THE FUNDAMENTAL GROUP OF A SURFACE, AND A THEOREM OF SCHREIER

BY

H. B. GRIFFITHS

The University, Birmingham, England

Introduction

Schreier proved in [8] that a finitely generated normal subgroup $U \neq \{1\}$ of a free group F is of finite index. This result was extended by Karrass and Solitar in [3], to the case when U is not necessarily normal, but contains a non-trivial normal subgroup of F. In Topology, the free groups occur as fundamental groups of surfaces with boundary, and we here extend the result still further (Theorem 6.1) to the case when F is the (non-abelian) fundamental group of any connected surface, with or without boundary, except for a Klein bottle. We use topological methods, and also the elements of Morse theory, although the latter could be eliminated. A sketch of this theory is included, however, partly for its intuitive appeal, and partly because the Morse theory picks out "stable" generators of the fundamental group, and therefore is helpful as a tool. Indeed, the author was able to use it quickly to prove Schreier's Theorem, and Theorem 3.3 below (that an open surface has free fundamental group), before knowing that proofs already existed in the literature. Our exposition is always from the point of view that it is the surface, rather than the group, which is the ultimate object of study; and we have perhaps laboured points that might irk the pure group-theorist.

2. Riemann surfaces

A Riemann surface (S, Φ) ([1], ch. II) is a 2-dimensional manifold S, for each point x of which there is a "co-ordinate chart" $(U, u) \in \Phi$ (where U is an open neighbourhood of x in S, and u is a homeomorphism of U onto an open subset of a plane), such that if U, V are overlapping charts, the homeomorphism

$$v \circ u^{-1}: u(U \cap V) \to v(U \cap V) \tag{i}$$

is conformal, with conformal inverse. The pair (S, Φ) is often denoted by S. It then makes sense to speak of harmonic functions on S; a real function f: $S \to R^1$ is "C^∞" resp. "harmonic" if and only if, for each chart $(U, u) \in \Phi$, $f \circ u^{-1}$ is an ordinary infinitely differentiable resp. harmonic function on $u(U)$. We therefore call $x \in S$ a non-degenerate critical point of f if and only if $u(x)$ is a non-degenerate critical point of $g = f \circ u^{-1}$ on $u(U)$; i.e. the derivatives g_x, g_y are both zero, and $g_{xx} g_{yy} - g_{xy}^2 \neq 0$ at $u(x)$. These definitions are independent of u, since the Jacobian of $v \circ u^{-1}$ in (i) is non-zero, while a harmonic function is C^∞.

Also x is a maximum, minimum, or saddle-point of f if the same is true for $u(x)$ and $f \circ u^{-1}$. The sets $\{f = c\}$ are the "levels" of f, and the value c is "critical" or "ordinary" according as $\{f = c\}$ contains a critical point or not. The "half space" S^c is defined by

$$S^c = \{x \mid f(x) \leqslant c\},$$

and its boundary ∂S^c is the level $\{f = c\}$. By an implicit function theorem, if c is ordinary, then ∂S^c is a 1-dimensional manifold.

A Riemann surface has always a Riemannian metric, by means of which we can define the orthogonal trajectories of (in particular) a C^∞ function f, (Morse [4] p. 150). These are curves τ, orthogonal to the ordinary levels of f and parametrised such that $f(\tau(t)) = t$ for each t in the domain of τ. There is a unique trajectory τ through each ordinary point x of f; and τ depends continuously on x. On the other hand, if x is, for example, an isolated saddle-point of f, then in a suitable chart, f is of the form $\alpha^2 - \beta^2$ at x; so there are exactly two trajectories

$$l(x), \quad r(x) \tag{ii}$$

corresponding to the portions $\beta \leqslant 0$, $\beta \geqslant 0$ of the β-axis, through x and lying in Int $(S^{f(x)} - x)$, and two others through x and lying in Int $(S^{(-f)(x)} - x)$. The basic lemmas of Morse Theory that we need here can then be stated as follows (Morse–Van-Schaak [6]).

Lemma A. *Let* [a, b] *be a closed interval of ordinary values of f on S. Then S^a is a deformation retract of S^b. If also* [a, b] $\subseteq$ [c, b], *where $c < a$ and $[c, b]$ contains no critical value of f, then*[1] $S^a \cong S^b$.

[1] $\cong$ denotes homeomorphism.

Remark. The proviso about c is automatically satisfied if S^a is compact, by the continuity of grad f. The existence of c is required in the proof of Lemma 2.2 in [4] VI.

COROLLARY A. *If $\gamma_1, \ldots, \gamma_m$ are a set of singular* 1*-cycles on S^a, linearly independent over the integers, then they are linearly independent on S^b.* (For S^a is a retract of S^b.)

LEMMA B. *Let c be an isolated critical value of f, and suppose that the critical points of f on S^c form a countable, isolated set of saddle points x_i, $i \in \Sigma$.*

Then for a sufficiently small $\varepsilon > 0$, $S^{c-\varepsilon}$ is a deformation retract of a set X, where

$$S^{c+\varepsilon} = X \cup \bigcup_{i \in \Sigma} \sigma_i.$$

and(1) (a) *each $\sigma_i \cong I^2$, $x_i \in$* Int *σ_i, $\sigma_i \cap X \cong \dot{I} \times I$;*

(b) *all the σ_i are mutually disjoint;*

(c) *the trajectory $l(x_i)$,* [see (ii) above] *travels into X from x_i through one of the two segments of $\sigma_i \cap X$, and $r(x_i)$ does so through the other.*

Remark. The proof is a direct consequence by induction on $i \in \Sigma$ of Theorem 4 of Morse–Van Schaak [6] p. 559 (where the result is proved for a set Σ with just one member).

COROLLARY B. *With the hypotheses of Lemma B, suppose that c is the only critical value of f, in the interval* [a, b] *of values, where $a < c < b$. Then the inclusion $S^a \subseteq S^b$ induces, among the singular homology groups:*

(1) *an epimorphism $H_0(S^a) \to H_0(S^b)$; and*

(2) *a monomorphism $H_1(S^a) \to H_1(S^b)$.*

(This follows by induction on $i \in \Sigma$, in view of the condition $\sigma_i \cap X \cong \dot{I} \times I$ in $B(a)$.)

