The Spheres of Sol

Matei P. Coiculescu and Richard Evan Schwartz [*]

February 24, 2020

Abstract

Let Sol be the 3-dimensional solvable Lie group whose underlying space is $\boldsymbol{R}^3$ and whose left-invariant Riemannian metric is given by

$$e^{-2z}dx^2 + e^{2z}dy^2 + dz^2.$$

Let $E : \boldsymbol{R}^3 \to \text{Sol}$ be the Riemannian exponential map. Given $V = (x,y,z) \in \boldsymbol{R}^3$, let $\gamma_V = \{E(tV)|t \in [0,1]\}$ be the corresponding geodesic segment. Let AGM stand for the arithmetic-geometric mean. We prove that γ_V is a distance minimizing segment in Sol if and only if

$$\text{AGM}\bigg(\sqrt{|xy|}, \frac{1}{2}\sqrt{(|x|+|y|)^2+z^2}\bigg) \leq \pi.$$

We use this inequality to precisely characterize the cut locus in Sol, prove that the metric spheres in Sol are topological spheres, and almost exactly characterize their singular sets.

1 Introduction

1.1 Background

Sol is one of the 8 Thurston geometries [**Th**], the one which uniformizes torus bundles which fiber over the circle with Anosov monodromy. Sol

[*] Supported by N.S.F. Grant DMS-1807320

has sometimes been the topic of studies in coarse geometry and geometric group theory. The deep and difficult work of A. Eskin, D. Fisher, and K. Whyte [**EFW**], a landmark of geometric group theory, shows that any quasi-isometry of Sol is boundedly close to an isometry. As another example, N. Brady [**B**] proves that lattices in Sol are not asynchronously automatic.

The metric geometry of Sol is intriguing and mysterious. Sol has two totally geodesic foliations by hyperbolic planes, meeting at right angles, but somehow the two foliations are "turned upside down" with respect to each other. This engenders a kind of topsy-turvy feel. Another complicating feature is that Sol has sectional curvatures of both signs, causing an interplay of focus and dispersion. A number of authors have studied the differential geometry of Sol, with an emphasis on mean curvature surfaces. See the work by R. López and M. I. Munteanu [**LM**] and the references therein.

In [**T**], M. Troyanov integrates the geodesic equations for Sol and gets explicit formulas for the geodesics in terms of elliptic integrals. He uses these expressions to determine what he calls the *horizon* of Sol: the topological space of equivalence classes of geodesics, where two geodesics are equivalent if they have finite Hausdorff distance. The horizon gives information about the large-scale organization of the Sol geodesics. This theme is further pursued by S. Kim in [**K**]. In [**BS**], A. Bölcskei and B. Szilágyi take a related approach to the geodesics in Sol, with the view towards drawing pictures of the spheres in Sol. Their paper has pictures of the spheres of radius 1 and 2.

Matt Grayson's 1983 Princeton PhD thesis [**G**] takes a different approach to studying the geodesics. Working in a special frame of reference, Grayson converts the geodesic flow on Sol to a particular Hamiltonian flow on the 2-sphere, and then gives a detailed, penetrating analysis of the geodesics in Sol. We think that Grayson had many of the ingredients needed to establish the results in our paper, but he doesn't quite go in that direction. In any case, [**G**] was a tremendous inspiration for us.

The Hamiltonian flow approach, which we also take, goes back at least to V. I. Arnold's work [**A**] on hydrodynamics. See also the book by V. I. Arnold and B. Khesin [**AK**]. In a related direction, A. V. Bolsinov and I. A. Taimanov [**BT**] use the same formalism to study the geodesic flow on a 3-dimensional solv-manifold and construct an integrable geodesic flow with positive topological entropy.

In a different direction, R. Coulon, E. A. Matsumoto, H. Segerman, and S. Trettel [**CMST**] recently made a virtual reality ray-tracing program for Sol. We can say, from firsthand experience, that this thing is amazing.

1.2 Main Results

The AGM, or *arithmetic-geometric mean*, is defined for $0 \leq \alpha_0 \leq \beta_0$, as follows. We iteratively define

$$\alpha_{n+1} = \sqrt{\alpha_n \beta_n}, \qquad \beta_{n+1} = \frac{\alpha_n + \beta_n}{2}. \tag{1}$$

Then

$$\text{AGM}(\alpha_0, \beta_0) = \lim_{n \to \infty} \alpha_n = \lim_{n \to \infty} \beta_n. \tag{2}$$

This definition gives a rapidly converging sequence. See [**BB**] for details.

Given $V = (x, y, z) \in \boldsymbol{R}^3$ we define

$$\mu(V) = \text{AGM}\bigg(\sqrt{|xy|}, \frac{1}{2}\sqrt{(|x| + |y|)^2 + z^2}\bigg). \tag{3}$$

Note that $\mu(V) = 0$ iff $xy = 0$. Also, $\mu(rV) = |r|\mu(V)$. The function μ is an "extension" of the AGM because $\mu(V) = \text{AGM}(|x|, |y|)$ when $z = 0$.

We equip Sol with the left invariant metric

$$e^{-2z}dx^2 + e^{2z}dy^2 + dz^2. \tag{4}$$

Given $V \in \boldsymbol{R}^3$ as above, we let $\gamma_V = \{E(tV) | t \in [0, 1]\}$ be the corresponding geodesic segment. Here E denotes the Riemannian exponential map.

We call V and γ_V *small*, *perfect*, or *large* whenever we have $\mu(V) < \pi$, $\mu(V) = \pi$, or $\mu(V) > \pi$, respectively. These notions have an interpretation in terms of Hamiltonian dynamics. The vector field

$$\Sigma(x, y, z) = (xz, -yz, -x^2 + y^2), \tag{5}$$

which is the symplectic gradient of the function $F(x, y, z) = xy$, encodes the geodesic flow on Sol in a way we will describe in §2.2. The geodesic segment γ_V corresponds to some integral curve σ_V of Σ. At least generically, γ_V is small if σ_V is embedded, perfect if σ_V makes precisely one closed loop, and large if σ_V winds more than once around a closed loop. More geometrically, γ_V is small, perfect, or large according as γ_V spirals less than, equal to, or more than once around its *Grayson cylinder*. See §5.2 for a discussion.

Theorem 1.1 (Main) *A geodesic segment in Sol is a distance minimizer if and only if it is small or perfect. That is, γ_V is a distance minimizing geodesic segment if and only if $\mu(V) \leq \pi$.*

The Main Theorem is a very compact way of writing a more extensive result, which we call the Cut Locus Theorem. We now describe this result. Let Π be the plane $Z = 0$. We define sets

$$\partial_0 M \subset \partial M \subset M \subset \boldsymbol{R}^3, \qquad \partial_0 N \subset \partial N \subset N \subset \mathrm{Sol}$$

as follows.

- Let $M \subset \boldsymbol{R}^3$ be the set of small vectors.
- Let $\partial M \subset \boldsymbol{R}^3$ be the set of perfect vectors.
- Let $\partial_0 M = \partial M \cap \Pi$.
- Let $\partial_0 N = E(\partial_0 M)$.
- Let ∂N be the complement, in Π, of the component of $\Pi - \partial_0 N$ that contains the origin.
- Let $N = \mathrm{Sol} - \partial N$.

Note the sets N and ∂N are defined entirely from the 1-dimensional set $\partial_0 N$. It turns out that $\partial_0 N$ is the disjoint union of 4 properly embedded curves, each diffeomophic to a line and the graph of a function in polar coordinates. See Lemma 3.1. Figure 3 in §3.2 shows one component of $\partial_0 N$ and the corresponding component of ∂N.

Theorem 1.2 (Cut Locus) *The following is true:*

1. *E induces a diffeomorphism from M to N.*
2. *E induces a 2-to-1 local diffeomorphism from $\partial M - \partial_0 M$ to $\partial N - \partial_0 N$.*
3. *E induces a diffeomorphism from $\partial_0 M$ to $\partial_0 N$.*

The Cut Locus Theorem gives ∂N as the cut locus of the identity in Sol.

Our main motivation for understanding the cut locus is to understand something about the spheres in Sol. We think that opinion had been divided as to whether or not the metric spheres in Sol are topological spheres. In §4.3 we deduce the following easy corollary of the Main Theorem.

