4 Low-Energy Transfers and Applications

EDWARD BELBRUNO
Department of Astrophysical Sciences, Princeton University; N.J. 08544-1000, U.S.A.

Contents

4.1 Introduction

The application of methods of dynamical systems theory to the field of astrodynamics has uncovered new types of low-energy trajectories that have many important applications. In particular, for the purpose of finding transfer trajectories from the Earth to lunar orbit. The mechanism used to obtain a transfer that is 'low energy' is called 'ballistic capture'. This is a process where a spacecraft is captured into lunar orbit without the use of rocket engines to slow down. The resulting transfer to the Moon is called a 'ballistic capture transfer'. Their property of being captured automatically into lunar orbit is completely different than that of the standard Hohmann transfer where substantial fuel must be used. This offers many advantages to the Hohmann transfer. In particular, they are substantially lower cost to use and operationally safer. The dynamical properties of the ballistic capture transfer are much more complicated than that of the Hohmann transfer, utilizing Newtonian four-body dynamics, as opposed to Newtonian two-body dynamics of the Hohmann transfer. The ballistic capture process itself is dynamically sensitive, but can be stabilized with a negligible maneuver.

A ballistic capture transfer was first operationally demonstrated in 1991 by the rescue of the Japanese spacecraft *Hiten* [2]. More recently, another type of ballistic capture transfer was used by ESAs spacecraft *SMART-1* [16, 17]. Their properties from the perspective of dynamical systems theory was investigated in 1994 [4], then by Marsden et al. [13]. Since then, a rigorous proof has been given showing that the ballistic process, in general, is chaotic in nature [4].

The use of low-energy trajectories has an interesting application on the origin of our own Moon. In a theory recently published by Belbruno and Gott [5], a class of low-energy transfers has shed light on the origin of the hypothetical Mars-sized object that slammed into the Earth 4 billion years ago to create the Moon. This is very briefly described in Section 4.5. For details, the reader should consult [5] (See also [10]).

Some popular survey articles on this material are cited in the bibliography. In the subject of astrodynamics, they are Refs. [8, 14, 16], and in the area of dynamical astronomy, Ref. [9]. Reference [4] provides a rigorous theoretical treatment of low-energy transfers. A more popular intuitive approach to the subject of chaos and low-energy transfers is given in [6].

4.2 Capture problem, models, and transfer types

In order to study transfers of spacecraft to the Moon, and other bodies, we define the *capture problem* to facilitate this.

A special four-body problem is generally defined between the spacecraft and three other planetary bodies. We will assume that the only forces acting on the spacecraft are the gravitational forces of the three bodies. We first consider a 'planar elliptic restricted three-body problem' between the particles P_1, P_2, P_3. That is, we assume that the spacecraft, labeled P_3, moves in the same plane as two planetary bodies P_1, P_2. The two planetary bodies move in prescribed mutually uniform elliptical Kelperian orbits about their common center of mass of eccentricity $e_{12} \approx 0$. We assume that P_1, P_2, P_3 move in a coordinate system Q_1, Q_2 which is inertial and centered at P_1 at the origin. The mass of P_1 is $m_1 > 0$; the mass of P_2 is $m_2 > 0$, where $m_2 \ll m_1$; and the mass of $P_3 = 0$. This latter assumption makes sense since P_1, P_2 are planetary sized bodies, and the mass of a spacecraft relative to them will be negligibly small.

A fourth mass point P_4, of mass $m_4 > 0$, is introduced. It is assumed to move about the center of mass point P_{cm} between P_1, P_2 in a uniform Keplerian ellipse of eccentricity $e_{124} \approx 0$. Since $m_2 \ll m_1$, then $P_{cm} \approx (0, 0)$ and P_4 approximately moves about P_1. We assume that the distance of P_4 from the center of mass of P_1, P_2 is much larger than the distance between P_1 and P_2, and that $m_1 \ll m_4$. The zero mass particle P_3 moves in the gravitational field generated by the assumed elliptic motions of P_1, P_2, P_4.

We refer to this model as a *planar elliptic restricted four-body problem*. It is shown in Figure 4.1. It could also be referred to as the *co-elliptic restricted four-body problem*. An example of this type of problem is where $P_1 =$ Earth, $P_2 =$ Moon, $P_3 =$ spacecraft, $P_4 =$ Sun. We will assume this labeling for the remainder of this chapter, although it should be noted that the results are not just limited to this choice of bodies.

If $e_{12} = e_{124} = 0$, then we refer to this as a co-circular restricted four-body problem. When $e_{12} = 0$ and we turn off the gravitational influence of P_4 by setting $m_4 = 0$, then this problem reduces to the *planar circular restricted three-body problem* between P_1, P_2, P_3. When $m_4 > 0$, and we turn off the gravitational influence of P_2 by setting $m_2 = 0$, then the circular restricted problem is obtained between P_1, P_3, P_4.

We define transfers from P_1 to P_2. In Figure 4.2 we just show P_1, P_2, which shows the conditions required for a transfer of P_3 from P_1 to P_2. P_4 is not shown.

Referring to Figure 4.2, the following assumptions are made:

- A1: The spacecraft, P_3, initially moves in a circular orbit about the Earth, P_1, of radius r_{13} as measured from the center of P_1.
- A2: A velocity increment magnitude ΔV_0 at the location $\mathbf{Q}_0$ at time t_0 on the circular orbit is added to the circular velocity $(Gm_1 r_{13}^{-1})^{\frac{1}{2}}$ so P_3 can transfer to the location $\mathbf{Q}_F$ near the Moon, P_2. Note that the vectors $\mathbf{Q}_0, \mathbf{Q}_F$ are in the coordinate system Q_1, Q_2,

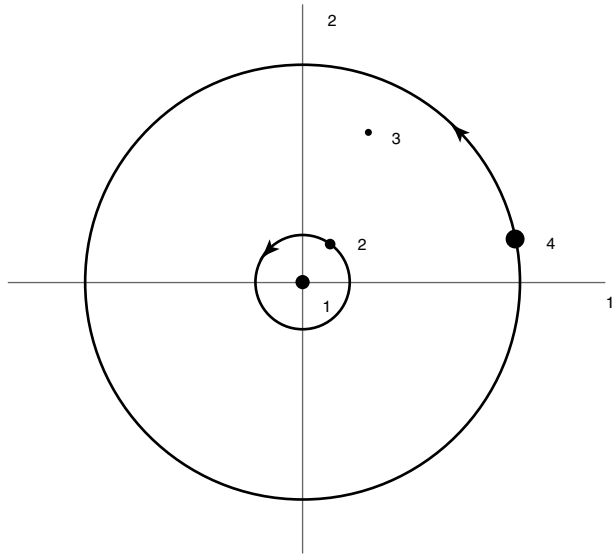

Fig. 4.1. Co-elliptic restricted four-body problem.

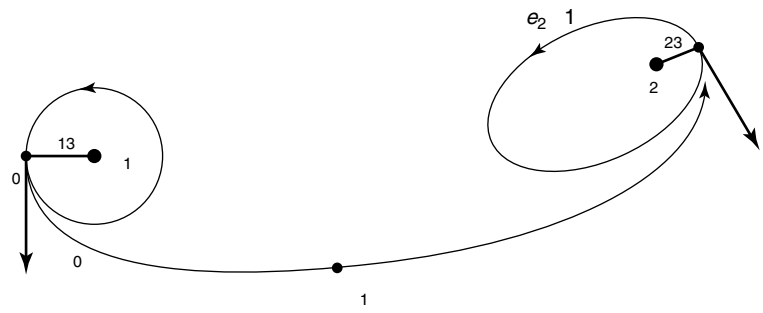

Fig. 4.2. The capture problem in inertial coordinates centered at P_1.

centered at the origin, P_1. G is the Newtonian gravitational constant, and r_{13} is the distance between P_1 and P_3.

