3. The algorithm

Notice that the cellular map is constructed from an equation of the form (1.1). We can apply the algorithm of the next section to this map, to get a new cellular map, from which we can extract a new equation of the form (1.1), without altering the given presentations, since at each step we can choose isomorphisms of the covering free groups which respect the relator up to conjugacy and inverse. Each element of S_1 will be transformed into a path in a labelled one-skeleton without changing the homotopy class in the surface underlying X_1 ; this amounts to choosing a new labelling for each element of S_1, which, in turn, amounts to choosing a new lifting at the free group level. This whole process will then give non-trivial group-theoretical information, although not so much as in the topological situation.

3. THE ALGORITHM

Throughout this section let $\beta\colon X_1 \to X_2$ be a cellular map of CW-surfaces.

Let V, E, and F denote the sets of vertices, edges, and faces, respectively, of X_2. We then have a diagram with V-faces, E-faces, and F-faces, as depicted in Figure 2.2.

The aim of this section is to alter β by composing it with various cellular homotopy equivalences of X_1 and X_2 (based on the operations of contracting, expanding, erasing, and subdividing), until we arrive at the minimum possible number of F-faces. These alterations of β can be viewed as homotopies, since one is free to imagine that there is a surface X underlying X_1 that has lines inscribed on it, and that these lines can be deformed continuously. Abusing notation then, we will say that the altered forms of β are *homotopic* to β.

3.1. CONSTRUCTION (type: subdivision). If X_2 has a loop e, we subdivide e by adding a new vertex v, and, in X_1, subdivide each e-edge, and each e-face, by adding a new v-vertex, and a new v-edge, respectively.

By our definition of CW-surface, X_2 has an edge. Thus we have the following.

3.2. CONDITION. *There is at least one edge in X_2, but there are no loops. Hence, in X_1, no E-edge is a loop, or equivalently, all loops are V-loops.*

We will not make any further adjustments to X_2, except in the case of a branched covering where we may have to apply Construction 3.30.

We now want to describe the basic configurations.

3.3. DEFINITIONS.

Suppose that w_1, w_2 are two distinct vertices of X_1, and d_1, d_2, d_3 are three distinct one-cells of X_1, such that d_1 is a loop at w_1, and d_2, d_3 join w_1 to w_2. Let g be a face of X_1 with boundary label $w_1, d_1, w_1, d_2, w_2, d_3^{-1}, w_1$. In this event we say that (the closure of) g is a *loop triangle*. We say that g is an *annular triangle* if d_1 is an orientable loop.

✠ We say that g is a *Möbius triangle* if d_1 is an unorientable loop.

Suppose further that β sends w_1 and d_1 to a vertex v of X_2, sends d_2, d_3 and g to a one-cell $e \in E^{\pm 1}$ of X_2, and sends w_2 to the other vertex u of e; in this event we say that g is a *loop e-triangle*. If d_1 is an orientable loop, we say that g is an *annular e-triangle*.

✠ If d_1 is an unorientable loop, we say that g is a *Möbius e-triangle*.

Two loop E-triangles are said to be *E-adjacent* if they have an E-edge in common.

By an *orientation-true prepinching*, we mean four consecutive E-adjacent triangles such that two pairs have orientable V-loops in common, and the V-loops have a common vertex, which determine a labelled CW-subcomplex as depicted in Figure 3.1(b). There are specified two edges which are incident to only one face; these edges are said to be the *boundary edges* of the prepinching. (We allow the possibility that the two boundary edges become identified in X_1, and in this case the prepinching is the whole diagram, a torus.) The remaining V- and E- edges are called the *interior* edges of the prepinching.

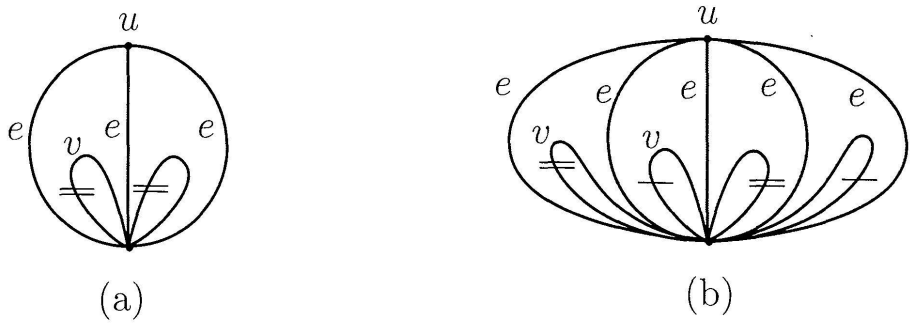

FIGURE 3.1

The two elementary types of prepinching

✠ By an *orientation-false prepinching* we mean two E-adjacent triangles having an unorientable V-loop in common, which determine a labelled CW-subcomplex as depicted in Figure 3.1(a). There are specified two edges which are incident to only one face; these edges are said to be the *boundary edges* of the prepinching. (We allow the possibility that the two boundary edges become identified in X_1, and in this case the prepinching is the whole diagram, a projective plane.) The remaining V- and E- edges are called the *interior* edges of the prepinching.

