The minimal volume of the plane.

B. H. Bowditch
Faculty of Mathematical Studies, University of Southampton,
Highfield, Southampton, SO9 5NH, Great Britain.
email: bhb@maths.soton.ac.uk

Abstract.

We give an account of the minimal volume of the plane, as defined by Gromov, and first computed by Bavard and Pansu. We also describe some related geometric inequalities.

0. Introduction.

The "minimal volume", as defined by Gromov [8], is an invariant associated to a smooth manifold. Gromov observed that the minimal volume of the plane, $\mathbf{R}^2$, is at most $2\pi(1+\sqrt{2})$, and conjectured that it was, in fact, equal to this. This was shown to be the case by Bavard and Pansu [1]. In this paper, we give some exposition of the geometry underlying this result, and offer a somewhat different proof of Bavard and Pansu's theorem.

Most of the material for this paper was prepared at the University of Melbourne, with the support of an A.R.C. Fellowship. It was written at the University of Aberdeen under an S.E.R.C. Research Assistantship. I would like to thank both institutions for their hospitality.

1. The minimal volume of a manifold.

We give a general outline of the ideas behind the notion of minimal volume. Beyond the main definition below, this section is not directly relevant to the rest of the paper.

Let M be a smooth riemannian manifold (without boundary). In [8], Gromov defined the *minimal volume* of M, $\min \mathrm{vol}(M)$, to be the infinum of all volumes $\mathrm{vol}(M,g)$ as g ranges over the set, $\mathcal{G}(M)$, of all complete riemannian metrics on M having sectional curvatures between -1 and 1. Thus, $\min \mathrm{vol}(M)$ might be zero, infinite, or finite and positive. In general, it is an interesting and difficult problem to compute the minimal volume, or even to decide between these three alternatives.

Suppose, for example, that M admits a finite-volume hyperbolic metric, g_{hyp} (constant curvature -1). Then it is conjectured that this metric attains the minimum volume, i.e. $\min \mathrm{vol}(M) = \mathrm{vol}(M, g_{hyp})$. This seems to be only known in dimension 2 (as we describe in Section 2). For some partial results in higher dimensions, see [7].

On the other hand, there are many examples where the minimal volume is zero. This is clearly true of tori (or manifolds finitely covered by tori)—just scale any flat metric so that the volume tends to zero. Also, an example of Berger shows how a 3-sphere may be "collapsed" down to a 2-sphere, starting with the spherical metric (constant curvature 1), and then uniformly shrinking the fibres of a Hopf fibration [4]. This may be done while keeping the curvature bounded, and so $\min \mathrm{vol}(S^3) = 0$. More generally, a similar argument may be applied to (orientable) Seifert-fibred 3-manifolds (those foliated by circles), and manifolds obtained by gluing together Seifert-fibred manifolds along 2-toroidal boundary components. In the latter case, the fibrations need not be compatible on different sides of the 2-tori. In fact, if we introduce enough 2-tori into such a 3-manifold, we can assume that all complementary pieces are homeomorphic to a surface times a circle.

Thus for example, a more direct way to see that $\min \mathrm{vol}(S^3) = 0$ is as follows. Take a 2-disc, and put on it any riemannian metric having curvature between -1 and 1, having area at most, say, 100, and having geodesic boundary of length ϵ. Now take a product with a circle of length ϵ to obtain a solid torus, and glue together two such solid tori so as to obtain a 3-sphere. As ϵ tends to 0, the 3-sphere "collapses", and the volume tends to 0. A similar, but more complicated argument [8] shows that $\min \mathrm{vol}(\mathbf{R}^3) = 0$. (This involves an "infinite" construction. In contrast to the 2-dimensional situation, $\mathbf{R}^3$ seems not to admit a finite-volume metric of bounded curvature and "cusp-like" end—c.f. Section 5.)

The 3-manifolds we have just described are said to admit "F-structures (of positive rank)". The definition of an F-structure is somewhat involved, but to first approximation it can be thought of as a decomposition of M into tori (or manifolds finitely covered by tori) of varying dimensions. Cheeger and Gromov [5,6] show that if $\min \mathrm{vol}(M) = 0$, then M admits an F-structure. Thus we imagine M collapsing down along the tori. This theorem holds in any dimension. There is a converse in dimensions 2 and 3.

In dimension 3, this is related to Thurston's geometrisation conjecture [16]. Suppose M^3 is a closed orientable 3-manifold with $\pi_2(M^3) = 0$. Then one can define the (possibly empty) "characteristic submanifold", M_0^3, of M^3 [10,11]. This is, in some sense, the "maximal" F-structured submanifold, and is well defined up to isotopy. If we take a volume-minimising sequence of metrics in $\mathcal{G}(M^3)$, then we imagine M^3 collapsing down on M_0^3, while the remainder $M^3 \setminus M_0^3$ becomes hyperbolic. For example, if M_0^3 consists of a single 2-torus, then M^3 gets stretched out along a tube which, in the limit, turns into two toroidal cusps. However, most of this remains in the realm of conjecture.

In general, in order to compensate for collapsing, it may be necessary to stretch M in an "orthogonal" direction. Thus the diameter may tend to ∞, and the volume need not tend to 0. Indeed, there are examples, in higher dimensions, of F-structured manifolds (which admit global collapses) for which the minimal volume is non-zero [5].

The question we have been considering can be reinterpreted as minimising the L^∞-norm of sectional curvature over all complete riemannian metrics on M of a fixed volume. One could similarly consider minimising the L^p-norm for some $p < \infty$. This is, in many ways, qualitatively similar, but seems more amenable to analytic methods. See for example [17], and the references contained therein, for some of the fruits of this approach.

2. Surfaces.

Since the higher-dimensional situation is evidently quite complicated, let's restrict attention to dimension 2. Most of this can be dismissed by applying the Gauß-Bonnet Theorem.