Combining (2) with Corollary A, we have the following useful "stability" property:

COROLLARY C. *Let the interval* [a, b] *contain only a finite number of critical levels of f, on each of which the critical points are countable, isolated saddle-points. Then any set $\gamma_1, \ldots, \gamma_m$ of singular* 1*-cycles, on S^a, linearly independent over the integers, remains linearly independent on S^b.*

Similarly, provided S^a is connected and contains the base point ω, we obtain an analogue for the fundamental groups based at ω:

[1] I denotes the unit interval with boundary $\dot{I}$; I^2 is the unit square.

COROLLARY D. *Under the hypotheses of Cor. C, let $g_1, \dots, g_m$ be a set L of loops on S^a based at ω. If L forms part of a free basis of $\pi_1(S^a)$, then it forms part of a free basis of $\pi_1(S^b)$.*

Expressed in this form, these results are useful when "lifting" a function into a covering space: see 5.2 below.

3. Topology of surfaces

Although the topological results are well known, and can be obtained by purely topological means, it is not without interest to carry out a study, using Morse theory, of the topological structure of the general Riemann surface S described in Sect. 2. The Morse theoretic treatment seems intuitively enlightening, especially since the existence of differentiable structure, and of the Green's function we use (= "potential due to a point charge, with earthed boundary") are intuitively very plausible.

Let then S be a Riemann surface, and let T be a compact region in S, whose boundary ∂T consists of a finite set of disjoint smooth Jordan curves. Fix a base-point $\omega \in \operatorname{Int} T$. Then T has a "Green's function" Γ with pole ω; that is, $\Gamma : T - \{\omega\} \rightarrow R^1$ is continuous, $\Gamma | \partial T$ is zero, Γ is harmonic on $\operatorname{Int} T - \{\omega\}$, while in a co-ordinate chart (U, u) round ω

$$\Gamma \circ u^{-1}(z) = \log |z| + V(z) \quad (u(0) = \omega), \tag{i}$$

where $V(z)$ is harmonic on U. (General Reference: Ahlfors–Sario [1].)

Now, for any $x \in T$, in any simply connected chart (V, v), $\Gamma \circ v^{-1}$ is harmonic on $v(V)$ and hence is the real part of a complex analytic function f. The critical points ζ of Γ then correspond to the zeros of the derivative of f, and therefore these are isolated. If ζ is degenerate, then by modifying Γ near ζ by adding a linear function, (see Morse) we obtain a function Γ' with only non-degenerate critical points, and these all saddle-points; and such that the critical values of Γ' are all distinct, say $c_1 < c_2 < \dots < c_n$, with corresponding distinct critical points $\zeta_1, \dots, \zeta_n$. That $\zeta_i \in \operatorname{Int} T$ follows immediately from applying the following [1] lemma, to a co-ordinate chart in S at each boundary point of ∂T.

3.1. LEMMA. *Let D be a disc in the plane, separated into components A, B by a differentiable arc ab joining the points $a, b \in D$. Let ϕ be a harmonic function on A, constant along ab, with $\phi < \phi\ (ab)$ on A. Then denoting by ν the normal from A to B*

$$\partial \phi / \partial \nu > 0 \quad \textit{at all points of } ab.$$

[1] Oral communication from W. K. Hayman.

Proof. Map A conformally, by a map θ onto a semi-circular disc Δ, in the lower half plane, to carry ab into the real axis. Then $\psi = \phi \circ \theta^{-1}$ is harmonic on Int Δ, and so can be continued across the axis, by the Reflection Principle. If $\partial \phi / \partial v$ were zero at $q \in \operatorname{Int}(ab)$ then $\partial \psi / \partial y$ would be zero at $u = \theta(q)$; now $\partial \psi / \partial x = 0$ at u, so ψ is the real part of a complex analytic function

$$f(z) = a + b(z-u)^2 + \text{higher order terms.} \tag{ii}$$

But we can always assume that ψ is zero on the real axis, so $\psi(v)$ is positive or negative in company with the imaginary part of v, because $\phi < \phi(ab)$ on A. This conflicts with the behaviour of f in (i). Hence $\partial \phi / \partial v$ is >0 at all points as required.

Continuing the discussion of Γ in (i) above, we see that if $|c|$ is sufficiently large, and $c < c_1$ (say $c \leqslant c_0 < c_1$) then the co-ordinate transformation

$$x' = x e^{V(x,y)}, \quad y' = y e^{V(x,y)}, \quad (V \text{ in (i)})$$

transforms the half-space T^c, with non-vanishing Jacobian, onto $\log |z| \leqslant c$; hence T^c is homeomorphic to a 2-cell, I^2. By Lemma A of Section 2:

$$T^c \cong I^2 \quad \text{if } c < c_1. \tag{iii}$$

Recall that the critical levels of Γ' on Int $T - \{\omega\}$ were $c_1 < \ldots < c_n$; let $\varepsilon = \frac{1}{2}$ min $((c_{i+1} - c_i), |c_n|)$, so that $\varepsilon > 0$. Then by Lemma B in Section 2, $T^{c_1+\varepsilon}$ is the union of a set Z together with a "strip" σ_1:

$$T^{c_1+\varepsilon} = Z \cup \sigma_1$$

where $T^{c_1-\varepsilon}$ is a deformation retract of Z, $\sigma_1 \cong I^2$; $\zeta_1 \in \operatorname{Int} \sigma_1$, and $\sigma_1 \cap Z$ is a pair of disjoint arcs σ_1', σ_1''.

Thus, by (iii), $T^{c_1+\varepsilon}$ is a disc to which a strip has been joined (without twisting) by its ends σ_1', σ_1'', and so (taking ω as base-point for the group), $\pi_1(T^{c_1+\varepsilon})$ is cyclic, and generated by an ω-based loop of the form $l r^{-1}$, where l, r are trajectories of the form 2 (ii). Applying Lemmas A and B in turn at each critical point ζ_i on the level c_i, we see inductively that $\pi_1(T^{c_i+\varepsilon})$ is free on i generators. Further, T is built up by adding strips $\sigma_1, \ldots, \sigma_n$ (one for each ζ_i) to the disc $D = T^c$ in (iii); and comparing Betti numbers R_0, R_1, we note the following facts.

