Transverse Heegaard Splittings

Hyam Rubinstein & Martin Scharlemann

1. Background and the Theorem

In [RS] we use Cerf theory to compare irreducible Heegaard splittings of the same irreducible non-Haken orientable 3-manifold. A critical part of the argument is the observation that any two Heegaard surfaces may be isotoped so that they intersect in a nonempty collection of simple closed curves, each of which is essential in both surfaces. Here we describe an analog to this theorem that applies to the Haken case. An eventual goal, not yet realized here, is a bound for the number of stabilizations needed to make two distinct Heegaard splittings equivalent. Such a bound is found, for the non-Haken case, in [RS].

All manifolds are assumed to be compact and orientable. It is a simple and standard exercise to show that, if S and T are closed incompressible surfaces in an irreducible 3-manifold M, then they can be isotoped so their intersection (if any) is a collection of simple closed curves, each of which is essential in both S and T. A Heegaard surface in M is as unlike an incompressible surface as possible. It is a surface that is not only compressible, but a surface that can be compressed entirely away—on both sides. Yet it is shown in [RS] that, if M is closed and non-Haken, then a pair of Heegaard surfaces behave something like a pair of incompressible surfaces: Any pair of Heegaard surfaces P and Q can be isotoped so that they intersect in a nonempty collection of simple closed curves, each of which is essential in both P and Q. The content here is in the word "nonempty", since it is obvious that P and Q can be made disjoint: Choose disjoint spines of handlebodies bounded by P and Q, then isotope P and Q near the respective spines. The purpose here is to extend this result, in a somewhat different form, to the case in which M may be Haken.

A *compression body* H is constructed by adding 2-handles to a (surface) $\times I$ along a collection of disjoint simple closed curves on (surface) $\times \{0\}$, and capping off any resulting 2-sphere boundary components with 3-balls. The component (surface) $\times \{1\}$ of ∂H is denoted $\partial_+ H$, and the surface $\partial H - \partial_+ H$ is denoted $\partial_- H$. If $\partial_- H = \emptyset$ then H is a *handlebody*. If $H = \partial_+ H \times I$ then H is called a *trivial* compression body. Define the *index* $I(H)$ of H to be $\chi(\partial_- H) - \chi(\partial_+ H)$.

Received December 20, 1995. Revision received July 16, 1996.
The first author was partially supported by a grant from the Australian Research Council. The second author was partially supported by NSF grant DMS 9203522.
Michigan Math. J. 44 (1997).

(Here $\chi(\emptyset) = 0$.) Note that $I(H)$ is nonnegative, and is trivial if and only if H is a trivial compression body or a solid torus.

The cores of the 2-handles defining H, extended vertically down through $(\partial_+ H) \times I$, are called a *defining set* of 2-disks for H. There is a dual picture: A *spine* for H is a properly imbedded 1-complex Σ such that H deformation retracts to $\Sigma \cup \partial_- H$. Such a spine can be constructed from a set of defining disks: The arc co-cores of the 2-handles, with the arc ends that lie on 2-spheres coned to the centers of the 3-balls and the other ends extended down to $\partial_- H$, are the edges of a spine. The retraction restricts to a map $\partial_+ H \to \Sigma \cup \partial_- H$ whose mapping cylinder is itself homeomorphic to H.

A *Heegaard splitting* $M = A \cup_P B$ of a 3-manifold consists of an orientable surface P in M, together with two compression bodies A and B into which P divides M; P itself is called the splitting surface. The genus of $A \cup_P B$ is defined to be the genus of P. The index $I(A \cup_P B)$ is defined to be $I(A) + I(B) = \chi(\partial M) - 2\chi(P)$. A Heegaard splitting of M can also be viewed as a handle structure on M in which the 1-handles are the co-cores of the defining 2-handles of A and the 2-handles are the defining 2-handles of B. A *stabilization* of the Heegaard splitting $A \cup_P B$ is the Heegaard splitting obtained by adding to A a regular neighborhood of a proper arc in B that is parallel in B to an arc in P. A stabilization has genus one larger than and, up to isotopy, is independent of the choice of arc in B, and is the same if the construction is done symmetrically to an arc in A instead.

Recall the following: If there are meridian disks D_A and D_B in A and B respectively so that ∂D_A and ∂D_B intersect in a single point in P, then $A \cup_P B$ can be obtained by stabilizing a lower-genus Heegaard splitting. We then say that $A \cup_P B$ is *stabilized.* If there are meridian disks D_A and D_B in A and B respectively so that ∂D_A and ∂D_B are disjoint in P, then $A \cup_P B$ is *weakly reducible*. If there are meridian disks so that $\partial D_A = \partial D_B$, then $A \cup_P B$ is *reducible.* It is easy to see that reducible splittings are weakly reducible and that (except for the genus-1 splitting of S^3) any stabilized splitting is reducible. It is a theorem of Casson and Gordon [CG] that if $A \cup_P B$ is a weakly reducible splitting then either M contains an incompressible surface or $A \cup_P B$ is reducible. It is a theorem of Haken [H] that any Heegaard splitting of a reducible 3-manifold is reducible, and it follows from a theorem of Waldhausen [W] that a reducible splitting of an irreducible manifold is stabilized.

A central point of [ST1] is that any irreducible Heegaard splitting $M = A \cup_P B$ can be broken up into a series of strongly irreducible splittings (see Figure 1). That is, we can begin with the handle structure determined by $A \cup_P B$ and rearrange the order of the 1- and 2-handles, so that ultimately

$$M = M_0 \cup_{F_1} M_1 \cup_{F_2} \cdots \cup_{F_m} M_m.$$

The 1- and 2-handles which occur in M_i provide it with a strongly irreducible splitting $A_i \cup_{P_i} B_i$, with $\partial_- A_i = F_i$ and $\partial_- B_{i-1} = F_i$ for $1 \le i \le m$; $\partial_- A_0 = \partial_- M$; and $\partial_- B_m = \partial_+ M$. Each component of each F_i is a closed incompressible surface of positive genus and, for any i, only one component of M_i (the *active* component) is not a product. The compression bodies $A_{\bowtie}$, $B_{\bowtie}$ and the splitting

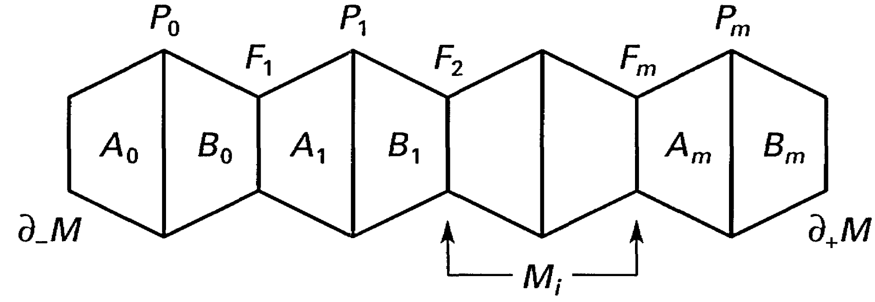

Figure 1

surface $P_{\bowtie}$ in this component are called the active components of A_i, B_i and P_i, respectively. None of the compression bodies A_i, B_{i-1} $(1 \le i \le m)$ is trivial. If $\partial_- A$ or $\partial_- B$ is compressible in M (so in particular M is ∂-reducible), then respectively A_0 or B_m is trivial (i.e., just a product). Such a rearrangement of handles will be called an *untelescoping* of the Heegaard splitting. It is easy to see that

$$I(A \cup_P B) = \sum_{i=0}^{m} I(A_i \cup_{P_i} B_i).$$

After the untelescoping we are again able to exploit strong irreducibility, so the version of [RS, 6.2] that remains true is one expressed in terms of untelescopings of the splittings.

