

A Classification of Automorphisms of Compact 3-Manifolds

Leonardo Navarro Carvalho* & Ulrich Oertel

October, 2005

Abstract

We classify isotopy classes of automorphisms (self-homeomorphisms) of 3-manifolds satisfying the Thurston Geometrization Conjecture. The classification is similar to the classification of automorphisms of surfaces developed by Nielsen and Thurston, except an automorphism of a reducible manifold must first be written as a suitable composition of two automorphisms, each of which fits into our classification. Given an automorphism, the goal is to show, loosely speaking, either that it is periodic, or that it can be decomposed on a surface invariant up to isotopy, or that it has a “dynamically nice” representative, with invariant laminations that “fill” the manifold.

We consider automorphisms of irreducible and boundary-irreducible 3-manifolds as being already classified, though there are some exceptional manifolds for which the automorphisms are not understood. Thus the paper is particularly aimed at understanding automorphisms of reducible and/or boundary reducible 3-manifolds.

Previously unknown phenomena are found even in the case of connected sums of $S^2 \times S^1$'s. To deal with this case, we prove that a minimal genus Heegaard decomposition is unique up to isotopy, a result which apparently was previously unknown.

Much remains to be understood about some of the automorphisms of the classification.

1 Introduction

There is a large body of work studying the mapping class groups and homeomorphism spaces of reducible and ∂ -reducible manifolds. Without giving references, some of the researchers who have contributed are: E. César de Sá, D. Gabai, M. Hamstrom, A. Hatcher, H. Hendriks, K. Johannson, J. Kalliongis, F. Laudenbach, D. McCullough,

*Supported by FAPESP fellowship 03/06914-2, Brazil.

A. Miller, C. Rourke. Much less attention has been given to the problem of classifying or describing individual automorphisms up to isotopy. The paper [10] shows that if M is a Haken manifold or a manifold admitting one of Thurston's eight geometries, then any homeomorphism of M is isotopic to one realizing the Nielsen number.

In this paper we use the word “automorphism” interchangeably to mean either a self-homeomorphism or its isotopy class.

The paper [16] gives a classification up to isotopy of automorphisms (self-homeomorphisms) of 3-dimensional handlebodies and compression bodies, analogous to the Nielsen-Thurston classification of automorphisms of surfaces. Indecomposable automorphisms analogous to pseudo-Anosov automorphisms were identified and called *generic*. Generic automorphisms were studied further in [3] and [4]. Our understanding of generic automorphisms remains quite limited.

The goal of this paper is to represent an isotopy class of an automorphism of any compact, orientable 3-manifold M by a homeomorphism which is either periodic, has a suitable invariant surface on which we can decompose the manifold and the automorphism, or is dynamically nice in the sense that it has invariant laminations which “fill” the manifold. We do not quite achieve this goal. In some cases we must represent an automorphism as a composition of two automorphisms each of which fit into the classification.

Throughout this paper M will denote a compact, orientable 3-manifold, possibly with boundary.

If an automorphism $f : M \rightarrow M$ is not periodic (up to isotopy), then we search for a *reducing surface*, see Section 3. This is a suitable surface $F \hookrightarrow M$ invariant under f up to isotopy. In many cases, we will have to decompose M and f by cutting M on the surface F . Cutting M on F we obtain a manifold $M|F$ (M -cut-on- F) and if we isotope f such that it preserves F , we obtain an induced automorphism $f|F$ (f -cut-on- F) of $M|F$.

If there are no reducing surfaces for f (and f is not periodic), then the goal is to find a “dynamically nice” representative in the isotopy class of f . This representative will have associated invariant laminations. Also, just as in the case of pseudo-Anosov automorphisms, we will associate to the dynamically nice representative a growth rate. A “best” representative should, of course, have minimal growth rate. In the case of automorphisms of handlebodies it is still not known whether, in some sense, there is a canonical representative.

The reducing surfaces used in the analysis of automorphisms of handlebodies and compression bodies, see [16], have a nice property we call *rigidity*: A reducing surface F for $f : M \rightarrow M$ is *rigid* if f uniquely determines the isotopy class of the induced automorphism $f|F$ on $M|F$. As we shall see, the analysis of automorphisms of many 3-manifolds requires the use of non-rigid reducing surfaces F . This means that there are many choices for the induced isotopy class of automorphisms $f|F$, and it becomes more difficult to identify a “canonical” or “best” representative of $f|F$. For example, if F cuts M into two components, then on each component of $M|F$, the induced

automorphism $f|F$ is not uniquely determined. If there are no reducing surfaces for $f|F$ on each of these components, and they are not periodic, we again seek a dynamically nice representative, and in particular, the best representative should at least achieve the minimal growth rate of all possible “nice” representatives. But now we search for the best representative not only from among representatives of the isotopy class of $f|F$, but from among all representatives of all possible isotopy classes of $f|F$. This means one may have to isotope f using an isotopy not preserving F to find a better representative of $f|F$.

There is a subtlety in the definition of reducing surface. One must actually deal with automorphisms of *pairs* or *Haken pairs*, of the form (M, V) where M is irreducible and V is an incompressible surface in ∂M . An automorphism $f : (M, V) \rightarrow (M, V)$ of a Haken pair is an automorphism which restricts on V to an automorphism of V . Even in the case of automorphisms of handlebodies, Haken pairs play a role: The kind of Haken pair which arises is called a compression body.

A *compression body* is a Haken pair (Q, V) constructed from a product $V \times [0, 1]$ and a collection of balls by attaching 1-handles to $V \times 1$ on the product and to the boundaries of the balls to obtain a 3-manifold Q . Letting $V = V \times 0$, we obtain the pair (Q, V) . Here V can be a surface with boundary. The surface V is called the *interior boundary* $\partial_i Q$ of the compression body, while $W = \partial Q - \text{int}(V)$ is called the *exterior boundary*, $W = \partial_e Q$. We regard a handlebody H as a compression body whose interior boundary V is empty. An automorphism of a compression body is an automorphism of the pair (Q, V) .

A reducing surface F for $f : (M, V) \rightarrow (M, V)$ is always f -invariant up to isotopy, but may lie entirely in ∂M . If $F \subseteq \partial M$, F is incompressible in M and f -invariant up to isotopy, and if $F \supset V$ up to isotopy, then F is a reducing surface. In other words, if it is possible to rechoose the f -invariant Haken pair structure for (M, V) such that V is replaced by a larger F to obtain (M, F) , then f is reducible, and F is regarded as a reducing surface, and is called a *peripheral reducing surface*.

We shall give a more complete account of reducing surfaces in Section 3.

An automorphism is (*rigidly*) *reducible* if it has a (rigid) reducing surface.

Returning to the special case of automorphisms of handlebodies, when the reducing surface is not peripheral, then the handlebody can actually be decomposed to yield an automorphism $f|F$ of $H|F$. When the reducing surface is peripheral, and incompressible in the handlebody, the reducing surface makes it possible to regard the automorphism as an automorphism of a compression body whose underlying topological space is the handlebody, and the automorphism is then deemed “reduced.”

We continue this introduction with an overview of our current understanding of automorphisms of handlebodies. Without immediately giving precise definitions (see Sections 3 and 5 for details), we state one of the main theorems of [16], which applies to automorphisms of handlebodies.

Theorem 1.1. *Suppose $f : H \rightarrow H$ is an automorphism of a connected handlebody. Then the automorphism is:*

- 1) rigidly reducible,
- 2) periodic, or
- 3) generic on the handlebody.

A *generic automorphism* f of a handlebody is defined as an automorphism which is not periodic and does not have a reducing surface (see Section 3 for the precise definition). One can show that this amounts to requiring that $f|_{\partial H}$ is pseudo-Anosov and that there are no closed reducing surfaces.

There is a theorem similar to Theorem 1.1, classifying automorphisms of compression bodies.

In the paper [16], the first steps were taken towards understanding generic automorphisms using invariant laminations of dimensions 1 and 2. The theory has been extended and refined in [3] and [4]. We state here only the theorem which applies to automorphisms of handlebodies; there is an analogous theorem for arbitrary compression bodies, though in that case no 1-dimensional invariant lamination was described. We shall remedy that omission in this paper.

If H is a handlebody, in the following theorem H_0 is a *concentric* in H , meaning that H_0 is embedded in the interior of H such that $H - \text{int}(H_0)$ has the structure of a product $\partial H \times I$.

Theorem 1.2. *Suppose $f : H \rightarrow H$ is a generic automorphism of a 3-dimensional handlebody. Then there is a 2-dimensional measured lamination $\Lambda \hookrightarrow \text{int}(H)$ with transverse measure μ such that, up to isotopy, $f((\Lambda, \mu)) = (\Lambda, \lambda\mu)$ for some $\lambda > 1$. The lamination has the following properties:*

- 1) Each leaf ℓ of Λ is an open 2-dimensional disk.
- 2) The lamination Λ fills H_0 , in the sense that the components of $H_0 - \Lambda$ are contractible.
- 3) For each leaf ℓ of Λ , $\ell - \text{int}(H_0)$ is incompressible in $H - \text{int}(H_0)$.
- 4) $\Lambda \cup \partial H$ is a closed subset of H .

There is also a 1-dimensional lamination Ω , transverse to Λ , with transverse measure ν and a map $\omega : \Omega \rightarrow \text{int}(H_0)$ such that $f(\omega(\Omega, \nu)) = \omega(\Omega, \nu/\lambda)$. The map f is an embedding on $f^{-1}(N(\Lambda))$ for some neighborhood $N(\Lambda)$. The statement that $f(\omega(\Omega, \nu)) = \omega(\Omega, \nu/\lambda)$ should be interpreted to mean that there is an isomorphism $h : (\Omega, \nu) \rightarrow (\Omega, \nu/\lambda)$ such that $f \circ \omega = \omega \circ h$.

We note that the laminations in the above statement are “essential” only in a rather weak sense. For example, lifts of leaves of Λ to the universal cover of H are not necessarily properly embedded. The map ω need not be proper either: if we regarded $w : \Omega \rightarrow H$ as a homotopy class, in some examples there would be much unravelling. The lamination Ω can also be regarded as a “singular” embedded lamination, with finitely many singularities.

Remark 1.3. The laminations Λ and Ω should be regarded as a pair of dual laminations. Both are constructed from a chosen “complete system of essential discs,” \mathcal{E} ,

see Section 5. One attempts to find the most desirable pairs Λ and Ω . For example, it is reasonable to require laminations with minimal λ , but even with this requirement it is not known whether the invariant laminations are unique in any sense.

One of the authors has shown in his dissertation, [3] and in the preprint [4], that it is often possible to replace the invariant laminations in the above theorem by laminations satisfying a more refined condition. The condition on the 2-dimensional lamination is called *tightness* and is defined in terms of both invariant laminations. He shows that invariant lamination having this property necessarily achieve the minimum eigenvalue λ (growth rate) among all (suitably constructed) measured invariant laminations. One remaining problem is to show that every generic automorphism admits a tight invariant lamination; this is known in the case of an automorphism of a genus 2 handlebody and under certain other conditions. The growth rate λ for a tight invariant lamination associated to a generic automorphism f is no larger than the growth rate for the induced pseudo-Anosov automorphism on ∂H , see [3]. In particular, this means that for a genus 2 handlebody a growth rate can be achieved which is no larger than the growth rate of the pseudo-Anosov on the boundary. We hope that the tightness condition on invariant laminations will be sufficient to yield some kind of uniqueness. The most optimistic goal (perhaps too optimistic) is to show that a tight invariant 2-dimensional lamination for a given generic automorphism $f : H \rightarrow H$ is unique up to isotopy.

In our broader scheme for classifying automorphisms of arbitrary compact 3-manifolds, an important test case is to classify up to isotopy automorphisms $f : M \rightarrow M$, where M is a connected sum of $S^2 \times S^1$'s. In calculations of the mapping class group, this was also an important case, see [12] and [13]. The new idea for dealing with this case is that there is *always* a surface invariant up to isotopy, a reducing surface which, as we shall see, is non-rigid. This invariant surface is the Heegaard splitting surface coming from the Heegaard splitting in which M is expressed as the double of a handlebody. We call this a *symmetric* Heegaard splitting. A remark by F. Waldhausen at the end of his paper [19] outlines a proof that every minimal genus Heegaard splitting of a connected sum of $S^2 \times S^1$'s is symmetric. We prove that there is only one symmetric splitting up to isotopy, and we call this the *canonical splitting* of M . We were surprised not to find this result in the literature, and it is still possible that we have overlooked a reference.

Theorem 1.4. *Any two symmetric Heegaard splittings of a connected sum of $S^2 \times S^1$'s are isotopic.*

Thus when an automorphism f of M is isotoped so it preserves F , it induces automorphisms of H_1 and H_2 , which are mirror images of each other. However, there is not a unique way of isotoping an automorphism f of M to preserve F , see [6] and Example 2.7. Thus the canonical splitting of a connected sum of $S^2 \times S^1$'s gives a class of examples of non-rigid reducing surfaces.

In the important case where M is a connected sum of k copies of $S^2 \times S^1$, if $f : M \rightarrow M$ is an automorphism, we next consider the best choice of representative of $f|_F$, which is the induced automorphism on $M|_F$, a pair of genus k handlebodies. From the symmetry of the canonical splitting, the induced automorphisms of the two handlebodies are conjugate, and we denote each of these by g .

To state our theorem about automorphisms of connected sums of $S^2 \times S^1$'s, we first mention that given a generic automorphism g of a handlebody H and an invariant 2-dimensional lamination as in Theorem 1.2, the lamination gives a quotient map $H \rightarrow G$ where G is a graph, together with a homotopy equivalence of G induced by g . Typically, the 2-dimensional lamination *cannot* be chosen such that the induced homotopy equivalence on the quotient graph is a train track map in the sense of [1]. In the exceptional cases when there exists a 2-dimensional invariant lamination so the quotient homotopy equivalence is a train track map, we say g is a *train track generic automorphism*.

Theorem 1.5 (Train Track Theorem). *If M is a connected sum of $S^2 \times S^1$'s, $M = H_1 \cup H_2$ is the canonical splitting, and $f : M \rightarrow M$ is an automorphism, then f is non-rigidly reducible on the canonical splitting surface, and can be isotoped so that f preserves the splitting and the induced automorphism g on H_1 is*

- 1) *periodic,*
- 2) *rigidly reducible, or*
- 3) *train track generic.*

In the above, not so much information is lost in passing to the induced automorphism on $\pi_1(H_1)$, which is not surprising in view of Laudenbach's calculation of the mapping class group of M . F. Laudenbach showed in [12] that the map from the mapping class group of M to $\text{Out}(\pi_1(M))$ has kernel \mathbb{Z}_2^n generated by rotations in a maximal collection of n disjoint non-separating 2-spheres in M .

Note that when the automorphism g is generic, we have not shown that there is a unique pair of invariant laminations such that the quotient is a train track map.

For a typical generic automorphism, any invariant 1-dimensional lamination has much self-linking and is somewhat pathological. Even for a train-track generic map, some of the self-linking and pathology may remain. We say a 1-dimensional lamination $\Omega \hookrightarrow M$ is *tame* if any point of Ω has a flat chart neighborhood in M .

Conjecture 1.6 (Tameness Conjecture). *If M is a connected sum of $S^2 \times S^1$'s, $M = H_1 \cup H_2$ is the canonical splitting, and $f : M \rightarrow M$ is a train-track generic automorphism, then f can be isotoped so that f preserves the splitting and the induced automorphism g on H_1 is train-track generic and has a tame 1-dimensional invariant lamination.*

To proceed, we must say something about automorphisms of compression bodies. Generic automorphisms of compression bodies are defined like generic automorphisms

of handlebodies; in particular, if $f : Q \rightarrow Q$ is an automorphism, the induced automorphism $\partial_e f$ on $\partial_e Q$ is pseudo-Anosov. Also, there should be no (suitably defined) reducing surfaces (see Section 3).

We will adopt some further terminology for convenience. A connected sum of $S^2 \times S^1$'s is a *sphere body*. Removing open ball(s) from a manifold is called *holing* and the result is a *holed manifold*, or a manifold with sphere boundary components. Thus holing a sphere body yields a *holed sphere body*. A *mixed body* is a connected sum of $S^2 \times S^1$'s and handlebodies. If we accept a ball as a special case of a handlebody, then a holed mixed body is actually also a mixed body, since it can be obtained by forming a connect sums with balls. Nevertheless, we distinguish a *holed mixed body*, which has sphere boundary components, from a mixed body, which does not.

It is often convenient to cap sphere boundary components of a holed manifold with balls, distinguishing the caps from the remainder of the manifold. Thus given a manifold M with sphere boundary components, we cap with balls, whose union we denote \mathcal{D} , to obtain a *spotted manifold* (with 3-dimensional balls). For classifying automorphisms, spotted manifolds are just as good as holed manifolds.

Suppose now that M is a mixed body. Again, there is a canonical Heegaard splitting $M = H \cup Q$, where H is a handlebody and Q is a compression body, see Theorem 2.4. Whereas in the decomposition of an irreducible manifold with compressible boundary the Bonahon characteristic compression body has exterior boundary in the boundary of the manifold, our compression body Q here has interior boundary $\partial_i Q = \partial M$. Again, this splitting lacks rigidity for the same reasons.

In Section 7, we undertake the study of automorphisms of mixed bodies. We prove a theorem analogous to Theorem 1.5 (the Train Track Theorem), see Theorem 7.11.

In a classification of automorphisms of compact 3-manifolds, the most difficult case is the case of an arbitrary reducible 3-manifold. We will assume the manifold has no sphere boundary components; if the manifold has boundary spheres, we cap them with balls. Let B be the 3-ball. Suppose that a (possibly holed) mixed body $R \neq B$ is embedded in M . We say that R is *essential* if any essential sphere in R or sphere component of ∂R is also essential in M . An automorphism f of M which preserves an essential mixed body R up to isotopy and which is periodic on $M - \text{int}(R)$ is called an *adjusting automorphism*. The surface ∂R is f -invariant up to isotopy, but we do not regard it as a reducing surface, see Section 3.

Note that an essential mixed body R in an irreducible manifold is a handlebody. Since an adjusting automorphism preserving R is isotopic to a periodic map on ∂R , it is then also isotopic to a periodic map on R , so such an adjusting automorphism is not especially useful.

For the classification of automorphisms of reducible compact 3-manifolds, we will use the following result, which is based on a result due to E. César de Sá [5], see also M. Scharlemann in Appendix A of [2], and D. McCullough in [15] (p. 69).

Theorem 1.7. *Suppose M has irreducible summands M_i , $i = 1, \dots, k$, and suppose \hat{M}_i is M_i with one hole, and one boundary sphere S_i . Suppose the \hat{M}_i are disjoint*

submanifolds of M , $i = 1, \dots, k$. Then the closure of $M - \cup_i \hat{M}_i$ is an essential holed sphere body \hat{M}_0 . (\hat{M}_0 is a holed connected sum of $S^2 \times S^1$'s or possibly a holed sphere.) If $f : M \rightarrow M$ is an automorphism, then there is an adjusting automorphism \bar{h} preserving an essential (possibly holed) mixed body $R \supset \hat{M}_0$ in M with the property that $g = \bar{h} \circ f$ is rigidly reducible on the spheres of $\partial \hat{M}_0$.

The automorphism g is determined by f only up to application of an adjusting automorphism preserving $\partial \hat{M}_0$.

Note that we can write f as a composition $f = \bar{h}^{-1} \circ g = h \circ g$, where $h = \bar{h}^{-1}$ is an adjusting automorphism and g preserves each \hat{M}_i .

The automorphism $\hat{g}_i = g|_{\hat{M}_i}$ clearly uniquely determines an automorphism g_i of M_i by capping (for $i = 0$, M_0 is \hat{M}_0 capped). On the other hand g_i determines \hat{g}_i only up to application of an adjusting automorphism in \hat{M}_i .

The main ingredients used in this result are *slide automorphisms* of M . There are two types, one of which is as defined in [15]. We shall describe them briefly in Section 4.

With a little work, see Section 4, we can reformulate Theorem 1.7 in terms of 4-dimensional compression bodies as follows. Let Q be a 4-dimensional compression body constructed from a disjoint union of products $M_i \times I$ and a 4-dimensional ball K_0 by attaching one 1-handle joining $\partial K_0 = M_0$ to $M_i \times 1$ and one 1-handle from ∂K_0 to itself for every $S^2 \times S^1$ summand of M . Then $\partial_e Q$ is homeomorphic with and can be identified with M , while $\partial_i Q$ is homeomorphic to and can be identified with the disjoint union of M_i 's, $i \geq 1$.

Corollary 1.8. *With the hypotheses of Theorem 1.7 and with the 4-dimensional compression body Q constructed as above, there is an automorphism $\bar{f} : Q \rightarrow Q$ of the compression body such that $\bar{f}|_{\partial_e Q} = f$. Also, there is an automorphism $\bar{h} : Q \rightarrow Q$ of the compression body such that $\bar{h}|_{\partial_e Q} = \bar{h}$, $\bar{h}|_{\partial_i Q} = id$. Also $\bar{h} \circ \bar{f}$ gives the adjusted automorphisms on $\partial_i Q$, which is the union of M_i 's, $i \geq 1$.*

In other words, the automorphism $f : M \rightarrow M$ is cobordant over Q to an automorphism of the disjoint union of the M_i 's via an automorphism \bar{h} of Q rel $\partial_i Q$.

We will refer to the automorphism \bar{h} as an *adjusting automorphism of the compression body Q* .

Before turning to the classification of automorphisms, we mention another necessary and important ingredient.

Theorem 1.9. *(F. Bonahon's characteristic compression body) Suppose M is a compact manifold with boundary. Then there is a compression body $Q \hookrightarrow M$ which is unique up to isotopy with the property that $\partial_e Q = \partial M$ and $\partial_i Q$ is incompressible in $\overline{M - Q}$.*

Bonahon's theorem gives a canonical decomposition of any compact irreducible 3-manifold with boundary, dividing it into a compression body and a compact irre-

ducible manifold with incompressible boundary. The interior boundary of the characteristic compression body is a rigid reducing surface.

Finally, in the classification, we often encounter automorphisms of *holed manifolds*. That is to say, we obtain an automorphism $f : \hat{M} \rightarrow \hat{M}$ where \hat{M} is obtained from M by removing a finite number of open balls. We have already observed that an automorphism of a holed manifold \hat{M} is essentially the same as an automorphism of manifold pair (M, \mathcal{D}) where \mathcal{D} is the union of closures of the removed open balls, i.e. $\mathcal{D} = M - \hat{M}$. An automorphism of a spotted manifold (M, \mathcal{D}) or holed manifold \hat{M} determines an automorphism of the manifold M obtained by forgetting the spots or capping the boundary spheres of \hat{M} . Clearly, there is a loss of information in passing from the automorphism \hat{f} of the spotted or holed manifold to the corresponding automorphism f of the unspotted or unholed manifold. The relationship is that \hat{f} defined on the spotted manifold is isotopic to f if we simply regard \hat{f} as an automorphism of M . To obtain \hat{f} from f , we compose with an isotopy h designed to ensure that $h \circ f$ preserves the spots. There are many choices for h , and two such choices h_1 and h_2 differ by an automorphism of (M, B) which can be realized as an isotopy of M . This means $h_1 \circ h_2^{-1}$ is an automorphism of (M, B) realized by an isotopy of M . In describing the classification, we will not comment further on the difference between automorphisms of manifolds and automorphisms of their holed counterparts. Of course any manifold with sphere boundary components can be regarded as a holed manifold. Much more can be said about automorphisms of spotted manifolds, but we will leave this topic to another paper.

Automorphisms of Irreducible and ∂ -irreducible Manifolds

The following outlines the classification of automorphisms of irreducible and ∂ -irreducible manifolds. If the characteristic manifold (Jaco-Shalen, Johannson) is non-empty or the manifold is Haken, we use [9], [11]. For example, if the characteristic manifold is non-empty and not all of M , then the boundary of the characteristic manifold is a rigid incompressible reducing surface. There are finitely many isotopy classes of automorphisms of the complementary “simple” pieces [11]. For automorphisms of Seifert-fibered pieces we refer to [17, 10]. The idea is that in most cases an automorphism of such a manifold preserves a Seifert fibering, therefore the standard projection yields an automorphism of the base orbifold. The article [10] gives a classification of automorphisms of orbifolds similar to Nielsen-Thurston’s classification of automorphisms of surfaces. It is then used for a purpose quite distinct from ours: their goal is to realize the Nielsen number in the isotopy class of a 3-manifold automorphism. Still, their classification of orbifold automorphisms is helpful in our setting. The idea is that invariant objects identified by the classification in the base orbifold lift to invariant objects in the original Seifert fibered manifold.