When σ_{i+1} is added to $D_i = D \cup \sigma_1 \cup \ldots \cup \sigma_i$, then

$$\left.\begin{aligned} R_0(\partial(D_i \cup \sigma_{i+1})) &= R_0(\partial D_i) + \alpha_i, \\ R_1(D_{i+1}) &= R_1(D_i) + 1, \end{aligned}\right\} \tag{iv}$$

where α_i is either 1 or -1. Let n^+, n^- denote the number of σ_i with α_i positive and α_i negative, respectively. Then by induction on (iv),

$$R_0(\partial T) = 1 + n^+ - n^-, \tag{v}$$

$$R_1(T) = n^+ + n^- = n, \tag{vi}$$

$$\pi_1(T, \omega) \textit{ is free on } n \textit{ generators.} \tag{vii}$$

Since $R_0(\partial T) \geqslant 1$, then (v) shows that $n^+ \geqslant n^-$; say

$$n^+ = n^- + d, \quad d = R_0(\partial T) - 1 \geqslant 0.$$

Note that, by (v) and (vi), *the topology of T determines the numbers n^+, n^-*, and hence the critical points of any approximation like Γ' to the Green function of T.

It is not hard to show, by induction, that T is homeomorphic to a "standard model" M consisting of the union of a disc D with n^+ holes h_i, and n^- "bridges" b_i joining the edge of h_i to ∂D, if $i \leqslant n^-$. From this model, it is easy to work out a homology basis and the associated intersection numbers; for, since a bridge b_i crosses h_i, then the intersection matrix has rank n^-. To summarise, we have the

3.2. THEOREM. *A compact Riemann surface with $\beta > 0$ boundary curves and first Betti number R_1, has a free fundamental group with R_1 generators, and is homeomorphic to a sphere with β holes and $\frac{1}{2}(R_1 + 1 - \beta)$ handles.*

Moreover, if the model M is such that ∂M is a circle γ, then $n^+ = n^-$ (by (v)) and we can obtain the standard closed surface V of genus $n^+ = p$ by adding a disc to ∂M along γ. By taking a suitable set of generators a_i, b_i of $\pi_1(M, \omega)$, $\gamma \cong [a_1, b_1][a_2, b_2] \ldots [a_p, b_p]$, so we see immediately that $\pi_1(V)$ has generators a_i, b_i $(1 \leqslant i \leqslant p)$ and relation $\gamma = 1$.

Next, let us indicate the corresponding analysis for a *non-compact* connected Riemann surface S which is countable, i.e. S can be expressed as a union

$$S = \bigcup_{n=0}^{\infty} S_n, \quad S_n \subseteq \operatorname{Int} S_{n+1},$$

where each S_n is a compact domain and sub-manifold of S with non-empty boundary. Such an S is called an *open surface*; its boundary is empty. Moreover, we can arrange that ∂S_n consists of *smooth* Jordan curves, such that

$$\textit{each component of } S_{n+1} - \operatorname{Int} S_n \textit{ has at least one} \\ \textit{boundary curve in } \partial S_n \textit{ and one in } \partial S_{n+1}. \tag{viii}$$

Let ω be a fixed point in Int S_0, and let Γ_0 be the Green's function for S_0, with pole ω. If $n > 0$, define Γ_n to be that harmonic function on $S_n - \operatorname{Int} S_{n-1}$ which is n on ∂S_n and $n-1$ on ∂S_{n-1} (this uses (viii)). By Lemma 3.1 the critical points of Γ_n lie in Int S_n; and they are finite in number because in a simply connected chart (U, u), $\Gamma_n \circ u^{-1}$ is locally the real part of a complex analytic function. By an argument mentioned above, we can modify each Γ_n so that its critical points are all saddle-points with distinct critical values.

Define a global function $\Gamma : S \to R^1$ by setting

$$\Gamma \,|\, (S_n - \operatorname{Int} S_{n-1}) = \Gamma_n, \quad (S_{-1} \textit{ empty}); \tag{ix}$$

then Γ is clearly continuous, and is differentiable except possibly on the levels $\{\Gamma = \text{integer}\}$. Such a level, $\{\Gamma = n\}$ is still "ordinary", however, by Lemma 3.1; for if $x \in \{\Gamma = n\}$, then a trajectory exists from x into $\{\Gamma < n\}$ and one into $\{\Gamma > n\}$, and these two join at x to make a continuous (if not differentiable) curve. Differentiability of trajectories is not required in the proof of Lemma A. Thus, if the critical levels of Γ are $c_1 < c_2 < \ldots$, corresponding to critical points $\zeta_1, \zeta_2, \ldots$, then arguing as for (iii) and using Lemmas A and B, we see as for T above that S is a union

$$S = D \cup \sigma_1 \cup \sigma_2 \cup \ldots \cup \sigma_n \cup \ldots, \quad \zeta_i \in \operatorname{Int} \sigma_i, \tag{*}$$

(which may terminate) of 2-cells D, σ_i, where $\sigma_{n+1} \cap (D \cup \sigma_1 \cup \ldots \cup \sigma_n)$ is a pair of arcs. Thus

$$\pi_1 (D \cup \sigma_1 \cup \ldots \cup \sigma_{n+1}) = \pi_1 (D \cup \sigma_1 \cup \ldots \cup \sigma_n) * F_1$$

(free product), where F_1 has just one generator, and this corresponds to σ_{n+1}. Hence $\pi_1 (S)$ is free (cf. [1], p. 102).

