$, is the distance from the origin to $\mathcal{L}_\theta$ (see Figure 2.1(a)). Equivalently, $p : S^1 \to \mathbb{R}$ is given by $p(\theta) = G(\mathbf{u}_\theta)$ where $G(z) = \max\{z \cdot w : w \in \Gamma\}$.

If $\mathcal{L}_\theta$ intersects Γ in an extremal point, let $\alpha(\theta)$ be this intersection. If $\mathcal{L}_\theta$ intersects Γ in a non-trivial segment, let $\alpha(\theta)$ be the extreme endpoint of the intersection in the counterclockwise direction. The arc length along Γ from $\alpha(0)$ to $\alpha(\theta)$ is denoted by $s(\theta)$. Note that s is nondecreasing, continuous from the right and

$$\alpha(\theta) - \alpha(0) = \int_0^\theta \mathbf{n}_\eta\, ds(\eta)$$

as a Riemann-Stieltjes integral. The points of discontinuity of s are exactly the countable set of θ for which $\mathcal{L}_\theta$ intersects Γ in a non-trivial segment.

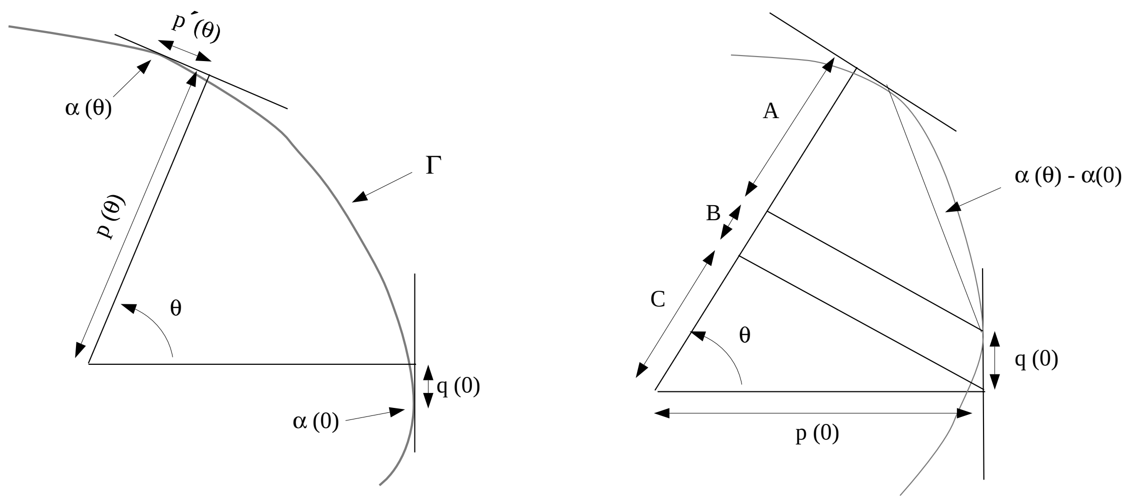

Figure 2.1: (a) The height function; (b) The geometry of equation (2.1).

A simple geometric argument (see Figure 2.1(b)) shows that

$$
\begin{aligned}
p(\theta) &= -\mathbf{n}_\theta \cdot (\alpha(\theta) - \alpha(0)) + q(0)\sin(\theta) + p(0)\cos(\theta) \\
&= \int_0^\theta \sin(\theta - \eta) ds(\eta) + q(0)\sin(\theta) + p(0)\cos(\theta). \qquad (2.1)
\end{aligned}
$$

In Figure 2.1(b) the three terms in this expression are labeled A, B and C, respectively. The quantity $q(0)$ is as pictured, explicitly, $q(0)$ is the signed distance from $\alpha(0)$ to $p(0)\,\mathbf{u}_0$ (the $q(0)$ shown in Figure 2.1(b) is positive). Note that the last two terms in (2.1) just reflect the position of the origin.

It follows from (2.1) that p is differentiable at all but a countable number of points where p' has a simple jump discontinuity and

$$
p'(\theta) = \int_0^\theta \cos(\theta - t) ds(t) - p(0)\sin(\theta) + q(0)\cos(\theta). \qquad (2.2)
$$

At points where p' has a jump discontinuity (those θ for which $\mathcal{L}_\theta$ intersects Γ in a segment) this formula may be interpreted in terms of right and left hand limits. In particular, $q(0) = p'(0)$ and p' has a geometric interpretation as the signed distance from $\alpha(\theta)$ to $p(\theta)\,\mathbf{u}_\theta$ (see Figure 2.1(a)). We shall be primarily interested in the case when Γ is strictly convex which corresponds to the condition that $s(\theta)$ is continuous and p is C^1.

If the arclength function $s(\theta)$ is differentiable, $\rho(\theta) := ds/d\theta$ is the radius of curvature of Γ, *i.e.* the radius of the osculating circle to Γ at the point $\alpha(\theta)$. Using (2.1) and (2.2),

$$
p'' + p = \rho \geq 0. \qquad (2.3)
$$

If the $s(\theta)$ is not differentiable, this equation may be interpreted in terms of distributions. For example, if Γ is a polygon, ρ is the sum of weighted Dirac delta functions. In any case, since s is nondecreasing, ρ exists almost everywhere. One interpretation of (2.3) is that the height function is given by a forced harmonic oscillator with forcing function given by the radius of curvature (in which case the last two terms in (2.1) reflect the initial conditions). Very informally, a closed curve is a point that has been forced outward by the radius of curvature. This observation will be crucial in connecting dual billiards with impact oscillators in §7.

Remark 2.1.1: For later reference we give a characterization due to J. Green of the type of periodic functions that can arise as height functions for a convex closed curve in the plane ([G1], [G2]). The characterization involves a generalized convexity. Let $S(x; A, B) = A\cos x + B\sin x$. A function $g : \mathbb{R} \to \mathbb{R}$ is said to be ***sub-sine*** if for each $x_1 < x_2 < x_1 + \pi$, $S(x; A_0, B_0) \geq g(x)$ for all $x \in [x_1, x_2]$, where A_0 and B_0 are the unique values for which $S(x_1; A_0, B_0) = g(x_1)$ and $S(x_2; A_0, B_0) = g(x_2)$. Green shows that a 2π-periodic function p is the height function of some convex closed curve if and only if p is sub-sine. This happens if and only if p can be written as in (2.1) for some nondecreasing function s with $\int \exp(i\theta)\, ds(\theta) = 0$. Further, a twice differentiable p is sub-sine if and only if $p'' + p \geq 0$.

In a similar vein note that if $s : [0, 2\pi) \to \mathbb{R}$ is any bounded, nondecreasing function with $\int \exp(i\theta)\, ds(\theta) = 0$ we can construct a Γ using (2.1) and (2.2). The resulting curve will have parameterization $\alpha(\theta) = R_\theta(p(\theta), p'(\theta-))$ where R_θ is rotation in the plane by the angle θ and $p'(\theta-)$ is the derivative at θ from the left. The curve Γ will be the convex hull of $\alpha(S^1)$. One can also use (2.1) to identify convex curves whose perimeter has length one with regular Borel probability measures ds on S^1 whose first Fourier coefficient vanishes. Atoms in the measures give flat spots (intervals in Γ); strictly convex curves correspond to measures without atoms.

§2.2 Envelope coordinates. The next step is to use Γ to give coordinates to the open (topological) annulus $U(\Gamma)$. For a point given in Euclidian coordinates $z \in U(\Gamma)$, let $\theta(z)$ be the unique angle θ with $z \in \mathcal{L}_\theta$ and $\alpha(\theta) - z$ is parallel to $\mathbf{n}_\theta$. A second coordinate for z is given by the distance from z to $\alpha(\theta)$ along $L_{\theta(z)}$ and is denoted $\ell(z)$ (see Figure 2.2). Note that the area elements are related by $dz = \ell\, d\ell\, d\theta = d\gamma\, d\theta$ where $\gamma = \ell^2/2$. Thus the map $z \mapsto (\theta, \gamma)$ is area preserving. The coordinates (θ, γ) will be called ***envelope coordinates.*** If Γ has corners, *i.e.* its radius of curvature function $\rho(\theta)$ vanishes on a nontrivial interval, the map $z \mapsto (\theta, \gamma)$ cannot be extended to a single valued function on Γ.

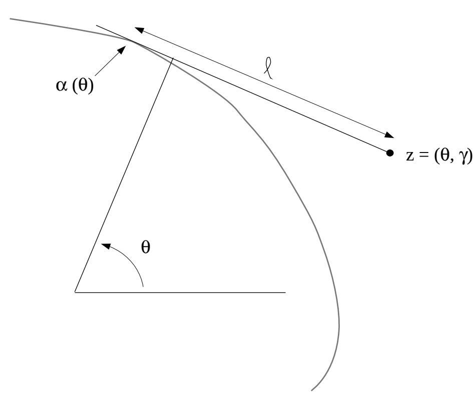

Figure 2.2: Envelope coordinates.

§3 Dual billiards yields a half-cylinder twist map

A geometric description of the dual billiard map arising from a convex curve Γ was given in the introduction. The purpose of this section is to show that under appropriate assumptions on Γ the dual billiards map in envelope coordinates is a half-cylinder twist map. It is geometrically clear that a dual billiards map preserves Euclidian area. The corresponding map in envelope coordinates is area-preserving because the change to envelope coordinates preserves area. The other main property that characterizes twist maps is that the image of a vertical ray is a graph over an arc in the circle. A vertical ray above the point $(\theta_0, 0)$ in envelope coordinates corresponds to the negative ray in $\mathcal{L}_{\theta_0}$ beginning at the point of tangency with Γ. The image of this ray under the dual billiards map is the positive ray in $\mathcal{L}_{\theta_0}$ that begins at the same point. Thus the condition that makes f a twist map is that this ray should hit each other $\mathcal{L}_\theta$ ($\theta_0 < \theta < \theta_0 + \pi$) in exactly one point. This follows from the convexity of Γ.

The dual billiards map on $U(\Gamma)$ is denoted Φ_Γ, or just Φ, and the corresponding map in envelope coordinates is denoted f_Γ or f, and its lift to $\tilde{C}$ as $\tilde{f}$.

§**3.1 The recursion relation from geometry.** Let Γ be a convex curve in the plane. It has associated functions α, p, etc. as defined in §2.1. Given two angles $\theta < \theta'$, let $\delta = \theta' - \theta$ and $\phi = \alpha(\theta') - \alpha(\theta)$. If $R = R(\theta, \theta')$, $L = L(\theta, \theta')$, ω_1, and ω_2 are as pictured in Figure 3.1, then the law of sines yields

$$\frac{\sin(\pi - \delta)}{|\phi|} = \frac{\sin(\omega_1)}{L} = \frac{\sin(\omega_2)}{R}. \tag{3.1}$$

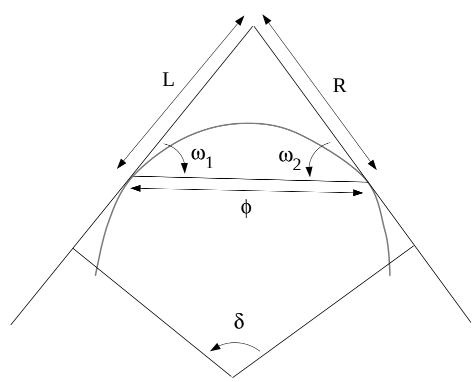

Figure 3.1: Geometric basis of the recursion relation.