- A3: A velocity increment magnitude ΔV_1 is applied at a time t_1, $t_0 < t_1 < t_F$; t_F is the arrival time at $\mathbf{Q}_F$.
- A4: A velocity increment magnitude ΔV_C is applied at $\mathbf{Q}_F$ in order that the two-body Keplerian energy E_2 between P_2, P_3 is negative or zero at $\mathbf{Q}_F$ so that at $t = t_F$ an oscillating ellipse of given eccentricity $0 \leq e_2 \leq 1$ is obtained of periapsis distance r_{23}. This defines an *instantaneous capture* at $\mathbf{Q}_F$ at time $t = t_F > t_0$ into an ellipse or parabola. We regard a parabola as an ellipse of infinite semimajor axis. $\mathbf{Q}_F$ represents the periapsis of the oscillating ellipse with respect to P_2 at distance r_{23}.

4.2.1 Remarks

Remark 1. The velocity increments $\Delta V_0, \Delta V_1, \Delta V_C$ are provided by firing the rocket engines of the spacecraft to impart a thrust, and hence a change in velocity. These increments are called ΔV's or *maneuvers*. In practice they cannot be achieved instantaneously, as we are assuming here, and depend on the magnitude of the ΔV. The engines may need to fire for a duration of a few seconds or several minutes. In general, modeling the ΔV's in an instantaneous or *impulsive* manner yields accurate modeling.

Remark 2. The term instantaneous capture in A4 also implies that for $t > t_F$ E_2 may become positive again. That is, P_3 is ejected right after being captured. It is generally the case that the ellipse shown in Figure 4.2 about the Moon may just exist when $t = t_F$. If ΔV_C is sufficiently large, then the capture ellipses can be stabilized for long times after $t = t_F$. In general, if it is desired to place a spacecraft several hundred kilometers from the surface of the Moon in a circular orbit after applying ΔV_C, then the orbit remains approximately circular for several months. Frequent ΔV's need to be applied by the spacecraft to maintain an orbit about the Moon; in general these are not stable due to nonuniformities of the mass distribution of the Moon and gravitational perturbations due to the Earth and Sun.

Remark 3. The term *osculating ellipse* in A4 means that the elliptical state at $t = t_F$ at $\mathbf{Q}_F$ may be unstable.

The *capture problem* is defined by the problem

$$\min\{\Delta V_0 + \Delta V_1 + \Delta V_C\}, \tag{4.1}$$

where the minimization is taken over all transfers from $\mathbf{Q}_0$ to $\mathbf{Q}_F$ and for assumptions A1–A4.

A solution of the capture problem for simplified assumptions with $\Delta V = 0$ is given by the Hohmann transfer.

4.2.2 Hohmann transfers

It is instructive to consider a typical Hohmann transfer between the Earth (P_1) and Moon(P_2). They have the property that at $\mathbf{Q}_F$, $\Delta V_C \gg 0$, or equivalently $E_2 \gg 0$, i.e., they are substantially hyperbolic with respect to the Moon at $\mathbf{Q}_F$. As we will see later, this is substantially different than ballistic capture transfers where $\Delta V_C = 0$ at $\mathbf{Q}_F$ and $E_2 \leq 0$. We only describe Hohmann transfers briefly here. They are described in detail throughout the astrodynamics literature.

The Hohmann transfer was developed by W. Hohmann in the early 1900s. Although his assumptions are oversimplifying in nature, they nevertheless lead to transfers from $\mathbf{Q}_0$ to $\mathbf{Q}_F$ which are very useful in practice, not just for the case $P_1 =$ Earth, $P_2 =$ Moon, $P_4 =$ Sun, but for transfers from the Earth to the other planets of our solar system.

Here, we discuss the Hohmann transfer that is relevant to Figure 4.2 for the Earth–Moon system. We assume that P_1, P_2 are in mutually circular orbits, i.e. $e_{12} = 0$.

The basic assumptions are the following: First, $m_4 = 0$ so that P_4 is not considered. Second, as P_3 transfers from $\mathbf{Q}_0$ to $\mathbf{Q}_F$, i.e., for $t_0 \le t \le t_F$, the gravity of P_2 is ignored, i.e., $m_2 = 0$. This yields a simple two-body problem between P_3, P_1, where then the two-body energy is then minimized. This gives one-half of a Kepler ellipse with periapsis at $\mathbf{Q}_0$ and apoapsis at $\mathbf{Q}_F$. This is the Hohmann transfer from $\mathbf{Q}_0$ to $\mathbf{Q}_F$. This ellipse arc has an eccentricity e_1. Upon arrival at $\mathbf{Q}_F$, the gravity of m_1 is ignored, $m_1 = 0$. m_2 is now assumed to be nonzero, and ΔV_C is computed relative to a two-body problem between P_3, P_2. $\mathbf{Q}_F$ is assumed to be on the far side of P_2 on the P_1–P_2 line. Breaking up a four-body problem into two disjoint two-body problems is an enormous simplification to the capture problem and dynamically is not correct. Nevertheless, these transfers change little when they are applied with full solar system modeling in many useful cases. This is because of the high energy associated with them. Their derivation is elegantly simple, and their usefulness is remarkable. They have paved the way for both human and robotic exploration of our solar system.

Of particular interest for applications considered later in this book is when r_{13}, r_{23} are relatively small numbers. Let km = kilometer, s = second. For example, $r_{13} = r_E + 200$ km, $r_{23} = r_M + 100$ km are typical radial distances used in applications of P_3 from P_1, P_2 at the locations $\mathbf{Q}_0, \mathbf{Q}_\mathrm{F}$, respectively, and at times $t = t_0, t_F$, respectively. The r_E and r_M represent the radii of the Earth and Moon, respectively. We will assume these values of r_{13}, r_{23} for the remainder of this paper for convenience. It is verified that $\Delta V_0 = 3.142$ km/s, $\Delta V_1 = 0$. $\Delta V_C = 0.200\,\mathrm{km/s}, 0.648\,\mathrm{km/s}$ for $e_2 = 0.95, 0$, respectively. Also, $t_F - t_0 = 5$ days. The transfer itself is nearly parabolic where $e_1 = 0.97$. Visually it would appear to be nearly linear. For Hohmann transfers in general, $E_2 \gg 0$ at $\mathbf{Q}_\mathbf{F}$. These values of r_{13}, r_{23} are the values that we desire for a solution of the capture problem.

It is verified that $E_2 > 0$ at $\mathbf{Q}_F$ for P_3, and this causes a large value of ΔV_C to occur. This property of $E_2 > 0$ is satisfied by Hohmann transfers. The reason $E_2 > 0$ follows from the fact that the magnitude V_F of the velocity vector at $\mathbf{Q}_\mathbf{F}$ of P_3 on the transfer at lunar periapsis, where the direction is in the same direction as the Moons orbit about the Earth, has the property that $V_F \ll V_M$, where V_M is the magnitude of the velocity of the Moon about the Earth. It turns out that under the given assumptions, $V_F = 0.176$ km/s and $V_M = 1.019$ km/s. This implies that $E_2 = 0.843\ \mathrm{km}^2/\mathrm{s}^2$. It is the discrepancy between V_F and V_M that yields a large value of ΔV_C of several hundred meters per second, depending on the value of e_2. The calculation of E_2 for a Hohmann transfer is estimated by noting that relative to P_2, the transfer is hyperbolic, with a hyperbolic periapsis at $\mathbf{Q}_\mathbf{F}$. The corresponding velocity at $r_2 = \infty$, called the hyperbolic excess velocity and labeled V_∞, is estimated by $V_\infty = V_M - V_F = 0.843$ km/s yielding $E_2 = (1/2)V_\infty^2$. The calculation of ΔV_C follows from a functional relationship it has with V_∞, or equivalently E_2.