In the following we let T^n be the n -torus, D^2 the 2-disc, I the interval (1-disc), $\mathbb{P}(2, 2)$ the orbifold with the projective plane as underlying space and two cone points

of order 2, and $M_{\mathbb{P}(2,2)}$ the Seifert fibered space over $\mathbb{P}(2,2)$.

Theorem 1.10 ([10]). *Suppose that M is a compact orientable Seifert fibered space which is not T^3 , $M_{\mathbb{P}(2,2)}$, $D^2 \times S^1$ or $T^2 \times I$. Given $f: M \rightarrow M$ there exists a Seifert fibration which is preserved by f up to isotopy.*

We discuss the classification of automorphisms of these four exceptional cases later. In the general case we consider the projection $M \rightarrow X$ over the fiber space X , which is an orbifold, and let the surface Y be the underlying space of X . If $f: M \rightarrow M$ preserves the corresponding fiber structure we let $\hat{f}: X \rightarrow X$ be the projected “automorphism” of X . An *automorphism of X* is an automorphism of Y preserving the orbifold structure. Let $\text{Sing}(X)$ be the set of singular points of X . Note that we are assuming that M is orientable therefore there are no reflector lines in $\text{Sing}(X)$, which then consists of isolated cone points. Moreover \hat{f} has to map a cone point to another of the same order. An isotopy of \hat{f} is an isotopy restricted to $X - \text{Sing}(X)$.

The classification of automorphisms of X is divided in three cases, depending on the sign of $\chi(X)$ (as an orbifold Euler characteristic). Below we essentially follow [10].

1. $\chi(X) > 0$. In this case, unless $\text{Sing}(X) = \emptyset$, the underlying space Y is either S^2 or \mathbb{P}^2 and X has a few singular points. Any map is isotopic to a periodic one.
2. $\chi(X) = 0$. For most orbifolds X any automorphism is isotopic to a periodic one. The only two exceptions are T^2 (without singularities) and $S^2(2,2,2,2)$, the sphere with four cone points of order 2.

If $X = T^2$ then an automorphism $\hat{f}: X \rightarrow X$ is isotopic to either i) a periodic one, ii) reducible, preserving a curve or iii) an Anosov map. In ii) the lift of an invariant curve in X yields a torus reducing surface for f in M (if the lift of the curve is a 1-sided Klein bottle we consider the torus boundary of its regular neighborhood). Such a torus may separate a $K \tilde{\times} I$ piece, which is the orientable I -bundle over the Klein bottle K . An automorphism of such a piece is isotopic to a periodic map. In iii) the invariant foliations lift to foliations on M invariant under f . The leaves are either open annuli or open Möbius bands.

If $X = S^2(2,2,2,2)$ then $X' = X - \text{Sing}(X)$ is a four times punctured sphere. Since $\chi(X') < 0$ the automorphism $\hat{f}|_{X'}$ is subject to Nielsen-Thurston’s classification, which isotopes it to be either i) periodic, ii) reducible or iii) pseudo-Anosov. As above, in ii) a reduction of \hat{f} yields a reduction of f along tori, maybe separating $K \tilde{\times} I$ pieces. In iii) we consider the foliations on X invariant under \hat{f} . They may have 1-prong singularities in $\text{Sing}(X)$. We can assume that there are no other singularities since $\chi(X) = 0$, with X covered by T^2 . But all cone points have order 2, therefore the lifts of the foliations to M are non-singular foliations invariant under f .

3. $\chi(X) < 0$. Also in this case $X' = X - \text{Sing}(X)$ is a surface with $\chi(X') < 0$, therefore $\hat{f}|_{X'}$ is subject to Nielsen-Thurston's classification. A reducing curve system for $\hat{f}|_{X'}$ is a reducing curve system for \hat{f} . As before, this yields a reducing surface for f consisting of tori. Some may be thrown out, in case there are parallel copies.

In the pseudo-Anosov case we consider the corresponding invariant laminations on X' , which lift to invariant laminations on M by vertical open annuli or Möbius bands.

In the above we did not say anything about $f: M \rightarrow M$ when $\hat{f}: X \rightarrow X$ is periodic. There is not much hope in trying to isotope f to be periodic, since there are homeomorphisms of M preserving each fiber but not periodic (e.g, "vertical" Dehn twists along vertical tori). What can be done is to write f as a composition $f = g \circ h$ where h is periodic and g preserves each fiber.

The outline above deals with Seifert fibered spaces M for which there is a fibering structure preserved by $f: M \rightarrow M$. Recall that this excludes T^3 , $M_{\mathbb{P}(2,2)}$, $D^2 \times S^1$ or $T^2 \times I$ 1.10, which we consider now. If $M = D^2 \times S^1$ then f is isotopic to a power of a Dehn twist along the meridional disc. If $M = T^2 \times I$ then f is isotopic to a product $g \times (\pm \text{Id}_I)$, where Id_I is the identity and $-\text{Id}_I$ is the orientation reversing isometry of I . Assume that $g: T^2 \rightarrow T^2$ is either periodic, reducible (along a curve) or Anosov.

In the case $M = T^3$ any automorphism is isotopic to a *linear automorphism* (i.e., an automorphism with a lift to $\widetilde{M} = \mathbb{R}^3$ which is linear). The study of eigenspaces of such an automorphism yield invariant objects. Roughly, such an automorphism may be periodic, reducible (preserving an embedded torus) or may have invariant 2-dimensional foliations, whose leaves may consist of dense open annuli or planes. In some cases it may also be relevant to further consider invariant 1-dimensional foliations.

In the case $M = M_{\mathbb{P}(2,2)}$ there is a cell decomposition which is preserved by any automorphism up to isotopy. It thus imply that any automorphism is isotopic to a periodic one.

This last case completes the outline of the classification for Seifert fibered spaces.

If the irreducible and ∂ -irreducible manifold M is closed and hyperbolic, it is known that any automorphism is periodic, see [18]. If the manifold is closed and spherical, recent results of D. McCullough imply the automorphisms are periodic, see [14].

The Classification

The following outline classifies automorphisms of 3-manifolds in terms of automorphisms of compression bodies, automorphisms of mixed bodies, and automorphisms of irreducible, ∂ -irreducible 3-manifolds. (Here we regard handlebodies as compression bodies.) In every case, we consider an automorphism $f: M \rightarrow M$.

i) If M is reducible use Theorem 1.7 and write $f = h \circ g$, where g is rigidly reducible on essential spheres, and h is non-rigidly reducible. The reductions yield automorphisms of holed irreducible 3-manifolds (the irreducible summands of M with one hole each), and of (holed) mixed bodies, which we consider in the following cases. The automorphism h also yields, after cutting on ∂R of Theorem 1.7), periodic automorphisms of (holed) irreducible manifolds with incompressible boundary.

Alternatively, using Corollary 1.8, the automorphism $f : M \rightarrow M$ is cobordant over a 4-dimensional compression body to an “adjusted” automorphism f_a of the disjoint union of the irreducible summands M_i .

ii) If M is irreducible, with $\partial M \neq \emptyset$ let Q be the characteristic compression body of Theorem 1.9, then f induces unique automorphisms on Q and $\overline{M - Q}$, see [2] and [16]. Q is a compression body and $\overline{M - Q}$ is irreducible and ∂ -irreducible, whose automorphisms we consider in following cases.

iii) If M is a holed sphere body (a connected sum of $S^2 \times S^1$'s with open balls removed), as is the case when the automorphism is obtained by decomposition from an adjusting automorphism for a different automorphism of a different 3-manifold, then we first cap the boundary spheres to obtain an induced automorphism of a sphere body. Then we use the Train Track Theorem (Theorem 1.5 and Theorem 7.11), to obtain a particularly nice representative of f . If the Tameness Conjecture, Conjecture 1.6, is true, then one can find an even nicer representative.

iv) If M is a (holed) mixed body, as is the case when the automorphism is an adjusting automorphism for a different automorphism of a different 3-manifold, then we use the canonical splitting surface (Theorem 2.4) as a non-rigid reducing surface for f which cuts M into a handlebody H and a compression body Q . In the handlebody, there is another non-rigid reducing surface (see Section 7), which makes it possible to cut the handlebody H into a compression body Q' isomorphic to Q , and a possibly disconnected handlebody H' , with a component of genus g corresponding to each genus g component of ∂M . Theorem 7.11 gives a best representative, yielding the simplest invariant laminations for f restricted to Q , Q' , and H' . The restrictions of f to Q and Q' are then automorphisms of compression bodies periodic on the interior boundary.

v) If M is a compression body, the automorphisms are classified by Theorem 1.1 and its analogue, see [16], which says the automorphism is periodic, rigidly reducible or generic. See Section 5 for a description of generic automorphisms.

vi) If M is irreducible and ∂ -irreducible, then the automorphism f is classified following the previous subsection.

—

If it were true that every automorphism of a reducible manifold had a closed reducing surface, possibly a collection of essential spheres, then decomposition on the reducing surface would make it possible to analyze the automorphism of the resulting manifold-with-boundary in a straightforward way, without the necessity of

first writing the automorphism as a composition of an adjusting automorphism and an automorphism preserving a system of reducing spheres. This would yield a dramatic improvement of our classification. For this reason, we pose the following:

Question 1.11. *Is it true that every automorphism of a reducible manifold admits a closed reducing surface, which cannot be isotoped to be disjoint from a system of essential spheres cutting M into holed irreducible manifolds? In particular, does every automorphism of a closed reducible 3-manifold admit a closed reducing surface of this kind?*

To understand the significance of this question, one needs a more detailed understanding of reducing surfaces. Assuming that an automorphism $f : M \rightarrow M$ has no reducing surfaces consisting of an f -invariant collection of essential spheres, a closed reducing surface F is required to have the property that $M|F$ is irreducible, see Section 3.

We briefly describe the much improved classification of automorphisms which would be possible if the answer to Question 1.11 were positive. If $f : M \rightarrow M$ is an automorphism and M is reducible, we first seek reducing surfaces consisting of essential spheres. If these exist, we decompose and cap the spheres. If there are no sphere reducing surfaces, we seek closed reducing surfaces. Decomposing on such a reducing surface gives an automorphism of an irreducible manifold with non-empty boundary, where the irreducibility of the decomposed manifold comes from the requirement that cutting on a reducing surface should yield an irreducible manifold. For the resulting irreducible manifold with boundary, we decompose on the interior boundary of the characteristic compression body as before, obtaining an automorphism of a manifold with incompressible boundary and an automorphism of a compression body. (Either one could be empty, and the compression body could have handlebody component(s).) Automorphisms of compression bodies are analyzed using the existing (incomplete) theory, the automorphism of the irreducible, ∂ -irreducible 3-manifold is analyzed as in the classification given above, using mostly well-established 3-manifold theory.

2 Canonical Heegaard splittings

Given a handlebody H of genus g we can double the handlebody to obtain a connected sum M of g copies of $S^2 \times S^1$, a genus g sphere body. We choose g non-separating discs in H whose doubles give g non-separating spheres in M . We then say that the Heegaard splitting $M = H \cup (M - \text{int}(H))$ is a *symmetric Heegaard splitting*. The symmetric splitting is characterized by the property that any curve in ∂H which bounds a disc in H also bounds a disc in $M - \text{int}(H)$.

The following is our interpretation of a comment at the end of a paper of F. Waldhausen, see [19].

Lemma 2.1. *(F. Waldhausen) Any minimal genus Heegaard splitting of a sphere body is symmetric.*

We restate a theorem which we will prove in this section:

(Theorem 1.4). *Any two symmetric Heegaard splittings of a sphere body are isotopic.*

One would expect, from the fundamental nature of the result, that this theorem would have been proved much earlier. A proof may exist in the literature, but we have not found it.

In the first part of this section, M will always be a genus g sphere body. An *essential system of spheres* is the isotopy class of a system of disjointly embedded spheres with the property that none of the spheres bounds a ball. The system is *primitive* if no two spheres of the system are isotopic. A system \mathcal{S} is said to be *complete* if cutting M on \mathcal{S} yields a collection of holed balls. There is a cell complex of weighted primitive systems of essential spheres in M constructed as follows. For any primitive essential system \mathcal{S} in M , the space contains a simplex corresponding to projective classes of systems of non-negative weights on the spheres. Two simplices corresponding to primitive essential systems \mathcal{S}_1 and \mathcal{S}_2 are identified on the face corresponding to the maximal primitive essential system \mathcal{S}_3 of spheres common to \mathcal{S}_1 and \mathcal{S}_2 . We call the resulting space the *sphere space* of M . In the sphere space, the union of open cells corresponding to complete primitive systems is called the *complete sphere space* of M .

Lemma 2.2. *The complete sphere space of a sphere body M is connected. In other words, it is possible to pass from any primitive complete essential system to any other by moves which replace a primitive complete essential system by a primitive complete subsystem or a primitive complete essential supersystem.*

Note: A. Hatcher proved in [8] that the sphere space is contractible. It may be possible to adapt the following proof to show that the complete sphere space is contractible.

Proof. Suppose $\mathcal{S} = \mathcal{S}_0$ and \mathcal{S}' are two primitive complete systems of essential spheres. We will abuse notation by referring to the union of spheres in a system by the same name as the system. Without loss of generality, we may assume \mathcal{S} and \mathcal{S}' are transverse.

Consider $\mathcal{S} \cap \mathcal{S}'$ and choose a curve of intersection innermost in \mathcal{S}' . Remove the curve of intersection by isotopy of \mathcal{S} if possible. Otherwise, surger \mathcal{S} on a disc bounded in \mathcal{S}' by the curve to obtain \mathcal{S}_2 . Now \mathcal{S}_2 is obtained from $\mathcal{S} = \mathcal{S}_0$ by removing one sphere, the surgered sphere S_0 of \mathcal{S}_0 , say, and adding two spheres, S_{21} and S_{22} , which result from the surgery. We can do this in two steps, first adding the two new spheres to obtain \mathcal{S}_1 , then removing the surgered sphere to obtain \mathcal{S}_2 . If one of the spheres S_{21} and S_{22} is inessential, this shows that the curve of intersection between $\mathcal{S} = \mathcal{S}_0$ and \mathcal{S}' could have been removed by isotopy. We note that S_{21} or S_{22} , or both, might be isotopic to spheres of \mathcal{S}_0 . If the new spheres are essential, we claim that \mathcal{S}_1 and

\mathcal{S}_2 are complete systems of essential spheres. It is clear that \mathcal{S}_1 is complete, since it is obtained by adding two essential spheres S_{21} or S_{22} to \mathcal{S}_0 . We then remove S_0 from \mathcal{S}_1 to obtain \mathcal{S}_2 . There is, of course a danger that removing an essential sphere from an essential system \mathcal{S}_1 yields an incomplete system. However, if we examine the effect of surgery on the complementary holed balls, we see that first we cut one of them on a properly embedded disc (the surgery disc), which clearly yields holed balls. Then we attach a 2-handle to the boundary of another (or the same) holed ball, which again yields a holed ball. Thus, \mathcal{S}_2 is complete. If \mathcal{S}_2 is not primitive, i.e. contains two isotopic spheres, we can remove one, until we are left with a primitive complete essential system.

Inductively, removing curves of intersection of \mathcal{S}_k with \mathcal{S}' either by isotopy or by surgery on an innermost disc in \mathcal{S}' as above, we must obtain a sequence ending with \mathcal{S}_{n-2} disjoint from \mathcal{S}' . Next, we add the spheres of \mathcal{S}' not in \mathcal{S}_{n-2} to \mathcal{S}_{n-2} to obtain \mathcal{S}_{n-1} , which is still complete. Finally we remove all spheres from \mathcal{S}_{n-1} except those of \mathcal{S}' to get $\mathcal{S}_n = \mathcal{S}'$. \square

A ball with r spots, or an r -spotted ball is a pair (K, E) where K is a 3-ball and E is consists of r discs in ∂K . Let P be a holed ball, with $r \geq 1$ boundary components say. Then we say $(K, E) \hookrightarrow (P, \partial P)$ is *standard* if the closure of $P - K$ in P is another r -spotted ball. In that case, the complementary spotted ball is also standard. In practical terms, a standard r -spotted ball is “unknotted.” Another point of view is the following: If $(K, E) \hookrightarrow (P, \partial P)$ is standard, then $(P, \partial P) = (K, E) \cup (K', E')$, where (K', E') complementary to (K, E) , is a “symmetric” splitting of $(P, \partial P)$, similar to a symmetric splitting of a sphere body.

Lemma 2.3. (a) *A standard embedded r -spotted ball (K, E) in an r -holed 3-sphere $(P, \partial P)$ ($r \geq 1$) is unique up to isotopy of pairs.*

(b) *Suppose (K, E) and (K', E') are two different standard spotted balls in $(P, \partial P)$ and suppose $\hat{E} \subseteq E$ is a union of spots in E , which coincides with a union of spots in E' . Then there is an isotopy rel \hat{E} of (K, E) to (K', E') .*

Proof. (a) We will use the following result. (It is stated, for example, in [15].) The (mapping class) group of isotopy classes of automorphisms of $(P, \partial P)$ which map each component of ∂P to itself is trivial. It is then also easy to verify that the mapping class group of $(P, \partial P)$ rel ∂P is trivial.

Suppose now that (K, E) and (K', E') are two different standard spotted balls in $(P, \partial P)$. We will construct an automorphism f of $(P, \partial P)$ by defining it first on K , then extending to the complement. We choose a homeomorphism $f : (K, E) \rightarrow (K', E')$. Next, we extend to $\partial P - E$ such that $f(\partial P - E) = \partial P - E'$. Since (K, E) is standard, the sphere $\text{fr}(K) \cup (\partial P - E)$ bounds a ball K^c (with interior disjoint from $\text{int}(K)$), and similarly $\text{fr}(K') \cup (\partial P - E')$ bounds a ball $(K')^c$. Thus we can extend f so that the ball K^c is mapped homeomorphically to $(K')^c$. Then f is isotopic to the identity, whence (K, E) is isotopic to (K', E') .

(b) To prove this, we adjust the isotopy of (a) in a collar of ∂P . To see that such an adjustment is possible, note that we may assume the collar is chosen sufficiently small so that by transversality the isotopy of (a) is a “product isotopy” in the collar, i.e., has the form $h_t \times \text{id}$ for an isotopy h_t of ∂P . The adjustment of the isotopy in the collar is achieved using isotopy extension. \square

Proof of Theorem 1.4. Any complete system of essential spheres \mathcal{S} gives a symmetric Heegaard splitting as follows. Each holed 3-sphere $(P_i, \partial P_i)$ obtained by cutting M on \mathcal{S} (where P_i has r_i holes, say) can be expressed as a union of two standard r_i -spotted balls (K_i, E_i) and (K_i^c, E_i^c) . We assemble the spotted balls (K_i, E_i) by isotoping and identifying the two spots on opposite sides of each essential sphere of the system \mathcal{S} , extending the isotopy to K_i and to K_i^c . The result is a handlebody $H = \cup_i K_i$ with a complementary handlebody $H^c = \cup_i K_i^c$.

We claim that H (depending on \mathcal{S}) is unique up to isotopy in M . First replace each sphere S of \mathcal{S} by $\partial N(S)$, i.e. by two parallel copies of itself. The result is that no holed ball obtained by cutting M on \mathcal{S} has two different boundary spheres identified in M . Suppose H' is constructed in the same way as H above. Inductively, we isotope $H' \cap P_i = K_i'$ to K_i using Lemma 2.3, extending the isotopy to H' . Having isotoped K_i' to K_i for $i < k$, we isotope K_k' to K_k in $P_k \text{ rel } \cup_{i < k} K_i' \cap \partial P_k$. This means we perform the isotopy rel spots on the boundary of P_k where previous K_i' 's have already been isotoped to coincide with K_i 's. Now it is readily verified that the H associated to the non-primitive \mathcal{S} obtained by replacing each sphere with two copies of itself is the same as the H associated to the original \mathcal{S} .

Next we must show that H does not depend on the choice of a complete system \mathcal{S} of essential spheres. By Lemma 2.2 it is possible to find a sequence $\{\mathcal{S}_i\}$ of complete systems of essential spheres such that any two successive systems differ by the addition or removal of spheres. It is then easy to verify that after each move, the handlebody H_i , constructed from \mathcal{S}_i as H was constructed from \mathcal{S} above, will be the same up to isotopy. For example, when removing an essential sphere from the system \mathcal{S}_i to obtain \mathcal{S}_{i+1} , two spotted balls are identified on a disc on the removed sphere, yielding a new spotted ball in a new holed sphere, and $H_{i+1} = H_i$. The reverse process of adding an essential sphere must then also leave $H_{i+1} = H_i$ unchanged. \square

We offer an alternative simpler proof of the theorem. Although it is simpler, this approach has the disadvantage that it does not appear to work for our more general result about canonical splittings of mixed bodies.

Alternate proof of Theorem 1.4. Let $M = H_1 \cup_{\partial H_1 = \partial H_2} H_2$ and $M = H_1' \cup_{\partial H_1' = \partial H_2'} H_2'$ determine two symmetric Heegaard splittings of M . Our goal is to build an automorphism $h: M \rightarrow M$ taking H_1 to H_1' (and hence H_2 to H_2') which is isotopic to the identity.

Choose any automorphism $f: H_1 \rightarrow H_1'$. It doubles to an automorphism $Df: M \rightarrow M$ which takes the first Heegaard splitting to the second. Now consider $Df_*: \pi_1(M) \rightarrow$

$\pi_1(M)$. It is a known fact that any automorphism of a free group is realizable by an automorphism of a handlebody (e.g., [7]). Let $g: H_1 \rightarrow H_1$ be such that $g_* = (Df_*)^{-1}$ and double g , obtaining $Dg: M \rightarrow M$. Then it is clear that $h = (Df \circ Dg): M \rightarrow M$ takes the first Heegaard splitting to the second. Also, $h_* = \text{Id}_{\pi_1(M)}$.

Now let \mathcal{S} be a complete system of spheres symmetric with respect to the splitting $M = H_1 \cup H_2$. From $h_* = \text{Id}_{\pi_1(M)}$ follows that h is isotopic to a composition of twists on spheres of \mathcal{S} [12]. But such a composition preserves H_1, H_2 (a twist on such a sphere restricts to each handlebody as a twist on a disc), therefore h is isotopic to an automorphism h' preserving H_1, H_2 . The isotopy from h to h' then takes the second splitting to the first one. \square

Our next task is to describe canonical Heegaard splittings of mixed bodies. Until further notice, M will be a mixed body without holes. We obtain a canonical splitting of a holed mixed body simply by capping the holes.

Theorem 2.4. *Let M be a mixed body. There is a surface F bounding a handlebody H in M with the property that every curve on F which bounds a disc in H also bounds a disc in $M - \text{int}(H)$, or it is boundary parallel in M . F and H are unique up to isotopy. $M - \text{int}(H)$ is a compression body. Also $\pi_1(H) \rightarrow \pi_1(M)$ is an isomorphism.*

The proof is entirely analogous to the proof of Theorem 1.4, but we must first generalize the definitions given above, for M as sphere body, to definitions for M a mixed body. Let M be a connected sum of handlebodies $H_i, i = 1, \dots, k, H_i$ having genus g_i , and N , which is a connected sum of n copies of $S^2 \times S^1$.

An *essential reducing system* or *essential system* for M is the isotopy class of a system of disjointly embedded spheres and discs with the property that none of the spheres bounds a ball and none of the discs is ∂ -parallel. The system is *primitive* if no two surfaces of the system are isotopic. A system \mathcal{S} is said to be *complete* if cutting M on \mathcal{S} yields a collection of holed balls P_i satisfying an additional condition which we describe below. Cutting on \mathcal{S} yields components of the form (P_i, S_i) , where S_i is the union of discs and spheres in ∂P_i corresponding to the cutting locus: if P_i is viewed as an immersed submanifold of M , then $\text{int}(S_i)$ is mapped to the interior of M . For the system to be complete we also require that S_i contain all but at most one boundary sphere of P_i . Thus at most one boundary component contains discs of S_i . There is a cell complex of weighted primitive systems of essential spheres and discs in M constructed as follows. For any primitive essential system \mathcal{S} in M , the space contains a simplex corresponding to projective classes of systems of non-negative weights on the spheres. Two simplices corresponding to primitive essential systems \mathcal{S}_1 and \mathcal{S}_2 are identified on the face corresponding to the maximal primitive essential system \mathcal{S}_3 of spheres common to \mathcal{S}_1 and \mathcal{S}_2 . We call the resulting space the *reducing space* of M . In the reducing space, the union of open cells corresponding to complete primitive systems is called the *complete reducing space* of M .

