Isoperimetric Functions of Groups and Computational Complexity of the Word Problem

J.-C. Birget, A. Yu. Ol'shanskii,
E. Rips, M. V. Sapir*

Abstract

We prove that the word problem of a finitely generated group G is in NP (solvable in polynomial time by a non-deterministic Turing machine) if and only if this group is a subgroup of a finitely presented group H with polynomial isoperimetric function. The embedding can be chosen in such a way that G has bounded distortion in H. This completes the work started in [26] and [24].

Contents

*The research of the first and the fourth authors was supported in part by an NSF grant. The research of the second author was supported in part by the Russian fund for fundamental research 96-01-420.

1 Introduction

A function $f : \mathbf{N} \to \mathbf{N}$ is called an *isoperimetric function* of a finite presentation $\mathcal{P} = \langle X \mid R \rangle$ of a group G if for every number n and every word w over X which is equal to 1 in G, $|w| \leq n$, there exists a van Kampen diagram over $\mathcal{P}$ whose boundary label is w and area $\leq f(n)$; in other words, w is a product in the free group of at most $f(n)$ conjugates of the relators from R. One can also define an isoperimetric function of a presentation $\mathcal{P}$ as an isoperimetric function of its Cayley complex (the Cayley graph equipped with 2-cells corresponding to relations of $\mathcal{P}$), that is a function bounding the filling area of a loop of a given perimeter in the Cayley complex (see [10], [11], [21], [7], [13], [4]). Isoperimetric functions of manifolds and 2-complexes were studied since ancient Greeks. Dehn probably was the first to notice a connection between them and the word problem in groups (this is how he solved the word problem in surface groups). But the concept of an isoperimetric function for groups (at least the name of this concept) probably first appeared in [10], [11].

The smallest isoperimetric function of a finite presentation $\mathcal{P}$ is called the Dehn function of $\mathcal{P}$ [7].

Let $f, g : \mathbf{N} \to \mathbf{N}$ be two functions. We write $f \preceq g$ if there exist non-negative constants a, b, c, d such that $f(n) \leq ag(bn) + cn + d$. All functions $g(n)$ which we are considering in this paper grow at least as fast as n. In this case $f(n) \preceq g(n)$ if and only if $f(n) \leq ag(bn)$ for some positive constants a, b. Two functions f, g are called *equivalent* if $f \preceq g$ and $g \preceq f$.

It is well known that Dehn functions corresponding to different finite presentations of the same group are equivalent (see [21] or [2], [7]). This allows us to speak about *the Dehn function* of a finitely presented group.

It is also known [7] that the Dehn function of a finitely presented group is recursive if and only if the group has decidable word problem. Moreover, a subset of the set of authors of this paper has proved in [26] (see the proof of Theorem 1.1 of [26]) that for every finitely presented group G with Dehn function $T(n)$ there exists a nondeterministic Turing machine $M(G)$ which solves the word problem in G and has time function equivalent to $T(n)$. Roughly speaking, this Turing machine takes a word w over the generators of G and just inserts relators of G. It stops and accepts w when it gets 1. Clearly this machine solves the word problem in every finitely generated subgroup of G as well. Therefore if a finitely generated group G is a subgroup of a finitely presented group with polynomial isoperimetric function then the word problem in G is in NP (i.e. it can be solved by a non-deterministic Turing machine with polynomial time function).

The drawback is that the word problem in a group G can be easy to solve but the Dehn function of G can be huge. A typical example is the Baumslag-Solitar group $G_{2,1} = \langle a, t \mid a^t = a^2 \rangle$ where $a^t = t^{-1}at$. This group has exponential Dehn function [13]. But this group is a subgroup of $GL(2, \mathbf{Q})$ ($t = \begin{pmatrix} 1/2 & 0 \\ 0 & 1 \end{pmatrix}$, $a = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$), so the word problem in $G_{2,1}$ can be solved in at most quadratic time: it is easy to see that the word problem of every finitely generated matrix group over a field of rational numbers can be solved in at most quadratic time by a deterministic Turing machine. In fact it is possible to solve the word problem there in time $n(\log n)^2(\log\log n)$ (this follows from the fact that the product of two n-digit numbers can be computed in time $n \log n \log\log n$ [18]). Using the construction in [26] Sapir showed (unpublished) that there exists a finitely presented group with word problem in NP and Dehn function not even primitive recursive.

Nevertheless, in this paper we prove that every finitely generated group G with word problem in NP can be embedded into a finitely presented group H with polynomial isoperimetric function. Thus if we can solve the word problem in G using a very clever and fast Turing machine, then we can use the simple-minded but almost as fast Turing machine $M(H)$ to solve the word problem in G. Moreover we prove that the group G can have *bounded length distortion* in H, that is the word length of any element of G in G is within a constant factor of the word length of this element in H. Here is the main result of this paper.

Theorem 1.1 *Let G be a finitely generated group with word problem solvable by a non-deterministic Turing machine with time function $\leq T(n)$ such that $T(n)^4$ is superadditive (that is $T(m+n)^4 \geq T(m)^4 + T(n)^4$ for every m, n). Then G can be embedded into a finitely presented group H with isoperimetric function equivalent to $n^2T(n^2)^4$ in such a way that G has bounded distortion in H. In particular, the word problem of a finitely generated group is in NP if and only if this group is a subgroup of a finitely presented group with polynomial isoperimetric function.*

Notice that this theorem strengthens the main result of [24] where it is proved that every finitely generated recursively presented group G can be embedded without distortion into a finitely presented group.

The class of finitely generated groups with word problem in NP is very large. We have mentioned that it includes all matrix groups over $\mathbf{Q}$. It also includes

- All finitely generated matrix groups over arbitrary fields: this follows from the fact that every finitely generated field is a finite extension of a purely transcendental extension of its simple subfield, and the fact that the word problem in the ring of polynomials over $\mathbf{Q}$ or $\mathbf{Z}/p\mathbf{Z}$ is solvable in polynomial time,

- Polycyclic and finitely generated metabelian groups because they are representable by matrices [27],

- Automatic groups (in particular, hyperbolic groups) [6],

- Groups of piecewise linear transformations of a line with finitely many rational singularities (including the R. Thompson group F),

- Every finitely generated subgroup of a diagram group [15],

- Every free Burnside group $B_{m,n}$ for sufficiently large odd exponent n (see, for example, Storozhev's argument in Section 28 of [23]).

This class is closed under free and direct products. It is easy to see using Magnus' embedding that for every normal subgroup N of a free finitely generated group F if F/N has word problem in NP (resp. P) then F/N' has word problem in NP (resp. P). Therefore every free group in the variety of all solvable groups of a given class has word problem in P. It is an interesting question whether this class also contains all one-relator groups. There are of course finitely generated groups with word problem not in NP, for example groups with undecidable word problem. But these groups and in some sense "artificial". So perhaps the class of groups with word problem in NP (which by Theorem 1.1 is the class of all subgroups of finitely presented groups with polynomial Dehn functions) can be considered as the class of "tame" groups.

An example of an embedding of one group into another where lengths are not distorted but areas are distorted can be found in [9]. Some examples of groups with big Dehn functions embeddable into groups with small Dehn functions can be found in [21] and [3]. Our theorem and Theorem 1.1 from [26] show that any finitely generated group can be embedded into a finitely presented group with bounded length distortion but with close to maximal possible area distortion. Indeed, Theorem 1.1 from [26] shows that the isoperimetric function of a group H containing a given group G cannot be smaller than the time complexity $T(n)$ of the word problem for G, and we show that G can be embedded into a finitely presented group with Dehn function at most $T(n^4)^4$.

For matrix groups our theorem implies that every such group is embedded with bounded length distortion into a finitely presented group with Dehn function at most $n^{10+\epsilon}$ for every $\epsilon > 0$. It is interesting to know the smallest Dehn function of a finitely generated group containing, for example, the Baumslag-Solitar group $G_{2,1}$. Is it possible to lower this bound for $G_{2,1}$ to n^3 or even n^2?

Using the proof of Theorem 1.1, in order to embed a finitely generated group with word problem in NP into a finitely presented group with polynomial isoperimetric function, one needs first construct a Turing machine which solves the word problem, then convert it into an S-machine, then convert the S-machine into a group. As a result the group we construct will have a relatively complicated set of relations. In some particular cases like the Baumslag-Solitar group $G_{2,1}$, the free Burnside groups $B_{m,n}$ and others, the authors can modify this construction and get simple presentations of groups with polynomial isoperimetric functions where these groups embed.

Theorem 1.1 is also interesting from the logic point of view. One can consider a group as a logical system where the defining relations are axioms, and the inference rules are constructing step by step van Kampen diagrams. The van Kampen diagrams are then *proofs* of their boundary labels. Then the Dehn function becomes the syntactic complexity of proofs. The computational complexity of the word problem is the complexity in the metaworld of the group theory. The embedding of groups becomes a conservative extention of theories.

With this vocabulary, the result means that the complexity of proofs in the outer world of groups (that is, in its metamathematical semantics) after appropriate conservative extention to a finitely axiomatized theory becomes the complexity of proofs in the syntactical sense (up to a polynomial correction).

In this formulation, one can ask a similar question for any logical system! Several results of this type for general logical systems can be found in [19]. Notice also that the semigroup analog of Theorem 1.1 was obtained in [5].

2 Preliminaries

2.1 Turing Machines and S-machines

Let $T(n)$ be a function such that $T(n)^4$ is superadditive. Let G be a group with finite symmetric set of generators $A = A^{-1}$, such that the word problem of G is decidable in time at most $T(n)$ by a nondeterministic Turing machine M. By a Turing Machine we always assume a non-deterministic multi-tape Turing machine (see [16] or [26]). Each tape is finite but can be expanded by adding cells at the right end of it. One of these tapes is called the *input tape*. Each tape has its own alphabet, the set of letters which can appear on this tape. A *configuration* of a Turing machine is the sequence of words written on the tapes, position of the head relative to the tapes, and the state of the head. In the *input configuration* of this machine a word is written on the input tape and all other tapes are empty. We shall assume that M has only one *stop configuration* c_0 and all tapes in this configuration are empty. The fact that M solves the word problem in G means that the alphabet of the input tape is A and for every word u in the alphabet A the machine can *accept* the input configuration with u written on the input tape (that is get to the stop state starting with this input configuration) if and only if u is equal to 1 in G.

Let M be a **nondeterministic Turing** machine with time complexity $\leq T(n)$ that decides if a word in the alphabet A is equal to 1 in G. In Proposition 4.1 of [26], it was proved that M is polynomially equivalent to a so called S-machine $\mathcal{S}(M)$. Let us repeat the definition of an S-machine from [26] and then explain in what sense $\mathcal{S}(M)$ is equivalent to M (see Proposition 2.1).

Roughly speaking, the difference between S-machines and ordinary Turing machines is that S-machines are almost "blind". They "see" letters written on the tape only when these letters are between two heads of the machine and the heads are very close to each other. If the heads are far apart, the machine does not see any letters on the tape.

We define S-machines as rewriting systems [17]. Let k be a natural number. A *hardware* of an S-machine $\mathcal{S}$ is a pair (Y, Q) where Y is a k-vector of (not necessarily disjoint) sets Y_i, Q is a $(k+1)$-vector of disjoint sets Q_i, $\overline{Q} = \bigcup Q_i$ and $\overline{Y} = \bigcup Y_i$ are also disjoint. The elements of $\overline{Y}$ are called *tape letters*, the elements of $\overline{Q}$ are called *state letters*. All these letters together form the *alphabet* of the machine $\mathcal{S}$.

With every hardware $\mathcal{S} = (Y, Q)$ we associate the *language of admissible words* $L(\mathcal{S}) = Q_1F(Y_1)Q_2...F(Y_k)Q_{k+1}$ where $F(Y_j)$ is the language of reduced group words in the alphabet Y_j. Notice that in every admissible word, there is exactly one representative of each Q_i and these representatives appear in this word in the order of the indices of Q_i. The $\overline{Y}$-letters which appear between a Q_i-letter and a Q_{i+1}-letter in an admissible word belong to Y_i, $i = 1, 2, ..., k$.

If $0 \leq i \leq j \leq k$ and $W = q_1u_1q_2...u_kq_{k+1}$ is an admissible word then the subword

$q_iu_i...q_j$ of W is called the (Q_i, Q_j)-subword of W $(i < j)$.

An S-machine with hardware $\mathcal{S}$ is a rewriting systems. The objects of this rewriting system are all admissible words.

The rewriting rules, or S-*rules*, have the following form:

$$[U_1 \to V_1, ..., U_m \to V_m]$$

where the following conditions hold:

Each U_i is a subword of an admissible word starting with a Q_ℓ-letter and ending with a Q_r-letter (where $\ell = \ell(i)$ must not exceed $r = r(i)$, of course).

If $i < j$ then $r(i) < \ell(j)$.

Each V_i is also a subword of an admissible word whose Q-letters belong to $Q_{\ell(i)} \cup ... \cup Q_{r(i)}$ and which contains a Q_ℓ-letter and a Q_r-letter.

If $\ell(1) = 1$ then V_1 must start with a Q_1-letter and if $r(m) = k+1$ then V_n must end with a Q_{k+1}-letter (so tape letters are not inserted to the left of Q_1-letters and to the right of Q_{k+1}-letters).

To apply an S-rule to a word W means to replace simultaneously subwords U_i by subwords V_i, $i = 1, ..., m$. In particular, this means that our rule is not applicable if one of the U_i's is not a subword of W. The following convention is important:

After every application of a rewriting rule, the word is automatically reduced. We do not consider reducing of an admissible word a separate step of an S-machine.

For example, if a word is

$$q_1aaq_2bq_3ccq_4$$

and $q_i \in Q_i$, $a \in Y_1$, $b \in Y_2$, $c \in Y_3$ and the S-rule is

$$[q_1 \to p_1a^{-1}, q_2bq_3 \to a^{-1}p_2b'q_3c], \tag{1}$$

where $p_1 \in Q_1$, $p_2 \in Q_2$, $b' \in Y_2$, then the result of the application of this rule is

$$p_1p_2b'q_3cccq_4.$$

With every S-rule τ we associate the inverse S-rule τ^{-1} in the following way: if

$$\tau = [U_1 \to x_1V_i'y_1,\ U_2 \to x_2V_2'y_2, ..., U_m \to x_mV_m'y_m]$$

where V_i' starts with a $Q_{\ell(i)}$-letter and ends with a $Q_{r(i)}$-letter, then

$$\tau^{-1} = [V_1' \to x_1^{-1}U_1y_1^{-1},\ V_2' \to x_2^{-1}U_2y_2^{-1}, ..., V_m' \to x_m^{-1}U_my_m^{-1}].$$

For example, the inverse of the rule (1) is

$$[p_1 \to q_1a, p_2b'q_3 \to aq_2bq_3c^{-1}].$$

It is clear that τ^{-1} is an S-rule, $(\tau^{-1})^{-1} = \tau$, and that rules τ and τ^{-1} cancel each other (meaning that if we apply τ and then τ^{-1}, we return to the original word).

The following convention is also important:

We always assume that an S-machine is symmetric, that is if an S-machine contains a rewriting rule τ, it also contains the rule τ^{-1}.

When an S-machine works, it applies its rules to admissible words. The sequence of these admissible words and the rules is called a *computation* of the S-machine. We define the history of a computation of an S-machine as the sequence (word) of rules used in this computation. A computation is called *reduced* if the history of this computation is reduced, that is if two mutually inverse rules are never applied next to each other.

As usual the *length* of a computation $W_1, ..., W_n$ is n, and the *area* is $\sum_i |W_i|$.

The following statement is a immediate corollary of Lemma 3.1 and Proposition 4.1 from [26]. Notice that Proposition 4.1 from [26] contains many properties of $\mathcal{S}(M)$ which we are not going to explicitely use in this paper, so we do not present them here. But we are going to use lemmas from [26] which are based on these properties.

Proposition 2.1 *Let M be a non-deterministic Turing machine with time function $T(n)$ accepting language $L \subseteq A^+$. Then there exists an S-machine $\mathcal{S}(M)$ and a 1-1 function σ from the set of configurations of M to the set of admissible words of $\mathcal{S}(M)$ which satisfy the following properties. Let c_0 be the stop configuration of M and denote $\sigma(c_0)$ by W_0. Then*

1. *A configuration c of the machine M is acceptable by M if and only if $\mathcal{S}(M)$ can take $\sigma(c)$ to W_0.*

2. *If u is an input word for the machine M, $|u| = n$, and $c(u)$ is the corresponding input configuration of M then $\sigma(c(u)) = z_0\alpha^n z_1 u z_2 \delta^n z_3 \omega^n z_4$ for some fixed letters α, δ, ω not occurring in u and some $\overline{Q}$-words z_0, z_1, z_2, z_3, z_4 which do not depend on u, such that the contents of these words are disjoint and do not contain letters from u or α, δ, ω.*

3. *If an admissible word W contains the subwords z_0, z_1,z_2, z_3, z_4, and the S-machine $\mathcal{S}$ takes W to W_0 then $W = \sigma(c(u))$ for some group word u over A.*

2.2 Groups Simulating S-machines

Now let M be a Turing machine solving the word problem in our group G. We can assume that the set of defining relations of G contains all relations of the form $aa^{-1} = 1$, $a \in A$. Let $\mathcal{S} = \mathcal{S}(M)$. Then $\mathcal{S}$ also solves the word problem in G in the following sense. For every group word u in the alphabet A, the word u is equal to 1 in G if and only if the S-machine $\mathcal{S}$ takes the word $\sigma(c(u))$ to W_0 (in this case we say that $\mathcal{S}$ accepts $\sigma(c(u))$).

Let Θ be the set of all rules of $\mathcal{S}$. In every pair of mutually inverse rules from Θ we choose one which we call *positive*. The set of all positive rules of $\mathcal{S}$ is denoted by Θ_+.

Let $\overline{Y}$ be the set of all tape letters of the machine which can appear in the admissible words of S and let $\overline{Q}$ be the set of all state letters of $\mathcal{S}$. Notice that $A \subseteq \overline{Y}$. Let us fix a large natural number N. Up to Section 7 it will be enough to assume that $N \geq 9$, the exact value of N will be given in Section 7. With every admissible word W of the S-machine $\mathcal{S}$ we associate the following word

$$\kappa(W) = \kappa_1 W \kappa_2 W^{-1} \kappa_3 W ... \kappa_{4N-1} W \kappa_{4N} W^{-1}.$$

Remark 1. Notice that this word will turn into the word $K(W)$ from [26] if we identify κ_{2N+i} with κ_i^{-1}, $i = 1, ..., 2N$.

Now we can define a group $G_N(\mathcal{S})$ similar to the one defined in [26].

The set of generators of $G_N(\mathcal{S})$ is $\overline{Y} \cup \overline{Q} \cup \Theta_+ \cup \{\kappa_1, ..., \kappa_{4N}\}$. The set of relators of $G_N(\mathcal{S})$ is divided into three groups.

1. Transition relations. These relations correspond to elements of Θ_+.

Let $\tau \in \Theta_+$, $\tau = [U_1 \to V_1, ..., U_p \to V_p]$. Then we include relations $U_1^\tau = V_1, ..., U_p^\tau = V_p$ into $\mathcal{P}_N(\mathcal{S})$. Here x^y stands for $y^{-1}xy$. If the letters of one of the component of the vector Q, say, Q_j, do not appear in any of the U_i then also include the relations $q_j^\tau = q_j$ for every $q_j \in Q_j$.