Theorem 1.3 (Sphere) *Metric spheres in Sol are topological spheres. For the sphere S_L of radius L centered at the identity in Sol the following holds.*

- *When $L < \pi\sqrt{2}$, the sphere S_L is smooth.*
- *When $L = \pi\sqrt{2}$, the sphere S_L is smooth except (perhaps) at the 4 points $(x, y, 0)$ where $|x| = |y| = \pi$.*
- *When $L > \pi\sqrt{2}$, the sphere S_L is smooth away from 4 disjoint arcs, all contained in the intersection of the plane $Z = 0$ and the set $|XY| = H_L^2$ for some $H_L > \pi$.*

We do not know whether the sphere $S_{\pi\sqrt{2}}$ is smooth at the 4 points $(x, y, 0)$ where $|x| = |y| = \pi$. The function $L \to H_L$ is defined by the following property.

$$L = \sqrt{8 + 8m}\mathcal{K}(m) \implies H_L = \frac{4\mathcal{E}(m)}{\sqrt{1-m}} - \sqrt{4 - 4m}\mathcal{K}(m). \tag{6}$$

Here $\mathcal{K}$ and $\mathcal{E}$ respectively are the complete elliptic integrals of the first and second kind, called `EllipticK` and `EllipticE` in Mathematica [**W**, p 774], and $m \in [0, 1)$ is the parameter used in Mathematica. See §5.2 for the definition of $\mathcal{K}$. One can derive Equation 6 from the formulas in [**G**] or [**T**], but we will not give a derivation because we do not need the formula for our proofs.

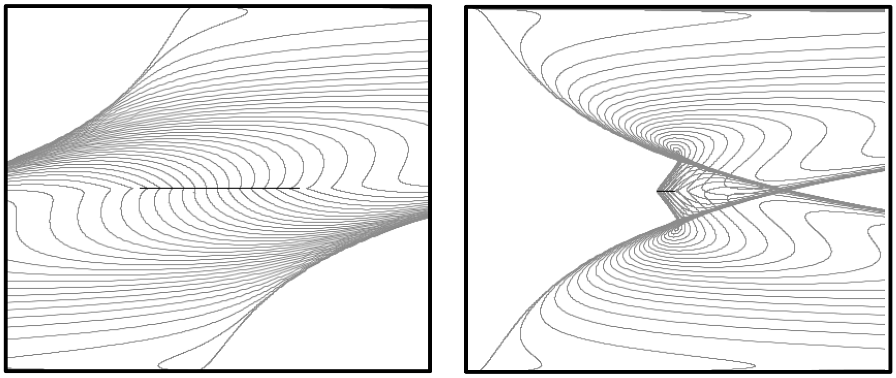

Figure 1: Two projections of the Sol metric sphere S_5.

Figure 1 shows two projections of a small portion of S_5. The black arc is one of the singular arcs mentioned in the Sol Sphere Theorem. The grey curves are images of lines of longitude under the exponential map. The Java program one of us wrote [**S**] generates these pictures and shows animations.

1.3 Proof Outline

We first recall several standard definitions from Riemannian geometry. See e.g. [**KN**, §8] for details. A geodesic segment is a *minimizer* if it is the shortest geodesic segment connecting its endpoints. It is a *unique minimizer* if it is the only such geodesic of minimal length connecting its endpoints. A geodesic segment γ_0 has a *conjugate point* if there is some nontrivial 1-parameter family γ_t of geodesics which vanishes to first order at 2 distinct points on γ_0, but not at all points along γ_0. The basic result is that if a geodesic segment is a minimizer, then every proper sub-segment is a unique minimizer without conjugate points. Call this the *restriction principle*. Now we can give the sketch.

Step 1: We call $V_+ = (x, y, z)$ and $V_- = (x, y, -z)$ *partners*. It turns out that V_+ is perfect if and only if V_- is perfect. Moreover, if $V_\pm$ is perfect, we prove that $E(V_+) = E(V_-)$. This is a surprising[1] result because the map $(x, y, z) \to (x, y, -z)$ is not an isometry of Sol. By the restriction principle (and a bit of fussing with the case $z = 0$), no large geodesic segment is a minimizer. We carry out this step in §2.

Step 2: This is the crucial step. We show that $E(M) \cap \partial N = \emptyset$. Hence $E(M) \subset N$. Let Π_+ be the portion of the plane $Z = 0$ above the X-axis and below the diagonal line $Y = X$. By symmetry it suffices to show that $E(M) \cap \partial N \cap \Pi_+ = \emptyset$. The set $E(M) \cap \Pi_+$ is a union of plane curves Ω_L with $L \in [\pi/2, \infty)$. Figure 3 from §3.2 shows some of these, in blue. We show that each such plane curve Ω_L is contained in the right triangle Δ_L shown (for $L = 5$) on the right side of Figure 3. The yellow set in Figure 3 is $\partial N \cap \Pi_+$. We prove the result that $\Omega_L \subset \Delta_L$, which we call the Bounding Triangle Theorem, by computing the differential equation satisfied by Ω_L and analyzing the behavior of this equation. The claim that $E(M) \cap \partial N = \emptyset$ follows readily. We carry out this step in §3.

Step 3: We show that $E(\partial M) \subset \partial N$. Combining this with step 2, we see that $E(\partial M) \cap E(M) = \emptyset$. The key point in showing that $E(\partial M) \subset \partial N$ is showing that E is injective on the closure of each component $\partial M - \partial_0 M$.

[1] We are not the first to notice this kind of phenomenon. [**K**, Lemma 4.1] is the less precise result that geodesics tangent to partner vectors meet "at some point". Sungwoon Kim proves this by analytic methods that differ from our more geometric approach.

This follows from our Corollary 2.10. We carry out this step in §4.1, though we prove Corollary 2.10 at the end of §2.

Step 4: Step 3 tells us that $E(M) \subset N$. Steps 1 and 3 tell us that if a geodesic segment γ is not a minimizer, then the actual minimizer γ^* with the same endpoints must also be perfect. The injectivity result in Step 3 then implies that γ and γ^* are the geodesic segments associated to partner perfect vectors, and hence have the same length, a contradiction. Hence, perfect geodesic segments are minimizers. We also carry out this step in §4.1.

Step 5: By the restriction principle, small geodesic segments are unique minimizers without conjugate points. Now we can say that the cut locus is ∂N. The rest of the proof is quite easy. We finish the proof of the Cut Locus Theorem in §4.2. At the end of §4 we deduce the Sphere Theorem from the Cut Locus Theorem, and then the Main Theorem from the Cut Locus Theorem and Equation 21.

1.4 Acknowledgements

We would like to thank ICERM for their fabulous Fall 2019 program, Illustrating Mathematics, during which this work was done. We would also like to thank Rémi Coulon, David Fisher, Bill Goldman, Alexander Holroyd, Boris Khesin Jason Manning, Greg McShane, Saul Schleimer, Henry Segerman, Sergei Tabachnikov, and Steve Trettel for many interesting discussions about Sol. We would also like to thank Matt Grayson for his great work on Sol.

2 Basic Structure

2.1 The Metric and its Symmetries

The underlying space for Sol is $\boldsymbol{R}^3$ and the group law is

$$(x,y,z)*(a,b,c)=(e^z a+x, e^{-z}b+y, c+z). \tag{7}$$

This is compatible with the left invariant metric on Sol given in Equation 4. For the sake of calculation, we mention two additional formulas:

$$(x,y,z)^{-1}=(-e^{-z}x,-e^{z}y,-z), \tag{8}$$

$$(x,y,z)^{-1}*(a,b,0)*(x,y,z)=(e^{-z}a,e^{z}b,0). \tag{9}$$

We identify $\boldsymbol{R}^3$ with the Lie algebra of Sol in such a way that the standard basis elements $(1,0,0)$, $(0,1,0)$, and $(0,0,1)$ respectively generate the 1-parameter subgroups $t\to(tx,0,0)$, $t\to(0,ty,0)$ and $t\to(0,0,tz)$. See §5.1 for a discussion of the left invariant vectorfields extending the standard basis elements.

Sol has 3 interesting foliations.

- The XY foliation is by (non-geodesically-embedded) Euclidean planes.
- The XZ foliation is by geodesically embedded hyperbolic planes.
- The YZ foliation is by geodesically embedded hyperbolic planes.

The complement of the union of the two planes $X=0$ and $Y=0$ is a union of 4 *sectors*. One of the sectors, the *positive sector*, consists of vectors of the form (x,y,z) with $x,y>0$. The sectors are permuted by the Klein-4 group generated by isometric reflections in the planes $X=0$ and $Y=0$. The Sol isometry $(x,y,z)\to(y,x,-z)$ also permutes the sectors. Because the coordinate planes $X=0$ and $Y=0$ are geodesically embedded, the Riemannian exponential map E carries each open sector of $\boldsymbol{R}^3$ into the same open sector of Sol. We will abbreviate this by saying that E is *sector preserving*.

There are 3 kinds of geodesics in Sol:

1. Certain straight lines contained in XY planes.
2. Hyperbolic geodesics contained in the XZ and YZ planes.
3. The rest. We call these *typical*.

We discuss the nature of typical geodesics in Sol in the next section.