A Riemann surface is always orientable. But if S is a non-orientable (connected) compact or open surface which is differentiable [i.e., the maps $v \circ u^{-1}$ in 2 (i), and their inverses, are required only to be differentiable], then S has an orientable double covering p: $T \to S$. The differentiable structure on S induces one on T, so that p becomes differentiable with nowhere-vanishing Jacobian. We construct a function Γ as in (ix) on T, and then define an analogue $\Gamma' : S \to R^1$ by

$$\Gamma' (x) = \Gamma (x_1) + \Gamma (x_2)$$

where $p^{-1} (x)$ consists of x_1 and x_2. Then Γ' is differentiable with a single pole, without maxima or minima, and possessing only saddle points for critical points. Hence the decomposition (*) is obtained as before, except that some of the cells σ_i may be twisted. Therefore if S is compact, we can derive generators and relations for $\pi_1(S)$ in the same way as for the orientable case discussed above; together with an analogue of Theorem 3.2. Also since every open surface has a differentiable structure, then by using 3 (vi) and its non-orientable analogue we have (cf. [1] p. 102):

3.3. THEOREM. *If S is a surface, then $\pi_1(S)$ is free provided either S is open, or S is compact with $\partial S \neq \emptyset$.*

3.4. COROLLARY (of proof). *If S is an open surface, with finite Betti number R_1, then S contains a compact manifold S_* with boundary, such that the injection*

$$j: \pi_1(S_*) \rightarrow \pi_1(S)$$

is an isomorphism. In fact S_* can be taken to be a half-space S^a of Γ.

(For, the number of strips σ_i in (*) is exactly k, the Betti number R_1; so we can take a to be any number $> c_k$.)

It would be interesting to know if the decomposition (*) could be derived by topological means, without the difficult job of having to triangulate S. For example, can S be given differentiable structure without a previous triangulation? And can a function like Γ be found without the complicated theories of harmonic functions or of polar functions (Morse [5])?

Further deductions from 3.3 are as follows.

3.5. THEOREM. *If S is an open surface with cyclic fundamental group, then S is either an unbounded annulus, or an unbounded Moebius strip.*

Proof. Since $\pi_1(S)$ is cyclic, the set S^a of 3.4 contains exactly one critical point of Γ. Therefore, by (*), S^a is the union of a disc and a strip and hence is either a bounded annulus A or Moebius band B. Choose $b > a$. By Lemma A of Sect. 2, $S^a \cong S^b$, so Int S^b is an unbounded annulus or Moebius band according as S^b is A or B. As described prior to 2 (ii), each $x \in S - S^b$ lies on a unique trajectory τ which meets ∂S^a, ∂S^b in unique points $r = \tau(a)$, $s = \tau(b)$ respectively, and $\Gamma(r) = a$, $\Gamma(s) = b$. Let $\phi: (a, \infty) \cong (a, b)$ be a homeomorphism, and define a homeomorphism $\psi: S \cong \text{Int } S^b$ by setting $\psi | S^a =$ identity, and otherwise $\psi(x) = t$, where t is on τ and $\Gamma(t) = \theta(\Gamma(x))$. Because τ depends continuously on x, then ψ and ψ^{-1} are continuous, as required. By the remarks above, S is therefore an unbounded annulus or Moebius strip, which completes the proof.

3.6. COROLLARY. *In* 3.4 *the components of* $S-S^a$ *are all unbounded annuli.*

Proof. We know by Lemma A, Sect. 2, that if $b>a$, there is a homeomorphism $\theta_a^b : S^b \cong S^a$, which in particular assigns to each point $x \in \partial S^b$ the point $\theta_a^b(x) \in \partial S^a$ where the unique trajectory through x meets $\{\Gamma = a\}$. Since such a trajectory passes through each $y \in S-S^a$, then y lies on a Jordan curve $J_{\alpha t}$, where $t=\Gamma(y)$, and $\theta_a^t(J_{\alpha t})$ is a component J_α of ∂S^a; since S^a is compact, then $1 \leqslant \alpha \leqslant N$, say. Hence $S-S^a$ is the union of sets

$$B_\alpha = \bigcup_{a<t<\infty} J_{\alpha t} \quad (1 \leqslant \alpha \leqslant N).$$

By uniqueness of trajectories, the curve $J_{\alpha t}$ through y is unique; hence the components of $S-S^a$ are the sets B_α, $1 \leqslant \alpha \leqslant N$. Each of these is a line bundle over a circle since the trajectory through y depends continuously on y; hence by 3.5, each B_α is an unbounded annulus if B_α is orientable, and possibly an unbounded Moebius strip otherwise. Now $S-S^a$ is open in S, so each component B_α is orientable if S is; thus B_α can be non-orientable only if S is.

But, if B_α were non-orientable, so would $B = \bigcup_{a,\leqslant t \leqslant a+1} J_{\alpha t}$, which is then a Moebius strip. But then the addition of B to S^a would introduce an element of period 2 in $H_1(S)$, since $B \cap S^a = J_\alpha$. This is impossible,(1) since $H_1(S)$ is free abelian. This completes the proof.

4. The centre of $\pi_1(S)$

Let S be a surface with base-point ω. We now consider covering surfaces of S, and use freely the general results of Hilton–Wylie ([2], ch. 6). In particular, we recall that if G is any subgroup of $\pi_1(S)$, then there is a connected covering space S_G with projection $p: S_G \to S$, and natural base-point $\omega_G \in p^{-1}(\omega)$, such that the induced homomorphism $p_*: \pi_1(S_G, \omega_G) \to \pi_1(S, \omega)$ has kernel zero, and image G. Moreover, p is a local homeomorphism, so S_G *is also a surface*, open if S is open; and if S is open or compact, then $\pi_1(S_G)$ is countable. For future reference we record

4.1. *If* S *is compact, then* G *is of finite index in* $\pi_1(S)$ *if and only if* S_G *is compact.*

Consequently we have from 3.3:

4.2. *If* S *is compact and* G *is of infinite index in* $\pi_1(S)$, *then* G *is free.*

We now derive some consequences of 4.2. It will be convenient to call a surface S "abelian" or not, according as $\pi_1(S)$ is abelian or not. The analysis in 3.2 and 3.5 shows that

[1] This impossibility was kindly pointed out by the referee.