Theorem 2.1 : *Suppose M^2 is a topologically finite surface, not diffeomorphic to $\mathbf{R}^2$. Then,*

$$\min \mathrm{vol}(M^2) = 2\pi|\chi(M^2)|,$$

where $\chi(M^2)$ is the Euler characteristic of M^2.

Proof : If $g \in \mathcal{G}(M^2)$, then the Gauß-Bonnet Theorem tells us that $\mathrm{area}(M^2, g) \geq 2\pi|\chi(M^2)|$. (If M^2 is non-compact, we need to observe that if $\mathrm{area}(M^2, g) < \infty$, then the ends of (M^2, g) are "cusp-like" in the sense that they can be cut off by arbitrarily short curves of bounded outward curvature. The arguments of Section 5 effectively show this. We can thus ignore the boundary term of the Gauß-Bonnet Theorem.) Now if $\chi(M^2) \neq 0$, then M^2 admits a metric of constant curvature ± 1 and area $\pm 2\pi\chi(M^2)$, so this bound is attained. We have already observed that the flat torus and klein bottle can be scaled to have arbitrarily small area, and it is a simple exercise to construct complete metrics on the annulus and möbius band of arbitrarily small area and bounded curvature. ◊

This leaves only the plane, $\mathbf{R}^2$, unaccounted for. In this case, Gromov suggested a candidate for achieving the minimal volume. It can be described as a surface of revolution in euclidean 3-space. Imagine taking a round sphere of radius 1, and a pseudosphere (of curvature -1), both embedded in 3-space. Now bring them together until they touch along some circle C (of length $\pi\sqrt{2}$). Now cut along C, and glue together the spherical cap and the unbounded portion of the pseudosphere (Figure 1). The resulting surface has area $2\pi(1 + \sqrt{2})$. Thus, Gromov conjectured that $\min \mathrm{vol}(\mathbf{R}^2) = 2\pi(1 + \sqrt{2})$. This was shown to be indeed the case by Bavard and Pansu [1].

Their proof proceeds by analysing the "isoperimetric profile" of a riemannian metric $g \in \mathcal{G}(\mathbf{R}^2)$. For this they appeal to compactness and regularity results for integral currents in dimension 2. One should imagine blowing larger and larger bubbles in $(\mathbf{R}^2, g)$, until they expand out the end of $\mathbf{R}^2$, eventually filling out the entire space. The isoperimetric profile measures how the length of the "surface" of a bubble varies with the area enclosed.

Here, we shall do the opposite. We start out the end, and work our way into the interior. Our approach avoids any direct use of geometric measure theory. We shall appeal to the Spherical Isoperimetric Inequality (Section 3). Admittedly, most proofs of this use some high-powered machinery, though if one is prepared to ignore some technical complications, one can give a fairly simple argument (see [14]).

Since it involves no extra work, we shall prove the more general result that if we constrain the sectional curvatures to lie between $-\kappa^2$ and 1, for some $\kappa > 0$, then we obtain a minimal area of $2\pi\left(1 + \frac{\sqrt{1+\kappa^2}}{\kappa}\right)$. This arises similarly, by connecting a spherical cap to a pseudosphere, scaled to have constant curvature $-\kappa^2$, along a circle of length $2\pi/\sqrt{1+\kappa^2}$.

We have been a bit vague about the category of metrics with which we are working. One can certainly make sense of curvature bounds for C^1 riemannian metrics, and indeed the extremal metrics described are only C^1. However, it will be technically convenient to assume that our metrics are at least C^2.

The theorem of Bavard and Pansu (in the case $\kappa = 1$) states:

Theorem 2.2 : *If $\kappa > 0$, and g is any complete riemannian metric on $\mathbf{R}^2$ having curvature between $-\kappa^2$ and 1, then*

$$\operatorname{area}(\mathbf{R}^2, g) \geq 2\pi \left(1 + \frac{\sqrt{1+\kappa^2}}{\kappa} \right).$$

In Section 8, we decribe a similar result for riemannian metrics on the disc with geodesic boundary (Theorem 8.1). This can also be interpreted in terms of metrics on the 2-sphere with injectivity radius less than π (Theorem 8.2). See also [1, Théorème 10] for an account of this.

3. The spherical isoperimetric inequality.

A key ingredient in the argument is the Spherical Isoperimetric Inequality, which we quote:

Theorem 3.1 : *Suppose that g is a riemannian metric on the disc, D, such that ∂D is rectifiable, and such that g has curvature ≤ 1 on the interior. Then*

$$L^2 \geq A(4\pi - A)$$

where $L = \operatorname{length}(\partial D)$ and $A = \operatorname{area}(D)$. $\diamondsuit$

Thus, for fixed $A < 4\pi$, the shortest possible boundary is obtained by taking a round circle (of length $\sqrt{A(4\pi - A)}$) bounding a spherical cap of area A.

The history of this result can be traced back a long way. Versions of the theorem, similar to that stated, were proven by Bol and Fiala in the 1940s. We refer to Osserman's articles [14,15] for expositions of this and related inequalities.

If we view the inequality as a question of spanning a curve of fixed length $L < 2\pi$ by a disc of curvature ≤ 1, then the following dichotomy arises. Either $\operatorname{area}(D) \leq A_-(L)$ or else $\operatorname{area}(D) \geq A_+(L)$, where $A_\pm(L) = 2\pi \pm \sqrt{4\pi^2 - L^2}$. Thus $A_-(L)$ and $A_+(L)$ are, respectively, the areas of the small and large spherical caps bounded by a round circle of length L on the spherical 2-sphere. Note that $A_-(L) < L < 2\pi < A_+(L)$.

