THE DOUBLE EXPONENTIAL THEOREM
FOR ISODIAMETRIC AND ISOPERIMETRIC FUNCTIONS

S. M. Gersten

§1. Isodiametric and Isoperimetric Functions

We shall give a new proof of the following result, first proved by D. E. Cohen [Co] by making use of results of Avenhaus and Madlener [A-M].

Theorem 2.5. *If $\mathcal{P}$ is a finite presentation and if $f : \mathbb{N} \to \mathbb{N}$ is an isodiametric function for $\mathcal{P}$, then there are constants $a, b > 0$ so that $n \mapsto a^{b^{f(n)+n}}$ is an isoperimetric function for $\mathcal{P}$.*

The relevant definitions of terms occurring in the statement are as follows. If $\mathcal{P} = \langle x_1, x_2, \dots, x_p \mid R_1, R_2, \dots R_q \rangle$ is a finite presentation, we shall denote by $G = G(\mathcal{P})$ the associated group; here $G = F/N$, where F is the free group freely generated by the generators $x_1, \dots, x_p$ and N is the normal closure of the relators. If w is an element of F (which we may identify with a reduced word in the free basis), we write $\ell(w)$ for the length of the word w and we write $w =_G 1$ to mean that w represents the identity element of G. We shall make use of the terminology of van Kampen diagrams [L-S, p. 235ff].

We write $\mathrm{Area}_{\mathcal{P}}(w)$ for the minimum number of faces (i.e 2-cells) in a van Kampen diagram with boundary label w. Equivalently, $\mathrm{Area}_{\mathcal{P}}(w)$ is the minimum number of relators or inverses of relators occurring in all expressions of w as a product (in F) of their conjugates. The function $f : \mathbb{N} \to \mathbb{N}$ is an *isoperimetric function* for $\mathcal{P}$ if, for all n and all words w with $\ell(w) \leq n$ and $w =_G 1$, we have $\mathrm{Area}_{\mathcal{P}}(w) \leq f(n)$. The minimal isoperimetric function for $\mathcal{P}$ is called its Dehn function.

If $\mathcal{D}$ is a van Kampen diagram with boundary label w, we choose the base point v_0 in the boundary of $\mathcal{D}$ corresponding to where one starts reading the boundary label w and one defines

$$\mathrm{Diam}_{v_0}(\mathcal{D}) = \max_{v \in \mathcal{D}^{(0)}} d_{\mathcal{D}^{(1)}}(v_0, v).$$

1991 *Mathematics Subject Classification.* 20E06, 20F05.

Key words and phrases. isoperimetric, isodiametric, finitely presented group.

This research was partially supported by NSF grant DMS 882-1749

Typeset by $\mathcal{A}\mathcal{M}\mathcal{S}$-TEX

Here $d_{\mathcal{D}^{(1)}}$ denotes the word metric for the 1-skeleton of $\mathcal{D}$, so that every edge has length 1. The function $f : \mathbb{N} \to \mathbb{N}$ is called an *isodiametric function* for $\mathcal{P}$ if, for all n and all reduced words w with $\ell(w) \leq n$ and $w =_G 1$, there exists a based van Kampen diagram $(\mathcal{D}, v_0)$ for w with $\mathrm{Diam}_{v_0}(\mathcal{D}) \leq f(n)$. An algebraic consequence of this is as follows. Let M denote the maximum length of a relator of $\mathcal{P}$. Let f be an isodiametric function for $\mathcal{P}$. If $w =_G 1$, then one can write in F,

$$w = \prod_{i=1}^{m} R_{j_i}^{\epsilon_i u_i},$$

where R_{j_i} is a relator, $\epsilon_i = \pm 1$, $u_i \in F$ and $\ell(u_i) \leq f(\ell(w)) + M$. Here we write $a^b = bab^{-1}$ for elements a and b in a group.

It is an elementary result [Ge, Lemma 2.2] that if f is an isoperimetric function for $\mathcal{P}$, then $n \mapsto Mf(n) + n$ is an isodiametric function. Theorem 2.5 provides a formula for an isoperimetric function, given an isodiametric function.