4.3. *If S is abelian then*

(a) *if S is compact and $\partial S = \emptyset$, then S is a sphere, torus, or real projective plane;*

(b) *if S is compact and $\partial S \neq \emptyset$ then S is a Moebius strip or an annulus;*

(c) *if S is open then S is an unbounded Moebius strip or annulus.*

Our main interest will therefore be in non-abelian surfaces. For brevity, if H is a subgroup of a group G, then $[G:H]$ will denote the index of H in G. Since we intend to prove Theorem 6.1 below, which is false when S is a torus, we also exclude here and now the case when S is a Klein bottle K, for K too is a counter-example as we now show. $\pi_1(K)$ has two generators a, b, with one relation $a b a^{-1} = b^{-1}$, so the cyclic subgroup β generated by b is normal since $x b x^{-1} = b^{\pm 1}$ for all $x \in \pi_1(K)$. Hence $\pi_1(K)/\beta$ is free cyclic, generated by a, so $[\pi_1(K):\beta] = \infty$, and therefore β is a finitely generated normal subgroup of $\pi_1(K)$ of infinite index. It can be shown that the centre of $\pi_1(K)$ is free cyclic, generated by a^2, in contrast to 4.4 below; a full treatment of the normal subgroups of $\pi_1(K)$ will appear elsewhere.

We now consider non-abelian surfaces $S \neq K$.

4.4. THEOREM. *If $S \neq K$ is a non-abelian surface, then the centre of $\pi_1(S)$ is trivial.*

Proof. If S is open, or if S is compact with boundary $\neq \emptyset$, then $P = \pi_1(S)$ is free; and since S is non-abelian, then P is freely generated by a set $\{g_i\}$ of at least two generators. Given $x \in P$, x has a unique reduced form $g_1^{\alpha_1} \dots g_r^{\alpha_r}$, and so if $i \neq r$, then $g_i x \neq x g_i$. Thus, the centre of P is $\{1\}$.

There remains the case when S is compact and $\partial S = \emptyset$. The centre C of P is therefore *not of finite index in P*; otherwise by 4.1, S_c is compact, abelian, and $\partial S_c = \emptyset$ whence by 4.3 (a), S_c is a sphere, torus, or real projective plane. But then by comparing Euler characteristics, S itself would be one of these three surfaces, contrary to the fact that S is non-abelian. Hence, $[P:C] = \infty$, so by 4.2, C is free; but C is abelian, so C is either trivial or cyclic infinite. If $C \neq \{1\}$, the abelianised group $(P/C)^A$ has rank $\geqslant 2$, for P has at least three generators. One of these generators at least, has no multiple $\neq 1$ in C, so [1] rank $(P/\{C,g\})^A \geqslant 1$, where $\{C,g\} = N$ denotes the normal subgroup of P generated by C and g. Therefore N is of infinite index in P, whence by 4.2, N is free; and N is not cyclic, by choice of g. Thus

[1] When S is a real projective plane with one handle, g is not completely arbitrary. Here P has 3 generators a, b, c, and the single relation $2c = 0$; so at most one of a, b, c has a non-zero multiple lying in the cyclic group C^A. Hence we can choose g so that either a or b is free in $(P/\{C, g\})^A$, thus ensuring a rank $\geqslant 1$.

the centre C_N of N is trivial, by the word-theoretic argument at the start of the proof. But $C \subseteq C_N$, so $C = \{1\}$, as required.

Remark. It would be interesting to have a geometric, as opposed to a word-theoretic, proof that a free non-cyclic group has a trivial centre.

4.5. COROLLARY. *If $S \neq K$ is non-abelian, no non-trivial cyclic subgroup of $P = \pi_1(S)$ is normal.*

Proof. If possible let $g \neq 1$ generate a cyclic normal subgroup G of P. Then for each $x \in P$, $xgx^{-1} \in G$, so $xgx^{-1} = g^{n(x)}$, say, for some non-zero integer $n(x)$. It follows that

$$n(xy) = n(x) \cdot n(y), \quad n(1) = 1, \tag{i}$$

whence $1 = n(xx^{-1})$, so $n(x) = \pm 1$. Let $H = \{x \mid x \in P \text{ and } n(x) = 1\}$; then by (i), H is normal in P and is either all of P or of index 2. In the first case, every element of P commutes with g, so G lies in the centre of P, which is impossible by 4.4. In the second, g commutes with every element in H, and $n(g) = 1$, so G lies in the centre of $H = \pi_1(S_H)$. Therefore S_H is a surface whose fundamental group has non-trivial centre, so either S_H is a Klein bottle (which is impossible since the Euler characteristic of S is not zero), or H is abelian and therefore either infinite cyclic (say Z) or $Z+Z$ or Z_2. If $H = Z$, then S_H is either a (possibly unbounded) annulus or Moebius strip, using 4.3, covering S twice. Hence S is the same so P is abelian contrary to the inequality $H \neq P$. If $H = Z+Z$, then S_H is a torus, covering S twice. Hence S is a torus, or Klein bottle, contrary to hypothesis. If $H = Z_2$, then P has just four elements, which is impossible since P is the fundamental group of a connected surface. Hence H cannot be of index 2 in P, so $H = P$ which we have seen to be impossible. Hence $G = \{1\}$. This completes the proof.

As we saw prior to 4.4, Cor. 4.5 is false when $S = K$.

The following is essentially a corollary of 3.3. *If S is a surface, not the real projective plane, then $\pi_1(S)$ contains no element of finite order.*

Proof. If G were a non-trivial finite cyclic subgroup of $\pi_1(S)$ then $\pi_1(S)$ could not be free so (by 3.3) S is compact, with $\partial S = \emptyset$. Since G is not free then S_G is compact, by 3.3. But then $\pi_1(S_G) \approx G$, and the only surface with fundamental group of this kind is the real projective plane, P^2. But then comparing Euler characteristics, $\chi(P^2) = g \cdot \chi(S)$ where $g = [\pi_1(S); G]$. Now $\chi(P^2) = 1$, so $g = \chi(S) = 1$, and $S \approx P^2$, contrary to hypothesis. This completes the proof.

