ISSN 1364-0380 (on line) 1465-3060 (printed)

Geometry & Topology
Volume 6 (2002) 889–904
Published: 21 December 2002

Attaching handlebodies to 3–manifolds

Marc Lackenby

Mathematical Institute, Oxford University
24–29 St Giles', Oxford OX1 3LB, UK

Email: `lackenby@maths.ox.ac.uk`

Abstract

The main theorem of this paper is a generalisation of well known results about Dehn surgery to the case of attaching handlebodies to a simple 3–manifold. The existence of a finite set of 'exceptional' curves on the boundary of the 3–manifold is established. Provided none of these curves is attached to the boundary of a disc in a handlebody, the resulting manifold is shown to be word hyperbolic and 'hyperbolike'. We then give constructions of gluing maps satisfying this condition. These take the form of an arbitrary gluing map composed with powers of a suitable homeomorphism of the boundary of the handlebodies.

AMS Classification numbers Primary: 57N10

Secondary: 57N16, 57M50, 20F65

Keywords: 3–manifold, handlebody, word hyperbolic

Proposed: Cameron Gordon
Seconded: Jean-Pierre Otal, Benson Farb

Received: 19 February 2002
Revised: 20 December 2002

1 Introduction

This paper deals with a generalisation of Dehn surgery. Instead of attaching solid tori to a 3–manifold with toral boundary components, we start with a 3–manifold with higher genus boundary components and glue on handlebodies. Our aim is to generalise well known surgery results to this setting. As is customary in surgery theory, we have to assume that the initial bounded 3–manifold satisfies certain generic topological hypotheses: it will be *simple*, which means that it is compact, orientable, irreducible, atoroidal, acylindrical, with incompressible boundary. Our first and main theorem is the following.

Theorem 1 *Let M be a simple 3–manifold with non-empty boundary. Then there is a finite collection $\mathcal{C}$ of essential simple closed curves on ∂M with the following property. Suppose that H is a collection of handlebodies, and that ϕ: $\partial M \to \partial H$ is a homeomorphism that sends no curve in $\mathcal{C}$ to the boundary of a disc in H. Then $M \cup_\phi H$ is irreducible, atoroidal, word hyperbolic and not Seifert fibred. Furthermore, the inclusion map of any component of H into $M \cup_\phi H$ induces an injection between their fundamental groups, and hence $\pi_1(M \cup_\phi H)$ is infinite.*

The set $\mathcal{C}$ we term the *exceptional curves*. They may be characterised in terms of the geometry of M. For, M has a complete finite volume hyperbolic structure, with totally geodesic boundary and possibly some cusps. This may be seen by applying Thurston's geometrisation theorem [11] to two copies of M glued via the identity map along their negative Euler characteristic boundary components. Let $N(\partial M_{\rm cusp})$ be a horoball neighbourhood of the cusps of M, such that, on each component of $\partial N(\partial M_{\rm cusp})$, the shortest Euclidean geodesic has length 1. It is shown in [2], for example, that $N(\partial M_{\rm cusp})$ is a product neighbourhood of the toral ends of M.

Theorem 2 *The set $\mathcal{C}$ consists of simple closed geodesics having length at most*

$$\frac{4\pi}{(1-4/\chi(S))^{1/4}-(1-4/\chi(S))^{-1/4}}$$

on boundary components S with genus at least two, and of closed Euclidean geodesics with length at most 2π on $\partial N(\partial M_{\rm cusp})$. Hence, there is an upper bound on the number of curves in $\mathcal{C}$ that depends only on the genus of ∂M and is otherwise independent of M.

Theorems 1 and 2 can be viewed as a generalisation of Thurston's hyperbolic Dehn surgery theorem [13] and its extension by Hodgson and Kerckhoff [8]. Thurston established the existence of a finite set $\mathcal{C}$ of slopes on a torally bounded hyperbolic 3–manifold, such that, provided these slopes are avoided when surgery is performed, the result is a 3–manifold with a hyperbolic structure. Hodgson and Kerckhoff provided a universal upper bound, independent of the 3–manifold, on the number of curves in $\mathcal{C}$ on each torus boundary component.

Theorems 1 and 2 should be compared with the main theorem of [12], due to Scharlemann and Wu. They considered the attachment of a single 2–handle to M. One is only allowed to attach it along a certain type of curve, known as a 'basic' curve, but this is not a serious restriction. They prove that the resulting manifold is hyperbolic, provided one avoids a finite set of curves on ∂M, up to isotopy. Attaching a handlebody can be performed by first gluing on 2–handles along basic curves and then Dehn filling. Thus, their result implies that, in a certain sense, most ways of attaching a handlebody give a hyperbolic manifold. The limitation of Scharlemann and Wu's procedure is that the 2–handles must be attached in sequence, and then be followed by the surgeries. Thus, the curves that the later 2–handles and surgeries must avoid depend on where the earlier 2–handles are attached. In Theorem 1, we attach the entire handlebody in a single step. This allows us to identify the exceptional curves at the outset.

Of course, Theorem 1 raises the problem of finding homeomorphisms ϕ with the required property. The approach we consider is to start with an arbitrary ϕ, and then modify this by applying powers of a homeomorphism $f\colon \partial H \to \partial H$. If f extends to a homeomorphism of H, then this will not change the resulting manifold. More generally, if some power of f extends to a homeomorphism, then only finitely many manifolds are created. We would like to avoid this situation. In fact, one must rule out a yet more general possibility: no power f^n of f (where $n \neq 0$) may *partially extend to* H. This means that there is a compression body R (other than a product) embedded in H with positive boundary ∂H, such that f^n extends to a homeomorphism of R. For, if ϕ were to map a curve C in $\mathcal{C}$ to the boundary of a disc in R, then $f^n\phi(C)$ would bound a disc in R for infinitely many n. However, if this condition is met, we shall show that this method does create infinitely many manifolds that satisfy the conclusions of Theorem 1.