2. Auxiliary relations.

These are all possible relations of the form $\tau x = x\tau$ where $x \in \overline{Y}$, $\tau \in \Theta_+$ and all commutativity relations of the form $\tau\kappa_i = \kappa_i\tau$, $i = 1, ..., 4N$, $\tau \in \Theta_+$.

3. The hub relation:

$$\kappa(W_0) = 1.$$

We also assume that all cyclic shifts and the inverses of these relations are in the set of defining relations of $G_N(\mathcal{S})$.

The remark we made above shows that the group $G_N(\mathcal{S})$ constructed here has a natural homomorphism onto the group $G_N(\mathcal{S})$ in [26]. This homomorphism is induced by the map $\kappa_{2N+i} \to \kappa_i^{-1}$. The following Proposition contains several properties of the group $G_N(\mathcal{S})$ from [26]. It is easy to check that all these properties are true for the group we deal with here. The proofs are exactly the same (almost identical), so we do not present them here. In fact we could avoid changing the presentation of $G_N(\mathcal{S})$ in this paper, but it would make the paper longer and more complicated, so we decided to make the change.

Remark 2. We can preserve these properties while changing the presentation of $G_N(\mathcal{S})$ even further: we can make different copies of the word W in the definition of $\kappa(W)$ written in disjoint alphabets. This is interesting because this would make the hub $\kappa(W_0)$ linear (each letter occurs only once). Then the group $G_N(\mathcal{S})$ would be an HNN extension of a free group with several stable letters, factorized by a linear relation. This may have nice geometric consequences. See also Remark 3 in Section 3.

Proposition 2.2 *The following properties of the group $G_N(\mathcal{S})$ hold:*

1. *The machine $\mathcal{S}$ takes an admissible word W to W_0 if and only if the word $\kappa(W)$ is equal to 1 in $G_N(\mathcal{S})$. This and part 1 of Proposition 2.1 imply that a word u in the alphabet A is equal to 1 in our original group G if and only if the word $\kappa(\sigma(c(u)))$ is 1 in $G_N(\mathcal{S})$.*

2. *The Dehn function of the group $G_N(\mathcal{S})$ is equivalent to $T(n)^4$.*

2.3 Van Kampen Diagrams, Bands and Annuli

In order to analyze van Kampen diagrams over group presentations we mainly use notation and definitions from Ol'shanskii [23], and from papers [26], [24].

In particular, we use auxiliary 0-edges, i.e. edges labeled by 1, and 0-cells, i.e. cells corresponding to the trivial relations of the form $a1^n a^{-1} 1^m$ where a is one of the generators, m and n are integers or to the relation 1^n (see [23] for details). Thus some edges in van Kampen diagrams can have label 1. One can insert 0-edges and 0-cells in a diagram in order to make a path, which touches itself, absolutely simple.

We call a van Kampen diagram *reduced* if it does contain 0-cells and 0-edges and does not have a pair of cells having a common edge and such that the label of the boundary of the subdiagram consisting of these two cells is equal to 1 in the free group (see Figure 1):

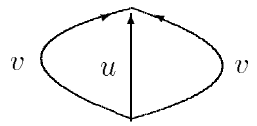

Fig. 1.

If a diagram has such a pair of cells then these two cells can be removed and the corresponding subdiagram can be replaced by a diagram without cells. This operation is called *canceling* a pair of cells.

If Δ is an (ordinary) van Kampen diagram then $\partial(\Delta)$ denotes its boundary.

We always assume that boundaries of van Kampen diagrams and boundaries of cells are oriented clockwise (this is an insignificant difference with [23]). The *contour* of a cell or a diagram is the union of the boundary and its inverse.

A path in the 1-skeleton of a diagram Δ is called *simple* if it does not cross itself after an arbitrary small deformation. A path is called *reduced* if it does not contain consecutive mutually inverse edges.

The length of a path p, denoted by $|p|$ is the number of non-zero edges in it.

We shall use the following operations which can be performed on arbitrary diagrams.

Taking the inverse (mirror image). Let Δ be a van Kampen diagram over a symmetric set of defining relations $\mathcal{P}$. Consider the mirror image Δ^{-1} of the graph Δ. Since $\mathcal{P}$ is symmetric, the graph Δ^{-1} is again a van Kampen diagram over $\mathcal{P}$. We call Δ^{-1} the *inverse of* Δ.

Composition. Let Δ_1 and Δ_2 be van Kampen diagrams over a presentation $\mathcal{P}$. Let $\partial(\Delta_1) = p_1 p_1'$, $\partial(\Delta_2) = p_2^{-1} p_2'$. Suppose that $\mathrm{Lab}(p_1) = \mathrm{Lab}(p_2)$ in the free group. Here $\mathrm{Lab}(p)$ is the label of the path p. Then by adding 0-cells we can make both paths p_1 and p_2 absolutely simple and the labels of p_1 and p_2 identical. After that we can glue Δ_1 and Δ_2 by identifying the corresponding edges of p_1 and p_2. This operation will be called the *composition* of Δ_1 and Δ_2.

Let Δ be any van Kampen diagram. The following definition is similar to the definition of a dual graph of a diagram [20], [23]. Fix a point in each of the cells in Δ and a point in the inside of each of the edges of Δ. For each cell π and each edge e on the boundary of this cell, fix a simple polygonal line $\ell(\pi, e)$ inside π_i which connects the point inside π_i and the point inside e. We can choose these lines in such a way that $\ell(\pi, e)$ and $\ell(\pi, e')$ do not have common points except for the fixed point inside π.

The next definition of a band in a diagram is crucial for our paper as it was in [26] and [24].

Let S be a set of letters and let Δ be a van Kampen diagram. Fix pairs of S-edges on the boundaries of some cells from Δ (we assume that each of these cells has exactly two S-edges).

Suppose that Δ contains a sequence of cells $(\pi_1, \dots, \pi_n)$ such that for each $i = 1, \dots, n-1$ the cells π_i and π_{i+1} have a common S-edge e_i and this edge belongs to the pair of S-edges fixed in π_i and π_{i+1}. Consider the line which is the union of the lines $\ell(\pi_i, e_i)$ and $\ell(\pi_{i+1}, e_i)$, $i = 1, \dots, n-1$. This polygonal line is called the *median* of this sequence of cells. Then our sequence of cells $(\pi_1, \dots, \pi_n)$ with common S-edges $e_1, \dots, e_{n-1}$ is called an *S-band* if the median is a absolutely simple curve or an absolutely simple closed curve (see Fig. 2).

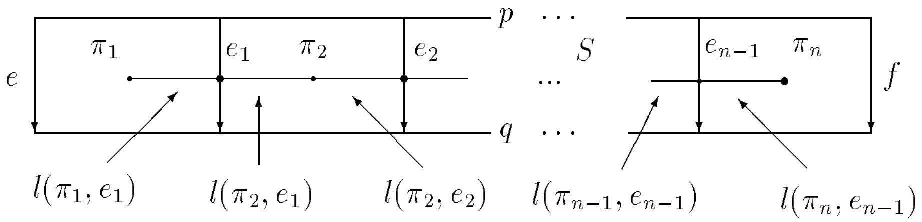

Fig. 2.

Notice that an S-band can contain no cells. In this case the band is called *empty.* An empty S-band consists of one S-edge and its median is empty.

We say that two bands *cross* if their medians cross. We say that a band is an *annulus* if its median is a closed curve. In this case the first and the last cells of the band coincide (see Fig. 3a).

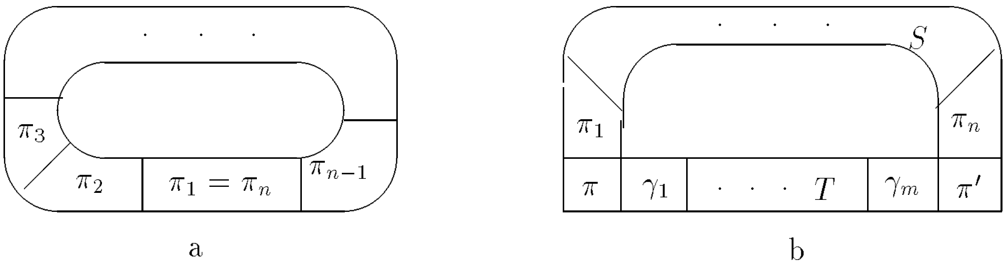

Fig. 3.

Let $\mathcal{B}$ be an S-band with common S-edges $e_1, e_2, \ldots, e_n$ which is not an annulus. Then the first cell has an S-edge e which forms a pair with e_1 and the last cell of $\mathcal{B}$ has an edge f which forms a pair with e_n. Then we shall say that e is the *start edge* of $\mathcal{B}$ and f is the *end edge of* $\mathcal{B}$. If p is a path in Δ then we shall say that a band starts (ends) on the path p if e (resp. f) belongs to p.

Let S and T are two disjoint sets of letters, let $(\pi, \pi_1, \ldots, \pi_n, \pi')$ be an S-band and let $(\pi, \gamma_1, \ldots, \gamma_m, \pi')$ is a T-band. Suppose that:

- the medians of these bands form a simple closed curve,
- on the boundary of π and on the boundary of π' the pairs of S-edges separate the pairs of T-edges,
- the start and end edges of these bands are not contained in the region bounded by the medians of the bands.

Then we say that these bands form an (S,T)-*annulus* and the curve formed by the medians of these bands is the *median* of this annulus (see Fig. 3b). Fig. 4 shows that a multiple intersection of an S-band and a T-band does not necessarily produce an (S,T)-annulus.

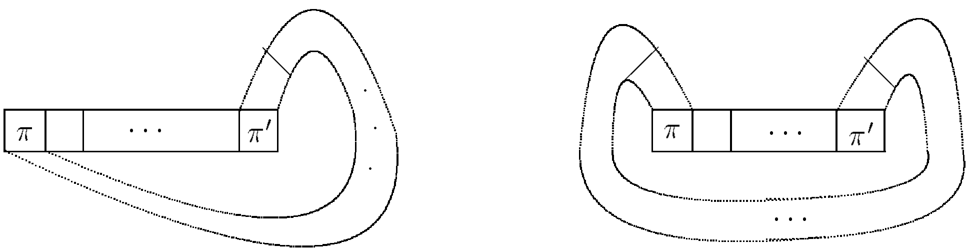

Fig. 4.

If ℓ is the median of an S-annulus or an (S,T)-annulus then the maximal subdiagram of Δ contained in the region bounded by ℓ is called the *inner diagram* of the annulus.

The union of cells of an S-band $\mathcal{B}$ forms a subdiagram. The boundary of this diagram, which we shall call the *boundary of the band*, has the form $e^{\pm 1}pf^{\pm 1}q^{-1}$ (recall that we trace boundaries of diagrams clockwise). Then we say that p is the *top path* of $\mathcal{B}$, denoted by $\mathbf{top}(\mathcal{B})$, and q is the *bottom path* of $\mathcal{B}$, denoted by $\mathbf{bot}(\mathcal{B})$.

If $\mathcal{B} = (\pi_1, \ldots, \pi_n)$ is an S-band then $(\pi_n, \pi_{n-1}, \ldots, \pi_1)$ is also an S-band which we shall call the *inverse of* $\mathcal{B}$.

We shall call an S-band *maximal* if it is not contained in any other S-band. If an S-band $\mathcal{W}$ starts on the contour of a cell π, does not contain π and is not contained in any other S-band with these properties then we call $\mathcal{W}$ a *maximal S-band starting on the contour of* π.

2.4 Known Facts About $G_N(\mathcal{S})$

Now let us return to our presentation of the group $G_N(\mathcal{S})$. Consider the following partition of the generating set X of this group:

$$\Theta_+ \cup Q_1 \cup \ldots \cup Q_k \cup \overline{Y} \cup \{\kappa_1\} \cup \ldots \cup \{\kappa_{4N}\}. \tag{2}$$

Recall that Q_i are the components of the vector of sets of state letters of the S-machine $\mathcal{S}$ and $\overline{Y}$ is the set of tape letters of this S-machine.

The fact that these subsets are disjoint follows from the definition of S-machines.

Let U be a block of this partition. In order to consider U-bands and annuli, we need to divide U-letters of some relation of $G_N(\mathcal{S})$ into pairs.

Every relation except for the hub contains a letter from Θ_+ and the inverse of this letter. These letters form a Θ-pair.

Every auxiliary relation containing a $\overline{Y}$-letter, contains a $\overline{Y}$-letter and the inverse of this letter. These letters form a $\overline{Y}$-pair.

Every auxiliary relation containing κ_i contains also κ_i^{-1}. These two letters form a κ_i-pair.

If a transition relation has a letter from Q_j then it has exactly two letters from $Q_j^{\pm 1}$ (this follows from the definition of an S-rule). They form a Q_j-pair.

For every κ_j-band the number j is called the *index* of this band.

For convenience we repeat here some auxiliary results from [26] which we shall use in this paper. Recall that the *area* of a diagram is the number of cells in this diagram.

Lemma 2.3 *(see* Lemma 7.1 in [26].*) If Δ contains a κ-annulus then one can get a new diagram with the same boundary label as Δ by removing all cells of the annulus, identifying the outer and the inner boundaries of the annulus and adding no new cells.*

Lemma 2.4 *(*Lemma 7.7 in [26].*) If Δ is reduced, does not contain hubs and $\partial(\Delta)$ has no Θ-edges then Δ does not have cells.*

Lemma 2.5 *(*Lemma 7.4 in [26].*) If Δ is reduced and does not contain hubs, then it does not have $(\overline{Q}, \Theta)$-annuli.*

Lemma 2.6 *(*Lemma 7.2 in [26].*) If Δ is reduced and does not have hubs then it does not have $\overline{Q}$-annuli.*

Lemma 2.7 *(*Lemma 7.6 in [26].*) If Δ is reduced and does not contain hubs then it does not contain Θ-annuli.*

Definition 2.8 *A disc is any reduced van Kampen diagram over the presentation of $G_N(\mathcal{S})$ which has exactly one hub and no Θ-edges on the contour.*

Lemma 2.9 *(*Proposition 10.1 in [26].*) The cyclically reduced boundary label of a disc is equal to $\kappa(w)$ for some admissible word w accepted by the S-machine $\mathcal{S}$. The area of the disc does not exceed big "O" of the area of the corresponding computation.*

Lemma 2.10 *(*see Lemma 11.1 in [26].*) If a diagram over the presentation of $G_N(\mathcal{S})$ contains discs then it has at least $4N-3$ κ-bands starting on the same hub and ending on the boundary of the diagram.*

Lemma 2.11 *(*See the proof of Lemma 11.1 in [26].*) Let Δ be a reduced van Kampen diagram over the presentation of $G_N(\mathcal{S})$ containing two hubs connected by two κ-bands. Then the number of hubs in Δ can be reduced by 2 without changing the boundary label of Δ.*

Lemma 2.12 *(*see Lemma 7.8 in [26].*) If Δ is reduced and contains no $\overline{Q}$-edges then it does not have $(\overline{Y},\Theta)$-annuli.*

2.5 Hyperbolic Graphs

Here we present some graph theoretic concepts and results from [24] and [26]. Let Γ be a finite plane 2-complex with vertices $v_0, ..., v_n$, $n \geq 1$. The vertex v_0 is called *external* and other vertices are called *internal*. We assume that there exists a curve from v_0 to the infinity without intersecting the edges of Γ (that is v_0 is indeed an external vertex of the plane graph Γ). We also assume that Γ does not have faces with 1 vertex v_i or bigons with vertices v_i, v_j, $i,j \geq 1$; there are also no 1-gons at the vertex v_0 but v_0 can be connected by several edges with an internal vertex v_i. Such a graph Γ will be called an ℓ-graph if the degree of every internal vertex is $\geq \ell \geq 6$. The next lemmas show that for $\ell \geq 7$ ℓ-graphs have *hyperbolic* properties. In particular the next Lemma shows that the "curvature" of ℓ-graphs is negative. It shows that the number of external edges of a hyperbolic graph is proportional to the number of internal edges.

Lemma 2.13 *(*See Lemma 3.2 in [24] or Lemma 11.5 in [26].*) In an ℓ-graph, there exists a vertex o of degree $d \geq \ell$ and $d-3$ consecutive (clockwise) edges $e_1, ..., e_{d-3}$ connecting o with the external vertex v_0, such that there are no vertices of Γ between e_i and e_{i+1}, $i = 1, ..., d-4$.*

Lemma 2.14 *(*See Lemma 3.4 in [24].*) Let $d_0, d_1, ..., d_n$ be degrees of vertices $v_0, v_1, ..., v_n$ of an ℓ-graph Γ. Then*

$$d_0 \geq 6 - 6n + \sum_{i=1}^{n} d_i.$$

Let two edges e and e' go out of some internal vertex v of an ℓ-graph Γ. Then all edges going out of v can be listed in the clockwise order as follows: $e, f_1, ..., f_{n_1} e', f'_1, ..., f'_{n_2}$. We say that the edges e and e' form a *big angle* if $n_1 \leq 2n_2$ and $n_2 \leq 2n_1$.

We call any path $e_1, ..., e_\alpha$ in Γ *weakly curved* if the vertices $(e_1)_+$,...,$(e_{\alpha-1})_+$ are internal and the edges e_i^{-1} and e_{i+1} form a big angle at the vertex $(e_i)_+ = (e_{i+1})_-$ for every $i = 1, ..., \alpha - 1$.

Lemma 2.15 *(*See Lemma 3.5 in [24].*). If a weakly curved path in an ℓ-graph starts in an internal vertex then it cannot be closed (if $\ell \geq 6$; see Fig. 5a).*

This lemma implies that any maximal weakly curved path in an ℓ-graph starts and ends in the external vertex.

Let x be a Jordan arc on the plane which has exactly two common points with Γ: vertices x_- and x_+. We shall call such an arc Γ-*normal.* Suppose that a simple path p in Γ passes through a vertex v which is equal to x_- or x_+. Then we can cyclically (clockwise) order all edges going out of v: $f_1, ..., f_d$ so that x comes to v between f_d and f_1. We say that the arc x is k-*separated* from p if the edges $f_1, ..., f_k$, $f_{d-k+1}, ..., f_d$ and their inverses do not occur in p.

Two maximal weakly curved paths p and p' in an ℓ-graph Γ are called *diverging* in their internal vertices v and v' if $v \neq v'$ and v and v' can be connected by a Γ-normal arc x which is 2-separated from p and p'.

Lemma 2.16 *(*See Lemma 3.6 in [24].*) Any two diverging paths in an ℓ-graph Γ do not have common internal vertices (see Fig. 5a).*

By Lemma 2.15 every maximal weakly curved path p in an ℓ-graph Γ is absolutely simple, so it divides the plane into two regions $\mathcal{O}_1$ and $\mathcal{O}_2$ (the inner and the outer regions). If for every vertex $v \neq v_0$ on p the number of edges going from v to $\mathcal{O}_1$ (resp. $\mathcal{O}_2$) is bigger by at least 8 than the number of edges going from v to $\mathcal{O}_2$ (resp. $\mathcal{O}_1$) then the path p is called an *oval* and we denote $\mathcal{O}_1$ (resp. $\mathcal{O}_2$) by $\mathcal{O}(p)$. It is easy to understand that if $\ell \geq 28$ then for every internal edge e of an ℓ-graph there exists an oval passing through e.

We call a simple Jordan curve Γ-*transversal* if it does not pass through the vertices of Γ and crosses the edges of Γ transversally. We say that a Γ-transversal curve *envelopes* an internal vertex v of Γ if it intersects consecutively edges $f_1, ..., f_s$ going out of v where $s-2$ is at least a half of the degree of v, and in the sectors between consecutive edges and the curve, there are no other vertices of Γ.