3.4. DEFINITION. The *measuring quadruple* of the diagram consists of the following non-negative integers:

(1) the number of F-faces,

(2) the number of E-faces which have boundary length at least four,

(3) the number of V-faces,

(4) the number of edges which are not interior edges in prepinchings.

The quadruples are ordered lexicographically reading from (1) to (4); this is a well-ordering.

All subsequent operations will reduce the measuring quadruple, and since the quadruples are well-ordered, the procedure must eventually stop.

3.5. CONSTRUCTIONS (type: subdivision).

(a) If, for some $v \in V$, the diagram contains a v-face, then we choose an $e \in E$ which is incident to v in X_2, and subdivide the v-face into e-triangles by adding a vertex and e-edges, as depicted in Figure 3.2 (a).

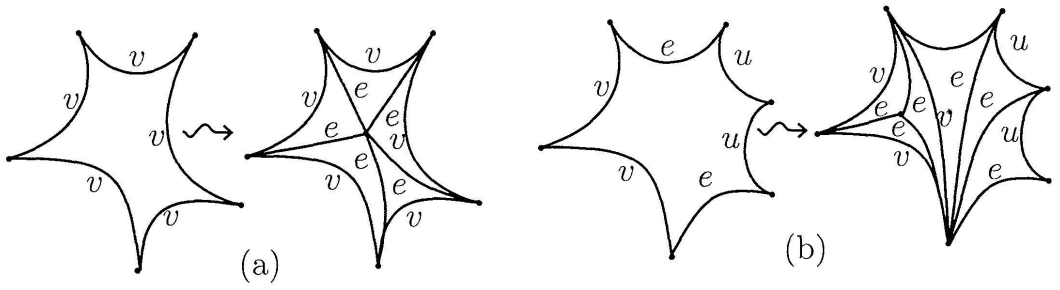

FIGURE 3.2
Subdividing into triangles

(b) If, for some $e \in E$, with vertices $u = \iota(e)$, $v = \tau(e)$, the diagram contains an e-face which has boundary length at least four, we subdivide the e-face into e-triangles as depicted in Figure 3.2 (b).

These operations reduce the measuring quadruple, since they reduce the second or third coordinate without affecting the preceding coordinates.

These are the first of several situations where we use subdivision, usually preceded by erasing, to express a homotopy between two maps which collapse a disc to a tree. Notice that it is important not to disturb the boundary of the disc, since we are not allowed to damage F-faces by collapsing E-edges.

We may now assume that we have the following.

3.6. CONDITION. *There are no V-faces.*
All E-faces have boundary length at most three.

Since an E-face has to have boundary length at least two, we then have only E-triangles and E-bigons, and no other E-faces.

Our next strategy is to eliminate some edges and faces.

3.7. CONSTRUCTION (type: collapsing). If the diagram contains a V-edge joining two distinct vertices of X_1, then we collapse the V-edge, and identify the two vertices.

This operation reduces the measuring quadruple, since it reduces the fourth coordinate, and does not increase any of the other coordinates.

We may therefore assume we have the following.

3.8. CONDITION. *All V-edges are loops.*

3.9. TERMINATING CASE. *If our diagram contains an E-face of boundary length two, and the two edges are identified in X_1, then X_1 is a sphere, and β is a degree zero map which collapses it to an edge which is not a loop, and we have the Normal Form 3.31(a).*

We may therefore assume we have the following.

3.10. CONDITION. *If an E-face has boundary length two, then the two edges are not identified in X_1.*

3.11. CONSTRUCTIONS (type: collapsing).

(a) If, for some $e \in E$, the diagram contains an e-face of boundary length two, whose edges are not identified in X_1, then the closure of the face is an e-disc, and we collapse it to an e-edge, as in Figure 3.3 (a).

(b) If the diagram contains an e-face of boundary length three, and two of the edges are identified in X_1, then, by Conditions 3.6 and 3.8, for some $v \in V$, the third edge is a v-loop, the e-face is a v-disc, and we then collapse this v-disc to a v-vertex, as in Figure 3.3 (b).

These operations reduce the measuring quadruple, since they reduce the fourth coordinate without affecting the preceding coordinates.

FIGURE 3.3
Two types of collapsing

Thus we may assume we have the following.

3.12. CONDITION. *Each E-face is a loop E-triangle.*

To summarize, in terms of closures in X_1, all E-edges are non-loops, all E-faces are loop triangles, there are no V-faces, and all V-edges are loops, and are incident to two loop E-triangles.