4.2.3 Ballistic capture transfers

A ballistic capture transfer is defined to be a solution of the capture problem where $\Delta V_C = 0$ at $\mathbf{Q}_\mathbf{F}$ for $t = t_F$. It will arrive at periapsis at $\mathbf{Q}_\mathbf{F}$ where E_2 is negative, and therefore it will have no V_∞. This enables capture where $\Delta V_C = 0$. Eliminating the V_∞ is the motivation for the construction of ballistic capture transfers. From our discussion

of the Hohmann transfer, this means that a ballistic capture transfer has to arrive at $\mathbf{Q_F}$ where the spacecraft's velocity approximately matches the velocity of the Moon about the Earth. We will see that a ballistic capture transfer going from $\mathbf{Q}_0$ to $\mathbf{Q}_F$ can be constructed with approximately the same value of ΔV_0 for a Hohmann transfer and also with $\Delta V_1 = 0$. We refer to a Hohmann transfer as *high energy* since V_∞ is significantly high, and a ballistic capture transfer is called *low energy* since the V_∞ is eliminated.

The basic idea behind finding ballistic capture transfers to the Moon, or any body, is to find a region about the Moon, in position-velocity space (i.e., phase space), where an object can be ballistically captured. When such a region is found, then one can try to find trajectories from the Earth that go to that region. Ballistic capture enables capture to occur in an natural way, where a spacecraft need not slow down using its engines. A spacecraft moving in this region about the Moon lies in the transition between capture and escape from the Moon. Its motion in this region is very sensitive—being both chaotic and unstable.

In the next section we study in more detail how to determine where ballistic capture can occur about the Moon.

4.3 Ballistic capture regions and transfers

To better understand the process of ballistic capture, and the transfers themselves, it is instructive to consider the planar circular restricted three-body problem between the spacecraft, Earth, and Moon.

This defines the motion of P_3 in the gravitational field generated by the uniform circular motion of P_1, P_2 in an inertial coordinate system. P_3 moves in the same plane of motion as P_1, P_2. The constant frequency ω of motion of P_1, P_2 about their common center of mass at the origin is normalized to 1. It is assumed that $m_3 = 0$, and $m_1 + m_2 = 1$. We set $m_1 = 1 - \mu, m_2 = \mu,\ \mu = m_2/(m_1 + m_2)$. In a rotating coordinate system x_1, x_2 which rotates with the same frequency ω, both P_1, P_2 are fixed. We normalize the distance between P_1, P_2 to be 1. Without loss of generality we place P_1 at $(\mu, 0)$ and P_2 at $(-1 + \mu, 0)$. We assume here that $m_2 \ll m_1$, or equivalently $\mu \ll 1$. With $P_1 =$ Earth, $P_2 =$ Moon, $\mu = 0.0123$.

The differential equations of motion for P_3 are given by

$$\begin{aligned} \ddot{x}_1 - 2\dot{x}_2 &= x_1 + \Omega_{x_1} \\ \ddot{x}_2 + 2\dot{x}_1 &= x_2 + \Omega_{x_2}, \end{aligned} \tag{4.2}$$

where $\dot{} \equiv \dfrac{\mathrm{d}}{\mathrm{d}t}, \Omega_x \equiv \dfrac{\partial \Omega}{\partial x}$,

$$\Omega = \frac{1-\mu}{r_1} + \frac{\mu}{r_2},$$

$r_1 =$ distance of P_3 to $P_1 = [(x_1 - \mu)^2 + x_2^2]^{\frac{1}{2}}$, and $r_2 =$ distance of P_3 to $P_2 = [(x_1 + 1 - \mu)^2 + x_2^2]^{\frac{1}{2}}$, see Figure 4.3. The right-hand side of Eq. (4.3) represents the sum of the radially directed centrifugal force $\mathbf{F_C} = (x, y)$ and the sum $\mathbf{F_G}$ of the gravitational forces due to P_1 and P_2.

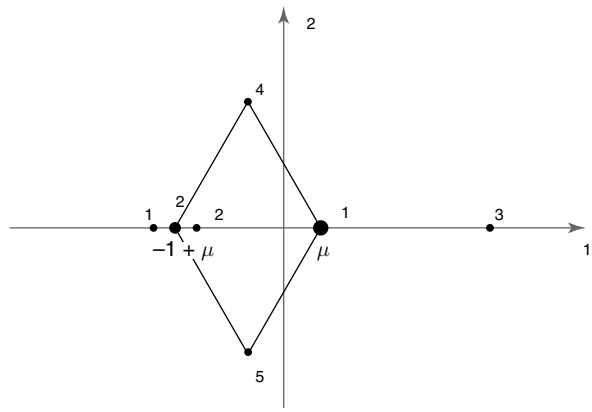

Fig. 4.3. Rotating coordinate system and locations of the Lagrange points.

The x_1 and x_2 are called barycentric rotating coordinates. If the coordinate systems were not rotating, then that defines a barycentric inertial coordinate system Q_1, Q_2. The transformation between x_1, x_2 and Q_1, Q_2 is given by a rotation matrix of rotational frequency 1. For notation we set $x = (x_1, x_2)$ and $Q = (Q_1, Q_2)$. x, Q are understood to be vectors.

An integral of motion for Eq. (4.2) is the Jacobi energy given by

$$J = -|\dot{x}|^2 + |x|^2 + \mu(1-\mu) + 2\Omega. \tag{4.3}$$

Thus

$$J^{-1}(C) = \{(x, \dot{x}) \in \mathbb{R}^4 \,|\, J = C,\ C \in \mathbb{R}\}$$

is a three-dimensional manifold in phase space for which the solutions of Eq. (4.2) which start on $J^{-1}(C)$ remain on it for all time. C is called the Jacobi constant. The additive term $\mu(1-\mu)$ occurring in Eq. (4.3) is present so that the values of C are normalized.