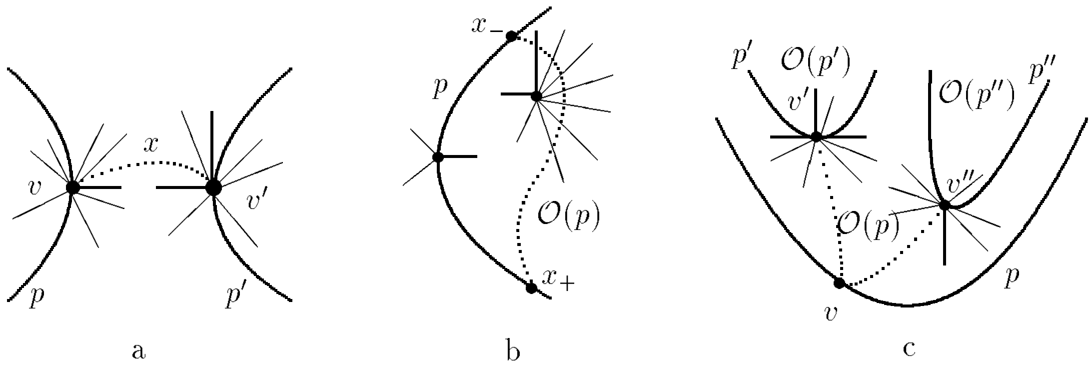

Fig. 5.

Lemma 2.17 *(*See Lemma 3.7 in [24].*). Let Γ be an ℓ-graph, and p be an oval in Γ. Suppose that a Γ-transversal curve x is contained in $\mathcal{O}(p)$ except for the two points x_- and x_+ which belong to p. Then either (1) x intersects the same edge of Γ twice and in the region bounded by this edge and x there are no other vertices of Γ or (2) x envelopes an internal vertex of Γ (see Fig. 5b).*

Consider an arbitrary oval p in an ℓ-graph Γ. Let v be an internal vertex on p. Let v' be another internal vertex belonging to the closure of the region $\mathcal{O}(p)$ and let p' be an oval passing through v'. If the ovals p and p' are diverging in v and v' and the corresponding Γ-normal arc x is 3-separated from p' and x is not in $\mathcal{O}(p')$ then we call the oval p' *derived* from the oval p in the vertices v and v'. By Lemma 2.16 v' cannot belong to p.

Lemma 2.18 (See Lemma 3.8 in [24]. *(1) In the above notation, the region $\mathcal{O}(p)$ contains the region $\mathcal{O}(p')$ defined by the derived oval p' .*

(2) Let p'' be another oval derived from p in vertices v and v'', $v' \neq v''$. Then the regions $\mathcal{O}(p')$ and $\mathcal{O}(p'')$ do not intersect (see Fig. 5c).

3 Higman Embedding

Now we are going to give the presentation of a finitely presented group $H_N(\mathcal{S})$ containing G which will have isopermetric function $n^2T(n^2)^4$.

This group will be constructed as a sequence of HNN extensions (similar to but simpler than) the well known Aanderaa construction in [25]).

The set of *generators* of $H_N(\mathcal{S})$ consists of all the generators of $G_N(\mathcal{S})$, together with the following new letters: ρ, d, and the set B which is in one-to-one correspondence with the set A (we assume that this correspondence preserves inverses).

The set of *relations* of $H_N(\mathcal{S})$ consists of relations of the previously given finite presentation of $G_N(\mathcal{S})$, together with the following new relations:

$\rho x = x\rho$ for all $x \in \{\alpha, \delta, \omega\} \cup A \cup \{z_j \mid 0 \leq j \leq 4\} \cup \{\kappa_i \mid 3 \leq i \leq 4N\}$;
$\rho^{-1}\kappa_1\rho = \kappa_1 d^{-1}$;
$\rho^{-1}\kappa_2\rho = d\kappa_2$;

$dx = xd$ for all $x \in \{\alpha, \delta, \omega\} \cup \{z_j \mid 0 \leq j \leq 4\}$;
$d^{-1}a_i d = a_i b_i$ for all $a_i \in A$ (where b_i is the letter in B that corresponds to a_i);

$a_i b_j = b_j a_i$ for all $a_i \in A$, $b_j \in B$, that is every element from A commutes with every element from B.

Let G_b be the copy of G over the generating set B (recall that B is in one-to-one correspondence with the original generating set A of G). From now on will simply write H for $H_N(\mathcal{S})$. We will show later (see Lemma 4.35 below) that the identity map on B induces an isomorphism between G_b and the subgroup $\langle B\rangle$ of H.

Remark 3. It can be easily deduced from lemmas from Section 4 that that H is obtained from $G_N(\mathcal{S})$ by a sequence of 3 HNN extensions (in the first extension the stable letters are letters from B, in the second extension d is the stable letter and in the third extension ρ is the stable letter; the reader can easily recover from the presentation of H the vertex groups and the edge groups of these extensions). Notice that if we used the version of $G_N(\mathcal{S})$ with linear word $\kappa(w) = \kappa_1 w_1 ... \kappa_{4N} w_{4N}$ where w_i are copies of w written in disjoint copies of the alphabet $\overline{Y} \cup \overline{Q}$ (as we discussed in Remark 2), then in the definition of $H_N(\mathcal{S})$ we could eliminate the letter d passing its responsibilities to ρ. That is ρ would still commute with all letters which could appear in w_i, $i \geq 2$, and with all k_i, $i = 1, ..., 4N$. It will also commute with copies of $z_0, ..., z_4$, α, δ, ω in w_1, but for every A-letter a which can appear in w_1 we would have $a^\rho = ab$. This presentation would reduce the number of HNN extensions needed to built $H_N(\mathcal{S})$ to 2 and it would slightly simplify the proofs. The disadvantage is that in this case we would have to explain in details how to rewrite lemmas in [26] using this new presentation of $G_N(\mathcal{S})$ and it would take too long. So we decided to use the above presentation of $H_N(\mathcal{S})$.

The following three lemma shows that the subgroup generated by B in H is a homomorphic image of G_b (the homomorphism is induced by the identity map on B).

Lemma 3.1 *The relations of G_b follow from the presentation of $H_N(\mathcal{S})$ (i.e. the identity map on B induces a homomorphism from G_b onto the subgroup $\langle B\rangle$ of $H_N(\mathcal{S})$).*

Proof. Let $u_b = 1$ be any relation in G_b. We are going to describe a van Kampen diagram with boundary u_b. Let $u = u_a$ be the copy of the word u_b rewritten in the alphabet A.

By Proposition 2.2 there exists a van Kampen diagram Δ with boundary label $\kappa(\sigma(c(u)))$. This word is a product of elements of the set $A \cup \{\alpha, \delta, \omega, z_0, z_1, z_2, z_3, z_4, \kappa_1, ..., \kappa_{4N}\}$. For each element x of this set there exists a relation of H of the form $x^\rho = y$ (where y is a word). Therefore there exists a ρ-annulus R whose inner boundary is also bounded by $\kappa(\sigma(c(u)))$. Let us glue Δ inside this annulus. Denote the resulting van Kampen diagram by Δ_1. Notice that all cells of R are commutativity cells (corresponding to relations $x^\rho = x$) except for two cells corresponding to the relations $\kappa_1^\rho = \kappa_1 d^{-1}$ and $\kappa_2^\rho = d\kappa_2$.

The cyclic word written on the boundary of Δ_1 has the form $Vd^{-1}\sigma(c(u))dW$ for some words V and W such that $V\sigma(c(u))W$ is a cyclic shift of $\kappa(\sigma(c(u)))$. The word $\sigma(c(u))$ is a product of words from $A \cup \{\alpha, \delta, \omega, z_0, z_1, z_2, z_3, z_4\}$. For each element x of this set we have a relation $x^d = y$ where $y = x$ if $x \in \{\alpha, \delta, \omega, z_0, z_1, z_2, z_3, z_4\}$ and $y = ab$ if $x = a \in A$. Since every element of A commutes with every element of B, we can form an annular diagram $\mathcal{A}$ with outer boundary labeled by the word $d^{-1}\sigma(c(u))d\sigma(c(u))^{-1}$ and the inner boundary labeled by u_b. Let us glue the diagrams Δ_1 and $\mathcal{A}$ along the path labeled by $d^{-1}\sigma(c(u))d$. The resulting annular diagram Δ_2 will have outer boundary labeled by $\kappa(\sigma(c(u)))$ and the inner boundary labeled by u_B.

Now draw diagram Δ_2 on the top half of a sphere with the north pole inside the disc bounded by the inner boundary of Δ_2 and the outer boundary of Δ occupying the equator of the sphere. Since the outer boundary of Δ_2 has the same label as the boundary of Δ, we can draw Δ^{-1} (the mirror image of Δ) on the bottom half of the sphere, and then glue two halves together. The stereographic projection of the resulting spherical graph on the plane gives us a van Kampen diagram over the presentation of H with boundary label u_b. The lemma is proved (see Fig. 6) $\square$

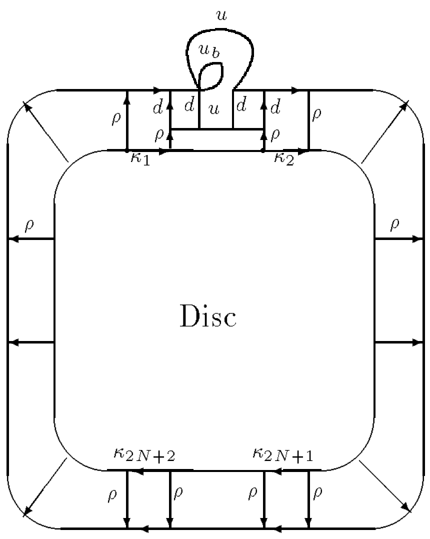

Fig. 6.

4 Forbidden Subdiagrams

In order to analyze the isoperimetric function of $H = H_N(\mathcal{S})$ we first use a new presentation, which we will call **the disc-based presentation** of H. Although H is finitely presented the disc-based presentation will be infinite (this help prove important lemmas from Section 4). The generators of the disc-based presentation are the same as those of the original presentation. The relators of the disc-based presentation consist of the finite set of original relators of H, together with the following new relators:

$\kappa(w) = 1$ for all admissible words w accepted by $\mathcal{S}$;

$u_b = 1$ for all u_b such that $u_b = 1$ in G_b and u_b is a cyclically reduced word.

Since by Proposition 2.2 and Lemma 3.1 these relations hold in H, the disc-based presentation is indeed a presentation of H. Notice that if a van Kampen diagram over the original finite presentation of $G_N(\mathcal{S})$ contains a disc (that is by Definition 2.8 a reduced diagram over the presentation of $G_N(\mathcal{S})$ which has exactly one hub and no Θ-edges on the contour) then by Lemma 2.9 there exists a cell of the disc-based presentation of H with boundary label equal in the free group to a cyclic shift of the boundary label of the disc. Therefore we can cut off the disc from our diagram and replace it by a diagram with one cell. All the lemmas in this section are about van Kampen diagrams relative to the disc-based presentation of H; and when we say "van Kampen diagram", we mean a van Kampen diagram over this presentation. Such a van Kampen diagram consists of *discs* with labels of the form $\kappa(w)$ ($= 1$ in $G_N(\mathcal{S})$), G_b-cells with labels of the form u_b ($= 1$ in G_b), and cells of the original finite presentation of H.

With every van Kampen diagram over the disc-based presentation we associate a 4-vector, called the *type* of this diagram. The first coordinate is the number of disc cells in this diagram, the second coordinate is the number of (ρ, κ)-cells, the third coordinate is the total number of κ-, Θ-, (d, B)-, G_b- and ρ-cells, excluding the discs and (ρ, κ)-cells, and the fourth coordinate is the number of (A, B)-cells (corresponding to the relations $a_i b_j = b_j a_i$). We order the types lexicographically. When we say that a diagram Δ is *minimal*, we always mean that the type of diagram is minimal among the types of all diagrams with the same boundary label. We shall say that the discs have the highest rank among all cells of a diagram, the (ρ, κ)-cells have smaller rank, the cells counted in the third coordinate of the type have rank smaller than the rank of (ρ, κ)-cells, and the (A, B)-cells have the smallest rank. Notice that we can view van Kampen diagrams over the disc based presentation as graded diagrams [23]. Similar grading is used in [24] and [26]. Notice also that by Proposition 2.2 and Lemma 2.9 we can add discs to the presentation of $G_N(\mathcal{S})$ without changing the group. This presentation will be called the *disc-based presentation* of $G_N(\mathcal{S})$.

The following Lemma shows that $G_N(\mathcal{S})$ is embedded into H.

Lemma 4.1 *Let Δ be a minimal diagram over the disc based presentation of H with boundary over the alphabet of generators of $G_N(\mathcal{S})$ then Δ is a diagram over the disc-based presentation of $G_N(\mathcal{S})$.*

Proof. Indeed, if we kill letters B, ρ and d in any relation of the presentation of H containing these letters, we get a trivial relation. Therefore if Δ contains a B-, ρ- or d-cell, then by replacing in Δ all B-, ρ-, d-edges with 0-edges, and then removing 0-cells, we can lower the type of the diagram. $\Box$

Notice that the proof of Lemma 4.1 shows that $G_N(\mathcal{S})$ is embedded in H and is a retract of H.

The rest of this section is similar to Section 2.4: we show that a minimal diagram over the disc-based presentation of H does not have certain bands and annuli.

In order to consider bands in diagrams over H we need to choose pairs of letters in every relation over the disc-based presentation of H. We shall preserve the pairs in the relations over the presentation of $G_N(\mathcal{S})$. In every κ-cell except for the discs, there are two κ-edges, they form a pair. So we can consider κ-bands. In (A, B)-, (ρ, A)- and (d, A)-relations, there are two $\overline{Y}$-letters, they form a pair. Thus we can include (B, A)-, (ρ, A)- and (d, A)-cells in $\overline{Y}$-bands. Every $\overline{Q}$-relation except for the discs contains two $\overline{Q}$-letters, they form a pair. So we can include $(d, \overline{Q})$- and $(\rho, \overline{Q})$-cells in $\overline{Q}$-bands. Every B-relation except G_b-relations and (d, B)-relations contain two B-letters, so we can consider B-bands (consisting of (A, B)-commutation cells). Every ρ-relation contain ρ and ρ^{-1}, these letters form a pair. So we can consider ρ-bands. Finally every d-relation except for (d, ρ)-relations contains d and d^{-1}, these letters form a pair. so we can consider d-bands.

In the rest of this section we shall often use the following obvious fact without references.

Lemma 4.2 *Let X denote κ, $\overline{Y}$, Θ, $\overline{Q}$, B, ρ or d. Then in every X-band in a reduced van Kampen diagram over the disc-based presentation of H, two neighbor cells can have* *only X-edges and $\overline{Y}$-edges in common. Therefore if an X-band does not have a Z-edge on its top or bottom path where $Z \neq X$ denotes κ, Θ, $\overline{Q}$, B, ρ or d, then the X-band does not have Z-cells.*

Lemma 4.3 *A minimal diagram over the disc-based presentation of H has no d-annuli.*

Proof. Suppose by contradiction that there is a d-annulus. Consider any d-annulus Σ. The outer or the inner boundary p of this annulus is a word over $\{\alpha, \omega, \delta\} \cup A \cup \{z_i \mid i = 0, \ldots, 4\}$. Therefore the label of p is a word over the generators of $G_N(\mathcal{S})$. By Lemma 4.1 the subdiagram Δ_1 bounded by p is over the disc-based presentation of $G_N(\mathcal{S})$. This immediately implies that p is the inner boundary of Σ.

If the subdiagram Δ_1 contains discs then by Lemma 2.10 Δ_1 contains a discs and $4N-3$ κ-bands starting on it which end on the boundary of Δ. But a d-band cannot have κ-cells. Therefore Δ_1 does not contain discs. Then by Lemma 2.4 Δ_1 does not have cells. Therefore the label of p is equal to 1 in the free group. This implies that p contains two neighbor edges with mutually inverse labels. The corresponding neighboring cells of Σ cancel which contradicts the minimality of Δ. $\Box$

Lemma 4.4 *In a minimal van Kampen diagram without κ-edges, there are no ρ-annuli.*

Proof. Suppose that a minimal diagram Δ without κ-edges contains a ρ-annulus Σ. Then Σ consists of commutativity cells (corresponding to the relations $\rho x = x\rho$. The inner and the outer boundaries of this annulus have the same labels. Therefore we can identify the inner and outer boundaries of Σ, removing all cells of Σ from Δ. This operation does not change the boundary label of Δ, but decrease the type of the diagram. $\Box$

Lemma 4.5 *In a reduced van Kampen diagram Δ over the disc-based presentation there are no (ρ, κ)-annuli, no (Θ, κ)-annuli and no (d, κ)-annuli.*

Proof. By contradiction, assume Δ contains a (ρ, κ_i)-annulus, a (Θ, κ_i)-annulus, or a (d, κ_i)-annulus for some i. Let Σ be an innermost one. So Σ is either a (ρ, κ_i)-annulus or a (Θ, κ_i)-annulus or a (d, κ_i)-annulus, and it has no other (ρ, κ_i)-annuli, (Θ, κ_i)-annuli, or (d, κ)-annuli in the subdiagram Δ_1 bounded by the inner boundary of Σ.

Let us first consider the case where Σ is a (Θ, κ_i)-annulus. Recall that the only cells which can occur in the κ_i-band are ρ-cells and Θ-cells. We claim that the κ_i-band of Σ contains no ρ-cells.

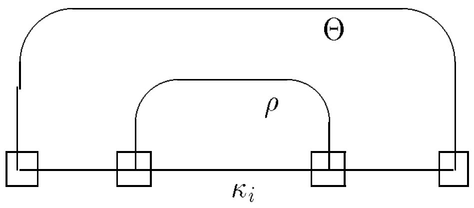

Fig. 7.

Indeed, if there were a ρ-cell on the κ_i-band, a ρ-band would start there, and it would form a (ρ, κ_i)-annulus with the κ_i-band of Σ (since a ρ-band cannot cross a Θ-band; see Fig. 7). But this would contradict the assumption that Σ is innermost.

Hence, the κ_i-band of Σ consists only of Θ-cells. If it contains not only the intersection cells then we would get a (Θ, κ_i)-annulus inside Δ_1 which again contradicts the assumption that Σ is innermost.

It follows that the κ_i-band consists of the two intersection cells. Then these cells cancel: they have a common κ_i-edge, and their Θ-edges have the same labels but point in the opposite directions (note that along one Θ-band the labels of the Θ-edges are the same everywhere; and the only (κ_i, Θ) cells are (κ_i, Θ)-commutation cells).

Let us consider the case where Σ is a (ρ, κ_i)-annulus. The κ_i-band of Σ contains no ρ-cells (except for the two intersection cells between the ρ-band and the κ_i-band of Σ). The proof is quite similar to the previous cases.

Indeed, if there were a ρ-cell on the κ_i-band, a ρ-band would start there, and it would form another (ρ, κ_i)-annulus with the κ_i-band (since two ρ-bands cannot cross). But this would contradict the assumption that Δ is inner-most. Similarly the κ_i-band cannot contain Θ-cells because Θ-bands and ρ-bands cannot intersect. Therefore the κ_i-band consists of two (adjacent) intersection ρ-cells. These two cells are adjacent and cancel (note that for a given i there exists only one kind of (ρ, κ_i)-cell). This contradicts the assumption that our diagram is reduced.