Lemma 2.5. *The complete reducing space of a mixed body M is connected. In other words, it is possible to pass from any primitive complete essential system to any other by moves which replace a primitive complete essential system by a primitive complete subsystem or a primitive complete essential supersystem.*

Proof. The proof is, of course, similar to that of Lemma 2.2. Suppose $\mathcal{S} = \mathcal{S}_0$ and \mathcal{S}' are two primitive complete systems. We will use the term “reducing surface” or just “reducer” to refer to a properly embedded surface which is either a sphere or a disc.

Consider $\mathcal{S} \cap \mathcal{S}'$ and choose a curve of intersection innermost in \mathcal{S}' . Remove the curve of intersection by isotopy of \mathcal{S} if possible. Otherwise, surger \mathcal{S} on a disc (half-disc) bounded in \mathcal{S}' by the curve to obtain \mathcal{S}_2 . Now \mathcal{S}_2 is obtained from $\mathcal{S} = \mathcal{S}_0$ by removing one reducer, the surgered reducer S_0 of \mathcal{S}_0 , say, and adding two reducers, S_{21} and S_{22} , which result from the surgery. We can do this in two steps, first adding the two new reducers to obtain \mathcal{S}_1 , then removing the surgered reducer to obtain \mathcal{S}_2 . If one of the reducers S_{21} and S_{22} is inessential, this shows that the curve of intersection between $\mathcal{S} = \mathcal{S}_0$ and \mathcal{S}' could have been removed by isotopy. Note that surgery on arcs is always between two different discs of the system, but surgery on closed curves may involve any pair of reducers (disc and sphere, two discs, or two spheres). We note that S_{21} or S_{22} , or both, might be isotopic to reducers of \mathcal{S}_0 . If the new reducers are essential, we claim that \mathcal{S}_1 and \mathcal{S}_2 are complete systems. It is clear that \mathcal{S}_1 cuts M into holed balls, since it is obtained by adding two reducers S_{21} or S_{22} to \mathcal{S}_0 . The property that cutting on reducers yields pairs (P_i, S_i) with at most one boundary component of P_i not in S_i is preserved. This is clearly true when adding a sphere. When adding a disc, a sphere of ∂P_i not contained in S_i is split, and so is P_i , yielding two holed balls with exactly one boundary sphere not contained in S_i . We then remove S_0 from \mathcal{S}_1 to obtain \mathcal{S}_2 . There is, of course a danger that removing an essential reducer from an essential system \mathcal{S}_1 yields a system which does not cut M into holed balls, or does not preserve the other completeness property. As before, if the surgery is on a disc, then the effect on the complementary (P_i, S_i) 's is first to cut a P_i on a disc with boundary in S_i . Clearly this has the effect of cutting P_i into two holed balls. If the surgery is on a disc of S_i , we obtain one new holed ball (P_j, S_j) with all of its boundary in S_j , and one holed ball with one boundary component not contained in S_j . Then we attach a 2-handle to an S_i of another (P_i, S_i) , which, as before, yields the required types of holed balls. If the surgery is on a half-disc, we first cut a holed ball (P_i, S_i) on a half disc, i.e. we cut on a disc intersecting the distinguished component of ∂P_i which is not in S_i . This yields the required types of holed balls. Then we attach a half 2-handle whose boundary intersects S_i in a single arc to a (P_i, S_i) , which has the effect of splitting a disc of S_i into two discs. Thus, \mathcal{S}_2 cuts M into holed balls of the desired type. If \mathcal{S}_2 is not primitive, i.e. contains two isotopic reducers, we can remove one, until we are left with a primitive complete essential system.

Inductively, removing curves of intersection of \mathcal{S}_k with \mathcal{S}' either by isotopy or by surgery on an innermost disc (half-disc) in \mathcal{S}' as above, we must obtain a sequence

ending with \mathcal{S}_{n-2} disjoint from \mathcal{S}' . Next, we add the reducers of \mathcal{S}' not in \mathcal{S}_{n-2} to \mathcal{S}_{n-2} to obtain \mathcal{S}_{n-1} , which is still complete. Finally we remove all reducers from \mathcal{S}_{n-1} except those of \mathcal{S}' to get $\mathcal{S}_n = \mathcal{S}'$. \square

Let P be a 3-sphere with $r \geq 1$ holes and let $S \subseteq \partial P$ be a union of discs and spheres such that each of $r - 1 \geq 1$ components of ∂P is a sphere of S , and the remaining sphere of ∂P either is also in S or contains a non-empty set of discs of S . ((P, S) is the type of pair obtained by cutting on a complete system.) If S contains discs, a *standard* r -spotted ball (K, E) in (P, S) is one with the property that each component of S contains a component of E , and the complement $\overline{P - K}$ is a spotted product, $(V \times I, F)$, where F is a union of discs in $V \times 1$. This means that $\overline{P - K}$ has the form $V \times I$, where V is a planar surface, $V \times 0$ is mapped to $\partial P - \text{int}(S)$, $\partial V \times I$ is mapped to the annular components of $S - \text{int}(E)$, and exactly one disc component of F is mapped to each sphere of S . We have already defined what it means for (K, E) to be standard in (P, S) if $\partial P = S$. Note that the spots of F are in one-one correspondence with the sphere components of S , and each is mapped to the complement of E in its sphere component.

Lemma 2.6. *Let P be a 3-sphere with $r \geq 1$ holes and let $S \subseteq \partial P$ be a union of discs and spheres such that each of $r - 1 \geq 1$ components of ∂P is a sphere of S , and the remaining sphere of ∂P either is also in S or contains a non-empty set of discs of S*

(a) *A standard embedded r -spotted ball (K, E) in (P, S) is unique up to isotopy of pairs.*

(b) *Suppose (K, E) and (K', E') are two different standard spotted balls in (P, S) and suppose $\hat{E} \subseteq E$ is a union of spots in E , which coincides with a union of spots in E' . Then there is an isotopy $\text{rel } \hat{E}$ of (K, E) to (K', E') .*

Proof. (a) The (mapping class) group of isotopy classes of automorphisms of (P, S) which map each component of S to itself is trivial. It is then also easy to verify that the mapping class group of (P, S) $\text{rel } S$ is trivial.

We may assume that S contains discs, otherwise we have already proved the result. Suppose now that (K, E) and (K', E') are two different standard spotted balls in (P, S) . We will construct an automorphism f of (P, S) by defining it first on K , then extending to the complement. Isotope E' to E extending the isotopy to K' . We choose a homeomorphism $f : (K, E) \rightarrow (K', E')$ sending a disc of E to the corresponding disc of E' . Next, we extend to $S - E$ such that $f(S - E) = S - E'$. Since (K, E) is standard, its complement has a natural structure as a spotted product, and the complement of K' has an isomorphic structure. Thus we can extend f so that the complementary spotted product K^c is mapped homeomorphically to $(K')^c$. Then f is isotopic to the identity, whence (K, E) is isotopic to (K', E') .

(b) To prove (b) we adjust the isotopy of (a) in a collar of \hat{E} in K . \square

Proof. (Theorem 2.4) Any complete system \mathcal{S} for M gives a Heegaard splitting as follows. In each holed 3-sphere (P_i, S_i) obtained by cutting M on \mathcal{S} we insert an s_i -spotted ball (K_i, E_i) , where s_i is the number of components of S_i . Each component of S_i contains exactly one disc E_i . We assemble the spotted balls (K_i, E_i) by isotoping and identifying the two spots on opposite sides of each essential sphere or disc of the system \mathcal{S} , extending the isotopy to K_i . The result is a handlebody $H = \cup_i K_i$ with a complementary handlebody $H^c = \cup_i K_i^c$.

We claim that H (depending on \mathcal{S}) is unique up to isotopy in M . First replace each sphere or disc S of \mathcal{S} by $\partial N(S)$, i.e. by two parallel copies of itself. The result is that no (P_i, S_i) obtained by cutting M on \mathcal{S} has two different disc or sphere components of S_i identified in M . Suppose H' is constructed in the same way as H above. Inductively, we isotope $H' \cap P_i = K'_i$ to K_i using Lemma 2.6, extending the isotopy to H' . Having isotoped K'_i to K_i for $i < k$, we isotope K'_k to K_k in P_k rel $\cup_{i < k} K'_i \cap \partial P_k$. This means we perform the isotopy rel spots on the boundary of P_k where previous K'_i 's have already been isotoped to coincide with K_i 's. Now it is readily verified that the H associated to the non-primitive \mathcal{S} obtained by replacing each sphere or disc with two copies of itself is the same as the H associated to the original \mathcal{S} .

Next we must show that H does not depend on the choice of a complete system \mathcal{S} . By Lemma 2.5 it is possible to find a sequence $\{\mathcal{S}_i\}$ of complete systems such that any two successive systems differ by the addition or removal of reducers. It is then easy to verify that after each move, the handlebody H_i , constructed from \mathcal{S}_i as H was constructed from \mathcal{S} above, will be the same up to isotopy. For example, when removing an essential reducer from the system \mathcal{S}_i to obtain \mathcal{S}_{i+1} , two spotted balls are identified on a disc on the removed reducing surface, yielding a new spotted ball in a new holed sphere, and $H_{i+1} = H_i$. The reverse process of adding an essential reducer must then also leave $H_{i+1} = H_i$ unchanged. \square

The canonical splitting surface for M a sphere body or a mixed body yields an automatic reducing surface for any automorphism of $f : M \rightarrow M$. However, this reducing surface is non-rigid. There are automorphisms of M isotopic to the identity which induce automorphisms not isotopic to the identity on the splitting surface. The non-rigidity (without our terminology) is well-known, but rather than giving the example described in [6], we describe a simpler example, which in some sense is the source of the phenomenon.

Example 2.7. Let M be a sphere body and let S be a non-separating sphere in M intersecting the canonical splitting surface in one closed curve. Let H be one of the handlebodies bounded by the splitting surface, so the splitting surface is ∂H . We describe an isotopy of H in M which moves H back to coincide with itself and induces an automorphism not isotopic to the identity on ∂H . In other words, we describe $f : M \rightarrow M$ isotopic to the identity, such that $f|_{\partial H}$ is not isotopic to the identity. Figure 1 shows the intersection of H with S . We have introduced a kink in

a handle of H such that the handle runs parallel to S for some distance. The ends of the handle project to (neighborhoods of) distinct points while the handle projects to a neighborhood of an arc α in S . There is an isotopy of α in S fixing endpoints, so the arc rotates about each point by an angle of 2π and the arc sweeps over the entire 2-sphere, returning to itself. We perform the corresponding isotopy of the 1-handle in H . Calling the result of this isotopy f , we note that $f|_{\partial H}$ is a double Dehn twist as shown. As one can easily imagine, by composing automorphisms of this type on different spheres intersecting H in a single meridian, one can obtain f such that f restricted to ∂H is a more interesting automorphism, e.g. a pseudo-Anosov, see [6].

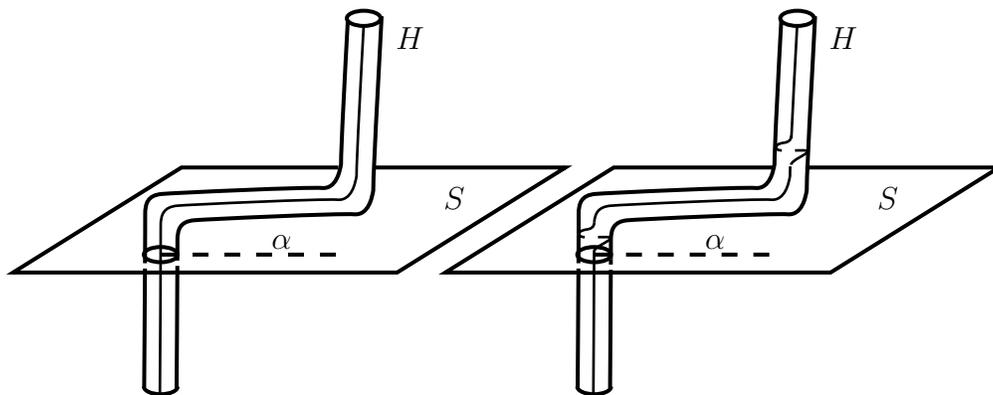


Figure 1: Non-rigidity of canonical splitting.

We shall refer to an isotopy of the type described above as a *double-Dehn isotopy*. If, as above, M is a sphere body or mixed body, and $f : M \rightarrow M$ is an automorphism, we shall use double-Dehn isotopies to choose a good representative of the isotopy class of $f : M \rightarrow M$ preserving the canonical splitting.

3 Reducing surfaces

The purpose of this section is to define reducing surfaces in the general setting of automorphisms of compact 3-manifolds. In addition, we review reducing surfaces which arise in the setting of automorphisms of handlebodies and compression bodies, [16], using the somewhat more inclusive definitions given here.

We will also consider manifolds with some additional structure; we will consider holed or spotted manifolds.

Suppose M is a compact 3-manifold without sphere boundary components, and $f : M \rightarrow M$ is an automorphism. If \mathcal{S} is an f -invariant (up to isotopy) union of essential spheres, then \mathcal{S} is a special kind of reducing surface. We can decompose M on \mathcal{S} to obtain $M|\mathcal{S}$ and cap boundary spheres with balls to obtain a *manifold* (\hat{M}, \mathcal{D}) with spots in the interior, where \mathcal{D} denotes the union of the capping balls,

and an induced automorphism $(\hat{M}, \mathcal{D}) \rightarrow (\hat{M}, \mathcal{D})$. We could equally well not cap boundary spheres after decomposing on \mathcal{S} and consider holed manifolds rather than spotted manifolds.

Definition 3.1. Let (M, \mathcal{D}) be a spotted 3-manifold with spots in the interior (\mathcal{D} a union of 3-balls disjointly embedded in the interior of M). We say a sphere is *essential* in (M, \mathcal{D}) if it is embedded in $M - \mathcal{D}$, it does not bound a ball in $M - \mathcal{D}$, and it is not isotopic in $M - \mathcal{D}$ to a component of $\partial\mathcal{D}$. The spotted manifold is *irreducible* if it has no essential spheres.

Clearly a connected spotted manifold (M, \mathcal{D}) is irreducible if and only if the underlying manifold M is irreducible and \mathcal{D} is empty or consists of a single ball.

Definition 3.2. Let (M, \mathcal{D}) be a spotted manifold with spots in the interior. Let $f : (M, \mathcal{D}) \rightarrow (M, \mathcal{D})$ be an automorphism of the spotted manifold. Then f is *reducible on spheres* if there is an f -invariant union \mathcal{S} of disjoint essential spheres in (M, \mathcal{D}) . \mathcal{S} is called a *sphere reducing surface*. *Decomposing* on \mathcal{S} and capping with balls gives a new spotted manifold with an induced automorphism. The caps of duplicates of spheres of the reducing surface are additional spots on the decomposed manifold.

Clearly by repeatedly decomposing on sphere reducing surfaces we obtain automorphisms of manifolds with interior spots which do not have sphere reducing surfaces.

Automorphisms of a spotted 3-manifold (M, \mathcal{D}) (with 3-dimensional spots) induce automorphisms of the underlying spotless 3-manifold M , and in the cases of interest to us it is relatively easy to understand the relationship between the two. For this reason, we can henceforth consider only 3-manifolds without 3-dimensional spots and without sphere reducing surfaces.

Definition 3.3. Suppose $f : M \rightarrow M$ is an automorphism, without sphere reducing surfaces, of a 3-manifold. Then a closed surface $F \hookrightarrow M$ which is f -invariant up to isotopy is called a *closed reducing surface* provided $M|F$ is irreducible. The reducing surface F is *rigid* if f induces uniquely an automorphism of the decomposed manifold $M|F$.

In particular, the canonical splitting surface for M a sphere body or mixed body is a (non-rigid) reducing surface for any automorphism of M .

The existence of any closed reducing surface for $f : M \rightarrow M$ (without sphere components) gives a decomposition into automorphisms of compression bodies and manifolds with incompressible boundary, as the following theorems show. We give a simple version for closed manifolds separately to illustrate the ideas.

Theorem 3.4. *Let M be a closed reducible manifold, $f : M \rightarrow M$ an automorphism, and suppose F is a closed reducing surface, with no sphere components. Then either*

- i) F is incompressible,*
- ii) there is a reducing surface F' such that the reducing surface $F \cup F'$ cuts M into a non-product compression body Q (possibly disconnected) and a manifold M' with incompressible boundary, or*
- iii) F is a Heegaard surface.*

Proof. We suppose M is a closed manifold, $f : M \rightarrow M$ is an automorphism and suppose F is a closed reducing surface. Cutting M on F we obtain $M|F$ which is a manifold whose boundary is not in general incompressible. Let Q be the characteristic compression body in $M|F$. This exists because $M|F$ is irreducible by the definition of a reducing surface. Then $f|F$ preserves Q and $F' = \partial_i Q$ is another reducing surface, unless it is empty. In case $\partial_i Q$ is empty, we conclude that Q must consist of handlebodies, and we have a Heegaard decomposition of M . Otherwise, letting $M' = \overline{M - Q}$, M' has incompressible boundary. \square

Here is a more general version:

Theorem 3.5. *Let M be a reducible manifold, $f : M \rightarrow M$ an automorphism, and suppose F is a closed reducing surface, with no sphere components. Then either*

- i) F is incompressible,*
- ii) there is a reducing surface F' such that the reducing surface $F \cup F'$ cuts M into a non-product compression body Q (possibly disconnected) and a manifold M' with incompressible boundary, or*
- iii) F is a Heegaard surface dividing M into two compression bodies.*

A priori, there is no particular reason why every automorphism $f : M \rightarrow M$ without sphere reducing surfaces should have a closed reducing surface. Recall we posed Question 1.11. In any case, if M is reducible, we deal only with automorphisms for which this is true. Therefore we henceforth consider only automorphisms of irreducible 3-manifolds of the types resulting from the previous two theorems: Namely, we must consider automorphisms of irreducible, ∂ -irreducible manifolds, and we must consider automorphisms of compression bodies (and handlebodies).

It might be interesting to learn more about all possible reducing surfaces associated to a given automorphism of a given manifold, especially for answering Question 1.11. In our classification, however, we often use reducing surfaces guaranteed by the topology of the 3-manifold: reducing surfaces coming from F. Bonahon's characteristic compression body and the Jaco-Shalen-Johannson characteristic manifold, and reducing surfaces coming from the canonical Heegaard splittings of sphere bodies or mixed bodies. To analyze automorphisms of compression bodies, one uses reducing surfaces not guaranteed by the topology, see [16].

The definition of reducing surfaces for irreducible, ∂ -irreducible manifolds should be designed such that the frontier of the characteristic manifold is a reducing surface. The following is a natural definition.

Definition 3.6. Suppose $f : M \rightarrow M$ is an automorphism of an irreducible, ∂ -irreducible compact manifold. Then an invariant surface F is a *reducing surface* if it is incompressible and ∂ -incompressible and not boundary-parallel.

It only remains to describe reducing surfaces for automorphisms of handlebodies and compression bodies, and this has already been done in [16]. We give a review here, taken from that paper.

Definitions 3.7. We have already defined a *spotless compression body* (Q, V) : The space Q is obtained from a disjoint union of balls and a product $V \times I$ by attaching 1-handles to the boundaries of the balls and to $V \times 1$; the surface $V \hookrightarrow \partial Q$ is the same as $V \times 0$. We require that V contain no disc or sphere components. When Q is connected and $V = \emptyset$, (Q, V) is a handlebody or ball. If (Q, V) has the form $Q = V \times I$ with $V \times 0 = V$, then (Q, V) is called a *spotless product compression body* or a *trivial compression body*. As usual, $\partial_e Q = \partial Q - \text{int}(V)$ is the *exterior boundary*, even if $V = \emptyset$.

A *spotted compression body* is a triple (Q, V, \mathcal{D}) where (Q, V) is a spotless compression body and $\mathcal{D} \neq \emptyset$ denotes a union of discs or “spots” embedded in $\partial_e Q = \partial Q - \text{int}(V)$.

A *spotted product* is a spotted compression body (Q, V, \mathcal{D}) of the form $Q = V \times I$ with $V = V \times 0$ and with $\mathcal{D} \neq \emptyset$ a union of discs embedded in $V \times 1$.

A *spotted ball* is a spotted compression body whose underlying space is a ball. It has the form (B, \mathcal{D}) where B is a ball and $\mathcal{D} \neq \emptyset$ is a disjoint union of discs in $\partial B = \partial_e B$.

An *I-bundle pair* is a pair (H, V) where H is the total space of an I -bundle $p : H \rightarrow S$ over a surface S and V is $\partial_i H$, the total space of the associated ∂I -bundle.

Remark 3.8. A spotless compression body or I -bundle pair is an example of a *Haken pair*, i.e. a pair (M, F) where M is an irreducible 3-manifold and $F \subset \partial M$ is incompressible in M .

Definition 3.9. If (Q, V) is a compression body, and $f : (Q, V) \rightarrow (Q, V)$ is an automorphism, then an f -invariant surface $(F, \partial F) \hookrightarrow (Q, \partial Q)$, having no sphere components, is a *reducing surface for f* , if:

- i) F is a union of essential discs, $\partial F \subset \partial_e Q$, or no component of F is a disc and:
- ii) F is the interior boundary of an invariant compression body $(X, F) \hookrightarrow (Q, V)$, with $\partial_e X \hookrightarrow \partial_e Q$ but F is not isotopic to V , or
- iii) $F \subset \partial Q$, F is incompressible, F strictly contains V , and (Q, F) has the structure of an I -bundle pair, or
- iv) F is a union of annuli, each having one boundary component in $\partial_e Q$ and one boundary component in $\partial_i Q$

In case i) we say F is a *disc reducing surface*; in case ii) we say F is a *compressional reducing surface*; in case iii) we say F is a *peripheral reducing surface*; in case iv) we say F is an *annular reducing surface*.

In case i) we *decompose* on F by replacing (Q, V) by the *spotted compression body* $(M|F, V, \mathcal{D})$, and replacing f by the induced automorphism on this spotted compression body. (\mathcal{D} is the union of duplicates of the discs in F .) An automorphism g of a spotted compression body (Q, V, \mathcal{D}) can be *decomposed* on a spot-removing reducing surface, which is $\partial_e Q$ isotoped to become a properly embedded surface. This yields an automorphism of a *spotted product* and an automorphism of a spotless compression body.

In case ii) we *decompose* on F by cutting on F to obtain $Q|F$, which has the structure $(X, F) \cup (Q', V')$ of a compression body, and by replacing f by the induced automorphism. The duplicate of F not attached to X belongs to $\partial_e Q'$. We throw away any trivial product compression bodies in (Q', V') .

In case iii) we *decompose on F* by replacing (Q, V) by (Q, F) .

In case iv), cutting on F yields another compression body with induced automorphism.

For completeness, if (H, V) is a spotless I -bundle pair, we define reducing surfaces for automorphisms $f : (H, V) \rightarrow (H, V)$. Any f -invariant non- ∂ -parallel union of annuli with boundary in V is such an *annular reducing surface*, and these are the only reducing surfaces.

Non-rigid reducing surfaces are quite common, as the following example shows.

Example 3.10. Let M be a mapping torus of a pseudo-Anosov automorphism ϕ , $M = S \times I / \sim$ where $(x, 0) \sim (\phi(x), 1)$ describes the identification. Let τ be an invariant train track for the unstable lamination of ϕ in $S = (S \times 0) / \sim$. Let $f : M \rightarrow M$ be the automorphism induced on M by $\phi \times \text{Identity}$. Then $N(\tau)$ is invariant under f up to isotopy and $F = \partial N(\tau)$ is a reducing surface for F . Isotoping τ and $F = \partial N(\tau)$ through levels $S_t = S \times t$ from $t = 0$ to $t = 1$ and returning it to its original position induces a non-trivial automorphism on $N(\tau)$. Hence F is non-rigid.

It is easy to modify this M by Dehn surgery operations such that $F = \partial N(\tau)$ becomes incompressible.

The results of Section 2 guarantee reducing surfaces in the special case where M is a mixed body without sphere boundary components, i.e. a connected sum of $S^2 \times S^1$'s and handlebodies, and this is what we use in the classification. The canonical splitting surfaces of Section 2 are examples of reducing surfaces such that applying Theorem 3.4 and Theorem 3.5 yields a Heegaard splitting.

We state the obvious:

Proposition 3.11. *If M is a mixed body, then there is a canonical Heegaard decomposition $M = H \cup Q$ on a reducing surface $F = \partial H$, which is unique up to isotopy. F is a non-rigid closed reducing surface for any automorphism $f : M \rightarrow M$.*

4 Adjusting automorphisms

In this section we will prove Theorem 1.7 and Corollary 1.8.