✠ We now turn our attention to unorientable V-loops.

✠ 3.13. CONSTRUCTION (Skora [16]; type: expand, erase, subdivide, collapse). If an unorientable V-loop is incident to two non E-adjacent E-faces, then we create an orientation-false prepinching by applying the steps depicted in Figure 3.4.

✠ This operation reduces the measuring quadruple, since it reduces the fourth coordinate without affecting the others.

✠ Thus we may assume the following.

✠ 3.14. CONDITION. *Each unorientable V-loop is the interior edge of an orientation-false pre-princhning.*

✠ This condition will enable us to ignore the unorientable V-loops, and treat them as if they were hidden within edges.

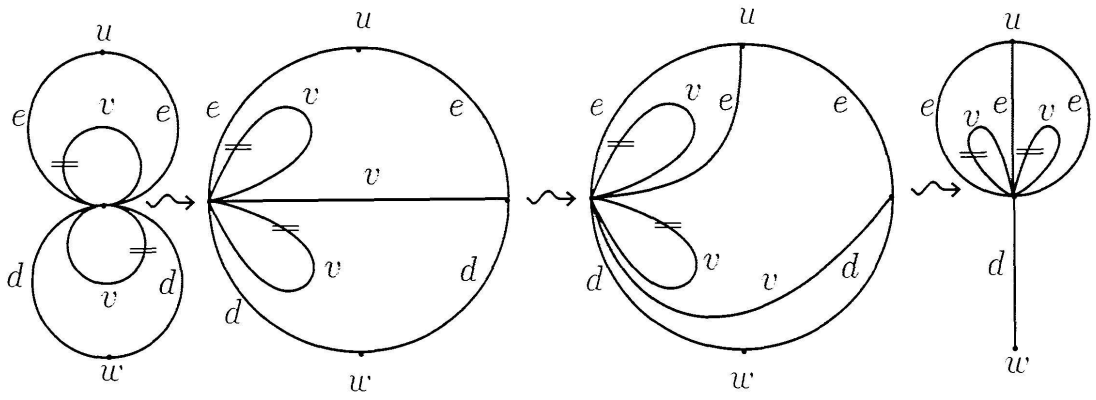

FIGURE 3.4
Normalizing a prepinching

We now want to examine how the orientable V-loops which meet at a vertex fit together.

3.15. DEFINITIONS. Let w be a vertex in the diagram, and denote its label by $v \in V$.

By a *labelled cycle around* w we mean a finite sequence

$$g_1 . d_1 . g_2 . \ldots . d_m . g_{m+1} = g_1 .$$

where each d_i is an edge with a distinguished incidence to w, but is not a v-loop, each g_i is a face with a distinguished incidence to d_i, and d_i, d_{i+1} become adjacent in g_i after omitting v-loops. That is, we are listing face-adjacent edges, except where E-triangles have loops at w in which case we treat the v-loop as a vertex, and pass from one E-edge to the other. Thus we are looking at the face-and-edge cycles around w which arise when we collapse to w all the v-loops at w.

There are various types of labelled cycles. For any e in E, an e-triangle can be E-adjacent to another e-triangle or to an F-face. By finiteness of the diagram, every e-triangle lies in a labelled cycle of e-triangles, as in Figure 3.5 (a), or in a sequence joining together two F-faces, as in Figure 3.5 (b). In the case of Figure 3.5 (a), we say that the e-faces form a *punctured e-sphere*.

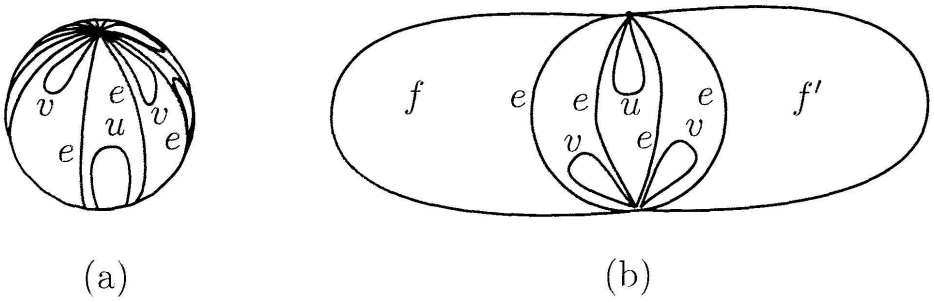

FIGURE 3.5
Adjacent e-triangles

✠ Notice that, by Condition 3.14, any unorientable loops are identified in neighbouring pairs.

The whole subsurface is collapsed to e by β. Thus we see that one possibility for a labelled cycle around w consists of e-triangles and e-edges.