Theorem 3 *Let M be a simple 3–manifold, let H be a collection of handlebodies, and let $\phi\colon \partial M \to \partial H$ be a homeomorphism. Let $f\colon \partial H \to \partial H$ be a homeomorphism, no power of which partially extends to H. Then, for infinitely many integers n, $M \cup_{f^n \circ \phi} H$ satisfies the conclusions of Theorem 1.*

An alternative method of finding suitable homeomorphisms f is to use the theory of pseudo-Anosov maps ([14], [3]). We can regard simple closed curves on ∂H, with counting measure, as elements of the space $PL(\partial H)$ of projective measured laminations on ∂H. Let $B(\partial H)$ be the closure in $PL(\partial H)$ of the set of curves on ∂H that bound discs in H. It is a theorem of Masur [10] that $B(\partial H)$ is nowhere dense in $PL(\partial H)$. Hence that the stable and unstable laminations of a 'generic' pseudo-Anosov homeomorphism $f\colon \partial H \to \partial H$ will not lie in $B(\partial H)$.

Theorem 4 *Let M be a simple 3–manifold, let H be a collection of handlebodies, and let $\phi\colon \partial M \to \partial H$ be a homeomorphism. Let $f\colon \partial H \to \partial H$ be a pseudo-Anosov homeomorphism whose stable and unstable laminations do not lie in $B(\partial H)$. Then for all but finitely many integers n, $M \cup_{f^n \circ \phi} H$ satisfies the conclusions of Theorem 1.*

The proof of Theorems 1 and 2 follows two papers: [9], which established the word hyperbolicity of certain surgered manifolds, and [7] by Hass, Wang and Zhou, which examined boundary slopes of immersed essential surfaces in hyperbolic 3–manifolds with totally geodesic boundary. Theorem 3 uses some new techniques, involving arrangements of discs in a handlebody and a delicate counting argument. Theorem 4 is an elementary application of the theory of pseudo-Anosov automorphisms.

This paper suggests many interesting areas for further research. Firstly, can the results be upgraded to deduce the existence of a metric that is hyperbolic or just negatively curved? Secondly, can Theorem 3 be strengthened so that the conclusion holds for all but finitely many n? Thirdly, can the techniques of this paper be generalised to analyse Heegaard splittings?

2 The main theorem

In this section, we will prove Theorems 1 and 2. Suppose therefore that M is a simple 3–manifold with non-empty boundary. Denote its toral boundary components by $\partial M_{\mathrm{cusp}}$ and the remaining boundary components by $\partial M_{\mathrm{geo}}$. We work with the complete finite volume hyperbolic structure on $M - \partial M_{\mathrm{cusp}}$ in which $\partial M_{\mathrm{geo}}$ is totally geodesic. Let $\mathcal{C}$ be the set of simple closed curves on ∂M as described in Theorem 2. The fact that there is an upper bound on the number of curves in $\mathcal{C}$ that depends only on the genus of ∂M is well known. A proof is given in [7] for example. Suppose that $\phi\colon \partial M \to \partial H$ is a homeomorphism that sends no curve in $\mathcal{C}$ to the boundary of a disc in H.

Our first step is to show that the inclusion map of any component of H into $M \cup_\phi H$ induces an injection between their fundamental groups. Consider an essential loop L in H, and suppose that it is homotopically trivial in $M \cup_\phi H$. There is then a map $f\colon D \to M \cup_\phi H$, where D is a disc, such that $f|_{\partial D}$ winds once around L. Homotope f a little, so that $f^{-1}(\partial H)$ is a collection of simple closed curves in the interior of D, and so that f is transverse to ∂H near these curves. We suppose that L and f have been chosen so that the number of these curves has been minimised.

Claim 1 *$f^{-1}(H)$ is a collection of discs and a collar on ∂D.*

If not, pick a curve L' of $f^{-1}(\partial H)$ that is innermost in D among curves that do not bound discs of $f^{-1}(H)$ and that are not parallel in $f^{-1}(H)$ to ∂D. It bounds a disc D'. If L' is homotopically trivial in H, we may modify f in D' so that it is mapped entirely to H, thus reducing $|f^{-1}(\partial H)|$, which contradicts the minimality of $|f^{-1}(\partial H)|$. If L' is homotopically non-trivial in H, we may work with D' instead of D. Again the minimality assumption is violated.

Claim 2 *The surface $F = f^{-1}(M)$ is homotopically boundary-incompressible in M.*

Recall from [9] that this means that no properly embedded essential arc in F can be homotoped in M, keeping its endpoints fixed, to an arc in ∂M. For, if there were such an arc α, we could perform a homotopy to f, taking a regular neighbourhood of α into H. There are two cases to consider: when the endpoints of α lie in distinct boundary components of F, and when they lie in the same component. In the first case, the result is to reduce $|f^{-1}(\partial H)|$, contradicting the minimality assumption.

In the second case, there are two subcases: either $\partial\alpha$ lies in the collar on ∂D or it does not. Suppose first that $\partial\alpha$ misses the collar. The arc separates F into two components, as F is planar. One of these, F' say, does not intersect the collar on ∂D. Let L' be the boundary component of F' that runs along the neighbourhood of α. Push L' a little into H. Since F' has fewer boundary components than F and L' is homotopically trivial in $M \cup_\phi H$, the minimality assumption implies that L' is homotopically trivial in H. We may therefore remove F' from F and replace it with a disc that maps to H. This reduces $|\partial F|$, which is a contradiction.