Remark. Isoperimetric functions were introduced into group theory by Gromov [Gr]. We introduced the term “Dehn function” in [Ge2] and developed techniques for calculating it. We also introduced the term “isodiametric function” [Ge1][Ge3] to generalize some conditions Casson had considered in connection with the Todd-Coxeter process in his study of when the universal cover of a closed aspherical irreducible 3-manifold is $\mathbb{R}^3$ [S-G]. It turned out subsequently that same notion had been considered earlier in [F-H-L] under a different name, as an alternative method of formulating the word problem for groups.

§2. The Double Exponential Theorem

Let S be a finitely generated subgroup of the finitely generated free group F, the latter equipped with a free basis $x_1, x_2, \ldots, x_p$. Let $y_1, y_2, \ldots, y_s$ be a set of generators for S. Then Stallings has shown how to produce effectively a free basis for the subgroup S [St] by a direct geometric method of folding; the result that S is free was first proved by Nielsen by a different algorithm [L-S, pp 4–7]. This last method was analyzed in terms of complexity theory in [A-M] and this analysis is the basis of Cohen's proof. We shall in contrast analyze Stallings' folds and thereby obtain the result. There are several reasons for doing this; Stallings' approach by folds gives a new way of looking at free groups which has already yielded new and powerful results (compare [B-H]). This approach had not previously been used to analyze Nielsen transformations (compare the proof of Lemma 2.2 below), and it seemed appropriate to make such an analysis in conjunction with obtaining a direct and geometric proof of Cohen's theorem.

Stallings represents F as $\pi_1(\Delta)$, where Δ is a 1-vertex graph, and represents the map $S \to F$ as the induced homomorphism on π_1 of a map $h : \Gamma \to \Delta$, where Γ is a bouquet of s-subdivided circles, where the i^{th} circle is mapped according to the word y_i; here the base point of Γ is the unique branch point of the graph. He then factors the map h as a composition of maps

$$\Gamma = \Gamma_1 \to \Gamma_2 \to \cdots \to \Gamma_n \to \Delta, \tag{1}$$

where the last map $\Gamma_n \to \Delta$ is an immersion and where each map $\Gamma_i \to \Gamma_{i+1}$ folds precisely one pair of edges.

We recall here that a fold $h : \Gamma \to \Gamma'$ of graphs is determined by two oriented edges $e, e' \in \Gamma$ with $\partial_0 e = \partial_0 e'$ and such that $e' \neq \bar{e}$, where $\bar{e}$ is the oppositely oriented edge to e, and is such that h is the quotient map identifying e with e'. Thus h makes only those identifications forced by this one; in particular one has $h(\partial_1 e) = h(\partial_1 e')$.

Choose a maximal tree T_n in Γ_n. This gives rise to a free basis for $\pi_1(\Gamma_n)$, where the base point is the image of that of Γ in the composition (1), by the familiar construction of choosing an oriented set of edges of $\Gamma_n - T_n$ and connecting these to the base point by geodesics in T_n. The image of this basis in $\pi_1(\Delta)$ is the desired free basis for S, since all folds are surjective on π_1 and since the immersion is injective on π_1. We call this basis a Stallings basis for S (note that it depends on the choice of folds and on T_n).

Lemma 2.1. *Given a fold $h : \Gamma \to \Gamma'$ of finite connected graphs, folding the oriented edges e, e', and given a maximal tree T' in Γ', there is a maximal tree T in Γ such that h maps the edges of $T - \{e, \bar{e}, e', \bar{e}'\}$ bijectively onto those of $T' - \{h(e), h(\bar{e})\}$.*