The manifold $J^{-1}(C)$ projected onto the physical (x_1, x_2)-plane form the *Hill regions*

$$\mathcal{H}(C) = \{x \in \mathbb{R}^2 \,|\, 2\widetilde{\Omega} - C \geq 0\},$$

where

$$\widetilde{\Omega} = \Omega + \frac{1}{2}|x|^2 + \frac{1}{2}\mu(1-\mu).$$

The particle P_3 is constrained to move in $\mathcal{H}(C)$, see Ref. [4]. The boundary of $\mathcal{H}(C)$ is given by the curves

$$\mathcal{Z}(C) = \{x \in \mathbb{R}^2 \,|\, 2\widetilde{\Omega} - C = 0\},$$

which are called *zero velocity curves*, because the velocity of P_3 vanishes there. The qualitative appearance of the Hill regions $\mathcal{H}(C)$ for different values of C are shown in

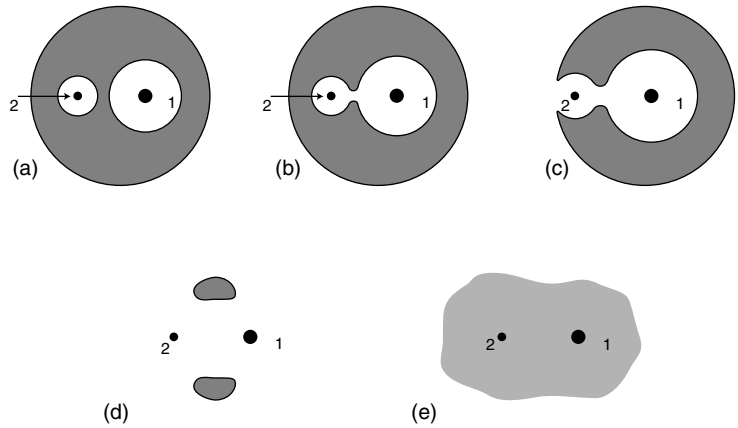

Fig. 4.4. Basic Hill's regions: Starting from the top three figures, left to right, C has the respective values, $C > C_2$, $C_1 < C < C_2$, $C \lesssim C_1$, and then the bottom two figures, left to right, $C_3 < C < C_1$, $3 < C < C_3$.

Figure 4.4. The particle P_3 cannot move in the hatched areas. The five values C_i are obtained by evaluating the function J at the five Lagrangian equilibrium points L_i of (4.2). The relative positions of the Lagrange equilibria are shown in Figure 4.3. The ones of interest in this paper are L_1 and L_2, which are unstable saddle center points [4]. The values C_i satisfy

$$C_4 = C_5 = 3 < C_3 < C_1 < C_2.$$

For $C < 3$, the Hill's region becomes the entire x_1, x_2-plane. Thus P_3 can move throughout the entire plane. When $C \lesssim C_2$, P_3 can pass between P_1 and P_2. For $C \geq C_1$ the Hill's region has two components. One is bounded, and the other is unbounded. When $C \lesssim C_1$, P_3 can move between the inner and outer Hill regions. C_2 represents the minimal energy for which P_3 can pass from P_1 to P_2. C_1 represents the minimal energy for which P_3 can pass between the bounded and unbounded components of the Hills regions.

It is noted that an approximation for C_1 and C_2 valid to three digits when $\mu \leq 0.01$, or four digits when $\mu \leq 0.001$, is

$$C_1 \approx 3 + 9\left(\frac{\mu}{3}\right)^{2/3} - 11\left(\frac{\mu}{3}\right), \quad C_2 \approx 3 + 9\left(\frac{\mu}{3}\right)^{2/3} - 7\left(\frac{\mu}{3}\right). \tag{4.4}$$

In inertial coordinates Eqs. (4.2) and (4.3), respectively, become,

$$\ddot{Q} = \Omega_Q, \tag{4.5}$$

$$\tilde{J} = -|\dot{Q}|^2 + 2(Q_1\dot{Q}_2 - Q_2\dot{Q}_1) + 2\Omega + \mu(1-\mu), \tag{4.6}$$

where

$$r_1(t) = \sqrt{(Q_1 + \mu c)^2 + (Q_2 + \mu s)^2},$$

$$r_2(t) = \sqrt{(Q_1 - (1-\mu)c)^2 + (Q_2 - (1-\mu)s)^2},$$

where $c \equiv \cos(t), s \equiv \sin(t)$. In these coordinates there is an explicit time dependence which is not the case in rotating coordinates.

Let $\phi(t) = (Q(t), \dot{Q}(t))$ be a solution of Eq. (4.5) for P_3. We assume at time $t = t_0$ it starts at some distance r_1 from P_1 and at time t_1 it is at a distance r_2 from P_2, $t_1 > t_0$. We are viewing $\phi(t)$ in the four-dimensional phase space. In position space it is given by $Q(t)$. As P_3 moves, $\tilde{J}(\phi(t)) = C$. We assume that no collision takes place so that $r_1 > 0, r_2 > 0$ along $\phi(t)$. Let $X = (X_1, X_2)$ be P_2-centered inertial coordinates.

Definition 4.3.1. The two-body Kepler energy of P_3 with respect to P_2 in P_2-centered inertial coordinates is given by

$$E_2(X, \dot{X}) = \frac{1}{2}|\dot{X}|^2 - \frac{\mu}{r_2} \tag{4.7}$$

where $r_2 = |X|$.

Definition 4.3.2. P_3 is *ballistically captured* at P_2 at time $t = t_1$ if

$$E_2(\phi(t_1)) \leq 0. \tag{4.8}$$

$\phi(t)$ is called a *ballistic capture transfer* from $t = t_0$ to $t = t_1$. If $E_2 \gtrsim 0$ at $t = t_1$ then P_3 is *pseudo-ballistically captured* at P_2.

It is noted that the notation $a \gtrsim b$ means that $a > b$ and $a - b = \delta \ll 1$.

When P_3 is ballistically captured with respect to P_2 the capture may lead to ballistic capture for all future time $t \geq t_1$, or this capture may be temporary where at a finite time $t = t_3 > t_2$, $E_2 > 0$. When this occurs then we say that P_3 has *ballistically escaped* P_2 for $t = t_3$. In this case the ballistic capture is *temporary*.

Ballistic capture can be stable or unstable whether it is temporary or not. By stability we mean *orbital stability*. That is, if the orbital elements of the motion change to a significant degree with very small changes in the initial conditions. If infinitesimally small changes in the initial conditions lead to predictably small changes in all the orbital elements for arbitrarily long time, then the motion is called stable, otherwise it is called unstable.

Temporary ballistic capture can be stable, so that although the Kepler energy is changing from negative to positive values, the orbital elements change in a small predictable way for all time. Likewise, ballistic capture for all time need not be stable. Thus whether or not the capture is temporary or not is not a good measure to describe the motion.

The key quantity to measure is the orbital stability. When ballistic capture is unstable we refer to it as *weak ballistic capture* or *weak capture* for brevity. A set where this occurs numerically can be estimated and is described in [4]. It can be analytically approximated by looking at the values of C of the Jacobi integral where the motion of ϕ has sufficiently high energy, where $C < C_1$. Also, we consider those points where $\dot{r}_2 = 0$. This defines a set W on the Jacobi integral surface $J^{-1}(C)$ in the coordinates $x, \dot{x}$,

Definition 4.3.3.

$$W = \{(x, \dot{x}) \in \mathbb{R}^4 | J = C, C < C_1, E_2 \leq 0, \dot{r}_2 = 0\}.$$

W is referred to as the *weak stability boundary*.

As is proven in [4] W on the three-dimensional surface $J^{-1}(C)$ is equivalent to a two-dimensional annular set about P_2. It is described by an explicit functional relationship

$$r_2 = f(\theta_2, e_2)$$

where θ_2 is the polar angle about P_2 in a P_2-centered rotating coordinate system, where f is periodic of period 2π in θ_2, and $0 \leq e_2 \leq 1$. This relationship can be conveniently written explicitly as

$$r_2 \approx \frac{(1 - e_2)\mu^{\frac{1}{3}}}{3^{\frac{5}{3}} - \frac{2}{3}\mu^{\frac{1}{3}}}$$

under the conditions that $C \lesssim C_1$ and $r_2 \gtrsim 0$.