Consider now the case of Figure 3.5 (b). Here we get a sequence, starting at an f-face, for some $f \in F$ with a distinguished incidence (f, i) to e, and ending in an f'-face, for some $f' \in F$ with a distinguished incidence (f', i') to e. We say that the f-face and the f'-face are *e-joined.* There are two possibilities. Either $(f, i) \neq (f', i')$, so they are the two faces with a distinguished incidence to e (in X_2), or $(f, i) = (f', i')$. In the former case, we say the f-face and the f'-face are *well* joined, and in the latter case we say they are *badly* joined. Notice that if they are badly joined, then the two F-faces are both f-faces, and they must be distinct f-faces, since if they are equal, then their distinguished e-edges must be equal, and these e-edges are then incident to zero or two e-faces and one f-face, which contradicts the surface property.

From Figure 3.5 (b), we see that the second possibility for a labelled cycle around w consists of F-faces joined together cyclically by E-triangles.

If the F-faces in a labelled cycle around w are well-joined, then it is easy to see that the corresponding cycle of labels in F is given by repeating the face cycle around v in X_2 an integral number of times; if, moreover, there is only one labelled cycle around w, then the number of times the face cycle around v is repeated will be called the *branching degree* at w.

If all the faces are F-faces, and all E-joined F-faces are well joined, then the corresponding map is a branched covering, since it becomes a covering if all the vertices are deleted.

Here the number of F-faces is a multiple of the size of F, and the quotient is the degree d of the map, and the map is then a d-fold branched covering.

If $d = 1$ then the map is a homeomorphism. If the branching degree at each vertex is 1, then the map is a covering.

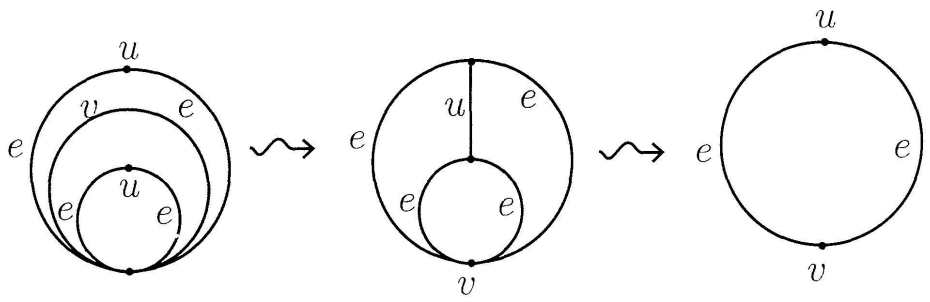

FIGURE 3.6
Collapsing

3.16. CONSTRUCTIONS. Suppose that two annular E-triangles have a common V-edge, and a common E-label, but do not lie in a prepinching.

Thus, for some $e \in E$, with vertices $u, v \in V$, two annular e-triangles have an orientable v-loop in common, and the v-loop is not an interior edge of a prepinching. We consider various possibilities for the intersection of (the closures of) the two annular e-triangles. Since the two e-triangles have an orientable loop in common, they get separated into different face cycles after collapsing the loop to a vertex, so they cannot have an e-edge in common. However, they may have a common u-vertex.

(a) (type: erasing, subdividing, collapsing). If the intersection of the two annular e-triangles is precisely the v-loop, then we erase the v-loop, and draw in a u-edge which is not a loop, and we can now collapse the resulting triangles, as depicted in Figure 3.6. Thus we take an annulus, which is a compact subsurface with two boundary components, and homotope it to a loop formed by two edges, in such a way that the boundaries are respected.

(b) (type: expanding, erasing, subdividing, collapsing). If the two annular e-triangles with a common v-loop also have a common u-vertex, then we create an orientation-true prepinching, as depicted in Figure 3.7. Thus we take a punctured torus, which is a compact subsurface with a single boundary component, and fit it into a sliced-open edge, in such a way that the boundary is respected.

Each of these two operations reduces the measuring quadruple, since it reduces the fourth coordinate without affecting the preceding coordinates.

We may therefore assume that we have the following.

3.17. CONDITION. *Each V-loop is either interior to a prepinching, or is an orientable V-loop incident to two E-faces with different E-labels.*

Moreover, all E-faces are loop triangles and there are no V-faces.