2.2 The Geodesic Flow

Let $G = \text{Sol}$. Let $S(G)$ denote the space of unit tangent vectors based at the origin in G. Given a unit speed geodesic γ, the tangent vector $\gamma'(t)$ is part of a left invariant vector field on G, and we let $\gamma^*(t) \in S(G)$ be the restriction of this vector field to $(0,0,0)$. In terms of left multiplication on G, we have the formula

$$\gamma^*(t) = dL_{\gamma(t)^{-1}}(\gamma'(t)). \tag{10}$$

In §5.1 we verify that γ^* satisfies the following differential equation.

$$\frac{d\gamma^*(t)}{dt} = \Sigma(\gamma^*(t)), \qquad \Sigma(x,y,z) = (+xz, -yz, -x^2+y^2). \tag{11}$$

This is the point of view taken in [**A**] and [**G**]. This system in Equation 11 is really just geodesic flow on the unit tangent bundle of Sol, viewed in a left-invariant reference frame. Our formula agrees with the one in [**G**] up to sign, and the difference of sign comes from the fact that our group law differs by a sign change from the one there.

This vector field Σ has Klein-4 symmetry and vanishes at the 6 points: $(0,0,\pm 1)$ and $(\pm 1/\sqrt{2}, \pm 1/\sqrt{2}, 0)$. The first two points are saddle singularities and the rest are elliptic. The geodesics corresponding to the elliptic singularities are straight (diagonal) lines in the plane $Z = 0$. The geodesics corresponding to the saddle singularities are vertical geodesics in the XZ and YZ planes. The geodesics corresponding to the flowlines connecting the saddle singularities lie in the XZ and YZ planes; these are all geodesics of the second kind. The rest of the geodesics are typical. The flowlines corresponding to the typical geodesics lie on closed loops.

Let us say more about these closed loops. Let $F(x,y,z) = xy$. The restriction of F to S^2 gives a function on the sphere. The *symplectic gradient* X_F is defined by taking the gradient of this function (on the sphere) and rotating it 90 degrees counterclockwise. Up to sign $X_F = \Sigma$. By construction, the flow lines of Σ lie in the level sets of F. Most of the level sets of F are closed loops. We call these *loop level sets*.

Each loop level set Λ has an associated *period* $L = L_\Lambda$, which is the time it takes a flowline – i.e., an integral curve – in Λ to flow exactly once around. Equation 21 below gives a formula. We can compare L to the length T of a geodesic segment γ associated to a flowline that starts at some point of Λ and flows for time T. We call γ *small*, *perfect*, or *large* according as $T < L$, or $T = L$, or $T > L$.

2.3 Concatenation

Let g be the flowline given by

$$g(t) = (x(t), y(t), z(t)), \qquad t \in [0, T]. \tag{12}$$

The corresponding geodesic segment is $Tg(0)$. This geodesic has length T. We call g *small*, *perfect*, or *large* according as the corresponding vector $Tg(0)$ is small, perfect, or large. We define

$$\Lambda_g = E(Tg(0)) \tag{13}$$

Here Λ_g is the far endpoint of the geodesic segment corresponding to g when this segment starts at the origin.

We use the notation $g = a|b$ to indicate that we are splitting the flowline g into sub-flowlines a and b. Here a is some initial part of g and b is the final part. It follows from the left invariant nature of the geodesics that

$$\Lambda_g = \Lambda_a * \Lambda_b \tag{14}$$

This is also a consequence of Equation 17 below.

While the elements Λ_a and Λ_b do not necessarily commute, their vertical displacements commute. This gives us

$$\pi_Z \circ \Lambda_g = \pi_Z \circ \Lambda_a + \Pi_Z \circ \Lambda_b. \tag{15}$$

Here Π_Z is projection onto the Z-coordinate. Equation 15 has a nice integral form:

$$\pi_Z(\Lambda_g) = \int_0^T z(t)\ dt. \tag{16}$$

Remark: Here is how we numerically simulate geodesics in Sol and reproduce the numerics in [**G**]. We choose equally spaced times

$$0 = t_0 < t_1 < ... < t_n = T,$$

and consider the corresponding points $g_0, ..., g_n$ along the flowline g. We then have

$$\Lambda_g = \lim_{n \to \infty} (\epsilon_n g_0) * ... * (\epsilon_n g_n), \qquad \epsilon_n = T/(n+1). \tag{17}$$

In practice, we first pick some large n and then use Euler's method to find approximations to $g_0, ..., g_n$. We then take the product in Equation 17.

Let us deduce some consequences from the equations above. We call g a *symmetric flowline* if the endpoints of g have the form $(x,y,+z)$ and $(x,y,-z)$. Let Π be the plane $Z=0$.

Lemma 2.1 *A small flowline g is symmetric if and only if $\Lambda_g \in \Pi$.*

Proof: If g is symmetric, then the integral in Equation 16 vanishes, by symmetry. Hence $\pi_Z(\Lambda_g)=0$. If b is a small flowline having both endpoints on the same side of Π then $\Pi_Z(\Lambda_b) \neq 0$ because the integrand in Equation 16 either is an entirely negative function or an entirely positive function. In general, if g is not symmetric then we can write $g=a|b|c$ where a,c are either symmetric or empty, and b lies entirely above or entirely below Π. But then $\pi_Z(\Lambda_g)=\pi_Z(\Lambda_b)\neq 0$ by Equation 15. ♠

Lemma 2.2 *If g is a perfect flowline then $\Lambda_g \in \Pi$. If g_1 and g_2 are perfect flowlines in the same loop level set, and $\Lambda_{g_j}=(a_j,b_j,0)$, then $a_1b_1=a_2b_2$.*

Proof: We can write $g=u|v$ where u and v are both small symmetric flowlines. But then by Equation 15 and Lemma 2.1,

$$\pi_Z(\Lambda_g)=\pi_Z(\Lambda_a)+\pi_3(\Lambda_b)=0+0=0.$$

Hence $\Lambda_g \in \Pi$ and we can write $\Lambda_g=(a,b,0)$. We can write $g_1=u|v$ and $g_2=v|u$ for suitable choices of small flowlines u and v. Then $\Lambda_{g_1}=\Lambda_a * \Lambda_b$ and Then $\Lambda_{g_2}=\Lambda_b * \Lambda_a$. Hence Λ_{g_1} and Λ_{g_2} are conjugate in Sol. The second statement now follows from Equation 9. ♠

Our Theorem 2.3 below strengthens [**K**, Lemma 4.1], but the method of proof is completely different. Let E be the Riemannian exponential map.

Theorem 2.3 *If V_+ and V_- are perfect partners, then $E(V_+)=E(V_-)$.*

Proof: Let $g_\pm \subset S(G)$ be the flowline corresponding to $V_\pm$. We can write $g_+=u|v$ and $g_-=v|u$ where u and v are small flowlines. Since V_+ and V_- are partners, we can take u and v both to be symmetric. But then the elements Λ_u and Λ_b both lie in the plane $Z=0$ and hence commute. Hence, by Equation 14, we have $E(V_+)=\Lambda_{g_+}=\Lambda_u*\Lambda_v=\Lambda_v*\Lambda_u=\Lambda_{g_-}=E(V_-)$. ♠

2.4 Large Geodesic Segments are not Minimizers

Now we complete Step 1 of our proof outline. Our result is essentially a corollary of Theorem 2.3, but we have to bring in some other results to handle special cases.

Lemma 2.4 *If $x, z > 0$ and (x, x, z) is perfect, then $E(x, x, z) = (h, h, 0)$ for some h.*

Proof: This is a result of [**G**]. Here is another proof. Let g be the flowline corresponding to (x, x, z). We can write $g = u|v$ where u is the flowline starting at (x, x, z) and ending at $(x, x, -z)$ and v is the flowline starting at $(x, x, -z)$ and ending at (x, x, z). Both u and v are small symmetric arcs. Also, the isometry $(x, y, z) \to (y, x, -z)$ swaps u and v. Hence $\Lambda_u = (\alpha, \beta, 0)$ and $\Lambda_v = (\beta, \alpha, 0)$. But then

$$E(x, x, y) = \Lambda_g = (\alpha, \beta, 0) * (\beta, \alpha, 0) = (h, h, 0),$$

with $h = \alpha + \beta$. ♠

Corollary 2.5 *A large geodesic segment is not a length minimizer.*

Proof: If this is false then, by the restriction principle, we can find a perfect geodesic segment γ, corresponding to a perfect vector $V = (x, y, z)$, which is a unique geodesic minimizer without conjugate points. If $z \neq 0$ we immediately contradict Theorem 2.3. If $z = 0$ and $|x| \neq |y|$ we consider the variation, $\epsilon \to \gamma(\epsilon)$, through same-length perfect geodesic segments $\gamma(\epsilon)$ corresponding to the vector $V_\epsilon = (x_\epsilon, y_\epsilon, \epsilon)$. The vectors V_ϵ and $V_{-\epsilon}$ are partners, so $\gamma(\epsilon)$ and $\gamma(-\epsilon)$ have the same endpoint. Hence, this variation corresponds to a conjugate point on γ, a contradiction.