Suppose the irreducible 3-manifold M has two splittings $A \cup_P B$ and $X \cup_Q Y$, as above, which have the respective untelescopings

$$M_0 \cup_{F_1} M_1 \cup_{F_2} \cdots \cup_{F_m} M_m$$

and

$$N_0 \cup_{G_1} N_1 \cup_{G_2} \cdots \cup_{G_n} N_n.$$

Here each M_i has the strongly irreducible splitting $A_i \cup_{P_i} B_i$ as described above, and each N_j has a similar strongly irreducible splitting $X_j \cup_{Q_j} Y_j$, with $\partial_- X_j = G_j$ and $\partial_- Y_{j-1} = G_j$ for $1 \le i \le n$. We will *not* assume that the partitioning of ∂M into $\partial_\pm M$ is the same in both splittings. Indeed, the correct viewpoint is that the splittings are of distinct manifolds M and N which happen to be diffeomorphic. Consistent with that viewpoint, we let $\partial_- N = \partial_- X_0 = \partial_- X$ and $\partial_+ N = \partial_- Y_n = \partial_- Y$. It will be convenient at times to let $F_0 = \partial_- M$, $F_{m+1} = \partial_+ M$, $G_0 = \partial_- N$, and $G_{n+1} = \partial_+ N$. Finally, define $P' = \bigcup_{0 \le i \le m} \{P_i\}$, $F = \bigcup_{1 \le i \le m} \{F_i\}$, $P^+ = P' \cup F$, $Q' = \bigcup_{0 \le i \le n} \{Q_i\}$, $G = \bigcup_{1 \le j \le n} \{G_j\}$, and $Q^+ = Q' \cup G$.

THEOREM 1.1. *Suppose that $A \cup_P B$ and $X \cup_Q Y$ are two irreducible Heegaard splittings of the same irreducible compact orientable 3-manifold M, and that P^+ and Q^+ are surfaces (described above) coming from untelescopings of $A \cup_P B$ and $X \cup_Q Y$, respectively. Then P^+ and Q^+ can be properly isotoped so that they are in general position and each curve of intersection is essential in both surfaces. Each P_i $(0 \le i \le m)$ and each F_i $(1 \le i \le m)$ intersects Q^+ nontrivially, and each Q_j $(0 \le j \le n)$ and each G_j $(1 \le j \le n)$ intersects P^+ nontrivially.*

2. Sweepouts and Their Structure

We begin the proof of Theorem 1.1 with some general considerations. Since F and G are incompressible, we can assume at the outset that F and G are in general position and that each component of intersection is essential in both surfaces. If $I(A \cup_P B) + I(X \cup_Q Y) = 0$ then they are both product splittings of $M = \partial_- M \times I$. Then $m = n = 0$ and the splitting surfaces P_0 and Q_0 are isotopic. Then, after a small isotopy, they can be made to intersect generically in a nontrivial collection of essential simple closed curves. The proof then proceeds by induction on $I(A \cup_P B) + I(X \cup_Q Y)$, but really induction is only needed to show part of the following simplifying lemma.

LEMMA 2.1.

(1) *No component of any P_i or Q_j is a torus.*
(2) *No F_i, $1 \leq i \leq m$ (resp. G_j, $1 \leq j \leq n$) has all its components isotopic in M to components of G (resp. components of F).*
(3) *No F_i, $1 \leq i \leq m$ (resp. G_j, $1 \leq j \leq n$) can be isotoped to be disjoint from Q^+ (resp. P^+).*

Proof. If any component were a torus, then M would be either (torus)$\times I$, $S^1 \times D^2$, or a Lens space ([ST1, Rule 7]); for each of these, any two irreducible splittings are isotopic [BO].

If all components of some F_i were isotopic in M to components of G, then consider the two manifolds $M^1 = \bigcup_{k=0}^{i-1} M_k$ and $M^2 = \bigcup_{k=i}^{m} M_k$. The untelescoped splittings of M and N restrict to untelescoped splittings of M^1 and M^2. These can be telescoped to give splittings $(A \cup_P B)^i$ and $(X \cup_Q Y)^i$ of M^i, with $I((A \cup_P B)^1) + I((A \cup_P B)^2) = I(A \cup_P B)$ and $I((X \cup_Q Y)^1) + I((X \cup_Q Y)^2) = I(X \cup_Q Y)$. Rearranging, we have

$$\begin{aligned} I((A \cup_P B)^1) + I((X \cup_Q Y)^1) + I((A \cup_P B)^2) + I((X \cup_Q Y)^2) \\ = I(A \cup_P B) + I(X \cup_Q Y). \end{aligned}$$

We know that each $I((A \cup_P B)^i) > 0$, so each $I((A \cup_P B)^i) + I((X \cup_Q Y)^i) < I(A \cup_P B) + I(X \cup_Q Y)$. The proof then follows by applying the inductive hypothesis to the splittings of each M^i.

This second condition implies the third, since any incompressible closed surface in a compression body H is parallel to $\partial_- H$. □

Attach to each F_i $(1 \leq i \leq m)$ spines Σ_{A_i} and $\Sigma_{B_{i-1}}$ of A_i and B_{i-1} in M_i and M_{i-1}, respectively. Attach to $\partial_- A_0 = \partial_- M$ and $\partial_- B_m = \partial_+ M$ spines Σ_{A_0} and Σ_{B_m} of A_0 and B_m, respectively. Similarly, attach spines Σ_{X_0} and Σ_{Y_n} to $\partial_\pm N$ and also to each G_j attach spines Σ_{X_j} and $\Sigma_{Y_{j-1}}$ of X_j and Y_{j-1} in N_j and N_{j-1}, respectively, $1 \leq j \leq n$. In each M_i, $0 \leq i \leq m$ (resp. N_j, $0 \leq j \leq n$), the regions between the spines is a product $P_i \times (0, 1)$ (resp. $Q_j \times (0, 1)$) and so can be swept out by P_i (resp. Q_j). Choose sweepouts in M and in N. Then, as in [RS], the sweepouts define an $I^{m+1} \times I^{n+1}$ parameterizing of positions of P' and Q'. Let I_i (resp. I_j) denote the interval which parameterizes the sweepout of M_i by P_i (resp. N_j by Q_j).