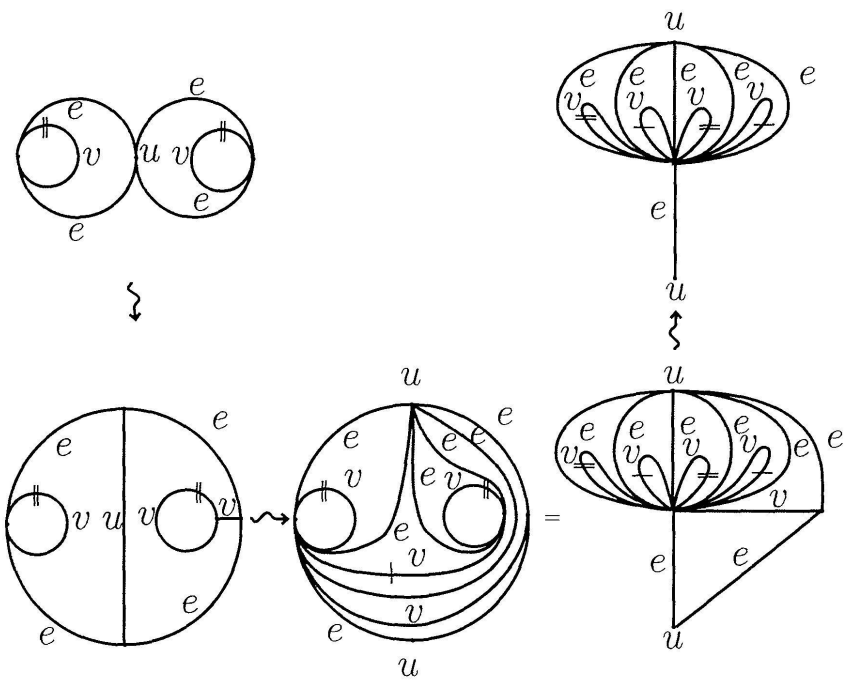

FIGURE 3.7
Normalizing a prepinching

3.18. CONSTRUCTION (type: expanding-collapsing). Suppose that, at some vertex of the diagram, there are two distinct labelled cycles having an E-label in common.

Thus there exist two e-edges incident to a vertex, and these two e-edges are separated into different labelled cycles by some orientable v-loop at the vertex, as in the left diagram in Figure 3.8.

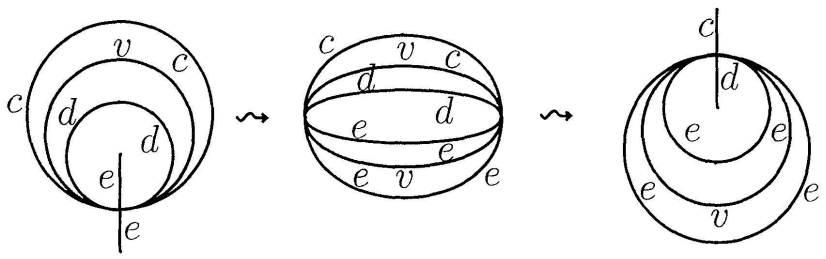

FIGURE 3.8
Readjusting

We expand the two e-edges into e-triangles, and then collapse the resulting c- and d-triangles to c- and d-edges, as in Figure 3.8, without disturbing the two boundary components at any stage. Possibly the interior e-edge equals one of the d-edges, and possibly, the exterior e-edge equals one of the c-edges; in this event, the corresponding identifications must be maintained throughout the operation.

Thus we have applied a homotopy which does not affect the measuring quadruple, and we can now apply the Constructions 3.16 (a) and (b), which will reduce the measuring quadruple.

Hence we may assume the following.

3.19. CONDITION. *At each vertex, distinct labelling cycles have disjoint label sets.*

3.20. CONSTRUCTION (type: expanding, erasing, subdividing, collapsing). Suppose there is a vertex w, such that there is only one labelling cycle around w, and there is a V-loop incident to w which is not interior to a prepinching.

Since collapsing the V-loop at w separates the face cycle around w into two disjoint cycles, and, by hypothesis, collapsing all the V-loops at w leaves a single face cycle, there must be some V-loop which connects up the two face cycles, giving us the situation depicted in the left diagram in Figure 3.9.

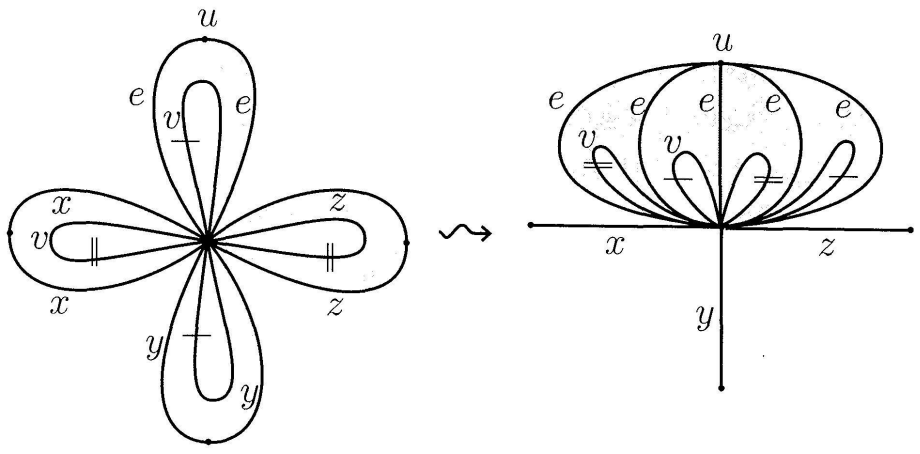

FIGURE 3.9
Normalizing a prepinching

It is now straightforward to homotope this configuration into an orientation-true prepinching, as depicted in Figure 3.9, without affecting the single boundary component.