It remains only to consider the segments connecting $(0, 0, 0)$ to $(t, \pm t, 0)$ with $|t| > \pi$. By symmetry it suffices to show that the segment connecting $(0, 0, 0)$ to $(t, t, 0)$ is not a distance minimizer when $t > \pi$. This is proved in [**G**]. For the sake of completeness, we give another proof. It follows from Equation 21 below that there are values $h \in (\pi, t)$ such that $(h, h, 0) = E(V)$ for some perfect vector V of the form (x, x, z) with $z \neq 0$. Hence, by the restriction principle, the segment connecting $(0, 0, 0)$ to $(t, t, 0)$ is not a distance minimizer. ♠

2.5 The Reciprocity Lemma

In this section we prove a technical result which is a crucial ingredient for Step 2 of our outline. We discovered this result experimentally. It does not appear in [**G**].

Lemma 2.6 (Reciprocity) *Let $V = (x, y, z)$ be any perfect vector. Then there some $\lambda \neq 0$ such that $E(V) = \lambda(y, x, 0)$.*

Proof: By symmetry it suffices to work in the positive quadrant. We write $\zeta = \zeta(t)$ for any quantity ζ which depends on t. Let $p = (x, y, z)$ be a flowline for the structure field Σ with initial conditions $x(0) = y(0)$ and (say) $z(0) > 0$. Let $(a, b, 0) = E(x, y, z)$. We want to show that $a/b = y/x$ for all t. We do this by showing that the two functions satisfy the same O.D.E. and have the same initial conditions. We get the same initial conditions by Lemma 2.4: we have $a(0)/b(0) = 1 = y(0)/x(0)$.

We get the O.D.E. for y/x using the definition of Σ and the product rule:

$$\frac{d}{dt}\frac{y}{x} = \frac{y'x - x'y}{x^2} = \frac{-yzx - xzy}{x^2} = -2z \times \frac{y}{x}. \tag{18}$$

Now we work out the O.D.E. satisfied by a/b. By definition,

$$\frac{d}{dt}\frac{a}{b} = \lim_{\epsilon \to 0} \frac{1}{\epsilon}\bigg(\frac{a(t+\epsilon)}{b(t+\epsilon)} - \frac{a(t)}{b(t)}\bigg).$$

Let $p(t, \epsilon)$ denote the minimal flowline connecting $p(t)$ to $p(t+\epsilon)$. Referring to the definition in §2.3, we have

$$\Lambda_{p(t,\epsilon)} \approx \epsilon(x, y, z). \tag{19}$$

Here the approximation means that we have equality up to order ϵ^2. Hence

$$(a(t+\epsilon), b(t+\epsilon), 0) = \Lambda^{-1}_{p(t,\epsilon)} * (a, b, 0) * \Lambda_{p(t,\epsilon)} \approx$$

$$(\epsilon x, \epsilon y, \epsilon z)^{-1} * (a, b, 0) * (\epsilon x, \epsilon y, \epsilon z) = (ae^{-\epsilon z}, be^{+\epsilon z}, 0).$$

The first equality is Equation 14. The approximation (to order ϵ^2) comes from Equation 19. The last equality is Equation 9. But then

$$\frac{d}{dt}\frac{a}{b} = \lim_{\epsilon \to 0} \frac{e^{-2\epsilon z} - 1}{\epsilon} \times \frac{a}{b} = -2z \times \frac{a}{b}. \tag{20}$$

Therefore a/b satisfies the same O.D.E. as does y/z. ♠

2.6 The Period Function

Let L_a be the period of the loop level set containing $U_a = (a, a, \sqrt{1-2a^2})$.

Lemma 2.7 *$dL_a/da < 0$.*

Proof: This is part of [**G**, Lemma 3.2.1], and it also follows from the formula

$$L_a = \frac{\pi}{\mathrm{AGM}(a, \frac{1}{2}\sqrt{1+2a^2})}. \tag{21}$$

We derive this formula in §5.2. ♠

Lemma 2.8 *Let $V_0 = (x, y, z)$ be a perfect vector with $x, y, z > 0$. Then E is a local diffeomorphism in a neighborhood of V_0.*

Proof: By the Inverse Function Theorem, this is the same as saying that the differential dE has full rank at V_0. Let S be the sphere in $\boldsymbol{R}^3$ centered at the origin and containing V_0. Let T_0 be the tangent plane to S at V_0. Let N_0 be the orthogonal complement of T_0. As is well known, the images $dE|_{V_0}(T_0)$ and $dE|_{V_0}(N_0)$ are orthogonal, and the latter space is 1-dimensional. So, we just need to show that $dE|_{V_0}(T_0)$ contains 2 linearly independent vectors. Below, the symbols $O(t)$ and $O(1)$ denote quantities which respectively are bounded below by positive constants times t and 1. Both our variations below consist of vectors all having the same length.

Let $V_t \in S$ be a curve of perfect vectors which moves at unit speed away from V_0 and which remain in a single loop level set. The projection of V_t into Π moves at speed $O(1)$ because $V_t \not\in \partial_0 M$. This point moves monotonically along a hyperbola. By the Reciprocity Lemma, $E(V_t)$ moves with speed $O(1)$ in Π. Hence $dE_{V_0}(T_0)$ contains a nonzero vector of the form $(a, b, 0)$.

Let Π be the plane $Z = 0$. Now let $V_t \in S$ be the curve of constant-length vectors moving at unit speed orthogonally to the loop level sets, with V_t a small vector for $t > 0$. Let g_t be the flowline corresponding to V_t. Let Θ_t be the loop level set containing g_t. Let h_t be the complementary flowline, so that $g_t|h_t$ is a perfect flowline in Θ_t. By Equation 16, we have $\pi_Z \circ E(V_t) = -\pi_Z(\Lambda_{h_t})$. Let $L(t)$ be the period of Θ_t. By Lemma 2.7 we have $dL/dt > 0$. Hence h_t travels for time $O(t)$. The distance from Π to h_t is $O(1)$. Therefore, by Equation 15, we have $|\pi_Z(\Lambda_{h_t})| = O(t)$. Hence $dE|_{V_0}(T_0)$ contains a vector (a', b', c') with $c' \neq 0$. ♠

2.7 The Holonomy Function

If V is a perfect vector, and $(a,b,0)=E(V)$, then we let $H(V)=\sqrt{|ab|}$. We call $H(V)$ the *holonomy invariant* of V. By Lemma 2.2, this only depends on the loop level set. Thus H is a function of L, the level set period. By Equation 21 we have $L \geq \pi\sqrt{2}$ and [2] $H(\pi\sqrt{2})=\pi$. The next result is stated on [**G**, p 78]. We give independent proofs.

Lemma 2.9 *$dH/dL \geq 0$, with strict inequality when when $L > \pi\sqrt{2}$. Also, H is a proper monotone increasing function of L.*

Proof: Let us first show that H is an unbounded function. Pick an arbitrary $R>0$ and let V be the shortest vector such that $E(V)=(R,R,0)$. Corollary 2.5 says that V is either small or perfect. In either case, there is some $\lambda \geq 1$ such that λE is perfect. Geodesic segments in the positive sector cannot be tangent to the coordinate planes $X=X_0$ or $Y=Y_0$. Hence $E(\lambda V)=(a,b,0)$ with $a,b \geq R$. Hence $H(\|\lambda V\|) \geq R$.

Now we know that H is unbounded. Suppose there is $L > \pi\sqrt{2}$ where $H'(L)=0$. Consider the perfect vectors $U_t=(x_t,x_t,z_t)$, with positive coordinates, such that $\|U_t\|=L+t$. By Lemma 2.4, we have $E(U_t)=(a_t,a_t,0)$. Since $H'(L)=0$ we have $da/dt(0)=0$. This shows that dE is singular at U_0. But this contradicts Lemma 2.8. Hence H' has just one sign on $(\pi\sqrt{2},\infty)$. Since H is unbounded, the sign must be positive. Since H is monotone and unbounded, H is proper. ♠

Our final result is not in [**G**].

Corollary 2.10 *The map E is injective on the set of perfect vectors having all non-negative coordinates.*

Proof: Since V_1 and V_2 have the same holonomy invariant, Lemma 2.9 implies $\|V_1\|=\|V_2\|$. But then Lemma 2.7 implies that $U_1=V_1/\|V_1\|$ and $U_2=V_2/\|V_2\|$ lie in the same loop level set. Hence $U_{11}U_{12}=U_{21}U_{22}$. But then $V_{11}V_{12}=V_{21}V_{22}$. Here U_{ij} and V_{ij} respectively are the jth coordinates of U_i and V_i. The Reciprocity Lemma says that $V_{12}/V_{11}=V_{22}/V_{21}$. Hence $V_{11}=V_{21}$ and $V_{21}=V_{22}$. Since $\|V_1\|=\|V_2\|$ we have $V_{13}=V_{23}$ as well. ♠

[2] The results of Grayson we mention here are stated in terms of $D=H\sqrt{2}$. Grayson states that $D \geq L$ and $D \sim \sqrt{2}\exp(L/4)$ as $L \to \infty$. (His formula for D in terms of L on p 75 line -7 has a typo – an extra factor of 2.)