Put the sweepouts in generic position. Since the P_i are pairwise disjoint, as are the Q_j, generic position is easy to interpret. Locally, sweepouts intersect just as would a single pair of sweepouts. This is the situation described in [RS]. In particular, points at which P' and Q' intersect nontransversally are isolated in P' and Q', and are either nondegenerate critical points or "birth–death" tangencies. Points at which P' and G or Q' and F intersect nontransversally are isolated nondegenerate critical points.

This local picture leads to the following global interpretation. Choose specific $0 \leq i \leq m$ and $0 \leq j \leq n$, and consider just the sweepouts of M_i and N_j by P_i and Q_j (respectively), parameterized by $I_i \times I_j$. The positions of P_i and Q_j during the sweepouts are either transverse (the generic situation), have a single tangency point (on codimension-1 strata of $I_i \times I_j$), have two simultaneous nondegenerate tangencies (these occur at isolated points in $I_i \times I_j$), or have a birth–death tangency, again at isolated points. This positioning of P_i and Q_j is, of course, unaffected by motion of the other surfaces, so these codimension-0, -1, and -2 strata become strata of the same codimension in $I^{m+1} \times I^{n+1}$. Moreover, the strata are "vertical" in the sense that they are simply products in the other $m + n$ directions. Any nongeneric position of P' and Q' in M corresponds in $I^{m+1} \times I^{n+1}$ to a point lying on one or more of these "primitive" codimension-1 or -2 strata, each of which corresponds to a distinct point of tangency between P' and Q'. If we extend these comments to intersections of P' and G and Q' and F then nothing fundamental changes; each point of tangency of (say) P_i with G produces a codimension-1 stratum $\{t\} \times I^{m+n+1}$, where $t \in I_i$ defines the level of P_i in its sweepout of M_i.

Consider again the subsquare $I_i \times I_j$. Any point on the codimension-0 or -1 strata in $I_i \times I_j$ can be moved by altering the sweepouts slightly near the corresponding tangency point of P_i with Q^+ (or P^+ with Q_j). Interpreted globally, this means we can assume that the primitive codimension-1 or -2 strata in $I^{m+1} \times I^{n+1}$ are transverse. Thus $I^{m+1} \times I^{n+1}$ is broken up into strata of arbitrary codimension. A codimension-q stratum corresponds to a positioning of P^+ and Q^+ so that there are simultaneously k nondegenerate tangencies of P^+ with Q^+ and l birth–death tangencies of P' with Q', $q = k + 2l$. Just as in [RS], strata associated to birth–death tangencies or to max-min–type (index 0 or 2) tangencies play almost no role in the argument.

The same generic positioning applies in any product subcube of $I^{m+1} \times I^{n+1}$: If I^p and I^q are subcubes of I^{m+1} and I^{n+1}, respectively (corresponding to a choice of p of the P_i and q of the Q_j), then the set of points for which there are k simultaneous nondegenerate tangencies between the two collections of components is of codimension k. For example, there is no point in which each of two of the P_i are simultaneously tangent to each of three of the Q_j, since this would require at least $2 \cdot 3 > 2 + 3$ tangency points. We formalize slightly with the following definition.

DEFINITION 2.2. A *region* is an open component of the top-dimensional strata in $I^{m+1} \times I^{n+1}$.

A *hyperplane* is the closed codimension-1 stratum corresponding to a single saddle tangency point of P^+ with Q^+. Distinct regions whose closures contain an open set of the same hyperplane are called *adjacent*.

A *hyperline* is the closed codimension-2 stratum corresponding to two simultaneous saddle tangencies between the same P_i and Q_j.

Note that a point on the boundary of a hyperplane lies either on $\partial(I^{m+1} \times I^{n+1})$ or on a codimension-2 stratum associated to a birth–death tangency at which the saddle tangency of the hyperplane is "born".

3. Coding How the Surfaces Intersect

Suppose W is a region. Then any point in W represents a positioning of P' and Q' so that the surfaces intersect transversally in a collection of simple closed curves. Label the component W with two labels, μ and ν. We will describe the label μ; the other label is defined symmetrically by reversing the roles of P and Q, i and j, etc. For one of the definitions it will be helpful first to observe that if, for some $0 \le i \le m$, all curves of $P_i \cap Q^+$ are inessential in both P_i and Q^+, then a spine of P_i is disjoint from Q^+. In fact, the rest of P_i can be isotoped, with support on disks bounded by curves of $P_i \cap Q^+$, to be disjoint from Q^+. In particular, the active component $P_{\bowtie}$ of P_i then lies in some X_j or Y_j.

DEFINITION 3.1. The label μ is an $(m+1)$-tuple for which each coordinate μ_i $(0 \le i \le m)$ is a subset of the following symbols: $?, a, b, A, B, !$. The choice of symbols is determined by the set of curves $P_i \cap Q^+$ as follows.

(1) If there is a curve in $P_i \cap Q^+$ that is inessential in Q^+ and bounds an essential disk in A_i (resp. B_i) whose only intersections with Q^+ are inessential in Q^+, then include the symbol A (resp. B) in μ_i.
(2) If all curves of $P_i \cap Q^+$ are inessential in both P_i and Q^+ then, as above, some subdisks of $P_{\bowtie}$ can be properly isotoped so that $P_{\bowtie}$ lies entirely in some X_j or Y_j. One of A_i or B_i then abuts $P_{\bowtie}$ on the opposite side of $P_{\bowtie}$ from Q_j in M_j. If it is A_i, include the symbol a; if it is B_i, include the symbol b.
(3) If all curves of $P_i \cap Q^+$ are inessential in P_i but at least one is essential in Q^+, include the symbol "?".
(4) If there is at least one curve of $P_i \cap Q^+$ that is essential in P_i, and if all curves that are essential in P_i are also essential in Q^+, then include the symbol "!" in μ_i.

Similarly, define ν as an $(n+1)$-tuple for which each coordinate ν_j $(0 \le j \le n)$ is a subset of the following symbols: $?, x, y, X, Y, !$.

PROPOSITION 3.2. *For any region $W \subset I^{m+1} \times I^{n+1}$, there is precisely one symbol in μ_i. Moreover, if the symbol is a then $i = 0$ and, after an isotopy of P_0 supported on disks bounded by components of $Q^+ \cap P_0$, every component of $Q^+ \cap A_0$*

is parallel in A_0 to a component of $\partial_- A_0$. Symmetric statements hold for label b, and for labels $\nu_j = x$ or y.