Suppose now that both endpoints of α lie in the collar on ∂D. We consider the two halves of F cut open along α. By the minimality assumption, the boundary curves of each are homotopically trivial in H. But then L is trivial in H, which is a contradiction.

Claim 3 *F is homotopically incompressible in M.*

This means that no homotopically non-trivial simple closed curve in F maps to a homotopically trivial curve in M. The argument is similar to that of Claims 1 and 2, but simpler, and so is omitted.

We now follow the argument of [7]. Let $N(\partial M_{\text{cusp}})$ be a horoball neighbourhood of the cusps of ∂M, such that on each component of $\partial N(\partial M_{\text{cusp}})$, the shortest geodesic has length 1. Let $N(\partial M_{\text{geo}})$ be the set of points at a distance at most U from ∂M_{geo}. By a theorem of Basmajian [1], if we take U to be

$$\frac{1}{4}\log\left(1-\frac{4}{\chi(S)}\right)=\sinh^{-1}\left(\frac{(1-4/\chi(S))^{1/4}-(1-4/\chi(S))^{-1/4}}{2}\right),$$

then $N(\partial M_{\text{cusp}}) \cup N(\partial M_{\text{geo}})$ will be a collar on ∂M. Denote $N(\partial M_{\text{cusp}}) \cup N(\partial M_{\text{geo}})$ by $N(\partial M)$.

Claim 4. *There is a least area surface in the homotopy class of $f\colon (F,\partial F)\to(M,\partial M)$. This is an immersion.*

If F were closed and M had no cusps, this would be Lemma 2 of [5], as F is homotopically incompressible, by Claim 3. This was extended to the case where F has boundary and M has cusps in Theorem 4.4 of [6]. In [6], though, it was assumed that $f_*\colon \pi_1(F)\to\pi_1(M)$ and $f_*\colon \pi_1(F,\partial F)\to\pi_1(M,\partial M)$ are injective. However, as explained in [5], these hypotheses can be weakened to the assumption that F is homotopically incompressible and homotopically boundary-incompressible.

We perform the homotopy in Claim 4. Let ∂F_{cusp} and ∂F_{geo} be the boundary components of F that map to ∂M_{cusp} and ∂M_{geo} respectively. Then $F-\partial F_{\text{cusp}}$ inherits a Riemannian metric with geodesic boundary. Its sectional curvature is the sum of the sectional curvature of M and the product of its principal curvatures (see Theorem 5.5 of [4], for example). The principal curvatures sum to zero, since F is minimal, and hence their product is non-positive. Therefore, the sectional curvature of F is at most -1.

We abuse notation slightly by denoting the component of ∂F parallel to L by L. Let $N(\partial F_{\text{cusp}})$ be those components of the inverse image of $N(\partial M_{\text{cusp}})$ in F that contain a component of ∂F_{cusp}. Let $N(\partial F_{\text{geo}})$ be the set of points in F with distance at most U from ∂F_{geo} in its intrinsic path metric. Denote $N(\partial F_{\text{cusp}}) \cup N(\partial F_{\text{geo}})$ by $N(\partial F)$.

Claim 5 *$N(\partial F)$ is a collar on ∂F.*

We start by showing that each component of $N(\partial F_{\text{geo}})$ is a collar. To see this, increase U from zero to its final value. Near zero, $N(\partial F_{\text{geo}})$ is clearly a collar. But suppose that as it expands, there is some point at which a self-tangency is created. Then there is a geodesic arc properly embedded in $N(\partial F_{\text{geo}})$ with endpoints perpendicular to ∂F_{geo}. This geodesic is the concatenation of two geodesic arcs which run from ∂F_{geo} to the point of self-tangency. Since F is negatively curved, this arc is essential in F. But it lies within $N(\partial M_{\text{geo}})$, which is a collar, and so can be homotoped into ∂M, keeping its endpoints fixed. This contradicts the fact that F is homotopically boundary-incompressible.

It is also clear that each component of $N(\partial F_{\text{cusp}})$ is a collar. For, otherwise, it contains a properly embedded arc that is essential in $N(\partial F_{\text{cusp}})$ and that has endpoints in ∂F_{cusp}. The fact that $N(\partial M_{\text{cusp}})$ is a collar on ∂M_{cusp} implies that this arc can be homotoped into ∂M. The arc is therefore inessential in F. Hence, some component of $\partial N(\partial F_{\text{cusp}})$ bounds a disc in F. Consider the lift of this disc to $\mathbb{H}^3$, which contains the universal cover of M. The boundary of the disc lies on a component of the inverse image of $\partial N(\partial M_{\text{cusp}})$, which is a horotorus. There is a natural projection from $\mathbb{H}^3$ onto this horotorus, and this reduces the area of the disc. This contradicts the assumption that F is least area.

Finally, $N(\partial F_{\text{cusp}})$ and $N(\partial F_{\text{geo}})$ must be disjoint. For the former lies in $N(\partial M_{\text{cusp}})$, whereas the latter lies in $N(\partial M_{\text{geo}})$, and these are disjoint. This proves the claim.

Claim 6 $\mathrm{Area}(N(\partial F_{\text{geo}})) \geq (\sinh U)\ (\mathrm{Length}(\partial F_{\text{geo}}))$.