The adjusting automorphisms are constructed from *slide automorphisms* of M . There are two types, one of which is as defined in [15], see the reference for details. For a fixed i , construct \check{M} by removing $\text{int}(\hat{M}_i)$ from M and capping the resulting boundary sphere with a ball B . Let $\alpha \subseteq \check{M}$ be an embedded arc meeting B only at its endpoints. One can “slide” B along α and back to its original location. By then replacing B with \hat{M}_i we obtain an automorphism of M which is the identity except in a holed solid torus T (a closed neighborhood of $\alpha \cup S_i$ in $M - \text{int}(\hat{M}_i)$). Call such an automorphism a *simple slide*. It is elementary to verify that $\hat{M}_0 \cup T$ is a mixed body.

We also need another type of slide which is not described explicitly in [15]. First, consider the classes of \hat{M}_i 's which are homeomorphic. For each class, fix an abstract manifold N homeomorphic to the elements of the class. For each \hat{M}_i in the class fix a homeomorphism $f_i: \hat{M}_i \rightarrow N$.

If $i \neq j$, and \hat{M}_i is homeomorphic to \hat{M}_j , construct \check{M} by removing $\text{int}(\hat{M}_i \cup \hat{M}_j)$ and capping the boundary spheres S_i, S_j with balls B_i, B_j . Let $\alpha, \alpha' \subseteq \check{M}$ be two disjoint embedded arcs in \check{M} linking S_i and S_j . One can slide B_i along α and B_j along α' simultaneously, interchanging them. Now remove the balls and re-introduce \hat{M}_i, \hat{M}_j , now interchanged. Denote by s the resulting automorphism of M . Isotope s further to ensure that it restricts to \hat{M}_i as $f_j^{-1} \circ f_i$ (and to \hat{M}_j as $f_i^{-1} \circ f_j$). Call such an automorphism an *interchanging slide*. The support of an interchanging slide is a closed neighborhood of $\hat{M}_i \cup \hat{M}_j \cup \alpha \cup \alpha'$. Note that a sequence of interchanging slides that fixes an \hat{M}_i restricts to it as the identity.

Remark 4.1. An argument in [15] makes use of automorphisms called *interchanges of irreducible summands*. Our interchanging slides are a particular type of such automorphisms.

Proof. (Theorem 1.7.) We start by noting that \hat{M}_0 is an essential holed sphere body.

Let $\mathcal{S} = \bigcup_i S_i$ and consider $f(\mathcal{S})$. The strategy is to use isotopies and both types of slides to take $f(S_i)$ to S_i . If only isotopies are necessary we make $R = \emptyset$, $h = \text{Id}_M$ and $f = g$ in the statement. Therefore suppose that some $f(S_i)$ is not isotopic to S_i . The first step is to simplify $f(\mathcal{S}) \cap \mathcal{S}$. The argument of [15] yields a sequence of simple slides (and isotopies) whose composition is σ satisfying $\sigma \circ f(\mathcal{S}) \cap \mathcal{S} = \emptyset$. We skip the details. We can then assume that $\sigma \circ f(\mathcal{S}) \subseteq \hat{M}_0$, otherwise some S_i would bound a ball. Note that σ preserves $K \subseteq M$, where K consists of \hat{M}_0 to which is attached a collection of disjoint 1-handles $J \subseteq \bigcup \hat{M}_i$ along \mathcal{S} . Each 1-handle consists of a neighborhood (in \hat{M}_i) of a component of $\alpha \cap \hat{M}_i$, where α is a path along which a simple slide is performed. Perturb these paths so that they are pairwise disjoint.

Now consider $\sigma \circ f(S_i) \subseteq \hat{M}_0$. It must bound an (once holed) irreducible summand on one side. It is then easy to check that $\sigma \circ f(S_i)$ is isotopic to an S_j . Indeed, $\sigma \circ f(S_i)$

separates \hat{M}_0 , bounding (in \hat{M}_0) on both sides a holed sphere body. On the other hand $\sigma \circ f(S_i)$ must bound (in M) an irreducible summand. Therefore in \hat{M}_0 it must bound an once-holed ball, whose hole corresponds to some S_j . We can then make $\sigma \circ f(S) = \mathcal{S}$.

If $f(S_i) = S_j$, $i \neq j$ we use a slide to interchange \hat{M}_i and \hat{M}_j . A composition ι of these interchanging slides yields $\iota \circ \sigma \circ f(S_i) = S_i$ for each i .

Now consider the collection of 1-handles $J = (\cup \hat{M}_i) \cap K$, where we recall that K is the union of the support of σ with \hat{M}_0 (K is preserved by σ). We note that there is some n such that, for any $i \geq 1$, $\iota^n|_{\hat{M}_i} = \text{Id}$. In particular, $\iota^n(J) = J$. By perturbing the paths along which the simple slides of σ are performed we can assume that $\iota^k(J) \cap J = \emptyset$ for any k , $1 \leq k < n$. Therefore $\cup_k \iota^k(J)$ consists of finitely many 1-handles. Let $R = K \cup (\cup_k \iota^k(J))$. Clearly R is preserved by ι . The restriction of ι to $\cup_i \hat{M}_i$ is periodic, therefore its restriction to $\overline{M - R} \subseteq (\cup_i \hat{M}_i)$ also is. Finally, R is a mixed body. Either ι or σ is not the identity by hypothesis, therefore $R \neq B$, the 3-ball. It is not hard to see that it is essential. Indeed, if $S \subseteq R$ is an essential sphere then an argument similar to the one above yields a sequence of slides τ which preserves R and takes S to $\tau(S) \subseteq \hat{M}_0$. Essentiality of S implies essentiality of $\tau(S)$ in R . Therefore $\tau(S)$ is essential in \hat{M}_0 (which is itself essential in M), and hence in M . But essentiality of $\tau(S)$ in M implies essentiality of S in M . It remains to consider S a sphere boundary component of R , but this is parallel to an $S_i \subseteq \mathcal{S}$, which is essential since M has no sphere boundary components.

Let $\hbar = \iota \circ \sigma$, $h = \hbar^{-1}$ and $g = \hbar \circ f$. The following are immediate from the construction:

- 1) R is preserved by \hbar (and hence by h).
- 2) $\hbar|_{\overline{M-R}} = \iota|_{\overline{M-R}}$, which is periodic.
- 3) g preserves each \hat{M}_i .

The other conclusions of the theorem are clear. □

Proof. (Corollary 1.8.) We follow the proof of Theorem 1.7. In that proof, a composition of simple slides and interchanging slides has the effect of moving $f(S_i)$ to S_i . A 4-dimensional handlebody P_0 is obtained from a 4-ball K_0 by attaching ℓ 1-handles. The boundary is a connected sum of ℓ $S^2 \times S^1$'s. The 4-dimensional compression body Q is obtained from a disjoint union of the products $M_i \times I$ and from P_0 by attaching a 1-handle to P_0 at one end and to $M_i \times 1$ at the other end, one for each $i \geq 1$. Then $\partial_e Q$ is the connected sum of the M_i 's and the sphere body ∂P_0 . The cocore of each 1-handle has boundary equal to an S_i , $i = 1, \dots, k$. Each S_i bounds a 3-ball E_i in Q . A simple slide in $\partial_e Q = M$ using some S_i can be realized as the (restriction to the) exterior boundary of an automorphism of Q as follows: Cut a 1-handle of Q on the ball E_i . This gives a lower genus compression body \hat{Q} with two 3-ball spots on its exterior boundary. We slide one of the spots along the curve α associated to the slide, extending the isotopy to \hat{Q} and returning the spot to its original position, then we reglue the pair of spots. This defines an automorphism of the compression

body fixing the interior boundary and inducing the slide on the exterior boundary. An interchanging slide can be realized similarly.

Suppose now that f is given as in the proof of Theorem 1.7, and we perform the slides in the theorem so that $\iota \circ \sigma \circ f(S_i) = S_i$ for each i . Capping each S_i in $\partial\hat{M}_i$ with a disc E_i to obtain M_i , and capping the other copy of S_i in the holed sphere body \hat{M}_0 with a disc E'_i , we obtain a disjoint union of the M_i 's and the sphere body $\partial P_0 = M_0$. The automorphism f induces an automorphism of this disjoint union, regarded as spotted manifolds. Thus we have an induced automorphism of a spotted manifold $f_s : ((\sqcup_i M_i \sqcup \partial P_0), \cup_i (E_i \cup E'_i)) \rightarrow ((\sqcup_i M_i \sqcup \partial P_0), \cup_i (E_i \cup E'_i))$. (f_s for f spotted.) If we ignore the spots, we have an automorphism f_a of $(\sqcup_i M_i \sqcup \partial P_0)$. (f_a for f adjusted.) These automorphisms yield an automorphism f_p of the disjoint union of the spotted products $M_i \times I$ with spots E_i in $M_i \times 1$ where f_p restricts to $(M_i \times 1, E_i \times 1)$ as f_s and to $M_i \times 0$ as f_a . (f_p for f product.) We can also extend f_s to f_p on all of $(P_0, \cup_i E'_i)$ so that the restriction of f_p to $(\partial P_0, \cup_i E'_i)$ is f_s . Now f_p is an automorphism of a disjoint union of products $M_i \times I$ with spots on one end of the product, disjoint union P_0 with spots on ∂P_0 . Gluing E_i to E'_i , we obtain Q and f_p yields an automorphism f_c of the compression body.

Now it only remains to apply the inverses of the automorphisms of Q realizing $\iota \circ \sigma$ on $\partial_e Q$ as a composition of slides of both kinds. This composition is \bar{h}^{-1} , and we have designed \bar{h} such that $\bar{h}^{-1} \circ f_c$ restricts to f on $\partial_e Q = M$. Then take $\bar{f} = \bar{h}^{-1} \circ f_c$.

Note that the support of the slides is actually in R of the previous proof. \square

5 One-dimensional invariant laminations

If $f : H \rightarrow H$ is a generic automorphism of a handlebody, there is an invariant 1-dimensional lamination for f described in [16], associated to an incompressible 2-dimensional invariant lamination of a type which we shall describe below, see Proposition 5.8. The invariant ‘‘lamination’’ is highly pathological, not even being embedded. One of our goals in this section is to describe various types of 1-dimensional laminations in 3-manifolds, ranging from ‘‘tame’’ to ‘‘wild embedded’’ and to the more pathological ‘‘wild non-embedded’’ laminations which arise as invariant laminations for automorphisms of handlebodies. All invariant 1-dimensional laminations we construct are dual to (and associated with) an invariant 2-dimensional lamination. In general, it seems most useful to regard the 1-dimensional and the 2-dimensional invariant laminations of a generic automorphism as a dual pair of laminations.

In the case of an automorphism of a compression body, only a 2-dimensional invariant lamination was described in [16]. In this section we shall also describe invariant 1-dimension laminations for (generic) automorphisms of compression bodies.

Definitions 5.1. A 1-dimensional *tame lamination* $\Omega \hookrightarrow M$ in a 3-manifold M is a subspace Ω with the property that Ω is the union of sets of the form $T \times I$ in *flat charts* of the form $D^2 \times I$ in M for a finite number of these charts which cover Ω . (T

is the *transversal* in the flat chart.) A tame lamination is *smooth* if the charts can be chosen to be smooth relative to a smooth structure on M .

A 1-dimensional *abstract lamination* is a topological space which can be covered by finitely many flat charts of the form $T \times I$, where T is a topological space, usually, in our setting, a subspace of the 2-dimensional disc.

A 1-dimensional *wild embedded lamination* in a 3-manifold M is an embedding of an abstract lamination. Usually wild will also mean “not tame.”

A 1-dimensional *wild non-embedded lamination* in a 3-manifold is a mapping of an abstract lamination into M .

We observe that in general the 1-dimensional invariant laminations constructed in [16] for $f : H \rightarrow H$, H a handlebody, are wild non-embedded, though they fail to be embedded only at finitely many *singular points* in H .

We will now construct invariant laminations for generic automorphisms of arbitrary compression bodies.

We will deal with a connected, spotless compression body (H, V) . “Spotless” means in particular that V contains no disc components. The surface $\partial H - \text{int}(V)$ will always be denoted W . Also, throughout the section, we assume $\text{genus}(H, V) > 0$, where $\text{genus}(H, V)$ denotes the number of 1-handles which must be attached to $V \times I$ to build Q .

Let $(H_0, V_0) \subseteq \text{int}(H)$ be a “concentric compression body” with a product structure on its complement. If the compression body has the form $V \times [0, 1]$ with handles attached to $V \times 1$, $H_0 = V \times [0, 1/2]$ with handles attached to $V \times 1/2$ in such a way that $H - H_0 \cup \text{fr}(H_0)$ has the structure of a product $W \times [0, 1]$. Here $V = V \times 0$ and $W = W \times 1$.

We shall sketch the proof of the following proposition, which is proven in [16].

Proposition 5.2. *Suppose $f : (H, V) \rightarrow (H, V)$ is generic automorphism of an 3-dimensional compression body. Then there is a 2-dimensional measured lamination $\Lambda \subseteq \text{int}(H)$ with transverse measure μ , such that $f(\Lambda, \mu) = (\Lambda, \lambda\mu)$, up to isotopy, for some stretch factor $\lambda > 1$. The leaves of Λ are planes. Further, Λ “fills” H_0 , in the sense that each component of $H_0 - \Lambda$ is either contractible or deformation retracts to V .*

The statement of Proposition 5.2 will be slightly expanded later, to yield Proposition 5.7.

An automorphism $f : (H, V) \rightarrow (H, V)$ is called *outward expanding* with respect to (H_0, V) if $f|_{W \times I} = f|_{W \times 0} \times h$, where $h : I \rightarrow I$ is a homeomorphism moving every point towards the fixed point 1, so that $h(1) = 1$ and $h(t) > t$ for all $t < 1$. We define $H_t = f^t(H_0)$ for all integers $t \geq 0$ and reparametrize the interval $[0, 1)$ in the product $W \times [0, 1]$ such that $\partial H_t = W \times t = W_t$, and the parameter t now takes values in $[0, \infty)$. Finally, for any $t \geq 0$ we define (H_t, V) to be the compression body cut from (H, V) by $W \times t = W_t$.

Since any automorphism $f : (H, V) \rightarrow (H, V)$ of a compression body, after a suitable isotopy, agrees within a collar $W \times I$ of $W \subseteq \partial H$ with a product homeomorphism, a further “vertical” isotopy of f within this collar gives:

Lemma 5.3. *Every automorphism of a compression body is isotopic to an outward expanding automorphism.*

Henceforth, we shall always assume that automorphisms f of compression bodies have been isotoped such that they are outward expanding.

Let $\mathcal{E} = \{E_i, i = 1 \dots q\}$. be a collection of discs essential in (H_0, V) , with $\partial E_i \subseteq W_0$, cutting H_0 into a product of the form $V \times I$, possibly together with one or more balls. Such a collection of discs is called a *complete* collection of discs. (When $V = \emptyset$, \mathcal{E} cuts H into one or more balls.) We abuse notation by also using \mathcal{E} to denote the union of the discs in \mathcal{E} . We further abuse notation by often regarding \mathcal{E} as a collection of discs properly embedded in H rather than in H_0 , using the obvious identification of H_0 with H . Thus, for example, we shall speak of W -parallel discs in \mathcal{E} , meaning discs isotopic to discs in W_0 . Corresponding to the choice of a complete \mathcal{E} , there is a dual object Γ consisting of the surface V with a graph attached to $\text{int}(V)$ at finitely many points, which are vertices of the graph. The edges correspond to discs of \mathcal{E} ; the vertices, except those on V , correspond to the complementary balls; and the surface V corresponds to complementary product. If $V = \emptyset$, and H is a handlebody, then Γ is a graph.

Now H_0 can be regarded as a regular neighborhood of Γ , when Γ is embedded in H_1 naturally, with $V = V \times 0$. By isotopy of f we can arrange that $f(\mathcal{E})$ is transverse to the edges of Γ and meets $H_0 = N(\Gamma)$ in discs, each isotopic to a disc of \mathcal{E} . Any collection \mathcal{E} of discs properly embedded in H_0 (or H) with $\partial \mathcal{E} \subseteq W_0$ (or $\partial \mathcal{E} \subseteq W$), not necessarily complete and not necessarily containing only essential discs, is called *admissible* if every component of $f(\mathcal{E}) \cap H_0$ is a disc isotopic to a disc of \mathcal{E} . We have shown:

Lemma 5.4. *After a suitable isotopy every outward expanding automorphism $f : (H, V) \rightarrow (H, V)$ admits a complete admissible collection $\mathcal{E} \subseteq H_0$ as above, where every $E_i \in \mathcal{E}$ is a compressing disc of W .*

We shall refer to admissible collections \mathcal{E} of discs $(E_i, \partial E_i) \hookrightarrow (H, W)$ as *systems* of discs. Sometimes, we shall retain the adjective “admissible” for emphasis, speaking of “admissible systems.” A system may contain discs which are not compressing discs of W . Also, even if a system contains these W -parallel discs, the definition of completeness of the system remains the same. We will say a system is *W-parallel* if every disc in the system is W -parallel.

Let P_j denote the holed disc $f(E_j) - H_0$. Let m_{ij} denote the number of parallel copies of E_i in $f(E_j)$, and let $M = M(\mathcal{E})$ denote the matrix (m_{ij}) , which will be called the *incidence matrix* for \mathcal{E} with respect to f . A system \mathcal{E} is *irreducible* if the incidence matrix is irreducible. In terms of the discs E_i , the system \mathcal{E} is irreducible if

for each i, j there exists a $k \geq 1$ with $f^k(E_j) \cap H_0$ containing at least one disc isotopic to E_i . It is a standard fact that a matrix M with non-negative integer entries has an eigenvector x with non-negative entries and that the corresponding eigenvalue $\lambda = \lambda(\mathcal{E})$ satisfies $\lambda \geq 1$. If the matrix is irreducible, the eigenvector is unique and its entries are positive.

It turns out that the lack of reducing surfaces for f in (H, V) is related to the existence of irreducible complete systems. The following are proven in [16].

Lemma 5.5. *Suppose $f : (H, V) \rightarrow (H, V)$ is an automorphism of a compression body, and suppose that there is an (admissible) system which is not complete and not W -parallel. Then there is a reducing surface for f .*

Proposition 5.6. *If the automorphism $f : (H, V) \rightarrow (H, V)$ is generic, then there is a complete irreducible system \mathcal{E} for f . Also, any non- W -parallel complete system \mathcal{E} has a complete irreducible subsystem \mathcal{E}' with no W -parallel discs. Further, $\lambda(\mathcal{E}') \leq \lambda(\mathcal{E})$, and $\lambda(\mathcal{E}') < \lambda(\mathcal{E})$ if \mathcal{E} contains W -parallel discs.*

We always assume, henceforth, that f is generic, so there are no reducing surfaces. Assuming \mathcal{E} is *any* (complete) irreducible system of discs for f in the compression body (H, V) , we shall construct a branched surface $B = B(\mathcal{E})$ in $\text{int}(H)$. First we construct $B_1 = B \cap H_1$: it is obtained from $f(\mathcal{E})$ by identifying all isotopic discs of $f(\mathcal{E}) \cap H_0$. To complete the construction of B we note that $f(B_1 - \text{int}(H_0))$ can be attached to ∂B_1 to obtain B_2 and inductively, $f^i(B_1 - \text{int}(H_0))$ can be attached to B_i to construct a branched surface B_{i+1} . Alternatively, B_i is obtained by identifying all isotopic discs of $f^i(\mathcal{E}) \cap H_j$ successively for $j = i - 1, \dots, 0$. Up to isotopy, $B_i \cap H_r = B_r$, $r < i$, so we define $B = \cup_i B_i$. The branched surface B is a non-compact branched surface with infinitely many sectors. (Sectors are completions of components of the complement of the branch locus.) Note that the branched surface B does not have boundary on W_i ; this follows from the irreducibility of \mathcal{E} . If \mathcal{E} is merely admissible, the same construction works, but B may have boundary on W_i , $i \geq 0$.

If x is an eigenvector corresponding to the irreducible system \mathcal{E} , each component x_i of the eigenvector x can be regarded as a weight on the disc $E_i \in \mathcal{E}$. The eigenvector x now yields an infinite weight function w which assigns a positive weight to each sector:

$$w(E_i) = x_i \text{ and } w(f^t(P_i)) = x_i/\lambda^{t+1} \text{ for } t \geq 1.$$

Recall that P_i is a planar surface, $P_i = f(E_i) - \text{int}(H_0)$. As is well known, a weight vector defines a measured lamination (Λ, μ) carried by B if the entries of the weight vector satisfy the *switch conditions*. This means that the weight on the sector in $H_t - H_{t-1}$ adjacent to a branch circle in W_t equals the sum of weights on sectors in $H_{t+1} - H_t$ adjacent to the same branch circle, with appropriate multiplicity if a sector abuts the branch circle more than once. It is not difficult to check that our

weight function satisfies the switch conditions, using the fact that it is obtained from an eigenvector for $M(\mathcal{E})$.

At this point, we have constructed a measured lamination (Λ, μ) which is f -invariant up to isotopy and fully carried by the branched surface B . Applying f , by construction we have $f(\Lambda, \mu) = (\Lambda, \lambda\mu)$. We summarize the results of our construction in the following statement, which emphasizes the fact that the lamination depends on the choice of non- W -parallel system:

Proposition 5.7. *Suppose $f : (H, V) \rightarrow (H, V)$ is a generic automorphism of a 3-dimensional compression body. Given an irreducible system \mathcal{E} which is not W -parallel, there exists a lamination (Λ, μ) , carried by $B(\mathcal{E})$, which is uniquely determined, up to isotopy of Λ and up to scalar multiplication of μ .*

The lamination Λ “fills” H_0 , in the sense that each component of the complement is either contractible or deformation retracts to V . Also, $\Lambda \cup W$ is closed.

It will be important to choose the best possible systems \mathcal{E} to construct our laminations. With further work, it is possible to construct an invariant lamination with good properties as described in the following proposition. This is certainly not the last word on constructing “good” invariant laminations, see [3].

Proposition 5.8. *Suppose $f : (H, V) \rightarrow (H, V)$ is a generic automorphism of a compression body. Then there is a system \mathcal{E} such that there is an associated 2-dimensional measured lamination $\Lambda \subseteq \text{int}(H)$ as follows: It has a transverse measure μ such that, up to isotopy, $f((\Lambda, \mu)) = (\Lambda, \lambda\mu)$ for some $\lambda > 1$. Further, the lamination has the following properties:*

- 1) *Each leaf ℓ of Λ is an open 2-dimensional disc.*
- 2) *The lamination Λ fills H_0 , in the sense that each component of $H_0 - \Lambda$ is contractible or deformation retracts to V*
- 3) *For each leaf ℓ of Λ , $\ell - \text{int}(H_0)$ is incompressible in $H - (\text{int}(H_0) \cup \text{int}(V))$.*
- 4) *$\Lambda \cup (\partial H - \text{int}(V))$ is closed in H .*

Given a “good” system \mathcal{E} , e.g. as in Proposition 5.8, we now construct a dual 1-dimensional invariant lamination. This was done in [16] for automorphisms $f : H \rightarrow H$ where H is a handlebody. We now construct an invariant lamination in the case of an automorphism $f : (H, V) \rightarrow (H, V)$ where (H, V) is a compression body.