3 Controlling Small Geodesic Segments

3.1 Preliminary Topological Information

The goal of this chapter is to prove that $E(M) \cap \partial N = \emptyset$, where M and ∂N are as in the Cut Locus Theorem. Here we gather some preliminary topological information. Given any set S, either in $\boldsymbol{R}^3$ or Sol, let S_+ denote the intersection of S with the positive sector. Also, let Π be the plane $Z = 0$. The properness statement in the next lemma justifies our definition of N as the closure of a certain union of components of $\Pi - \partial_0 N$.

Lemma 3.1 *The $\partial_0 N_+$ is the graph of a function in polar coordinates, diffeomorphic to $\boldsymbol{R}$, and properly embedded in Π.*

Proof: The set $\partial_0 M_+$ is the graph of a smooth function in polar coordinates. By Equation 21, the function is

$$f(\theta) = \frac{\pi}{\mathrm{AGM}(\sin(\theta), \cos(\theta))}. \tag{22}$$

The polar defining function g for $\partial_0 N_+$ is

$$g(\theta) = \sqrt{2/\sin(2\theta)} \times H(f(\theta)) \geq \frac{\pi\sqrt{2}}{\sqrt{\sin(2\theta)}}. \tag{23}$$

Here H is the holonomy function from Lemma 2.9. The statements in the lemma follow from this formula.

Here we derive Equation 23. Let $V = (x, y, 0) \in \partial_0 M_+$ be the vector which makes angle θ with the X-axis. By the Reciprocity Lemma we have $E(V) = \lambda(y, x, 0)$. This gives us

$$g(\theta) = \|E(V)\| = \lambda\sqrt{x^2 + y^2}, \qquad H(f(\theta)) = H(\|V\|) = \lambda\sqrt{xy}.$$

We also have the trig identity:

$$\frac{\sqrt{x^2 + y^2}}{\sqrt{xy}} = \sqrt{2/\sin(2\theta)},$$

Equation 23 comes from these relations and a bit of algebra. The inequality in Equation 23 comes from Lemma 2.9 and the fact $H(\pi\sqrt{2}) = \pi$. ♠

3.2 Small Symmetric Flowlines

We want to show that $E(M) \cap \partial N = \emptyset$. Let $M_+^{\rm symm} \subset M$ denote those vectors having all positive coordinates which correspond to small symmetric flowlines in the sense of Lemma 2.1. Let Π be the plane $Z = 0$.

Lemma 3.2 *Suppose that $E(M) \cap \partial N \neq \emptyset$. Then $E(M_+^{\rm symm}) \cap \partial N_+ \neq \emptyset$.*

Proof: Let $V = (x, y, z) \in M$ be such that $E(V) \in \partial N$. By symmetry, it suffices to assume that $x, y, z \geq 0$. Since E is sector-preserving, we must have $E(V) \in \partial N_+$. If $x = 0$ then $E(V)$ lies in the plane $X = 0$, a set which is disjoint from ∂N_+. Hence $x > 0$. Similarly, $y > 0$. If $x = y$ and $z = 0$ then $E(V) = (x, x, 0)$ and $x < \pi$. Since the minimum holonomy invariant is π, the vector $E(V)$ is too short to land in ∂N.

We have ruled out all the possibilities which correspond to vectors *not* associated to small flowlines in the sense of §2.3. So, V is associated to a small flowline. Since $\partial N_+ \subset \Pi$, we must have $E(V) \in \Pi$. But then, V is associated to a small *symmetric* flowline, by Lemma 2.1. In this case, we must have $z > 0$ because the endpoints of small symmetric flowlines are partner points in the sense of §2.3. So, $V \in M_+^{\rm symm}$, as claimed. ♠

By Equation 21, every vector in $M_+^{\rm symm}$ has length L for some $L > \pi\sqrt{2}$. Let Θ_L^+ denote those points in the (unique in the positive sector) loop level set of period L having all coordinates positive. Every element of $M_+^{\rm symm}$ corresponds to a small symmetric flowline starting in Θ_L^+ for some $L > \pi\sqrt{2}$.

The Canonical Parametrization: The set Θ_L^+ is an open arc. We fix L and we set $\ell = L/2$. Let $p_0 = (x(0), y(0), 0) \in \Theta_L \cap \Pi$ be the point with $x(0) > y(0)$. We then let

$$p_t = (x(t), y(t), z(t)) \tag{24}$$

be the point on Θ_L^+ which we reach after time $t \in (0, \ell)$ by flowing *backwards* along the structure field Σ. That is

$$\frac{dp}{dt} = (x', y', z') = -\Sigma(x, y, z) = (-xz, +yz, x^2 - y^2). \tag{25}$$

Here and below we use x' to stand for dx/dt, etc.

The Associated Flowlines: We let $\overline{p}_t$ be the partner of p_t, namely

$$\overline{p}_t = (x(t), y(t), -z(t)). \tag{26}$$

We let g_t be the small symmetric flowline having endpoints p_t and $\overline{p}_t$. Since the structure field Σ points downward at p_0, the symmetric flowline g_t starts out tiny and increases all the way to a perfect flowline as t increases from 0 to ℓ. We the limiting perfect flowline g_ℓ. Figure 2 shows the symmetric flowlines g_t as t increases. The arrows indicate the direction of the structure field flow.

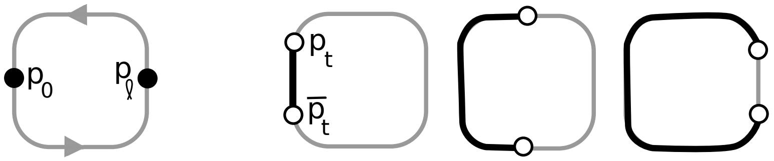

Figure 2: Increasingly long symmetric flowlines.

The Associated Plane Curves: Let $V_t \in M_+^{\rm symm}$ be the vector corresponding to g_t. (To reconcile our notation, we recall from §2.3 that $E(V_t) = \Lambda_{g_t}$.) Define

$$\Lambda_L(t) := E(V_t) = (a(t), b(t), 0) \qquad t \in (0, \ell]. \tag{27}$$

These curves are the main objects of interest to us.

Lemma 3.3 *$\Lambda_L(\ell) \in \partial_0 N_+$, and $0 < b(\ell) < a(\ell)$.*

We have $\Lambda_L(\ell) \in \partial_0 N_+$ because g_ℓ is perfect and starts at $(x(\ell), y(\ell), 0)$. Note that $x(\ell) = y(0)$ and $y(\ell) = x(0)$. Hence $x(\ell) < y(\ell)$. The reciprocity Lemma, applied to the perfect vector V_ℓ, now gives $0 < b(\ell) < a(\ell)$. ♠

We have $E(M_+^{\rm symm}) \cap \partial N_+ = \emptyset$ provided that

$$\Lambda_L(0, \ell) \cap \partial N_+ = \emptyset, \qquad \forall L > \pi\sqrt{2}. \tag{28}$$

So all we have to do is establish Equation 28.

Equation 28 looks true numerically. The left side of Figure 3 shows part of $\partial_0 N_+$ and $\Lambda_L(0, \ell)$ for various values of L. The right side of Figure 3 shows part of ∂N_+ in yellow and focuses on the curve Λ_5. The right side also shows the triangle Δ_5, where Δ_L is the triangle with vertices $(0, 0, 0)$, $(a(\ell), 0, 0)$ $(a(\ell), b(\ell), 0)$.