This reduces the measuring quadruple in the fourth coordinate, without affecting the other coordinates.

Hence we may assume the following.

3.21. CONDITION. *At each vertex with a single labelling cycle, all V-loops are interior to prepinchings.*

3.22. TERMINATING CASE. *If there is no F-face then we have the following situation.*

All the faces are E-triangles, and the diagram is formed by amalgamating punctured E-spheres along the V-loops, and the E-spheres which meet at a vertex have distinct E-labels.

The algorithm now terminates, as we have the Normal Form 3.31(a).

Hence we may assume the following.

3.23. CONDITION. *The diagram has at least one F-face.*

3.24. CONSTRUCTION (type: erasing and subdividing). Suppose that two F-faces are badly E-joined, as in the left diagram of Figure 3.10.

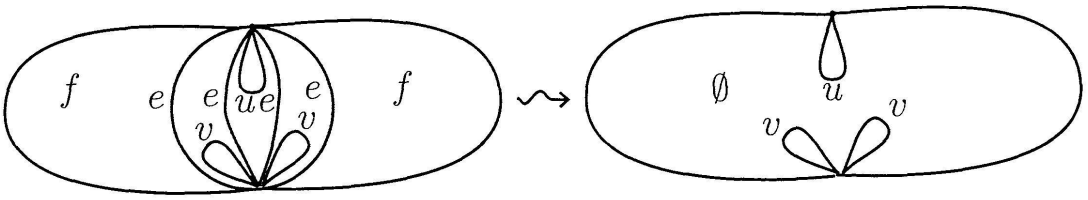

FIGURE 3.10
Starting over

Here we abandon all the progress we have made. We erase all the E-edges involved, to join up distinct faces, and so obtain a disc with a boundary label which determines a trivial element of $\pi X_2^{(1)}$, as in the second diagram of Figure 3.10. Now we apply Construction 2.7 to fill in the disc with V- and E-faces, and so get a new diagram representing a map homotopic to β.

This procedure reduces the first coordinate of the measuring quadruple by two.

The foregoing construction disturbs the conditions we have obtained so far, and we return to Construction 3.5.

Repeating the procedure up to this stage a finite number of times, we eventually eliminate all pairs of F-faces which are badly joined. Thus we may assume the following.

3.25. CONDITION. *Each E-joined pair of F-faces is well-joined. Hence the only faces are E-triangles forming prepinchings, and F-faces. Moreover, the branching degree is defined at each vertex.*

Proof. By Condition 3.23, there is at least one F-face.

Let us consider a vertex w incident to an F-face, and look at the labelled cycle around w containing the F-face. Since F-faces are well joined, we see, as in Definition 3.15, that as we run through the labelled cycle, the labels run through the face-and-edge cycle around v, where $v \in V$ is the label of w. Thus every edge incident to v occurs as a label in the labelled cycle around w. It follows, from Condition 3.19, that there is only one labelled cycle around w. Now, by Condition 3.21, all the V-loops at w are interior to prepinchings. Hence all the faces incident to w are E-triangles forming prepinchings and F-faces. It follows that the compact subsurface formed by the F-faces and the E-triangles occurring in prepinchings is closed under edge adjacency, so is the whole surface.

Hence the branching degree is defined at each vertex. □

3.26. TERMINATING CASE. *If the branching degree is 1 at each vertex then after the pinching, consisting of collapsing to edges the prepinching regions which are as depicted in Figure 3.1, we have a diagram in which all faces are F-faces, and all E-joined F-faces are well joined, and the branching degree at each vertex is 1, so it is a diagram representing a covering.*

The algorithm terminates since we have the Normal Form 3.31(b).

Hence we may assume the following.

3.27. CONDITION. *The branching degree is at least two at some vertex w.*

Let $v \in V$ denote the label of w. Since w has branching degree at least 2, the face labels around w run at least twice through the faces round v. Thus we can choose an $f \in F$ with a distinguished incidence to v, and choose the first two terms of the labelled face sequence around w with this label. Clearly

these two faces with distinguished vertex are distinct. Choose an $e \in E$ in the boundary cycle of f next to the distinguished occurrence of v. Thus we have the situation occurring in the left diagram in Figure 3.11, and the two e-edges are distinct.