One way to express this dichotomy is as follows. We say that a rectifiable closed curve of length $< 2\pi$ is *shrinkable* if it can be homotoped to a point in such a way that all the intermediate curves are also rectifiable and of length $< 2\pi$. It turns out that $\operatorname{area}(D) \leq A_-(L)$ if and only if ∂D is shrinkable in D (see [2]). We shall not formally need the notion of shrinkability in the proof, though it seems an intuitively useful idea to keep in mind.

Suppose now that g is a complete riemannian metric of curvature ≤ 1 on $\mathbf{R}^2$, and that $\gamma \subseteq \mathbf{R}^2$ is a simple closed rectifiable curve of length $< 2\pi$. Then γ bounds a disc $D(\gamma)$ and an annulus $R(\gamma)$.

Definition : We say that γ is *essential* if $\text{area}(D(\gamma)) \geq 2\pi$.

We see that equivalent ways to say that γ is essential are:
(1) $\text{area}(D(\gamma)) \geq A_+(L)$,
(2) $\text{area}(D(\gamma)) > A_-(L)$,
(3) γ is not shrinkable in $D(\gamma)$,
and, in fact [2],
(4) γ is not shrinkable in $\mathbf{R}^2$.
We only formally need the equivalence with (1) and (2).

4. The idea of the proof.

Suppose g is a compete riemannian metric on $\mathbf{R}^2$ with curvature between $-\kappa^2$ and 1, and with $\text{area}(\mathbf{R}^2, g) < \infty$.

The first step is to show that the end of $(\mathbf{R}^2, g)$ is "cusp-like" (Section 5). In particular, we define a *horofunction* $h : \mathbf{R}^2 \longrightarrow [0, \infty)$. This is a proper 1-lipshitz map, with the property that for any $x \in \mathbf{R}^2$, there is a geodesic ray $\alpha_x : [0, \infty) \longrightarrow \mathbf{R}^2$ with $\alpha_x(0) = x$ and $h(\alpha_x(t)) = h(x) + t$ for all $t \in [0, \infty)$.

Definition : A *horocycle at level* t is the boundary of a component of $h^{-1}[0, t)$.

Thus a horocycle is a closed subset of the level set $h^{-1}t$.

Lemma 4.1 (Section 6) : *A horocycle is a Jordan curve.*

Thus, a horocycle, γ, bounds a disc $D(\gamma)$ (the closure of a component of $h^{-1}[0, t)$) and an annulus $R(\gamma)$. In fact:

Lemma 4.2 (Section 6) : *A horocycle γ is rectifiable, and* $\text{length}(\gamma) \leq \kappa \text{area}(R(\gamma))$.

This only makes essential use of the lower curvature bound $-\kappa^2$.

Now, if γ is an essential horocycle (Section 3) of length $L < 2\pi$, then (from the upper curvature bound) we obtain

$$\text{area}(D(\gamma)) \geq A_+(L).$$

Thus,

$$\text{area}(\mathbf{R}^2, g) \geq \frac{1}{\kappa} L + A_+(L).$$

In fact, we claim:

Proposition 4.3 (Section 7) : *Given any $L \in (0, 2\pi)$, there is some essential horocycle of length L.*

On substituting $L = 2\pi/\sqrt{1+\kappa^2}$, we arrive at the result $\text{area}(\mathbf{R}^2, g) \geq 2\pi \left(1 + \frac{\sqrt{1+\kappa^2}}{\kappa}\right)$, and Theorem 2.2 is proved.

We shall prove Proposition 4.3 by applying an intermediate value theorem to a certain function, f, of the parameter $t \in [0, \infty)$. Note that for any t, there can only be finitely many essential horocycles at level t. Define $f(t)$ to be the length of the longest essential horocycle. If there is no essential horocycle, set $f(t) = 2\pi$. Clearly, $f(0) = 2\pi$. From the cusp-like nature of the end we obtain:

Lemma 4.4 (Section 7) : *As $t \to \infty$, $f(t) \to 0$.*

As t decreases, the growth of f is bounded by an exponential function coming from the lower curvature bound. However, f need not be continuous. Its value may fall suddenly due typically to a horocycle dividing into two or more "components". Even so, f must eventually attain the value of 2π, and hence any intermediate value. Intuitively, this can be thought of in terms of shrinkability. More formally, we need to verify that when an essential horocycle splits up, at least one of the components arising is essential.

To give the idea, suppose that at some "critical" time t, the level set $h^{-1}t$ contains a "figure-of-eight" curve which represents a "horocycle" $\gamma = \gamma_1 \cup \gamma_2$ splitting into two horocycles γ_1 and γ_2. We can think of $D(\gamma)$ as the union of $D(\gamma_1)$ and $D(\gamma_2)$. Let $A_i = \text{area}(D(\gamma_i))$ and $L_i = \text{length}(\gamma_i)$. If γ_1 and γ_2 are inessential, we have

$$\text{area}(D(\gamma)) = A_1 + A_2 \leq A_-(L_1) + A_-(L_2) \leq L_1 + L_2 = \text{length}(\gamma) < 2\pi,$$

and so γ is inessential. Of course, the general situation may be much more complicated than this, so we will have to approach the matter more formally.

The property of f that we require is:

Lemma 4.5 (Section 7) : *If $t, u \in [0, \infty)$, then $f(u - t) \leq e^{\kappa t} f(u)$.*

5. The horofunction.

Suppose g is a complete riemannian metric on $\mathbf{R}^2$, and d is the induced path-metric.

Definition : By a *geodesic ray*, α, based at $x \in \mathbf{R}^2$, we mean a path $\alpha : [0, \infty) \longrightarrow \mathbf{R}^2$ such that $\alpha(0) = x$, and $d(\alpha(t), \alpha(u)) = |t - u|$ for all $t, u \in [0, \infty)$.