We slightly extend the definition of W for the case of pseudo-ballistic capture and where we need not require $\dot{r}_2 = 0$. This set is labeled W_H, and is given by

$$W_H = \{(x, \dot{x}) \in \mathbb{R}^4 | J = C, C < C_1, E_2 \gtrsim 0 \ (i.e.\ e_2 \gtrsim 1)\}.$$

The set

$$\tilde{W} = W \cup W_H$$

is called the *extended* weak stability boundary. W_H represents points with respect to P_2 which are slightly hyperbolic and have $C < C_1$.

Numerical simulations indicate that the motion of trajectories with initial conditions on $\tilde{W}$ are generally unstable. In Theorem B in the next section shows that this is indeed the case due to the existence of a chaotic motion associated with $\tilde{W}$. W is referred to as the *weak stability boundary*, and $\tilde{W}$ is a hyperbolic extension of that set.

4.3.1 Method of determining ballistic capture transfers, and their properties

A method for finding ballistic capture transfers is to assume that your spacecraft is already at the extended weak stability boundary of the Moon at the desired capture distance r_{23} at the point $\mathbf{Q_F}$. This will give precise values of the velocity the spacecraft will have when it arrives at the Moon. Then one can employ a 'backwards integration method', where the trajectory is integrated backwards in time. Since the capture state at the Moon is unstable, tiny variations in the ballistic capture state can be used to target the trajectory in backwards time to have a periapsis with respect to the Earth at the point $\mathbf{Q_0}$. In another method, one can perform a 'forward algorithm', and start at the Earth at periapsis at $\mathbf{Q_0}$, and by varying specific control variables, target to the desired ballistic capture state at $\mathbf{Q_F}$. The details of this are described in [4].

The first numerical demonstration of the construction of a ballistic capture transfer was in 1986. This was for the Lunar Get Away Special (LGAS) mission study [1]. The numerical simulation was done for a low thrust spacecraft that was designed to be released from low Earth orbit from a Get Away Special cannister in the cargo bay of the space shuttle. After slowly spiraling out of Earth orbit with the solar electric ion engines for 1.5 years (using 3000 spirals), it reached a sufficiently large distance from the Earth where it shut off its engines, and moved on a ballistic capture transfer to the Moon over the north lunar pole, where it arrived in ballistic capture. It then turned its engines back on and took several months to gradually spiral down to the desired altitude at low lunar orbit at 100 km altitude.

The first operational demonstration of a ballistic capture transfer occurred a few years later. In 1991, a special ballistic capture transfer was used by E. Belbruno and J. Miller to resurrect a failed Japanese lunar mission, and get their spacecraft *Hiten* to the Moon since it had almost no fuel [2, 4, 16]. It took three months to reach the Moon instead of the 3 days a Hohmann transfer takes. The spacecraft first goes out about 1.5 million km from the Earth, then falls back to the Moon for ballistic capture. The transfer *Hiten* used is shown in Figure 4.5. The small elliptic orbit shown in the lower third quadrant is just a phasing orbit, and the transfer starts from near the Earth at the end of the phasing orbit.

Another name for a general ballistic capture transfer is a 'weak stability boundary (WSB) transfer'. *Hiten* is using an 'exterior' WSB transfer since it travels outside the orbit of the Moon. If a WSB transfer stays inside the Moon's orbit, it is called an 'interior' WSB transfer. *LGAS* used an interior transfer. ESAs *SMART-1* mission was

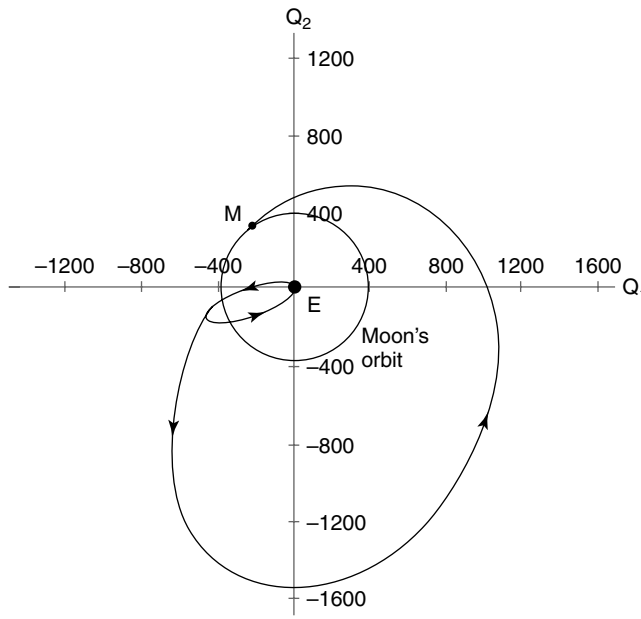

Fig. 4.5. The exterior ballistic capture transfer used by *Hiten*.

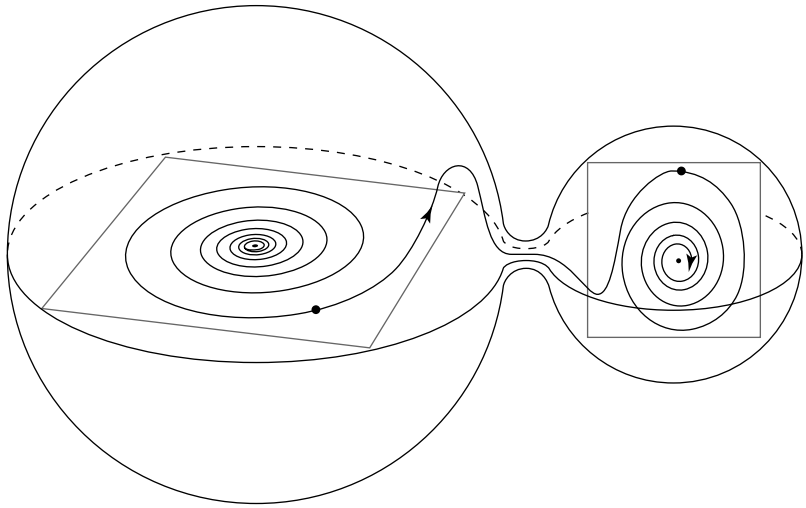

Fig. 4.6. Representation of complete transfer to low circular lunar orbit.

inspired by that design, and their spacecraft arrived at the Moon in November 2004. An illustration, not done accurately to scale, is shown in Figure 4.6. The trajectory lies within a three-dimensional Hills region about the Earth and Moon. The interior transfer itself in Figure 4.6 starts about 100000 from the Earth after the spiraling has stopped and the engines have been shut off. It ends when the transfer arrives at ballistic capture at approximately 30000 km over the north lunar pole. Over a period of a few months, using its engines, it slowly spirals down to low lunar orbit. The interior transfer has the disadvantage that it cannot start closer than approximately 60000 km from the Earth. The exterior transfer solves that problem.

The exterior WSB transfer is particularly important since it can be designed from any altitude from the Earth and go to any altitude at the Moon, and saves substantial fuel as compared to a Hohmann transfer. It promises to have important applications to future lunar missions due to its cost savings in bringing payloads into lunar orbit. The exterior WSB transfer's dynamics is complicated and is described in detail in Ref. [4]. We discuss a few of its properties here because of its interesting dynamics.