Proof. Consider the curves of intersection of P_i with Q^+. If all are inessential in P_i, then we choose labels a, b or "?" and the definitions make clear that these possibilities are mutually exclusive. If at least one is essential in P_i then consider the collection of all such essential curves, and consider how they lie in Q^+. If each is essential in Q^+ then the only possible label is "!". If one is inessential in Q^+, consider an innermost such component c in Q^+: c is essential in P_i and bounds a disk in Q^+ that intersects P_i only in inessential components. Since F is incompressible, we may isotope the disk so that it is disjoint from F and then isotope its interior to remove any remaining inessential curves of intersection with P_i. Then the disk lies in either A_i or B_i, and we accordingly choose either label A or B. Labels A and B cannot both occur, since the splitting of M_i is strongly irreducible. This verifies that precisely one symbol appears in μ_i.

Suppose that $\mu_i = a$ and $i \neq 0$, so F_i lies in the interior of M. This means (by Definintion 3.1(2)) that we can isotope disks in P_i so that P_i becomes disjoint from Q^+, so that the active component $P_{\bowtie}$ lies in X_j, say, and so that $A_{\bowtie}$ lies on the opposite side of $P_{\bowtie}$ from Q_j in X_j. All components of A_i except the one $A_{\bowtie}$ containing $P_{\bowtie}$ are just products. Let $F_{\bowtie} = (F_i \cap A_a)$. Exploiting this structure of A_i, isotope all of $F_i - F_{\bowtie}$ and a spine of $F_{\bowtie}$ very near P_i and hence into $M - Q^+$. Since $F_{\bowtie}$ and Q_j lie on opposite sides of $P_{\bowtie}$ in X_j, it follows that the components of $F_{\bowtie} \cap Q^+$ that are outermost in $F_{\bowtie}$ all lie in G_j. Since both $F_{\bowtie}$ and G_j are incompressible, it follows that $F_{\bowtie}$ can also be isotoped into X_j so all of $F_i \subset (M - Q^+)$. But this would violate Lemma 2.1(3).

Finally, suppose that $\mu_i = a$ and $i = 0$, so $F_i = \partial_- M$ and, after the isotopy above, P_0 becomes disjoint from Q^+. No inactive (product) component of A_0 can contain an active component of Q^+, since the product splitting is essentially the only irreducible splitting of a product [ST2]. So, after the isotopy, every component of Q^+ contained in an inactive component of A_0 is parallel to a component of $\partial_- A_0 = \partial_- M$. Since $A_{\bowtie}$ lies on the opposite side of $P_{\bowtie}$ from Q_j in X_j (resp. Y_j), either $j = 0$ (resp. $j = n$) and $A_{\bowtie} \subset X_0$ (resp. Y_n), or $j > 0$ (resp. $j < n$) and so G_j (resp. G_{j+1}) is disjoint from F_0. In the former case, after the isotopy, $Q^+ \cap A_{\bowtie} = \emptyset$. In the latter case, after the isotopy, either all of $A_{\bowtie}$ lies in X_j or else G_j lies in $A_{\bowtie}$. If all of $A_{\bowtie}$ lies in X_j (which is possible only if $\partial_M \cap A_{\bowtie} = \partial_- A_{\bowtie} = \emptyset$, that is, if $A_{\bowtie}$ is a handlebody) then again $Q^+ \cap A_{\bowtie} = \emptyset$. If G_j lies in $A_{\bowtie}$ then it is an incompressible surface in a compression body, so G_j is parallel in $A_{\bowtie}$ to $\partial_- A_{\bowtie}$. Then any component of Q^+ between G_j and $\partial_- A_{\bowtie}$ must be parallel to $\partial_- A_{\bowtie}$. □

4. How Labels Can Change Near a Point in $I^{m+1} \times I^{n+1}$

PROPOSITION 4.1. *Suppose V and W are adjacent regions. If $\mu_i(V) = A$ or a, then $\mu_i(W)$ cannot be B or b. Symmetric statements hold for labels $\nu_j(V) = X$ or x and $\nu_j(W) = Y$ or y.*

Proof. Passing from V to W represents passing P^+ through a single saddle tangency with Q^+. This cannot change the label μ_i from A to B or from a to b (see [RS, Cor. 5.1 and Cor. 5.2]).

It also cannot change the label from b to A (or, symmetrically, from a to B). The argument is analogous to that of [RS, Lemmas 4.5 and 5.3]. Such a change of labels would mean that B_m could be made disjoint from Q^+ and yet P_m would have a ∂-reducing disk in $A_m - Q^+$. That is, both $B_{\bowtie}$, the active component of B_m, and the ∂-reducing disk lie in a single X_j or Y_j, say Y_j (necessarily Y_n or X_0 if $\partial_- B_{\bowtie} \neq \emptyset$). Attach to $B_{\bowtie}$ a maximal collection of essential 2- and 3-handles in the complement of Q^+. The resulting 3-manifold B' has boundary a surface $F'_{\bowtie}$ lying entirely in Y_j. $F'_{\bowtie}$ must be incompressible in Y_j (see Figure 2). Indeed, a compressing disk cannot lie outside B' by definition of B', and it cannot lie in B' since the splitting of M_m by P_m is strongly irreducible [CG]. Hence each component of $F'_{\bowtie}$ must be parallel to a component of G_{j+1}. If $\partial_- B_{\bowtie} \neq \emptyset$ so $j = n$ and if $G_{n+1} \subset \partial_+ N$, then B' is $\partial_- B_{\bowtie} \times I$ split by $P_{\bowtie}$. This is impossible because nontrivial splittings of any such product are reducible [ST2]. If $\partial_- B_{\bowtie} = \emptyset$ then $F'_{\bowtie}$ is the entire boundary of B' so it cannot be parallel to G_{j+1}, which does not bound in Y_j. □

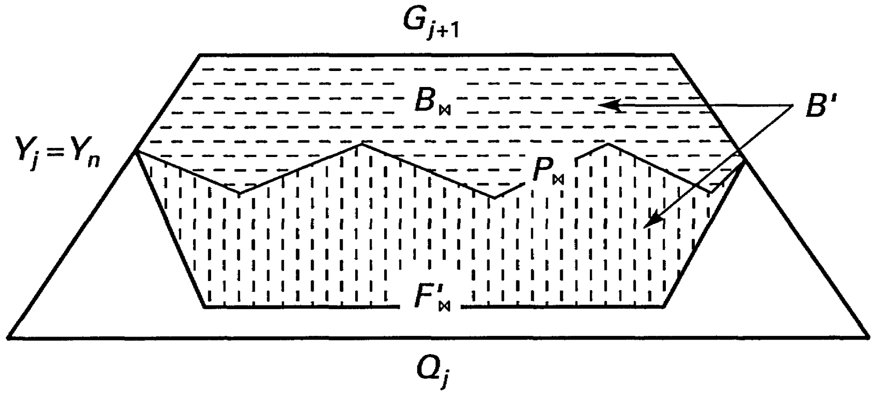

Figure 2

DEFINITION 4.2. Any region in $I^{m+1} \times I^{n+1}$ whose closure contains $z \in I^{m+1} \times I^{n+1}$ is called a *region at z.*