Finally suppose that Σ is a (d, κ_i)-annulus. The d-band of Σ cannot contain κ-cells except for the intersection cells. A ρ-band in Δ_1 starting on the boundary of the κ_i-band of Σ cannot intersect the d-band. This as before implies that the κ_i-band of Σ cannot contain ρ-cells except for the intersection cells. Similarly it cannot contain Θ-cells (because a Θ-band cannot intersect with a d-band). Therefore again the κ_i-band consists of two intersection cells that cancel. $\square$

Lemma 4.6 *A reduced diagram over the disc-based presentation cannot contain κ-annuli.*

Proof. Indeed, by Lemma 4.5 a κ-annulus cannot contain Θ-cells or ρ-cells, so it is empty. $\square$

Lemma 4.7 *In a reduced diagram over the disc-based presentation there are no B-annuli.*

Proof. A B-annulus consists only of (A, B)- commutation cells, so the labels of the inner and outer boundaries of the annulus are equal, and we can remove the annulus as in Lemma 4.4. $\square$

Lemma 4.8 *In a minimal diagram without κ-edges, there are no $(\overline{Q}, \rho)$-annuli, no $(\overline{Q}, d)$-annuli and no $(\overline{Q}, \Theta)$-annuli.*

Proof. Let Σ be an inner-most $(\overline{Q}, \rho)$-, $(\overline{Q}, d)$-, or $(\overline{Q}, \Theta)$-annulus. Let $\mathcal{Q}$ be the $\overline{Q}$-band and let $\mathcal{B}$ be the ρ-, d, or Θ-band in this annulus.

Suppose first that $\mathcal{Q}$ contains just two cells (the intersection cells with $\mathcal{B}$). Then these two ρ-, d-, or Θ cells have a common $\overline{Q}$-edge. Therefore these cells cancel. This contradicts the minimality of our diagram.

Now suppose that $\mathcal{Q}$ contains more than two cells. The inner subdiagram of Σ (the subdiagram bounded by the inner boundary of this annulus) cannot contain ρ-edges or d-edges by Lemmas 4.3 and 4.4.

Therefore the cells of $\mathcal{Q}$ which are not the intersection cells with $\mathcal{B}$ are Θ-cells. A maximal Θ-band $\mathcal{B}'$ starting on the contour of $\mathcal{Q}$ cannot cross $\mathcal{B}$ because there are no (Θ, d)- or (Θ, ρ)-cells. Therefore we have a $(\overline{Q}, \Theta)$-annulus Σ' contained inside the subdiagram bounded by the annulus $\mathcal{R}$. This contradicts the assumption that Σ is innermost. □

Lemma 4.9 *In a minimal diagram without discs there are no $\overline{Q}$-annuli.*

Proof. Suppose that Σ is a $\overline{Q}$-annulus in a minimal diagram Δ. Let Δ' be the subdiagram bounded by the external boundary of Σ. By Lemma 4.8 Σ does not contain d-cells or ρ-cells. Therefore the label of the boundary of Δ' is a word over the generators of $G_N(\mathcal{S})$. Thus Δ' is a diagram over the presentation of $G_N(\mathcal{S})$. But by Lemma 2.6 a diagram over the presentation of $G_N(\mathcal{S})$ cannot contain $\overline{Q}$-annuli. □

Lemma 4.10 *In a minimal diagram without discs there are no Θ-annuli.*

Proof. Let Σ be a Θ-annulus in a minimal diagram Δ. Let Δ' be the subdiagram bounded by the outer boundary of Σ. Then the boundary label of Δ' is a word over the generators of $G_N(\mathcal{S})$, so Δ' is a diagram over the presentation of $G_N(\mathcal{S})$, and it remains to use Lemma 2.7 □

Definition 4.11 Let Φ be a Θ- or ρ-band, and let Π be a disc. We say that Φ **touches** Δ if the bottom path $\mathbf{bot}(\Phi)$ has at least 2 common κ-edges with the contour of Π. Let κ_i (resp. κ_j) be the label of the first (resp. last) common κ-edge of Φ and $\partial(\Pi)$ (counted along the path $\mathbf{bot}(\Phi)$. Let p_1 be the shortest subpath of $\mathbf{bot}(\Phi)$ which contains these κ-edges and let p_2 be the subpath of the contour of Π with the same ends as p_1 such that the subdiagram bounded by p_1 and p_2 does not contain the disc Π. Then the subdiagram bounded by p_1 and p_2 will be denoted by $\Delta(\Phi, \Pi)$ (see Fig. 8).

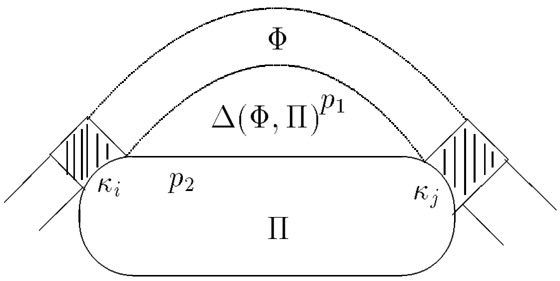

Fig. 8.

Lemma 4.12 *Suppose that a Θ-band Φ touches a disc Π in a minimal diagram Δ and $\Delta(\Phi,\Pi)$ does not contain discs. Then the subdiagram $\Delta(\Phi,\Pi)$ has no cells.*

Proof. Let us use the notation from the Definition 4.11. Every maximal κ-band in $\Delta(\Phi,\Pi)$ starting on p_2 must end on p_1. By Lemma 4.5 none of these κ-bands contain ρ-cells or Θ-cells (otherwise there would be (ρ,κ)- or (Θ,κ)-annuli. Every maximal κ-band starting on p_1 must end on p_2 because otherwise we would have a (Θ,κ)-annulus. Therefore each of the κ-bands starting on the boundary of $\Delta(\Phi,\Pi)$ is empty. Therefore $\Delta(\Phi,\Pi)$ is a disjoint union of several subdiagrams Δ_i, $i=1,...,k$, connected by κ-edges (bridges): for every $i=1,...,k-1$, a κ-edge connects Δ_i and Δ_{i+1}. The boundary of each Δ_i contains no κ-edges.

Consider one of the subdiagrams Δ_i. The label of the boundary of this diagram is a word over the generators of $G_N(\mathcal{S})$. Therefore by Lemma 4.1 Δ_i is a diagram over $G_N(\mathcal{S})$. The contour of Δ_i does not contain Θ-edges, and Δ_i does not contain discs by the assumption, so by Lemma 2.4 Δ_i does not contain cells. $\Box$

The next construction plays a very important role both in [26] and in [24]. We repeat it here.

Definition 4.13 Moving Θ-bands. Suppose that in a minimal diagram Δ a Θ-band $\mathcal{R}$ touches a disc Π. By Lemma 4.12 then the bottom path of $\mathcal{R}$ has a common subpath with the contour of Π starting and ending with κ-edges. Let p be the maximal common subpath with this property, so that $\mathbf{bot}(\Theta)=qpq'$, $\partial(\Pi)=pp_1$. Without loss of generality we can assume that the label $\mathrm{Lab}(p)$ of the path p has the form $\kappa_i w^{\pm 1}\kappa_{i+1}w^{\mp 1}\kappa_{i+2}..\kappa_j$ where w is an admissible word accepted by the S-machine $\mathcal{S}$ (it could be also the inverse of this word, but then we could take the mirror image of Δ instead of Δ). Then for some word V we have that $\mathrm{Lab}(p)V$ is a cyclic shift of $\kappa(w)=\mathrm{Lab}(\partial(\Pi))$. There exists a Θ-band $\mathcal{R}'$ with the bottom path labeled by the word V and the Θ-edges having the same labels as in $\mathcal{R}$. Let $\mathcal{R}''$ be the subband of $\mathcal{R}$ with bottom path p, so $\mathcal{R}=\mathcal{R}_1\mathcal{R}''\mathcal{R}_2$. Let e be the start edge of $\mathcal{R}''$ and let e' be the end edge of $\mathcal{R}''$. Cut the diagram Δ along the path $e^{-1}p_1e'$. We can fill the resulting hole by gluing in the Θ-band $\mathcal{R}'$ and the mirror image $\overline{\mathcal{R}}'$ of $\mathcal{R}'$. The new diagram Δ' that we obtain this way will have two Θ-bands instead of the old Θ-band $\mathcal{R}$. The first is $\mathcal{R}_1(\overline{\mathcal{R}}')^{-1}\mathcal{R}_2$ (the inverse band $(\overline{\mathcal{R}}')^{-1}$ differs from $\overline{\mathcal{R}}'$ by the order of cells) and the second one is $\mathcal{R}''\mathcal{R}'$. The second Θ-band is an annulus which touches Π along its inner boundary. If we replace the disc Π be the corresponding van Kampen diagram over the presentation of $G_N(\mathcal{S})$, we see that the subdiagram Π' bounded by the outer boundary of the annulus $\mathcal{R}''\mathcal{R}'$ is a diagram over the presentation of $G_N(\mathcal{S})$ with exactly one hub and no Θ-edges on the boundary. By Definition 2.8 Π' is a disc. We replace it by one cell of the disc-based presentation. Then we reduce the resulting diagram. Notice that this construction amounts to enlarging the disc Π and moving the band $\mathcal{R}$ through the disc (the band $\mathcal{R}_1(\overline{\mathcal{R}}')^{-1}\mathcal{R}_2$ touches the disc Π' along the κ_{j+1}-, ...,κ_{i-1}-edges) and then reducing the resulting diagram (see Fig. 9).

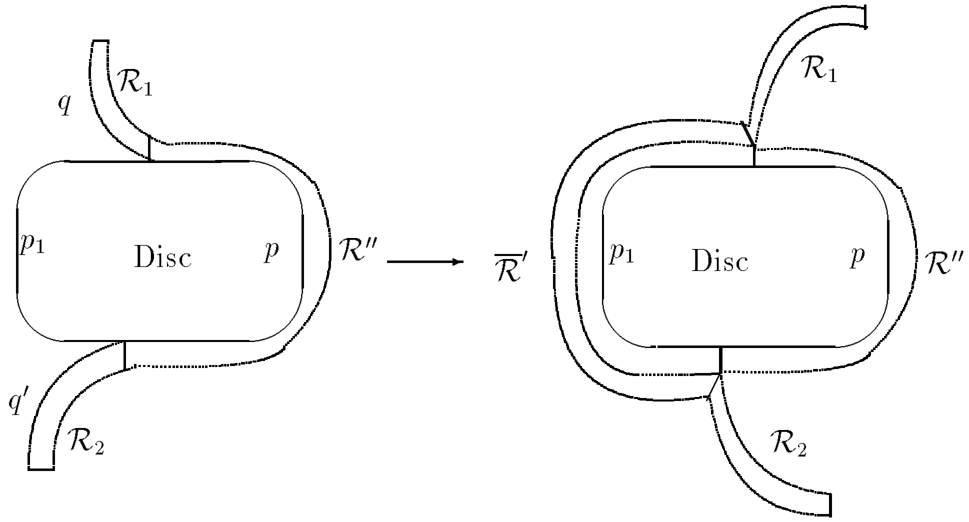

Fig. 9.

This construction has several important applications. For example, suppose that the band $\mathcal{R}$ had more than $2N$ common κ-edges with the contour of the disc Π and there are no other discs between $\mathcal{R}$ and Π. Then the Θ-band moving construction reduces the number of Θ-cells in the diagram without increasing the number of cells of bigger ranks. Indeed, every maximal $(\overline{Y}, \overline{Q})$-part of the boundary of a disc has the same length by Lemma 2.9. Therefore the Θ-band moving construction reduces the number of Θ-cells. On the other hand the Θ-band moving construction replaces one disc Π by another disc Π' which does not increase the number of discs in the diagram, and it does not touch (κ, ρ)-cells and cells of the same rank as Θ-cells). This implies the following statement.

Lemma 4.14 *In a minimal diagram there is no Θ-band $\mathcal{R}$ which touches a disc Π in more than $2N$ edges and such that $\Delta(\mathcal{R}, \Pi)$ contains no discs.*

We will be using this construction more later.

Now we are going to define an analogous construction for ρ-bands. It is similar to a construction in [24].

Lemma 4.15 *Let $W = \sigma(c(u))$ where u is a word in the alphabet A such that $u = 1$ in the group G. Then the equalities $W = \rho W \rho^{-1}$ and $W = dWd^{-1}$ can be deduced from the presentation of $H_N(\mathcal{S})$ without using κ-relations.*

Proof. By Proposition 2.1 (4) $W \equiv z_0\alpha^n z_1 u z_2 \delta^n z_3 \omega^n z_4$ and u_b is equal to 1 in the group G_b. Therefore using the non-κ-relations of $H_N(\mathcal{S})$ we can transform the word dWd^{-1} into $z_0\alpha^n z_1 u u_b z_2 \delta^n z_3 \omega^n z_4$. Now the subword u_b can be deleted by using the G_b-relation $u_b = 1$.

The first statement of the lemma follows from ρ-relations of $H_N(\mathcal{S})$. □

Lemma 4.16 *Suppose that a ρ-band Σ touches a disc Π in a minimal diagram Δ along two consecutive κ-edges whose indices differ from 1 and 2. Suppose further that the subdiagram $\Delta(\Sigma,\Pi)$ contains no discs. Then the label of $\partial(\Pi)$ has the form $\kappa(\sigma(c(u)))$ where $u=1$ in G.*

Proof. We use the notation of Definition 4.11. Since p_1 and p_2 do not contain ρ- or Θ-edges, by Lemmas 4.4 and 4.10 the subdiagram $\Delta(\Sigma,\Pi)$ contains no ρ- or Θ-cells and also (by assumption) contains no discs. Since the subdiagram $\Delta(\Sigma,\Pi)$ contains no discs, it contains no κ-edges except the two κ-edges where $\mathbf{bot}(\Sigma)$ touches Π (these κ-edges are common for p_1 and p_2). Since the indices of these κ-edges differ from 1 and 2, we can conclude that the paths p_1 and p_2 contain no d-edges. Therefore by Lemma 4.3 $\Delta(\Sigma,\Pi)$ contains no d-edges.

Therefore this subdiagram contains no $\overline{Q}$-cells (because every $\overline{Q}$-cell is either a disc or a Θ-cell, or a d-cell or a ρ-cell). Therefore each $\overline{Q}$-edge of p_2 also belongs to p_1. By Lemma 4.8, every $\overline{Q}$-edge of p_1 belongs to p_2. Since every $(\rho,\overline{Q})$-relation has the form $\rho z_i = z_i\rho$ where $i=0,1,2,3,4$, we can conclude that the word V obtained from the label of p_2 by removing all non-$\overline{Q}$-letters, is a product of z_i, $i\in\{1,2,3,4\}$. In the case where Δ has only one hub, Δ is a disc (see Definition 2.8 and Lemma 2.9). By the definition of S-machines, the label of p_2 contains one representative from each component Q_i of the vector Q, and these representatives are placed in w in the order of indices of Q_i. Since words z_j contain no common letters and $z_0z_1z_2z_3z_4$ contains representatives of all Q_i, each Q_i represented precisely once, we can conclude that $V\equiv z_0\dots z_4$. Now the statement of our lemma follows from Proposition 2.1 (part 3) and Proposition 2.2 (part 1). $\square$

Lemma 4.17 *Let Δ be a van Kampen diagram over the disc-based presentation. Suppose that Δ consists of a disc Π with boundary label $\kappa(\sigma(c(u)))$ for some word u equal to 1 in G and ρ-cells $\rho_1,\dots,\rho_n$ attached to Π along $n\le 4N$ κ-edges e_i, $i\in\{1,\dots,n\}$ and attached to Π in a similar manner (that is their ρ-edges either all point toward Π or all point away from Π). Let S be the set of the indices of the κ-edges e_i, $i=1,\dots,n$. Then there exists another diagram Δ' with the same boundary label as Δ, consisting of a disc with the same boundary label as Π, $4N-n$ (ρ,κ)-cells $\pi_{i+1},\dots,\pi_{4N}$ attached to it and a number of cells of smaller ranks such that the indices of κ-edges of the cells $\pi_{n+1},\dots,\pi_{4N}$ form a complement to the set S in the set $\{1,\dots,4N\}$.*

Proof. Let $W=\sigma(c(u))$. Consider an auxiliary diagram Δ_1 obtained by attaching to Δ new (ρ,κ)-cells π_{n+1}, ...,π_{4N} in the same manner as the other cells π_i, $i\le n$ (that is their ρ-edges point toward Π or away from Π, as in $\pi_1,\pi_2,\dots,\pi_n$). The boundary of the diagram Δ_2 has label $\prod\kappa_iU_i$ where $U_1\equiv(\rho d)^{\pm1}W(\rho d)^{\mp1}$ (depending on the orientation of ρ-edges), $U_j\equiv\rho^{\pm1}W^{\pm1}\rho^{\mp1}$ for $j>1$.

By Lemma 4.15, for every $i=1,\dots,4N$, to the paths labeled by U_i, we can attach a number of cells of ranks smaller than the rank of κ-cells, and obtain a diagram Δ_3 whose boundary label is the same as the label of the disc Π.

Thus the diagram Δ_3 can be obtained by gluing Δ_1 and a diagram Δ_0 consisting of $4N-n$ (ρ,κ)-cells whose indices form a complement to S, and several cells of smaller ranks. This implies that, conversely, there exists a diagram Δ' with the same boundary label as Δ_1, which can be obtained by gluing Π with the mirror image Δ_0^{-1} of the diagram Δ_0. $\square$

The construction defined in the proof of Lemma 4.17 will be called the *ρ-band moving construction.*

The next Lemma gives the first application of this construction.

Lemma 4.18 *Suppose that a ρ-band Σ touches a disc Π along n κ-edges in a minimal diagram Δ. Then $n < 4N - 6$.*

Proof. By contradiction, suppose that Σ touches Π along $n \geq 4N - 6$ κ-edges. Since $N \geq 6$, two of these edges have consecutive indices (mod $4N$) which differ from 1 and 2. By Lemma 4.15 the boundary label of Π is $\kappa(\sigma(c(u)))$ for some word u which is equal to 1 in G.

Consider a subdiagram Δ_1 of Δ consisting of the disc Π with $n \geq 4N - 6$ (ρ, κ)-cells π_i from Σ attached to it in the same manner (all ρ-edges of these cells are pointing either toward the disc or away from it). By Lemma 4.17 we can find a diagram Δ' which consists of the disc Π, $4N - n$ (ρ, κ)-cells and a number of cells of smaller ranks, and which has the same boundary label as Δ_1. Since $4N - n < n$, the type of the diagram Δ' is smaller than the type of Δ_1. Replacing subdiagram Δ_1 in Δ with Δ' will make the type of the whole diagram smaller which contradicts the minimality of Δ. $\square$

With a van Kampen diagram Δ over the disc-based presentation we can associate the following graph $\Gamma(\Delta)$, which we call the **disc graph**. The internal vertices are the discs, the external vertex is taken outside the diagram, the edges are the κ-bands with indices ≥ 4 that start on discs, any κ-band with index ≥ 4 that starts on a disc and ends on the boundary will be considered as an edge connecting this disc with the external vertex. This graph is similar to the graphs considered in [26] and [24] (see Section 2.5).

Definition 4.19 A diagram Δ over the disc-based presentation of H is called *normal* if it does not contain two discs Π_1 and Π_2 connected by two κ-bands $\mathcal{B}_1$, and $\mathcal{B}_2$ with consecutive (mod $4N$) indices not equal 1 and 2, and the region between $\mathcal{B}_1$ and $\mathcal{B}_2$ does not contain discs.