The exterior transfer at first appears to resemble a standard Hohmann bi-elliptic transfer. However, the looks are deceiving. Such a transfer, which is analogous to the one *Hiten* used, can leave the Earth at any altitude. If it leaves at low Earth orbit at an altitude of say 200 km, it will need approximately the same ΔV as a Hohmann transfer. But that is where the similarity stops. It takes about 1.5 months to reach the apoapsis at about 1.5 million km. While moving in this region, the gravitational forces of the Earth and Sun approximately balance as the spacecraft moves. It is actually moving in a weak stability boundary region about the Earth with the Sun in this case as the larger perturbing body. As the spacecraft arcs around and falls back towards the Moon, no maneuver is required to fall back towards the Moon. This is completely different than the bi-elliptic transfer

which requires a 0.250 km/s maneuver to do this. As the spacecraft falls back towards the Moon, the Sun is positioned in such a way so as to slow down the spacecraft as it approaches the Moon. In this way it can arrive with a velocity that approximately matches the Moons about the Earth. It will approach the Moon from outside the Moon's orbit. If the Jacobi constant C is just slightly less than C_1, the Hill's region opens slightly near the L_1 location, and the trajectory can pass though, passing close to the invariant manifolds associated with the Lyapunov orbit about the L_1 location. This is seen in Figure 4.7. It then passes into the Hill's region about the Moon and to low lunar orbit to weak capture. As is described in [4], and described below, the structure of the phase space where weak capture occurs is very complicated, and consists of an infinite set of intersecting invariant manifolds. It is important to point out that C need not be just be slightly less than C_1. This condition poses a large constraint that the trajectory needs to approach the Moon via the tiny opening near L_1. This condition also generally gives rise to transfers with times of flight on the order of 120 days. If we allow more generally that C could be substantially less than C_1, say $C < 3$, then the spacecraft can move anywhere in the physical space near the Moon, and is not constrained to pass near the location of L_1. This is because the Hill's region becomes the entire physical space and the zero velocity curves bounding the motion, no longer exist. Then the trajectory can be ballistically captured near the Moon by approaching the Moon, in general, from any direction. The time of flight also decreases to approximately 90 days.

Since ballistic capture transfers arrive at P_2 where $E_2 < 0$, they save substantial ΔV required to place P_3 into a capture orbit about P_2 relative to a Hohmann transfer. For

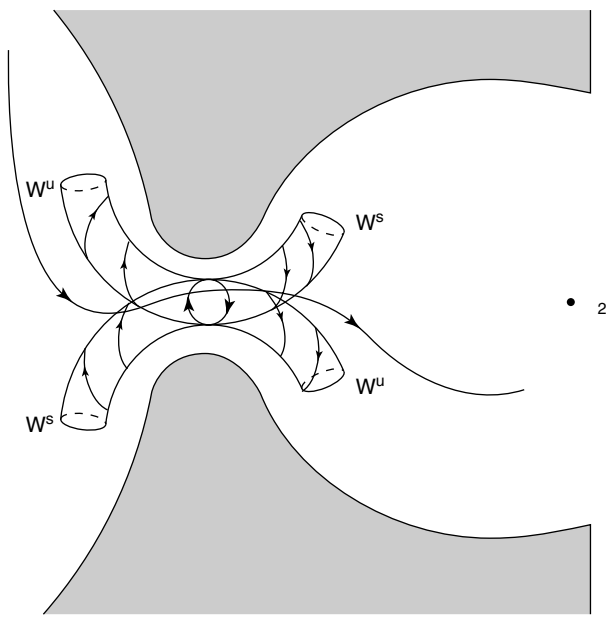

Fig. 4.7. Capture dynamics with $C \lesssim C_1$.

example, in the case of going into circular orbit of 100 km altitude, they save approximately 25% in ΔV, and to achieve an elliptic lunar orbit with a 100 km altitude, they require zero ΔV, where the osculating eccentricity is approximately 0.95 at the time of capture. It is remarked that the savings of 25% in ΔV is very significant, and can double the payload that one can place into low circular lunar orbit. At a cost of approximately 1 million dollars per pound to bring anything into lunar orbit, this savings is significant. Another advantage of WSB transfers is that their capture at the Moon is gradual, and not as risky as the Hohmann transfer which must perform a large capture maneuver in a short time span. A WSB transfer just gradually drifts into capture in a slow fashion and is much less risky.

It is instructive to consider other types of capture and see how ballistic capture may be related to them. A type of capture which is defined in a completely different way than ballistic capture is called permanent capture. This is topologically defined whereas ballistic capture is locally defined analytically.

Definition 4.3.4. P_3 is permanently captured in forward time with respect to both P_1, P_2 if

$$\lim_{t \to -\infty} |Q(t)| = \infty$$

and

$$|Q(t)| < a < \infty$$

as $t \to \infty$ where a is a finite constant. An analogous definition is given for permanent capture in backwards time.

Thus, for permanent capture, the particle P_3 comes in from infinity and remains bound to P_2 *for all time*. Permanent capture does not imply ballistic capture since while P_3 is bound to P_2, E_2 need not be negative. Permanent capture is an unstable process.

From this we can define another type of capture where a particle comes in from infinity, remains bounded for a finite period of time, then goes out to infinity as time goes to infinity. Thus the motion of P_3 is bounded for only a finite period of time. We call this *temporary capture* which is different from 'temporary ballistic capture' we defined previously.

Permanent capture has been studied from a mathematical perspective, and it can be proven to occur only for a set of measure zero in the phase space, and thus it is very unlikely.

It turns out that points exist on $\tilde{W}$ that also lead to permanent capture. Moreover, there exists a region on $\tilde{W}$ whose points lead more generally to chaotic motion of which permanent capture is of one type. This is described in the next section.

4.4 Chaos and weak capture

Consider a solution $Q(t)$ to (4.5) for P_3. $Q(t)$. It can be proven that under special conditions $Q(t)$ will perform chaotic motion. This chaotic motion occurs on a special set in phase space called a hyperbolic invariant set. We describe briefly the types of motions that can occur.

The orbits are near parabolic orbits, and lie between bounded and unbounded motion with respect to P_1 or P_2. Thus, they are in the transition between capture and escape from the P_1, P_2-system. For $\mu = 0$, they are Keplerian parabolic trajectories of P_3 about P_1, with Jacobi energy $|C| = 2\sqrt{2}$. A positive value of C represents direct motion about P_1, and a negative value of C represents retrograde motion.

The orbits start at a reference time $t = 0$ transversal to the Q_1-axis, slightly beyond the Moon, P_2, where $r_2 \gtrsim 0$ and move out to near infinity. For $\mu \ll 1$ the orbits appear nearly parabolic in appearance. They will in general fall back to the Q_1-axis crossing it again for $r_2 \gtrsim 0$ and then move out to infinity again. Then P_3 will fall by P_2 again passing slightly beyond the Moon, etc. This motion can repeat forever. It is also possible that while this oscillatory motion is occurring it is periodic in nature, or it can eventually escape and never return to the Q_1-axis. Or it can start from infinitely far from P_2 and then keep passing slightly beyond P_2 as it crosses the Q_1 axis while repeatably passing out to near infinity at a bounded distance, for all future time. This would correspond to permanent capture. In general many other types of motions can occur that pass slightly beyond the Moon.

The dynamics of this motion can be observed most easily by cutting the near parabolic oscillatory motion by a two-dimensional section Σ_t in phase space, at a given times t, where Σ_t is on the Q_1-axis, where $r_2 \gtrsim 0$, and for $\mu \ll 1$. A given orbit ψ will cut the axis at a sequence of times $t_k, k = 1, \ldots$, where

$$t_k < t_{k+1},$$

where $Q_2(t_k) = 0$. Set

$$s_k = \left[\frac{t_{k+1} - t_k}{2\pi}\right], \tag{4.9}$$

where $[a]$ is the largest integer $k \leq a$, for $a \in \mathbb{R}$. Thus, s_k gives a measure of the number of complete revolutions the primaries P_1, P_2 make (since they have period 2π) in the time it takes P_3 to make two passes through $Q_1 = 0$. The s_k can be used to define bi-infinite sequences

$$s = (\ldots s_{-2}, s_{-1}, s_0; s_1, s_2, \ldots).$$

Let S define the space of all such sequences.