LEMMA 4.3. *Suppose $z \in I^{m+1} \times I^{n+1}$ and $0 \leq i \leq m$. Labels a and b cannot both occur among the labels μ_i of regions at z. A similar statement holds for labels x and y in any ν_j, $0 \leq j \leq n$.*

Proof. z represents a positioning of P' and Q' with perhaps many points of tangency. According to Definition 3.1(2) and Proposition 3.2, if a occurs as a label for a region at z then $i = 0$ and some nearby generic positioning of the active component $P_{\bowtie}$ puts a spine $P_{\bowtie}$ in some X_j (or Y_j). Similarly, if b occurs as a label then $i = n$ (so $n = 0$) and (via Definition 3.1(2)) some nearby generic positioning puts a spine of $P_{\bowtie}$ into Y_j (not X_j, since Q_j lies on opposite sides of $P_{\bowtie}$ in the two positionings). But P_0 and Q_j can have at most two simultaneous tangencies, and, in order to be able to push the spine across Q_j with two critical points, the genus of $P_{\bowtie}$ can be at most one. But this violates Lemma 2.1(1). □

LEMMA 4.4. *Suppose $z \in I^{m+1} \times I^{n+1}$ and $0 \leq i \leq m$. Labels a and B cannot both occur among the labels μ_i of regions at z. A similar statement holds for labels A and b or for labels x, Y or X, y in any ν_j, $0 \leq j \leq n$.*

Proof. Again we know that $i = 0$ and, for P_0 and Q^+ determined by z, there is some nearby positioning of P_0 that puts a spine of the active component $P_{\bowtie}$ of P_0 in some X_j (or Y_j). Another nearby positioning creates a curve c_b in $P_0 \cap Q^+$ that is inessential in Q^+ and bounds an essential disk in B_0. It follows that c_b must in fact be in $P_{\bowtie} \cap Q_j$ and at z there are at most two saddle tangencies of $P_{\bowtie}$ with Q', both at Q_j.

Consider what happens as we move across the one or two tangencies of $P_{\bowtie} \cap Q_j$ that change the label a to B. We have already shown in Proposition 4.1 that at least two saddle tangencies are required. In the region labeled a, all curves of intersection of Q_j with $P_{\bowtie}$ are trivial. Two band moves are made to get us to the region labeled B, so one of the resulting circles is inessential in Q_j and essential in $P_{\bowtie}$. If the two bands connect together three distinct curves, the result would not be essential in $P_{\bowtie}$ (see Figure 3(i)). If one connects two curves and the other is joined to the same curve, then the latter band must be inessential in Q_j but essential in $P_{\bowtie}$ (see Figure 3(ii)). The second band takes us from label a to B just with this single saddle, contradicting Proposition 4.1. Finally, suppose both bands are attached to the same curve. The new circle c_b created is inessential in Q_j and could not have been created with a single band move. Since Q_j is not a torus (by Lemma 2.1(1); see Figure 3(iii)), it follows that both bands must be inessential in Q_j, and so c_b is but one of three new curves created by adding the handles (see Figure 3(iv)). At least one of the bands must be essential in $P_{\bowtie}$. After attaching just this band we would get an essential curve in $P_{\bowtie}$ that is inessential in Q_j. This corresponds to a labeling of this region by either $\mu_i = A$ or B, and it would be adjacent to regions labeled $\mu_i = a$ and B. Hence either labeling contradicts Proposition 4.1. □

5. Labels Near a Hyperline

LEMMA 5.1. *Suppose $z \in I^{m+1} \times I^{n+1}$ and $0 \leq i \leq m$. Let V and W be regions at z with $\mu_i(V) = a$ or A and $\mu_i(W) = b$ or B.*

(1) *z lies on a hyperline $\hbar$ representing two simultaneous tangencies of P_i with some Q_j.*
(2) *In one of the quadrants of $\hbar$, all regions at z have $\mu_i = A$. In the opposite quadrant all regions at z have $\mu_i = B$, and in the other two quadrants all regions have label $\mu_i = !$.*
(3) *If V is any region at z, then in each of the quadrants of $\hbar$ there is a region U at z so that $\mu_k(U) = \mu_k(V)$ (all $k \neq i$), $\nu_k(U) = \nu_k(V)$ (all $k \neq j$), and $\mu_i(U) = !$.*