Theorem 5.9. *Suppose $f : (H, V) \rightarrow (H, V)$ is a generic automorphism of a compression body with invariant 2-dimensional measured lamination (Λ, μ) satisfying the incompressibility property. There is a 1-dimensional abstract measured non-compact 1-dimensional lamination (Ω, ν) , and a map $\omega : \Omega \rightarrow H_0 - V$ such that $f(\omega(\Omega, \nu)) = \omega(\Omega, \nu/\lambda)$. The map ω is an embedding on $\omega^{-1}(N(\Lambda))$ for some neighborhood $N(\Lambda)$. The statement that $f(\omega(\Omega, \nu)) = \omega(\Omega, \nu/\lambda)$ should be interpreted to mean that there is an isomorphism $h : (\Omega, \nu) \rightarrow (\Omega, \nu/\lambda)$ such that $f \circ \omega = \omega \circ h$.*

Proof. . As in [16] we define H_n, \mathcal{E}_n (for $n \in \mathbb{Z}$) to be $f^n(H_0)$ and $f^n(\mathcal{E})$ respectively. We choose a two-dimensional invariant lamination (Λ, μ) with the incompressibility property. It can be shown, as in [16], that Λ can be put in Morse position without centers with respect to a height function t whose levels $t \in \mathbb{Z}$ correspond to $\partial_e H_t$. It can also be shown as in [16] that for $n > t$ $f^n(\mathcal{E}) \cap H_t$ consists of discs belonging to a system \mathcal{E}_t in H_t . This implies that for all real t , (H_t, V) is a compression body homeomorphic to (H, V) , and we can take a quotient of (H_t, V) which gives Γ_t , a graph attached to V , the quotient map being $q_t : H_t \rightarrow \Gamma_t$. More formally, let H_t^Λ denote the completion of the complement of Λ in H_t . Points of the graph in Γ_t correspond to leaves of the 2-dimensional lamination Λ_t in (H_t, V) , or they correspond to balls in H_t^Λ . Points of $V \subseteq \Gamma_t$ correspond to interval fibers in the product structure for the product component of H_t^Λ . We also clearly have maps $\omega_{st} : \Gamma_s \rightarrow \Gamma_t$ such that if $i_{st} : H_s \rightarrow H_t$ is the inclusion, then $q_t \circ i_{st} = \omega_{st} \circ q_s$. Then the inverse limit of the maps ω_{st} , for integer values of $t = s + 1$ say, yields a 1-dimensional lamination limiting on V . The 1-dimensional lamination is the desired lamination. More details of the ideas in this proof can be found in the proof in [16] of the special case when $V = \emptyset$. \square

We next investigate the 1-dimensional invariant lamination Ω further. In [16] it is pointed out that (in the case of an automorphism of a handlebody) Ω has, in general, much self-linking. An issue which was not addressed is whether the leaves of Ω might be “knotted.” This sounds unlikely, and in fact knotting does not occur, but it is not even clear what it means for a leaf to be knotted. To clarify this notion, we first make some further definitions.

Definition 5.10. A *spotted ball* is a pair (K, \mathcal{D}) where K is a ball and \mathcal{D} is a collection of discs (or *spots*) in ∂K . A *spotted product* is a triple (K, V, \mathcal{D}) of the form $K = V \times I$ with $V = V \times 0$ and with \mathcal{D} a collection of discs in the interior of $V \times 1$. The spotted product can also be regarded as a pair (K, R) , where $R = \partial K - \text{int}(V) - \text{int}(\mathcal{D})$. If (K, \mathcal{D}) is a spotted ball we let the corresponding $V = \emptyset$. If (K, V, \mathcal{D}) is a spotted ball or connected product, (where $V = \emptyset$ if K is a ball) we say that it is a *spotted component*.

Spotted balls occur naturally in our setting as follows: Suppose H is a handlebody. Define \widehat{H}_1 to be H_1 cut open along \mathcal{E}_1 , or $H_1|_{\mathcal{E}_1}$. We let $\Gamma_t \cap \widehat{H}_1$ (or $H_t \cap \widehat{H}_1$) denote $\Gamma_t|_{\mathcal{E}_1}$ (or $H_t|_{\mathcal{E}_1}$). The disc system \mathcal{E}_1 determines a collection of spots $\widehat{\mathcal{E}}_1 \subseteq \partial \widehat{H}_1$ (2 spots for each disc of \mathcal{E}_1). If K is a component of \widehat{H}_1 and $\mathcal{D} = K \cap \widehat{\mathcal{E}}_1$ then it is clear that (K, \mathcal{D}) is a spotted ball. If we make similar definitions for a compression body (H, V) , then we also obtain spotted product components K in \widehat{H}_1 .

In general, for $t \leq 1$, the triple $(K_t, V_t, \mathcal{D}_t)$ will represent a spotted component where K_t is a component of $H_t \cap \widehat{H}_1$ and $\mathcal{D}_t = K_t \cap \widehat{\mathcal{E}}_1$. Usually a spotted ball (K_t, \mathcal{D}_t) is a ball-with-two-spots, corresponding to arc components L_t of $\Gamma_t \cap \widehat{H}_1$. The other spotted balls have more spots, K_t corresponding to star components L_t of $\Gamma_t \cap \widehat{H}_1$.

We denote $L_t \cap \partial K_t$ by ∂L_t . It is clear that the spots of \mathcal{D}_t correspond to points of ∂L_t or to the edges of L_t . We also note that $L_t \subseteq K_t$ is uniquely defined, up to isotopy (rel \mathcal{D}_t) by the property that it is a star with a vertex in each spot and that it is “unknotted”. We say L_t is *unknotted* in K_t if it is isotopic in K_t (rel ∂K_t) to ∂K_t .

When we are dealing with a compression body ($V_t \neq \emptyset$), we also have component(s) of L_t consisting of a component X of V with some edges attached at a single point of X . In fact, it is often useful to consider the union of star graphs \hat{L}_t , which is the closure of $L_t - V$ in L_t . These components of L_t correspond to connected spotted products K_t and the edges correspond to spots. The natural equivalence classes for the L_t 's are isotopy classes (rel \mathcal{D}_t), and for the \hat{L}_t 's isotopy classes (rel $\mathcal{D}_t \cup X$). Given K_t corresponding to a spotted product, there are many such equivalence classes, even if restricted to “unknotted” \hat{L}_t 's. Here our \hat{L}_t is unknotted if it can be isotoped rel \mathcal{D}_t to $V \times 1$.

We will often use “handle terminology.” H_t is constructed from the 0-handles K_t (or spotted products), with 1-handles corresponding to edges of Γ_t attached to the spots.

Definition 5.11. If Ω is the 1-dimensional dual lamination associated to a 2-dimensional invariant lamination Λ for a generic automorphism $f : (H, V) \rightarrow (H, V)$, which in turn is associated to a complete system \mathcal{E} , then a leaf ℓ of Ω is *unknotted* if every component of $\ell \cap \hat{H}_1$, contained in a component K , say of \hat{H}_1 , is ∂ -parallel in K , or isotopic to an arc embedded in ∂K , whether K is a spotted ball or a spotted product. In a spotted product (K, V, \mathcal{D}) , we require that the arc be parallel to $R = \partial K - \text{int}(V) - \text{int}(\mathcal{D})$.

Theorem 5.12. *Let $f : (H, V) \rightarrow (H, V)$ be a generic automorphism of a compression body. The 1-dimensional dual lamination Ω associated to a system \mathcal{E} is unknotted in (H, V) .*

To prove the theorem, we prove some lemmas from which the theorem follows almost immediately. We first make a few more definitions.

If a spotted ball K_t has k spots then we can embed a polygonal disc T_t with $2k$ sides in $R_t = \partial K_t - \text{int}(\mathcal{D}_t)$, see Figure 2. Alternate sides of T_t are identified with embedded subarcs of each component of \mathcal{D}_t in ∂K_t . We let $P_t = \partial K_t - \text{int}(T_t)$. Then the pair (P_t, \mathcal{D}_t) is a *spinal pair* for (K_t, \mathcal{D}_t) . From another point of view, P_t is a disc embedded in ∂K_t containing \mathcal{D}_t and with ∂P_t meeting the boundary of each disc in \mathcal{D}_t in an arc. The reason we call (P_t, \mathcal{D}_t) a spinal pair is that there is a homotopy equivalence from (P_t, \mathcal{D}_t) to $(L_t, \text{closed edges of } L_t)$, which first collapses K_t to P_t , then replaces the discs of \mathcal{D}_t by edges, and $P_t - \mathcal{D}_t$ by a central vertex for the star graph L_t . Thus the spinal pair is something like a spine. There is of course a similar homotopy equivalence between (K_t, \mathcal{D}_t) and L_t .

If $(K_t, V_t, \mathcal{D}_t)$ is a spotted product, we let in $P_t = \partial K_t - \text{int}(V_t)$ (which is just $V_t \times 1$ when K_t is given the product structure $K_t = V_t \times I$) and the *spinal pair* for

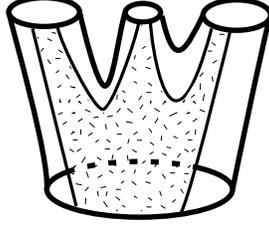


Figure 2: Polygonal disc. The complement is P_t .

$(K_t, V_t, \mathcal{D}_t)$ is (P_t, \mathcal{D}_t) . In the case of spotted products, there is no choice for P_t . Again, there is an obvious homotopy equivalence between the spinal pair and Γ_t , taking discs of \mathcal{D}_t to closed edges.

Finally, it will be convenient later in the paper to have another notion, used mostly in the case of a spotted ball (K_t, \mathcal{D}_t) . A *subspinal pair* for $(K_t, V_t, \mathcal{D}_t)$ is a pair $(\check{P}_t, \check{\mathcal{D}}_t)$ embedded in a spinal pair (P_t, \mathcal{D}_t) , such that $\check{\mathcal{D}}_t$ is (the union of) a subcollection of the discs in \mathcal{D}_t , and such that $\check{P}_t \hookrightarrow P_t$ is an embedded subdisc whose boundary meets each boundary of a disc in $\check{\mathcal{D}}_t$ in a single arc. Exceptionally, we also allow a subspinal pair consisting of a single disc component D of \mathcal{D}_t . In this case the subspinal pair is (D, D) .

Definition 5.13. Fix $s \leq t$ and consider a spotted component (K_s, \mathcal{D}_s) contained in a spotted component (K_t, \mathcal{D}_t) . We say that (K_s, \mathcal{D}_s) is ∂ -parallel to (K_t, \mathcal{D}_t) if there exist spinal pairs P_s and P_t such that there is an isotopy of (K_s, \mathcal{D}_s) in (K_t, \mathcal{D}_t) such that, after the isotopy, (P_s, \mathcal{D}_s) is contained in (P_t, \mathcal{D}_t) . Then we say the spinal pair (P_s, \mathcal{D}_s) is *parallel* to the spinal pair (P_t, \mathcal{D}_t) . Similarly, we say a *subspinal pair* $(\check{P}_s, \check{\mathcal{D}}_s)$ is *parallel* to (P_t, \mathcal{D}_t) if there is an isotopy from the first pair into the second. Finally, we also say that the graph $(L_s, \partial L_s)$ is *parallel* to (P_t, \mathcal{D}_t) if it can be isotoped in (K_t, \mathcal{D}_t) to (P_t, \mathcal{D}_t) .

We will need one more variation of the definition. In Section 7 we will have a spotted product K_t with closed interior boundary and contained in a handlebody $K_t \cup H'$, with H' a handlebody; the interior boundary of K_t is “capped” with a handlebody H' . Then we speak of $(L_s, \partial L_s)$, (P_s, \mathcal{D}_s) , or $(\check{P}_s, \check{\mathcal{D}}_s)$ being *parallel in* $K_t \cup H'$ to (P_t, \mathcal{D}_t) .

Lemma 5.14. *For $s < t \leq 1$, suppose K_t is a spotted component of $H_t \cap \widehat{H}_1$; suppose (P_t, \mathcal{D}_t) is a spinal pair for K_t ; and suppose K_s is a spotted component of $H_s \cap \widehat{H}_1$ with $K_s \subseteq K_t$. (If K_t is a spotted product, it might be capped by H' on its interior boundary.)*

Then:

- i) If K_s is a spotted ball, and the star graph L_s is parallel to (P_t, \mathcal{D}_t) (in $K_t \cup H'$) then there exists a spinal pair (P_s, \mathcal{D}_s) parallel to (P_t, \mathcal{D}_t) (in $K_t \cup H'$).*
- ii) If K_s is a spotted ball, and the embedded star subgraph \check{L}_s of L_s with ends in discs of $\check{\mathcal{D}}_s \subseteq \mathcal{D}_s$ is parallel to (P_t, \mathcal{D}_t) in K_t (in $K_t \cup H'$) then there exists a subspinal*

pair $(\check{P}_s, \check{\mathcal{D}}_s)$ parallel to (P_t, \mathcal{D}_t) in K_t (in $K_t \cup H'$).

iii) If K_s is a spotted product, and the embedded star subgraph \check{L}_s of \hat{L}_s with ends in discs of $\check{\mathcal{D}}_s \subseteq \mathcal{D}_s$ is parallel to (P_t, \mathcal{D}_t) in K_t (in $K_t \cup H'$) then there exists a subspinal pair $(\check{P}_s, \check{\mathcal{D}}_s)$ parallel to (P_t, \mathcal{D}_t) in K_t (in $K_t \cup H'$). In particular, if $\check{L}_s = \hat{L}_s$, then we obtain a subspinal pair $(\hat{P}_s, \mathcal{D}_s)$ including all spots of K_s .

iv) If K_t is a spotted ball, a choice of isotopy of pairs of $(L_t, \partial L_t)$ in (K_t, \mathcal{D}_t) to ∂K_t corresponds to a choice of a spinal pair P_t .

v) If K_t is a connected spotted product, a choice of isotopy of pairs of $(\hat{L}_t, \partial \hat{L}_t)$ in (K_t, \mathcal{D}_t) to (P_t, \mathcal{D}_t) corresponds to a choice of a subspinal pair $(\hat{P}_t, \mathcal{D}_t)$, including all spots, for K_t .

Proof. i) The argument is the same regardless K_t is capped with H' or not. Isotope L_s to $L \subseteq P_t$ (possibly passing through H'). Now note that a regular neighborhood $N(L)$ of L in K_t is a spotted ball isotopic to K_s , the component of $H_s \cap \hat{H}_t$ corresponding to (and containing) L_s . We let $P = N(L) \cap \partial K_t$, then $(P, P \cap \mathcal{D}_t)$ is a spinal pair for the spotted ball $(N(L), N(L) \cap \mathcal{D}_t)$. We can use the inverse of the isotopy taking L_s to L to obtain an isotopy taking $N(L)$ to K_s . This isotopy takes $(P, P \cap \mathcal{D}_t)$ to a spinal pair (P_s, \mathcal{D}_s) in K_s . By construction, (K_s, \mathcal{D}_s) is parallel to (K_t, \mathcal{D}_t) as desired and we have proved the isotopy of $(L_s, \partial L_s)$ to (P_t, \mathcal{D}_t) determines a P_s such that (P_s, \mathcal{D}_s) is isotopic to (P_t, \mathcal{D}_t) .

ii), iii), The proofs are very similar to the proof of i).

iv) If we consider L_t as a core graph of K_t , an isotopy of $(L_t, \partial L_t)$ in (K_t, \mathcal{D}_t) to ∂K_t gives a graph $(L, \partial L) \subseteq (\partial K_t, \mathcal{D}_t)$. Again we let $P = N(L) \cap \partial K_t$ containing \mathcal{D}_t to form a spinal pair (P, \mathcal{D}_t) for the spotted ball $(N(L), \mathcal{D}_t)$. $N(L)$ is the same as K_t up to isotopy, and $N(L) \cap \partial K_t$ gives a spinal pair for $N(L)$, so we have a spinal pair for K_t . It is easy to recover the \hat{L}_t and L_t from the spinal pair.

v) The proof is similar to the proof of iv). □

The following lemma analyzes the “events” which change the L_t 's corresponding to spotted balls or products K_t .

Lemma 5.15. *Recall K_t is a component of $H_t \cap \hat{H}_1$ and L_t is the corresponding component of $\Gamma_t \cap \hat{H}_1$. Suppose s and t are consecutive regular values, in the sense that there is exactly one critical value between them. Then one of the following holds (see Figure 3):*

1. *If the corresponding critical point is not contained in K_t then $L_t = \Gamma_t \cap K_t$ is the same as $L_s = \Gamma_s \cap K_t$, otherwise*
2. *L_t is obtained from two components L_s and L'_s by isotoping an edge of L_s to an edge of L'_s (in the complement of the remainder of $\Gamma_s \cap K_t$) and identifying the two edges, or*

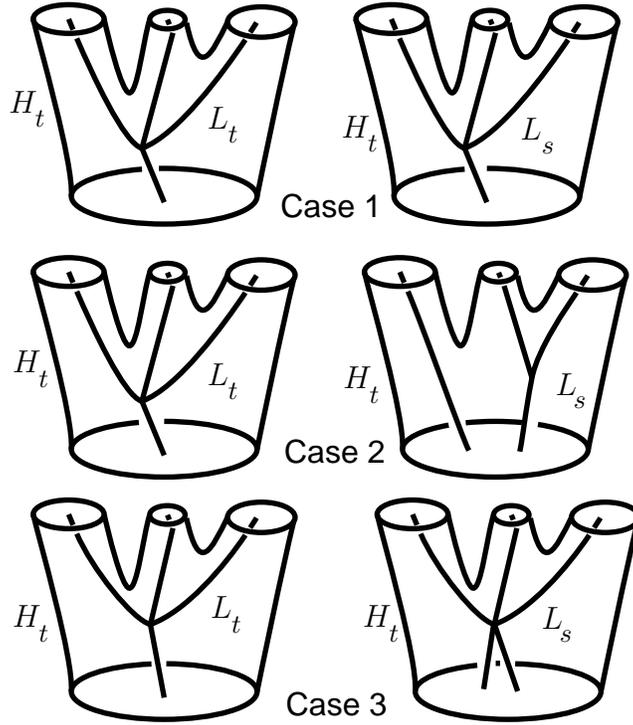


Figure 3: Event in H_t .

3. L_t is obtained from a component L_s of $\Gamma_s \cap \widehat{H}_1$ by isotoping an edge of L_s to another edge of L_s (in the complement of the remainder of $\Gamma_s \cap K_1$) and identifying the two edges.

Figure 3 shows the events in case K_t is a spotted ball.

Proof. Recall that in the context of Theorem 5.12 we are assuming that the discs \mathcal{E}_1 are in Morse position without centers with respect to the height function in $H_1 - \text{int}(H_0)$. The lemma follows from the fact that there is exactly one critical value between s and t for the Morse height function on the system of discs in H_1 . \square

We classify critical values according to the three cases of Lemma 5.15.

In the following statement “consecutive” regular values are again regular values separated by exactly one critical value.

Lemma 5.16. *For $s < t \leq 1$ consecutive regular values, suppose K_t is a spotted component of $H_t \cap \widehat{H}_1$ and suppose K_s is a spotted component of $H_s \cap \widehat{H}_1$ with $K_s \subseteq K_t$. Then for any choice of spinal pair (P_t, \mathcal{D}_t) for (K_t, \mathcal{D}_t) there is a choice of spinal pair (P_s, \mathcal{D}_s) for (K_s, \mathcal{D}_s) such that the spinal pair (P_s, \mathcal{D}_s) is parallel to the spinal pair (P_t, \mathcal{D}_t) , and hence (K_s, \mathcal{D}_s) is ∂ -parallel in (K_t, \mathcal{D}_t) .*

Proof. We first consider the case that K_t is a spotted ball, and we choose a spinal pair (P_t, \mathcal{D}_t) .

We are assuming that the discs \mathcal{E}_1 are in Morse position without centers with respect to the height function in $H_1 - \text{int}(H_0)$. We assume s and t are “consecutive” regular values, in the sense that there is exactly one critical value between them, and that the corresponding critical point lies in K_t . Lemma 5.15 classifies the events corresponding to the critical value: A component L_s of $\Gamma_s \cap K_t$ is obtained from a component L_t of $\Gamma_t \cap K_t$ as described in the lemma. In Case 1, the component L_s coincides with the corresponding L_t . In particular $L_s \subseteq L_t$ in this case. In Case 2, where L_s is isotopic to a subspace of L_t , which is split, we can isotope it so that $L_s \subseteq L_t$. In Case 3, L_s is obtained from L_t by splitting the edge e , and we can isotope L_s so that it is contained in L_t away from a neighborhood $N(e) \subseteq K_t$ of e .

But L_t is clearly *parallel to* P_t , in the sense that there exists an isotopy of $(L_t, \partial L_t)$ in (K_t, \mathcal{D}_t) taking $(L_t, \partial L_t)$ to (P_t, \mathcal{D}_t) . Therefore, in Cases 1 and 2, when $L_s \subseteq L_t$, the parallelism of L_t realizes the parallelism of L_s and by Lemma 5.14, we obtain a spinal pair (P_s, \mathcal{D}_s) which is parallel to the spinal pair (P_t, \mathcal{D}_t) .

We deal with Case 3 as follows. We can regard the isotopy of L_t into P_t as a composition of two isotopies: first isotope a neighborhood $L_t \subseteq K_t$ (containing e) into a neighborhood $N(P_t) \simeq P_t \times I$. We may assume that L_t is taken to a horizontal slice $P_t \times \{x\}$. The isotopy of L_t is completed by vertical projection. Now we can assume $L_s \subseteq K_t$ is contained in L_t away from $N(e)$. After the first isotopy we then have that $(L_s - N(e)) \subseteq P_t \times \{x\}$. But we can isotope $L_s \cap N(e)$ (rel $\partial N(e)$) so that $L_s \subseteq P_t \times \{x\}$. Vertical projection gives $L_s \subseteq P_t$. Again, by Lemma 5.14 we obtain a spinal pair (P_s, \mathcal{D}_s) which is parallel to the spinal pair (P_t, \mathcal{D}_t) .

Now we consider the case that K_t is a spotted product and K_s is a spotted ball. There is no choice for the spinal pair (P_t, \mathcal{D}_t) . Again, analysis of the cases of Lemma 5.15 shows that in each case L_s can be isotoped to P_t , so by Lemma 5.14 we again get a spinal pair (P_s, \mathcal{D}_s) which is parallel to (P_t, \mathcal{D}_t) .

Finally, if both K_t and K_s are spotted products, then there is no choice for either spinal pair (P_t, \mathcal{D}_t) or (P_s, \mathcal{D}_s) . The surface P_s is isotopic to P_t and \mathcal{D}_s represents a subcollection of the discs of \mathcal{D}_t . Hence (P_s, \mathcal{D}_s) is automatically parallel to (P_t, \mathcal{D}_t) \square

Proposition 5.17. *For all $s < t \leq 1$, for any component K_t of $H_t \cap \widehat{H}_1$ and for any component K_s of $H_s \cap \widehat{H}_1$ with $K_s \subseteq K_t$, K_s is ∂ -parallel in K_t .*

Proof. There is a sequence t_i , $1 \leq i \leq n$ where $t_1 = s$, $t_n = t$ and with consecutive values t_i separated by a single critical value. By Lemma 5.16, given a spinal pair $(P_{t_n}, \mathcal{D}_{t_n})$ for K_{t_n} , there is a parallel spinal pair $(P_{t_{n-1}}, \mathcal{D}_{t_{n-1}})$ for $K_{t_{n-1}}$. We use induction with decreasing i : If we have found a spinal pair $(P_{t_i}, \mathcal{D}_{t_i})$ for K_{t_i} which is parallel to $(P_{t_{i+1}}, \mathcal{D}_{t_{i+1}})$ for $K_{t_{i+1}}$, then we can find a spinal pair $(P_{t_{i-1}}, \mathcal{D}_{t_{i-1}})$ for $K_{t_{i-1}}$ which is parallel to the pair $(P_{t_i}, \mathcal{D}_{t_i})$ for K_{t_i} . This then proves we have a sequence of parallel spinal pairs. It is easy to see that parallelism for spinal pairs is transitive, whence we obtain the statement of the lemma. \square

Proof of Theorem 5.12. For a leaf ℓ of Ω , every component of $\ell \cap \widehat{H}_1$ is the core of a spotted ball K_s with two spots, for some s . K_s is contained in some component K of \widehat{H}_1 . Letting $t = 1$ in the previous lemma, we see that K_s is ∂ -parallel in K_t , which implies the arc of ℓ in \widehat{H}_1 is ∂ -parallel, hence unknotted. \square

6 Automorphisms of sphere bodies, the Train Track Theorem

Our goal in this section is to prove Theorem 1.5.

We recall from the Introduction that g is *train track generic* if it is generic on H_1 and the corresponding invariant 2-dimensional lamination determines, through a quotient map $H \rightarrow G$ on a graph G , a homotopy equivalence which is a train track map.

Our strategy for proving Theorem 1.5 is similar to that of [1] for improving homotopy equivalences of graphs. Given an automorphism $f : M \rightarrow M$ we first isotope it so that it preserves the canonical splitting $M = H \cup H'$ (see Section 2 and Theorem 1.4). Let $g = f|_H : H \rightarrow H$. By Theorem 1.1 g is either periodic, reducible or generic. If it is periodic or reducible the proof is done, so assume that it is generic. We consider the corresponding 2-dimensional lamination and homotopy equivalence of a graph. If it is a train-track generic map, the proof is also done, so assume otherwise. This means that a power of the induced homotopy equivalence on the graph admits “back tracking”, in the sense of [1]. In that context, there is a sequence of moves that simplify the homotopy equivalence. Similarly in our setting there are analogous moves that realize this simplification. The result of this simplification is an automorphism g' of H which is either periodic, reducible or generic. In the last case, the growth rate of g' is smaller than that of g . Proceeding inductively, this process yields the proof of the theorem. In general g and g' , which are automorphisms of H , will not be isotopic. The key here is that they are both restrictions of automorphisms f, f' of M to the handlebody H . These f and f' will be isotopic (through an isotopy that does not leave invariant the splitting, see e.g. Example 2.7).