Such a ray must exist for all $x \in \mathbf{R}^2$, and we write α_x for some choice of ray based at x. Thus α_x need not vary continuously in x. However, given any $x, y \in \mathbf{R}^2$, either the images of α_x and α_y are disjoint, or else one is contained in the other.

Let's now assume that $(\mathbf{R}^2, g)$ has curvature ≤ 1, and finite area.

Let α be any geodesic ray. Clearly, the injectivity radius at $\alpha(t)$ must tend to 0 as $t \to \infty$. For $n \in \mathbf{N}$, let β_n be a shortest non-constant geodesic loop based at $\alpha(n)$. Thus β_n is a simple closed curve, and bounds a disc $D_n \subseteq \mathbf{R}^2$. By Gauß-Bonnet, $\mathrm{area}(D_n) \geq \pi$. Since $\mathrm{area}(\mathbf{R}^2) < \infty$, we can suppose that $D_m \subseteq D_n$ whenever $m \leq n$. We conclude:

Lemma 5.1 : *There is an exhaustion of $\mathbf{R}^2$ by a nested sequence of discs $(D_n)_{n \in \mathbf{N}}$ such that* $\mathrm{length}(\partial D_n) \to 0$ *as* $n \to \infty$. $\diamondsuit$

Now, given any $x \in \mathbf{R}^2$, set

$$h(x) = \lim_{t \to \infty} (t - d(x, \alpha(t))) + \text{constant}.$$

Lemma 5.1 shows that this limit exists. The constant is chosen so that we can assume, for convenience that $h(x) \geq 0$ for all $x \in \mathbf{R}^2$. We see that h is in fact a 1-lipshitz proper map of $\mathbf{R}^2$ to $[0, \infty)$. Moreover, for any $x \in X$ and $t \in [0, \infty)$, we have $h(\alpha_x(t)) = h(x) + t$.

We refer to a h as a *horofunction* on $\mathbf{R}^2$. (Any two horofunctions defined in such a way will differ by an additive constant.)

6. Horocycles.

This is a somewhat technical section. We shall appeal to some basic planar topology such as the Jordan Curve Theorem (see [13]), as well as some standard comparison theorems of riemannian geometry such as those of Toponogov (see [4] or [12]). We also use (implicitly) the fact that a path-connected hausdorff space is arc-connected.

If K is a compact connected metric space such that $K \setminus \{x, y\}$ is disconnected for any pair of distint points $x, y \in K$, then either K is a point, or it is homeomorphic to a circle [9, Theorem 2-28]. We arrive at the following characterisation of Jordan curves.

Lemma 6.1 : *Suppose that $K \subseteq \mathbf{R}^2$ is compact and that $\mathbf{R}^2 \setminus K$ has precisely two components U_1 and U_2. Suppose that every point of K is accessible from both U_1 and U_2. Then, K is a Jordan curve (i.e. homeomorphic to a circle).*

To say that $x \in K$ is *accessible* from U_i means that there is a continuous path $\beta : [0, 1] \longrightarrow \mathbf{R}^2$ such that $\beta(0) = x$ and $\beta((0, 1]) \subseteq U_i$.

Proof : Given any distinct $x, y \in K$, there is a Jordan curve γ such that $\gamma \cap K = \{x, y\}$ and $\gamma \cap U_i \neq \emptyset$ for $i = 1, 2$. Apply the Jordan Curve Theorem. $\diamondsuit$

Now let's suppose that g is a complete, finite-area riemannian metric on $\mathbf{R}^2$ with curvature between $-\kappa^2$ and 1. Let d be the induced path metric. We use $N(x, r)$ to denote the closed metric r-neighbourhood of x.

If $x, y \in h^{-1}t$ are distinct, then the rays α_x and α_y are disjoint. If x and y are close enough together, they will be joined by a unique geodesic β, and the path $\alpha_x \cup \beta \cup \alpha_y$ will be a properly embedded line. This line divides $\mathbf{R}^2$ into two closed *triangular regions* T_1 and T_2.

Proof of Lemma 4.1 : We aim the show that each horocycle at level t is a Jordan curve. Let $F = h^{-1}[t,\infty)$. Thus F is a closed connected neighbourhood of infinity in $\mathbf{R}^2$, and so a component, U, of $\mathbf{R}^2 \setminus F$ is an open disc. We want to see that its closure, $\bar{U}$, is a closed disc, or equivalently, that ∂U is a Jordan curve.

We first show that $V = \mathbf{R}^2 \setminus \bar{U}$ is connected, by joining every point $x \in V$ to infinity by a path in V. If $h(x) \geq t$, then α_x will do the trick, so we suppose that $h(x) = u < t$. We can assume that there is some point $y \in V \cap h^{-1}u \setminus \{x\}$ close to x. (Move x a bit if necessary.) Let T_1 and T_2 be the triangular regions described above. Now neither α_x nor α_y meets U. Thus U and hence $\bar{U}$ lies inside one of these regions, say T_1. We can now connect x to infinity by a path in T_2. This shows that V is connected.

Now $\mathbf{R}^2 \setminus \partial U = U \cup V$, and clearly every point $x \in \partial U \subseteq h^{-1}t$ is accessible (by a geodesic segment) from V. Thus, by Lemma 6.1, it suffices to show that x is accessible from U. We claim, in fact, that for all $\epsilon > 0$ sufficiently small (smaller than the convexity radius), any two points of $U \cap N(x,\epsilon/3)$ can be joined by a path in $U \cap N(x,\epsilon)$. Accessibility then follows by taking a sequence of points of U converging to x.