The wedge (or cross product) between two vectors in the plane is $\mathbf{w} \wedge \mathbf{v} = w_1 v_2 - v_2 w_1 = |w||v| \sin(\omega)$ where ω is the angle from $\mathbf{w}$ to $\mathbf{v}$. Using this notation and (3.1) we have

$$L(\theta, \theta') = \frac{|\phi| \sin(\omega_1)}{\sin(\pi - (\theta' - \theta))} = \frac{\mathbf{n}_\theta \wedge \int_\theta^{\theta'} \mathbf{n}_\eta \, ds(\eta)}{\sin(\theta' - \theta)} = \frac{\int_\theta^{\theta'} \sin(\eta - \theta) \, ds(\eta)}{\sin(\theta' - \theta)}, \tag{3.2}$$

and a similar calculation yields

$$R(\theta, \theta') = \frac{\int_\theta^{\theta'} \sin(\theta' - \eta) \, ds(\eta)}{\sin(\theta' - \theta)}. \tag{3.3}$$

Thus using the geometric definition of the dual billiard map a triple of points $(\theta, \theta', \theta'')$ are the angles of an orbit of dual billiards if and only if $0 < \theta' - \theta < \pi$, $0 < \theta'' - \theta' < \pi$, and

$$L(\theta, \theta') = R(\theta', \theta'') > 0. \tag{3.4}$$

§3.2 The generating function and the boundary maps. Next we define the functions h and b_i that are the generating function and boundary maps for f_Γ. If we let

$$h(x, x') = \frac{1}{2}\int_x^{x'} (L(x,y))^2 \, dy = \frac{1}{2}\int_x^{x'} (R(y,x'))^2 \, dy \tag{3.5}$$

then $h_1 = -R^2/2$, $h_2 = L^2/2$, and the condition that $h_1(x', x'') + h_2(x, x') = 0$ is equivalent to (3.4).

To define the boundary maps, let $b_0(x) = \inf\{z : z > x \text{ and } \rho(z) > 0\}$. Perhaps put more simply, $b_0(x) = x$ if x is not in an interval in which ρ identically vanishes. If x is in such an interval, $b_0(x)$ is the right endpoint of that interval. The function b_0 is clearly continuous from the right, nondecreasing, satisfies $b_0(x + 2\pi) = b_0(x) + 2\pi$ and has rotation number equal to zero. Let $b_\infty(x) = x + \pi$.

Let us assume now that Γ is a convex curve whose arclength function $s(\theta)$ is Lipschitz. In this case the radius of curvature ρ exists almost everywhere, is bounded and satisfies $\rho \geq 0$. The derivatives at points where ρ exists are:

$$\begin{aligned}
R_1(x,x') &= -\rho(x) + \cot(x'-x)R(x,x') \\
R_2(x,x') &= \frac{L(x,x')}{\sin(x'-x)} \\
L_1(x,x') &= -\frac{R(x,x')}{\sin(x'-x)} \\
L_2(x,x') &= \rho(x') - \cot(x'-x)L(x,x').
\end{aligned} \tag{3.6}$$

It is now a simple matter to check that h and the b_i satisfy the necessary properties to define a half-cylinder twist map.

Theorem 3.1: *If Γ is a convex curve whose arclength function $s(\theta)$ is Lipschitz, then its dual billiard map f_Γ in envelope coordinates is a half-cylinder twist map with rotation set equal to* $[0, 1/2]$.

Remark 3.2.1: The function $h(\theta, \theta')$ has the geometric interpretation as the area bounded by Γ and the supporting lines $\mathcal{L}_\theta$ and $\mathcal{L}_{\theta'}$ ([Ms4]) (see Figure 3.2). This area is the same as the geometric interpretation of the generating function given in §1.1 because the change from Euclidian to envelope coordinates preserves area, The variational formulation also has a nice geometric interpretation in dual billiards. For example, to find a period three orbit of the dual billiards map, one finds the circumscribed triangle of smallest area (see Figure 3.3, left). Theorem 1.1(a) says that we can find a monotone periodic orbit of all rotation numbers. This corresponds to finding circumscribed polygons of each "rotation

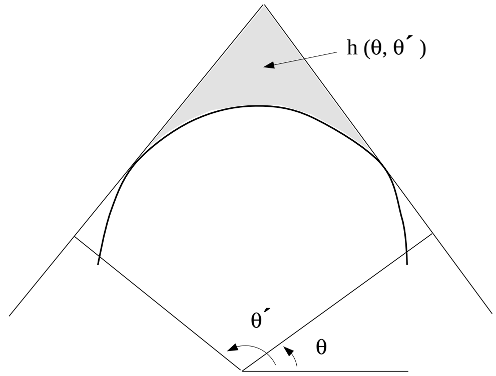

Figure 3.2: The geometric interpretation of the action.

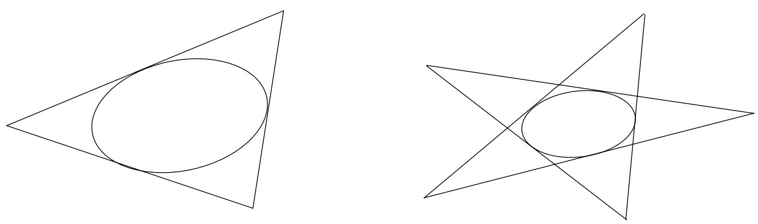

Figure 3.3: Using variational methods to find monotone periodic orbits with rotation number 1/3 (left) and 2/5 (right).

type" (see Figure 3.3, right). Further, there are the irrational analogs of these sets for each irrational in the rotation set.

Remark 3.2.2: If the curve Γ is strictly convex, then its arclength function $s(\theta)$ is continuous, and the half-cylinder twist map f_Φ on C is also. By inspecting (3.6) we see that this is sufficient to make h_{12} continuous, even though h_{11}, etc. may not exist. If s is C^r, then so is f_Γ.

If s is Lipschitz, Dh will be also, and so f_Φ is locally Lipschitz. However, as $|x-x'| \to 0$, $h_{12} \to 0$, and so (1.2) yields that $\tilde{f}$ is never Lipschitz, no matter how smooth s is. On the other hand, $\tilde{f}$ has bounded twist near the bottom boundary because h_{11} and h_{22} are bounded there. In fact, as $|x-x'| \to 0$, $h_{11} \to 0$, so the slope of the image of a vertical tangent goes to zero near the bottom edge.

Note that Φ_Γ always extends continuously to Γ by making it the identity there, but f_Γ will not have a continuous extension to $S^1 \times \{0\}$ if Γ has corners (nontrivial intervals where ρ vanishes). This is why discontinuous boundary maps b_0 had to be allowed in the definition of half-cylinder twist map. Note also that as $x'-x \to \pi$ (*i.e.* as $\gamma \to \infty$), $h_{22} \to \infty$ so the image of vertical tangent is near vertical and the twist of f_Γ goes to zero.

§4 Invariant circles for dual billiards.

As noted in the introduction, the question of the existence of invariant circles is central in twist map theory. An invariant circle for a dual billiards map is geometrically connected to the given convex curve via an equal area construction. Recall that in this paper the phrase "invariant circle" always means a homotopically nontrivial circle.

§**4.1 Existence of invariant circles.** The first theorem gives sufficient conditions on the curve Γ so that its dual billiards map has invariant circles near to Γ and near infinity. The Hausdorff distance between two compact sets, X and Y, is denoted $d(X,Y)$.

Theorem 4.1: (R. Douady) *Let Γ be a convex curve in the plane with a radius of curvature function $\rho(\theta)$ that is C^r.*

(a) If $r > 4$, then for each $N > 0$ the set of Φ_Γ-invariant circles Ω with $d(\Omega,\Gamma) > N$ has positive Lebesgue measure.

(b) If $r > 5$ and $\rho > 0$, then for each $\epsilon > 0$ the set of Φ_Γ-invariant circles Ω with $d(\Omega,\Gamma) < \epsilon$ has positive Lebesgue measure.

The basic idea of the proof in [D] (see also [Ms2]) is to extend the half-cylinder twist map f associated with Φ_Γ to the boundaries of C^* and then use a normal form result based on Herman's version of the curve translation theorem ([H]). In the compactification of the half-cylinder, f does not preserve a finite measure near infinity. However, it does satisfy the circle intersection property, so the curve translation theorem can be used to obtain invariant circles. The condition $\rho > 0$ in (b) is needed to extend f smoothly to $\mathbb{R} \times \{0\}$ (*cf.* Theorem 4.3(a) below).

§**4.2 Invariant circles and area envelopes.** The next result concerns the inverse problem for invariant circles, *i.e.* given a closed curve Γ_1 can you find a convex curve Γ_0 so that Γ_1 is an invariant circle for Φ_{Γ_0}. The solution involves a geometric construction called the area envelope.

Given a simple closed curve Γ_1 in the plane and a number a less than half the area enclosed by Γ_1, for each angle θ we can find a unique oriented line $\mathcal{L}_\theta$ in the direction u_θ so that the area to the right of the line and inside Γ_1 is equal to a. Denote the envelope of these lines as $AE(\Gamma_1, a)$ (see Figure 4.1). If $AE(\Gamma_1, a)$ is convex, it is an easy geometric exercise to show that the locus of the midpoints of the chords of the lines $\mathcal{L}_\theta$ in Γ_1 is the area envelope.♯ The next result is a consequence of this geometric fact. It appeared first in [D].

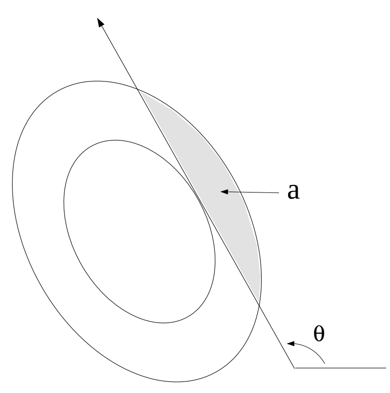

Figure 4.1: An area envelope.

♯ The author learned this from R. Ticciati who "discovered" it during a High School project in 1967

Theorem 4.2: (R. Douady) *If* Γ *is a convex curve with a Lipschitz arclength function, then a closed curve* $\Gamma_1 \subset U(\Gamma)$ *is invariant under the dual billiards map* Φ_Γ *if and only if* $\Gamma = AE(\Gamma_1, a)$ *for some* a.