5. Covering spaces

Next, we need to gather some results from the general theory of covering spaces. Let $\phi:(A,a)\rightarrow(B,b)$ be a regular covering map of based spaces, so that

$$G=\pi_1(B,b)/\phi_*\pi_1(A,a)$$

acts (anti-isomorphically) as a group of covering transformations of A, without fixed points. Thus, following 4.5, there exist compact non-abelian surfaces S, such that $\pi_1(S)$ *possesses finitely generated normal subgroups N of finite index* (hence $N\neq\{1\}$). For any non-abelian surface S' possesses at least one finite group G of automorphisms without fixed-point (see e.g. Nielsen [7] p. 95), and S' is a g-leaved regular covering space of the orbit space S'/G, where $g=$ order of G. Hence S'/G is a compact surface S whose fundamental group P contains a normal subgroup N isomorphic to $\pi_1(S')$ such that P/N is anti-isomorphic to G. Hence N is normal in P, finitely generated, and incidentally not free (cf. 4.2).

Returning to the general $\phi:A\rightarrow B$ above, we recall that if $w\in B$, then $\phi^{-1}(w)=\{w_\alpha\}$, where, given w_α, w_β, there exists $g\in G$ such that $g(w_\alpha)=w_\beta$. Moreover (by definition), w has a "ϕ-canonical" neighbourhood W such that $\phi^{-1}(W)$ consists of a set of neighbourhoods U_α in A, one for each w_α, with the property that $\phi_\alpha=\phi\,|\,U_\alpha$ is a homeomorphism onto W.

5.1. Lemma. *With $g(w_\alpha)=w_\beta$ as above, we have $g(U_\alpha)=U_\beta$, provided W is path-connected.*

Proof. Let $u_\alpha\in U_\alpha$ and let λ: $I,\dot{I}\rightarrow W,(w,\phi_\alpha u_\alpha)$ be a path from w to $\phi_\alpha u_\alpha$. Then $\phi_\alpha^{-1}\circ\lambda=\lambda_\alpha$, $\phi_\beta^{-1}\circ\lambda=\lambda_\beta$ are paths in U_α, U_β from w_α to u_α and w_β to $u_\beta=\phi_\beta^{-1}(\phi_\alpha u_\alpha)$ respectively. But then λ_β and $g\circ\lambda_\alpha$ are paths covering λ, and issuing from w_β; therefore they are equal. Thus $g(u_\alpha)=u_\beta$, so $g(U_\alpha)\subseteq U_\beta$. Similarly $g^{-1}(U_\beta)\subseteq U_\alpha$, so $g(U_\alpha)=U_\beta$ as claimed.

5.2. At this point, we remark that we could prove at least the original form of the Schreier Theorem mentioned in the Introduction ($N=U$, $\pi_1(S)$ free in Theorem 6.1 below) directly using the above theory and that of Sect. 2. This was done in the first draft of this paper, and we sketch the method briefly, since it has some intrinsic interest. Let E be an open plane disc with $n>0$ holes, let p: $E_N\rightarrow E$ be the covering associated with $N\subseteq\pi_1(E)=F$, the free group on n generators; and suppose $[F:N]=\infty$. Since E is a Riemann surface, the local homeomorphism p induces a Riemann structure in E_N. Hence the function $\Gamma_N=\Gamma'\circ p$, induced by the function Γ' of 3 (ix), has isolated critical points, all saddle points lying above those of Γ'; and the critical levels of Γ_N are those of Γ'. Also Γ_N has a discrete infinity of poles lying above that of Γ'. If any loop in E, of the form $l\,r^{-1}$ in 2 (ii), were to lift as a loop λ in E_N, then λ would consist of pieces of trajectory of Γ_N, and so λ would, together with its iterates under F/N, form an infinite set of linearly independent cycles in E_N, by Cor C, Sect. 2, – on the assumption $[F:N]=\infty$. Hence it would follow that $H_1(E_N)$ is infinitely generated, contradicting the fact that $\pi_1(E_N)\approx N$ and N is finitely generated. Hence λ is not a loop, so E_N is simply connected and $N=\{1\}$. This method does not unfortunately work so well when E is replaced by a compact surface, and we alter our tactics as follows.

We consider the following situation. Let X be a locally connected, locally compact, pathwise connected space based at ω; let

$$X_N \xrightarrow{\alpha} X_U \xrightarrow{\beta} X \tag{i}$$

be covering spaces corresponding to the subgroups $N\subseteq U\subseteq G=\pi_1(X,x)$. Suppose that N is normal in G, so that the groups of covering transformations:

$$U/N=\mathfrak{U}\subseteq\mathfrak{G}=G/N$$

act on X_N without fixed-points.

5.3. Theorem. *If $\mathfrak{G}/\mathfrak{U}$ is countably infinite, and $A\subseteq X_N$, $B\subseteq X_U$ are compact, then for all but a finite number of $\mathfrak{g}\in\mathfrak{G}$,*

$$\alpha(\mathfrak{g}\,A)\cap B=\emptyset.$$

Proof. There is a natural $(1-1)$ correspondence

$$G/U\approx(G/N)\,/\,(U/N),$$

so we have coset decompositions

$$\left.\begin{aligned} G/U&=g_1\,U\cup g_2\,U\cup\ldots\cup g_n\,U\cup\ldots,\\ \mathfrak{G}/\mathfrak{U}&=\mathfrak{g}_1\,\mathfrak{U}\cup\mathfrak{g}_2\,\mathfrak{U}\cup\ldots\cup\mathfrak{g}_n\,\mathfrak{U}\cup\ldots;\end{aligned}\right\} \tag{ii}$$

where $\mathfrak{g}_n$ is the image of g_n under the natural homomorphism $G\to\mathfrak{G}$, and (therefore) the set $K=\{g_n\}$ is infinite.