To prove the claim, note first that if $p \in F \cap \partial N(x,\epsilon)$, then the ray α_p does not meet $N(x,\epsilon/3)$ (otherwise $h(x) > h(p) \geq t$). Suppose then that $y, z \in U \cap N(x,\epsilon/3)$. Join y to z by an arc β in U. Suppose $J \subseteq \beta$ is a component of $\beta \setminus N(x,\epsilon/3)$. Join the endpoints of J by an arc in $N(x,\epsilon/3)$ to form a Jordan curve C which bounds a disc D. Note that if $p \in F \cap \partial N(x,\epsilon)$, then α_p cannot meet either J or $N(x,\epsilon/3)$, and hence cannot meet C. It follows that $D \cap F \cap \partial N(x,\epsilon) = \emptyset$. Thus $D \cap \partial N(x,\epsilon) \subseteq U$ (since $\partial U \subseteq F$). We can now replace (by induction on complexity) all those parts of J which venture outside $N(x,\epsilon)$ by segments of $\partial N(x,\epsilon)$ lying in U. We eventually end up joining y to z by a path in $U \cap N(x,\epsilon)$, as claimed. $\Diamond$

Lemma 6.2 : *Given $u, t_0 \in [0,\infty)$ and $\mu > 1$, there is some $\delta > 0$ such that if $x, y \in h^{-1}u$ and $d(x,y) \leq \delta$, and $t \in [0,t_0]$, then $d(\alpha_x(t),\alpha_y(t)) \geq e^{-\mu\kappa t} d(x,y)$.*

Proof : Choose any $\mu' \in (1,\mu)$. Choose $t_1 > t_0$, large, depending on t_0, μ, μ', κ. Choose $\eta > 0$, small, depending on μ', κ. Choose $\delta > 0$ less than the injectivity radius on $h^{-1}[u, u+t_1]$, and small, depending on η, μ', κ. The properties required of t_1, η, δ will become apparent in the course of the proof.

Suppose then that $x, y \in h^{-1}u$, and that $d(x,y) = r < \delta$. Set $j(v) = d(\alpha_x(v), \alpha_y(v))$ for $v \in [0,\infty)$. Suppose, for contradiction, that there is some $t \in [0,t_0]$ with $j(t) < re^{-\mu\kappa t}$. Thus $t > 0$. Let $[a,b]$ be the maximal subinterval of $[0,t_1]$ containing t, such that $j(v) \leq r$ for all $v \in [a,b]$. Thus, $j(a) = r$, and either $b = t_1$ or $j(b) = r$.

We chose δ less than the injectivity radius, and so, for all $v \in [a,b]$, there is a unique geodesic segment, β_v, joining $\alpha_x(v)$ to $\alpha_y(v)$. Moreover, β_v varies smoothly in v. Let $\theta(v)$ and $\phi(v)$ be the angles made by β_v with $\alpha_x([v,\infty))$ and $\alpha_y([v,\infty))$ respectively. Now, if δ is small enough, in relation to η, then (using Toponogov's comparison theorem) we have $\theta(v) \geq \frac{\pi}{2} - \eta$ and $\phi(v) \geq \frac{\pi}{2} - \eta$. Also (since j is non-increasing at a) we can deduce that $\theta(v) + \phi(v) \leq \pi + 2\eta$ (again provided that δ is sufficiently small). Thus $\theta(v), \phi(v) \in \left[\frac{\pi}{2} - \eta, \frac{\pi}{2} + \eta\right]$. Now, if we choose η and δ sufficiently small (depending on μ'), then by standard comparison arguments, we see that j must satisfy a differential inequality $\frac{d^2j}{dv^2} \leq \mu' \kappa j$ on $[a,b]$. (The infinitesimal case as $\mu' \to 1$ and $\eta, \delta \to 0$ is the Rauch Comparison Theorem.) Provided we have chosen t_1 large enough, depending on t_0, μ, μ', κ, the boundary conditions imply that $j(v) \geq re^{-\mu\kappa v}$ for all $v \in [a,b]$, in particular, for $v = t$. ♢

Continuing with the same line of argument, we deduce:

Lemma 6.3 : *Given $u \in [0,\infty)$ and $\sigma > 1$, there exists $\delta > 0$ such that if $x, y \in h^{-1}u$ and $d(x,y) < \delta$, then $d(x,y) \leq \sigma\kappa \mathrm{area}(T)$, where T is either one of the triangular regions bounded by α_x, α_y and the geodesic segment joining x to y.*

Proof : Choose $\mu > 1$ close to 1, and t_0 large, both depending on σ and κ. We apply Lemma 6.2. Let $r = d(x,y) < \delta$. In this case we can assume that $d(\alpha_x(t), \alpha_y(t))$ is small for all $t \in [0,t_0]$ (since we can assume that area(T) is small). Thus, for all such t, the angles $\theta(t)$ and $\phi(t)$, as in the proof of Lemma 6.2, can be assumed to lie between $\frac{\pi}{2} - \eta$ and $\frac{\pi}{2} + \eta$ for small $\eta > 0$. We get a lower bound for area(T) of the form $r\nu(\eta,\delta) \int_0^{t_0} e^{-\mu\kappa t} dt$, where $\nu(\eta,\delta)$ tends to 1 as η and δ tend to 0. For suitable δ, μ, t_0, and hence η and $\nu(\eta,\delta)$, we can arrange that $\nu(\eta,\delta) \int_0^{t_0} e^{-\mu\kappa t} dt \geq 1/\sigma\kappa$. ♢

Proof of Lemma 4.2 : Suppose γ is a horocycle at level u. Given $\sigma > 1$, choose δ as in Lemma 6.3. Suppose $x_1, \dots x_n$ are cyclically ordered on γ and that $d(x_i, x_{i+1}) < \delta$ for all i. Set $\alpha_i = \alpha_{x_i}$. For each i, choose the triangular region T_i bounded by α_i, α_{i+1} and the geodesic segment joining x_i to x_{i+1}, so that the interiors of all the T_i are disjoint away from γ. It is conceivable that the T_i may overlap in a δ-neighbourhood of γ, but such overlaps can be eliminated on further subdivision of γ. We can thus assume that the interiors of the T_i are disjoint. Now $\bigcup_{i=1}^n T_i \subseteq N(R(\gamma), \delta)$ and so, by Lemma 6.3,

$$\sum_{i=1}^n d(x_i, x_{i+1}) \leq \sum_{i=1}^n \sigma\kappa \mathrm{area}(T_i) \leq \sigma\kappa \mathrm{area}(N(R(\gamma)), \delta)).$$

Letting $\sigma \to 1$ and $\delta \to 0$, we deduce

$$\mathrm{length}(\gamma) \leq \kappa \mathrm{area}(R(\gamma)).$$

♢

7. The intermediate value theorem.

In this section, we define and describe the properties of the function f mentioned in Section 4, and thereby conclude the proof of Theorem 2.2.