Choosing orthogonal co-ordinates for the surface $N(\partial F_{\text{geo}})$, the metric is given by

$$ds^2 = du^2 + J^2(u,v)dv^2,$$

where $J(u,v) > 0$ and $J(0,v) = 1$. The curves where v is constant are geodesics perpendicular to the boundary, and the curve $u = 0$ lies in the boundary. By the calculations in [7],

$$\frac{\partial J}{\partial u} \geq 0, \qquad \frac{\partial^2 J}{\partial u^2} \geq J,$$

and hence $J(u,v) \geq \cosh(u)$. Therefore,

$$\begin{aligned}\mathrm{Area}(N(\partial F_{\text{geo}})) &= \int_{v=0}^{\mathrm{Length}(\partial F_{\text{geo}})} \int_{u=0}^{U} J(u,v)\ du\ dv \\ &\geq \int_{v=0}^{\mathrm{Length}(\partial F_{\text{geo}})} \int_{u=0}^{U} \cosh u\ du\ dv \\ &= (\sinh U)\ (\mathrm{Length}(\partial F_{\text{geo}})).\end{aligned}$$

Give ∂H the Riemannian metric on $\partial N(\partial M_{\text{cusp}}) \cup \partial M_{\text{geo}}$. Each component of ∂F, except the one parallel to L, is mapped to the boundary of a disc in H, by Claim 1. It is essential in ∂H by Claim 3. So, by assumption, it has length more than the bound in Theorem 2 if it is simple. The non-simple case follows from the following claim.

Claim 7 *Each shortest essential closed curve in ∂H that is homotopically trivial in H is simple.*

Let $\tilde{H}$ be the universal cover of H. Then $\partial\tilde{H}$ is the cover of ∂H corresponding to the kernel of $\pi_1(\partial H) \to \pi_1(H)$. In particular, it is a regular cover. The closed curves in $\partial\tilde{H}$ are exactly the lifts of curves in ∂H that are homotopically trivial in H. The shortest of these, C say, that is essential in $\partial\tilde{H}$ is a geodesic. It must be simple, for otherwise, it can be modified at a singular point to reduce its length.

We claim that C must be disjoint from its covering translates. For if C and some translate C' were to intersect transversely, they would do so at least twice, since both are homologically trivial in $\tilde{H}$. Pick two points of intersection, which divide C and C' each into two arcs. Glue the shorter of the two arcs in C (or either of these arcs if they have the same length) to the shorter of the two in C', and then smooth off to form a shorter closed curve in $\partial\tilde{H}$. If this is essential in $\partial\tilde{H}$, we have a contradiction. If not, then we may homotope the longer of its two arcs onto the other, and then smooth off. This reduces the length of C or C', which again is a contradiction.

If C and C' were to intersect non-transversely, then they would be equal. Hence, C would project to a multiple of a simple closed curve C'' in ∂H. If we apply the Loop Theorem to a regular neighbourhood of C'', we see that C'' must also bound a disc in H. But this is a shorter curve than C, which is a contradiction.

Hence, C projects homeomorphically to a simple closed curve in ∂H which is shortest among essential curves that are trivial in H. This proves the claim.

Putting together Claims 6 and 7, we obtain

$$\text{Area}(N(\partial F_{\text{geo}})) \geq 2\pi\,|\partial F_{\text{geo}} - L|.$$

Let $\text{Length}(\partial F_{\text{cusp}})$ be the length of the curves $\partial N(\partial F_{\text{cusp}})$ on $\partial N(\partial M_{\text{cusp}})$. By Claim 5, each of these curves has the same slope as the corresponding component of ∂F. Hence, each (except possibly L) has length more than 2π, by assumption. Thus, using a well known argument, which can be found in the proof of Theorem 4.3 of [6], for example, we obtain the following.

Claim 8 $\mathrm{Area}(N(\partial F_{\mathrm{cusp}})) \geq 2\pi\,|\partial F_{\mathrm{cusp}} - L|$.

Since F is a minimal surface in a hyperbolic manifold, its sectional curvature K is at most -1. So, applying Gauss–Bonnet to F:

$$\begin{aligned}2\pi(2-|\partial F|) = 2\pi\chi(F) = \int_F K\ dA &\leq -\mathrm{Area}(F) \\ &\leq -\mathrm{Area}(N(\partial F_{\mathrm{cusp}})) - \mathrm{Area}(N(\partial F_{\mathrm{geo}})) \leq -2\pi(|\partial F|-1),\end{aligned}$$

which is a contradiction. We therefore deduce that the inclusion of each component of H into $M \cup_\phi H$ is π_1–injective.

A very similar argument can be used to show that $M \cup_\phi H$ is irreducible. For otherwise, M would contain a properly embedded incompressible boundary-incompressible planar surface, each boundary component of which would extend to a disc in H. (To make this argument work, we need the fact that the inclusion of H into $M \cup_\phi H$ is π_1–injective.) We can then apply the above analysis to this surface.

It remains to show that $M \cup_\phi H$ is word hyperbolic. For this implies that $\pi_1(M \cup_\phi H)$ contains no rank two abelian subgroup, and hence that $M \cup_\phi H$ is atoroidal. The fact that $\pi_1(M \cup_\phi H)$ is infinite then gives that $M \cup_\phi H$ is not Seifert fibred.

Pick a Riemannian metric g on $M \cup_\phi H$, in which H is an ϵ–neighbourhood of a graph, for small $\epsilon > 0$. We wish to establish a linear isoperimetric inequality for g. So, consider a loop L that is homotopically trivial, and which therefore forms the boundary of a mapped-in disc $f\colon D \to M \cup_\phi H$. Using a small homotopy, we may move L off H. We can ensure that this changes the length of L by a factor that is bounded independently of L. We can also ensure that the area of the annulus realizing the homotopy is similarly bounded.