Since β in (i) is a covering map, each $x\in X$ has a neighbourhood W such that $\beta^{-1}(W)$ is a disjoint union of neighbourhoods P_i in X_U, one P_i for each $x_i\in\beta^{-1}(x)$, and the x_i being in $(1-1)$ correspondence with G/U, hence with K. Moreover, $\beta_i = \beta | P_i$ is a homeomorphism of P_i on W. Also by considering $\gamma = \beta \circ \alpha$ in (i), and choosing W smaller if necessary, we can suppose, by 5.1, that W is γ-canonical, so that

$$\gamma^{-1}(W) = \{\mathfrak{g}\, Q\},$$

for one fixed $Q \subseteq X_N$, as $\mathfrak{g}$ runs through $\mathfrak{G}$. We can express $\gamma^{-1}(W)$ in two ways, first as

$$\gamma^{-1}(W) = \bigcup_{g_n \in K} \{(\mathfrak{g}_n \mathfrak{u}) \cdot Q\} \quad (\mathfrak{u} \in \mathfrak{U}), \tag{iii}$$

(where for clarity, the action of $\mathfrak{G}$ on X_N is denoted by a dot), and second as

$$\gamma^{-1}(W) = \alpha^{-1}(\beta^{-1} W) = \bigcup_{g_n \in K} \alpha^{-1}(P_n), \tag{iv}$$

where we recall that each P_n corresponds bi-uniquely to g_n. We reconcile (iii) and (iv) by observing that $\alpha^{-1}(P_n)$ is a disjoint union of *some* of the sets $\mathfrak{g} \cdot Q$ (since they are connected) and then showing more precisely that

$$\alpha^{-1}(P_n) \text{ *is the disjoint union of the sets* } \mathfrak{g}_n \mathfrak{u} \cdot Q, \tag{v}$$

$$\text{*as* } \mathfrak{u} \text{ *runs through* } \mathfrak{U}.$$

We first prove

$$\alpha((\mathfrak{g}_n \mathfrak{u}) \cdot Q) = \alpha\,((\mathfrak{g}_n \mathfrak{u}') \cdot Q) \quad \textit{for all } \mathfrak{u},\ \mathfrak{u}' \in \mathfrak{U}. \tag{a}$$

For $(\mathfrak{g}_n \mathfrak{u}) \cdot Q = \mathfrak{u} \cdot (\mathfrak{g}_n \cdot Q)$ since $\mathfrak{G}$ acts anti-isomorphically on X_N; and since $\mathfrak{u}$ is a covering transformation relative to α, then

$$\alpha((\mathfrak{g}_n \mathfrak{u}) \cdot Q) = \alpha(\mathfrak{u} \cdot \mathfrak{g}_n \cdot Q) = \alpha(\mathfrak{g}_n \cdot Q)$$

from which (a) follows at once. As a kind of converse, we now prove

If $$\alpha(\mathfrak{g} \cdot Q) \cap \alpha(\mathfrak{g}' \cdot Q) \neq \emptyset, \tag{b}$$

then $$\mathfrak{g} \equiv \mathfrak{g}' \ \textit{mod}\ \mathfrak{U}.$$

By the observation above, $\alpha(\mathfrak{g} \cdot Q)$ lies in some P_k, and therefore since the P's are mutually disjoint, then $\alpha(\mathfrak{g} \cdot Q)$, $\alpha(\mathfrak{g}' \cdot Q)$ each lie in P_k. By the hypothesis in (b), there exist $\mathfrak{g}, q' \in Q$ such that$\alpha(\mathfrak{g} \cdot q) = \alpha(\mathfrak{g}' \cdot q')$. Hence there exists $\mathfrak{u} \in U/N = \mathfrak{U}$ such that $\mathfrak{u} \cdot (\mathfrak{g} \cdot q)$; and then by 5.1, $\mathfrak{g}' \cdot Q = \mathfrak{u} \cdot (\mathfrak{g} \cdot Q)$. Therefore $\mathfrak{g}' = \mathfrak{g}\,\mathfrak{u}$ since $\mathfrak{G}$ acts anti-isomorphically on X_N, without fixed points. This proves (b), and then (v) is a direct consequence of (a) and (b) together.

We now apply (v) to prove the Theorem. We are given that $B \subseteq X_U$ is compact. Hence

given $\mathfrak{g} \in \mathfrak{G}$ *there exists an integer* $j = j(\mathfrak{g}, W)$ *such that* (vi)

$$\alpha(\mathfrak{g}_m \cdot (\mathfrak{g} \cdot Q)) \cap B = \emptyset$$

provided $m > j$.

(Otherwise for some fixed $\mathfrak{g} \in \mathfrak{G}$ and for each integer i, there exists $m = m(i) > i$, $g_m \in K$ and $x_i \in \mathfrak{g} \cdot Q$ such that $y_i = \alpha(\mathfrak{g}_m \cdot x_i) \in B$; but the different products $\mathfrak{g}\,\mathfrak{g}_m$ are incongruent mod $\mathfrak{U}$, so by (b) above the sets $\alpha(\mathfrak{g}_{m(i)} \cdot \mathfrak{g} \cdot Q)$ lie in mutually disjoint sets $P_{k(i)}$ where $P_{k(r)} \neq P_{k(s)}$ if $r \neq s$. Hence the points y_i form an infinite set without limit point, contrary to the fact that $\{y_i\}$ lies in the compact set B.)

Finally, consider the given compact $A \subseteq X_N$. It can be covered by a finite family of open sets of the form $\mathfrak{g}_{rs} \cdot Q_r = (\mathfrak{g}_{n(r,s)} \mathfrak{u}_{rs}) \cdot Q_r$, where $g_{n(r,s)} \in K$, the Q_r corresponding to different open sets W_r in X. Then by (a),

$$\alpha(\mathfrak{g}_{rs} \cdot Q_r) = \alpha(\mathfrak{g}_{n(r,s)} \cdot Q_r)$$

and so by (vi) we consider

$$j_* = \max_{r,s} \, (j(\mathfrak{g}_{n(r,s)}, W_r)).$$

By definition of the j's, if $m > j_*$, then

$$\begin{aligned} \alpha(\mathfrak{g}_m \cdot A) \cap B &\subseteq [\alpha(\mathfrak{g}_m \cdot \bigcup_{r,s} (\mathfrak{g}_{rs} \cdot Q_r)] \cap B \\ &\subseteq [\bigcup_{r,s} \alpha(\mathfrak{g}_m \cdot \mathfrak{g}_{rs} \cdot Q_r)] \cap B \\ &= \bigcup_{r,s} [\alpha(\mathfrak{g}_m \cdot \mathfrak{g}_{rs} \cdot Q_r) \cap B] \\ &= \emptyset \text{ by (vi).} \end{aligned}$$

Thus $\{\mathfrak{g}_m, m > j_*\}$ will do for the infinity of $\mathfrak{g} \in \mathfrak{G}$ required in 5.2.