Suppose $u \in [0,\infty)$. If there is an essential horocycle at level u, we define $f(u)$ to be the length of the longest such. Otherwise, we set $f(u) = 2\pi$. Note that $f(0) = 2\pi$.

We show (Lemma 4.4) that $f(u) \to 0$ as $u \to \infty$.

Proof of Lemma 4.4 : By Lemma 5.1, there is a compact exhaustion of $\mathbf{R}^2$ by discs D_n such that length$(\partial D_n) \to 0$. From the Spherical Isoperimetric Inequality (Section 3), we can suppose that area$(D_n) \geq 2\pi$ for all n. Given any $\epsilon > 0$, we can find n such that area$(\mathbf{R}^2 \setminus D_n) < \epsilon/\kappa$, and then $u(\epsilon)$ such that $D_n \subseteq h^{-1}[0, u(\epsilon)]$. It follows that for all $u > u(\epsilon)$, there is some horocycle γ such that $D_n \subseteq D(\gamma)$. We see that γ is the unique essential horocycle at level u, and so (by Lemma 4.2) we have $f(u) = \text{length}(\gamma) \leq \kappa\text{area}(R(\gamma)) < \epsilon$. ♢

Lemma 7.1 : *Suppose γ is a horocycle at level u, and that Γ is a set of horocycles at level $u - t$ $(t > 0)$ such that $\gamma' \subseteq D(\gamma)$ for all $\gamma' \in \Gamma$. Then*

$$\sum_{\gamma' \in \Gamma} \text{length}(\gamma') \leq e^{\kappa t}\text{length}(\gamma).$$

Proof : Let's suppose, first, that Γ consists of a single element γ'. Given $\mu > 1$, and $t_0 = t$, choose δ as in Lemma 6.2. Let $x_1, \ldots, x_n \in \gamma'$ be a cyclically ordered sequence of points of γ' such that $d(x_i, x_{i+1}) \leq \delta$ for all i. Thus, writing $y_i = \alpha_{x_i}(t)$ we have $d(y_i, y_{i+1}) \geq e^{-\mu\kappa t} d(x_i, x_{i+1})$ for all i. Since $\gamma' \subseteq D(\gamma)$, we see easily that $y_i \in \gamma$. In fact, from the topology of the situation, the points y_i must be cyclically ordered around γ. Thus,

$$\sum_{i=1}^{n} d(x_i, x_{i+1}) \leq e^{\mu\kappa t} \sum_{i=1}^{n} d(y_i, y_{i+1}) \leq e^{\mu\kappa t}\text{length}(\gamma).$$

Letting $\mu \to 1$ and $\delta \to 0$, we find that length$(\gamma') \leq e^{\kappa t}$length$(\gamma)$.

Now suppose that Γ consists of a finite number, m, of horocycles, $\gamma^1, \ldots, \gamma^m$. The argument is similar. Given $\mu > 0$, choose $\delta > 0$, suppose $(x_i^j)_i$ are cyclically ordered on γ^j with $d(x_i^j, x_{i+1}^j) \leq \delta$, and set $y_i^j = \alpha_{x_i^j}(t)$. This time, the order of the points $(y_i^j)_i^j$ on γ can be obtained, combinatorially, by cutting the circles γ^j, and splicing them together to form γ. (Figure 2.) We need to make at most $2m - 2$ cuts. In other words, all but at most $2m - 2$ adjacent pairs of points x_i^j correspond to adjacent points y_i^j on γ. Thus

$$\sum_{j=1}^{m} \sum_{i} d(x_i^j, x_{i+1}^j) \leq e^{\mu\kappa t}\text{length}(\gamma) + (2m - 2)\delta.$$

Letting $\mu \to 1$ and $\delta \to 0$, we obtain

$$\sum_{j=1}^{m} \text{length}(\gamma^j) \leq e^{\kappa t}\text{length}(\gamma).$$

The result now follows also for Γ infinite. ♢

Proof of Lemma 4.5 : Suppose $u \in [0, \infty)$ and $t > 0$. We want to show that $f(u-t) \leq e^{\kappa t} f(u)$.

Note that if γ' is any essential horocycle at level $u-t$, then $\gamma' \subseteq D(\gamma)$ for some essential horocycle γ at level u. Thus, by Lemma 7.1, $\text{length}(\gamma') \leq e^{\kappa t}\text{length}(\gamma) \leq e^{\kappa t} f(u)$. We thus need to show that if $e^{\kappa t} f(u) < 2\pi$, then there is indeed at least one essential horocycle at level $u-t$.

We claim that there is some function $\epsilon : (0, 2\pi) \longrightarrow (0, \infty)$ such that if $u_1, u_2 \in [0, \infty)$ with $0 < u_2 - u_1 \leq \epsilon(L)$, and if γ is an essential horocycle at level u_2 of length at most L, then $D(\gamma)$ contains an essential horocycle at level u_1.