The eventual general pattern of intersection points on Σ_{t_k} from all such orbits ψ takes on the appearance of that shown in Figure 4.8. This is called a *hyperbolic network* associated with a *transverse homoclinic point* r. The intersection of the invariant manifolds W^s, W^u associated to a hyperbolic equilibrium point p of the return map on Σ_t eventually forms a dense set of points, forming a Cantor set. Each of these points has a direction where the motion moves away from the point under iteration and another where it moves towards the point under iteration. This is analogous to a saddle point, except that the points of a hyperbolic network need not be equilibrium points. Thus the motion is very unstable. This Cantor set is called a hyperbolic network we label Λ. A motion defined on a hyperbolic network is called *chaotic*. This network is also called a hyperbolic invariant set [4].

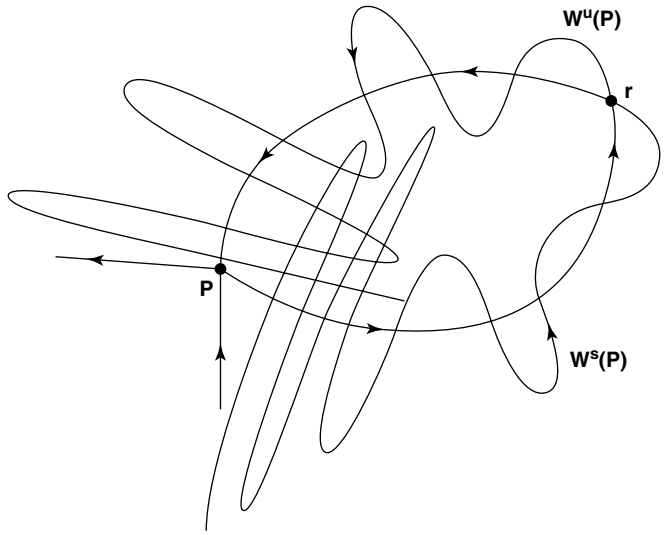

Fig. 4.8. Transverse homoclinic point **r** and hyperbolic network.

The existence of this type of dynamics near parabolic motion in another version of the three-body problem is proven to exist by Moser [15]. The case described above for the restricted problem was proven by Xia [19].

In order to describe the dynamics on Λ, the sequences of S can be used. It is remarkable that it can be proven that prescribing *any* sequence $s \in S$, a motion will exist for the restricted problem passing near to the Moon, where $|C| - 2\sqrt{2} \gtrsim 0$, $\mu \ll 1$. The condition on C means that the motion is near to parabolic for $\mu \ll 1$.

Theorem A. For $|C| - 2\sqrt{2} \gtrsim 0$, $\mu \ll 1$, $r_2 \ll 1$, there exists an integer $m = m(\mu, C)$ such that for any sequence $s \in S$ with $s_k \geq m$ there corresponds a solution of (4.5).

The sequences keep track of the itinerary of the motion of P_3 as it repeatably passes through the points of Λ.

Thus, to every bi-infinite sequence there corresponds an actual solution of the planar circular restricted three-body problem, which is near parabolic and oscillates in a chaotic fashion, repeatably passing very near to P_2 at a distance r_2. This gives an infinite variety of possible motions. Let's see what different sequences say. If the sequence is unbounded, then successive t_k become unbounded. This implies the solution takes so long to come back to $Q_2 = 0$, it in fact is becoming unbounded, but it has infinitely many zeros. This is an unbounded oscillatory solution. A periodic orbit would give rise to a repeating sequence, e.g., $s = \{\ldots, 1, 2, 1, 2, 1, 2, \ldots\}$

A permanent capture orbit which comes in from ∞ corresponds to a sequence terminating on the left with ∞, and then performs infinitely many bounded oscillations for all future time,

$$s = (\infty, \ldots s_k, s_{k+1}, \ldots),$$

where s_k are bounded. Temporary capture is defined by sequences which begin and end with ∞.

Theorem A can be applied to the problem of weak capture. It is very interesting that this is the case, and it is proven [4] that a subset of Λ exists on $\tilde{W}$.

Theorem B. A subset of the hyperbolic network Λ exists on the set W_H of the extended weak stability boundary $\tilde{W}$, which gives rise to the same chaotic motions in the bi-infinite sequence space S as described in Theorem A.

Thus weak capture is chaotic, and can lead to infinitely many possible motions including permanent capture. This fact as interesting applications to the possible construction of permanent capture transfers for spacecraft and also for the permanent capture of small bodies such as comets, asteroids or Kuiper belt objects about the Sun or a planet.

It is remarked that the chaos proven to exist in the weak capture process in Theorem B is associated to near parabolic motion which moves far from P_2. This is done by utilizing the transverse intersection of invariant manifolds associated with parabolic motion. Weak capture can also to studied by studying the possible transverse intersection of the invariant manifolds associated with the Lyapunov orbits about L_1, L_2. This is not yet analytically proven in a general manner; however, there exists an interesting numerically assisted proof of the transverse intersection of these invariant manifolds for some selected parameter values of μ and for $C \lesssim C_1$ [12].

4.5 Origin of the Moon

An outstanding question in astronomy is to understand where the Moon came from. One of the first theories to try to answer this is called the 'sister planet theory'. It proposed that the Moon formed together with the Earth as sister planets, in the solar nebula of gas and dust from which all the planets formed about 4 billion years ago. However, there are some inconsistencies with this. One is the fact that a large iron core is absent in the Moon, and present in the Earth, giving the Earth and Moon different densities, which are 5.5 grams/cm^3 and 3.3 grams/cm^3, respectively. Another theory is that the Moon was formed from beyond the Earth's orbit, and was captured into orbit about the Earth. If this were the case then the Earth and Moon would have different abundances of oxygen isotopes. This is inconsistent with the fact that the Earth and Moon have identical abundances.

A generally accepted theory which explains the differences in iron and the oxygen isotope abundances among other things is called the "impactor theory". It was formulated by W. Hartmann and D. Davis [11], and A. Cameron and W. Ward [7]. It proposes that after the Earth had already formed 4 billion years ago, a giant Mars-sized object smashed into the Earth. When it hit, it formed the Moon from iron poor mantel material debris primarily from the impactor, and also from the Earth, both of which already had iron cores. The Moon coalesced from this material. The iron core of the impactor was deposited into the iron core of the Earth. This explains the iron deficiency of the Moon. This theory also proposes that the impactor formed at the same 1 AU distance that the Earth is from the Sun. This explains the identical oxygen isotope abundances.

Numerical simulations from this theory show conditions of impact that an object the size of Mars would have to have to form the Moon from the resulting debris. They show that this object would have to approach the Earth with a relatively slow velocity—nearly parabolic with respect to the Earth, with velocities only a few hundred meters per second.

A fundamental question to ask is—*Where did this Mars-sized impactor come from?*

In 2001, Richard Gott described his theory to me to explain where the impactor may have come from. He proposed that at the time of the solar nebula from which the Earth was formed, there was so much debris flying around the Sun that it could have settled near the stable equilateral Lagrange points L_4, L_5 with respect to the Earth and Sun. Since these locations are stable, debris arriving there with a small relative velocity could remain trapped there. As more and more debris arrives, it could start to coalesce and a massive body could start to grow. Given several million years a large Mars-sized body could result.