In the light of the sketch above, it is clear that Theorem 1.5 will follow from the result below.

Theorem 6.1. *Let $f : M \rightarrow M$ be an automorphism preserving the canonical splitting $M = H \cup H'$. Assume that $g = f|_H$ is generic and not train track. Then f is isotopic to an f' preserving the splitting such that $g' = f'|_H$ is either periodic, reducible or generic with growth rate smaller than g .*

There are two technical ingredients in proving Theorem 6.1. The first is to prove the existence of “half-discs”, which are geometric objects that represent some of the homotopy operations performed in the algorithm of [1] (these operations are “folding” and “pulling-tight”). In special situations, when the half-discs are “unholed,” we can

use them to realize these homotopies by isotopies. These isotopies are described in [16] (a “down-moves” corresponds to a “fold” and a “diversion” to a “pulling-tight”). We can then proceed to the next step of the algorithm remaining in the same isotopy class, simplifying the automorphism by either reducing the growth rate or turning it into a periodic or reducible automorphism.

The second ingredient is an operation called “unlinking”; the aim is to isotope f so that we can obtain these special unholed half-discs from general ones. As noted above, this makes it possible to proceed with the algorithm. An unlinking preserves the canonical splitting but, unlike the other moves, it is an isotopy which does not leave invariant the canonical splitting. In the unlinking process we will decompose the given half-disc into “rectangles” and a “smaller” half-disc.

We have sketched the proof of Theorem 6.1 (which implies Theorem 1.5). We now turn to details. We fix $f: M \rightarrow M$ preserving the splitting $M = H \cup H'$ and consider its restriction $g: H \rightarrow H$. We assume that g is generic. Now consider the construction of the invariant laminations, determining H_t , \mathcal{E}_t and Γ_t for $t \in \mathbb{R}$, as in Section 5. We also use the notions of spotted balls as in Section 5, and we work in the space \widehat{H}_1 , i.e., $H_1|_{\mathcal{E}_1}$ as before. In general, for $t \leq 1$, the pair (K_t, \mathcal{D}_t) will represent a spotted ball where K_t is a component of $H_t \cap \widehat{H}_1$ and $\mathcal{D}_t = K_t \cap \widehat{\mathcal{E}}_1$.

Definitions 6.2. A *half-disc* is a triple (Δ, α, β) where Δ is a disc and $\alpha, \beta \subseteq \partial\Delta$ are complementary closed arcs. We may abuse notation and say that Δ is a half-disc.

Let $s < t$. A *half-disc for H_s in H_t* is a half-disc (Δ, α, β) satisfying:

- (i) $\Delta \subseteq K_t$ is embedded for some K_t ,
- (ii) β is contained in a spot D_t of K_t ,
- (iii) $\alpha \subseteq \partial K_s$ for some $K_s \subseteq K_t$ and $\Delta \cap K_s = \alpha$,
- (iv) α *connects* (i.e., intersects the boundaries of) two distinct spots of (K_s, \mathcal{D}_s) .

We refer to α , the arc in ∂K_s as the *upper boundary of the half-disc for H_s in H_t* , $s < t$, denoted by $\partial_u \Delta$, and β as its *lower boundary*, denoted by $\partial_l \Delta$.

Similarly, a *rectangle* is a triple (R, γ, γ') where R is a disc and $\gamma, \gamma' \subseteq \partial R$ are disjoint closed arcs. We may abuse notation and say that R is a rectangle.

Let $s < t$. A *rectangle for H_s in H_t* is a rectangle $(R_s, \gamma_s, \gamma_t)$ satisfying:

- (a) $R_s \subseteq K_t$ is embedded for some K_t ,
- (b) $\gamma_s \subseteq \partial K_s$ for some $K_s \subseteq K_t$ and $R_s \cap K_s = \gamma_s$,
- (c) γ_s connects distinct spots of (K_s, \mathcal{D}_s) ,
- (d) $\gamma_t \subseteq \partial K_t - \text{int}(\mathcal{D}_t)$,
- (e) $\partial R_s - (\gamma_s \cup \gamma_t) \subseteq \mathcal{D}_t$.

We note that a half-disc Δ (for H_s) may be *holed*, in the sense that H_s may intersect the interior of Δ . If this intersection is essential then Δ also intersects H_0 . We define *holed rectangles* similarly.

Definition 6.3. Let $0 \leq s < t \leq 1$ and Δ a half-disc for H_s in H_t . Suppose that $s = t_0 < t_1 < \dots < t_k = t$ is a sequence of regular values. We say that Δ is *standard with respect to $\{t_i\}$* if there is $0 \leq m < k$ such that the following holds. The half-disc Δ is the union of rectangles R_{t_i} , $s \leq t_i \leq t_{m-1}$, and a half-disc Δ_{t_m} (see Figure 4) with the properties that:

1. each R_{t_i} , $s \leq t_i < t_m$, is a rectangle for H_{t_i} in $H_{t_{i+1}}$,
2. the holes $\text{int}(R_{t_i}) \cap H_{t_i}$ are essential discs in H_{t_i} ,
3. Δ_{t_m} is an unholed half-disc for H_{t_m} in $H_{t_{m+1}}$.

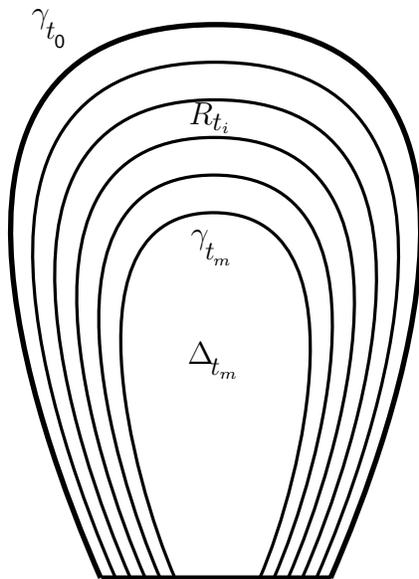


Figure 4: Standard half disc.

We want to prove the following:

Proposition 6.4. *Let $0 = t_0 < \dots < t_k = 1$ be a complete sequence of regular values. If a component of $H_0 \cap \widehat{H}_1$ intersects a disc of $\widehat{\mathcal{E}}_1$ in two distinct components D_0, D_1 , then there exists a half-disc Δ for H_0 connecting D_0 and D_1 which is standard with respect to $\{t_i\}$.*

To prove the proposition, we need:

Lemma 6.5. *Suppose that there exist spinal pairs P_{t_i} such that (K_{t_i}, P_{t_i}) is parallel to $(K_{t_{i+1}}, P_{t_{i+1}})$. If a component K_0 intersects a disc of \mathcal{D}_1 in two distinct components D_0, D_1 , then there exists a half-disc for H_0 connecting D_0 and D_1 which is standard with respect to $\{t_i\}$.*

Proof. Since $P_{t_0} = P_0$ is a spinal pair it intersects both D_0, D_1 . Let $\gamma_{t_0} \subseteq P_{t_0}$ be an arc connecting D_0 to D_1 .

For any i , let $\gamma_{t_i} \subseteq P_{t_i}$ be an embedded arc with endpoints in distinct spots of K_{t_i} . Consider an isotopy realizing the parallelism between (K_{t_i}, P_{t_i}) and $(K_{t_{i+1}}, P_{t_{i+1}})$, taking P_{t_i} to a subsurface of $P_{t_{i+1}}$. The arc γ_{t_i} is taken to an arc $\gamma_{t_{i+1}} \subseteq P_{t_{i+1}}$. It is clear that the isotopy yields, via the Loop Theorem, a rectangle $(R_{t_i}, \gamma_{t_i}, \gamma_{t_{i+1}})$ for H_{t_i} in $H_{t_{i+1}}$. Here we note that indeed $R_{t_i} \cap K_{t_i} = \gamma_{t_i}$, as needed. It can happen, though, that H_{t_i} intersects R_{t_i} in other components. In this case there exists another $K'_{t_i} \neq K_{t_i}$ contained in $K_{t_{i+1}}$. Consider then the graph L'_{t_i} corresponding to K'_{t_i} and perturb R_{t_i} so that $L'_{t_i} \cap R_{t_i}$ is transverse. Now make $K'_{t_i} \cap R_{t_i}$ consist of essential discs (in H_{t_i}).

From γ_{t_0} defined in the beginning of the proof we use the process described above to proceed inductively, obtaining rectangles $(R_{t_i}, \gamma_{t_i}, \gamma_{t_{i+1}})$. We do that for as long as γ_{t_i} has endpoints in distinct spots of $K_{t_{i+1}}$ (and hence in distinct spots of K_{t_i}).

It is clear that eventually γ_{t_i} will have both endpoints in the same spot of the corresponding $K_{t_{i+1}}$. To see this simply note that all γ_{t_j} 's are parallel (as pairs $(\gamma_{t_j}, \partial\gamma_{t_j})$ in (K_1, \mathcal{D}_1)) and hence have endpoints in the same spot of K_1 by hypothesis.

So let γ_{t_m} be the first such arc which has both endpoints in the same spot of $K_{t_{m+1}}$, say D . In this case we still build $\gamma_{t_{m+1}} \subseteq P_{t_{m+1}}$ through parallelism. Now consider $\beta \subseteq (\partial D \cap \partial P_{t_{m+1}})$ with $\partial\beta = \partial\gamma_{t_{m+1}}$. Then $\beta \cup \gamma_{t_{m+1}}$ bounds a disc $D' \subseteq P_{t_{m+1}}$. As before, the parallelism between γ_{t_m} and $\gamma_{t_{m+1}}$ defines a rectangle $(R_{t_m}, \gamma_{t_m}, \gamma_{t_{m+1}})$. We consider $\Delta_{t_m} = R_{t_m} \cup D'$ and push (Δ_{t_m}, β) into $(K_{t_{m+1}}, D)$ as a pair. Now $(\Delta_{t_m}, \gamma_{t_m}, \beta)$ is a half-disc for H_{t_m} in $H_{t_{m+1}}$.

From the construction it is clear that if R_{t_m} is unholed then so is Δ_{t_m} . We claim that that is the case. Indeed, recall that $\{t_i\}$ is complete and that $\gamma_{t_{m+1}}$ is the first γ_{t_i} to have both endpoints in the same spot of $K_{t_{i+1}}$. Therefore K_{t_m} may be regarded as obtained from $K_{t_{m+1}}$ by the event pictured Figure 3, Case 3. In particular, there is no other $K'_{t_m} \neq K_{t_m}$ in $K_{t_{m+1}}$. Therefore R_{t_m} is unholed, so the same holds for Δ_{t_m} .

It is now easy to see that $\Delta = (\bigcup_{t_0 \leq t_i \leq t_{(m-1)}} R_{t_i}) \cup \Delta_{t_m}$ is 1) a half-disc for H_0 , 2) connects D_0 and D_1 and 3) is standard with respect to $\{t_i\}$. \square

Remark 6.6. Note from the construction of Δ above that the rectangles R_{t_i} are holed only if there are two distinct K_{t_i}, K'_{t_i} contained in $K_{t_{i+1}}$, as pictured in Figure 3, Case 2.

Proof of Proposition 6.4. Follows directly from the result above, Theorem 5.12 and Lemma 5.17. \square

We wish to use a standard half disc to obtain a simplified representative of the automorphism.

Let Δ be a half-disc for H_0 , with upper boundary on some K_0 . If its interior did not intersect H_0 , the half-disc Δ could be used to drag H_0 so it intersects \mathcal{E}_1 in fewer discs, but recall that Δ may be holed. A solution would be to isotope 1-handles of H_0 , to avoid Δ . One might worry that the handles that make holes in Δ may be “linked” with K_0 . More precisely, it may be impossible to isotope the handles of H_0 in H (leaving K_0 fixed) in order to avoid holes in Δ . We emphasize that the word “linked” is abused of here. The property it describes is not intrinsic to the K_0 ’s in K_1 , and rather relative to the half-disc Δ . In any case, these “essential links” do appear in the setting of handlebodies (see [16, 3, 4]). But if we allow the isotopies to run through the whole manifold M , not just H , the unlinking is always possible. The idea is to use isotopies of the kind described in Example 2.7.

More specifically we shall describe an isotopy of the type in Example 2.7 and how it can be used to unlink. We call an isotopy of this type *an unlinking move* or just an *unlinking*.

At this point we do not worry about the automorphism $f: M \rightarrow M$. Although f is the object of interest, we will intentionally disregard it initially, thinking instead of the various $H_t \subseteq H$, disc systems and sphere systems, as geometric objects embedded in M . We will bring f back into the picture later in the argument.

As usual, let K be a component of \widehat{H}_1 . Let $0 \leq s \leq 1$ and consider a component K_s of $H_s \cap K$. Recall that a spot D_s of (K_s, \mathcal{D}_s) corresponds to a disc $E_s \subseteq H_s$. Such a disc extends to an essential sphere $S_s \hookrightarrow M$ with the following properties: 1) $S_s \cap H_s = E_s$ and 2) for $s < t \leq 1$, $S_s \cap H_t$ consists of a single disc. We say that such a sphere is *dual to H_s* , or just a *dual sphere* in case the value of s is clear.

The unlinking move will be performed along a rectangle (coming from a standard half-disc) and a sphere. Let K be a component of \widehat{H}_1 . Fix $0 \leq s < t \leq 1$, with exactly one critical point between s and t as usual. Consider a component K_s of $H_s \cap K$ and the component K_t of $H_t \cap K$ containing K_s . In this situation, most spots K_t contain exactly one spot of K_s , see Figure 3. We shall take advantage of this fact. Consider a rectangle $(R_s, \gamma_s, \gamma_t)$ for H_s in H_t with $R_s \subseteq K_t$. Let K_s be the component that contains $\gamma_s \subseteq \partial R_s$. Assume that γ_s connects distinct spots of K_s (which is the case if the rectangle comes from a standard half-disc). Finally, since most spots of K_t contain exactly one spot of K_s , it is reasonable also to suppose that at least one of the two sides of R_s in a spot of K_t lies in a spot containing only one spot of K_s . If this is the case, we can assume that the dual sphere S_s containing D_s has the property that $S_s \cap K_t$ is a spot of K_t . In other words, we are assuming that S_s is also dual to H_t , which is a strong restriction. Thus, $S_s \cap R_s \subseteq \partial R_s$, a key property we need for the unlinking move.

Consider a component C of $H_s \cap R_s$ distinct from γ_s . Assume that C is an essential disc in H_s , which is the case if the rectangle comes from a half-disc in

standard position. Let K'_s be the component containing C . The move that we will describe isotopes K'_s along R_s and S_s , see Figure 5. More precisely, consider an embedded path $\sigma \subseteq R_s - \text{int}(H_s)$ connecting C to D_s . We first isotope C along σ so it is close to $D_s \subseteq S_s$. This move changes K'_s in a neighborhood $N \subseteq K'_s$ of C . We note that the condition imposed on S_s that $S_s \cap K_t$ is a spot of K_t implies that $\sigma \cap S_s = \sigma \cap D_s \subseteq \partial\sigma$. Hence the move described does not introduce intersections of K'_s with S_s . So S_s is preserved as a dual sphere and the disc $\check{D}_s = S_s - \text{int}(D_s)$ does not intersect H_s (except at $\partial\check{D}_s = \partial D_s$).

We now regard N as the neighborhood of an arc α transverse to R_s and close to an arc $\alpha' \subseteq \partial D_s$ (this α' also intersecting R_s). Let $\beta' \subseteq \partial D_s$ be the closed complement of α' .

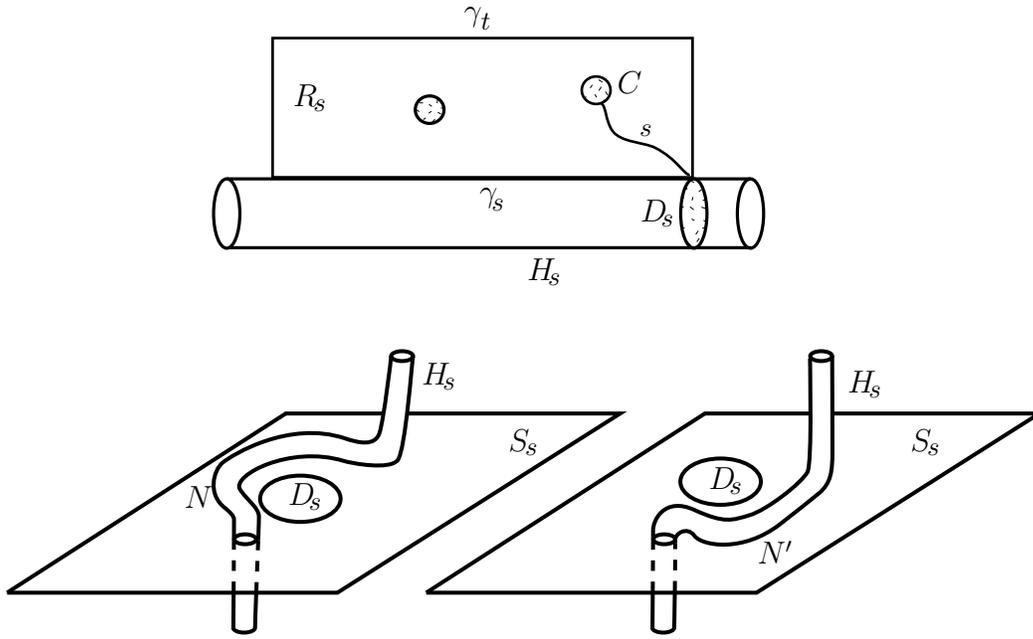


Figure 5: Unlinking move.

Now recall that the disc \check{D}_s is disjoint from H_s . We can push the arc α close to α' and isotope it along \check{D}_s to an arc $\beta \subseteq K_t - \text{int}(H_s)$ close to β' . More precisely, consider an arc $\beta \subseteq \text{int}(K_t) - H_s$ such that 1) $\alpha \cap \beta = \partial\alpha = \partial\beta$ and that 2) the circle $\alpha \cup \beta$ is parallel in $K_t - \text{int}(H_s)$ to $\partial D_s = \alpha' \cup \beta'$. Then α is isotopic to β (along the disc \check{D}_s). If N' is a neighborhood of β we can isotope N to N' . But $\beta \cap R_s = \emptyset$ then also $N' \cap R_s = \emptyset$. Clearly, this operation does not change H_s away from N , therefore such an operation reduces $H_s \cap R_s$. We call this move an *unlinking along R_s and S_s* . We also say that this unlinking is *performed in H_t* . We emphasize that H_t remains unchanged and that the resulting H_s still is contained in H_t .

The following summarizes properties of an unlinking that are either clear or were already mentioned above.

Lemma 6.7. *Consider H'_s obtained from H_s by an unlinking along R_s and S_s realized in H_t . Then:*

- (i) $|H'_s \cap R_s| < |H_s \cap R_s|$,
- (ii) $H_t \subseteq H_1$ remains unchanged and $H'_s \subseteq H_t$ is isotopic to H_s in M , and
- (iii) the spots D_s of K_s also remain unchanged, therefore $H'_s \cap \mathcal{E}_1 = H_s \cap \mathcal{E}_1$.

Proposition 6.8. *Let $0 = t_0 < t_1 < \dots < t_{k-1} < t_k = 1$ be a complete increasing sequence of regular values. Suppose that Δ is a half-disc for $H_0 \subseteq H_1$, standard with respect to $\{t_i\}$. Then there exists a sequence of unlinkings taking H_0 to $H'_0 \subseteq H_1$ for which Δ is an unholed half-disc. Moreover $H'_0 \cap \mathcal{E}_1 = H_0 \cap \mathcal{E}_1$.*

Proof. Let $(R_{t_i}, \gamma_{t_i}, \gamma_{t_{i+1}})$, $t_0 \leq t_i \leq t_{m-1}$ and Δ_{t_m} be the rectangles and half-disc contained in Δ determined by $\{t_j\}$. For each of the rectangles R_{t_i} we will perform a sequence of unlinkings along R_{t_i} and a carefully chosen sphere S_{t_i} attached to one side of R_{t_i} . The effect will be to unlink K_{t_i} in the $K_{t_{i+1}}$ containing R_{t_i} . The unlinkings will be performed successively using induction on i , this time with i increasing. At the end of the inductive process, all rectangles will be unholed, leading to an unholed Δ (recall that Δ_{t_m} is unholed from the start and this property will be preserved).

As we have mentioned before, an important part of the argument will be aimed at the following. Each R_{t_i} is contained in a $K_{t_{i+1}}$ and intersects a K_{t_i} . It is necessary to find a spot D_{t_i} of K_{t_i} which 1) intersects R_{t_i} and 2) extends to a dual sphere S_{t_i} with the property that $S_{t_i} \cap K_{t_{i+1}} = D_{t_{i+1}}$ for some spot $D_{t_{i+1}}$ of $K_{t_{i+1}}$ (see Lemma 6.9 below). This means that one side of R_{t_i} intersects a spot of $K_{t_{i+1}}$ containing just one spot of K_{t_i} . Recall that this property is necessary to perform an unlinking.

Let K be the component of \widehat{H}_1 that contains Δ and let K_0 be the component of $H_0 \cap K$ that contains the upper boundary α of Δ . We assume that α intersects two distinct 1-handles of H_0 and thus a 0-handle. If it intersects just one 1-handle (hence no 0-handle) then we replace a disc of \mathcal{E}_0 , the one corresponding to this 1-handle, by two parallel copies, which has the effect of introducing a 0-handle in K_0 . This amounts to choosing a new system \mathcal{E} a, and there are choices in how $f(\mathcal{E}_0) = \mathcal{E}_1$ intersects H_0 in discs parallel to discs of \mathcal{E}_0 , i.e. choices for the incidence matrix, see Section 5.

Our modification ensures that each $K_{t_i} \supseteq K_0$, $t_i > 0$, corresponds to a 0-handle, with 1-handles giving distinct spots on ∂K_t . Therefore each rectangle R_{t_i} has the following property: the spots of K_{t_i} that $\gamma_{t_i} \subseteq \partial R_{t_i}$ connects correspond to distinct discs of \mathcal{E}_{t_i} . That property will be important in finding the sphere S_{t_i} intersecting $K_{t_{i+1}}$ at a spot.

Consider two consecutive terms $s < t$ of the sequence (so there is precisely one singular value between them). We need the following.

Lemma 6.9. *A holed rectangle R_s intersects a spot of K_s with the property that the corresponding dual sphere intersects K_t in a single spot.*

Proof. It clearly suffices to show that R_s intersects a spot D_s with the following property. If D_t is the spot of K_t containing D_s then $K_s \cap D_t = D_s$.

Recall that we are assuming that no disc of $\mathcal{E}_1 \subseteq H_1$ is represented twice as a spot of a single component K_1 of \widehat{H}_1 .

There are three cases. The case analysis is exactly the same as in the proof of 5.17, pictured in Figure 3. The simplest and most common is Case 1, where a single component K_s of $H_s \cap K$ is contained in K_t and with the same isotopy type (in K) as K_t . Another possibility is Case 2; that there are two distinct K_s 's in K_t . In Case 3, there is a single K_s in K_t but K_s is not isotopic to K_t . In Cases 1 and 3, there is only one K_s in K_t , so in fact there is no real need for unlinking; there may be a rectangle, but there is no danger that it is holed by a K'_s in K_t , see Remark 6.6. Therefore we only need to deal with Case 2.

In Case 2 there are two distinct components of $H_s \cap K$ contained in K_t : K_s and K'_s . Here one spot D_t of K_t contains precisely one spot of each of K_s, K'_s . Denote these by D_s and D'_s respectively. The other spots of K_t contain a single spot, of either K_s or K'_s . Now all spots of K_s, K'_s other than D_s, D'_s extend to distinct spots of K_t . But each of K_s, K'_s contains just one of these exceptional spots. Recalling that γ_s connects two distinct spots, at least one will have the desired property. The illustration of Case 2 in Figure 3 is somewhat misleading, since it is not easy to picture a holed rectangle there. Figure 6 is more realistic. It shows a rectangle R_s essentially intersected by K'_s .

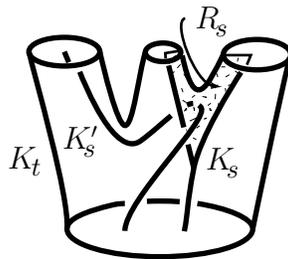


Figure 6: Linking and the unlinking move.