Recall that the change to envelope coordinates is area-preserving. Thus if one considers the half-cylinder twist map corresponding to Φ_Γ, the quantity $A(x)$ (defined in §1.4) associated to the invariant circle is exactly the appropriate area a for which $\Gamma = AE(\Gamma_1, a)$. Thus Theorem 4.2 is a consequence of the more general fact given in Proposition 1.3. The appendix contains a remark on an analogous result for inner billiards.

Even if Γ_1 is convex, its area envelope need not be (*eg.* an area envelope inside an equilateral triangle). With the appropriate interpretation of chord, the area envelope will still be the locus of the midpoints. Area envelopes are discussed in more detail in [FT] and [GK].

§**4.3 The non-existence of invariant circles.** The next result gives sufficient conditions on Γ that imply the non-existence of invariant circles near Γ. These results are the analogs for dual billiards of results for billiards due to Mather and Hubacher. Conditions on the curvature in billiards play the role of the conditions on the radius of curvature in dual billiards.

The proof of part (a) goes exactly like that of Mather in [M1], and indeed, Theorem 1.1(c) is taken from that paper. The proof of (b) is unlike that of [Hr] in that we use estimates involving the recursion relation of the twist map, rather than geometric arguments. Part (a) is also contained in [GK]. That paper also contains estimates based on the geometry of Γ for the sizes of regions in $U(\Gamma)$ that do not contain invariant circles.

Theorem 4.3: *Let* Γ *be a convex curve in the plane with Lipschitz arclength function* $s(\theta)$, *and* $J \in S^1$ *is an open interval. If*

(a) the arclength function $s(\theta)$ *is* C^1 *on* J, $\rho > 0$ *at the endpoints of* J, *and for some* $\hat\theta \in J$, $\rho(\hat\theta) = 0$,

or

(b) there is a $\hat\theta \in J$ *so that* s *is* C^1 *on* $J - \hat\theta$, $\rho > 0$ *on* J, *and* ρ *has a jump discontinuity at* $\hat\theta$,

then there exists a neighborhood of Γ *that contains no homotopically nontrivial* Φ_Γ*-invariant circles.*

Proof: The proof will use the half-cylinder twist map f associated with Γ that has lift $\tilde f$. Let h, b_0 and $\mathcal{D}$ be the generating function, lower boundary map and the domain of h, respectively, and $\hat x$ is a lift of $\hat\theta$.

If $\tilde\Omega \subset \tilde C$ is the lift of a homotopically nontrivial f-invariant circle, then by Theorem 1.1(b), $\tilde\Omega$ is the graph of a Lipschitz function $u : \mathbb{R} \to (0, \infty)$. Using the derivative formulas (3.6) and Theorem 1.1(d), for almost every x_0,

$$2\rho(x_0) > \cot(x_1 - x_0)R(x_0, x_1) + \cot(x_0 - x_{-1})L(x_{-1}, x_0) \tag{4.1}$$

where $(x_i, \gamma_i) = f^i(x_0, u(x_0))$.

Assume first that the situation in (a) happens. If there were a Φ-invariant circles arbitrarily close to to Γ, then by Theorem 1.1(b), we could find the lift of an invariant circle $\tilde{\Omega}$ arbitrarily close to $\mathbb{R}\times\{0\}$ in the Hausdorff topology, in which case $\max\{|x_1-x_0|, |x_0-x_{-1}|\} < \pi/2$ and $\{x_{-1}, x_0, x_1\} \subset J$ for all x_0 sufficiently close to $\hat{x}$. In this case the right hand side of (4.1) is positive, continuous and bounded away from zero while the left hand side is a continuous function that vanishes at $x_0 = \hat{x}$. Thus the equation cannot hold almost everywhere for x_0 in a neighborhood of $\hat{x}$, a contradiction.

Now assume the situation given in (b). Let $\lim_{x\to\hat{x}^\pm} \rho(x) := \rho^\pm$ and assume without loss of generality that $\rho^+ > \rho^-$ and $|J| < \pi/2$. The principle estimate needed is the following: $\forall\epsilon > 0$, $\exists\delta_1 > 0, \forall N > 0$, $\exists\delta_2 = \delta_2(\delta_1, N)$ so that $|x_{-1} - x_0| < \delta_1$, $\delta_1/N < |x_0 - x_1| < \delta_1$ and $0 < \hat{x} - x_0 < \delta_2$ implies

$$\max\{|\bar{L}(x_{-1}, x_0) - \rho^-|, |\bar{R}(x_0, x_1) - \rho^+|, |\rho(x_0) - \rho^-|\} < \epsilon$$

where $\bar{L}(x_{-1}, x_0) = \cot(x_0 - x_{-1})L(x_{-1}, x_0)$ and $\bar{R}(x_0, x_1) = \cot(x_1 - x_0)R(x_0, x_1)$.

To prove this first note that if (x_{-1}, x_0, x_1) is a segment of an allowable configuration for h, $\{x_{-1}, x_0, x_1\} \in J$, and $\hat{x} > x_0$, then

$$\begin{aligned}
|\bar{R}(x_{-1}, x_0) - \rho^+| &= |\cot(x_1 - x_0)\int_{x_0}^{x_1} \Big(\frac{\sin(x_1-\eta)}{\sin(x_1-x_0)}\rho(\eta) - \frac{\tan(x_1-x_0)}{(x_1-x_0)}\rho^+\Big)\, d\eta| \\
&\le \frac{1}{(x_1-x_0)}\int_{x_0}^{x_1} |\rho(\eta) - \rho^+|\, d\eta \\
&\le \frac{1}{(x_1-x_0)}\big((\hat{x} - x_0)\sup\{|\rho(\eta) - \rho^+| : \eta \in [x_0, \hat{x})\} \\
&\qquad + (x_1 - \hat{x})\sup\{|\rho(\eta) - \rho^+| : \eta \in (\hat{x}, x_1]\}\big)
\end{aligned}$$

and

$$\begin{aligned}
|\bar{L}(x_{-1}, x_0) - \rho^-| &\le \frac{1}{(x_0 - x_{-1})}\int_{x_{-1}}^{x_0} |(\rho(\eta) - \rho^-)|\, d\eta \\
&\le \sup\{|\rho(\eta) - \rho^-| : \eta \in [x_{-1}, x_0]\}.
\end{aligned}$$

Now given $\epsilon > 0$, since ρ is continuous and bounded on $J - \hat{x}$ there is an $\delta_1 > 0$ so that $0 < x_1 - \hat{x} < \delta_1$ implies

$$\sup\{|\rho(\eta) - \rho^+| : \eta \in (\hat{x}, x_1]\} < \epsilon/2$$

and in addition, $0 < \hat{x} - x_{-1} < 2\delta_1$ implies

$$\sup\{|\rho(\eta) - \rho^-| : \eta \in (x_{-1}, x_0)\} < \epsilon.$$

Now given N, pick $\delta_2 < \delta_1$ so that

$$\frac{N\delta_2}{\delta_1}\sup\{|\rho(\eta) - \rho^+| : \eta \in [x_0, \hat{x})\} < \epsilon/2.$$

With these choices, note that $\hat{x} - x_{-1} = x_0 - x_{-1} + \hat{x} - x_0 < \delta_1 + \delta_2 < 2\delta_1$, and thus principle estimate follows.

Continuing the proof of the lemma, pick $\epsilon > 0$ so that $\rho^+ - \rho^- > 4\epsilon$ and let δ_1 be as in the principle estimate. If there are f-invariant circles arbitrarily near $S^1 \times \{0\}$, then there is a lift of an invariant circle for which $0 < x_i - x_{i-1} < \delta_1$ for all $x_i \in J$ where $x_i = \pi_1(\tilde{f}^n(z))$ with z an element of the lift of the invariant circle. Fix one such circle Ω. Since it contains no fixed points by Theorem 1.1(b) there is an N with $\delta_1/N < x_i - x_{i-1} < \delta_1$ for all pairs on Ω. Using this N, find δ_2 as in the principle estimate. Then the principle estimate along with (4.1) implies that for almost all x_0, in particular for some x_0 with $0 < \hat{x} - x_0 < \delta_2$, $0 > \bar{L}(x_{-1}, x_0) + \bar{R}(x_0, x_1) + 2\rho(x_0) > \rho^+ - \rho^- - 4\epsilon > 0$, a contradiction. □

Remarks:

4.3.1: Neither the condition of part (a) nor (b) excludes the existence of invariant circles outside neighborhoods of Γ. Indeed, the area envelope construction of §4.2 allows one to construct an invariant circle for dual billiards on a convex curve whose radius of curvature function has any prescribed *local* behavior. In the case of billiards, the analog of part (a) (a point where the curvature vanishes) excludes all invariant circles, while the analog of part (b) does not.

4.3.2: The Birkhoff-Mather stability theorem relates the non-existence of invariant circles with the existence of orbits with certain limit behavior. The theorem states that if Ω_1 and Ω_2 are invariant circles for a twist map, then there are no other invariant circles in the annulus bounded by Ω_1 and Ω_2 if and only if there is a point z whose α- and ω-limit sets are contained in Ω_1 and Ω_2, respectively. Mather's result is proven in [M4] for a specific class of twist maps, but the arguments are quite robust and almost certainly apply to half-cylinder twist maps.

Thus, for example, when there are no Φ_Γ-invariant circles near Γ there is an orbit that converges to Γ in forward time. This orbit is distinguished from the crash orbits of the next section by the fact that it converges to all of Γ, not a single point on Γ. There will also be an orbit that converges to Γ under backward iteration, so it may be viewed as "emerging" from Γ.

The question of stability depends on the behavior at infinity. In the Birkhoff-Mather theorem, one of the invariant circles can be located at infinity. Thus if there are no invariant circles near infinity, there is an orbit that escapes, *i.e.* goes to infinity under forward iteration. As noted in the introduction one of the most important outstanding problems in the theory of dual billiards is the existence of a convex curve whose dual billiards map has such an orbit.

4.3.3: Remark 1.2.1 contains the main geometric idea underlying the proof of Theorem 4.3. The conditions given in (a) or (b) ensure that a vertical tangent tumbles past the vertical in two iterates. To obtain a stronger result on the non-existence of invariant circles, one would need to use more iterates. However, for the half-cylinder twist maps coming from dual billiards, the twist goes to zero at infinity. It thus requires more and more iterates to tumble a tangent as one approaches infinity. This partially explains the difficulty in obtaining a Γ that yields no invariant circles near infinity. As noted in Remark 4.3.1, no local condition on Γ can eliminate these invariant circles.

§5 Crash orbits

A *crash orbit* for dual billiards is an orbit that converges under forward iteration to a point on the convex curve Γ. The first theorem states that there exist convex curves with differentiable radius of curvature functions ρ for which there are crash orbits. However, the second theorem says that if ρ is positive and has bounded derivative, then there do not exist crash orbits. These are analogs of billiards results due to Halpern ([Ha]). The method of proof here is different, although the analog of Hapern's construction would also suffice.

We give a fairly explicit ρ and then use the criterion of Angenent from Theorem 1.2 to get the existence of the crash solutions. The simplest version of the example has a radius of curvature function near zero that looks roughly like $1 + (1/2)x^2 \sin(1/x^2)$. Note that this is a standard example of a function whose derivative exist everywhere, but the derivative is not bounded.