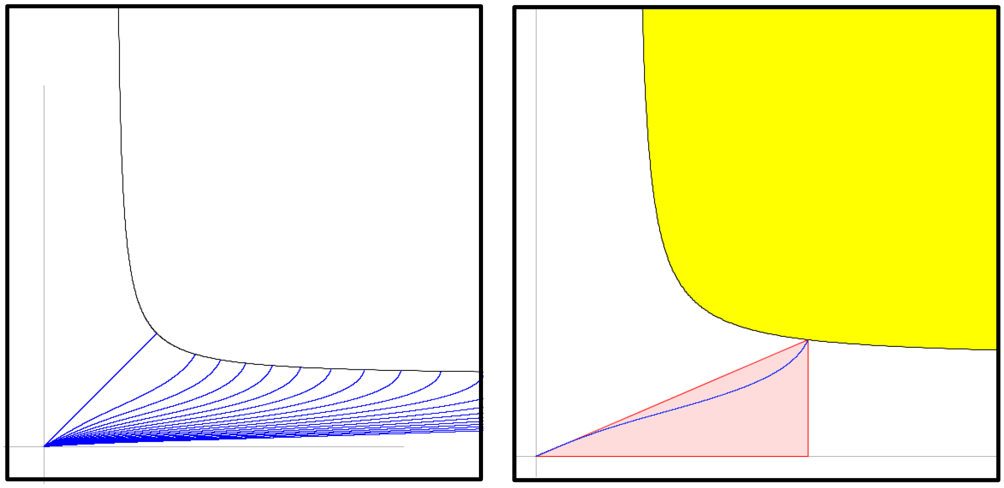

Figure 3: $\partial_0 N_+$ (black), ∂N_+ (yellow), Λ_L (blue), and Δ_L (red).

Theorem 3.4 (Bounding Triangle) *$\Lambda_L(0,\ell) \subset \mathrm{interior}(\Delta_L)$ for all L.*

The Bounding Triangle Theorem gives us what we need to establish Equation 28. We need one preliminary result.

Lemma 3.5 *If $(a,b,0) \in \partial_0 N_+$ is such that $0 < b < a$, and Δ is the solid triangle with vertices $(0,0,0)$, $(a,0,0)$ and $(a,b,0)$, then $\partial_0 N_+ \cap \Delta = \{(a,b,0\})$.*

Proof: Let ρ be the ray in Π_+ starting at the origin and going through $(a,b,0)$. Let h be the positive component of the hyperbola $XY = ab$. Let us trace out $\partial_0 N_+$ starting at the angle $\theta = \pi/4$ and decreasing θ. Because we are tracing out a graph in polar coordinates, we are separated from Δ by ρ until we hit $(a,b,0)$. But then, by Lemma 2.9, we are separated from Δ by h thereafter. So, we just brush past Δ, hitting $(a,b,0)$ and missing the rest of it. ♠

Corollary 3.6 *$E(M) \cap \partial N = \emptyset$.*

Proof: By Lemma 3.5, interior of Δ_L is disjoint from ∂N_+. Hence Equation 28 is true. Equation 28 combines with Lemma 3.2 to finish the proof. ♠

3.3 Proof of The Bounding Triangle Theorem

Now we prove the Bounding Triangle Theorem. Recall that

$$p_t = (x(t), y(t), z(t)), \qquad \Lambda_L(t) = (a(t), b(t), 0). \tag{29}$$

Here $a, b, x, y, z > 0$ on $(0, \ell)$. In particular, $\Omega_L(0, \ell)$ avoids the bottom side of Δ_L. We will show that $a' > 0$ on $(0, \ell)$. This implies that $\Omega_L(0, \ell)$ is the graph of a function and hence avoids the vertical side of Δ_L. We define

$$f(t) = \phi(t) - \phi(\ell), \qquad \phi(t) = \frac{b(t)}{a(t)}. \tag{30}$$

We will show that $f < 0$ on $(0, \ell)$. This means that $\Omega(0, \ell)$ avoids the diagonal side of Δ_L as well. So, the two inequalities $a' > 0$ and $f < 0$ on $(0, \ell)$ imply the Bounding Triangle Theorem.

The First Inequality: Compare the proof of the Reciprocity Lemma. We write $g_{t+\epsilon} = u|g_t|v$, where u is the flowline connecting $p_{t+\epsilon}$ to p_t and v is the flowline connecting $\overline{p}_t$ to $\overline{p}_{t+\epsilon}$. We have

$$(a', b', 0) = \Lambda_L'(t) = \lim_{\epsilon \to 0} \frac{\Lambda_L(t+\epsilon) - \Lambda(t)}{\epsilon},$$

$$\Lambda_L(t+\epsilon) \approx (\epsilon x, \epsilon y, \epsilon z) * (a, b, 0) * (\epsilon x, \epsilon y, -\epsilon z).$$

The approximation is true up to order ϵ^2 and $(*)$ denotes multiplication in Sol. A direct calculation gives $a' = 2x + az$ and $b' = 2y - bz$. Since $a, x, z > 0$ on $(0, \ell)$ we have $a' > 0$ on $(0, \ell)$.

Continuous Extension: The function f extends continuously to 0 and $f(0) = 0$. (Geometrically, the diagonal edge of Δ_L is tangent to Λ_L at the origin. Compare Figure 3.) To see this, we use L'Hopital's rule:

$$\phi(0) = \lim_{t \to 0} \frac{2y(t) - b(t)z(t)}{2x(t) + a(t)z(t)} = \frac{y(0)}{x(0)} = \frac{x(\ell)}{y(\ell)} =^* \frac{b(\ell)}{a(\ell)} = \phi(\ell). \tag{31}$$

The starred equality is the Reciprocity Lemma, which applies to the perfect vector V_ℓ. The function $\psi = ab' - ba'$ also extends continuously to 0: We have $\psi(0) = 0$ because $a(0) = b(0) = 0$ and a', b' do not blow up at 0.

The Second Inequality: We have $f(0)=f(\ell)=0$. If $f\geq 0$ somewhere on $(0,\ell)$ then f has a local maximum at some $t_0\in(0,\ell)$. We have $f'(t_0)=0$ and $f''(t_0)\leq 0$. Recalling that $\psi=ab'-ba'$, we have

$$\psi=ab'-ba'=a^2\phi'=a^2f', \qquad \psi'=2aa'f'+a^2f''. \tag{32}$$

Hence $\psi(t_0)=0$ and $\psi'(t_0)\leq 0$. Recalling Equation 25, we have

$$a'=2x+za, \quad b'=2y-zb, \quad x'=-xz, \quad y'=yz, \quad z'=x^2-y^2.$$

From all this and from the fact that $\psi'=ab''-ba''$, we get

$$a''=(+x^2-y^2+z^2)a, \quad b''=(-x^2+y^2+z^2)b, \quad \psi'=2ab(y^2-x^2). \tag{33}$$

Note that $y^2-x^2=-z'$ is negative on $(0,\ell/2)$, zero at $\ell/2$, and positive $(\ell/2,\ell)$. (Compare Figure 2.) Hence ψ' has these same properties. Hence $t_0\leq\ell/2$. But ψ is negative on $(0,\ell/2]$ because $\psi(0)=0$ and ψ' is negative on $(0,\ell/2)$. Hence $\psi(t_0)<0$. This contradiction establishes the second inequality. Our proof is done.

4 The Main Results

4.1 Separating Small and Perfect Vectors

Here we carry out Step 3 of the outline. Let E be Riemannian exponential map. Let Π be the plane $Z = 0$. Let $\mathcal{M}$ be the component of $\partial M_+ - \partial_0 M_+$ which contains vectors with all coordinates positive. Let $\mathcal{N} = \partial N_+ - \partial_0 N_+$.

Lemma 4.1 *$E(\mathcal{M}) \subset \mathcal{N}$.*

Proof: By Corollary 2.10, the map E is injective on $\mathcal{M} \cup \partial_0 M_+$. At the same time, $E(\partial_0 M_+) = \partial_0 N_+$. Hence

$$E(\mathcal{M}) \subset \Pi - \partial_0 N_+. \tag{34}$$

By definition, $\mathcal{N}$ is one of the components of the $\Pi - \partial_0 N_+$. Therefore, since $\mathcal{M}$ is connected, the image $E(\mathcal{M})$ is either contained in $\mathcal{N}$ or disjoint from $\mathcal{N}$. By Lemma 2.4 and Lemma 2.9 we have $E(V) \in \mathcal{N}$ where $V \in \mathcal{M}$ has the form (x, x, z) and $\|V\|$ is large. So, we have containment rather than disjointness. ♠

Corollary 4.2 *$E(\partial M) \cap E(M) = \emptyset$.*

Proof: Up to Sol symmetries, every vector in ∂M lies either in $\mathcal{M}$ or in $\partial_0 M$. By definition, $E(\partial_0 M) = \partial_0 N \subset \partial N$. So, by the previous result, we have $E(\partial M) \subset \partial N$. By Corollary 3.6 we have $E(M) \cap \partial N = \emptyset$. Combining these two statements gives the result. ♠

Theorem 4.3 *Perfect geodesic segments are length minimizing.*

Proof: Suppose $V_1 \in \partial M$ and $E(V_1) = E(V_2)$ for some V_2 with $\|V_2\| < \|V_1\|$. We take V_2 to be the shortest vector with this property. By symmetry we can assume both V_1 and V_2 have all coordinates non-negative. By Corollary 2.5, we have $V_2 \in M \cup \partial M$. By Corollary 4.2 we have $V_2 \in \partial M$. But then $V_1 = V_2$ by Corollary 2.10. This is a contradiction. ♠

4.2 Proof of the Cut Locus Theorem

The results above identity ∂M as the cut locus of the identity of Sol – we will justify this momentarily. So, our next lemma seems redundant, given standard properties of the cut locus in a Riemannian manifold. But we include the proof just to be sure.