We shall prove later that every minimal diagram is normal. Before that we need to show that the disc graph of a normal diagram is hyperbolic in the sense of Section 2.5.

Lemma 4.20 *The disc graph of a minimal van Kampen diagram has no one-gons.*

Proof. Since a disc has only one κ_i-edge for any i, there are obviously no one-gons in the disc graph. $\square$

Lemma 4.21 *The disc graph corresponding to a minimal normal van Kampen diagram contains no bigons.*

Proof. Indeed suppose that the disc graph of a minimal diagram Δ contains a bigon. The bigon consists of two discs Π_1 and Π_2 connected by two κ-bands. Let i and j be the indices of these κ-bands. Then either i and j are consecutive numbers ≥ 4 or $i = 4N, j = 4$ (the case $j = 4$, $i = 4N$ is, of course, similar). In the first case Lemma 4.23 immediately gives a contradiction because our diagram is normal.

In the second case notice that since the bigon must bound a face in the disc graph, there are no discs between the edges of the bigon. Consider the maximal κ_3-band starting on the contour of Π_1. This band cannot intersect the κ_{4N}-band or the κ_4-band. Therefore it ends up on the contour of Π_2. Therefore the discs Π_1 and Π_2 are connected by a κ_3-band and by a κ_4-band, a contradiction with the assumption that our diagram is normal. $\Box$

The following Lemma follows from Lemma 2.13.

Lemma 4.22 *For every minimal normal diagram Δ with at least one disc there exists a disc Π with $4N-6$ consecutive (modulo $4N$) κ-bands $\mathcal{B}_1, \ldots, \mathcal{B}_{4N-6}$ starting on $\partial(\Pi)$ and ending on the boundary of Δ. For every such disc let $\Psi_\Delta(\Pi)$ be the subdiagram of Δ bounded by $\mathbf{top}(\mathcal{B}_1)$, $\mathbf{bot}(\mathcal{B}_{4N-6})$, $\partial(\Delta)$ and $\partial(\Pi)$, which contains $\mathcal{B}_1$, $\mathcal{B}_{4N-6}$ and does not contain Π (there is only one subdiagram in Δ satisfying these conditions). Then there exists a disc Π such that $\Psi_\Delta(\Pi)$ does not contain discs.*

Lemma 4.23 *Suppose that a diagram Δ over the disc-based presentation contains two discs Π_1 and Π_2 connected by two κ-bands $\mathcal{B}_1$ and $\mathcal{B}_2$ (both start on the contour of Π_1 and end on the contour of Π_2) whose indices are consecutive numbers (mod $4N$) not equal to 1 or 2. Suppose also that the diagram Δ_1 bounded by $\mathbf{top}(\mathcal{B}_1)$, $\mathbf{bot}(\mathcal{B}_2)$ and the boundaries of Π_1 and Π_2 ($\Pi_1, \Pi_2 \notin \Delta_1$) is reduced and normal (we do not assume that Δ is reduced or normal). Then the number of discs in the diagram can be reduced by 2 without changing the boundary label of Δ.*

Proof. Assume by contradiction that there exist two discs connected by two κ-bands $\mathcal{B}_1$, $\mathcal{B}_2$ and satisfying the conditions of the lemma. Let us denote the discs by Π_1 and Π_2. It is clear that the indices of the κ-bands must be consecutive numbers modulo $4N$. We can assume that Π_1, Π_2, $\mathcal{B}_1$ and $\mathcal{B}_2$ are chosen in such a way that the diagram Δ_1 is innermost.

Let us first consider the case where there are no ρ-edges on the two κ-bands $\mathcal{B}_1$ and $\mathcal{B}_2$. Then Δ_1 has boundary label over $G_N(\mathcal{S})$. Then by Lemma 4.1 Δ_1 is a diagram over the disc based presentation of $G_N(\mathcal{S})$.

By replacing the discs in $\Delta_1 \cup \{\Pi_1, \Pi_2\}$ by diagrams over the (original finite) presentation of $G_N(\mathcal{S})$, and by applying Lemma 2.11, we can conclude that the number of discs in Δ can be reduced by 2 without changing the boundary label of Δ. The new diagram has the same boundary label as Δ but the number of discs is smaller by 2.

Now suppose that one of the κ-bands $\mathcal{B}_1$ and $\mathcal{B}_2$ contains ρ-cells. The diagram Δ_1 does not contain discs. Indeed, since this diagram is normal, by Lemma 4.22 there exists a disc inside Δ' with $4N-6$ consecutive external edges. At most two of the corresponding κ-bands can end on Π_1 or Π_2 (otherwise we would have two discs connected by two κ-bands and the diagram bounded by the discs and the bands smaller than Δ_1). The other

$4N-8>1$ κ-bands must intersect $\mathcal{B}_1$ or $\mathcal{B}_2$ which is impossible because two κ-bands cannot intersect.

Since by Lemma 4.5, Δ cannot contain (ρ,κ)- or (Θ,κ)-annuli, every maximal Θ-band and every maximal ρ-band crossing $\mathcal{B}_1$ (resp. $\mathcal{B}_2$) must cross $\mathcal{B}_2$ (resp. $\mathcal{B}_1$).

Let e_1, e_2 be the start edges of the bands $\mathcal{B}_1$ and $\mathcal{B}_2$. Suppose that the cell π of $\mathcal{B}_1$ attached to e_1 is a (Θ,κ)-cell. Then the maximal Θ-band $\mathcal{T}$ containing π crosses $\mathcal{B}_2$. The intersection cell of $\mathcal{T}$ and $\mathcal{B}_2$ must be attached to e_2 because otherwise the Θ- or ρ-band containing the cell of $\mathcal{B}_2$ attached to e_2 forms a (Θ,κ)- or (ρ,κ)-annulus with $\mathcal{B}_2$. By Lemma 4.12 the diagram between $\mathcal{T}$ and Π_1 is empty (because as we saw it does not contain discs), and we can move the Θ-band $\mathcal{T}$ over the disc Π_1 enlarging the disc Π_1 (see Fig. 10). Let Δ' be the new diagram and Π_1' be the enlarged disc. In the (reduced) diagram Δ', discs Π_1' and Π_2 are still connected by two κ-bands $\mathcal{B}_1'$ and $\mathcal{B}_2'$ which are shorter than the bands $\mathcal{B}_1$ and $\mathcal{B}_2$.

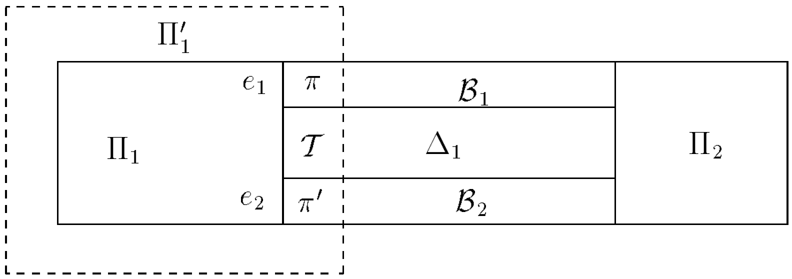

Fig. 10.

Now let us assume that the cell π is a ρ-cell. Then as in the previous paragraph, the intersection π' of the ρ-band $\mathcal{T}$ containing π with the κ-band $\mathcal{B}_2$ must be a cell attached to e_2. Therefore the band $\mathcal{T}$ touches Π_1 along two consecutive κ-edges whose indices are not 1 or 2. Then by Lemma 4.16 the label of the boundary of Π_1 is $\kappa(\sigma(c(u)))$ for some word u which is equal to 1 in the group G. Consider the subdiagram of Δ consisting of the disc Π_1 and two cells π and π'. By Lemma 4.17 we can apply the ρ-band moving construction and replace $\Delta_1\cup\{\Pi_1,\Pi_2\}$ by a diagram Δ_1' with the same boundary label such that Δ_1' consists of the disc Π_1, $4N-2$ (ρ,κ)-cells attached to it along the $4N-2$ κ-edges not including e_1 and e_2 and a number of cells of smaller ranks. Let us replace Δ_1' for $\Delta_1\cup\{\Pi_1,\Pi_2\}$ in Δ. Denote the new diagram by Δ_2. The diagram Δ_1' contains only two discs Π_1' and Π_2 connected by proper sub-bands $\mathcal{B}_1'$ and $\mathcal{B}_2'$ of the bands $\mathcal{B}_1$ and $\mathcal{B}_2$. After reducing the subdiagram Δ' bounded by the boundaries of Π_1 and Π_2, $\mathbf{top}(\mathcal{B}_1')$, $\mathbf{bot}(\mathcal{B}_2')$ we still have two discs Π_1' and Π_2 connected by two κ-bands with consecutive indices not equal to 1 and 2, which are shorter than $\mathcal{B}_1$ and $\mathcal{B}_2$. The subdiagram between $\mathcal{B}_1'$ and $\mathcal{B}_2'$ is normal because it is disc-free. Notice that the number of discs in the whole diagram does not change. Continuing in this manner, we shall make both κ-bands connecting our discs empty. But this situation has been considered before. $\square$

Lemma 4.23 immediately implies

Lemma 4.24 *Every minimal diagram is normal.*

Proof. By contradiction let Δ be a minimal non-normal diagram with the smallest possible number of cells. Then Δ contains two discs Π_1 and Π_2 connected by two κ-bands $\mathcal{B}_1$ and $\mathcal{B}_2$ with consecutive indices (distinct from 1 and 2). Since Δ contains minimal possible number of cells, the subdiagram of Δ bounded by Π_1, Π_2, $\mathcal{B}_1$, $\mathcal{B}_2$ and not containing Π_1, Π_2 is normal. Then by Lemma 4.23 we get a contradiction with the minimality of Δ. □

Lemma 4.25 *In a minimal van Kampen diagram Δ over the disc-based presentation there are no ρ-annuli and no Θ-annuli.*

Proof. Suppose by contradiction that there is a ρ-annulus or a Θ-annulus in our diagram. Let us pick an inner-most one and let us call it Σ.

Assume first that there are no discs inside the diagram Δ_1 bounded by the outer boundary of Σ.

Then Σ cannot be a Θ-annulus by Lemma 4.10.

If Σ is a ρ-annulus then there are no κ-edges inside the diagram bounded by the annulus (otherwise there would be either (ρ,κ)-annuli ruled out by Lemma 4.5 or κ-annuli ruled out by Lemma 4.6). Now we can get a contradiction by applying Lemma 4.4.

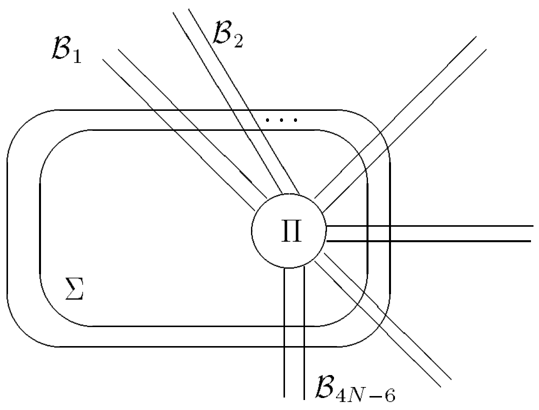

Fig. 11.

Let us now assume that the subdiagram Δ_1 bounded by the outer boundary of the annulus Σ contains discs. The discs inside Δ_1 form a disc graph, and by Lemma 4.22 at least one disc Π in this graph has $4N-6$ consecutive κ-bands $\mathcal{B}_1, ..., \mathcal{B}_{4N-6}$ starting on the contour of Π and ending on the boundary of Δ_1 and such that the subdiagram $\Psi_{\Delta_1}(\Pi)$ is disc-free (see Fig. 0). By Lemma 4.5 every maximal ρ- or Θ-band crossing $\mathcal{B}_1$ must cross all $4N-6$ bands $\mathcal{B}_1, ..., \mathcal{B}_{4N-6}$. Take the ρ- or Θ-band Σ' which touches Π at the start edge of $\mathcal{B}_1$. Then the intersection cell of Σ' and $\mathcal{B}_2$ must contain the start edge of $\mathcal{B}_2$ because otherwise we would have a (ρ,κ) or (Θ,κ) annulus ruled out by Lemma 4.5. Continuing in this manner, we conclude that the intersection cell of Σ' and $\mathcal{B}_i$, $i = 1, ..., 4N-6$ contains the start edge of $\mathcal{B}_i$. Therefore Σ' touches Π in $4N-6 \geq 2N$ κ-edges. Since $\Psi_{\Delta_1}(\Pi)$ is disc-free, we can use Lemmas 4.14 and 4.18 and get a contradiction. □

Lemma 4.26 *In a minimal van Kampen diagram there are no $\overline{Q}$-annuli.*

Proof. Assume by contradiction that there is a $\overline{Q}$-annulus $\mathcal{Q}$. Let Δ_1 be the diagram bounded by the outer boundary of $\mathcal{Q}$. Then by Lemma 4.22, Lemma 4.24 and Lemma 4.6 Δ_1 cannot contain κ-edges (since a $\overline{Q}$-band cannot intersect a κ-band). It remains to apply Lemma 4.9. $\square$

Lemma 4.27 *In a minimal van Kampen diagram there are no $(\overline{Q},\Theta)$-annuli.*

Proof. Notice that if there were a disc inside the subdiagram Δ_1 bounded by a $(\overline{Q},\Theta)$-annulus, then the Θ-band of this annulus would intersect $4N-6$ consecutive κ-bands starting on the contour of a disc (Lemma 4.22 and Lemma 4.24) which is impossible by Lemma 4.14 (this lemma applies to the Θ-band which intersects the $4N-6$ κ-bands and is the closest to the disc). Therefore Δ_1 does not contain discs. Now the statement follows from Lemma 4.8. $\square$

Lemma 4.28 *In a minimal van Kampen diagram there are no $(\overline{Q},\rho)$-annuli and no $(\overline{Q},d)$-annuli.*

Proof. By Lemma 4.8 it is enough to show that the subdiagram Δ_1 bounded by the outer boundary of an innermost such annulus Σ cannot have κ-edges.

If the diagram Δ_1 contains discs then by Lemmas 4.22 and 4.24 there would exist $4N-6$ κ-bands starting on the same disc and intersecting the d- or ρ-band of our annulus (the $\overline{Q}$-band cannot intersect with κ-bands), and the subdiagram $\Psi_{\Delta_1}(\Pi)$ is disc-free. But a d-band can intersect only with κ_1- and κ_2-bands and $4N-6>2$. So our annulus cannot be a $(\overline{Q},d)$-annulus. But if Σ is a $(\overline{Q},\rho)$-annulus, we can get a contradiction as in the proof of Lemma 4.25.

Thus Δ_1 contains no discs. Then by Lemma 4.6 if Δ_1 contains a κ-edge, the maximal κ-band of Δ_1 containing this edge would form a (ρ,κ)- or (d,κ)-annulus with the ρ- or d-band of Σ. But this has been ruled out by Lemma 4.5. $\square$

Lemma 4.29 *Consider a minimal van Kampen diagram Δ. Then in Δ, there are no $(\overline{Y},\rho)$-annuli, no $(\overline{Y},d)$-annuli, and no $(\overline{Y},B)$-annuli.*

Proof. Let us consider a $(\overline{Y},\rho)$-, $(\overline{Y},d)$-, or $(\overline{Y},B)$-annulus Σ and suppose that Σ is an innermost annulus of the corresponding form (that is if Σ is a $(\overline{Y},\rho)$-annulus then the diagram bounded by the outer boundary of this annulus does not contain other $(\overline{Y},\rho)$-annuli, etc.). Let Δ_1 be the subdiagram bounded by the outer boundary of the annulus. Let $\mathcal{Y}$ be the $\overline{Y}$-band of the annulus and let $\mathcal{T}$ be the d-, ρ- or B-band of the annulus. Then as in the proof of Lemma 4.28 we prove that Δ_1 contains no κ-edges (like $\overline{Q}$-bands, a $\overline{Y}$-band cannot cross a κ-band).

Suppose that $\mathcal{T}$ is a ρ-band. Then $\mathcal{T}$ cannot contain κ-cells, as we proved before. It also cannot contain $\overline{Q}$-cells other than the two intersection cells, otherwise there would be a $(\rho,\overline{Q})$-annulus ruled out by Lemma 4.28 (recall that a $\overline{Q}$-band cannot cross a $\overline{Y}$-band). This and the fact that Δ does not contain $\overline{Q}$-annuli (Lemma 4.26) imply that Δ_1 does not contain $\overline{Q}$-edges at all (otherwise Δ_1 would contain a $\overline{Q}$-annulus ruled out by Lemma 4.26). If $\mathcal{T}$ contains $\overline{Y}$-cells distinct with the intersection cells then the maximal $\overline{Y}$-bands starting on the boundary of $\mathcal{T}$ cannot end on a cell inside Δ_1 (since as we just proved, Δ_1 does not contain $\overline{Q}$-cells). This implies that these $\overline{Y}$-bands must form $(\overline{Y}, \rho)$-annuli with $\mathcal{T}$. This contradicts the assumption that our annulus is innermost. Thus $\mathcal{T}$ does not contain $\overline{Y}$-cells. Therefore $\mathcal{T}$ contains exactly two intersection cells. These cells must cancel which contradicts the minimality of Δ. This implies that Δ does not contain $(\overline{Y}, \rho)$-annuli.

Suppose that $\mathcal{T}$ is a d-annulus. If $\mathcal{T}$ contains $\overline{Q}$-cells, then Δ contains $(\overline{Q}, d)$-annulus, which is impossible by Lemma 4.28. Therefore $\mathcal{T}$ does not contain $\overline{Q}$-cells. This and the absence of $\overline{Q}$-annuli (Lemma 4.26) implies that Δ_1 does not contain $\overline{Q}$-cells. Therefore $\mathcal{T}$ cannot contain $\overline{Y}$-cells except for the intersection cells (otherwise as in the case when $\mathcal{T}$ is a ρ-band, we would get a contradiction with the assumption that Σ is innermost). Since every d-cell is either a $\overline{Q}$-cell or a $\overline{Y}$-cell, we can conclude that $\mathcal{T}$ consists of two intersection cells that cancel, a contradiction. This proves that Δ does not have $(\overline{Y}, d)$-annuli.

Finally assume that $\mathcal{T}$ is a B-band. Every cell in $\mathcal{T}$ is an (A, B)-commutation cell. A maximal $\overline{Y}$-band in Δ_1 starting on the boundary of $\mathcal{T}$ cannot end in Δ_1 because Δ_1 cannot contain $\overline{Q}$-cells (otherwise we would have a $\overline{Q}$-annulus which is impossible). Therefore every such $\overline{Y}$-band must form a $(\overline{Y}, B)$-annulus with $\mathcal{T}$ which contradicts the assumption that Σ is innermost. Therefore $\mathcal{T}$ consists of two intersection cells that cancel. This proves that Δ does not have $(\overline{Y}, B)$-annuli. □

Lemma 4.30 *In a minimal van Kampen diagram, a $\overline{Y}$-band cannot start and end on the same disc.*

Proof. Suppose there is a $\overline{Y}$-band $\mathcal{Y}$ that starts and ends on the same disc Π of Δ.

Let Δ_1 be the subdiagram of Δ bounded by the top or the bottom path of $\mathcal{Y}$ and the boundary of the disc, such that Δ_1 contains the $\overline{Y}$-band but does not contain Π.