The existence of the crash orbits implies that the dual billiards map has no invariant circles near Γ. However, as in Remark 4.3.1, one can construct examples with crash orbits and invariant circles far away using the area envelope construction.

§5.1 Examples with crash orbits.

Theorem 5.1: *There exists a convex curve Γ with a differentiable radius curvature function such that the corresponding dual billiards map has crash orbits.*

Proof: We use the Angenent criterion of Theorem 1.2 with a modified form of the recursion relation. Define $\bar{\Delta}(x_{-1}, x_0, x_1) = R(x_0, x_1) - L(x_{-1}, x_0)$ where R and L are as in (3.2) and (3.3). It is clear that solutions, subsolutions, etc. of this recursion relation will be the same as those given by the Δ defined in (1.5) with h given by (3.5).

We will construct a subsolution $\underline{c}$ with $c_n \nearrow 0$, and a supersolution $\underline{w}$ with $w_n \nearrow w < \infty$, so that $\underline{c} \leq \underline{w}$. By Theorem 1.2, there exists a solution $\underline{x}$ with $\underline{c} \leq \underline{x} \leq \underline{w}$, and since any solution is of necessity monotone increasing, x_n converges as $n \to \infty$. Now by (1.1), there exists some (x_0, γ_0) with $x_i = \pi_1(\tilde{f}^i(x_0, \gamma_0))$. Then using (3.2) or (3.3), $\gamma_n \to 0$, and so $\tilde{f}^i(x_0, \gamma_0)$ converges to a point on $\mathbb{R} \times \{0\}$, thus the corresponding orbit in $U(\Gamma)$ converges to a point of Γ under the dual billiards map.

Given a sequence a_n (to be specified later) with $n \in \mathbb{N}$ such that $a_n \nearrow 0$ and $a_n \in [-.1, 0)$, let $\delta_n = (a_{n+1} - a_n)/2$ and $b_n = a_n + \delta_n$. Pick a $0 < c < .001$, $k > 1$, and for $n = 1, 2, \ldots$, define $\rho : \mathbb{R} \to \mathbb{R}$ by

$$\rho(x) = \begin{cases} 1 - c a_n^k & \text{for } x \in [b_{n-1}, a_n) \\ 1 + c a_n^k & \text{for } x \in [a_n, b_n) \\ 1 & \text{for } x \in [0, 2\pi - b_0) \end{cases}$$

and extend ρ so that it is 2π-periodic.

Now let

$$r(x) = \frac{\cos(x) - \cos(2x)}{\sin(2x)}$$

and

$$s(x) = \frac{1 - \cos(x)}{\sin(2x)}.$$

Note that at points where the denominator is zero, r and s can be continued to real analytic functions with $s(x) + r(x) = \tan(x)$.

If $r_n = r(\delta_n)$ and $s_n = s(\delta_n)$, a simple calculation yields that

$$\begin{aligned} R_n &:= R(a_n, a_{n+1}) = (1 + ca_n^k)r_n + (1 - ca_{n+1}^k)s_n \\ L_{n-1} &:= L(a_{n-1}, a_n) = (1 + ca_{n-1}^k)s_{n-1} + (1 - ca_n^k)r_{n-1}. \end{aligned}$$

Our goal is to show that $R_n - L_{n-1} \geq 0$ for sufficiently large n. Assuming $\delta_{n-1} - \delta_n > 0$, this is equivalent to

$$\begin{aligned} & c\,(\frac{a_n^k r_n - a_{n+1}^k s_n - a_{n-1}^k s_{n-1} + a_n^k r_{n-1}}{\delta_{n-1} - \delta_n}) \\ & > \frac{s_{n-1} + r_{n-1} - s_n - r_n}{\delta_{n-1} - \delta_n} \\ & = \frac{\tan(\delta_{n-1}) - \tan(\delta_n)}{\delta_{n-1} - \delta_n} \end{aligned} \tag{5.1}$$

which as $n \to \infty$ goes to $\tan'(0) = 1$.

Now specify $a_n = -n^b$ for some $b < 0$. In this case,

$$\begin{aligned} -bn^{b-1} > \delta_n > -b(n+1)^{b-1} \\ b(b-1)(n+1)^{b-2} < \delta_{n-1} - \delta_n < b(b-1)(n-1)^{b-2}. \end{aligned}$$

Using Taylor's theorem, for sufficiently small positive δ, $r(\delta) > 3\delta/4$ and $-s(\delta) > -\delta/2$. Thus the left hand side of (5.1) is larger than

$$c(\frac{\frac{3}{4}n^{kb}((n+1)^{b-1} + n^{b-1}) - \frac{1}{2}(n+1)^{kb}n^{b-1} - \frac{1}{2}(n-1)^{(k+1)b-1}}{(1-b)(n-1)^{b-2}}),$$

which goes to infinity if $kb + 1 > 0$. Thus for these choices of k and b, $R_n > L_{n-1}$ for $n \geq N$ for some sufficiently large N.

Now pick $\hat{\rho}$ that is C^∞ except at 0 and is close enough to ρ to ensure that $\underline{a}$ is still a subsolution (for $n \geq N$) for the recursion relation defined using $\hat{\rho}$.

To construct the subsolution $\underline{c}$, let $\gamma_N = h_2(a_{N-1}, a_N)$ and for $i \leq 0$, $c_{N+i} = \tilde{f}^i(a_N, \gamma_N)$ and for $i \geq 0$, $c_{N+i} = a_{N+i}$. By construction, $\Delta(c_{i-1}, c_i, c_{i+1}) = 0$ if $i \leq N-1$ and $\Delta(c_{i-1}, c_i, c_{i+1}) \geq 0$ if $i \geq N$, and thus $\underline{c}$ is a subsolution.

To construct a supersolution $\underline{w}$ with $\underline{w} > \underline{c}$ begin by letting $v_i = \pi_1(\tilde{f}^i(a_N, \gamma_N))$ where (a_N, γ_N) is defined in the previous paragraph. Now if the sequence v_i is bounded above it is a crash orbit and we are done, otherwise let $i_0 = \inf\{i : v_i > 0\}$ and $w_{N-j} = v_{i_0 - j}$ for $j \geq 0$. Define w_{N+j} for $j \geq 0$ inductively as follows. Assuming w_{n-1} and w_n have been obtained, pick $w_n < w_{n+1} < w_N + \sum_{i=0}^{n+1-N} \frac{1}{2^{i+3}}$ so that $0 < \hat{R}(w_n, w_{n+1}) < \hat{L}(w_{n-1}, w_n)$, where $\hat{L}$ and $\hat{R}$ are defined by (3.2) and (3.3) using $\hat{\rho}$ for the radius of curvature. This is possible because $\hat{R}(x, x') \to 0$ as $|x' - x| \to 0$.

Now note that $\Delta(w_{k-1}, w_k, w_{k+1})$ is less than zero for $k \geq N$ and equal to zero for $k < N$, and thus $\underline{w}$ is a subsolution. By construction, $\underline{c} \leq \underline{w}$ and $\lim w_n \leq w_N + 1/4$. □

§**5.2 Nonexistence of crash solutions.**

Theorem 5.2: *If* Γ *is a convex curve in the plane with a radius of curvature function* $\rho > 0$ *that is differentiable and the derivative is bounded, then the dual billiards map has no crash orbits.*

Proof: Assume to the contrary that the dual billiards map has a crash orbit. Lift the corresponding half-cylinder twist map to the universal cover and say the crash orbit is (x_n, γ_n), where we assume without loss of generality that $x_n \nearrow 0$.

Using the recursion relation given by (3.4) and the mean value theorem for integrals we can find $\eta_{n-1} \in (x_{n-1}, x_n)$ and $\tau_n \in (x_n, x_{n+1})$ with

$$\rho(\eta_{n-1}) \tan(\delta_{n-1}/2) = \rho(\tau_n) \tan(\delta_n/2)$$

where $\delta_n = x_{n+1} - x_n$, and thus

$$(\rho(\eta_{n-1}) - \rho(\tau_n)) \tan(\delta_{n-1}/2) = \rho(\tau_n)(\tan(\delta_n/2) - \tan(\delta_{n-1}/2)).$$

Now by construction, $\tau_n - \eta_{n-1} < \delta_n + \delta_{n-1}$ and if $\delta > 0$ is small, $\delta < \tan(\delta) < 2\delta$. Using the mean value theorem for derivatives, there is a $\sigma_n \in [\eta_{n-1}, \tau_n]$ with

$$\begin{aligned} |\rho'(\sigma_n)| &= |\frac{\rho(\tau_n) - \rho(\eta_{n-1})}{\tau_n - \eta_{n-1}}| \\ &= |\frac{\rho(\tau_n)(\tan(\delta_n/2) - \tan(\delta_{n-1}/2))}{(\tau_n - \eta_{n-1}) \tan(\delta_{n-1}/2)}| \\ &\geq |\frac{\rho(\tau_n)(\delta_n/2 - \delta_{n-1})}{(\delta_n + \delta_{n-1})\ \delta_{n-1}}| \\ &= |\frac{\rho(\tau_n)}{2}(\frac{\frac{\delta_n}{\delta_{n-1}} - 2}{\delta_{n-1} + \delta_n})|. \end{aligned} \tag{5.2}$$

Now since ρ is continuous, as $n \to \infty$, $\rho(\tau_n) \to \rho(0) > 0$; since $\sum \delta_n$ converges, $\delta_n \to 0$ and $\limsup |\delta_n / \delta_{n-1}| \leq 1$. Thus the last expression in (5.2) diverges, contradicting the boundedness of ρ'. $\square$

§**6 The normalized action of the derivative**

In this section we make some geometric observations about the action of the derivative of the dual billiards map on tangent directions. Since our aims are primarily descriptive, we will adopt a more informal tone than in previous sections.

§**6.1 Computation of the action of the derivative.** Let $\beta(t) = (\beta_1(t), \beta_2(t))$ be a parameterization of Γ *by arc length*, and let us assume initially that β is twice differentiable.

Fix a point $z \in u(\Gamma)$ and say its envelope coordinates are (θ, γ). Let t and l be such that $\beta(t) = \alpha(\theta)$ and $\ell = \sqrt{2\gamma}$. Thus $\beta(t)$ is the point on Γ used for the dual billiards shot, and ℓ the distance from z to $\beta(t)$. From inspection of Figure 6.1 we have

$$2\beta(t) = z + \Phi(z) \tag{6.1}$$

$$z = \beta(t) - \ell\beta'(t). \tag{6.2}$$

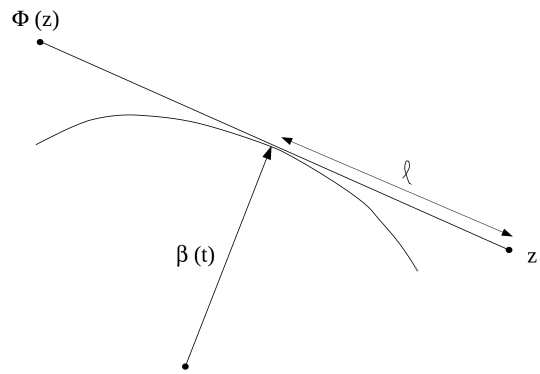

Figure 6.1: Some geometry of dual billiards.