Lemma 4.4 *The map E is a proper map from M to N.*

Proof: Note first that $E(M) \subset N$ because Sol $= N \cup \partial N$ and $E(M) \cap \partial N = \emptyset$ by Corollary 3.6. Suppose $\{V_n\}$ is a sequence in M which exits every compact subset of M. Since vectors in M correspond to distance minimizing geodesics, we have $\|E(V_n)\| \to \infty$ when $\|V_n\| \to \infty$. If $\|V_n\|$ remains bounded than $V_n \to \partial M$. By continuity $E(V_n) \to \partial N$ in this case. Hence $E(V_n)$ exits every compact subset of N. Hence $E : M \to N$ is proper. ♠

Proof of Statement 1: We know already that a geodesic segment is a length minimizer if and only if it is small or perfect. By the restriction principle mentioned in the introduction, small geodesic segments are unique length minimizers and they have no conjugate points. Hence $E : M \to N$ is an injective, proper, local diffeomorphism. But this implies that $E : M \to N$ is also surjective and hence a diffeomorphism. ♠

Proof of Statement 2: Let $\mathcal{M}$ and $\mathcal{N}$ be as in the previous section. By symmetry it suffices to prove that E is a diffeomorphism from $\mathcal{M}$ to $\mathcal{N}$. Note first that $E(\mathcal{M}) \subset \mathcal{N}$ by Lemma 4.1. To see that $\mathcal{M}$ is a smooth surface, note that $\mathcal{M}$ is an open subset of $\mu^{-1}(\pi)$. Given that $\mu(tV) = t\mu(V)$ for $t > 0$ we see that any positive value, including π, is a regular value for the smooth function μ. Hence $\mathcal{M}$ is a smooth surface. We now see that $E : \mathcal{M} \to \mathcal{N}$ is a local diffeomorphism by Lemma 2.8, injective by Corollary 2.10, and proper by an argument just like the one given in Lemma 4.4. All these properties together imply that $E : \mathcal{M} \to \mathcal{N}$ is a diffeomorphism. ♠

Proof of Statement 3: By symmetry it suffices to consider the map $E : \partial_0 M_+ \to \partial_0 N_+$. By definition, $E(\partial_0 M_+) = \partial_0 N_+$. This map is surjective by definition, injective by Corollary 2.10, and a local diffeomorphism by the Reciprocity Lemma. Hence $E : \partial_0 M_+ \to \partial_0 N_+$ is a diffeomorphism. ♠

4.3 Proof of The Sphere Theorem

Let S_L denote the sphere of radius L centered at the origin in $\boldsymbol{R}^3$. Let $\mathcal{S}_L$ denote the metric sphere of radius L centered at the origin of Sol. When $L < \pi\sqrt{2}$, we have $S_L \subset M$ and so $E : S_L \to \mathcal{S}_L$ is a diffeomorphism.

Let $T = \{(x, y, 0)|\ |x| = |y| = \pi\}$. When $L = \pi\sqrt{2}$, we have $S_L \subset M \cup T$. The map E is a homeomorphism when restricted to $M \cup T$ and the identity on T. Hence $\mathcal{S}_L = E(S_L)$ is a topological sphere. Since E is smooth on M, we see that $\mathcal{S}_L$ is smooth away from the 4 points of T.

Now we get to the interesting case. Let $L > \pi\sqrt{2}$. Define

$$S'_L = S_L \cap (M \cup \partial M). \tag{35}$$

The space S'_L is a 4-holed sphere. The boundary $\partial S'$ consists of 4 loops, each contained in ∂M, each homothetic to the loop level set of period L, each having holonomy invariant H_L. It follows from the Cut Locus Theorem that $\mathcal{S}_L = E(S'_L)$ and that E is a diffeomorphism when restricted to $S'_L - \partial S'_L$. On $\partial S'_L = S'_L \cap \partial M$, the map E is a 2-to-1 folding map which identifies partner points within each component. Thus, we see that $\mathcal{S}_L$ is obtained from a 4-holed sphere by gluing together each boundary component (to itself) in a 2-to-1 fashion. This reveals $\mathcal{S}_L$ to be a topological sphere. Also, $\mathcal{S}_L$ is smooth away from $E(\partial S'_L)$. This latter set lies in the union of 4 planar arcs satisfying $Z = 0$ and $|XY| = H_L^2$.

4.4 Proof of The Main Theorem

Let $V = (x, y, z)$ be a vector such that $V/\|V\|$ lies in a loop level set. For our formula we will take $x, y > 0$. The other cases follow from symmetry. If we define the quantity a by the formula

$$a^2 = \frac{xy}{\|V\|^2} = \frac{xy}{x^2 + y^2 + z^2}, \tag{36}$$

then $V/\|V\|$ lies in the same loop level set as $U_a = (a, a, \sqrt{1 - 2a^2})$. By the Cut Locus Theorem, V corresponds to a distance minimizing geodesic if and only if $\|V\| \leq L_a$, the period of the loop level set containing U_a. So, by Equation 21, and the Cut Locus Theorem, V correponds to a distance minimizing geodesic if and only if

$$\pi \geq \|V\| \times \mathrm{AGM}(a, \frac{1}{2}\sqrt{1 + 2a^2}) = \mu(V).$$

This completes the proof of the Main Theorem.

5 Technical Calculations

5.1 The Structure Field

In this section we derive Equation 11. The derivation is a bit different from the one on [**G**, pp 62-65]. Let $\{e_1, e_2, e_3\}$ denote the standard Euclidean orthonormal basis. Let E_j be the left invariant vector field which agrees with e_j at $(0,0,0)$. The triple $\{E_1, E_2, E_3\}$ is a left-invariant orthonormal framing of Sol. If we express the derivative γ' of a unit speed geodesic γ in terms of our left-invariant framing, namely

$$\gamma'(t) = \sum u_i(t) E_i,$$

then Equation 11 describes the evolution of the coefficients. For convenience, we have set $x(t) = u_1(t)$ and $y(t) = u_2(t)$ and $z(t) = u_3(t)$.

Let ∇ denote the covariant derivative for Sol. The fact that γ is a geodesic means that the covariant derivative of γ' along γ vanishes. That is,

$$0 = \nabla_{\gamma'}(\gamma') = \sum_i \frac{du_i}{dt} E_i + \sum_{ij} u_i u_j \nabla_{E_j} E_i. \tag{37}$$

Parallel translation along any curve contained in a totally geodesic plane Π preserves the unit normals to Π along that curve, and thus the covariant derivative of that unit normal along the curve vanishes. Hence $\nabla_{E_j} E_i = 0$ for $(j,i) = (1,2), (2,1), (3,1), (3,2)$. Also, $\nabla_{E_3} E_3 = 0$ because the curves integral to E_3 are geodesics. Below we will show that

$$\nabla_{E_1} E_1 = +E_3, \qquad \nabla_{E_1} E_3 = -E_1, \qquad \nabla_{E_2} E_2 = -E_3, \qquad \nabla_{E_2} E_3 = +E_2. \tag{38}$$

Plugging all this information into Equation 37, we get

$$0 = \left(\frac{du_1}{dt} - u_1 u_3\right) E_1 + \left(\frac{du_2}{dt} + u_2 u_3\right) E_2 + \left(\frac{du_3}{dt} + u_1^2 - u_2^2\right) E_3. \tag{39}$$

This is equivalant to Equation 11.

It only remains to establish Equation 38. Since E_1, E_3 are parallel to the totally geodesic plane $x_2 = 0$ and form an orthonormal framing of this plane, and since parallel translation along the curves integral to E_1 is an isometry, there is some constant λ such that $\nabla_{E_1}E_1 = \lambda E_3$ and $\nabla_{E_1}E_3 = -\lambda E_1$. By left invariance, we have $\lambda = \Gamma^3_{11}(0,0,0)$, the Christoffel symbol with respect to $\{e_1, e_2, e_3)$, evaluated at $(0,0,0)$. Let g^{ij} be the (ij)th entry of g^{-1}. Using the facts that, at $(0,0,0)$,

$$g^{31} = 0, \qquad g^{32} = 0, \qquad g^{33} = 1, \qquad \frac{dg_{1i}}{dx_1} = 0, \qquad \frac{dg_{11}}{dx_3} = -2,$$

we have

$$\Gamma^3_{11}(0,0,0) = \frac{1}{2}\sum_{i=1}^{3} g^{3i}\left(\frac{dg_{1i}}{dx_1} + \frac{dg_{1i}}{dx_1} - \frac{dg_{11}}{dx_i}\right) = 1.$$

This deals with the first two equalities in Equation 38. The last two have similar treatments, and indeed follow from the first two and the existence of the isometry $(x_1, x_2, x_3) \to (x_2, x_1, -x_3)$.