□

We continue the proof of Proposition 6.8. We are considering two consecutive terms $s < t$ of the sequence. Suppose that for any $t_i < s$ the rectangle R_{t_i} is unholed. Suppose also that this process of removing holes did not alter H_s . Therefore the spheres dual to H_s are preserved and so the conclusions of Lemma 6.9 above still hold for R_s .

The goal now is to prove the induction step and remove components (holes) of $H_s \cap R_s$ using unlinkings. We shall do this while also preserving H_t . Consider K_s containing γ_s . The arc $\gamma_s \subseteq \partial R_s$ connects two spots of K_s . As mentioned in the

previous paragraph, at least one of these has the property that the corresponding sphere S_s intersects K_t at a spot, therefore an unlinking along R_s and S_s is possible.

We consider components of $H_s \cap R_s$ disjoint from γ_s . Note that these are essential discs in H_s . Indeed, Δ was standard in the beginning. This was true before any unlinking moves were done because Δ was standard. But our inductive procedure ensures that H_s remains unchanged as we do unlinkings in lower levels, therefore the property is preserved. We perform the unlinking move to remove such a disc component of $H_s \cap R_s$. As stated in Lemma 6.7, this reduces $|H_s \cap R_s|$. The idea is to continue performing unlinking moves inductively until $H_s \cap R_s = \gamma_s$. Here again we note that unlinkings preserve spots of K_s and K_t (Lemma 6.7), therefore the property in the conclusion of Lemma 6.9 is preserved.

Lemma 6.7 also ensures that each unlinking does not introduce intersections of H_0 with the other rectangles R_{t_i} , $t_i < s$, and that Δ is preserved as a half-disc. It also shows that the resulting $H_0 \subseteq H_s$ is contained in H_1 and that $H_0 \cap \mathcal{E}_1$ is left unchanged. We now have all R_{t_i} for $t_i \leq s$ unholed, preserving the other desired properties of the conclusion of the proposition. This proves the induction step.

We recall that Δ_{t_m} is unholed originally. Since unlinkings along R_{t_i} , $t_i < t_m$ preserve H_{t_m} , Δ_{t_m} remains unholed, completing the proof. \square

So far, we have introduced the unlinking move and used it to obtain an unholed half-disc out of a holed one. As mentioned before, we omitted any reference to the automorphism $f: M \rightarrow M$, which is the object of interest. We now use Proposition 6.8 to obtain the desired unholed half-disc by isotopies of f .

The reader may have anticipated the idea for dealing with f : A sequence of unlinkings is isotopic to the identity, hence the composition is realizable by an isotopy of f . The catch here is that it is not obvious that the resulting automorphism preserves some important features of the original one. Among these is the property of preserving the canonical Heegaard splitting of M , restricting to an automorphism of H . It is also needed that this handlebody automorphism is outward expanding with respect to H'_0 and that the growth rate does not increase (in fact, it will remain unchanged).

Recall that handlebodies $H \subseteq H'$ are “concentric” if $H' - \text{int}(H) \simeq (\partial H \times I)$. The main technical lemma in this discussion is the following:

Lemma 6.10. *In the situation described in Proposition 6.8 we have that $H'_0 \subseteq H_1$ are concentric.*

Proof. With the same complete increasing sequence $0 = t_0 < t_1 < \dots < t_{k-1} < t_k = 1$ used in the proof of Proposition 6.8, suppose s, t are consecutive terms of the complete sequence. Suppose we can show that an unlinking move replacing H_s by H'_s in H_t leaves H'_s concentric in H_t . Then the concentricity of each H_{t_i} in $H_{t_{i+1}}$ is preserved by the unlinking move, and inductively, we then show that all the sequence of unlinkings preserves concentricity.

To prove that an unlinking move replacing H_s by H'_s in H_t leaves H'_s concentric in H_t , we first observe that the concentricity of H_s in H_t follows from the fact that for all but one K_t , there is just one component $K_s \subseteq K_t$ of $H_s \cap K_t$. For all but two components K_t , the unique $K_s \subseteq K_t$ is concentric, i.e. there is a product structure for $K_t - K_s$. The two exceptional K_t 's are illustrated in Figure 3, on the right in Cases 2 and 3. The transition to a picture before the event, replacing H_s by H'_s , which intersects every K_t concentrically, is made by adding a neighborhood of a pinching half-disc, as illustrated in Figure 7.

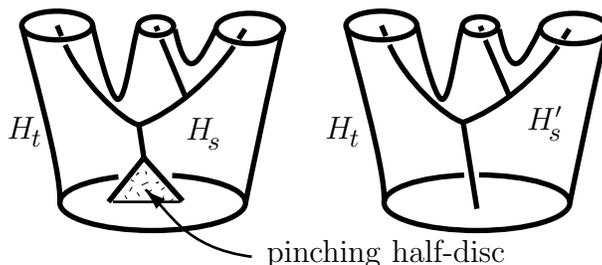


Figure 7: The pinching half-disc.

Now we need only observe that the unlinking move does not affect the pinching half-disc. Namely, the rectangle R_s can easily be chosen to be disjoint from the pinching half-disc. Thus after the unlinking move, we can still add a neighborhood of the pinching half disc to obtain an H'_s perfectly concentric in each K_t . \square

Proof of Theorem 6.1. Let G be the quotient graph determined by the 2-dimensional lamination invariant under g . Then g determines a homotopy equivalence $h_g: G \rightarrow G$. By hypothesis, h_g is not train-track. Consider the algorithm of Bestvina-Handel to simplify the homotopy equivalence. It clearly suffices to show that we can carry out the algorithm using isotopies of f that preserve the canonical splitting.

The following moves of [1] are realizable by isotopies of g (and hence isotopies of f): *collapses*, *valence-one* and *valence-two homotopies* and *subdivisions*. We refer the reader to [16] for details on these isotopies.

We shall consider the other moves, *folds* and *pulling-tights*, more carefully. For a fold, two edges e, e' of G incident to the same vertex are mapped to the same edge. In the context of $g: H \rightarrow H$ this corresponds to the following situation. There is a component K of H_1 and a $K_0 \subseteq H_0 \cap K$ with the property that two distinct spots D_0, D'_0 of K_0 are contained in the same spot D_1 of K , see Figure 8.

The spots D_0, D'_0 correspond to the edges e, e' and D_1 to the edge to which they are mapped. By Proposition 6.4, there exists a standard half-disc Δ for H_0 connecting these two spots.

If Δ is unholed, the fold is realizable by a *down-move* [16], which is an isotopy of g . Roughly speaking, such a move changes g so that the disc corresponding to the spot D_1 , which is the image of a disc $E \subseteq \mathcal{E}_0$, is isotoped along Δ . The two discs of

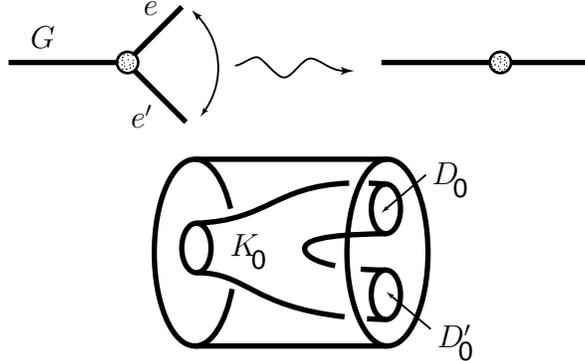


Figure 8: A fold in the quotient graph and the corresponding H_0 in H_1 .

\mathcal{E}_0 corresponding to the spots D_0, D'_0 are replaced by their band-sum along the upper boundary of Δ . Such a move realizes the desired fold.

If Δ is holed then we use Proposition 6.8 and obtain H'_0 . Lemma 6.7 assures that H_1 (and the canonical splitting) are preserved. Lemma 6.10 shows that $H'_0 \subseteq H$ and H are concentric, so the resulting automorphism of M can be adjusted to preserve the canonical splitting $M = H \cup H'$. Moreover the resulting restriction $g': H \rightarrow H$ can be made outward expanding with respect to H'_0 (see Section 5). The assertion of Proposition 6.8 that $H'_0 \cap \mathcal{E}_1 = H_0 \cap \mathcal{E}_1$ implies that the original disc system yields invariant laminations with the same growth rate. In fact more holds. Let $H \rightarrow G'$ be the new quotient map and $h_{g'}: G' \rightarrow G'$ the new homotopy equivalence. Then $h_{g'}$ and h_g are conjugate by a homeomorphism $G' \rightarrow G$. Therefore the half-disc Δ still corresponds to the fold being considered. But now Δ is not holed so we fall in the previous case and realize the fold.

For a pulling-tight the argument is similar, so we will skip details. In this case there exists an edge of G in whose interior h_g fails to be locally injective. As before, in the setting of the handlebody H that corresponds to a $K_0 \subseteq K$ with two distinct spots in the same spot of K (in this case K_0 will have just two spots). So, here again, that corresponds to a standard half-disc Δ . If it is not holed we perform a *diversion* [16], which realizes the pulling-tight move. If it is holed we remove the holes with unlinkings and return to the previous case. \square

Remark 6.11. We can say more about the resulting f' and $g' = f'|_H$ in the conclusion of Theorem 1.5. Bestvina-Handel's algorithm stops when either 1) the homotopy equivalence admits no back-tracking and the corresponding incidence matrix is irreducible or 2) when the incidence matrix, after collapsing of invariant forests, is reducible. In this last case g' is either reducible or periodic. In the first case, if the growth $\lambda = 1$ then g' may be isotoped to permute (transitively) the discs of \mathcal{E}_0 . This means that a power $(g')^k$ fixes \mathcal{E}_0 . Therefore $(f')^k$ fixes the sphere system $\mathcal{S}_0 \subseteq M$ corresponding to \mathcal{E}_0 . By reducing $(f')^k$ along \mathcal{S}_0 we obtain automorphisms of holed

balls preserving each boundary component, therefore isotopic to the identity. This proves that $(f')^k$ is isotopic to a composition of twists on spheres of \mathcal{S}_0 . In particular $(f')^{2k}$ is isotopic to Id_M (although in general g' will not be periodic). If $\lambda > 1$ then g' may be either reducible or generic. If it is generic, as stated in the theorem, it is train-track. If it is reducible, we believe we can prove that it is a lift of a pseudo-Anosov automorphism of a surface F to the total space of an I -bundle over F (in this case f restricted to the splitting surface ∂H is reducible). The proof of this last statement is not written formally yet.

7 Automorphisms of mixed bodies

In this section we suppose M is a mixed body. Automorphisms of holed mixed bodies can be understood by the same method. In fact, recall from the introduction that the “difference” between an automorphism of a holed or spotted manifold and an unspotted manifold is an automorphism of the spotted manifold which can be realized by an isotopy of the unspotted manifold M .

Recall from Theorem 2.4 that there exists a canonical closed surface $F \subseteq M$ separating M into a handlebody H and a compression body Q . The splitting $M = H \cup_F Q$ has the property that every curve in F bounding a disc in Q also bounds a disc in H . A curve in F bounding a disc in H may either bound a disc in Q or be parallel to $\partial_i Q = \partial M$. As in the special case treated in Section 6 we can use the canonicity of F to isotope an automorphism $f: M \rightarrow M$ to preserve the splitting, so it induces restricted automorphisms on Q and H . Here, again, F is not “rigid” (e.g., a double twist on a sphere S intersecting F on an essential curve is isotopic to the identity, see Example 2.7).

But, in this case, there will be another canonical surface $V \subseteq H$ which can also be made invariant. From another point of view, we shall prove that the restriction of f to H is rigidly reducible. More specifically, we shall prove that if f preserves F , then $f|_H$ always preserve a compression body $Q' \subseteq H$ with $\partial_e Q' = \partial H$ and $\partial_i Q' \neq \emptyset$ (see Section 1). Roughly speaking, Q' is the “mirror image” of Q through F and V is the mirror image of $\partial M = \partial_i Q$.

Proposition 7.1. *Let f be an automorphism of M . Then f can be isotoped to preserve the canonical splitting $M = H \cup_F Q$. Moreover, f can be further isotoped rel F to preserve a surface $V \subseteq H$ with the following properties. It separates H into a (possibly disconnected) handlebody H' and a compression body Q' such that a curve in F bounds a disc in Q if and only if it bounds a disc in Q' . Then f preserves Q , Q' and H' . Also, $f|_Q$ and $f|_{Q'}$ are conjugate, and f restricted to V (and hence H') is unique, in the sense that it depends only on the isotopy class of f , and not on $f|_H$.*

To prove the proposition above we consider a compression body $Q' \subseteq M$ with the following properties: 1) $Q' \subseteq H$, 2) $\partial_e Q' = \partial H = F$ and 3) a curve in F bounds a

disc in Q' if and only if it bounds a disc in Q . We can regard $Q \cup Q'$ as the double of Q along $\partial_e Q$.

Lemma 7.2. *The compression body Q' is well-defined up to isotopy. In particular, so is $V = \partial_i Q'$.*

Proof. Consider a complete system of essential discs \mathcal{E} in Q i.e., $\mathcal{C} = \partial\mathcal{E} \subseteq F$ and \mathcal{E} cuts Q open into a union of a product $(\partial_i Q) \times I$ and balls. Then \mathcal{C} also bounds discs $\mathcal{E}' \subseteq H$. Let $Q' \subseteq H$ be the compression body determined by F and \mathcal{E}' . More precisely, consider a neighborhood $N \subseteq H$ of $F \cup \mathcal{E}'$. Let $B \subseteq H$ be the union of ball components of $\overline{H - N}$. Then $Q' = N \cup B$.

We need to prove “symmetry”, i.e. that curves of F bound discs in Q' if and only if they also bound in Q . Here we require that the system \mathcal{E} is minimal, in the sense that the removal of any discs yields a system which is not complete. It is clear that any essential disc is contained in some such system. It is a standard fact that any pair of such minimal disc systems in a compression body are connected by a sequence of *disc slides* (see e.g. [2], appendix B). But a disc slide in any of the compression bodies Q, Q' corresponds to a disc slide in the other whose results coincide in F . We need not consider ∂ -parallel discs in the compression bodies, so the proof that Q and Q' are symmetric is complete.

The above argument with disc slides, together with irreducibility of H , also proves that the construction of Q' does not depend on \mathcal{E} , implying its uniqueness. \square

Let $f: M \rightarrow M$ be an automorphism of a mixed body. Denote by \mathcal{A}_f the class of automorphisms $f': M \rightarrow M$ which are isotopic to f and preserve $Q, Q', H' \subseteq M$.

Lemma 7.3. *If $f, f' \in \mathcal{A}_f$ then $f|_V$ and $f'|_V$ are isotopic (hence also $f|_{H'}$ and $f'|_{H'}$ are isotopic).*

Proof. Let $h = (f^{-1} \circ f') \in \mathcal{A}_{\text{Id}}$. Therefore $h|_{\partial M} = \text{Id}_{\partial M}$.

It is a simple and known fact that an automorphism of a compression body is uniquely determined by its restriction to the exterior boundary. Recall that F is the exterior boundary of both Q and Q' . Symmetry of Q and Q' implies that $h|_{Q \cup Q'}$ may be regarded as the double of $h|_Q$. Now $\partial_i Q = \partial M$, then $h|_{\partial_i Q} = \text{Id}$. Therefore $h|_{\partial_i Q'} = \text{Id}$, completing the proof. \square

Proof of Proposition 7.1. Use Lemma 7.2 and note the following. By cutting Q' along the cocores of its 1-handles it is easy to see that $V = \partial_i Q'$ separates H into Q' and a handlebody $H' = \overline{H - Q'}$. This proves the conclusions on the existence of the canonical decomposition. The conclusions on $f|_V$ and $f|_{H'}$ follow from Lemma 7.3. The conclusion on conjugacy of $g = f|_Q$ and $g' = f|_{Q'}$ is clear from the symmetry of Q and Q' . \square

Remark 7.4. We emphasize that while V is a rigid reducing surface for $f|_H$ it is not a rigid reducing surface for f , even with unique $f|_{H'}$ and $f|_V$. The reason is that f need not restrict uniquely to the complement $Q \cup Q'$.

Recall from the introduction that some of the automorphisms $f: M \rightarrow M$ being considered arise from adjusting automorphisms of general reducible 3-manifolds. In this case $f|_{\partial M}$ is periodic, permuting components. The argument above in Lemma 7.3 shows, then, that in this case f will always be periodic on the (possibly disconnected) handlebody $H' \subseteq H \subseteq M$. But we also consider cases without the periodic hypothesis on $f|_{\partial M}$. Here there is a constraint on $f|_{\partial M}$. To see this recall from the argument that $f|_{\partial M}$ determines $f|_V$ uniquely. But $V = \partial H'$, therefore $f|_V$ is a surface automorphism that extends over the handlebody H' , which is a constraint.

The goal is to find a nice representative for the class of an automorphism $f: M \rightarrow M$. Use Proposition 7.1 to decompose $f: M \rightarrow M$ into automorphisms $f|_Q$, $f|_{Q'}$ and $f|_{H'}$. The first two are conjugate and they will be dealt with using unlinkings analogous to those of Section 6. The last is dealt with in [16] (see also Sections 1 and 5).

We now proceed to find desirable restrictions of f to the canonical compression bodies Q , Q' . The two restrictions are conjugate so it suffices to deal with one of them. We choose Q' because $Q' \cup H'$ is a handlebody, what shall be of a technical relevance. So consider $g = f|_{Q'}$. We need to define a convenient notion of “back-tracking” for an automorphism of Q' . Here suppose that $g: Q' \rightarrow Q'$ is generic. Similarly to what is done in Section 6, consider concentric $Q_i \subseteq Q'$, disc systems $\mathcal{E}_i \subseteq Q_i$, laminations Λ and Ω , and growth rate λ as in Section 5. Also assume that Λ has the incompressibility property. Consider $\widehat{Q}_1 = Q_1|\mathcal{E}_1$ and let K_1 be a component. Recall that $(K_1, V_1, \mathcal{D}_1)$ denotes either a spotted ball or spotted product, with spots determined as duplicates of \mathcal{E}_1 .

The new phenomenon for back-tracking in Q' is roughly the following. If one follows a leaf of the mapped 1-dimensional lamination it may enter the product part of Q' through a spot D and also leave through D . Such a leaf may “back-track” in Q' , in the sense that it can be pulled tight (homotopically) in Q' to reduce intersections with that spot. Other (portions of) leaves may not admit pulling tight in this way, being somehow linked with $\partial_i Q'$. But among these some can still be pulled tight if they are allowed to pass through the handlebody H' (recall that $H = Q' \cup H'$).

Definition 7.5. Suppose that $\sigma \subseteq \Omega$ is an arc such that $\sigma' = \omega(\sigma) \subseteq Q_0 \subseteq Q'$ has the following properties: 1) $\partial\sigma'$ is contained in a single disc E_0 of the system $\mathcal{E}_0 \subseteq Q_0$, and 2) σ' is isotopic in H (rel $\partial\sigma'$) to an arc in E_0 . If there exists such an arc we say that Ω *back-tracks*. In this case we say that $g = f|_{Q'}$ *admits back-tracking in H* or just *admits back-tracking*.

Remark 7.6. Property 1) above is intrinsic to Q' while property 2) is not, since it refers to the entire handlebody H .

Lemma 7.7. *If Ω back-tracks then there exists an integer $i \leq 0$ such that Q_i has the following properties. There is a component K_i of $Q_i \cap K_1$ which is a twice spotted ball (in other words, K_i is contained in a 1-handle of Q_i), both contained in the same spot D_1 of \mathcal{D}_1 . Moreover if (P_i, \mathcal{D}_i) is the unique spinal pair for Q_i , then it is parallel in $K_1 \cup H'$ to the subspinal pair (D_1, D_1) for $(K_1, V_1, \mathcal{D}_1)$, where D_1 is a disc of \mathcal{D}_1 .*

We skip the proof, which follows directly from unknottedness and parallelism (see Section 5).

Remark 7.8. The converse is obviously true.

In the situation of the lemma above we say that Q_i back-tracks in Q_1 . More generally, we say that Q_i back-tracks in Q_{j+1} , $i < j$, if $g^{-j}(Q_i)$ back-tracks in $g^{-j}(Q_{j+1}) = Q_1$.

Definition 7.9. If $f \in \mathcal{A}_f$ and $g = f|_{Q'}$ is generic and admits no back-tracking in H then we say that g is *train-track generic in H* , or just *train-track generic*.

The importance of considering train-track generic automorphisms comes from the result below. Its proof is an adaptation of an argument of [4].

Theorem 7.10. *Let f be an automorphism of a mixed body M . If $g = f|_{Q'}$ is train-track generic in H then its growth rate is minimal among the restrictions to Q' of all automorphisms in \mathcal{A}_f .*

Proof. We prove the contrapositive, so assume that there exists $f' \in \mathcal{A}_f$ such that $g' = f'|_{Q'}$ has growth λ' less than the growth λ of g .

Consider the original disc system $\mathcal{E}_0 \subseteq Q_0$, determining a dual object Γ_0 . For simplicity denote it by Γ . Recall that $\Gamma = \hat{\Gamma} \cup V$, where $\hat{\Gamma}$ is a graph with one vertex on each component of V . Consider the mapped-in 1-dimensional lamination $\omega: \Omega \rightarrow Q_0$ (see Section 5, Theorem 5.9). Recall that ν denotes its transverse measure. Use (Ω, ν) to give weights $\hat{\nu}$ to the edges of Γ . More precisely, if $E_i \subseteq \mathcal{E}_0$ is the disc corresponding to an edge e_i of Γ then the weight on e_i is $\int_{(\omega^{-1}(E_i))} \nu$. For simplicity, call the pair $(\Gamma, \hat{\nu})$ a *weighted graph*, although it is really $\hat{\Gamma}$ which is a weighted graph.

Extend \mathcal{E}_0 to a disc system $\mathcal{E} \subseteq Q'$ through the product in $Q' - \text{int}(Q_0)$. Apply powers of g^{-1} to Γ . One can “push forward” the weights on Γ to $\Gamma_n = g^{-n}(\Gamma)$ through g^{-n} . More precisely, the weight on the image $g^{-n}(e)$ of an edge e is the same weight as that on e . Denote the resulting weighted graph as $(\Gamma_n, \hat{\nu})$. It follows that

$$\frac{\mathcal{E} \bullet (\Gamma_n, \hat{\nu})}{\lambda^n} = c > 0 \tag{1}$$

is constant, where $\mathcal{E} \bullet (\Gamma_n, \hat{\nu})$ denotes the sum of points of intersection $\mathcal{E} \cap \Gamma_n$ weighted by $\hat{\nu}$.

Now consider f' and $g' = f'|_{Q'}$, with corresponding Q'_0 , disc system $\mathcal{E}'_0 \subseteq Q'_0$ with dual graph Γ' and growth λ' . As before, the corresponding 1-dimensional lamination determines a weighted graph $(\Gamma', \hat{\nu}')$.

But Q_0, Q'_0 are concentric in Q' , therefore one can apply an ambient isotopy rel V taking Γ to some $\bar{\Gamma} \subseteq Q'_0$. We can also assume that $\bar{\Gamma}$ does not “back-track” in the 1-handles of Q'_0 , in the sense that no vertex of $\bar{\Gamma}$ is contained in such a 1-handle K and that the edges of $\bar{\Gamma}$ are transverse to the foliation of K by cocores. The isotopy from Γ to $\bar{\Gamma}$ pushes forward $\hat{\nu}$ to yield $(\bar{\Gamma}, \hat{\nu})$.

We are interested in applying powers of $(g')^{-1}$ to $\bar{\Gamma}$. Denote $\bar{\Gamma}_n = (g')^{-n}(\bar{\Gamma})$. By the construction of $\bar{\Gamma} \subseteq Q'_0$,

$$\frac{\mathcal{E} \cdot (\bar{\Gamma}_n, \hat{\nu})}{(\lambda')^n} < \bar{C}$$

is bounded. This is true because the growth rate of intersections with any essential disc with the images under $(g')^{-n}$ of *any* weighted graph is λ' . Since $\lambda' < \lambda$,

$$\frac{\mathcal{E} \cdot (\bar{\Gamma}_n, \hat{\nu})}{(\lambda)^n} \rightarrow 0. \quad (2)$$

By combining (1) and (2) it follows that there exists N such that

$$\frac{\mathcal{E} \cdot (\bar{\Gamma}_N, \hat{\nu})}{\lambda^N} < \frac{\mathcal{E} \cdot (\Gamma_N, \hat{\nu})}{\lambda^N} \quad (3)$$

But recall that there is an isotopy rel V taking Γ to $\bar{\Gamma}$. There is also an isotopy (not necessarily rel V) taking $g^{-N}(\bar{\Gamma})$ to $\bar{\Gamma}_N$ (f and f' , which restrict to Q' as g and g' , are isotopic). Therefore there is an isotopy ι taking Γ_N to $\bar{\Gamma}_N$. Note that this is an isotopy in M , not just Q' .