Adopting the classical (and often confusing) notation of using the same symbol for a coordinate and a coordinate change, $z = (x, y) = (x(t, \ell), y(t, \ell))$, and subscripts to denote differentiation, (6.2) yields

$$M := \begin{pmatrix} x_t & x_l \\ y_t & y_l \end{pmatrix} = \begin{pmatrix} \beta_1' - \ell\beta_1'' & \beta_1' \\ \beta_2' - \ell\beta_2'' & \beta_2' \end{pmatrix}.$$

Note that $det(M) = \ell(\beta' \wedge \beta'') = \ell\kappa(t)$, where $\kappa(t)$ is the curvature at the point $\beta(t)$ (the curvature is the reciprocal of the radius of curvature ρ). Thus inverting M we get

$$\begin{pmatrix} t_x & t_y \\ \ell_x & \ell_y \end{pmatrix} = \frac{1}{\kappa\ell}\begin{pmatrix} \beta_2' & -\beta_1' \\ -\beta_2 + \ell\beta_2'' & \beta_1 - \ell\beta_1'' \end{pmatrix}. \tag{6.4}$$

Differentiating (6.1) yields

$$\begin{aligned} D\Phi + Id &= 2\begin{pmatrix} \beta_1' & 0 \\ \beta_2' & 0 \end{pmatrix}\begin{pmatrix} t_x & t_y \\ \ell_x & \ell_y \end{pmatrix} \\ &= \frac{2}{\kappa\ell}\begin{pmatrix} \beta_1'\beta_2' & -(\beta_1')^2 \\ (\beta_2')^2 & -\beta_1'\beta_2', \end{pmatrix} \end{aligned}$$

and thus acting on a vector $\mathbf{u}$,

$$(D\Phi + Id)[\mathbf{u}] = \frac{2}{\kappa\ell}(\mathbf{u} \wedge \beta')\,\beta'. \tag{6.5}$$

We now need some elementary geometry of the wedge product. Let $\mathbf{w}$ be a unit vector, $\mathbf{u}$ and $\mathbf{v}$ be vectors, and a, b, c be positive scalars so that $b\mathbf{w} = a\mathbf{v} + c\mathbf{u}$. If A represents the area of the triangle formed by the vectors then,

$$2A = c\mathbf{u} \wedge b\mathbf{w} = b\mathbf{w} \wedge a\mathbf{v} = c\mathbf{u} \wedge a\mathbf{v}$$

and so

$$2A = \frac{b^2(\mathbf{w} \wedge \mathbf{v})(\mathbf{w} \wedge \mathbf{u})}{\mathbf{v} \wedge \mathbf{u}}. \tag{6.6}$$

Now wedging (6.5) by $D\Phi[u]$ on the left yields

$$D\Phi[\mathbf{u}] \wedge \mathbf{u} = \frac{2}{\kappa\ell}(\mathbf{u} \wedge \beta')(D\Phi[\mathbf{u}] \wedge \beta') \tag{6.7}$$

Thus if B is the area of the shaded region in in Figure 6.2, *i.e.* B is the area of the triangle bounded by the line connecting z to $\Phi(z)$, the line in the direction $\mathbf{u}$ passing through the point z, and the line in the direction of $D\Phi[\mathbf{u}]$, then using (6.6) and (6.7),

$$B = \frac{(2\ell)^2}{2} \frac{(\beta' \wedge D\Phi[\mathbf{u}])(\beta' \wedge \mathbf{u})}{D\Phi[u] \wedge \mathbf{u}} = \kappa\ell^3. \tag{6.8}$$

Equation (6.8) contains the same information as Proposition 2.2 in [GK].

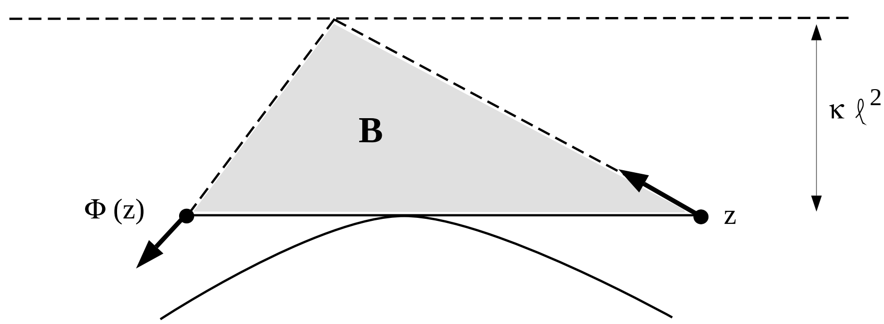

Figure 6.2: Geometry of the derivative.

§6.2 Geometric computation of the derivative. To geometrically compute the action of the derivative of the dual billiards map on a tangent direction, first fix a point z and its image $\Phi(z)$ as in Figure 6.2. Now draw a line parallel to that from z to $\Phi(z)$ that is a distance $\kappa\ell^2$ from that line, where κ is the curvature at the point of tangency. Call this new line the ***bounce line***. Given a unit tangent vector $\mathbf{v}$ based at z, draw a line in the direction of $\mathbf{v}$ until it hits the bounce line. The line from this point of intersection to the point $\Phi(z)$ will give the direction of the vector $D\Phi[\mathbf{v}]$. This is because the shaded region in Figure 6.2 has area equal to $\kappa\ell^3$ as required by (6.8).

If Γ has a point with $\rho(\theta) = 0$, by taking limits one sees that the action on tangent directions is a bounce off infinity. Thus when the dual billiards shot uses a corner of Γ, the action of $D\Phi$ is rigid rotation by π.

Remark 6.2.1: One motivation for seeking a geometric understanding the action of the derivative is to assist in the finding of strictly invariant cone fields and thus (perhaps) obtain positive Lyapunov exponents as in [W]. This theory does not seem to be well understood in the case of infinite area as is the case in dual billiards. One could also search for these cone fields in bounded regions, *eg.* between a given curve and a second curve inside constructed using the area envelope construction. The existence of positive exponents is particularly interesting in light of the physical interpretation of dual billiards given in §7.

§6.3 Comparison with a hyperbola. Having understood geometrically the action of the derivative we are now in a position to interpret the tumbling tangent criterion of Remark

1.2.1. Again, pick a point z. Figure 6.3(a) shows the action of the second iterate of a dual billiards map on a tangent direction based at $\Phi^{-1}(z)$. There are three supporting lines involved in the iteration, call them $\mathcal{L}_{-1}, \mathcal{L}_0$, and $\mathcal{L}_1$. The half-ray in $\mathcal{L}_{-1}$ that begins at Γ and points downward corresponds to a vertical line in envelope coordinates. Thus the initial vector chosen corresponds to the initial vertical tangent in Figure 1.2. Its first iterate under $D\Phi$ yields a parallel vector pointed in the opposite direction. For the next iterate we use the procedure of the previous subsection. Draw a bounce line at height $\kappa\ell^2$ and compute the next image. The condition that this vector has tumbled beyond the vertical is that the image is on the side of $\mathcal{L}_1$ nearest Γ. In the figure the initial vector has tumbled by the third iterate.

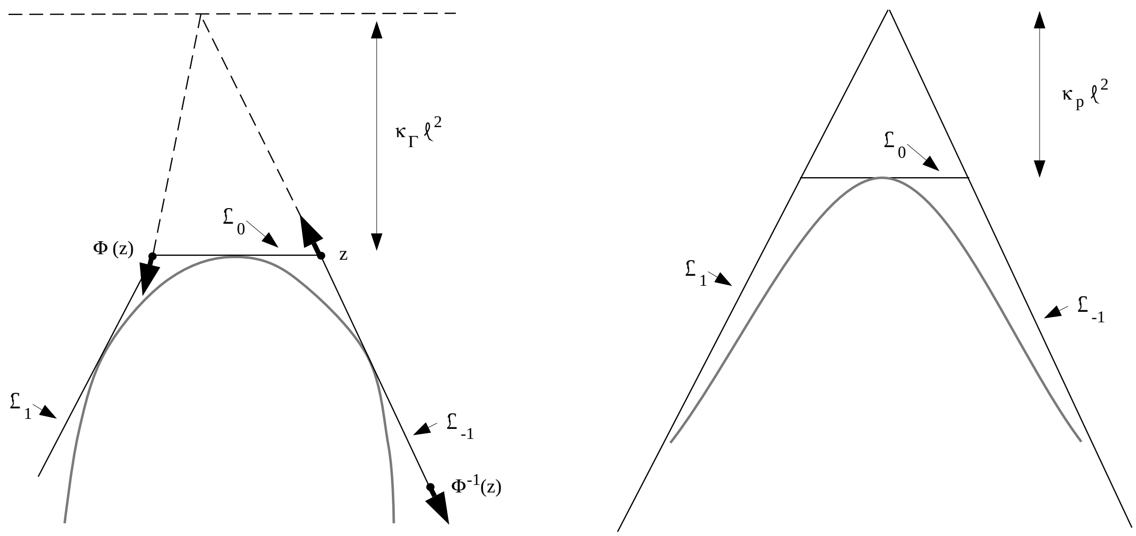

Figure 6.3: (a) Tumbling tangents in dual billiards; (b) The comparison hyperbola.

It is geometrically clear now that when the curvature is infinite at the point of tangency of $\mathcal{L}_0$ (*i.e.* the radius of curvature vanishes) one always gets tumbling tangents for points z near Γ, and thus no invariant circles near Γ. It is also clear why the same argument does not work for z that are far away from Γ.

To get a more refined criterion, construct a hyperbola as the area envelope inside the pair of lines $\mathcal{L}_{-1}$ and $\mathcal{L}_1$ using an area equal to that of the triangle bounded by the $\mathcal{L}_{-1}, \mathcal{L}_0$, and $\mathcal{L}_1$ (see Figure 6.3(b)). This hyperbola will be tangent to $\mathcal{L}_0$ at the same point that Γ is. Now if Ψ is the dual billiards map of this hyperbola, the image of tangent vector based at z that points in the direction of $\mathcal{L}_{-1}$ will be a vector based at $\Phi(z)$ that points in the direction of $\mathcal{L}_1$. Thus by §6.1, the point $\mathcal{L}_{-1} \cap \mathcal{L}_1$ is a distance $\kappa_p \ell^2$ from $\mathcal{L}_0$ where κ_p is the curvature of the hyperbola at the point of tangency. Recalling the geometric condition for a tumbling tangent for Φ one sees that the tangent tumbles if and only if the bounce line for Γ is higher than the bounce line for the hyperbola, *i.e.* if $\kappa_\Gamma > \kappa_p$. A simple calculation shows that this condition is identical to (4.1).