By the argument of Claims 1, 2 and 3 and using the fact that the inclusion of H into $M \cup_\phi H$ is π_1–injective, we may homotope f, keeping it fixed on L, so that afterwards $F = f^{-1}(M)$ is homotopically incompressible and homotopically boundary-incompressible in M, and so that $f^{-1}(H)$ is a collection of discs.

Let h be the hyperbolic metric on $M - \partial M_{\mathrm{cusp}}$. Since M is compact, the Riemannian manifolds $(M - \mathrm{int}(N(\partial M_{\mathrm{cusp}})), h)$ and (M, g) are bi-Lipschitz equivalent, with constant c_1, say. Also, there is a map $(N(\partial M_{\mathrm{cusp}}), h) \to (H, g)$, collapsing the cusps to solid tori, that increases areas by at most a factor c_1^2, say. Initially, we will measure lengths and areas in the h metric. Note that, by construction, L is disjoint from $N(\partial M_{\mathrm{cusp}})$.

We may assume that L is homotopically non-trivial in M. For otherwise it is trivial to construct a disc in M bounded by L with area linearly bounded by the length of L. (Indeed, the fact that closed curves in hyperbolic space have this property is the motivation behind Gromov's theory of hyperbolic groups.) Also, we will, for the moment, assume that L is not homotopic to a curve in $\partial M_{\rm cusp}$. Hence it has a geodesic representative $\overline{L}$. We may realize the free homotopy between L and $\overline{L}$ with a mapped-in annulus A having area at most $c_2\mathrm{Length}(L)$, for some constant c_2 depending only on h. We may assume that A has least possible area.

Claim 9 *Each component of $A \cap N(\partial M_{\rm cusp})$ is either a disc disjoint from $\overline{L}$, or a disc containing a single component of $\overline{L} \cap N(\partial M_{\rm cusp})$.*

Note that $A \cap N(\partial M_{\rm cusp})$ is disjoint from L, and does not contain a core curve of A. Suppose that the claim is not true. Then we can find a component of $A \cap N(\partial M_{\rm cusp})$ which is not a disc or which intersects $\overline{L}$ more than once. In the former case, $A \cap N(\partial M_{\rm cusp})$ has a boundary component that is a simple closed curve bounding a disc in A with interior disjoint from $N(\partial M_{\rm cusp})$. However, we could then homotope this disc into $N(\partial M_{\rm cusp})$ to reduce the area of A, which is a contradiction. Thus, each component of $A \cap N(\partial M_{\rm cusp})$ is a disc. If one component intersects $\overline{L}$ in more than one arc, then the sub-arc of $\overline{L}$ between these two arcs does not lie wholly in $N(\partial M_{\rm cusp})$, but can be homotoped into $N(\partial M_{\rm cusp})$. This contradicts the fact that $\overline{L}$ is a geodesic. This proves the claim.

Homotope F (minus those components that map to $\partial M_{\rm cusp}$) to a least area surface in $M - \partial M_{\rm cusp}$, with one boundary component mapping to $\overline{L}$, and the remainder mapping to ∂M. Divide the boundary components of F into three subsets $\overline{L}$, $\partial F_{\rm cusp}$ and $\partial F_{\rm geo}$. Note that $F \cup A$ is now a planar surface with one boundary component mapping to L and the remaining components mapping to curves in ∂M that bound discs in H. This surface will extend to a disc in $M \cup_\phi H$ with area linearly bounded by the length of L, establishing the required linear isoperimetric inequality.

We define $N(\partial F_{\rm cusp})$ and $N(\partial F_{\rm geo})$ as before. Claim 5 now reads that $N(\partial F_{\rm cusp}) \cup N(\partial F_{\rm geo})$ is a collar on $\partial F_{\rm cusp} \cup \partial F_{\rm geo}$, which may have non-empty intersection with $\overline{L}$. However, the calculations in Claims 6 and 8 now need modification.

Further divide $\partial F_{\rm geo}$ into two subsets: $\partial F_{\rm thin}$ and $\partial F_{\rm thick}$. The former is the set of points x in $\partial F_{\rm geo}$ such that, when a perpendicular geodesic is emitted from x in F, it meets $\overline{L}$ within a distance U. Let $\partial F_{\rm thick}$ be the remainder of $\partial F_{\rm geo}$. Then the argument of Claim 6 gives that

$$\mathrm{Area}(N(\partial F_{\rm geo})) \geq (\sinh U)\ (\mathrm{Length}(\partial F_{\rm thick})).$$

By Claim 9, $N(\partial F_{\rm cusp}) \cup (A \cap N(\partial M_{\rm cusp}))$ is a collar on $\partial F_{\rm cusp}$ and possibly some discs. Using these collars, we may associate to each component of $\partial F_{\rm cusp}$ a curve on $\partial N(\partial M_{\rm cusp})$, namely the relevant boundary component of $N(\partial F_{\rm cusp}) \cup (A \cap N(\partial M_{\rm cusp}))$. We define the length of $\partial F_{\rm cusp}$ to be the length of these curves. Then, Claim 8, suitably modified, gives that

$$\mathrm{Area}(N(\partial F_{\rm cusp}) \cup A) \geq \mathrm{Length}(\partial F_{\rm cusp}),$$

and hence that

$$\mathrm{Area}(N(\partial F_{\rm cusp})) \geq \mathrm{Length}(\partial F_{\rm cusp}) - c_2 \ \mathrm{Length}(L).$$

Therefore,

$$\begin{aligned} \mathrm{Area}(F) &\geq \mathrm{Area}(N(\partial F_{\rm geo})) + \mathrm{Area}(N(\partial F_{\rm cusp})) \\ &\geq (\sinh U)\ (\mathrm{Length}(\partial F_{\rm thick})) \\ &\qquad + \mathrm{Length}(\partial F_{\rm cusp}) - c_2 \ \mathrm{Length}(L). \end{aligned} \tag{1}$$