The same argument as in the proof of the previous lemmas shows that Δ_1 contains no κ-edges (recall that a $\overline{Y}$-band cannot cross a κ-band).

The subdiagram Δ_1 contains no ρ-annuli (by Lemma 4.4), and no d-annuli (Lemma 4.3). Since (by Lemma 4.29) there are no $(\overline{Y}, \rho)$-, $(\overline{Y}, d)$-, or $(\overline{Y}, B)$-annuli, Δ_1 contains no ρ-, d- or B-edges (recall that the boundary of a disc does not have ρ-, d- or B-edges).

Therefore Δ_1 is a diagram over the presentation of $G_N(\mathcal{S})$. There are no $\overline{Q}$-edges inside Δ. Indeed, there are no $\overline{Q}$-annuli by Lemma 2.6, a $\overline{Q}$-band cannot cross a $\overline{Y}$-band, and a $\overline{Q}$-band cannot start and end on the boundary of the same disc because no two $\overline{Q}$-edges on the path between any two consecutive κ-edges on the boundary of a disc have mutually inverse labels).

Now by Lemmas 2.12 and 2.7, Δ_1 cannot have Θ-edges either. Therefore Δ_1 is empty. In particular, the $\overline{Y}$-band $\mathcal{Y}$ is empty, which means that the label of the disc is not reduced, a contradiction with the definition of a disc (indeed the label of a disc is $\kappa(w)$ where w is an admissible word for $\mathcal{S}$. □

Lemma 4.31 *In a minimal van Kampen diagram there are no $(\Theta, \overline{Y})$-annuli.*

Proof. Suppose, by contradiction, that there is a $(\Theta, \overline{Y})$-annulus. Let Σ be the innermost one. Then the same argument as in the proof of Lemma 4.28 shows that the inner diagram of Σ does not contain discs. The Θ-band of this annulus cannot contain $\overline{Q}$-cells because the inner diagram Δ_1 of Σ does not contain terminal cells for $\overline{Q}$-bands (discs) and because Lemma 4.27 rules out $(\overline{Q}, \Theta)$-annuli. The subdiagram Δ_1 does not contain $\overline{Q}$-cells because Σ is innermost. This Θ-band also cannot contain $\overline{Y}$-cells (except for the intersection cells) because our annulus is innermost and because the inner diagram of Σ does not contain terminal cells for $\overline{Y}$-bands (discs and $\overline{Q}$-cells) . Therefore the Θ-band consists of two intersection cells that cancel. $\square$

Lemma 4.32 *In a minimal van Kampen diagram, there is no $\overline{Y}$-annuli.*

Proof. Indeed, a $\overline{Y}$-annulus cannot contain discs in the inner diagram (Lemmas 4.22 and 4.24) and other κ-cells (by Lemma 4.6). Therefore this inner diagram cannot contain terminal cells for d-bands. Thus by Lemmas 4.29 and 4.31 an $(\overline{Y})$-annulus cannot have $(\overline{Y}, \rho)$-, $(\overline{Y}, d)$-, or $(\Theta, \overline{Y})$-cells. Therefore it consists of (A, B)-cells. Therefore the inner boundary and the outer boundary of the annulus have equal labels. Thus we can remove the annulus by identifying its inner and outer boundaries. This contradicts the minimality of our diagram. $\square$

The following Lemma is similar to Lemma 11.2 in [24] and Lemma 11.8 from [26].

Lemma 4.33 *Let $\mathcal{R}_1$,...,$\mathcal{R}_n$ be maximal $\overline{Y}$-bands starting on a path p where p is either a $\overline{Y}$-subpath of the boundary of a disc Π or a subpath of $\partial(\Delta)$ containing no κ-edges and no $\overline{Q}$-edges. Suppose that the end edges of all $\mathcal{R}_i$ are on the contours of $(\overline{Q}, \Theta)$-cells belonging to the same Θ-band $\mathcal{T}$. Then $n \leq c$ where c is two times the maximal number of $\overline{Y}$-letters in a $(\overline{Q}, \Theta)$-relation.*

Proof. Indeed, if $n > c$ then there are three $\overline{Y}$-bands, say, $\mathcal{R}_1$, $\mathcal{R}_2$, $\mathcal{R}_3$ starting on p and ending on three different cells π_1, π_2 and π_3 of $\mathcal{T}$. We can assume that π_2 is between π_1 and π_3. Consider the minimal subdiagram Δ_1 of our diagram containing $\overline{Y}$-bands $\mathcal{R}_1, \mathcal{R}_2, \mathcal{R}_3$, the minimal subpath of the path p containing the starting edges of $\mathcal{R}_1$, $\mathcal{R}_2$, $\mathcal{R}_3$, and the part of the band $\mathcal{T}$ between π_1 and π_3 (see Fig. 12). Then Δ_1 has no κ-edges on its contour. Therefore Δ_1 does not contain discs by Lemmas 4.22 and 4.24. Therefore the maximal $\overline{Q}$-band $\mathcal{Q}$ in Δ_1 containing π_2 divides Δ_2 into two parts (that is if we delete the $\overline{Q}$-edges from $\mathcal{Q}$, the diagram Δ_2 will fall into two pieces). The subpath of the path p containing the start edges of $\mathcal{R}_1, \mathcal{R}_2, \mathcal{R}_3$ is contained in one of these parts since it does not contain $\overline{Q}$-edges. The cells π_1 and π_3 belong to different parts because $\mathcal{Q}$ cannot intersect $\mathcal{T}$ twice (Lemma 4.27). Since the π_1 and π_3 are connected with the cells on p by $\overline{Y}$-bands, one of these bands must intersect $\mathcal{Q}$ which is impossible (a $\overline{Q}$-band cannot cross a $\overline{Y}$-band). $\square$

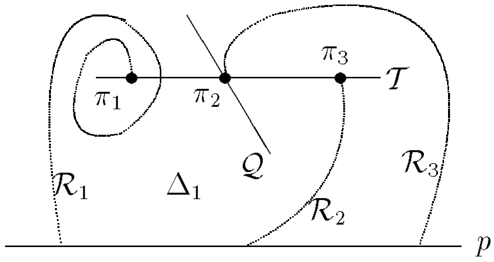

Fig. 12.

Lemma 4.34 *Let Δ be a minimal diagram and let $\partial(\Delta) = pq$ where p is a B-path. Then at most 2 different maximal B-bands in Δ starting on p end on contours of cells of the same $\overline{Y}$-band.*

Proof. The proof is similar to the proof of Lemma 4.33: one needs only replace there $\overline{Q}$ by d and $\overline{Y}$ by B and notice that the minimal subdiagram Δ_1 considered in the proof of Lemma 4.33 does not contain cells which are terminal for d-bands (i.e. (κ, ρ)-cells). □

Lemma 4.35 *The homomorphism from G_b onto the subgroup $\langle B \rangle$ induced by the identity map on B (see Lemma 3.1) is an isomorphism.*

Proof. Consider a minimal diagram Δ over the disc-based presentation of H with boundary label u over B. By the van Kampen Lemma it is enough to proof that $u = 1$ in G_b.

Since the boundary label of Δ does not contain κ-edges, Δ does not contain discs (Lemmas 4.22 and 4.24). Since the word u does not contain letter ρ, Δ does not contain ρ-edges (Lemma 4.25). Since u does not contain letter d and Δ does not contain d-annuli (Lemma 4.3), Δ does not contain d-edges (maximal d-bands in Δ cannot start or end on cells in Δ because Δ does not contain ρ-cells). Similarly we can exclude Θ- and $\overline{Q}$-edges. Therefore $\overline{Y}$-bands in Δ cannot start (end) on cells from Δ. Since u does not contain letters from $\overline{Y}$, Δ cannot contain $\overline{Y}$-edges. Therefore all cells in Δ are G_b-cells, so $u = 1$ in G_b by the van Kampen lemma. □

Lemma 4.36 *In a minimal van Kampen diagram, no B-band can start and end on G_b-cells.*

Proof. Assume, by contradiction, that there is a B-band $\mathcal{B}$ that starts and ends on G_b-cells. This B-band consists of (A, B)-commutation cells. There are two possibilities: the two G_b-cells at the ends of the B-band could be different, or they could be the same.

Let us first consider the case of a B-band $\mathcal{B}$ that starts and ends on the same G_b-cell π. The inner diagram Δ_1 bounded by π and $\mathcal{B}$ has boundary label over $A \cup B$. Since $A \subset G_N(\mathcal{S})$, $G_N(\mathcal{S})$ is a retract of H (see Lemma 4.1 and a remark after the proof of this lemma), and B is contained in the kernel of this retraction, we can conclude that the A-word u and the B-word v written on the boundary of Δ_1 are equal to 1 in H. By Lemma 2.4 and Lemma 4.1 A is a set of free generators of $\langle A\rangle$ (we can apply these lemmas because a diagram with boundary label over A cannot have discs by Lemma 4.22 and 4.24). Therefore $u = 1$ in the free group. Therefore if $\mathcal{B}$ is not empty it must contain a pair of (A, B)-cells that cancel. Thus $\mathcal{B}$ is empty which means that π touches itself. Then the subdiagram Δ_1 of Δ has boundary label v. Let Δ' be the smallest subdiagram of Δ containing the G_b-cell π. Then Δ' contains Δ_1 and π. The boundary label of Δ' is a word w over B. By the van Kampen Lemma $w = 1$ in H. Then by Lemma 4.35 $w = 1$ in G_b, so we can extract Δ' from Δ and replace it by a diagram with one G_b-cell. If Δ_1 contains cells, thius operation would reduce the type of Δ which is impossible since Δ is minimal. Therefore Δ_1 contains no cells. But this means that the boundary label of π is not a cyclically reduced word which contradicts the definition of a G_b-cell. Thus this case is impossible.

Let us now consider a B-band that starts and ends on two different G_b-cells; the B-band consists entirely of (A, B)-commutation cells, or is empty (consists of just one edge). Let bu_b be the label (over $B^{\pm 1}$) of one of the G_b-cells, and let $b^{-1}v_b$ be the label of the other G_b-cell; b is the common B-letter of all the (A, B)-cells of the B-band; let w_a (over A) be the label of the top (or bottom) path of the B-band (it may be equal to 1, if the B-band is empty).

The label of the boundary of the subdiagram consisting of the two G_b-cells and the B-band is $u_b w_a v_b w_a^{-1}$ (and is $= 1$ in H). Modulo (A, B)-commutativity, $u_b w_a v_b w_a^{-1}$ is equal to $u_b v_b$ (hence this is also $= 1$ in H). Thus, in our van Kampen diagram we can remove the two G_b-cells and the B-band, and replace them by the following subdiagram, whose boundary label is still $u_b w_a v_b w_a^{-1}$: a G_b-cell with boundary label $u_b v_b$ and (A, B)-commutation cells, that fill in the space between the G_b-cell and the closed path labeled by $u_b w_a v_b w_a^{-1}$. See Fig. 13. Thus we can lower the type of Δ which contradicts its minimality.

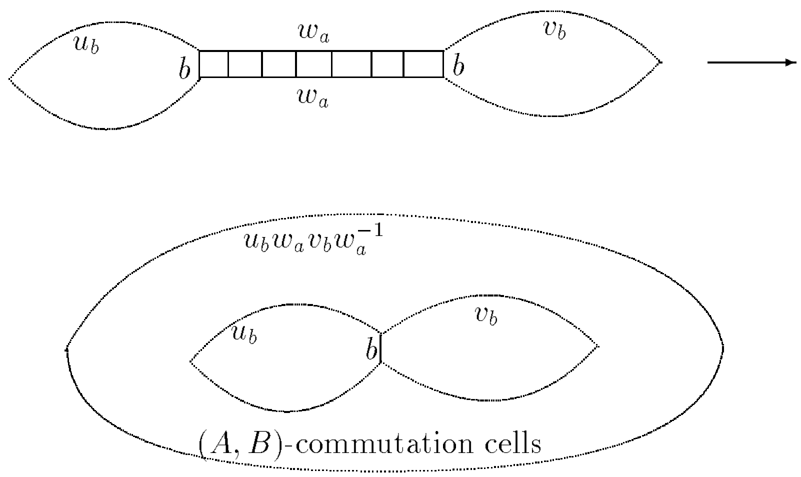

Fig. 13.

5 Estimating Areas in the Disc-based Presentation

Fix a minimal diagram Δ over the disc-based presentation of $H_N(\mathcal{S})$. Let n be the length of the boundary of Δ. We will assume that $n > 3$. We are going to estimate the area of Δ in terms of n. Let k be the number of components of the vector Y of the S-machine $\mathcal{S}$.

Lemma 5.1 *The number of discs in Δ is $O(n)$, and the number of maximal κ-bands is $O(n)$.*

Proof. Let Γ be the disc graph of Δ, and let $v_0, v_1, ..., v_m$ be the vertices of this graph, where v_0 is the external vertex. Without loss of generality we can assume that $m \geq 1$. Then as we know Γ is an ℓ-graph with $\ell = 4N-3$. Let $d_0, d_1, ..., d_m$ be the degrees of vertices $v_0, v_1, ..., v_m$. Then by Lemma 2.14 we have that

$$d_0 \geq 6 - 6m + \sum_{i=1}^{m} d_i.$$

Notice $d_i = 4N-3$ for $i \geq 1$. Therefore

$$d_0 \geq 6 - 6m + (4N-3)m.$$

Since $N \geq 2$ we deduce that $d_0 \geq m$. Notice that obviously $d_0 \leq n$. Therefore the number m of discs in Δ does not exceed n.

It is well known (see for example Lemma 10.2 from [23]) that in any planar graph without 1-gons and bigons the number of vertices n_0 and the number of edges n_1 satisfy the following inequality: $n_1 \leq 3n_0$. Therefore the number of maximal κ-bands starting on discs is at most $6n$ (we consider a band and its inverse as different bands). Other κ-bands start and end on the boundary of Δ, and their number is at most n. Since by Lemma 4.6 Δ does not have κ-annuli, the total number of maximal κ-bands in Δ is at most $7n$. $\square$

In the rest of this section, we are going to estimate the number of non-disc and non-G_b-cells in Δ and the total perimeter of all discs and G_b-cells. The idea is the following. Every cell except for discs and G_b-cells is an intersection of two bands. For example, every (Θ, Q)-cell is an intersection of a Θ-band and a Q-band. So first we compute the number of different types of maximal bands in Δ using the absence of Q-annuli, Θ-annuli, etc. proved in the previous section. Then we count the number of intersections of two bands. There we use the fact that there are no annuli formed by bands of different types.

Remark 4. Notice that a multiple intersection of a maximal S-band $\mathcal{X}$ and a maximal T-band $\mathcal{Y}$ (where S and T are components of our subdivision of the set of generators of $H_N(\mathcal{S})$) does not formally imply that $\mathcal{Y}$ and $\mathcal{T}$ form a (S,T)-annulus. Indeed, $\mathcal{Y}$ can start on the contour of a cell π, and $\mathcal{T}$ can go as a spiral around π and intersect $\mathcal{Y}$ many times as in Fig. 14. In this case $\mathcal{Y}$ and $\mathcal{T}$ do not form a (S,T)-annulus because the start edge of $\mathcal{Y}$ belongs to the region bounded by the medians of $\mathcal{Y}$ and $\mathcal{T}$ (see the definition of an (S,T)-annulus).

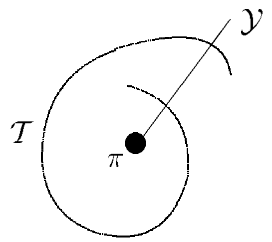

Fig. 14.

But it is easy to see that a multiple intersection does imply the existence of (S,T)-annulus if $\mathcal{T}$ does not start or end on a contour of a cell. Indeed let $\mathcal{Y}=\mathcal{Y}_1\mathcal{Y}_2\mathcal{Y}_3$, $\mathcal{T}=\mathcal{T}_1\mathcal{T}_2\mathcal{T}_3$ where $\mathcal{Y}_2$ and $\mathcal{T}_2$ are parts of $\mathcal{Y}$ and $\mathcal{T}$ between two consecutive intersections, and let π_1 and π_2 be these intersections (so these cells belong to both $\mathcal{Y}_2$ and $\mathcal{T}_2$). Without loss of generality assume that π_1 precedes π_2 along both $\mathcal{Y}$ and $\mathcal{T}$. Then the medians of $\mathcal{Y}_2$ and $\mathcal{T}_2$ form a region Δ_1 on the plane, and the cell π is inside this region. Notice that the part $\mathcal{T}_1$ of $\mathcal{T}$ cannot escape from Δ_1 without crossing $\mathcal{Y}_2$ (it cannot cross itself). If $\mathcal{T}_1$ crosses $\mathcal{Y}_2$ then the intersection cell must be between π_1 and π_2, so $\mathcal{T}_1$ and $\mathcal{Y}_2$ form a (S,T)-annulus. But by our assumption, $\mathcal{T}$ does not start on a contour of a cell, so $\mathcal{T}_1$ must escape from Δ_1, a contradiction. We are going to use this remark several times later.

Lemma 5.2 *The number of maximal ρ- or Θ-bands is $O(n)$. The number of maximal $\overline{Q}$-bands is $O(n)$.*

Proof. By Lemma 4.25 every ρ-band and every Θ-band has to start and end on the contour of Δ. This proves the first statement of the lemma.

The $\overline{Q}$-bands all start and end on the outer boundary of the van Kampen diagram or on discs. Since there are at most n maximal $\overline{Q}$-bands starting on the boundary and there are $O(n)$ discs (by Lemma 5.1), and each disc has a constant number of $\overline{Q}$-edges on the boundary (the number k of components of vector Q multiplied by $4N$), the second statement of the lemma follows. □

Lemma 5.3 *The number of (ρ,κ)-cells in Δ is $O(n^2)$ The number of (Θ,κ)-cells is also $O(n^2)$.*

Proof. A ρ-band cannot cross the same κ_i-band more than once: this follows from the fact that there are no (ρ,κ)-annuli (Lemma 4.5) and the fact that a maximal ρ-band does not start on the contour of a cell (see Remark 4). Thus, each ρ-band has $O(n)$ intersections with κ-bands (by Lemma 5.1). Since there are $O(n)$ maximal ρ-bands (Lemma 5.2), the total number of intersections of ρ-bands and κ-bands is $O(n^2)$.