§7 Dual billiards and impact oscillators

In this section we show that the dynamics of a dual billiards map associated with a convex curve is identical to the dynamics of a certain impact oscillator. As in the previous section, our purpose is primarily descriptive so we maintain a somewhat informal tone.

§7.1 The impact oscillator. The first ingredient in the impact oscillator is a periodically moving wall. Its position at time t is given by $p(t)$, and we require that $p(t+2\pi) = p(t)$, p is continuous, and p is sub-sine as defined in Remark 2.1.1. As noted there, if p is twice differentiable this happens if and only if

$$\ddot{p} + p := \rho \geq 0. \tag{7.1}$$

Alternatively, we can start with a periodic ρ with $\int \rho(t)\exp(it)\,dt = 0$, pick initial conditions for $p(0)$ and $\dot{p}(0)$, and then solve (7.1) for p.

The second ingredient in the impact oscillator is a particle whose position is given by $x(t)$. Except at collision this particle is a simple harmonic oscillator with period 2π, *i.e.* it satisfies

$$\ddot{x} + x = 0. \tag{7.2}$$

The collisions of the particle with the wall are assumed to be perfectly elastic and the wall has infinite mass; so if t_c is a time when $x(t_c) = p(t_c)$, then

$$-(\dot{x}_{before}(t_c) - \dot{p}(t_c)) = \dot{x}_{after}(t_c) - \dot{p}(t_c). \tag{7.3}$$

A ***solution*** of the impact oscillator is a function $x(t) \geq p(t)$ that satisfies (7.2) except when $x(t) = p(t)$ (*i.e.* collisions) at which time one starts a new solution to (7.2) using the initial conditions dictated by (7.3). We also require that $\dot{p} > \dot{x}$ at collision. Thus if ρ vanishes on a nontrivial interval J, we do not consider an $x(t)$ with $x(t) = p(t)$ for $t \in J$ a solution. Note that since p is sub-sine, if $\dot{p} > \dot{x}$ at one collision the same inequality holds at the next collision, *i.e.* the situation in Figure 7.1 *does not* occur. From a more physical point of view, the fact that p is sub-sine implies that there are no so-called graze solutions ([BD]). In these solutions the particle approaches the wall, touches it for an instance matching the wall's position and velocity, and then pulls away without a collision. The absence of these solutions implies that the return time to the wall is a continuous function of $\dot{x}(t_c)$ and t_c.

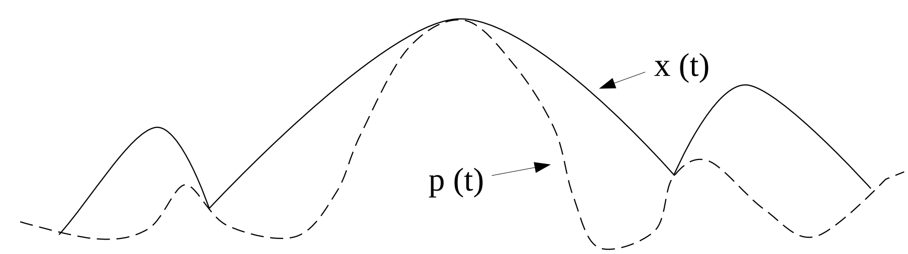

Figure 7.1: Solution ruled out by sub-sine hypothesis.

§7.2 Pictorial connection of the two systems. Fix a convex curve Γ in the plane with radius of curvature function ρ, and assume for simplicity that the origin is interior to Γ.

Call the orthogonal projection of Γ onto the x-axis I_0 (see Figure 7.2(a)). The distance of the origin to the right endpoint of I_0 is $p(0)$ as defined in §2.1. Now imagine rotating the plane clockwise about the origin while keeping the direction of the x-axis fixed. If I_θ is the projection of Γ onto the x-axis after we have rotated by an angle θ, then the distance from the origin to the right endpoint of I_θ is $p(\theta)$. We will think of the segments I_θ as the shadow of Γ and we will rotate at unit speed and identify the angle θ with the time t. It follows from (2.3) that the right endpoint of the shadow is evolving according to $\ddot{p} + p = \rho$. Thus the end of the shadow is moving like the wall in the impact oscillator.

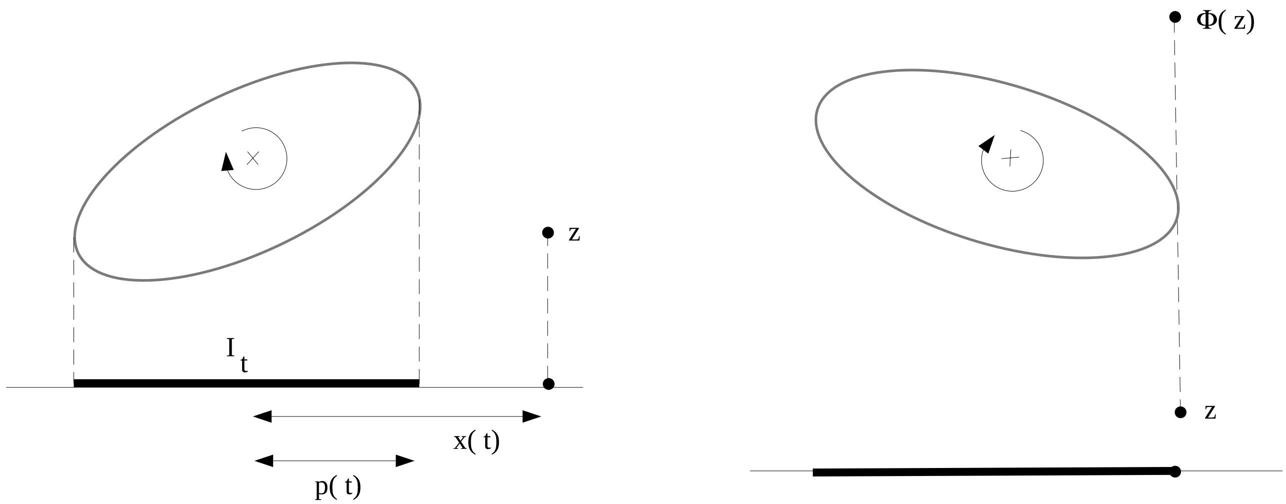

Figure 7.2: (a)Pre-collision in the rotating plane model; (b)Collisions in the rotating plane model.

To introduce the particle, pick a point z_0 exterior to Γ and monitor the evolution of its shadow (its orthogonal projection onto the x-axis) as we rotate the plane. The point moves in a circle and so if its shadow is given by $x(t)$, then $\ddot{x} + x = 0$.

Now for the collisions. Assume that at time t, $x(t) > p(t)$ as in Figure 7.2(a). As we continue rotating, at some time t_c, $x(t_c)$ will equal $p(t_c)$, *i.e.* the shadow of the point and that of Γ will collide. This will be when the supporting line to Γ that contains z_0 is vertical. Since the shadows of z_0 and Γ are taking the role of the particle and the wall in the impact oscillator, when the shadows collide we should act on the point z_0 in such a way that its shadow rebounds from the shadow of Γ with the appropriate velocity. It is clear from Figure 7.2(b), that the correct action is to reflect z_0 around the point of tangency to the Γ to obtain a new point $\Phi(z_0)$. As the plane continues to rotate, the shadow of $\Phi(z_0)$ will evolve away from the shadow of Γ as if it had a perfectly elastic collision. The map Φ just defined is clearly identical to the dual billiards map associated with Γ.

§7.3 Connecting the generating functions. To get a more precise connection between dual billiards and the impact oscillator we connect the generating function of the half-cylinder twist map associated with dual billiards with an action integral of an impact oscillator equivalent to that of §7.1.

First we derive a new formula for the generating function of dual billiards. If $B(\theta_1, \theta_2)$ is the area of the sector of Γ between the θ_1 and θ_2 (see Figure 7.3), then using Stokes theorem

$$\begin{aligned} B(\theta_1, \theta_2) &= \frac{1}{2}\int_{\theta_1}^{\theta_2} \alpha(\theta) \wedge \mathbf{n}_\theta \; ds(\theta) \\ &= \frac{1}{2}\int_{\theta_1}^{\theta_2} p(\theta)\rho(\theta) \; d\theta. \end{aligned}$$

Recalling that $\rho = p + p''$, integrating by parts yields

$$B(\theta_1, \theta_2) = \frac{1}{2}\left(\int_{\theta_1}^{\theta_2} p^2 - (p')^2 \; d\theta + p(\theta_2)p'(\theta_2) - p(\theta_1)p'(\theta_1)\right). \tag{7.4}$$

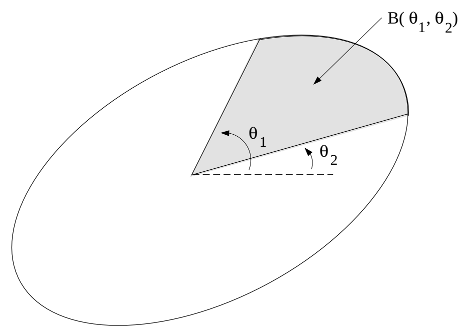

Figure 7.3: The region $B(\theta_1, \theta_2)$.

We now move the origin to the point $Q = \mathcal{L}_{\theta_1} \cap \mathcal{L}_{\theta_2}$ and let $\eta(\theta)$ be the distance from Q to $\mathcal{L}_\theta$. With this new origin the area computed as $B(\theta_1, \theta_2)$ is the generating function of the half-cylinder twist map associated with the dual billiards map (see Remark 3.2.1). In a slight abuse of notation we will treat this as a function of θ rather than x and denote it $h_\Gamma(\theta_1, \theta_2)$. To use (7.4) we note that from the point of view of the point Q, the radius of curvature of Γ is $-\rho(\theta)$, and so using the fact that $\eta(\theta_1) = \eta(\theta_2) = 0$, we get

$$h_\Gamma(\theta_1, \theta_2) = \frac{1}{2}\int_{\theta_1}^{\theta_2} (\eta')^2 - \eta^2 \; d\theta. \tag{7.5}$$

where the function $\eta(\theta)$ is the (unique!) solution to

$$\eta'' + \eta = -\rho$$

with initial conditions $\eta(\theta_1) = \eta(\theta_2) = 0$. For future reference note that from the geometric interpretation of the derivative of the height function (see §2.1), $\eta'(\theta_1) = R(\theta_1, \theta_2)$ and $\eta'(\theta_2) = -L(\theta_1, \theta_2)$.

To connect this action to the impact oscillator above we define for that system a new variable $\eta(t) = x(t) - p(t)$; thus η measures the distance between the particle and the wall. The collision rule is now

$$-\dot{\eta}_{before}(t_c) = \dot{\eta}_{after}(t_c), \tag{7.6}$$

and motion away from collision is governed by

$$\ddot{\eta} + \eta = -\rho. \tag{7.7}$$

It will be useful to think of (7.6) and (7.7) as defining a new impact oscillator that is dynamically equivalent to the original one. In this new system, η is the position of a particle governed by (7.7) that has perfectly elastic collisions with a stationary wall.