5.2 Grayson's Cylinders and Period Formula

Let $U_a = (a, a, \sqrt{1-2a^2})$ and let L_a be the period of the loop level set containing U_a. The following result bundles together some of the results on [**G**, pp 67-75].

Proposition 5.1 *When $a \in (0, \sqrt{2}/2)$ and $r \in \boldsymbol{R}$, we have $E(rU_a) \in C_a$, where*

$$C_a = \{(x,y,z) | w^2 + \cosh 2z = \frac{1}{2a^2}\}, \qquad w = \frac{x-y}{\sqrt{2}}.$$

The geodesic segment corresponding to the perfect vector L_aU_a winds once around C_a. Moreover,

$$L_a = \int_a^{t_a} \frac{4dt}{\sqrt{1-2a^2\cosh 2t}}, \qquad t_a = \frac{1}{2}\cosh^{-1}\left(\frac{1}{2a^2}\right). \tag{40}$$

One can deduce from symmetry and from Proposition 5.1 that every typical geodesic lies on some cylinder isometric to C_a, and that a typical geodesic segment is small, perfect, or large according as it winds less than, equal to, or greater than once around the cylinder that contains it.

Our one remaining goal is to prove Equation 21. For this we don't need Proposition 5.1 but we do need Equation 40. For the sake of completeness, we essentially repeat the proof given on [**G**, p 68]. In our derivation, the symbol $\cdot$ denotes a quantity we don't need to compute.

Lemma 5.2 *Equation 40 is true.*

Proof: Let u denote the flow line for the structure field Σ corresponding to the vector $\frac{1}{4}L_aU_a$. Referring to Equation 11 the flowline u starts at U_a and ends the first time it reaches Π, the plane $Z=0$. The loop level sets are level sets of the function $F(x,y,z)=xy$, and they lie on the unit sphere. Hence

$$u=S([0,t_a]), \qquad S(t)=(ae^t, ae^{-t}, \sqrt{1-2a^2\cosh 2t}). \tag{41}$$

Referring to Equation 11, the two quantities $S'(t)$ and $\Sigma(S(t))$ are scalar multiples. Setting $S(t)=(x_t,\cdot,z_t)$, and noting that $dx_t/dt=x_t$, we have

$$S'(t)=(x_t,\cdot,\cdot)=(1/z_t)\times(x_tz_t,\cdot,\cdot)=(1/z_t)\times\Sigma(S(t)). \tag{42}$$

Let γ be the geodesic corresponding to u. Let $\gamma(t)$ be the point of γ corresponding to $S(t)$. By definition, the unit tangent field $\boldsymbol{T}(t)$ along $\gamma(t)$ lies in the same left invariant vector field as $S(t)$. By the Chain Rule and Equation 42,

$$\frac{d\gamma}{dt}(t)=\frac{1}{z_t}\boldsymbol{T}(t). \tag{43}$$

By symmetry and by definition, L_a is 4 times the length of the geodesic segment γ just considered. Noting that $\|\boldsymbol{T}(t)\|=1$, and integrating Equation 43, we have

$$L_a=4\ \text{Length}(\gamma)=4\int_a^{t_a}\left\|\frac{d\gamma}{dt}\right\|dt=4\int_a^{t_a}\frac{dt}{z_t}=\int_a^{t_a}\frac{4dt}{\sqrt{1-2a^2\cosh(t)}}.$$

This completes the proof ♠

5.3 The AGM Period Formula

Now we manipulate Equation 40 until it is equivalent to Equation 21. Using the relations

$$\cosh(2t)=2\sinh^2(t)+1, \qquad m=\frac{1-2a^2}{1+2a^2}, \qquad \mu=\sqrt{\frac{m}{1-m}}=\frac{\sqrt{1-2a^2}}{2a},$$

we see that Equation 40 is equivalent to the following:

$$L_a=\frac{4}{\sqrt{1+2a^2}}\times I_a, \qquad I_a=\frac{1}{\sqrt{m}}\int_a^{\sinh^{-1}(\mu)}\frac{dt}{\sqrt{1-(\sinh(t)/\mu)^2}}. \tag{44}$$

To get further we relate this expression to something more classical. Let

$$\mathcal{K}(m) = \mathcal{F}(\pi/2, m), \qquad \mathcal{F}(\phi, m) := \int_a^\phi \frac{d\theta}{\sqrt{1 - m\sin^2\theta}}. \tag{45}$$

These quantities respectively are called the complete and incomplete elliptic integrals of the first kind.

Lemma 5.3 $I_a = \mathcal{K}(m)$.

Proof: This is related to Equation 19.7.7 in the Electronic Handbook of Mathematical Functions. The substitution

$$u = \tan^{-1}\sinh(t), \qquad du = dt/\cosh(t) = dt\cos(u)$$

gives

$$I_a = \frac{1}{\sqrt{m}} \times \mathcal{F}(\tan^{-1}(\mu), \frac{1}{m}).$$

The substitution $t = \sin(\theta)$ gives

$$I_a = \frac{1}{\sqrt{m}} \int_a^{\sqrt{m}} \frac{dt}{\sqrt{(1-t^2)(1-t^2/m)}}, \qquad \mathcal{K}(m) = \int_a^1 \frac{dt}{\sqrt{(1-t^2)(1-mt^2)}}.$$

The substitution $u = t/\sqrt{m}$ converts I_a into $\mathcal{K}(m)$. ♠

See e.g. [**BB**] for a proof of the following classic identity:

$$\mathcal{K}(m) = \frac{\pi/2}{\mathrm{AGM}(\sqrt{1-m}, 1)}, \qquad m \in (0,1). \tag{46}$$

Combining Lemma 5.3 and Equation 46, we get Equation 21:

$$L_a = \frac{4}{\sqrt{1+2a^2}} \times \frac{\pi/2}{\mathrm{AGM}(1, \sqrt{1-m})} = \frac{\pi}{\mathrm{AGM}(a, \frac{1}{2}\sqrt{1+2a^2})}.$$

6 References

[**A**] V. I. Arnold, *Sur la géométrie différentielle des groupes de Lie de dimension infinie et ses applications à l'hydrodynamique des fluides parfaits.* Ann. Inst. Fourier Grenoble, (1966).

[**AK**] V. I. Arnold and B. Khesin, *Topological Methods in Hydrodynamics*, Applied Mathematical Sciences, Volume 125, Springer (1998)

[**B**] N. Brady, *Sol Geometry Groups are not Asynchronously Automatic*, Proceedings of the L.M.S., 2016 vol 83, issue 1 pp 93-119

[**BB**] J. M. Borwein and P. B. Borwein, *Pi and the AGM*, Monographies et Études de la Société Mathématique du Canada, John Wiley and Sons, Toronto (1987)

[**BS**] A. Bölcskei and B. Szilágyi, *Frenet Formulas and Geodesics in Sol Geometry*, Beiträge Algebra Geom. 48, no. 2, 411-421, (2007).

[**BT**], A. V. Bolsinov and I. A. Taimanov, *Integrable geodesic flow with positive topological entropy*, Invent. Math. **140**, 639-650 (2000)

[**CMST**] R. Coulon, E. A. Matsumoto, H. Segerman, S. Trettel, *Noneuclidean virtual reality IV: Sol*, math arXiv 2002.00513 (2020)

[**EFW**] D. Fisher, A. Eskin, K. Whyte, *Coarse differentiation of quasi-isometries II: rigidity for Sol and Lamplighter groups*, Annals of Mathematics 176, no. 1 (2012) pp 221-260

[**G**], M. Grayson, *Geometry and Growth in Three Dimensions*, Ph.D. Thesis, Princeton University (1983).

[**K**] S. Kim, *The ideal boundary of the Sol group*, J. Math Kyoto Univ 45-2 (2005) pp 257-263

[**KN**] S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry, Volume 2*, Wiley Classics Library, 1969.

[**LM**] R. López and M. I. Muntaenu, *Surfaces with constant curvature in Sol geometry*, Differential Geometry and its applications (2011)

[**S**] R. E. Schwartz, *Java Program for Sol*, download (in 2019) from http://www.math.brown.edu/∼res/Java/SOL.tar

[**T**] M. Troyanov, *L'horizon de SOL*, Exposition. Math. 16, no. 5, 441-479, (1998).

[**Th**] W. P. Thurston, *The Geometry and Topology of Three Manifolds*, Princeton University Notes (1978). (See
http://library.msri.org/books/gt3m/PDF/Thurston-gt3m.pdf
for an updated online version.)

[**W**] S. Wolfram, *The Mathematica Book, 4th Edition*, Wolfram Media and Cambridge University Press (1999).