To see that the result follows from the claim, suppose, for contradiction, that $t_0 = \inf\{v > 0 \mid f(u-v) = 2\pi\} \leq t$. Then $L = e^{\kappa t_0} f(u) \leq e^{\kappa t} f(u) < 2\pi$. Choose u_1, u_2 so that $u - t_0 - \frac{\epsilon(L)}{2} < u_1 < u - t_0 < u_2 < u - t_0 + \frac{\epsilon(L)}{2}$. Thus $f(u_1) < e^{\kappa t_0} f(u) = L$, and so $f(u_2) < 2\pi$. This contradicts the definition of t_0.

To prove the claim, fix $L \in (0, 2\pi)$, and choose $\epsilon < \frac{1}{\kappa} \log \frac{1}{2}\left(1 + \frac{2\pi}{L}\right)$, and such that $\text{area}(N(\beta, \epsilon)) < \frac{1}{2}(2\pi - L)$ for any rectifiable curve, β, of length $\leq 2\pi$. Let Γ be the set of all horocycles at level u_1 contained in $D(\gamma)$. By Lemma 7.1, we have

$$\sum_{\gamma' \in \Gamma} \text{length}(\gamma') \leq e^{\kappa(u_2 - u_1)}\text{length}(\gamma) \leq e^{\kappa\epsilon} L < \frac{1}{2}(L + 2\pi).$$

Now, if each $\gamma' \in \Gamma$ were inessential, we would have $\text{area}(D(\gamma')) \leq A_-(\text{length}(\gamma')) \leq \text{length}(\gamma')$, and so $\text{area}\left(\bigcup_{\gamma' \in \Gamma} D(\gamma')\right) \leq \frac{1}{2}(L + 2\pi)$. Now, $D(\gamma) \subseteq N(\gamma, \epsilon) \cup \bigcup_{\gamma' \in \Gamma} D(\gamma')$, and so $\text{area}(D(\gamma)) < \frac{1}{2}(2\pi - L) + \frac{1}{2}(2\pi + L) = 2\pi$. Thus γ would be inessential. This proves the claim. ♢

Proof of Proposition 4.3 : Note that $f(0) = 2\pi$. If $L < 2\pi$, let $t = \inf\{u > 0 \mid f(u) < L\}$. From Lemma 4.5, we see that $f(t) = L$. ♢

Proof of Theorem 2.2 : By Proposition 4.3, Lemma 4.2 and Theorem 3.1, we have $\text{area}(\mathbf{R}^2) \geq \frac{1}{\kappa} L + A_+(L)$ for all $L \in (0, 2\pi)$. Substitute $L = 2\pi/\sqrt{1 + \kappa^2}$. ♢

8. Generalities.

We can apply similar arguments to obtain a result about riemannian metrics on the disc, D^2:

Theorem 8.1 : *Suppose that g is a riemannian metric on the disc D^2 such that the boundary, ∂D^2, is geodesic, and such that g has curvature between $-\kappa^2$ and 1 on the interior. If* $\text{length}(\partial D^2) = 2\pi\lambda$ *for $\lambda < 1$, then*

$$\text{area}(D^2, g) \geq 2\pi\left(1 + \frac{1}{\kappa}\sqrt{(1 + \kappa^2)(1 - \lambda^2)}\right).$$

Thus Theorem 2.2 can be viewed as the limiting case as $\lambda \to 0$. Here we use "geodesic" in the riemannian sense to mean that ∂D^2 has zero extrinsic curvature.

The bound of the theorem is sharp, as the following construction demonstrates. Set $r = \frac{1}{\kappa}\cosh^{-1}\left(\frac{1}{\lambda}\sqrt{\frac{1+\lambda^2\kappa^2}{1+\kappa^2}}\right)$, and let N be a one-sided r-neighbourhood of a biinfinite geodesic in the plane of constant curvature $-\kappa^2$. Let τ be a hyperbolic isometry which translates the geodesic through a distance of $2\pi\lambda$. Let F be the annulus N/τ. Now, F has area $\frac{2\pi}{\kappa}\sqrt{\frac{1-\lambda^2}{1+\kappa^2}}$, and ∂F consists of a closed geodesic of length $2\pi\lambda$, and a curve of length $L = 2\pi\sqrt{\frac{1+\lambda^2\kappa^2}{1+\kappa^2}}$ and inward curvature $\kappa\sqrt{\frac{1-\lambda^2}{1+\lambda^2\kappa^2}}$. We now take a large spherical cap bounded by a round circle of length L (and hence outward curvature $\kappa\sqrt{\frac{1-\lambda^2}{1+\lambda^2\kappa^2}}$) and of area $A_+(L) = 2\pi\left(1+\kappa\sqrt{\frac{1-\lambda^2}{1+\kappa^2}}\right)$. Joining these together, we get the required metric on the disc, of area $2\pi\left(1+\frac{1}{\kappa}\sqrt{(1+\kappa^2)(1-\lambda^2)}\right)$.

Proof of Theorem 8.1 (sketch) : The argument proceeds along similar lines to that of Theorem 2.2. Geodesic rays are replaced by shortest paths to the boundary, ∂D^2. The "horofunction" h is defined by $h(x) = T - d(x, \partial D^2)$ (where the constant T is chosen so that $h \geq 0$). We define "horocycle" in the same way. Thus, $\partial D^2 = h^{-1}T$ is the unique horocycle at level T. The Gauß-Bonnet Theorem tells us that area$(D^2) \geq 2\pi$, and so ∂D^2 is essential. If $u \in [0,T]$ and γ is a horocycle at level u, then length$(\gamma) \leq \kappa(\coth\kappa(T-u))$area$(R(\gamma))$, where $R(\gamma)$ is the annulus lying between γ and ∂D^2. (This reduces to Lemma 4.2 as $T \to \infty$, and can be proved by similar arguments.) Also, for $t, u \in [0,T]$, we have $f(u-t) \leq \frac{\cosh\kappa(t+T-u)}{\cosh\kappa(T-u)} f(u)$ (which reduces to Lemma 4.5 as $T \to \infty$). In particular $f(T-t) \leq 2\pi\lambda\cosh\kappa t$. (Note that, for small $t > 0$, $h^{-1}t$ is a smooth essential horocycle.) These results imply that f attains every value between $2\pi\lambda$ and 2π.