Now between impacts this system has Lagrangian $L(x, \dot{x}, t) := \dot{x}^2/2 - x^2/2 - x\rho(t)$ and Hamiltonian $H(q, p, t) := q^2/2 + p^2/2 + p\rho(t)$. If $t_1 < t_2 < t_1 + \pi$, we define

$$h_I(t_1, t_2) = \min \int_{t_1}^{t_2} L(x, \dot{x}, t)\, dt$$

where the minimum is taken over all paths $x(t)$ with $x(t_1) = 0$ and $x(t_2) = 0$. Thus

$$h_I(t_1, t_2) = \int_{t_1}^{t_2} L(\eta, \dot{\eta}, t)$$

where $\eta = \eta(t; t_1, t_2)$ is the (unique!) solution to $\ddot{\eta} + \eta = -\rho$ with boundary values $\eta(t_1) = 0 = \eta(t_2)$.

Now note that h_I is the Hamilton principal function (see Remark 7.3.1) with the starting and ending configuration variables fixed at zero, thus

$$\begin{aligned} \partial_1 h_I(t_1, t_2) &= H(\eta(t_1), \dot{\eta}(t_1), t_1) = \dot{\eta}(t_1)^2 \\ \partial_2 h_I(t_1, t_2) &= -H(\eta(t_2), \dot{\eta}(t_2), t_2) = -\dot{\eta}(t_2)^2. \end{aligned} \tag{7.8}$$

Recalling the collision rule (7.6), we see that a sequence of times (t_i) will be the sequence of a collision times of a solution to the oscillator precisely when it is gives stationary configuration as in §1.2 for the variational problem defined using h_I. But exactly as in (7.4) one has

$$\int_{t_1}^{t_2} \eta\rho = \int_{t_1}^{t_2} \eta^2 - \dot{\eta}^2,$$

and so $h_\Gamma(t_1, t_2) = -h_I(t_1, t_2)$.

In conclusion, given an impact oscillator described by (7.6) and (7.7), define Φ_I as the return map to collisions, *i.e.* $\Phi_I : (t_0, (\dot{\eta}(t_0))^2) \mapsto (t_1, (\eta'(t_1))^2)$ where t_0 and t_1 are the times at consecutive collisions; we have shown that Φ_I is identical to the half-cylinder twist map map associated with dual billiards on a convex curve with radius of curvature function ρ.

Remarks:

7.3.1: If H is a Hamiltonian that satisfies the Legendre condition, then Hamilton's principal function is

$$S(q_0, q_1, t_0, t_1) = \int_{t_0}^{t_1} L(x, \dot{x}, t)$$

where L is the corresponding Lagrangian and $x(t)$ is a solution to Hamilton's equations (or equivalently an extremal of the integral) with $x(t_0) = q_0$ and $x(t_1) = q_1$ (see *eg.* chapter 9 in [Gs]). For S to be well-defined this solution needs to be unique. The differential of the principal function is $dS = p\ dq - H\ dt$. What was defined as h_I above is the restriction of S to $0 = q_0 = q_1$. In [M5] Moser uses a similar construction, but restricts S to $t_0 = 0$ and $t_1 = 1$ to obtain the generating function of the time-one map of the solution flow to a time-periodic Hamiltonian. He shows that any twist map of the compact annulus can be thus obtained. This raises the question of which twist maps of the half-cylinder can arise from impact oscillators.

7.3.2: One can also find solutions to the impact oscillator by considering broken paths consisting of a collection of functions $x_i(t)$ and times t_i so that $x_i(t_i) = 0 = x_i(t_{i+1})$ and $x_i > 0$, otherwise. Solutions are then extremals for $\sum L(x_i, \dot{x}_i, t)$.

In this context introducing the generating function h_I is analogous to the strategy of broken geodesics in the variational theory of geodesics (see page 330 of [Gl] for some history). For extremals of the fixed endpoint problem the introduction of h_I replaces an infinite dimensional problem with a finite dimensional one. Crucial to the application of this method is the fact that that there is a unique extremal from $x = 0$ to $x = 0$ which begins at a time t_0 and ends at t_1, where $t_0 < t_1 < t_0 + \pi$.

7.3.3: Yet another way of connecting the dual billiards map to an impact oscillator can be obtained through a kind of regularization of the collisions with the wall. Solutions to the equation

$$\ddot{\eta} + \eta = -\mathrm{sgn}(x)\rho(t) \tag{7.9}$$

behave as if they pass through the wall instead of colliding. Upon passage through the wall the sign of the forcing is changed. Solutions of (7.9) for $x < 0$ are the reflections through the origin of those for $x > 0$. We can use (7.9) to generate a flow on $\mathbb{R}^2 \times S^1$ via

$$\begin{aligned} \dot{x} &= y \\ \dot{y} &= -x - \mathrm{sgn}(x)\rho(\tau) \\ \dot{\tau} &= 1 \end{aligned}$$

The return map of the flow to the open annulur cross section defined by $x = 0$, $y > 0$ will be the second iterate dual billiards map on the Γ with radius of curvature ρ. But note that the map will be in (θ, ℓ) coordinates and will thus not be area-preserving.

The correct choice of second coordinate (*i.e.* the one that makes the map area-preserving) can be seen to be dictated by (7.8) and Hamilton-Jacobi theory as in Remark 7.3.1. From yet another point of view we can make the $1\frac{1}{2}$-degree of freedom system given by (7.9) into an autonomous 2-degree of freedom solution in the usual way by letting $H^*(q, p, E, \tau) = H(q, p, \tau) - E$, where now $H(q, p, t) = p^2/2 + q^2/2 + |q|\rho(t)$. Letting the τ variable be on a circle, and restricting attention to $H^* = 0$, the return map in (τ, E) coordinates to $q = 0, p > 0$ will be the second iterate of dual billiards map. Note that this map is area preserving because the flow preserves volume and since $q = 0$ on the chosen cross section, the E coordinate there is just $p^2/2$.

7.3.4: In [Kl], [SV] and [GS], it is shown that dual billiards on rational polygons are stable, *i.e.* there do not exist any unbounded orbits. Recalling §2.1, this implies that a class of impact oscillators as defined by (7.6) and (7.7) whose forcing functions are contained in a certain collection of linear combinations of delta functions have no unbounded solutions.

7.3.5: The physical content of the fact that the return map to the wall using the $(t, \dot{\eta}^2)$ coordinates is a twist map is simply that when two points are started at the same time at the wall, the one with greatest initial velocity will take the longest to return to the wall.

Appendix

The appendix is a remark on invariant circles for (inner) billiards. As is the case with dual billiards (§4.2), the solution to the inverse problem for the existence of invariant circles is a special case of Proposition 1.3. The remark rests on a elementary computation, but it does not seem to be widely known. Billiards is the subject of a large body of work (*cf.* [T3]); we only briefly describe the system here. Billiards as a twist map is considered in [M1], [D] and [Me2].

If Γ is an oriented convex curve in the plane, the billiards map is defined using a point particle moving freely in the region bounded by Γ. The particle has perfectly elastic collisions with the curve, so the angle of incoming motion with the tangent to the curve is equal to the angle of the outgoing direction. The billiards map just keeps track of collisions with the curve. The map can be described using coordinates t, the arc length parameter along the curve, and ω, the angle between an outgoing line of motion and the tangent to the curve. The billiards map is then $(t_0, \omega_0) \mapsto (t_1, \omega_1)$, where (t_0, ω_0) represents an initial point on the curve and an outgoing direction, and (t_1, ω_1) represents the point on the curve at the next collision and its outgoing direction after the collision (see Figure A.1). This defines a homeomorphism of the annulus $S^1 \times [0, \pi]$.

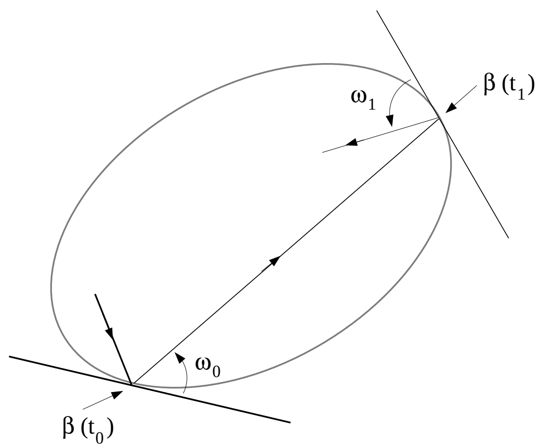

Figure A.1: The billiards system.

An invariant circle of this annulus map corresponds to a geometric object called a caustic. For simplicity we restrict to the case of convex caustics. A convex caustic is a convex curve inside of Γ with the property that an outgoing path that is tangent to Γ_1 on one pass must also be tangent on passes after all future (and past) collisions. The inverse problem for the existence of caustics is the following: given a convex curve Γ_1, does there exist another convex curve Γ so that Γ_1 is a caustic for billiards on Γ? The solution to this problem involves the string body construction. Fix the curve Γ_1 and a length of string that is longer than the perimeter of Γ_1. Loop the string around Γ_1 and put a pencil in the loop, pull the string tight and move the pencil around Γ_1 keeping the string tight. The pencil will draw out a curve Γ that is characterized by the constancy of the length $B + C + D$ as shown in

Figure A.2. It is clear that Γ will have Γ_1 as a caustic. The converse is also true; if Γ has a caustic Γ_1, then Γ may be obtained as a string body of Γ_1. Put another way, given any Γ_2 inside Γ we can define a quantity called the ***string parameter*** $L(t) = B(t) + C(t) + D(t)$ where the quantities B, C, and D may vary with the point $\beta(t)$. The curve Γ_2 is a caustic if and only if $L(t)$ is constant. This result is attributed to Minasian in [S]; it also appears in [P] (see also [Tr]).

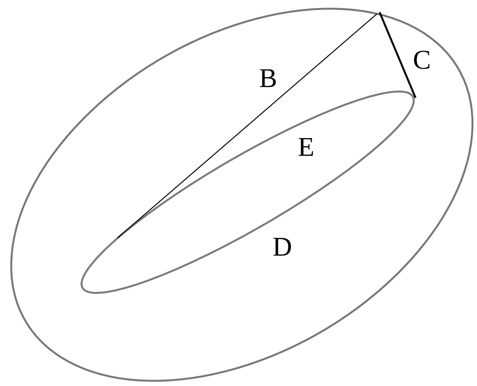

Figure A.2: The string parameter.

With the proper choice of coordinates, the annulus homeomorphisms arising from billiards on convex curves gives a class of area preserving twist maps that have a theory similar to that described in §1. In particular, there is a generating function for the homeomorphism, and one can find orbits using variational techniques. The analog of Theorem 1.1 is true as well as much else. Our purpose here is just to remark on one aspect of the theory, namely, the connection of the area function $A(t)$ defined in §1.4 with the string body construction.