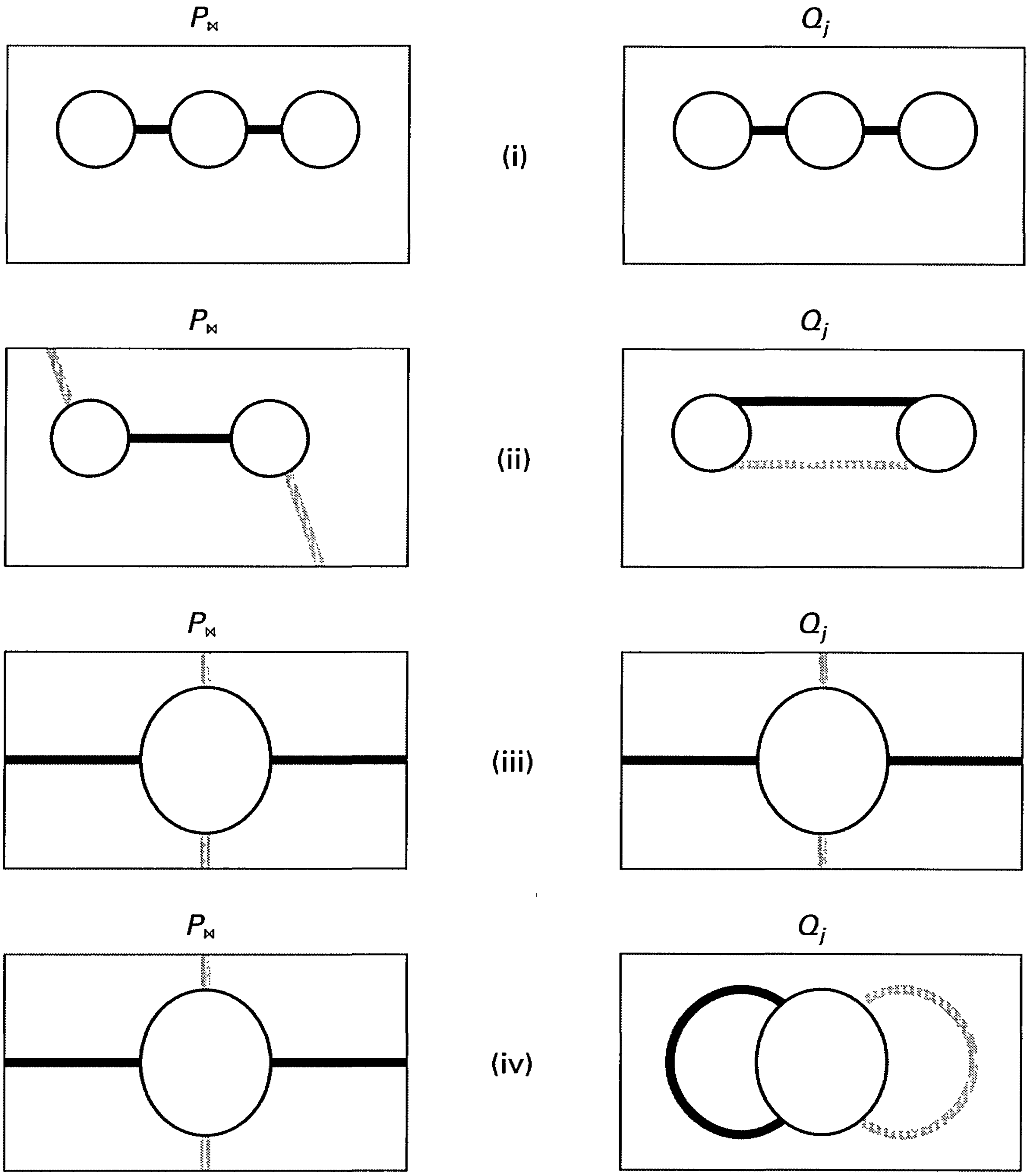

Figure 3

A symmetric statement holds if z lies in the closure of two regions, one with some $v_j = x$ or X and the other with $v_j = y$ or Y.

Proof. Following Lemmas 4.3 and 4.4, we may suppose that $\mu_i(V) = A$ and $\mu_i(W) = B$; z represents a positioning of P' and Q^+ with perhaps many points of tangency. Since $\mu_i(V) = A$, P' and Q^+ can be made generic near some of these tangency points so that a curve c_a of intersection of P_i with some Q_j is essential in P_i and bounds a disk in A_i. Similarly, some tangencies can be made generic to create a curve c_b satisfying the conditions for $\mu_i = B$. We know that neither c_a nor c_b are nonsingular curves at z itself because if (say) c_a were then we could move slightly into W without destroying c_a and thereby show that both labels occur in W. This would violate Proposition 3.2. Indeed, unless c_a and c_b lie in the same component of $P_i \cap Q^+$ at z, we could create them simultaneously by small isotopies near each tangency. It follows that they do in fact lie in the same component, and hence in a single component of Q^+. But we have seen above that P_i and a fixed Q_j can have at most two simultaneous tangencies (and P_i and a fixed G_j not even two), and these tangencies correspond to a hyperline. This verifies (1).

We have also seen (Proposition 4.1) that passing through a single tangency cannot change the label A to the label B, so c_a and c_b must be obtained by resolving two tangencies of a single component of $P_i \cap Q_j$ at z. If there is any other component of $P_i \cap Q^+$ at z which resolves into a circle satisfying the requirements for label A or B, then we could again find a region with both labels. It follows that opposite quadrants of the hyperline have labels $\mu_i = A$ and $\mu_i = B$ in every region at z and that, in the other two quadrants, no such region has $\mu_i = A$ or B. It further follows (from Lemma 4.4) that $\mu_i \neq a$ or b in any region at z. So each region at z in the other two quadrants has label $\mu_i =\,!$ or ?.

We adapt the argument of [RS, 5.6–5.7] to understand the behavior of $P_i \cap Q_j$ in the four quadrants of the hyperline. Among the curves of $P_i \cap Q_j$ determined by one quadrant (called the north) is a single component c_n to which bands corresponding to the two saddles are attached. In each of the two adjacent quadrants (the east and west) is a pair of curves in $P_i \cap Q_j$ obtained by attaching one of the two bands. We denote the pairs respectively as $c_{e\pm}$ and $c_{w\pm}$. In the remaining quadrant (the south), each of the pair of curves $c_{e\pm}$ and $c_{w\pm}$ are banded together by one of the saddles to produce either three curves or one curve of $P_i \cap Q_j$, depending on how the bands are situated. We call this curve (these curves) c_s. We know that $c_a \in c_{e\pm}$ if and only if $c_b \in c_{w\pm}$ and symmetrically. Also $c_a = c_n$ if and only if $c_b \in c_s$ and symmetrically.

CLAIM 5.1.1. *If a region at z in the northern quadrant has $\mu_i =\,!$ or ?, then the ends of both bands attached to c_n lie on the same side of c_n in P_i and c_s consists of three curves.*

Proof. In this case, $c_n \neq c_a$ or c_b. After attaching either band to c_n, one of the resulting curves (either in $c_{e\pm}$ or in $c_{w\pm}$) must be essential in P_i but not in Q_j, because one band creates c_a and the other creates c_b. If in P_i the ends of the band corresponding to one saddle tangency lie on the opposite side of c_n from the ends of the band corresponding to the other, then the compressing disks in A_i and B_i would be disjoint, contradicting strong irreducibility of M_i. Even if the ends of both bands lie on the same side of c_n and c_s is a single curve, a curve from $c_{e\pm}$ would still intersect a curve from $c_{w\pm}$ in at most one point, which again contradicts the strong irreducibility of M_i. □

CLAIM 5.1.2. *No region at z has label $\mu_i =$?.*

Proof. We first show that the northern quadrant cannot have $\mu_i =$?. If c_n were inessential in P_i then attaching one band would give two parallel curves. Hence, since one bounds a disk in A_i, so would the other and we may regard either as c_a. Similarly, attaching the other band alone would produce two parallel curves, each bounding a disk in B_i, and we may regard either as c_b. But then, after attaching both, we get three curves (Claim 5.1.1), so c_a and c_b are disjoint. This contradicts strong irreducibility. We conclude that c_n must be essential in P_i, so no region in the northern quadrant can have label $\mu_i = ?$. Similarly, no region in the southern quadrant can have $\mu_i = ?$.

If a region in the eastern (or, symmetrically, the western) quadrant had $\mu_i = ?$, then the curves $c_{e\pm}$ would both be inessential in P_i. Banding them together would produce another component inessential in P_i, contradicting the label A or B in the northern quadrant. □

This verifies part (2) of Lemma 5.1.