Regard Q_{-N} as a neighborhood of Γ_N . Extend the isotopy ι to Q_{-N} , taking it to a neighborhood \bar{Q}_N of $\bar{\Gamma}_N$. Define $\bar{\omega} = \iota \circ \omega: \Omega \rightarrow Q_{-N}$. The inequality (3) is restated in terms of laminations as

$$\int_{(\bar{\omega})^{-1}(\mathcal{E})} \nu < \int_{\omega^{-1}(\mathcal{E})} \nu.$$

Now let $\mathcal{S} \subseteq M$ be a “dual sphere system” corresponding to \mathcal{E} , i.e. $\mathcal{S} \cap Q' = \mathcal{E}$. The inequality above clearly implies

$$\int_{(\bar{\omega})^{-1}(\mathcal{S})} \nu < \int_{\omega^{-1}(\mathcal{S})} \nu. \quad (4)$$

The isotopy ι determines a map $p: \Omega \times I \rightarrow M$ such that $p|_{\Omega \times \{0\}} = \omega$ and $p|_{\Omega \times \{1\}} = \bar{\omega}$. Perturb it to put it in general position. Consider $p^{-1}(\mathcal{S})$. Let $L \subseteq \Omega$ be a leaf that embeds through ω . Then $(L \times I) \subseteq (\Omega \times I)$ is a product $\mathbb{R} \times I$ and $(L \times I) \cap p^{-1}(\mathcal{S})$ consists of embedded closed curves and arcs. Ignore closed curves

for the moment. Let A_L be the set of such arcs α intersecting the lower boundary $L \times \{0\}$. Recall that the set of leaves that do not embed has measure zero. If for all leaves L that embed and all $\alpha \in A_L$ the other endpoint of α is in $L \times \{1\}$ then

$$\int_{(\bar{\omega})^{-1}(S)} \nu \geq \int_{\omega^{-1}(S)} \nu,$$

contradicting inequality (4). Therefore there exists L that embeds and $\alpha \in A_L$ with both endpoints at $L \times \{0\}$.

Let α be an edgemost such arc. Then there is an arc $\beta \subseteq (L \times \{0\})$ such that $\partial\beta = \partial\alpha$. Let D be the half-disc bounded by $\alpha \cup \beta$ in $L \times I$. Then $p|_D$ is a map which sends α to a sphere S , $S \subseteq \mathcal{S}$, and embeds β in the leaf L . We can homotope the map $p|_D$ rel β to obtain a map $d : D \rightarrow M$ such that $d(\alpha) \subseteq E_0 \subseteq S$, where $E_0 \subseteq \mathcal{E}_0$ is the disc in S . Thus one has $d(\alpha \cup \beta) \subseteq Q_0$.

Recall the Heegaard surface F separates M into Q and H . Let d be transverse to F and let $\Delta = d^{-1}(Q) \subseteq D$. On Δ redefine d to be its mirror image $r \circ d$, where r is the reflection map $r : Q \rightarrow Q'$. Then homotope d by pushing slightly into Q' . By post-composing d with an isotopy along the I -fibers of the product $Q' - \text{int}(Q_{-1})$ one obtains d mapping into $Q_0 \cup H'$.

Using the irreducibility of $Q_0 \cup H'$ (which is a handlebody) we remove closed curves from $d^{-1}(\mathcal{E}_0)$, which will then consist of α . By unknottedness (Theorem 5.12), the embedded image of $d|\beta \subseteq \Omega$, which lies in a leaf, is parallel to the subspinal pair (D_0, D_0) for some spotted component $(K_0, V_0, \mathcal{D}_0)$, where D_0 is a duplicate of E_0 , proving that Ω back-tracks. \square

Theorem 7.11 (Compression Body Train Track Theorem). *Let $f : M \rightarrow M$ be an automorphism of a mixed body. Then there exists $f' \in \mathcal{A}_f$ such that $g = f'|_{Q'}$ is*

- 1) *periodic,*
 - 2) *rigidly reducible or*
 - 3) *train-track generic in H .*
- On the other pieces, Q and H' , we have:*
- 4) *$f'|_Q$ is isotopic to a conjugate of g , and*
 - 5) *the class of $f'|_{H'}$ does not depend on $f' \in \mathcal{A}_f$.*

The conclusions concerning $f'|_Q, f'|_{H'}$ are proved in Proposition 7.1. We proceed now to prove that one can always find $f' \in \mathcal{A}_f$ such that $f'|_{Q'}$ is either periodic, reducible or train-track generic, i.e. admits no back-tracking. To do so we need to perform unlinkings analogous to those of Section 6. As before, assume that $g = f'|_{Q'}$ is generic and that the leaves of the 2-dimensional lamination are in Morse position with respect to the height function determined by the product structure in $Q' - \text{int}(Q_0)$. We refer the reader to Section 5 and [16] for details.

For $t \leq 1$ consider spotted balls or products (K_t, \mathcal{D}_t) determined by $Q_t \cap \widehat{Q}_1$.

As before, to define linking we need half-discs. Let (Δ, α, β) be a half-disc and (R, γ, γ') a rectangle (Definitions 6.2).

Definitions 7.12. Let $s < t$. A *half-disc for Q_s in Q_t* is a half-disc (Δ, α, β) satisfying:

- (i) $\Delta \subseteq K_t \cup H'$ is embedded for some K_t ,
- (ii) β is contained in a spot D_t of K_t ,
- (iii) $\alpha \subseteq \partial K_s$ for some $K_s \subseteq K_t$ and $\Delta \cap K_s = \alpha$,
- (iv) α connects two distinct spots of (K_s, \mathcal{D}_s) .

As before we refer to α as the *upper boundary* of Δ , denoted by $\partial_u \Delta$, and β as its *lower boundary*, denoted by $\partial_l \Delta$.

A *rectangle for Q_s in Q_t* is a rectangle $(R_s, \gamma_s, \gamma_t)$ satisfying:

- (a) $R_s \subseteq K_t \cup H'$ is embedded for some K_t ,
- (b) $\gamma_s \subseteq \partial K_s$ for some $K_s \subseteq K_t$ and $R_s \cap K_s = \gamma_s$,
- (c) γ_s connects distinct spots of (K_s, \mathcal{D}_s) ,
- (d) $\gamma_t \subseteq \partial K_t - \text{int}(\mathcal{D}_{t_{i+1}}) - \text{int}(V_t)$ (recall that when K_t is a spotted ball then $V_t = \emptyset$),
- (e) $\partial R_s - (\gamma_s \cup \gamma_t) \subseteq \mathcal{D}_t$.

Note that a half-disc Δ in Q_t has the property that $\partial \Delta \cap \partial_i Q' = \emptyset$. Similarly $\partial R \cap \partial_i Q' = \emptyset$, where R is a rectangle in Q_t .

Also, holes now may appear due not just to intersections with Q_s but also with H' .

Definition 7.13. Let $0 \leq s < t \leq 1$ and Δ a half-disc for Q_s in Q_t . Suppose that $s = t_0 < t_1 < \dots < t_k = t$ is a sequence of regular values. We say that Δ is *standard with respect to $\{t_i\}$* if there is $0 \leq m < k$ such that the following holds. The half-disc Δ is the union of rectangles R_{t_i} , $s \leq t_i \leq t_{m-1}$, and a half-disc Δ_{t_m} (see Figure 4) with the properties that:

1. each R_{t_i} , $s \leq t_i < t_m$, is a rectangle for Q_{t_i} in $Q_{t_{i+1}}$,
2. the holes $\text{int}(R_{t_i}) \cap Q_{t_i}$ are essential discs in $Q_{t_m} \cup H'$, each consisting of either an essential disc in Q_{t_m} or of a vertical annulus in the product part of Q_{t_m} union an essential disc in H' ,
3. Δ_{t_m} is an unholed half-disc for K_{t_m} in $H_{t_{m+1}}$.

The following is the analogue of Proposition 6.4 needed here.

Proposition 7.14. *Let $0 = t_0 < \dots < t_k = 1$ be a complete sequence of regular values. Suppose that a component K_0 of $Q_0 \cap \widehat{Q}_1$ intersects a disc D of $\widehat{\mathcal{E}}_1$ in two distinct components D^0, D^1 . If an arc $\sigma \subseteq K_0$ having an endpoint at each D^i is parallel (rel $\partial\sigma$) in $\widehat{Q}_1 \cup H'$ to an arc $\sigma' \subseteq D$ then there exists a half-disc Δ for Q_0 connecting D^0 and D^1 which is standard with respect to $\{t_i\}$.*

Proof. The proof is similar to that of Proposition 6.4, using Theorem 5.12, Lemma 5.16 and techniques of Lemma 6.5.

If K_1 is a spotted ball then all K_{t_i} 's containing σ also are spotted balls and the result follows directly from the proof of Lemma 6.5. In fact, suppose that t_m is the biggest t_i with the property that D^0 and D^1 are contained in distinct spots $D_{t_i}^0, D_{t_i}^1$ of K_{t_i} . If this K_{t_m} is a spotted ball then $K_{t_{m+1}}$ also is a spotted ball, as one can verify by considering the possible events described in Lemma 5.15 changing K_{t_m} to $K_{t_{m+1}}$. The argument of Lemma 6.5 still works.

We can therefore assume that K_{t_m} is a spotted product. In this case $K_{t_{m+1}}$ is also a spotted product. Let D' be the spot of $K_{t_{m+1}}$ that contains both $D_{t_m}^0, D_{t_m}^1$. There is a single arc $\gamma_{t_{m+1}} \subseteq D' - \text{int}(D_{t_m}^0 \cup D_{t_m}^1)$ connecting $D_{t_m}^0$ to $D_{t_m}^1$ (up to isotopy in $D' - \text{int}(D_{t_m}^0 \cup D_{t_m}^1)$, rel $\partial(D_{t_m}^0 \cup D_{t_m}^1)$). Recall that there are unambiguous spinal pairs $(P_{t_m}, \mathcal{D}_{t_m}), (P_{t_{m+1}}, \mathcal{D}_{t_{m+1}})$ for $K_{t_m}, K_{t_{m+1}}$ respectively, and that they are parallel. We use this parallelism to build $\gamma_{t_m} \subseteq P_{t_m}$ isotopic to $\gamma_{t_{m+1}}$. Note that the parallelism determines an unholed half-disc $(\Delta_{t_m}, \gamma_{t_m}, \gamma_{t_{m+1}})$ using the Loop Theorem.

Now proceed inductively “downwards”, building $\gamma_{t_{i-1}}$ from γ_{t_i} in the analogous manner for as long as $K_{t_{i-1}}$ is a spotted product. More precisely, to each spotted product K_{t_i} corresponds a spinal pair $(P_{t_i}, \mathcal{D}_{t_i})$. The parallelism of spinal pairs $(P_{t_{i-1}}, \mathcal{D}_{t_{i-1}})$ and $(P_{t_i}, \mathcal{D}_{t_i})$ yields $\gamma_{t_{i-1}}$ from γ_{t_i} . Note that this parallelism defines a rectangle $(R_{t_{i-1}}, \gamma_{t_{i-1}}, \gamma_{t_i})$, which is also holed only if there is a spotted ball $K'_{t_{i-1}} \subseteq K_{t_i}$ (i.e. in a situation analogous to Case 2 in Figure 3). In any case, $\gamma_{t_{i-1}}$ has the property of connecting the spots $D_{t_{i-1}}^0$ and $D_{t_{i-1}}^1$ that contain D^0, D^1 .

If no K_{t_j} containing σ is a spotted ball then K_0 is a spotted product and we obtain an arc $\gamma_0 \subseteq P_0$. But γ_0 clearly connects the spots D^0 and D^1 of K_0 . Hence the union of all rectangles R_{t_i} 's and the half-disc Δ_{t_m} yields the desired standard half-disc. This takes care of the case when there is no spotted ball K_{t_j} containing σ .

Therefore assume that σ is contained in a spotted ball K_{t_k} . Assume that the spotted ball K_{t_k} is maximal, in the sense that it is contained in a spotted product $K_{t_{k+1}}$. We also assume from the construction above that there is an arc $\gamma_{t_{k+1}} \subseteq P_{t_{k+1}}$ isotopic (rel D') to $\gamma_{t_{m+1}}$. The original arc σ determines an embedded arc $\sigma_{t_k} \subseteq L_{t_k}$. Now let H'' be the component of H' which intersects $K_{t_{k+1}}$. Then $H_{t_{k+1}} = K_{t_{k+1}} \cup H''$ is a (spotted) handlebody. We first note that $\gamma_{t_{k+1}}$ is isotopic (rel D') in $H_{t_{k+1}}$ to σ . Therefore σ_{t_k} is isotopic in $H_{t_{k+1}}$ to $\gamma_{t_{k+1}}$. But since σ_{t_k} is a subgraph of Γ_{t_k} , we can apply Lemma 5.14 to obtain a subspinal pair $(\check{P}_{t_k}, \check{\mathcal{D}}_{t_k})$ (a disk with two spots) which is parallel in $H_{t_{k+1}}$ to $(P_{t_{k+1}}, \mathcal{D}_{t_{k+1}})$.

The parallelism described above determines $\gamma_{t_k} \subseteq P_{t_k}$. It also determines a rectangle $(R_{t_k}, \gamma_{t_k}, \gamma_{t_{k+1}})$ which may be holed. In this case the holes come from intersections with H'_{t_k} , which is the spotted handlebody corresponding to the spotted product component $K'_{t_k} \subseteq K_{t_{k+1}}$. As usual, these holes may be adjusted to be essential discs either in Q_{t_k} or in H'_{t_k} , each of these last consisting of a disc in H' and a vertical annulus in K'_{t_k} .

Now enlarge $(\check{P}_{t_k}, \check{D}_{t_k})$ to a spinal pair for K_{t_k} . We use Lemma 5.16 to obtain spinal pairs (P_{t_i}, D_{t_i}) for each $t_i < t_k$. We now proceed as in Lemma 6.5. Recalling, consider a path $\gamma_{t_0} = \gamma_0 \subseteq P_0$ connecting D^0 and D^1 . Now use parallelism to go “upwards”, obtaining $\gamma_{t_i} \subseteq P_{t_i}$ for any $t_i < t_k$. In the same way $\gamma_{t_{k-1}}$ determines through parallelism an arc γ'_{t_k} in P_{t_k} . Now $\gamma_{t_k}, \gamma'_{t_k} \subseteq P_{t_k}$ both connect the same spots of K_{t_k} therefore they are isotopic (rel ∂D_{t_k}) in P_{t_k} . We adjust them to coincide, obtaining $\gamma'_{t_k} = \gamma_{t_k}$.

To finalize, let Δ be the union of the rectangles R_{t_i} 's and the half-disc Δ_{t_m} , which is the desired standard half-disc in this last case. \square

Remark 7.15. Note from the argument that, as in the sphere body case, a rectangle R_{t_i} is holed only when $K_{t_i} \subseteq K_{t_{i+1}}$ is in a situation like Case 2 of Figure 3.

The next step is to describe the unlinking move in this setting of mixed bodies. Suppose that there is a half-disc Δ for Q_0 in Q_1 , which we assume to be in standard position with respect to a complete sequence of regular values $\{t_i\}$. The goal is to remove holes of the rectangles R_{t_i} and the half-disc Δ_{t_m} . Let $s < t$ be consecutive terms of the sequence and consider $R_s \subseteq K_t$. If K_t is a spotted ball then the operation works precisely as in Section 6.

The new situation appears when K_t is a spotted product, and even there the argument is essentially the same, so we will be brief. Recall from Section 6 that the main condition for the unlinking to be possible was that one side of the holed rectangle R_{t_i} needed to intersect a spot of $K_{t_{i+1}}$ in a spot containing a single spot D of K_{t_i} . If that is the case, as before, we consider the sphere S in the mixed body M which is dual to D . The holes of R_{t_i} are removed by “isotoping them” along R_{t_i} close to S and then along $S - \text{int}(D)$. Here the only new feature is that a hole may correspond to a disc $E \subseteq Q_{t_i} \cup H'$.

Let N be a neighborhood (in $Q_{t_i} \cup H'$) of a component E of $R_{t_i} \cap (Q_{t_i} \cup H')$. One can consider it as a neighborhood of an arc α either in H' (in case E is a disc intersecting H') or in Q_{t_i} (otherwise). Isotope α (rel $\partial\alpha$) along R_{t_i} bring it close to D . Now perform the isotopy along $S - \text{int}(D)$ as before. Note that this isotopy does not introduce intersections of α with R_{t_i} , for S intersects R_{t_i} only at its boundary. Call the composition of these isotopies ι and let $\alpha' = \iota(\alpha) \subseteq K_{t_i}$. Extend ι to N and let $N' = \iota(N) \subseteq K_{t_i}$, a neighborhood of α' .

We separate two cases. If $\alpha \subseteq Q_{t_i}$ the unlinking move is done, as in Section 6. If $\alpha \subseteq H'$ note that α and α' are isotopic (rel $\partial\alpha$) as arcs in $K_{t_{i+1}} \cup H'$. Consider an ambient isotopy κ of $K_{t_{i+1}} \cup H'$ such that $\kappa(\alpha') = \alpha$ and fixing $\partial K_{t_{i+1}} \cup H'$. Adjust κ

so that $\kappa \circ \iota(N) = N$. Now replace K_{t_i} by $\kappa(K_{t_i})$, so H' is preserved. This completes the description of the operation, which we call an *unlinking along R_{t_i} and S* . Clearly such an unlinking reduces the number of holes in R_{t_i} .

Proposition 7.16. *Let $0 = t_0 < t_1 < \dots < t_{k-1} < t_k = 1$ be a complete sequence of regular values. Suppose that Δ is a half-disc for $Q_0 \subseteq Q_1$, standard with respect to $\{t_i\}$. Then there exists a sequence of unlinkings taking Q_0 to $Q'_0 \subseteq Q_1$ for which Δ is an unholed half-disc. Moreover $Q'_0 \cap \mathcal{E}_1 = Q_0 \cap \mathcal{E}_1$.*

Sketch of proof. The proof is essentially the same as that of Proposition 6.8. We just note that also here, as Section 6, a rectangle R_{t_i} will be holed only if $K_{t_i} \subseteq K_{t_{i+1}}$ is in a situation like Case 2 of Figure 3 (see Remark 7.15 above). Therefore one can always find a spot of $K_{t_{i+1}}$ containing a single spot of K_{t_i} and containing a side of R_{t_i} . A finite sequence of unlinkings in $K_{t_{i+1}}$ yields an unholed rectangle R_{t_i} . As before, the final unholed half-disc Δ is obtained by induction on i . The other conclusion of the theorem clearly also holds here. \square

The last ingredient needed for the proof of Theorem 7.11 is concentricity. The following is the analogue of Lemma 6.10 needed in this setting.

Lemma 7.17. *In the situation described in Proposition 7.16 we have that $Q'_0 \subseteq Q_1$ are concentric.*

Sketch of proof. As in Section 6 the rectangles R_{t_i} in the original half-disc Δ will be holed only in a situation like Case 2 of Figure 3 (see Remark 7.15). Therefore, as in Lemma 6.10, an unlinking does not interfere with the pinching half-discs. The argument is completed as before. \square

Proof of Theorem 7.11. As usual, the proof is analogous to that of Theorem 1.5 (which was done as the proof of Theorem 6.1).

Consider $f \in \mathcal{A}_f$. If $g = f|_{Q'}$ is reducible or periodic, the proof is over, so assume that g is generic, with growth rate λ . If it is not train-track generic there is an integer $m < 1$ such that Q_m back-tracks in Q_1 (Lemma 7.7). Let $K_m \subseteq K_1$ be the spotted ball in a K_1 (spotted ball or product) as in the lemma and assume that m is the greatest such value. Let D_m^0, D_m^1 be the two spots of K_m as in the lemma, contained in a single spot D of K_1 . Consider the increasing sequence of consecutive integers $m = i_0, i_1, \dots, i_{1-m} = 1$. Then there are two spots D^0, D^1 of $K_{i_{(2-m)}} = K_0$ containing D_m^0, D_m^1 respectively (hence $D^0, D^1 \subseteq D$). The parallelism in the conclusion of Lemma 7.7 gives an arc $\sigma \subseteq \partial K_0$ as in the statement of Proposition 7.14, yielding a half-disc Δ for Q_0 in Q_1 connecting D^0 and D^1 .

Choose a complete sequence $0 = t_0 < \dots < t_k = 1$ of regular values and assume that Δ is in standard position with respect to $\{t_i\}$. Apply Proposition 7.16 to obtain Δ unholed. Lemma 7.17 assures that an unlinking preserves Q', H' . This operation does not change the growth rate λ of g .

Skipping details, one proceeds as in [16] and Theorem 1.5, using Δ to realize down-moves or diversions, respective analogues of folds and pulling tights, as in [1] (note that “splittings” may be needed as a preparation for these moves). After performing a down-move the resulting automorphism g' may be either periodic, reducible or generic. In the first two cases the proof is done, so assume it is generic. The operation increases the greatest value m in the beginning of the proof, therefore repetition of this process has to stop eventually (when $m = 0$).

After performing a diversion, which is the case when $m = 0$, the resulting g' may also be periodic, reducible or generic. Again assuming it is generic its growth λ' is strictly smaller than λ , therefore repetition of this process also has to stop.

The procedure above, which can be carried on for as long as g' is generic admitting back-tracking, yields g' either periodic, reducible or generic without back-tracking in H , completing the proof for g' .

We recall that the conclusions on $f'|_Q, f'|_{H'}$ follow from Proposition 7.1. \square

References

- [1] M. Bestvina and M. Handel. Train tracks and automorphisms of free groups. *Ann. Math.*, 135:1–51, 1992.
- [2] F. Bonahon. Cobordism of automorphisms of surfaces. *Ann. Sci. École Norm. Sup. (4)*, 16(2):237–270, 1983.
- [3] L. Navarro de Carvalho. Generic automorphisms of handlebodies. Dissertation, Rutgers University, 2003.
- [4] L. Navarro de Carvalho. Tightness and efficiency of irreducible automorphisms of handlebodies. preprint, available at arXiv:math.GT/0408353, 2004.
- [5] E. César de Sá. Automorphisms of 3-manifolds and representations of 4-manifolds. Dissertation, Warwick University, 1977.
- [6] Albert Fathi and François Laudenbach. Difféomorphismes pseudo-Anosov et décomposition de Heegaard. *C. R. Acad. Sci. Paris Sér. A-B*, 291(6):A423–A425, 1980.
- [7] H. B. Griffiths. Automorphisms of a 3-dimensional handlebody. *Abh. Math. Sem. Univ. Hamburg*, 26:191–210, 1963/1964.
- [8] Allen Hatcher. Homological stability for automorphism groups of free groups. *Comment. Math. Helv.*, 70(1):39–62, 1995.

- [9] W. Jaco and P. Shalen. Seifert fibered spaces in 3-manifolds. *Mem. Amer. Math. Soc.*, 220:1–192, 1979.
- [10] B. Jiang, S. Wang, and Y. Wu. Homeomorphisms of 3-manifolds and the realization of Nielsen number. *Comm. Anal. Geom.*, 9(4):825–877, 2001.
- [11] K. Johannson. *Homotopy equivalences of 3-manifolds with boundaries*. Number 761 in LNM. Springer Verlag, 1979.
- [12] F. Laudenbach. Sur les 2-sphères d’une variété de dimension 3. *Ann. of Math.*, 97(2):57–81, 1973.
- [13] François Laudenbach. *Topologie de la dimension trois: homotopie et isotopie*. Société Mathématique de France, Paris, 1974. With an English summary and table of contents, Astérisque, No. 12.
- [14] D. McCullough. Isometries of elliptic 3-manifolds. available ArXiv math.GT/0010077.
- [15] D. McCullough. Mappings of reducible 3-manifolds. *Geometric and Algebraic Topology, Banach Center Publications*, 18:61–76, 1986.
- [16] U. Oertel. Automorphisms of 3-dimensional handlebodies. *Topology*, 41:363–410, 2002.
- [17] P. Scott. The geometries of 3-manifolds. *Bull. London Math. Soc.*, 15:401–487, 1983.
- [18] W. Thurston. The geometry and topology of 3-manifolds. notes, 1978.
- [19] Friedhelm Waldhausen. Heegaard-Zerlegungen der 3-Sphäre. *Topology*, 7:195–203, 1968.