Also,

$$\mathrm{Length}(L) \geq \mathrm{Length}(\overline{L}) \geq \mathrm{Length}(\overline{L} \cap N(\partial M_{\rm geo})) \geq \mathrm{Length}(\partial F_{\rm thin}). \tag{2}$$

Let c_3 (respectively, c_4) denote the minimal length of a geodesic on $\partial N(\partial M_{\rm cusp})$ (respectively, $\partial M_{\rm geo}$) with length more than 2π (respectively, $2\pi/\sinh(U)$), and let c_5 be $\min\{c_3, c_4 \sinh(U)\}$, which is more than 2π. Then, since each curve of $\partial F_{\rm cusp}$ and $\partial F_{\rm geo}$ has length more than the bound in Theorem 2, we know that

$$\begin{aligned} &\mathrm{Length}(\partial F_{\rm cusp}) + (\sinh U)(\mathrm{Length}(\partial F_{\rm geo})) \\ &\qquad \geq c_5(|\partial F| - 1) > -c_5 \ \chi(F) \geq c_5 \ \mathrm{Area}(F)/2\pi. \end{aligned} \tag{3}$$

The final inequality is an application of Gauss–Bonnet. So, adding $(2\pi/c_5)$ times (3), and $(\sinh U)$ times (2), to (1), and cancelling the area term, we get:

$$\begin{aligned} &(c_2 + \sinh U)\mathrm{Length}(L) \\ &\qquad \geq (1 - 2\pi/c_5)\,(\mathrm{Length}(\partial F_{\rm cusp}) + (\sinh U)(\mathrm{Length}(\partial F_{\rm geo})))\,, \end{aligned}$$

which we summarise as

$$\mathrm{Length}(L) \geq c_6 \ \mathrm{Length}(\partial F_{\rm geo} \cup \partial F_{\rm cusp}),$$

for some positive constant c_6 independent of L. So, by equation (3),

$$\mathrm{Length}(L) \geq c_7 \ \mathrm{Length}(\partial F_{\rm geo} \cup \partial F_{\rm cusp}) + c_8 \ \mathrm{Area}(F), \tag{4}$$

where c_7 and c_8 are positive constants independent of L.

Recall that we assumed earlier that L was not homotopic in M to a curve in $\partial M_{\rm cusp}$. We may now drop this assumption, since in this case, we can easily find a surface F satisfying the above inequality. One method is as follows. Let L' be the curve in $\partial N(\partial M_{\rm cusp})$ that is homotopic to L, and that is a geodesic in the Euclidean Riemannian metric on $\partial N(\partial M_{\rm cusp})$. Then L' has length less than that of L. The homotopy between L and L' is realized by an annulus A', say. Identify A' with $S^1 \times I$, and homotope each arc $\{*\} \times I$, keeping its endpoints fixed, to a geodesic. It is not hard to calculate that the area of A' after this homotopy is at most $(\mathrm{Length}(L) + \mathrm{Length}(L')) \leq 2\ \mathrm{Length}(L)$. Now let F be the union of A and the vertical annulus above L' in $N(\partial M_{\rm cusp})$. Then, the area of F is at most $3\ \mathrm{Length}(L)$. Also, the length of $\partial F_{\rm geo} \cup \partial F_{\rm cusp}$ is the length of L', which is less than that of L. This verifies (4) in this case.

Changing (4) to the metric g, we get

$$\mathrm{Length}(L, g) \geq (c_7/c_1^2)\mathrm{Length}(\partial F_{\rm geo} \cup \partial F_{\rm cusp}, g) + (c_8/c_1^3)\mathrm{Area}(F, g).$$

Now, each component of H has free fundamental group, and hence is word hyperbolic. So, the curves $\partial F_{\rm geo} \cup \partial F_{\rm cusp}$ bound discs in H with area at most $c_9\,\mathrm{Length}(\partial F_{\rm geo} \cup \partial F_{\rm cusp}, g)$, for some constant c_9. Attaching these discs to F, and then attaching the annulus A between L and $\overline{L}$, we create a disc bounded by L with area at most $c_{10}\,\mathrm{Length}(L, g)$, for some constant c_{10} independent of L. This establishes the required linear isoperimetric inequality, and hence the proof of Theorems 1 and 2 is complete.

3 Modifying the gluing map by powers of a homeomorphism

In this section, we prove Theorem 3. Therefore let $\phi\colon \partial M \to \partial H$ be some homeomorphism, and let $f\colon \partial H \to \partial H$ be a homeomorphism, no power of which partially extends to H. Let $\mathcal{C} = \{C_1, \dots, C_{|\mathcal{C}|}\}$ be the exceptional curves on ∂M. Suppose that Theorem 3 does not hold. Then, there is some finite subset $\mathcal{S}$ of $\mathbb{Z}$, and a function $i\colon \mathbb{Z} - \mathcal{S} \to \{1, \dots, |\mathcal{C}|\}$ such that for each $n \in \mathbb{Z} - \mathcal{S}$, $f^n\phi(C_{i(n)})$ bounds a disc D_n in H.

We wish to analyse these discs D_n in H, and we therefore introduce a definition. A *disc arrangement* is a collection of properly embedded discs in general position in a 3–manifold. Two discs arrangements are *equivalent* if there is a homeomorphism between their regular neighbourhoods taking one set of discs to the other. We term this a *relative equivalence* if the homeomorphism is the identity on the boundary of the discs. A *double curve* is a component of intersection between two of the discs. The *boundary curves* of a disc arrangement are simply the boundaries of the discs.