CLAIM 5.1.3. *The two bands are attached to c_n on the same side in one of P_i or Q_j and on the opposite side of c_n in the other.*

Proof. Suppose the northern quadrant has $\mu_i = A$ (or B) and consider how the bands are attached at c_n. If both bands were attached with ends on the same side of c_n then the resulting curve(s) c_s could be made disjoint from c_n. Since $c_a = c_n$ and $c_b \in c_s$, this would contradict strong irreducibility. It follows that the bands are attached on opposite sides in P_i. On the other hand, c_n is inessential in Q_j; if either band were on the inside of the disk c_n bounds then this would force a label $\mu_i = A$ in the eastern or western quadrant. Thus, both bands lie on the same side of c_n in Q_j.

Suppose the northern quadrant has $\mu_i = !$. From Claim 5.1.1 we can conclude that the bands are attached to c_n on the same side in P_i. But since A and B lie on opposite sides of P_i, in order to create inessential curves in Q_j bounding disks in A and B, the bands must lie on the opposite side of c_n in Q_j. □

Claim 5.1.3 means that the normal orientations of P_i and Q_j agree at one saddle point and disagree at the other. Choose axes $(I_i)_z$ and $(I_j)_z$ through z and let $(I \times I)_z = (I_i)_z \times (I_j)_z$. The hyperplane H_+ corresponding to the tangency where the normal orientations of P_i and Q_j agree will intersect $(I \times I)_z$ in a line of positive slope; the other hyperplane H_- will intersect in a line of negative slope. These lines, together with the axes $(I_i)_z$ and $(I_j)_z$ themselves, divide the square $(I \times I)_z$ into octants and in each quadrant determined by the axes $(I_i)_z$ and $(I_j)_z$ in $(I \times I)_z$ there will be an octant over which every region at z has label $\mu_i = !$ (see Figure 4). Any hyperplane at z other than $H_\pm$ is a product in the I_i and I_j direction, so it will intersect $(I \times I)_z$ in one of the axes and will have a normal vector projecting to a vector in $(I \times I)_z$ that is parallel to the axes. It follows that, if we ignore the hyperplanes $H_\pm$, any region at z in the complement of the other hyperplanes will project in $(I \times I)_z$ to some union of the four quadrants cut out by the axes. So any of the original regions V is adjacent, across one of $H_\pm$ to a region U with $\mu_i = !$. But crossing a hyperplane $H_\pm$ changes only labels μ_i and/or ν_j. This verifies part (3) and so concludes the proof of Lemma 5.1. □

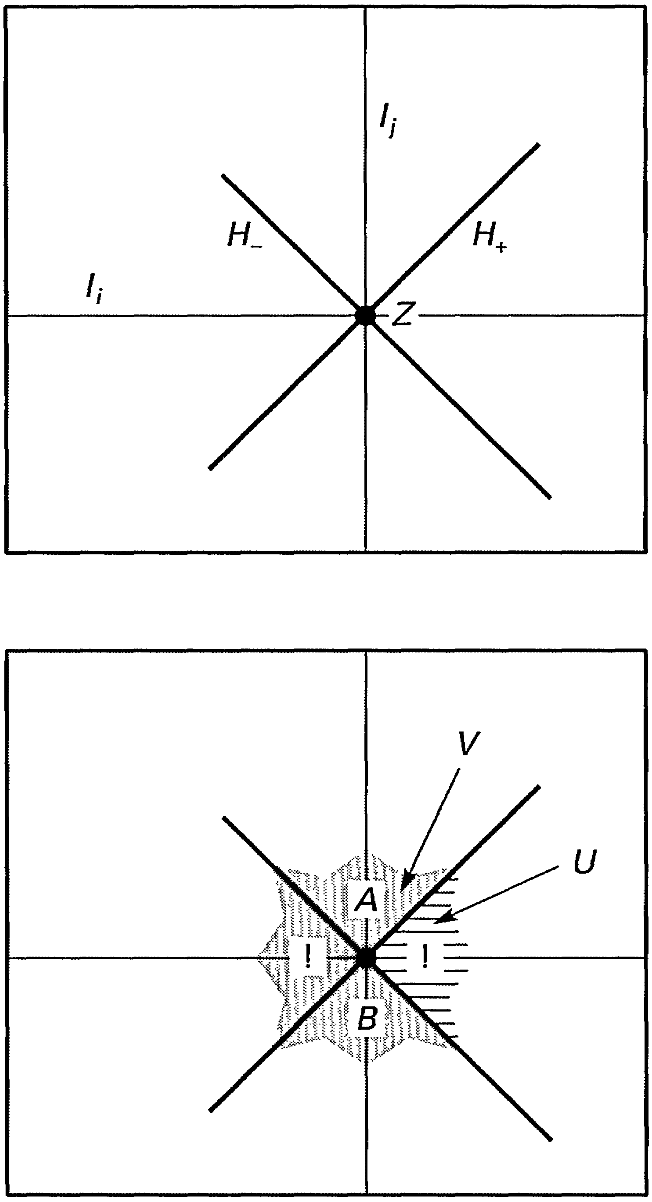

Figure 4

6. An Acyclic Map and its Kakutani Fixed Point

Our next goal will be to define a particular multivalued map $T\colon I^{m+1} \times I^{n+1} \to I^{m+1} \times I^{n+1}$ so that the image of any point is a union of faces of $I^{m+1} \times I^{n+1}$ and is contractible. We begin by defining T in a region W by using the labeling described above. For each $0 \leq i \leq m$ let $S_i = \{1\}$ if $\mu_i = a$ or A, $S_i = \{0\}$ if $\mu_i = b$ or B, and $S_i = [0, 1]$ if $\mu_i = !$ or ?. Define T_j similarly for each $0 \leq j \leq n$. Then, for any region W, define $T(W) \subset I^{m+1} \times I^{n+1}$ to be the product of all the S_i and T_j, a face of $I^{m+1} \times I^{n+1}$. For $z \in I^{m+1} \times I^{n+1}$ an arbitrary point, define $T(z)$ to be $\bigcup_{z \in \bar{W}} T(W)$. For any subcube $C = I^p \times I^q \subset I^{m+1} \times I^{n+1}$, define $T_C(z)$ to be the projection of $T(z)$ onto C.

LEMMA 6.1. *For any subcube C, $T_C(z)$ is contractible.*

Proof. The proof is by induction on the dimension of C. If C is just an interval I_i or I_j then it follows from Claim 5.1.2 that $T_C(z)$ is either an endpoint or the whole interval. So suppose the lemma is true for all subcubes of lower dimension than C. We may as well assume that I_i $(0 \leq i \leq p)$ and I_j $(0 \leq j \leq q)$ are the factors of C.

Case 1: Some label μ_i $(i \leq p)$ or ν_j $(j \leq q)$ is the same in every region at z.