Proof. The argument is a case-by-case analysis, where the cases are determined by coincidences among the vertices $\partial_0 e, \partial_1 e$, and $\partial_1 e'$ and whether or not $h(e) \in T'$. The only case that needs careful consideration is when the set $\partial_0 e, \partial_1 e, \partial_1 e'$ consists of 3 distinct vertices and $h(e) = h(e') \notin T'$. Let p' be the geodesic in T' from $\partial_0 h(e)$ to $\partial_1 h(e)$ in this case and lift p' to a path p in Γ beginning at $\partial_0 e$. This is possible since p' is an arc which meets $\partial_1 h(e)$ only at its terminal point. We have $\partial_1 p \in \{\partial_1 e, \partial_1 e'\}$. We may assume without loss of generality that $\partial_1 p = \partial_1 e'$. Define T to be the smallest subgraph of Γ containing the edges of $h^{-1}(T' - \{h(e), h(\bar{e})\}) \cup \{e, \bar{e}\}$. One checks that T is a maximal tree of Γ.

In the situation of Lemma 2.1, choose a base point $\star$ in Γ and take its image under h as the base point for Γ' (the image will also be denoted $\star$). The trees T, T' and base points determine free bases for $\pi_1(\Gamma), \pi_1(\Gamma')$.

Lemma 2.2. *In this situation, each free basis element of $\pi_1(\Gamma')$ is the product of at most 3 of the images of basis elements of $\pi_1(\Gamma)$ (or their inverses).*

Proof. The only 'difficult' cases are when $h(e) \notin T'$ but one of $e, e' \in T$, for in all other situations each basis element of $\pi_1(\Gamma')$ is the image of precisely one basis element of $\pi_1(\Gamma)$. We will work out one of these 'difficult' cases.

Assume then that e, e' are the oriented edges which are folded and $e \in T$ but $h(e) \notin T'$ (note that $e' \notin T$ in this situation, as the argument for Lemma 2.1 shows). Let x be an oriented edge of Γ which is not in T, so the corresponding basis element of $\pi_1(\Gamma)$ is uxu', where $u = [\star, \partial_0 x]_T,\ u' = [\partial_1 x, \star]_T$ and where $[v, v']_T$ denotes the geodesic in T between the vertices v, v'.

There are several cases, depending on whether e or $\bar{e}$ occur in u or u'. We shall consider in detail the case where e, but not $\bar{e}$, occurs in u, whereas neither e nor $\bar{e}$ occurs in u', the other cases involving similar computations. Then the basis element determined by $h(x)$ in $\pi_1(\Gamma')$ is $\gamma := [\star, \partial_0 h(x)]_{T'} \cdot x \cdot h(u')$. But

$$[\star, \partial_0 h(x)]_{T'} \sim h([\star, \partial_1 e']_T \cdot \bar{e}' \cdot [\partial_0 e', \star]_T \cdot [\star, \partial_0 e]_T \cdot e \cdot [\partial_1 e, \partial_0 x]_T),$$

since $h(e) = h(e')$ and since the path $[\partial_0 e', \star]_T$ is inverse to the path $[\star, \partial_0 e]_T$, neither containing e; here "$\sim$" denotes end point fixed homotopy. Since $[\star, \partial_0 e]_T \cdot e \cdot [\partial_1 e, \partial_0 x]_T \sim [\star, \partial_0 x]_T$, it follows that $\gamma = h(\alpha)h(\beta)$, where $\alpha = [\star, \partial_1 e']_T \cdot \bar{e}' \cdot [\partial_0 e', \star]_T$ and $\beta = [\star, \partial_0 x]_T \cdot x \cdot [\partial_1 x, \star]_T$. Thus, in this case γ is the product of the images under h of two basis elements (or their inverses) of $\pi_1(\Gamma)$.

If both u and u' had contained one of the pair of edges $e, \bar{e}$, we would have had γ expressed as the product of 3 such images of basis elements (or their inverses). It follows that in all cases, a basis element of $\pi_1(\Gamma')$ is the product of at most 3 images of basis elements (or their inverses) from $\pi_1(\Gamma)$. This completes the proof of Lemma 2.2.

We now return the situation of representing the subgroup $S < F$ by means of folds and an immersion. We may now use Lemma 2.1 to find maximal trees T_i in Γ_i satisfying the condition of the conclusion there for each fold $\Gamma_i \to \Gamma_{i+1}$. The base point for Γ_i is taken as the image of that of Γ_1. This gives rise to bases for all the groups $\pi_1(\Gamma_i)$.