Lemma 1 *Fix a finite collection of embedded (not necessarily disjoint) simple closed curves in general position in a surface S. Suppose that these form the boundary curves of a disc arrangement in some irreducible 3–manifold bounded by S. Then we may isotope these discs keeping their boundaries fixed, so that afterwards, they belong to one of only finitely many relative equivalence classes of disc arrangements, and so that a regular neighbourhood of these discs and S is homeomorphic to a punctured compression body with positive boundary S.*

Proof This is by induction on partition number, which we define to be the smallest number of subsets into which we may partition the collection of discs, such that if two discs belong to the same subset, they are disjoint. When the partition number is one, the lemma is trivial, and the induction is started. We now prove the inductive hypothesis. Let $\mathcal{D}$ denote the collection of discs, and let $\mathcal{D}_0$ be one of the subsets in a partition of $\mathcal{D}$ that realizes the partition number.

First perform an isotopy to the discs not in $\mathcal{D}_0$ to remove any simple closed double curves that lie in $\bigcup \mathcal{D}_0$. This is achieved by an elementary innermost curve argument.

We now perform isotopies on the discs not in $\mathcal{D}_0$, supported in a regular neighbourhood of $\bigcup \mathcal{D}_0$, so that afterwards any two double curves in $\bigcup \mathcal{D}_0$ intersect in at most one point. For if not, two double curves form a bigon in some disc in $\mathcal{D}_0$. We may remove an innermost bigon by an isotopy. Note that this introduces no simple closed double curves in $\bigcup \mathcal{D}_0$. After this isotopy, there are only finitely many possibilities for the double curves in $\bigcup \mathcal{D}_0$, up to ambient isotopy fixed on the boundary of $\bigcup \mathcal{D}_0$.

Throughout these isotopies, discs which were initially disjoint have remained so. Thus, the partition number has not increased. Cut the ambient manifold along $\bigcup \mathcal{D}_0$ to a give a 3–manifold with boundary S', say. The remaining discs of $\mathcal{D}$ are cut up to form a disc arrangement $\mathcal{D}'$ in the cut-open 3–manifold. Its boundary curves arise from the boundary curves of $\mathcal{D} - \mathcal{D}_0$ and from the double curves in $\bigcup \mathcal{D}_0$. Thus, there are only finitely many possibilities for these boundary curves.

Now, $\mathcal{D}'$ has lower partition number than the original $\mathcal{D}$. Hence, inductively, $\mathcal{D}'$ may be isotoped, keeping its boundary fixed, so that afterwards, it belongs to one of only finitely many relative equivalence classes of disc arrangements. Thus, the same is true of $\mathcal{D}$. Also, a regular neighbourhood of $S' \cup \bigcup \mathcal{D}'$ is a punctured compression body with positive boundary S', and so a regular neighbourhood of $S \cup \bigcup \mathcal{D}$ is as required. □

We will always take any given finite subcollection of the discs D_n to be one of the disc arrangements as in Lemma 1.

We define the *complexity* of a closed orientable surface to be twice its genus, minus the number of its non-spherical components. If it is a union of 2–spheres or it is empty, we take its complexity to be zero. Thus, complexity is a non-negative integer, with the property that if the surface is compressed, its complexity goes down. So, if one compression body R' is embedded within another, R, and they have the same positive boundary, then the complexity of $\partial_- R$ is at most the complexity of $\partial_- R'$. This is an equality if and only if $R - R'$ is a regular neighbourhood of a subsurface of $\partial_- R$.

Let $s = \max\{n : n \in \mathcal{S}\}$. For integers $s < a \leq b$, we define $H(a,b)$ to be the submanifold of H formed by taking a regular neighbourhood of $\partial H \cup \bigcup_{a \leq n \leq b} D_n$, and then filling in any 2–sphere boundary components with 3–balls. This is a compression body with positive boundary ∂H. Let $h(a,b)$ be the complexity of its negative boundary. Note that if $a \leq a'$ and $b \leq b'$, then $h(a,b)$ and $h(a',b')$ are at least $h(a,b')$.

We term two pairs of integers (a,b) and (a',b') *comparable* if $f^{a'-a}$ takes the curves $\partial D_a, \dots, \partial D_b$ to the curves $\partial D_{a'}, \dots, \partial D_{b'}$, and this extends to an equivalence of disc arrangements. Note that in this case, $f^{a'-a}$ extends to a homeomorphism of $H(a,b)$ to $H(a',b')$.

We claim that we can find comparable pairs of integers (a,b) and (a',b'), where $a < a'$, such that $h(a,b) = h(a,b') = h(a',b')$. Theorem 3 follows very quickly. For, $H(a,b')$ is then ambient isotopic to both $H(a,b)$ and $H(a',b')$. So $f^{a'-a}$ extends to a homeomorphism of $H(a,b)$ to itself. Hence this power of f partially extends to H.

Suppose the claim were not true. Then we will show by induction that, for each positive integer j, there are non-negative integers $p(j)$ and $m(j)$ and a sequence $\{k(j,n) : n \geq s\}$ such that for each $n \geq s$,

(i) $h(k(j,n), k(j,n) + p(j)) \leq 2\ \mathrm{genus}(\partial H) - j$;

(ii) $0 \leq k(j,n) - n \leq m(j)$.

This will lead to a contradiction, since $h(a,b)$ is non-negative, but (i) implies that it is negative for large j.