However, here was a problem with his theory. How could it be demonstrated that the impactor could leave the L_4 (or L_5) neighborhood and impact the Earth? Back of the envelope calculations showed that collision would be unlikely since the impactor would likely fly by the Earth at high relative velocities of several kilometers per second, and easily miss the Earth.

The solution was found by calculating a WSB region about L_4 (or L_5) where a small massive object would first move captured in neighborhoods about these points and move in a horseshoe orbit—i.e., moving in an Earth-like orbit and oscillating back and forth, between counterclockwise and clockwise motions (without moving 360 degrees about the Sun), and not moving past the Earth. They would very gradually gain energy, and hence velocity with respect to the Sun, by the resulting interactions with the small planetesimals in the solar nebula at the time. Eventually, the impactor would grow in size, and move beyond the Earth, and the motion would bifurcate from the oscillating horseshoe motion to a non-oscillating cycling motion, repeatably flying closely by the Earth. This cycling motion is called 'breakout', and it is chaotic in nature. This is determined in the restricted problem by fixing a direction of motion at L_4 (or L_5) and gradually increasing the velocity until breakout occurs. This gives a parametric set of critical velocities about L_4 (or L_5) depending on direction, yielding a WSB about L_4 (or L_5). This is seen in Figure 4.9. It is

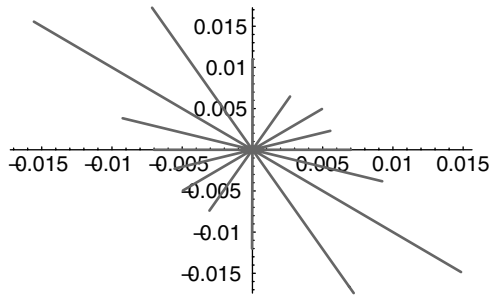

Fig. 4.9. Critical velocity magnitudes as a function of direction at L_4 giving rise to escape, or, equivalently, 'breakout' from L_4. (A velocity value of 1 corresponds to the velocity of the Earth about the Sun.)

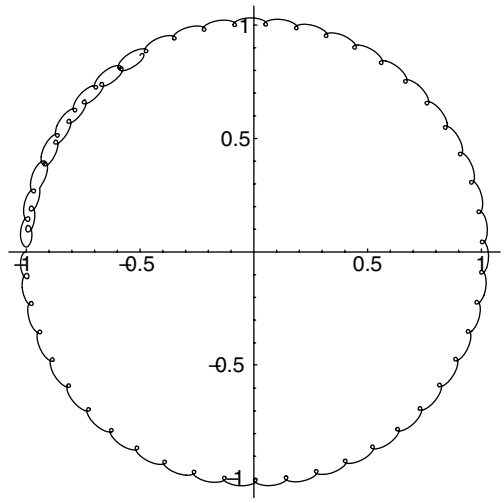

Fig. 4.10. A near parabolic creeping collision orbit emanating from L_4 in the third quadrant, and colliding with the Earth 57.32 years later near -1 on the x-axis. It first moves downward towards the Earth in a counterclockwise direction, then reverses its direction and moves nearly 360 degrees in a clockwise direction about the Sun at the origin, where it collides with the Earth.

shown in Ref. [5] that the likelihood of collision was very high, and that this process is preserved with more accurate modeling of the solar system. An actual collision orbit is shown in Figure 4.10.

The description presented here is very brief. The detailed exposition of this theory is given in Ref. [5]. Also, see Ref. [10]. The theory for the formation of the Mars impactor at L4/L5 as presented in Ref. [5] is currently being applied to the formation of some of the moons of Saturn, and to Saturn's rings in a collaboration with Gott, Vanderbei, and Belbruno [10, 18].

Acknowledgements

I would like to thank Pini Gurfil for helpful suggestions to put this paper into the desired form. This work was supported by grants from NASA.

References

1. Belbruno, E.A. (1987). Lunar Capture Orbits, A Method of Constructing Earth-Moon Trajectories and the Lunar GAS Mission, in: *Proceedings of AIAA/DGLR/JSASS Inter. Propl. Conf.* AIAA Paper No. 87-1054, (May 1987).
2. Belbruno, E.A. and Miller, J. (1993). Sun-perturbed earth-to-moon transfers with ballistic capture. *J. Guid., Control and Dynamics* **16**(4), July–August, 770–775.

3. Belbruno, E.A. (2004). Existence of Chaos Associated with Weak Capture and Applications, in *Astrodynamics, Space Missions, and Chaos* (E. Belbruno, D. Folta, and P. Gurfil eds.), **1017**, *Annals of the New York Academy of Sciences*, pp. 1–10.
4. Belbruno, E.A. (2004). *Capture Dynamics and Chaotic Motions in Celestial Mechanics*, Princeton University Press.
5. Belbruno, E.A. and Gott III, J.R. (2005). Where Did the Moon Come From? *Astronomical Journal*, **129** (4), March, pp. 1724–1745.
6. Belbruno, E.A. (2006). *Fly Me to the Moon*, Princeton University Press (planned for publication on December 15, 2006).
7. Cameron A.G.W. and Ward, W.R. (1976). The Origin of the Moon, in *Proc. Lunar Planet. Sci. Conf. 7th*, pp. 120–122.
8. Case, J. (2004). Celestial Mechanics Theory Meets the Nitty-Gritty of Trajectory Design, (book review of [4]), **37**(6), *SIAM News*, July–August, pp. 1–3.
9. Chown, M. (2004). The Planet that Stalked the Earth, (Featured cover story), *New Scientist*, August 14, pp. 26–30.
10. Gott III, J.R. (2006). Lagrange L4/L5 Points and the Origin of our Moon and Saturn's Moons and Rings, in *Astrodynamics and Its Applications* (E. Belbruno ed.), New York Academy of Sciences, Annals, V 1065.
11. Hartmann, W.K. and Davis, D.R. (1975). Satellite-sized Planetesimals and Lunar Origin, *Icarus*, **24**, pp. 504–515.
12. Koon, W.S., Lo, M.W., Marsden, J.E. and Ross, S.D. (2000). Heteroclinic Connections Between Periodic Orbits and Resonance Transitions in Celestial Mechanics, *Chaos*, **10**, pp. 427–469.
13. Koon, W.S., Lo, M.W., Marsden, J.E. and Ross, S.D. (2000). Shoot the Moon, *AAS/AIAA Astrodynamics Specialist Conference*, Florida, Paper Number AAS 000-166.
14. Klarreich, E. Navigating Celestial Currents (Featured cover story), *Science News*, **167**(16), (April 16, 2005), 250–252.
15. Moser, J. (1973). *Stable and Random Motions in Dynamical Systems*, Princeton University Press, V 77, Annals of Mathematics.
16. Osserman, R. (2005). *Mathematics of the Heavens*, *Notices of the American Mathematical Society (AMS)*, **52**(4), April, 417–424.
17. Racca, G. (2003). New Challanges to Trajectory Design by the Use of Electric Propulsion and Other Means of Wandering in the Solar System, *Celestial Mechanics and Dynamical Astronomy*, **85**, pp. 1–24.
18. Vanderbei, R. (2006). *Horsing Around on Saturn*, in *Astrodynamics and Its Applications* (E. Belbruno ed.), New York Academy of Sciences, Annals, V 1065.
19. Xia, Z. (1992). Melnikov Method and Transversal homoclinic Points in the Restricted Three-Body problem, *JDE* **96**, pp. 170–184.