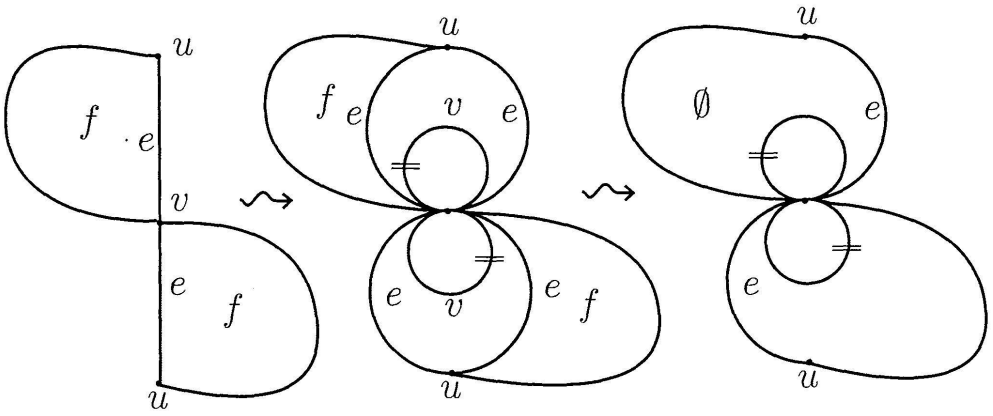

FIGURE 3.11
Starting over

✠ 3.28. CONSTRUCTION (Skora [16]; type: erase-subdivide). Suppose the diagram has at least one unorientable V-loop.

✠ We claim that unorientable V-loops are highly mobile, in the sense that two Möbius E-triangles, attached along an unorientable V-loop (and possibly an E-edge) forming a Möbius band (that is, a punctured projective plane), can slide around; this sliding has the same effect as cutting open a pair of incident edges (resp. an edge) to form a loop, which is then identified with the boundary of the Möbius band, while the reverse of such an operation is performed somewhere else. This can be shown using Construction 3.13, and its reverse, and similar arguments, and we will not go into details since they are straightforward. These operations do not affect the first coordinate of the measuring quadruple, which is the coordinate the operation will eventually reduce. Thus we can move one of the unorientable V-loops into the two e-edges, as depicted in the middle diagram of Figure 3.11.

✠ Now we have two f-faces, and two e-faces, which give four distinct faces, and we erase the three edges along which they are joined, to get an open disc, as in Figure 3.11, and the boundary label is $\partial f, e, e^{-1}, \partial f^{-1}$, which determines a trivial element in the free groupoid $\pi X_2^{(1)}$. Notice that the boundary cycle has a repeated vertex which causes the closure of the disc to be attached to itself with a twist, and there may be other boundary identifications. We now apply Construction 2.7 to subdivide the disc into V- and E-faces.

✠ The first coordinate of the measuring quadruple drops by two.

✠ As happened after Construction 3.24, we have to return to Construction 3.5 and repeat all the steps. Since the measuring quadruple is reduced in the first coordinate, eventually it reaches a stage where it cannot drop any more. Now we have the following.

✠ 3.29. CONDITION. *The diagram has no unorientable V-loop.*

Once the algorithm arrives here, it stops, and we do not consider the measuring quadruple any more. We now perform a tidying operation, which alters F, and increases the number of F-faces.

3.30. CONSTRUCTION (Edmonds [3]; type: subdivide X_2, erase-subdivide-relabel). Suppose that there is an orientable V-loop, and hence a pair of orientable V-loops in a prepinching region, by Condition 3.25.