Lemma 2.3. *Each Stallings basis element of S is the product of at most 3^n of the original generators $y_1, y_2, \ldots, y_s$ (or their inverses).*

Proof. This follows from Lemma 2.2 by an induction.

Lemma 2.4. *Let the word w in the generators $x_1^{\pm 1}, x_2^{\pm 1}, \ldots, x_p^{\pm 1}$ for F represent an element of the subgroup S. Then w is the product of at most $\ell(w)$ Stallings basis elements (or their inverses).*

Proof. We may assume that w is reduced, since reduction only decreases the number of letters in a word. Represent w as a reduced path in the 1-vertex graph Δ. Since $w \in S$, this path is the image of a unique reduced *circuit* u in Γ_n at the base point. Note that $\ell(w) = \ell(u)$. The homotopy class of u in $\pi_1(\Gamma_n)$ can be written as the product of k Stallings basis elements (or their inverses), where k is the number of those edges of u which are *not* in T_n. Thus $k \leq \ell(u) = \ell(w)$ and the proof is complete.

Theorem 2.5. *If $\mathcal{P}$ is a finite presentation and if $f : \mathbb{N} \to \mathbb{N}$ is an isodiametric function for $\mathcal{P}$, then there are constants $a, b > 0$ so that $n \mapsto a^{b^{f(n)+n}}$ is an isoperimetric function for $\mathcal{P}$.*

Proof. Let $\mathcal{P} = \langle x_1, x_2, \ldots, x_p \mid R_1, R_2, \ldots, R_q \rangle$ be a finite presentation for the group $G = G(\mathcal{P})$ and let C be the Cayley graph. Observe that the number of vertices in the ball of radius r center at 1 in C is at most $(2p)^r$. Let f be an isodiametric function for $\mathcal{P}$ and let w be a word in the generators with $\ell(w) = n$ and such that w represents the identity in G. By definition we have $w \in N_{f(n)}$, where $N_k < N \lhd F$, and N is the normal closure of the relators and N_k is the subgroup generated by all elements $R_i^u = uR_iu^{-1}$, $1 \leq i \leq q$, with $\ell(u) \leq k$. Thus the number of generators for $N_{f(n)}$ is at most $B^{f(n)}$, where B is a constant.

By Lemma 2.3, each Stallings generator for the subgroup $N_{f(n)}$ is a product of at most 3^{sM} of the original generators for the subgroup, where $s \leq B^{f(n)}$ and where M is the maximum length of one of the generators R_i^u of $N_{f(n)}$. The reason is that at most sM folds are needed in the Stallings folding algorithm. Note that $M \leq Df(n) + E$, where D, E are constants.

It follows that each Stallings basis element for $N_{f(n)}$ is the product of at most $3^{(Df(n)+E)B^{f(n)}}$ of the original generators for $N_{f(n)}$ (or their inverses). On the other hand, by Lemma 2.4, $\ell(w)$ is an upper bound for the number of Stallings basis elements (or their inverses) needed in a factorization of w. It follows that

$$\mathrm{Area}_{\mathcal{P}}(w) \leq n \cdot 3^{(Df(n)+E)B^{f(n)}} \leq a^{b^{f(n)+n}},$$

where a, b are constants. This completes the proof.

Remark. The reader will have noticed that the sum $f(n) + n$ only appears at the last step. It appears to be necessary, since we do not know whether an isodiametric function can grow sublinearly. Indeed, unlike the corresponding situation with isoperimetric functions (where a finitely presented group is word hyperbolic if and only if it possesses a linear isoperimetric function), we do not know how to characterize those groups possessing linear isodiametric functions; these include many of the groups that occur naturally in geometry and in combinatorial group theory [Ge1].

Remark. If f is an isodiametric function for $\mathcal{P}$, it is reasonable to ask whether there is an isoperimetric function of the form $n \mapsto a^{f(n)+n}$. Easy examples, like $\langle x, y \mid yx = x^2y \rangle$ (which possesses a linear isodiametric function, but whose Dehn function is exponential), show that this is the best one can expect in general.