The proof for (Θ,κ)-cells is very similar: by Lemma 4.5 there are no (Θ,κ)-annuli and a maximal Θ-band cannot start on the contour of a cell. □

Lemma 5.4 *The number of κ-cells in Δ is $O(n^2)$.*

Proof. A κ-cell in our presentation is either a disc (we have $O(n)$ of them, by Lemma 5.1), or a (κ,Θ)-cell (there are $O(n^2)$ of them by Lemma 5.3), or a (κ,ρ)-cell (there are $O(n^2)$ of them by Lemma 5.3). Adding these bounds gives the result. □

Lemma 5.5 *The number of maximal d-bands in Δ is $O(n^2)$.*

Proof. Since Δ does not contain d-annuli (Lemma 4.3), a d-band starts either on the contour of Δ (which has n edges) or on the contour of a (ρ, κ)-cell (of which there are $O(n^2)$, by Lemma 5.3). $\square$

Lemma 5.6 *The number of $(\overline{Q}, \Theta)$-cells in Δ is $O(n^2)$.*

Proof. The proof is similar to the proof of Lemma 5.3. By Lemma 4.27 there are no $(\overline{Q}, \Theta)$-annuli and a maximal Θ-band does not start on the boundary of a cell. Thus, each $\overline{Q}$-band has $\leq n$ intersections with Θ-bands (Lemma 5.2). Since the number of maximal $\overline{Q}$-bands is $O(n)$ (Lemma 5.2) the total number of intersections of Θ- and $\overline{Q}$-bands is $O(n^2)$. $\square$

Lemma 5.7 *The number of $(\overline{Q}, \rho)$-cells in Δ is $O(n^2)$.*

Proof. There are no ρ-annuli (Lemma 4.25), and no $\overline{Q}$-annuli (Lemma 4.26). The number of maximal ρ-bands is $O(n)$ and the number of maximal $\overline{Q}$-bands is $O(n)$, by Lemma 5.2. Since there are no $(\rho, \overline{Q})$-annuli (Lemma 4.28), each ρ-band intersects each $\overline{Q}$-band at most once. Since a maximal ρ-band cannot start on the contour of a cell, there are $O(n^2)$ intersections. $\square$

Lemma 5.8 *The number of $(\overline{Q}, d)$-cells is $O(n^3)$.*

Proof. The numbers of $\overline{Q}$-bands and d-bands are respectively $O(n)$ and $O(n^2)$ (Lemma 5.2 and Lemma 5.5). By Lemma 4.28 there are no $(d, \overline{Q})$-annuli. We cannot apply Remark 4 now because a maximal Q-band and a maximal d-band can start on the boundary of a cell. So we should be more careful in proving that a maximal $\overline{Q}$-band $\mathcal{Y}$ and a maximal d-band $\mathcal{T}$ cannot have a multiple intersection. Suppose that $\mathcal{Y}$ starts on a disc Π, and $\mathcal{T}$ starts on a ρ-cell π (see Fig.15).

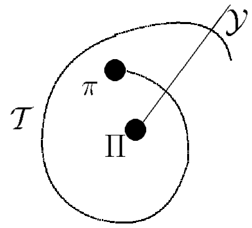

Fig. 15.

Let π_1 and π_2 be consecutive intersections of $\mathcal{Y}$ and $\mathcal{T}$. Then as in Remark 4 let $\mathcal{Y} = \mathcal{Y}_1\mathcal{Y}_2\mathcal{Y}_2$ and $\mathcal{T} = \mathcal{T}_1\mathcal{T}_2\mathcal{T}_3$ where $\mathcal{Y}_2$ and $\mathcal{T}_2$ are parts of $\mathcal{Y}$ and $\mathcal{T}$ between π_1 and π_2. Consider the subdiagram Δ_1 of Δ bounded by the medians of $\mathcal{Y}_2$ and $\mathcal{T}_2$. Then if $\mathcal{Y}_2$ and $\mathcal{T}_2$ do not form an $(\overline{Q}, d)$-annulus, the start or end edge of $\mathcal{T}_2$ and $\mathcal{Y}_2$ must belong to Δ_1. Therefore $\mathcal{Y}$ and $\mathcal{T}$ must start or end on contours of cells belonging to Δ_1 (none of them can escape from Δ_1 without forming a $(\overline{Q}, d)$-annulus). Without loss of generality we can assume that both the disc Π and the ρ-cell π are inside Δ_1. But the contour of Δ_1 cannot contain κ-edges because a $\overline{Q}$-band and a d-band do not contain κ-cells. Therefore by Lemma 4.22 Δ_1 cannot contain discs, a contradiction. Therefore there are no multiple intersections of a $\overline{Q}$-band and a d-band. The statement of the lemma now follows immediately. □

Lemma 5.9 *The number of $\overline{Q}$-cells is $O(n^3)$.*

Proof. In the presentation we use, every $\overline{Q}$-cell is either a disc (we counted $O(n)$ in Lemma 5.1), or a $(\overline{Q}, \rho)$-cell (we have $O(n^2)$ of them by Lemma 5.7), or a $(\overline{Q}, d)$-cell (we have $O(n^3)$), by Lemma 5.8), or a $(\overline{Q}, \Theta)$-cell (we counted $O(n^2)$ in Lemma 5.6). Adding these bounds yields the result. □

Lemma 5.10 *The total number of maximal $\overline{Y}$-bands and therefore the sum of the perimeters of all the discs are $O(n^2)$.*

Proof. By Lemma 4.32 Δ does not contain $\overline{Y}$-annuli. Therefore a maximal $\overline{Y}$-band can start on the contour of Δ or on a $(\Theta, \overline{Q})$-cell or on a disc. The number of maximal $\overline{Y}$-bands starting on the contour of Δ is at most n, so we need to estimate only the number of other maximal $\overline{Y}$-band. Let $y(\Delta)$ be the number of maximal $\overline{Y}$-bands which start on discs in Δ. Let $x(\Delta)$ be the number of maximal $\overline{Y}$-bands which start and end on discs of Δ. Notice that we count every band which start and end on discs and the inverse of this band as two separate bands when we compute $x(\Delta)$ and $y(\Delta)$.

Let us prove by induction on the number of discs in Δ that $x(\Delta) \leq y(\Delta)/2$.

By Lemma 4.22 Δ contains a disc Π and $4N - 6$ consecutive κ-bands $\mathcal{B}_1, ..., \mathcal{B}_{4N-6}$ starting on this disc and ending on the boundary of Δ and such that the diagram $\Psi_\Delta(\Pi)$ is disc-free. Create a diagram Δ_1 by removing from Δ the subdiagram $\Psi_\Delta(\Pi)$ and the disc Π. The number of discs in Δ_1 is fewer than in Δ by 1. Let z be the number of $\overline{Y}$-edges on the path on the boundary of Π between two consecutive κ-edges.

Let p be the maximal subpath of the boundary of Π which is not contained in $\Psi_\Delta(\Pi)$. Notice that $|p|$ contains $7z$ $\overline{Y}$-edges. The number $x(\Delta_1)$ of maximal $\overline{Y}$-bands in Δ_1 which start and end on the contours of discs is $x(\Delta)$ minus the number of those maximal $\overline{Y}$-bands one of whose ends is on p and the other end - on another disc. Therefore

$$x(\Delta_1) \geq x(\Delta) - 14z.$$

On the other hand, the number $y(\Delta_1)$ of the maximal $\overline{Y}$-bands which start on discs of Δ_1 is $y(\Delta) - 4Nz$ (we subtract from $y(\Delta)$ the number of all maximal $\overline{Y}$-bands which start on Π).

Since Δ_1 has fewer discs than Δ, we can conclude that

$$x(\Delta_1) \leq y(\Delta_1)/2.$$

Therefore

$$x(\Delta) - 14z \leq (y(\Delta) - 4Nz)/2.$$

Since $N \geq 7$, we conclude that

$$x(\Delta) \leq y(\Delta)/2.$$

This implies that $y(\Delta)$ does not exceed twice the number of maximal $\overline{Y}$-bands in Δ which start on a disc and end on the boundary of Δ or on a $(\overline{Q}, \Theta)$-cell. Since the number of $(\overline{Q}, \Theta)$-cells in Δ is $O(n^2)$ by Lemma 5.6, the number of maximal $\overline{Y}$-bands starting on $(\overline{Q}, \Theta)$-cells is $O(n^2)$ (we have only finitely many $(\overline{Q}, \Theta)$-relations). Therefore $y(\Delta) = O(n^2)$ and the total number of maximal $\overline{Y}$-bands in Δ is also $O(n^2)$. $\Box$

Lemma 5.11 *The number of ρ-cells is $O(n^3)$.*

Proof. We already proved that the number of (ρ, κ)-cells and $(\rho, \overline{Q})$-cells is $O(n^2)$ (Lemmas 5.3 and 5.7). The only other ρ-cells are the $(\rho, \overline{Y})$-cells. The number of ρ-bands is $O(n)$ (Lemma 5.2), and the number of $\overline{Y}$-bands is $O(n^2)$ (Lemma 5.10). Since there are no $(\rho, \overline{Y})$-annuli (Lemma 4.29), and a maximal ρ-band does not start or end on the contour of a cell, it follows that each $(\rho, \overline{Y})$-cell corresponds to a distinct pair of a maximal ρ-band and a maximal $\overline{Y}$-band. Thus, there are at most $O(n^3)$ ρ-cells. $\Box$

Lemma 5.12 *The number of d-cells is $O(n^4)$.*

Proof. We already proved that the number of $(d, \overline{Q})$-cells is $O(n^3)$ (Lemma 5.8), and the number of κ-cells is at most $O(n^2)$ (Lemma 5.4). The only other d-cells are the $(d, \overline{Y})$-cells. The number of maximal $\overline{Y}$-bands is $O(n^2)$ (Lemma 5.10), and the number of d-bands is $O(n^2)$ (Lemma 5.5). There are no $(d, \overline{Y})$-annuli (Lemma 4.29). A similar argument as in Lemma 5.8 shows that a maximal $\overline{Y}$-band cannot intersect a maximal d-band more than once. Indeed, by Remark 4 the only "bad" case is when both bands start or end on cells inside the region bounded by the medians of these bands. The cell where d-band starts or ends is a κ-cell. But the maximal κ-band containing this cell cannot cross the $\overline{Y}$-band or the d-band. This implies that the number of $(d, \overline{Y})$-cells is $O(n^4)$. $\Box$

Lemma 5.13 *The number of $(\overline{Y}, \Theta)$-cells is $O(n^3)$.*

Proof. We saw already that the number of $(\overline{Q}, \Theta)$-cells is $O(n^2)$ (Lemma 5.6), so we only have to count the $(\overline{Y}, \Theta)$-cells that have no $\overline{Q}$-edges.

We already proved that there are no Θ-annuli (Lemma 4.25), and no $\overline{Y}$-annuli (Lemma 4.32). We also saw that the number of Θ-bands is $O(n)$ (Lemma 5.2), and the number of maximal $\overline{Y}$-bands is $O(n^2)$ (Lemma 5.10). Since there are no $(\overline{Y}, \Theta)$-annuli (Lemma 4.31), and a maximal Θ-band cannot start on the contour of a cell, every Θ-band intersects every maximal $\overline{Y}$-band at most once. Hence the number of $(\overline{Y}, \Theta)$-cells which do not contain $\overline{Q}$-edges (commutation cells) is $O(n^3)$. $\Box$

Lemma 5.14 *The number of Θ-cells is $O(n^3)$.*

Proof. A Θ-cell is either a $(\overline{Y}, \Theta)$-cell (by Lemma 5.13, there are $O(n^3)$ of these), or a $(\overline{Q}, \Theta)$-cell (by Lemma 5.6, there are $O(n^2)$ of these), or a (κ, Θ)-cell (by Lemma 5.3, there are $O(n^2)$ of these). Adding these bounds yields the Lemma. $\Box$

Lemma 5.15 *The number of maximal B-bands is $O(n^4)$ and the perimeter of each G_b-cell is $O(n^2)$.*

Proof. By Lemma 4.36, no B-band can both start and end on a G_b-cell. Thus every maximal B-band must have at least one end either on the outer boundary of the van Kampen diagram (this accounts for at most n maximal B-bands), or on a d-cell. This accounts for at most $O(n^4)$ maximal B-bands, by Lemma 5.12. Thus there are at most $O(n^4)$ maximal B-bands.

By Lemma 4.34 for every G_b-cell there are at most 3 maximal B-bands starting on the contour of this cell and ending on the contour of the same $\overline{Y}$-band. Notice that every edge on the boundary of a G_b-cell is the start edge of a B-band. Since by Lemma 5.10 the total number of maximal $\overline{Y}$-bands is $O(n^2)$, the perimeter of each G_b-cell is $O(n^2)$. □

Lemma 5.16 *The number of $(B, \overline{Y})$-cells and the total number of $\overline{Y}$-cells are bounded by $O(n^6)$.*

Proof. We saw (Lemmas 5.15 and 5.10) that there are $O(n^4)$ maximal B-bands and $O(n^2)$ maximal $\overline{Y}$-bands. Moreover, we cannot have $(\overline{Y}, B)$-annuli by Lemma 4.29. Unfortunately, to exclude multiple intersections, we cannot use Remark 4 here because a $\overline{Y}$-band can start on a disc or a $\overline{Q}$-cell and a maximal B-band can start on a G_b-cell or on a d-cell. Suppose that a maximal $\overline{Y}$-band $\mathcal{Y}$ intersects a maximal B-band $\mathcal{T}$ more than once. Then as in the proof of Lemma 5.8 we can form a subdiagram Δ_1 bounded by medians of a subband $\mathcal{Y}_2$ of $\mathcal{Y}$ and a subband $\mathcal{T}_2$ of $\mathcal{T}$. Then Δ_1 must contain either a disc or a $\overline{Q}$-cell. The first option is impossible by Lemma 4.22 because the boundary of Δ_1 cannot contain κ-edges. The second option is also impossible because the boundary of Δ_1 does not contain $\overline{Q}$-edges and there are no $\overline{Q}$-annuli in the diagram Δ (by Lemma 4.9).

Therefore, there are at most $O(n^6)$ $(B, \overline{Y})$-cells.

Every $\overline{Y}$-cell is either a $(B, \overline{Y})$-cell (there are at most $O(n^6)$ as we just proved), or a $\overline{Q}$-cell (there are at most $O(n^3)$ of those, by Lemma 5.9), or a ρ-cell (there are at most $O(n^3)$ of those, by Lemma 5.11), or a d-cell (there are at most $O(n^4)$ of those, by Lemma 5.12), or a Θ-$\overline{Y}$-cell (there are at most $O(n^3)$ of those, by Lemma 5.13). Summing these bounds yields the Lemma. □

6 Estimating the Isoperimetric Function of $H_N(\mathcal{S})$

If a word w over the generators of H is equivalent to 1 in H then there exists a minimal van Kampen diagram over the disc-based presentation with boundary label w. Let $n = |w|$ be the perimeter of this van Kampen diagram.

The number of κ-, $\overline{Q}$-, ρ-, d-, Θ-, and $\overline{Y}$-cells in this van Kampen diagram is at most $O(n^6)$ (by Lemmas 5.4, 5.9, 5.11, 5.12, 5.14, 5.16).

The only other cells in the diagram are the discs and the G_b-cells. By Proposition 2.2 every disc of perimeter p can be filled in by a $G_N(\mathcal{S})$-diagram with area at most $O(T(O(p))^4)$. Moreover, by Lemma 5.10, the sum of the perimeters of all the discs in our diagram is $O(n^2)$, for some positive constant c. Hence by superadditivity of T^4 we have: the total area of all the discs in the diagram, when replaced by $G_N(\mathcal{S})$-diagrams is $O(T(O(n^2))^4)$.

Every G_b-cell with perimeter p can be replaced by a diagram with the same boundary label, consisting of two discs (of perimeter $O(p)$ each), and $O(p^2)$ cells over the original presentation of H ($O(p)$ ρ-cells and d-cells, and $O(p^2)$ (A, B) commutation cells); this follows from the proof of Lemma 3.1 and Fig. 6. Let us fix one of these diagrams for each G_b-cell in Δ. In Lemma 5.15, we proved that the perimeter of each G_b-cell does not exceed C_1n^2 for some constant C_1. Since the total perimeter of all G_b-cells does not exceed the number of maximal B-bands in Δ, Lemma 5.15 shows that this total perimeter does not exceed C_2n^4 for some constant C_2.

Then the set of all G_b-cells can be divided into subsets M_1,...,M_t where $t \le C_3n^2$, such that the sum of perimeters of G_b-cells in each subset is $\le C_1n^2$.

Since as we saw, every G_b-cell with perimeter m can be filled by a diagram (as in Fig. 6) with area at most $C_4T^4(C_5m)$ (see Proposition 2.2) for some constants C_4 and C_5, and since T^4 is a superadditive function, the sum of areas of diagrams filling the G_b-cells in each group M_i does not exceed $C_4T(C_5C_1n^2)^4$. Therefore the total sum of areas of these diagrams for all G_b-cells in Δ does not exceed $C_3C_4n^2T(C_5C_1n^2)^4$ which is equivalent to $n^2T(n^2)^4$.

Thus, we can transform a minimal diagram Δ with boundary label w into a diagram over the original presentation of $H_N(\mathcal{S})$ with the same boundary label and area bounded by a function equivalent to $n^2T(n^2)^4$. Since the total area of discs is equivalent to $T(n^2)^4$ and the number of other cells is $O(n^6)$, we conclude that $n^2T(n^2)^4$ is an isoperimetric function of the group $H_N(\mathcal{S})$ (since $T(n)$ growth at least linearly).

Remark 5. Our construction of group $H_N(\mathcal{S})$ can be viewed as a substantial development of the Novikov-Boone-Higman-Aanderaa construction (with improvements by Britton, Collins and Miller)[25]. The original construction (and all other known constructions for the Higman embedding theorem) contains Baumslag-Solitar relations $s^{-1}xs = x^2$ which lead to exponential lower bounds for the isoperimetric function of the groups. In [26] and here we eliminated the Baumslag-Solitar relations by replacing arbitrary Turing machines with S-machines. There exists another approach based on replacing Baumslag-Solitar relations with relations which define groups with polynomial Dehn functions. One could use, for example, embeddings of Baumslag-Solitar groups and similar groups into groups with polynomial Dehn functions. One could use also the R. Thompson group F [22] which has polynomial Dehn function [14] and an injective but not surjective endomorphism of linear growth. This approach will most probably give upper bounds for isoperimetric functions bigger than $n^2T(n^2)^4$.

7 Length Distortion

In this section we shall prove that the natural embedding of G into H (identifying G with G_b) has bounded length distortion. This means that there exists a constant $C > 0$ such that for every element $z \in G_b$ its length in G_b is at most C times its length in $H_N(\mathcal{S})$.

7.1 Shadows

Let us fix $N = 9ck$ where c is the constant from Lemma 4.33 (notice that c is twice the maximal number of $\overline{Y}$ letters in relations of the finite presentation of $G_N(\mathcal{S})$ and does not depend on N) and k is the number of components of the vector $\overline{Q}$ of the S-machine $\mathcal{S}$. Then the degree of every internal vertex of $\Gamma(\Delta)$ is $36ck - 3 > 33ck$. By Lemmas 4.20 and 4.21 $\Gamma(\Delta)$ is an ℓ-graph for $\ell = 33ck$. Consider an oval p passing through an internal vertex Π. By Lemma 2.15 p contains exactly two external edges $\mathcal{B}$ and $\mathcal{B}'$ and defines a region $\mathcal{O}(p)$. The external edges of p cut a subpath $\overline{p} = \overline{p}(\Pi, p) \subset \mathcal{O}(p)$ of the boundary $\partial(\Delta)$. (The end edges of the corresponding κ-bands $\mathcal{B}$ and $\mathcal{B}'$ do not belong to $\overline{p}$.) The path $\overline{p}$ is called the *shadow* of the disc Π determined by the oval p.

Let W be a word over the alphabet of generators of $H_N(\mathcal{S})$. Then $|W|_y$ denotes the number of $\overline{Y}$-letters in $|W|$. Let $|W|_{y,\theta}$ (resp. $|W|_{y,b,\theta}$), mean the number of $\overline{Y}$-letters (resp. $(\overline{Y} \cup B)$-letters) plus $4c$ times the number of Θ-letters in W. Thus we count every Θ-letter with weight $4c$.

The following Lemma is similar to Lemma 11.3 of [24].

Lemma 7.1 *Let W be a word written on the shadow $\overline{p}$ of a disc Π of a minimal diagram Δ. Let the label of Π be $\kappa(u)$ for some word u. Then $2|u|_{y,\theta} = 2|u|_y \leq |W|_{y,\theta}$.*

Proof. Induction on the number of vertices from $\Gamma(\Delta)$ belonging to $\mathcal{O}(p)$.