Suppose then that γ is an essential horocycle at level $u = T - t$, of length $L \in [2\pi\lambda, 2\pi)$. Now $L \leq f(u) \leq 2\pi\lambda\cosh\kappa t$, and so area$(R(\gamma)) \geq \frac{\tanh\kappa t}{\kappa}L \geq \frac{L}{\kappa}\sqrt{1-\left(\frac{2\pi\lambda}{L}\right)^2} = \frac{1}{\kappa}\sqrt{L^2-(2\pi\lambda)^2}$. Thus, we have area$(D^2, g) \geq A_+(L) + \frac{1}{\kappa}\sqrt{L^2-(2\pi\lambda)^2}$. On substituting $L = 2\pi\sqrt{\frac{1+\lambda^2\kappa^2}{1+\kappa^2}}$, we obtain area$(D^2, g) \geq 2\pi\left(1+\frac{1}{\kappa}\sqrt{(1+\kappa^2)(1-\lambda^2)}\right)$ as required. $\diamondsuit$

We can apply this result metrics on the 2-sphere, S^2.

Suppose that g is a riemannian metric on S^2 of curvature ≤ 1. If (S^2, g) contains a closed geodesic (in the riemannian sense) of length less than 2π, then, following Charney and Davis [3], we define the *systole*, sys(S^2, g), of (S^2, g), to be the length of a shortest closed geodesic. Such a shortest geodesic is necessarily embedded. The systole seems to be natural quantity to associate to such a metric. For example, suppose $0 < \lambda < 1$. Then, the statement that sys$(S^2, g) \leq 2\pi\lambda$ (i.e. that there exists a closed geodesic of length at most $2\pi\lambda$) is equivalent to either of the statements:

(1) The injectivity radius of (S^2, g) (at some point) is at most $\pi\lambda$, or

(2) There is a non-shrinkable closed curve of length at most $2\pi\lambda$.

("Shrinkable" is defined in Section 3.) For proofs, see [3,2].

The following result reduces to [1, Théorème 10] when $\kappa = 1$:

Theorem 8.2 : *Suppose g is a riemannian metric on S^2, with curvature between $-\kappa^2$ and 1. Suppose* $\mathrm{sys}(S^2, g) \leq 2\pi\lambda$ *for some $\lambda < 1$. Then,*

$$\mathrm{area}(S^2, g) \geq 4\pi \left(1 + \frac{1}{\kappa} \sqrt{(1+\kappa^2)(1-\lambda^2)} \right).$$

Proof : The shortest closed geodesic divides S^2 into two discs. Apply Theorem 8.1. ♢

Again, this inequality is sharp. The bound is attained by taking the disc described after Theorem 8.1, and doubling it in its boundary.

References.

[1] C.Bavard, P.Pansu, *Sur le volume minimal de* $\mathbf{R}^2$: Ann. Scient. Éc. Norm. Sup. **19** (1986) 479–490.

[2] B.H.Bowditch, *Notes on locally CAT(1) spaces* : preprint, Aberdeen (1992).

[3] R.Charney, M.Davis, *Singular metrics of non-positive curvature on branched covers of riemannian manifolds* : preprint, Ohio State (1990).

[4] J.Cheeger, D.G.Ebin, *Comparison theorems in riemannian geometry* : North-Holland (1975).

[5] J.Cheeger, M.Gromov, *Collapsing riemannian manifolds while keeping their curvature bounded I* : J. Diff. Geom. **23** (1986) 309–346.

[6] J.Cheeger, M.Gromov, *Collapsing riemannian manifolds while keeping their curvature bounded II* : J. Diff. Geom. **32** (1990) 269–298.

[7] S.Gallot, *Volume minimal des variétés hyperboliques: un théorème local et un resultat global* : in "Séminaire de théorie spectrale et géométrie" Vol 7, Grenoble—St. Martin d'Hères, (1988) 35–52.

[8] M.Gromov, *Volume and bounded cohomology* : Publ. Math. I.H.E.S. No. 56 (1982) 5–99.

[9] J.G.Hocking, G.S.Young, *Topology* : Addison-Wesley (1961).

[10] W.H.Jaco, P.B.Shalen, *Seifert fibered spaces in 3-manifolds* : Amer. Math. Soc. Memoirs No. 220 (1979).

[11] K.Johannson, *Homotopy equivalences of 3-manifolds with boundary* : Springer Lecture Notes in Mathematics No. 761, Springer-Verlag (1979).

[12] W.Klingenberg, *Riemannian geometry* : de Gruyter Studies in Mathematics No. 1, de Gruyter (1982).

[13] E.E.Moise, *Geometric topology in dimensions 2 and 3* : Graduate Texts in Mathematics No. 47, Springer (1977).

[14] R.Osserman, *Bonnesen-style inequalities* : Amer. Math. Monthly **86** (1977) 1–29.

[15] R.Osserman, *The isoperimetric inequality* : Bull. Amer. Math. Soc. **84** (1978) 1182–1238.

[16] W.P.Thurston, *Three dimensional manifolds, Kleinian groups and hyperbolic geometry* : Bull. Amer. Math. Soc. **6** (1982) 357–381.

[17] D.Yang, *Convergence of riemannian manifolds with integral bounds on curvature I* : Ann. Scient. Éc. Norm. Sup. **25** (1992) 77-105.