Remark. It is shown in [A-M] that the complexity of the Nielsen reduction algorithm is bounded by a polynomial function $p(s, M)$ of the number s of generators and the maximal length M of a generator of S. Cohen deduces from this that the area of a Nielsen basis element is bounded by $2^{p(s,M)}$. In our case, the complexity of

Stallings' folding algorithm is at most sM (once the trees T_i are determined), from which we deduce that the area of a Stallings basis element is at most 3^{sM}. It would appear then that the two procedures are not the same and it might be of interest to clarify their connection. We suspect it is analogous to the relationship between Whitehead automorphisms and Nielsen automorphisms.

§3. Remarks on Whitehead Automorphisms

We return to the general situation described in Lemmas 2.1 and 2.2 of a single fold $h : \Gamma \to \Gamma'$. If h is a homotopy equivalence, it induces a bijection of edges of $\Gamma - T$ with those of $\Gamma' - T'$. This makes it possible to identify $\pi_1(\Gamma)$, $\pi_1(\Gamma')$ with the free group F with basis $x_1, x_2, \ldots, x_p$ (where p is the rank of $\pi_1(\Gamma)$) as follows. Choose an orientation $\mathcal{O}$ on the oriented edges of $\Gamma - T$, so $|\mathcal{O} \cap \{\eta, \bar{\eta}\}| = 1$ for each oriented edge η of $\Gamma - T$ and choose an ordering of $\mathcal{O}$. Let $\mathcal{O}' = h(\mathcal{O})$ with the induced ordering. Then $\mathcal{O}, \mathcal{O}'$ determine bases A, A' for $\pi_1(\Gamma), \pi_1(\Gamma')$ with respect to the base points and trees T, T'. This gives "markings" $F \to \pi_1(\Gamma), F \to \pi_1(\Gamma')$, so that $\{x_1, x_2, \ldots, x_p\}$ maps to A by an order preserving bijection, and similarly for A'. In terms of these markings, a calculation as in the proof of Lemma 2.2[†] shows that the isomorphism $h_* : \pi_1(\Gamma) \to \pi_1(\Gamma')$ corresponds to the Whitehead automorphism $W(h)$ of F of type 2; that is, there is a commutative diagram

$$\begin{tikzcd}
F \arrow[r, "\cong"] \arrow[d, "W(h)"] & \pi_1(\Gamma) \arrow[d, "h_*"] \\
F \arrow[r, "\cong"] & \pi_1(\Gamma').
\end{tikzcd}$$

The "multiplier" of $W(h)$ (in the terminology of [L-S, p. 31]) is determined as follows. The base point $\star$ of T determines a partial order on the vertices, where $v < v'$ for v, v' vertices of T if and only if $v \in [\star, v']_T$. The multiplier of $W(h)$ is the element of $\{x_1^{\pm 1}, x_2^{\pm 1}, \ldots, x_p^{\pm 1}\}$ corresponding to e' (resp. $\bar{e}'$) if $\partial_1 e < \partial_0 e$ (resp. $\partial_0 e < \partial_1 e$) (observe that since $e \in T$, one must have here either $\partial_0 e < \partial_1 e$ or $\partial_1 e < \partial_0 e$).

Observe that an immediate consequence of this discussion is the fact, which is usually deduced from Nielsen's results [L-S], that the Whitehead automorphisms generate $\mathrm{Aut}(F)$. The immersion at the last step in the folding process, applied to the graph for the image of a free basis for F under an automorphism, is an isomorphism of graphs, and hence is a Whitehead automorphism of type 1 (i.e. a signed permutation of the given free basis for F).