By the definition of oval at least $1/2(36ck - 3) + 4 > 18ck$ edges which are going out of the vertex Π, go in $\mathcal{O}(p)$. The word u contains $k_u \leq k$ maximal $\overline{Y}$-subwords (some of which may be empty). If u does not contain $\overline{Y}$-subwords then there is nothing to prove. So let us assume that u contains non-empty $\overline{Y}$-subwords.

Thus the total number of maximal $\overline{Y}$-subpaths of $q = \partial(\Pi) \cap \mathcal{O}(p)$ is at least $18ckk_u$. Ignoring the subpaths between the first two and the last two edges going from Π into $\mathcal{O}(p)$, we are left with at least $14ckk_u > 14ck$ middle subpaths. Let us denote $K = 14ck$ of these middle subpaths by $q_1, ..., q_K$. Let $\Delta(p)$ be the subdiagram of Δ consisting of all cells intersected by the oval p and those which are in $\mathcal{O}(p)$. Consider 2 cases.

(1) The number N_θ of maximal Θ-bands in the diagram $\Delta(p)$ is at least $\frac{1}{2c}|u|_y$. Notice that each of these bands must start or end on the shadow $\overline{p}$. Indeed otherwise by Lemma 2.17 we either have a (Θ, κ)-bigon ruled out by Lemma 4.5 or the median of the Θ-band would envelope a disc Π of $\Gamma(\Delta)$. This would lead to a contradiction with Lemma 4.14 as in the proof of Lemma 4.25. Therefore $|W|_{y,\theta} \geq 4cN_\theta \geq 2|u|_y$.

(2) Now suppose that $N_\theta < \frac{1}{2c}|u|_y$. Recall that by Lemma 4.25 Δ does not contain Θ-annuli. Consider the set Ω of maximal $\overline{Y}$-bands starting on $q_1, ..., q_K$. Notice that these bands cannot intersect κ-bands, so they are inside $\Delta(p)$. By Lemma 4.33, for every i from 1 to $K = 14ck$, the number of maximal $\overline{Y}$-bands starting on q_i which end up on cells of the same Θ-band is at most c. Therefore at most $14c^2k$ bands from Ω can have their end edges on cells belonging to the same Θ-band. Hence the number of bands from Ω ending on Θ-cells cannot be larger than $\frac{14c^2k}{2c}|u|_y = 7ck|u|_y$ because of the restriction on the number N_r of maximal Θ-bands. Therefore at least $14ck|u|_y - 7ck|u|_y = 7ck|u|_y$ bands from Ω end up on discs inside the diagram $\Delta(p)$ or on $\overline{p}$. (Notice that these $\overline{Y}$-bands cannot end on $\partial(\Pi)$ by Lemma 4.30.) Let us denote this set of $\overline{Y}$-bands by Ω_1.

Consider two bands $\mathcal{T}_1$ and $\mathcal{T}_2$ from Ω_1 which end on the same disc Π' (if such a disc exists). In the area bounded by the medians of these bands, there are no discs by Lemma 4.22. Therefore $\mathcal{T}_1$ and $\mathcal{T}_2$ start on q and end on $\overline{Y}$-parts of the boundary of Π' separated by at most 1 edge of the disc graph, that is by at most 4 κ-edges (recall that κ_1-, κ_2-, and κ_3-bands are not edges of the disc graph). This and the fact that the degree of every internal vertex of a disc graph is ≥ 28 allows us do define a derived oval p' passing through Π' (using the median of any $\overline{Y}$-band connecting Π and Π'; see Fig. 16a).

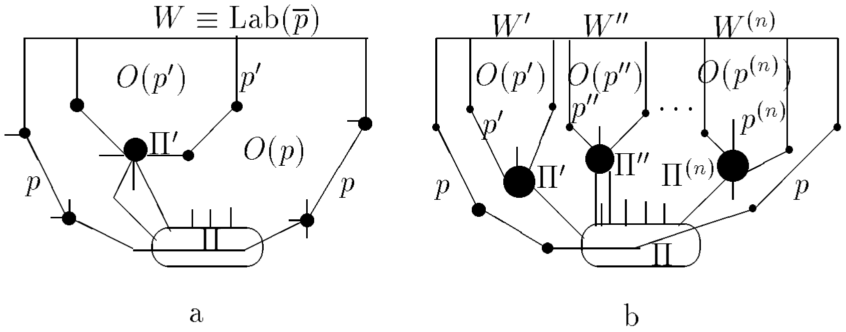

Fig. 16.

Let $p', ..., p^{(n)}$ be all ovals constructed for bands from the subset Ω_2 of Ω_1 consisting of all bands from Ω_1 which end on discs $\Pi', ..., \Pi^{(n)}$ (see Fig. 16b). From Lemma 2.18, it follows that the shadows of these discs do not intersect and are inside the shadow of Π. The end edges of the bands from $\Omega_3 = \Omega_1 \backslash \Omega_2$ are on the shadow $\overline{p}$ and cannot be on the shadows of the discs $\Pi', ..., \Pi^{(n)}$. Therefore for the labels $W', ..., W^{(n)}$ of these shadows we have:

$$|W'|_{y,\theta} + ... + |W^{(n)}|_{y,\theta} + card(\Omega_3) \leq |W|_{y,\theta}. \tag{3}$$

By Lemma 2.18 each of the regions $\mathcal{O}(p'), ..., \mathcal{O}(p^{(n)})$ contains fewer discs than $\mathcal{O}(p)$. By the induction hypothesis,

$$2|u^{(i)}|_y \leq |W^{(n)})|_{y,\theta} \tag{4}$$

for $i = 1, ..., n$ and word $u^{(i)}$ written on the boundary of Π' between two consecutive κ-letters. At the same time

$$5(|u'|_y + ... + |u^{(n)}|_y) \geq card(\Omega_2) \geq 7ck|u|_y - card(\Omega_3) \tag{5}$$

since different bands from Ω_2 end on no more than 5 different $(\overline{Y} \cup \overline{Q})$-parts of the boundary of a disc. From the inequalities 3-5, it follows that

$$2|u|_y \leq \sum_{i=1}^{n} \frac{10}{7}|u^{(i)}|_y + \frac{2}{7}card(\Omega_3) \leq \sum_{i=1}^{n} |W^{(i)}|_{y,\theta} + card(\Omega_3) \leq |W|_{y,\theta}.$$

□

7.2 Bounded Distortion

Let T be the subgroup of $H_N(\mathcal{S})$ generated by all generators of $H_N(\mathcal{S})$ except $\kappa_1, ..., \kappa_{4N}$ and generators from $\overline{Q}$. Then $G_b = \langle B \rangle$ is inside T. First we show that G_b has bounded distortion in T.

Lemma 7.2 *Let Δ be a minimal diagram whose boundary label is a word in generators of T. Let $\partial(\Delta) = pq$ where $u = \mathrm{Lab}(p)$ is a reduced word in the alphabet B. Then the length of the element u in G_b (that is the length of a geodesic which represents u in the Cayley graph of G_b) does not exceed the length of $|\mathrm{Lab}(q)|_{y,b}$.*

Proof. Let us assume that Δ has the smallest type among all minimal diagrams with $\partial(\Delta) = pq$ where $\mathrm{Lab}(p)$ is reduced and equal to u in G_b.

By Lemma 4.22 the diagram Δ contains no discs since its boundary contains no κ-edges.

Consider any maximal B-band $\mathcal{T}$ which starts on p. This band can end either on p or on a G_b-cell or on q or on a d-cell.

In the first case we can get a contradiction as in the proof of Lemma 4.36: the A-word written on the top path of $\mathcal{T}$ is 1 in the free group, so either there are two cells in this band that cancel (which contradicts the minimality of Δ) or the B-band is empty (which contradicts the assumption that $\mathrm{Lab}(p)$ is reduced).

In the second case, again as in the proof of Lemma 4.36, we can replace the subdiagram consisting of p, the band $\mathcal{T}$ and the G_b-cell by a new subdiagram with the same boundary label and one G_b-cell which is attached to p (this operation adds a number of (A, B)-commutation cells). Now we can remove the G_b-cell replacing p by a new B-path p' whose label is a reduced word which is equal to u in G_b. The new diagram will have smaller type than Δ but $\partial(\Delta') = p'q$ where $\mathrm{Lab}(p')$ is reduced and equal to u in G_b, a contradiction with the assumption that Δ has the smallest type among all such diagrams.

Thus we can assume that every maximal B-band starting on p ends either on q or on the contour of a d-cell.

Let Ω be the set of all maximal B-bands starting on p and let Ω_1 be the subset of Ω consisting of all bands ending on a d-cell. Then by Lemma 4.34 at most two bands of Ω end on cells of the same maximal $\overline{Y}$-band. Since Δ does not contain $\overline{Y}$-annuli (Lemma 4.32), and $\overline{Q}$-cells (Lemma 4.26), every maximal $\overline{Y}$-band starts and ends on $\overline{Y}$-edges of the boundary of Δ (recall that Δ contains no discs) we have a one to one correspondence between bands from Ω_1 and a subset of $\overline{Y}$-edges of q.

Thus the number of edges of p is smaller than the number of $\overline{Y} \cup B$-edges on q. $\square$

Lemma 7.3 *Let u be an element of G_b and let v be a word in generators of $H_N(\mathcal{S})$. Suppose that $u = v$ in $H_N(\mathcal{S})$. Then the length of u in G_b does not exceed $|v|_{y,b,\theta}$.*

Proof. Let Δ be the diagram of the smallest type with the following property: (P) $\partial(\Delta) = pq^{-1}$ where $\mathrm{Lab}(p)$ is a word over the generators of T and $\mathrm{Lab}(p) = u$ in T, $\mathrm{Lab}(q) \equiv v$. We know by Lemma 7.2 that $|\mathrm{Lab}(p)|_{y,b} \geq |u|$. So it is enough to show that

$|\mathrm{Lab}(p)|_{y,b} \leq |v|_{y,b,\theta}$.

No edge in p can belong to the contour of a $\overline{Y}$- or G_b-cell because otherwise by removing this cell we decrease the type of Δ preserving the property (P). Therefore every B-edge on p also belongs to q. Thus if p contains B-edges then by removing this edge we break our diagram into two parts; these parts will have smaller ranks than Δ, and we can consider them separately. Therefore we can assume that p does not contain B-edges and in general p and q do not have any edges in common.

Thus every $\overline{Y}$-edge on p belongs to the contour of either a $(\overline{Q}, \Theta)$-cell or a disc.

By Lemma 4.25 Δ does not contain Θ-annuli. Consider two cases.

(1) The number N_θ of maximal Θ-bands in Δ is at least $\frac{1}{4c}|\mathrm{Lab}(p)|_y$ where c is the constant from Lemma 4.33. Suppose that a Θ-band $\mathcal{T}$ starts and ends on p. Then the same argument as in the proof of Lemma 4.25 shows that the subdiagram Δ_1 bounded by $\mathcal{T}$ and p cannot contain κ-edges. It cannot contain $\overline{Q}$-edges also because p does not contain $\overline{Q}$-edges and Lemma 4.27 rules out $(\overline{Q}, \Theta)$-annuli. Therefore the top and the bottom paths of $\mathcal{T}$ are words over the generators of T. Thus removing the subdiagram Δ_1 from Δ would give us a diagram with smaller type still satisfying property (P). Therefore no Θ-band in Δ starts and ends on p. Thus q contains at least $\frac{1}{4c}|\mathrm{Lab}(p)|_y$ Θ-edges. Since the weight of each of these edges is $4c$, we get the desired inequality $|\mathrm{Lab}(p)|_{y,b} \leq |\mathrm{Lab}(q)|_{y,b,\theta}$.

(2) Now suppose that $N_\theta < \frac{1}{4c}|\mathrm{Lab}(p)|_y$. By Lemma 4.33 we deduce that at most c $\overline{Y}$-edges of p can belong to cells of the same Θ-band. Therefore the number of $\overline{Y}$-edges of p belonging to Θ-cells is at most $\frac{1}{4}|\mathrm{Lab}(p)|_y$.

Let Ω be the set of $\overline{Y}$-edges of p which belong to discs. We have established that $|\Omega| \geq \frac{3}{4}|\mathrm{Lab}(p)|_y$.

If some edges of Ω belong to the contour of the same disc then they cannot be separated on this disc by κ-edges. To prove that, consider the subdiagram between the disc and p. The κ-bands which start on the boundary of this subdiagram would have nowhere to end (there are no discs in this subdiagram by Lemmas 4.22 and 4.24).

Therefore the $\overline{Y}$-edges of p belong to one $(\overline{Y} \cup \overline{Q})$-part of the boundary of any disc Π which touches p. Therefore there exists an oval z in the disc graph of Δ passing through Π such that no $\overline{Y}$-edge on p belongs to the region $\mathcal{O}(z)$. In this case the shadow of the disc Π is contained in q because p contains no κ-edges. By Lemma 7.1 we have the following inequality for the label Z of this shadow:

$$|Z|_{y,\theta} \geq 2|w_\Pi|_y \tag{6}$$

where w_Π is the word written on the boundary of Π between two consecutive κ-edges.

The oval z can be chosen in such a way that every $\Gamma(\Delta)$-normal arc x with end on an edge belonging to Π and p is 2-separated from z. Then ovals z_1 and z_2 constructed in this way for two such discs Π_1 and Π_2 will be diverging because the edges e_1 and e_2 where these discs are touching p can be connected by suitable $\Gamma(\Delta)$-normal arc x. (Recall that there are no κ-edges between edges e_1 and e_2 on p.)

By Lemma 2.16 ovals z_1 and z_2 do not intersect in internal vertices of the disc graph $\Gamma(\Delta)$. Since the arc x is outside regions $\mathcal{O}(z_1)$ and $\mathcal{O}(z_2)$, these regions also do not intersect. Therefore the shadows q_1 and q_2 of z_1 and z_2 do not intersect.

Therefore we can sum up the inequalities 6 for all discs touching p and get

$$|\mathrm{Lab}(q)|_{y,\theta} \geq 2\sum |w_\Pi|_y \geq 2\ \frac{3}{4}|\mathrm{Lab}(p)|_y \geq |\mathrm{Lab}(p)|_y$$

as desired. □.

Lemmas 7.2 and 7.3 complete the proof that G_b has bounded length distortion in $H_N(\mathcal{S})$.

References

[1] S. Aanderaa. A proof of Higman's embedding theorem using Britton extensions of groups. In *Word Problems, Decision Problems and the Burnside problem in group theory*, pages 1–18, North-Holland, Publ. Corp., Amsterdam, London, 1973.

[2] J. Alonso, Inégalités isopérimétriques et quasi-isométries, C. R. Acad. Sci. Paris Série 1, 311 (1990), 761-764.

[3] G. Baumslag, M. R. Bridson, C. F. Miller III, and H. Short, "Finitely presented subgroups of automatic groups and their isoperimetric functions", J. London Math. Soc. (2) 56 (1997), no. 2, 292–304.

[4] G. Baumslag, C. F. Miller III, and H. Short. Isoperimetric inequalities and the homology of groups. Invent. Math., 113:531–560, 1993.

[5] J. C. Birget, "The complexity of the word problem for semigroups and the Higman embedding theorem", *International J. of Algebra and Computation* (to appear).

[6] Epstein, David B. A. et al., *Word processing in groups*, Jones and Bartlett, Boston, MA, 1992.

[7] S. M. Gersten, "Dehn functions and l_1-norms for finite presentations", in *Algorithms and Classification in Combinatorial Group Theory* (G. Baumslag, C. F. Miller, editors), MSRI Publications 23 (1992), Springer-Verlag.

[8] S. M. Gersten, "Isoperimetric functions of groups and exotic cohomology", in *Combinatorial and Geometric Group Theory*, (A.J. Duncan, N.D.Gilbert, J. Howie, editors), Edinburgh, 1993, London Mathematical Soc. Lecture Notes Series, 204, 1995.

[9] S.M. Gersten, "Preservation and distortion of area in finitely presented groups", GAFA vol 6 (1996) 301–345.

[10] M. Gromov, "Hyperbolic manifolds (according to Thurston and Jorgensen)". Bourbaki Seminar, Vol. 1979/80, pp. 40–53, Lecture Notes in Math., 842, Springer, Berlin-New York, 1981

[11] M. Gromov, "Hyperbolic manifolds, groups and actions", Riemann surfaces and related topics: Proceedings of the 1978 Stony Brook Conference (State Univ. New York, Stony Brook, N.Y., 1978), pp. 183–213, Ann. of Math. Stud., 97, Princeton Univ. Press, Princeton, N.J., 1981.

[12] M. Gromov, "Hyperbolic groups", in: *Essays in group theory* (S.M. Gersten, editor), MSRI Series 8, Springer-Verlag, 1987.

[13] M. Gromov, "Asymptotic invariants of infinite groups", in: *Geometric Group Theory, Volume 2* (G. A. Niblo and M. A. Roller, eds.), London Mathematical Society Lecture Notes Series 182, 1993.

[14] Victor Guba, " Polynomial upper bounds for the Dehn function of R.Thompson's group F", J. Group Theory, to appear.

[15] V. Guba and M. Sapir, *Diagram groups*, Memoirs of AMS, November 1997, v. 130, # 620.

[16] J. Hopcroft, J. Ullman, *Introduction to Automata Theory, Languages, and Computation*, Addison-Wesley, 1979.

[17] O. Kharlampovich, M. Sapir, *Algorithmic problems in varieties*, IJAC, 1995.

[18] Donald E. Knuth, *The art of computer programming. Vol. 2*, Second edition, Addison-Wesley Publishing Co., Reading, Mass., 1981.

[19] Jan Krajíček, *Bounded arithmetic, propositional logic, and complexity theory*, Cambridge University Press, Cambridge, 1995.

[20] R. Lyndon, P. Schupp, *Combinatorial Group Theory*, Springer-Verlag, 1977.

[21] K. Madlener, F. Otto, "Pseudo-natural algorithms for the word problem for finitely presented monoids and groups", *J. Symbolic Computation* 1 (1985) 383-418.

[22] R. McKenzie and R. J. Thompson. An elementary construction of unsolvable word problem in group theory. In W. W. Boon, F. B. Cannonito, and R. C. Lyndon, editors, *Word Problems*, pages 457–478, North-Holland, Publ. Corp., Amsterdam, London, 1973.

[23] A. Yu. Ol'shanskii, *The Geometry of Defining Relations in Groups*, Kluwer Academic Publishers, 1991.

[24] A. Yu. Ol'shanskii, "On distortion of subgroups in finitely presented groups", Mat. Sb., 1997, V.188, N 11, 51-98.

[25] J. Rotman, *An Introduction to the Theory of Groups*, Allyn & Bacon, 3rd edition, 1984.

[26] M. V. Sapir, J. C. Birget, E. Rips, "Isoperimetric and isodiametric functions of groups", submitted.

[27] D. Segal, *Polycyclic groups*, Cambridge Univ. Press, Cambridge, 1983.

Jean-Camille Birget
Department of Computer Science
University of Nebraska-Lincoln
birget@cse.unl.edu

Alexander Yu. Olshanskii
Department of Higher Algebra
MEHMAT. Moscow State University
olsh@nw.math.msu.su

Eliyahu Rips
Department of Mathematics
Hebrew University of Jerusalem
rips@math.huji.ac.il

Mark V. Sapir
Department of Mathematics
Vanderbilt University
http://www.math.vanderbilt.edu/~msapir