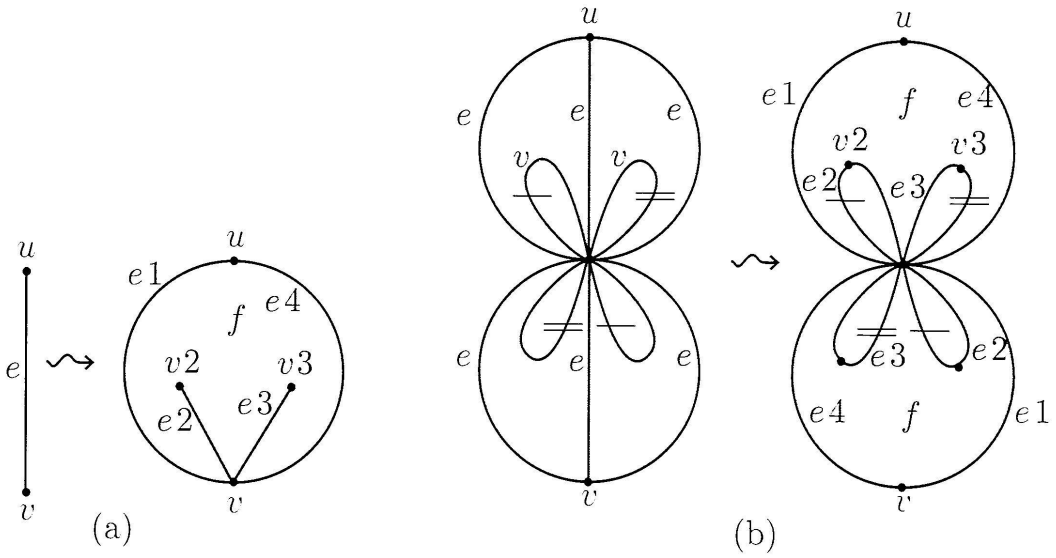

FIGURE 3.12
Changing prepinchings to get a branched cover

It follows from Construction 3.16, and its reverse, that pairs of orientable V-loops are highly mobile, and we can move them into the pair of e-edges incident to w that is given by Condition 3.27, as in the left diagram in Figure 3.12 (b). Now we can subdivide a face of X_2 incident to e, to create a new face f, and new faces and vertices, as in the right diagram in Figure 3.12 (a), or, equivalently, expand the edge e into a disc. We expand each e-edge in

X_1 to a disc in a corresponding manner, with the exception of the six e-edges occurring in the left diagram in Figure 3.12 (b), where we erase, subdivide edges, and relabel, as indicated in Figure 3.12 (b).

We see that all F-faces are well E-joined, and the branching degree is unchanged at each old vertex, and that, at the two added vertices, the branching degree is 2.

Repeating this operation once for every prepinching region, we eliminate all the V-loops, and obtain a diagram in which all faces are F-faces, and all E-joined F-faces are well joined, so it is a diagram representing a branched covering.

The algorithm terminates since we have the Normal Form 3.31 (c).

The foregoing algorithm has proved the main result.

3.31. NORMAL FORM THEOREM (Kneser [11], Edmonds [3], Skora [16]). *If $\beta: X_1 \to X_2$ is a cellular map of CW-surfaces, then there exists a cellular map $\beta': X_1' \to X_2'$ of CW-surfaces homotopic to β whose diagram satisfies one of the following.*

(a) (Degree zero) *The diagram is a union of punctured E-spheres with bouquets of V-loops at the two poles, and the V-loops are identified in pairs in such a way that no two distinct E-spheres with a vertex in common have the same E-label. The identifications have the property that, at each vertex, there is only one face-and-edge cycle, but there can be various labelled cycles. If X_1 is orientable, the identifications can be chosen to respect orientations of the spheres. After collapsing the spheres to edges, there are no faces, and all incident edges have distinct E-labels, yielding an immersion of graphs. Here β has degree zero.*

(b) (Pinching followed by covering) *After pinching, consisting of collapsing the prepinchings to edges, all faces are F-faces, all edge-adjacent pairs are well E-joined, and the branching degree at each vertex is 1, yielding a d-fold covering for some positive integer d. Here β has degree d.*

(c) (Branched covering) *All faces are F-faces, and all edge-adjacent pairs are well E-joined, yielding a d-fold branched covering for some positive integer d. Here β has degree d.*

3.32. COROLLARY (Kneser [10], [11]). *Any cellular map of CW-surfaces of degree 1 is homotopic to a (possibly trivial) pinching.*

The following is an interesting illustration of Theorem 3.31.

3.33. EXAMPLE: SELF-MAPS OF THE REAL PROJECTIVE PLANE. By considering the Puppe exact sequence [12, p. 238], [13, p. 3] associated to a map $S^1 \to S^1$ of degree 2, one finds that each pointed homotopy class of maps from a real projective plane to a real projective plane is determined by its degree, and the possible values are 0, 2, and the odd positive integers. In particular, the same holds for the (unpointed) homotopy classes of maps.

A degree zero map is given by collapsing the source surface to a point. This is of type (a).

A degree two map is given by collapsing an unorientable loop to a point to obtain a two-sphere, and then composing with a double covering of the projective plane. This is an orientation-false pinching composed with a double covering, so is of type (b).

An odd positive integer degree map is given by taking an odd positive degree covering of one Möbius band by another, and then collapsing the boundaries to points. This is a branched covering with a single branch point, so is of type (c).

In the usual way, the homotopy classes of self-maps of the real projective plane form a monoid under composition; to calculate composites one need only calculate the degree, and that can be done easily, even using the algorithm given here. Thus we can identify each homotopy class with its degree, and examine the binary operation induced by composition. We find that the monoid is obtained from the usual multiplicative monoid of non-negative integers by identifying two distinct non-negative integers if and only if they are even and are equal modulo 4.

4. HOMOMORPHISMS OF SURFACE GROUPS

Throughout this section, let $\alpha\colon G_1 \to G_2$ be a homomorphism of infinite surface groups, and $G_1 = \langle S_1 \mid r_1 \rangle$, $G_2 = \langle S_2 \mid r_2 \rangle$ be surface group presentations.

4.1. REVIEW. The arguments of Sections 2, 3 give us a method for finding a normal form for α, and hence for calculating the degree of α.

Let us itemize the steps performed.