The induction starts with $j = 1$, where we take $m(1) = p(1) = 0$ and $k(j,n) = n$. We now prove the inductive step. Each interval $[k(j,n), k(j,n) + p(j)]$ corresponds to the boundary of discs. There are a finite number of choices ($|\mathcal{C}|^{p(j)+1}$) for the boundary curves of these discs. By Lemma 1, we may assume that these discs form one of only finitely many (d, say) disc arrangements.

Thus, we assign to each integer $k(j,n)$ one of these d colours. If $k(j,n')$ and $k(j,n'')$ are given the same colour, then the pairs $(k(j,n'),k(j,n')+p(j))$ and $(k(j,n''),k(j,n'')+p(j))$ are comparable.

For any $n \geq s$, consider the values of $k(j,n')$ as n' varies between n and $n+(m(j)+1)d$. By (ii) at most $(m(j)+1)$ different n' can give the same value of $k(j,n')$. Therefore, the $k(j,n')$ take more than d different values. Therefore, we can find integers n' and n'' in this interval coloured with the same colour, such that $k(j,n') \neq k(j,n'')$. Say that $k(j,n') < k(j,n'')$.

We let

$$\begin{aligned} k(j+1,n) &= k(j,n') \\ p(j+1) &= (m(j)+1)d + m(j) + p(j) \\ m(j+1) &= (m(j)+1)d + m(j). \end{aligned}$$

We need to check that (i) and (ii) hold for $j+1$. Since we are assuming that the claim is not true, it must be the case that $h(k(j,n'),k(j,n'')+p(j))$ is strictly less than at least one of $h(k(j,n'),k(j,n')+p(j))$ and $h(k(j,n''),k(j,n'')+p(j))$. Hence, by (i), $h(k(j,n'),k(j,n'')+p(j)) \leq 2\ \text{genus}(\partial H) - j - 1$. So,

$$\begin{aligned} h(k(j+1,n),k(j+1,n)+p(j+1)) &= h(k(j,n'),k(j,n')+p(j+1)) \\ &\leq h(k(j,n'),k(j,n'')+p(j)) \\ &\leq 2\ \text{genus}(\partial H) - j - 1, \end{aligned}$$

since

$$\begin{aligned} &k(j,n') + p(j+1) - k(j,n'') - p(j) \\ &\qquad \geq n' + (m(j)+1)d + m(j) + p(j) - n'' - m(j) - p(j) \geq 0. \end{aligned}$$

This verifies (i) for $j+1$.

To verify (ii), note that

$$k(j+1,n) = k(j,n') \geq n' \geq n,$$

$$k(j+1,n) = k(j,n') \leq n' + m(j) \leq n + (m(j)+1)d + m(j) = n + m(j+1).$$

This establishes the claim, and hence completes the proof of Theorem 3.

We conclude with a proof of Theorem 4. So, let $f\colon \partial H \to \partial H$ be a pseudo-Anosov homeomorphism. Suppose that, for infinitely many n, $M \cup_{f^n \circ \phi} H$ fails to satisfy the conclusions of Theorem 1. Then, for these n, $f^n\phi(C_{i(n)})$ bounds a disc D_n in H, for some exceptional curve $C_{i(n)}$. We may pass to a monotone subsequence where $C_{i(n)}$ is some fixed curve C. But $f^n\phi(C)$ tends in $PL(\partial H)$ to the stable or unstable lamination of f, according to whether the sequence is decreasing or increasing. Hence, this lamination lies in the closed set $B(\partial H)$. This completes the proof of Theorem 4.

References

[1] **A Basmajian**, *Tubular neighbourhoods of totally geodesic hypersurfaces in hyperbolic manifolds*, Invent. Math. 117 (1994) 207–225

[2] **S Bleiler**, **C Hodgson**, *Spherical space forms and Dehn filling*, Topology 35 (1996) 809–833.

[3] **A Fathi**, **F Laudenbach**, **V Poenaru** et al, *Travaux de Thurston sur les surfaces*, Asterisque 66–67 (1979)

[4] **S Gallot**, **D Hulin**, **J Lafontaine**, *Riemannian geometry*, Springer–Verlag (1991)

[5] **J Hass**, *Minimal surfaces in manifolds with S^1 actions and the simple loop conjecture for Seifert fibred spaces*, Proc. Amer. Math. Soc. 99 (1987) 383–388

[6] **J Hass**, **H Rubinstein**, **S Wang**, *Boundary slopes of immersed surfaces in 3–manifolds*, J. Differential Geom. 52 (1999) 303–325

[7] **J Hass**, **S Wang**, **Q Zhou**, *On finiteness of the number of boundary slopes of immersed surfaces in 3–manifolds*, Proc. Amer. Math. Soc. 130 (2002) 1851–1857

[8] **C Hodgson**, **S Kerckhoff**, *Universal bounds for hyperbolic Dehn surgery*, preprint

[9] **M Lackenby**, *Word hyperbolic Dehn surgery*, Invent. Math. 140 (2000) 243–282

[10] **H Masur**, *Measured foliations and handlebodies*, Ergod. Th. Dynam. Sys. 6 (1986) 99–116

[11] **J Morgan**, *On Thurston's uniformization theorem for three-dimensional manifolds*, in The Smith conjecture (New York, 1979) 37–125, Pure Appl. Math., 112, Academic Press.

[12] **M Scharlemann**, and **Y-Q Wu**, *Hyperbolic manifolds and degenerating handle additions*, J. Austral. Math. Soc. 55 (1993) 72–89

[13] **W Thurston**, *The geometry and topology of three-manifolds*, Princeton (1980)

[14] **W Thurston**, *On the geometry and dynamics of diffeomorphisms of surfaces*, Bull. Amer. Math. Soc. 19 (1989) 417–431