To describe the billiards map more carefully, begin by letting $\beta(t) = (\beta_1(t), \beta_2(t))$ be a parameterization of convex curve Γ *by arc length*. For simplicity, assume that Γ has perimeter 2π, and so $t \in S^1$. The generating function for the billiards map is the opposite of the length of the chord between impacts with the curve,

$$h(t_0, t_1) = -\|\beta(t_1) - \beta(t_0)\|.$$

Thus, for example, one can obtain period three orbits by finding a triangle with edges on Γ that has *maximal* perimeter among all such triangles.

In accord with (1.1) we differentiate h to get the appropriate second coordinates to make billiards an area-preserving twist map,

$$h_1(t_0, t_1) = \frac{\beta(t_1) - \beta(t_0)}{\|\beta(t_1) - \beta(t_0)\|} \cdot \beta'(t_0) = \cos(\omega(t_0, t_1))$$

where $\omega(t_0, t_1)$ is the angle labeled ω_0 in Figure A.1. Thus the appropriate second coordinate is the opposite of the cosine of the angle between the outgoing ray and the tangent line to the curve. Billiards defines a twist map because $h_{12} < 0$. The billiards on a convex curve Γ then defines an area-preserving twist map $\psi_\Gamma : S^1 \times [-1, 1] \to S^1 \times [-1, 1]$.

Now assume that Γ has a caustic, *i.e.* there exists a ψ_Γ-invariant circle. By the appropriate version of Birkhoff's Theorem (Theorem 1.1(b), see [M1]), an invariant circle is the graph of a Lipschitz function $u : S^1 \to [-1,1]$. The function whose graph is the lift of the invariant circle to $\mathbb{R} \times [-1,1]$ will also be denoted u. The coordinates on both S^1 and $\mathbb{R}$ will be called t.

Examining the geometric situation one sees that a caustic Γ_1 defines a one-parameter family of angles each associated with a point on Γ. The angle at the point $\beta(t)$ is denoted $\nu(t)$. The point of tangency to Γ_1 of the outgoing ray from $\beta(t)$ is denoted $\sigma(t)$ and the advance map along Γ is $g(t)$ (see Figure A.3).

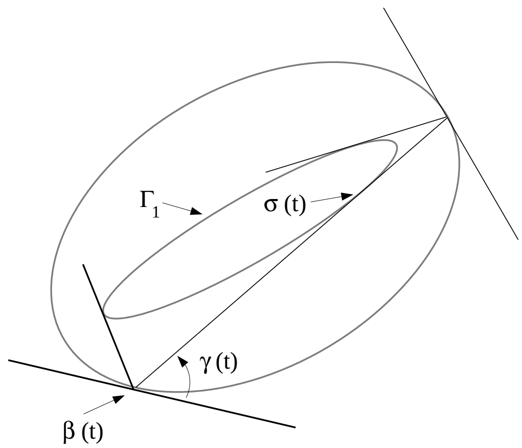

Figure A.3: Functions associated with a caustic.

Using the connection of $u(t)$ to these geometrically defined quantities,

$$u(t) = -\cos(\nu(t)) = -\beta'(t) \cdot \frac{\sigma(t) - \beta(t)}{\|\sigma(t) - \beta(t)\|},$$

and so

$$\frac{d}{dt}(\|\sigma(t) - \beta(t)\|) = u(t) + \frac{\sigma(t) - \beta(t)}{\|\sigma(t) - \beta(t)\|} \cdot \sigma'(t).$$

The area function (§1.4) associated with the invariant circle is

$$\begin{aligned} A(t) &= \int_t^{g(t)} u(s)\, ds - h(t, g(t)) \\ &= -\int_t^{g(t)} \frac{\sigma(s) - \beta(s)}{\|\sigma(s) - \beta(s)\|} \cdot \sigma'(s)\, ds + \int_t^{g(t)} \frac{d}{ds}(\|\sigma(s) - \beta(s)\|)\, ds - h(t, g(t)) \\ &= -\int_t^{g(t)} \|\sigma'(s)\|\, ds + \|\sigma(g(t)) - \beta(g(t))\| \\ &\qquad - \|\sigma(t) - \beta(t)\| + \|\beta(g(t)) - \beta(t)\| \end{aligned}$$

where we used the fact that $\frac{\sigma(s)-\beta(s)}{\|\sigma(s)-\beta(s)\|}$ is the unit vector in the direction of $\sigma'(s)$.

The first term in the final expression for $A(t)$ is the opposite of the length along Γ_1 between the two tangent rays (this length is labeled E in Figure A.2). If the perimeter of Γ_1 is P, this first term is thus $D - P$. The second and third terms in the expression for $A(t)$

are B and C as shown in Figure A.2, and so $A(t) = L(t) - P$. Thus using the appropriate version of Proposition 1.3, the string parameter is constant if and only if Γ_1 is a caustic for Γ.

Remark A.1: Note that in the billiards systems the generating function is always negative. The geometric interpretation of the generating function as an area in the annulus (Remark 1.1.1) is still valid, but now the area is below the t-axis. The area function in the case of billiards is a positive number giving the area below the invariant circle and between a vertical arc and its image.

References

[A] Angenent, S., Monotone recurence relations, their Birkhoff orbits and topological entropy, *Ergod. Th. & Dynam. Sys.*, **10**, 1990, 15–41.

[Bg] Bangert, V., Mather sets for twist maps and geodesics on tori, *Dyn. Rep.*, **1**, 1988, 1–56.

[B+H] Boyland, P. and Hall, G. R., Invariant circles and the order structure of periodic orbits in monotone twist maps, *Topology*, **26**, 21-36, 1987.

[BD] Budd, C. and Dux, F., Intermittency in impact oscillators close to resonance, *Nonlinearity*, **7**, 1994, 1191–1224.

[Dv] Devaney, R., *An Introduction to Chaotic Dynamical Systems,* Addison-Wesley, 1989.

[D] Douady, R., Thèse de Troisième Cycle, Univ. Paris 7, 1982.

[E] Eggleston, H., *Convexity*, Cambridge University Press, 1958.

[FT] Fuchs, D. and Tabachnikov, S., Segments of equal area, *Quantum*, **2**, 1992.

[Gs] Goldstein, H., *Classical Mechanics*, second edition, Addison–Wesley, 1980.

[Gl] Golé, C., Periodic orbits for Hamilitonian systems in cotangent bundles, *Trans. AMS*, **343**, 1994, 327–347.

[G1] Green, J., Sets subtending a constant angle on a circle, *Duke Math. Jour.*, **17**, 1950, 263–267.

[G2] Green, J., Length and area of a convex curve under affine transformation, *Pac. J. Math.*, **3**, 1953, 393–402.

[GK] Gutkin, E. and Katok, A., Caustics for innner and outer billiards, preprint, 1994.

[GS] Gutkin, E. and Simanyi, N., Dual polygonal billiards and necklace dynamics, *Comm. Math. Phys.*, **143**, 1991, 431–450.

[Ha] Halpern, B., Strange billiard tables, *TAMS*, **232**, 1977, 297–305.

[H] Herman, M., Sur les courbes invariantes par les difféomorphismes de l'anneau (volume 1), *Astérisque*, **103–104**, 1983.

[Hr] Hubacher, A., Instability of the boundary in the billiard ball problem, *Comm. Math. Phys.*, **108**, 1987, 483–488.

[K] Katok, A., Some remarks on the Birkhoff and Mather twist theorems, *Erg. Th. & Dynam. Syst.*, **2**, 1982.

[Kl] Kolodziej, R., The antibilliard outside a polygon, *Bull. Pol. Acad. Sci*, **37**, 1989, 163–168.

[Mc] MacKay, R., Introduction to the dynamics of area presrving maps, *Physics of Particle Accelerators*, ed. by Month, M. and Dienes, M., Am. Inst. Phys. Conf. Proc., **153**, vol. 1, 1987, 534–602.

[MP] MacKay, R. and Percival, I., Converse KAM: Theory and practise, *Comm. Math. Phys.*, **98**, 1985, 469–512.

[M1] Mather, J., Glancing billiards, *Ergod. Th. & Dynam. Sys.*, **2**, 1982, 397–403.

[M2] Mather, J., Nonexistence of invariant circles, *Ergod. Th. & Dynam. Sys.*, **4**, 1982, 301–309.

[M3] Mather, J., Letter to R. MacKay, 1984.

[M4] Mather, J., Variational construction of orbits of twist homeomorphisms, *Jour. A.M.S.*, **4**, 1991, 207–263.

[M5] Mather, J., Variational construction of connecting orbits, *Ann. Inst. Fourier, Grenoble*, 43, 1993, 1349–1386.

[MH] Mayer, K. and Hall, G. R., *Introduction to Hamilitonian Dynamical Systems and the N-body Problem*, Applied Mathematical Sciences, **94**, Springer-Verlag, 1991.

[Me1] Meiss, J., Symplectic maps, variational principles, and transport, *Rev. Mod. Phys.*, **64**, 1992, 795–848.

[Me2] Meiss, J., Cantori for the stadium billiard, *Chaos*, **2**, 1992, 267–272.

[Ms1] Moser, J., Is the solar system stable?, *Math. Intell.*, **1**, 1978, 65–71.

[Ms2] Moser, J., *Stable and Random Motions in Dynamical Systems*, Ann. of Math. Stud., Princeton University Press, **77**, 1973.

[Ms3] Moser, J., Recent developments in the theory of Hamiltonian dynamical systems, *SIAM Rev.*, **28**, 1986, 459–485.

[Ms4] Moser, J., Breakdown of stability, *Nonlinear Dynamics Aspaects of Particle Accelorators*, ed. by Jowett, J., Springer LNIP, **247**, 1986, 492–518.

[Ms5] Moser, J., Monotone twist maps and the calculus of variations, *Ergod. Th. & Dynam. Sys.*, **6**, 1986, 401–413.

[P] Poritsky, H., The billiard ball problem on a table with convex boundary – a illustrative problem, *Ann. Math.*, **51**, 1950, 446–470.

[SV] Shaidenko, A. and Vivaldi, F., Global stability of a class of discontinuous dual billiards, *Comm. Math. Phys.*, **110**, 1987, 625–640.

[S] Sinai, Y., *Introduction to ergodic theory*, Princeton Univ. Press, 1976.

[T1] Tabachnikov, S., Dual biliards, *Russ. Math. Surv.*, **48**, 1993, 75–102.

[T2] Tabachnikov, S., On the dual billiard problem, *Adv. in Math.*, to appear.

[T3] Tabachnikov, S., Billiards, University of Arkansas Preprint UofA-R-69, 1994.

[Tr] Turner, P., Convex caustics for billiards in $\mathbb{R}^2$ and $\mathbb{R}^3$, *Convexity and related combinatorial geometry*, Marcel Dekker, 1982, 85–105.

[W] Wojtkowski, M., Principles for the design of billiards with nonvanishing Lyapunov exponents, *Comm. Math. Phys.*, **105**, 1986, 391–441.