Then, with no loss, assume this label is μ_p and let C' denote the subcube of C in which the factor I_p is dropped. Then $T_{C'}(z)$ is contractible, by induction. But, depending on whether μ_p is always A or a, B or b, "?" or "!", we have $T_C(z)$ is either the face $T_{C'}(z)\times\{1\}$, $T_{C'}(z)\times\{0\}$, or $T_{C'}(z)\times I_p$. In any case it is contractible.

Case 2: Some label μ_i $(i \leq p)$ is never a or A (or never b or B) in every region at z. (Or the symmetric case for label ν_j.)

Assume the label is μ_p and define C' as above. Again we know that $T_{C'}(z)$ is contractible. Let L be the subcomplex of $T_{C'}(z)$ that is the union of all faces of $T_{C'}(z)$ coming from regions for which μ_p is "?" or "!". Then $T_C(z) = (T_{C'}(z) \times \{0\}) \cup (L \times I_p)$. This clearly deformation retracts to $(T_{C'}(z) \times \{0\})$ and so is contractible.

Case 3: Among the regions at z, every label μ_i $(i \leq p)$ is somewhere a or A and somewhere b or B, and every label ν_j $(j \leq q)$ is somewhere x or X and somewhere y or Y.

Choose any $i \leq p$. Then, according to Claim 5.1.1, z lies on a hyperline representing two simultaneous tangencies of P_i with some Q_j $(0 \leq j \leq n)$. If $j > q$, so j does not occur among the coordinates of C then, according to Claim 5.1.3, $T_C(z) = T_{C'}(z) \times I_p$ and so is contractible. So suppose also that $j \leq q$ is a coordinate in C. This means that there is a region at z for which $\nu_j = X$ and one where $\nu_j = Y$. This implies a double tangency of Q_j with some P_k and, since $1+2 < 2\times 2$, k must be i. That is, Lemma 5.1 can be applied to the same pair of hyperplanes, this time using labels X and Y instead of A and B. The argument of [RS, Lemma 5.7] shows that the labels X and Y must occur in the same two quadrants as A and B, so the other two quadrants are labeled $\mu_i = \nu_j =$!. Let C'' be the subcube of C in which both factors I_i and I_j are dropped. Then it follows from Claim 5.1.3, now applied both to I_i and I_j, that $T_C(z) = T_{C''}(z) \times I_i \times I_j$, which is contractible. □

LEMMA 6.2. *The multivalued map $T\colon I^{m+1} \times I^{n+1} \to I^{m+1} \times I^{n+1}$ is closed.*

Proof. Suppose $(z, z') \notin \mathrm{graph}(T)$. Then z' is not in any face of $I^{m+1} \times I^{n+1}$ that is the image of a region at z. Since faces are closed, there is a neighborhood of z' which is also disjoint from these faces, just as there is a neighborhood of z which is disjoint from any region whose closure does not contain z. Then the product of the two neighborhoods is a neighborhood of (z, z') in $I^{m+1} \times I^{n+1} \times I^{m+1} \times I^{n+1}$ that is disjoint from $\mathrm{graph}(T)$. □

THEOREM 6.3. *There is a region for which all labels μ_i and ν_j are "!".*

Proof. According to the Eilenberg–Montgomery generalization of Kakutani's theorem [EM, Thm. 1], the map T has a fixed point z. That is, there is a $z \in I^{m+1} \times I^{n+1}$

such that $z \in T(z)$. We first show that there is a region adjacent to z for which all labels μ_i and ν_j are either "!" or "?". Equivalently, we will show that $T(z) = I^{m+1} \times I^{n+1}$. Indeed, we will show that any region W for which $z \in T(W)$ has $T(W) = I^{m+1} \times I^{n+1}$. Suppose not, so $z \in T(W)$ and $T(W)$ is a proper face. Then we can assume (with no loss of generality) that, for some i, $T_{I_i}(W) = S_i(W) = 0$ and so the ith coordinate z_i of z is also zero. Then, by definition of the sweepout, a point in W very close to z corresponds to a positioning of P_i very near to $F_i \cup \Sigma_{A_i}$. In such a position, $P_i \cap Q^+$ consists of $F_i \cap Q^+$, for which all curves which are essential in F_i are essential in Q^+, together with meridian circles of any points in $\Sigma_{A_i} \cap Q^+$. In particular, $\mu_i(W) = A$, !, or ?, so $S_i(W) = 1$ or I, a contradiction.

In W, then, each μ_i and ν_j is "!" or "?". Suppose that some $\mu_i(W) = ?$ (or $\nu_j(W) = ?$). This means that, for any point in W, all curves of the corresponding $P_i \cap Q^+$ are inessential in P_i but at least one is essential in Q^+. Then an innermost such curve in P_i bounds a disk in some X_j or Y_j. But this would imply that $\nu_j = X$, a contradiction. We conclude that in W, all labels are "!". □

Proof of Theorem 1.1. A point in the region given by Theorem 6.3 corresponds to a positioning of $P^+ \cap Q^+$ in which each curve in $P^+ \cap Q^+$ is either essential in both P^+ and Q^+ or inessential in both. Moreover, each set $P_i \cap Q^+$ and $P^+ \cap Q_j$ contains at least one essential curve. A standard innermost disk argument can be used to remove, by an isotopy, all inessential curves of $P^+ \cap Q^+$ without disturbing the essential curves. □

References

[BO] F. Bonahon and J. P. Otal, *Scindements de Heegaard des espaces lenticulaires,* Ann. Sci. École Norm. Sup. (4) 16 (1983), 451–466.

[CG] A. Casson and C. McA. Gordon, *Reducing Heegaard splittings,* Topology Appl. 27 (1987), 275–283.

[EM] S. Eilenberg and D. Montgomery, *Fixed point theorems for multi-valued transformations,* Amer. J. Math. 68 (1946), 214–222.

[H] W. Haken, *Some results on surfaces in 3-manifolds,* Studies in modern topology, pp. 34–98, Prentice-Hall, Englewood Cliffs, NJ, 1968.

[RS] H. Rubinstein and M. Scharlemann, *Comparing Heegaard splittings of non-Haken 3-manifolds,* Topology 35 (1996), 1005–1026.

[ST1] M. Scharlemann and A. Thompson, *Thin position for 3-manifolds,* Contemp. Math., 164, pp. 231–238, Amer. Math. Soc., Providence, RI, 1994.

[ST2] ———, *Heegaard splittings of (surface)* $\times$ *I are standard,* Math. Ann. 295 (1993), 549–564.

[W] F. Waldhausen, *Heegaard-Zerlegungen der 3-Sphäre,* Topology 7 (1968), 195–203.

H. Rubinstein
Department of Mathematics
University of Melbourne
Parkville, Victoria 3052
Australia

M. Scharlemann
Department of Mathematics
University of California
Santa Barbara, CA 93106