This completes the proof of Theorem 5.2.

6. Generalisation of Schreier's theorem

In this section we generalise the theorem of Schreier, mentioned in the introduction, and include also the generalisation of Karrass and Solitar [3].

6.1. THEOREM. *Let* $S \neq K$ *be a non-abelian surface with base-point* ω. *Let* $N \subseteq U \subseteq P = \pi_1(S, \omega)$ *be non-trivial groups, such that* U *is finitely generated and* N *is normal in* P. *Then* U *is of finite index in* P.

Proof. We can suppose that $\partial S = \emptyset$, since $\pi_1(S - \partial S) \approx \pi_1(S)$. If we deny 6.1, then $[P : U] = \infty$. Let

$$S_N \xrightarrow{\alpha} S_U \xrightarrow{\beta} S$$

be the associated covering spaces of N and U relative to the base-point $\omega \in S$; thus $\omega_U \in \beta^{-1}(\omega), \omega_N \in \alpha^{-1}(\omega_U)$ are the base-points of S_U, S_N. Certainly, since $[P:U]=\infty$, then S_U, S_N are both open surfaces, using 4.1 and the above assumption that $\partial S = \emptyset$. Therefore, by 3.4, there is a compact half-space $B \subseteq S_U$ such that the injection $j:\pi_1(B) \rightarrow \pi_1(S_U)$ is an isomorphism. All the fundamental groups involved are countable, by a remark preceding 4.1, so in the notation of 5.2 (ii), $\mathfrak{G}/\mathfrak{U}$ is countably infinite, since it has cardinal $[P:U]$. Now $N \neq \{1\}$, by hypothesis, so N has at least two independent generators, by 4.5, and these can be represented by ω_U-based loops λ, μ in S_U. Since j above is an isomorphism, we can assume λ, μ to be in B. They lift into ω_N-based loops λ', μ' in a compact subset A of $\alpha^{-1}(B) \subseteq S_N$. Hence we can apply 5.2, to find a covering transformation $g \in \mathfrak{G}$ such that $\alpha(gA) \cap B = \emptyset$. Therefore, setting $\omega' = \alpha(g(\omega_N))$, the ω'-based loops $\alpha(g\lambda')$, $\alpha(g\mu')$ lie in the same component K of $S_U - B$. But by choice of B as a half-space, we can invoke 3.6, to assert that $\pi_1(K)$ is cyclic infinite; thus there exist integers p, q such that $(\alpha(g\lambda'))^p \simeq (\alpha(g\mu'))^q$ rel ω' in K. By the lifting homotopy theorem, $(g\lambda')^p \simeq (g\mu')^q$ rel $g(\omega_N)$ in S_N, so $\lambda'^p \simeq \mu'^q$ rel ω_N in S_N since g is a homeomorphism. Hence $\lambda^p \simeq \mu^q$ rel ω in B since j above is an isomorphism. But then λ, μ cannot represent independent generators of N, and we have a contradiction.

This completes the proof.

Example. The commutator subgroup of P is of infinite index, hence free, since $H_1(S)$ is infinite if S is non-abelian.

We conclude by recording the following theorem, whose proof is identical with that of Theorem 2 of Karrass and Solitar [3].

6.2. Theorem. *Let U be a finitely generated subgroup of $P = \pi_1(S)$, S non-abelian, $S \neq K$. Then $[P:U] < \infty$ if and only if U contains the (normal) subgroup N_d of all d-th powers of elements of P, for some d. (Then $d = [P:U]$!)*

Added in proofs. While the paper was in the press, the author's attention was drawn to the paper by L. Greenberg, "Discrete Groups of Motions", *Canadian J. Math.*, 12 (1960), 415–426. There, Theorem 4 includes our Theorem 6.1, but it is proved by quite different methods. However, a conversation with A. M. Macbeath showed how to extend our 6.1 to the case when P is what Greenberg calls a "non-quasi-abelian" group of motions of the hyperbolic plane, as follows. The result of Fox quoted on p. 415 of Greenberg's paper shows that P contains a subgroup Q of finite index, and such that no transformation in Q has fixed points; hence Q is the fundamental group of a surface, which is non-abelian since P is non-quasi-abelian. Hence 6.1 can be applied directly to $U \cap Q$, $N \cap Q$, the latter being non-trivial since P contains no finite normal subgroup $\neq 1$. Thus $U \cap Q$ has finite index in Q. whence $[Q: U]$ is finite. More generally, the argument applies to any group P without finite non-trivial normal subgroups and possessing a subgroup like Q. It is hoped to extend these ideas in a later paper.

References

[1]. Ahlfors, L. V. & Sario, L., *Riemann Surfaces*, Princeton, 1960.

[2]. Hilton, P. J. & Wylie, S., *Homology Theory*, Cambridge, 1960.

[3]. Karrass, A. & Solitar, D., Note on a Theorem of Schreier. *Proc. Amer. Math. Soc.*, 8 (1957), 696.

[4]. Morse, M., *The Calculus of Variations in the Large*. A. M. S. Colloqu. Publications, Vol. XVIII.

[5]. Morse, M., The existence of non-degenerate polar functions on manifolds. *Ann. of Math.*, 71 (1960), 352–383.

[6]. Morse, M. & Van Schaak, G. B., The critical point theory under general boundary conditions. *Ann. of Math.*, 35 (1934), 545–571.

[7]. Nielsen, J., Über fixpunktfreie topologischer Abbildungen geschlossener Flächen. *Math. Ann.*, 81 (1920), 94–96.

[8]. Schreier, O., Die Untergruppen der freien Gruppen. *Abh. Math. Sem. Univ. Hamburg*, 5 (1928).

Received July 17, 1962