We can also make some deductions about the word metric on $\mathrm{Aut}(F)$. Let A denote the set of Whitehead automorphisms, which we have just seen form a (finite, symmetric) set of generators, and let Λ be the corresponding Cayley graph,

---

[†] In effect, what we were doing in Lemma 2.2 was calculating the inverse of h_*.

equipped with the word metric. We denote the distance of an automorphism ϕ from the identity by $||\phi||$. In addition, let $c(\phi) = \sum_{i=1}^{p} \ell(\phi(x_i))$, which we have seen is a bound on the complexity of the folding algorithm applied to the generators $\{\phi(x_i) \mid 1 \leq i \leq p\}$ for F.

Proposition 3.1. *One has for all $\phi \in F$ the inequalities*

$$||\phi|| \leq c(\phi) \leq p \cdot 3^{||\phi||}.$$

Proof. We may assume that $\operatorname{rank}(F) = p \geq 1$. Observe that if we take a bouquet of subdivided circles for the generators $\phi(x_i)$, $1 \leq i \leq p$, for F, then after folding we must be left with exactly p circles in the one-vertex graph Δ. This means that the number of folds is exactly $c(\phi) - p \leq c(\phi) - 1$. From the remarks at the beginning of this section, we deduce that ϕ can be expressed as the product of at most $c(\phi)$ Whitehead automorphisms, where at most $c(\phi) - 1$ of these are of type 2 and at most one is of type 1, in the terminology of [L-S]. This gives the inequality $||\phi|| \leq c(\phi)$.

For the second inequality, let $\phi = \alpha_1 \alpha_2 \ldots \alpha_n$, where $n = ||\phi||$ and where α_i are Whitehead automorphisms. We have $\ell(\alpha_i(x_j)) \leq 3$, whence $\ell(\phi(x_j)) \leq 3^n$, and $c(\phi) \leq p \cdot 3^{||\phi||}$. This completes the proof.

Remark. Simple examples show that an exponential bound in the second inequality in the preceding proposition is the best one can expect in general. The usual proof [M-K-S] that the Nielsen automorphisms generate $\operatorname{Aut}(F)$ does not give immediately an estimate on the word metric, although it should be possible to deduce one from the analysis of [A-M].

References

[B-H] M. Bestvina and M. Handel, *Train tracks and automorphisms of free groups*, to appear in Annals of Math.

[A-M] J. Avenhaus and K. Madlener, *The Nielsen reduction and P-complete problems in free groups*, Theoretical Computer Science **32** (1984), 61–76.

[Co] D. E. Cohen, *Isodiametric and isoperimetric inequalities for group presentations*, preprint, Queen Mary and Westfield College, London, 1991.

[F-H-L] W. J. Floyd, A. H. M. Hoare, and R. C. Lyndon, *The word problem for geometrically finite groups*, Geometriae Dedicata **20** (1986), 201–207.

[Ge1] S. M. Gersten, *Isoperimetric and isodiametric functions of finite presentations*, to appear in the proceedings of the conference on geometric group theory, University of Sussex, 1991.

[Ge2] ———, *Dehn functions and ℓ_1-norms of finite presentations*, Algorithms and classification in combinatorial group theory (G. Baumslag and C. F. Miller, III, eds.), MSRI series vol. 23, Springer-Verlag, 1991.

[Ge3] ———, *Isodiametric and isoperimetric inequalities in group extensions*, preprint, Univ. of Utah, 1991.

[Gr] M. Gromov, *Hyperbolic groups*, Essays in group theory (S. M. Gersten, ed.), MSRI series vol. 8, Springer-Verlag, 1987.

[L-S] R. C. Lyndon and P. E. Schupp, *Combinatorial group theory*, Springer-Verlag, 1977.

[M-K-S] W. Magnus, A. Karrass, and D. Solitar, *Combinatorial group theory*, Interscience (John Wiley), 1966.

[St] J. R. Stallings, *Topology of finite graphs*, Inv. Math. **71** (1983), 551–565.

[S-G] J. R. Stallings and S. M. Gersten, *Casson's idea about 3-manifolds whose universal cover is* $\mathbb{R}^3$, to appear in Int. Jour. Alg. and Computation.

Mathematics Department, University of Utah, Salt Lake City, UT 84112
E-mail address: `gersten@math.utah.edu`