Statistics of Intersections of Curves on Surfaces

Rachel Zhang
Parkway South High School

Mentor: Professor Moira Chas, SUNY

PRIMES-USA

Abstract

Each orientable surface with nonempty boundary can be associated with a planar model, whose edges can then be labeled with letters that read out a surface word. Then, the curve word of a free homotopy class of closed curves on a surface is the minimal sequence of edges of the planar model through which a curve in the class passes. The length of a class of curves is defined to be the number of letters in its curve word.

We fix a surface and its corresponding planar model. Fix a free homotopy class of curves ω on the surface. For another class of curves c, let $i(\omega, c)$ be the minimal number of intersections of curves in ω and c. In this paper, we show that the mean of the distribution of $i(\omega, c)$, for random curve c of length n, grows proportionally with n and appraoches $\mu(\omega) \cdot n$ for a constant $\mu(\omega)$. We also give an algorithm to compute $\mu(\omega)$ and have written a program that calculates $\mu(\omega)$ for any curve ω on any surface. In addition, we prove that $i(\omega, c)$ approahces a Gaussian distribution as $n \to \infty$ by viewing the generation of a random curve as a Markov Chain.

1 Introduction

The study of closed curves up to deformation on surfaces is related to many other studies, including the structure of a Lie Algebra [1], periodic geodesics in non-Euclidean geometries [2], and the pattern of continued fractions, which are generalized by curves on the triple punctured sphere [3]. Deformation classes of curves also have applications in the statistics of primes [4], which relate to famous open problems in number theory such as the Twin Prime Conjecture. A property of curves that is fundamental to these related studies is a curve's geometric intersections. In this paper, we study the distribution of the minimal number of intersections of curves in two free homotopy classes of curves.

Surfaces and curves can be associated with *words*, or sequences of letters, in the following manner: On any orientable surface S with nonempty boundary, consider a maximal set of disjoint arcs with their endpoints on the boundary, such that removing these arcs does not render S disconnected. Both sides of each arc are associated with a distinct letter x and $\overline{x}$. Then to any free homotopy class of curves on S, we assign a *curve word*, defined by the shortest ordered sequence of the letters of edges of the arcs that a curve in the class enters. The *length* of a class of curves is the number of letters in its curve word.

Properties of curve words that imply intersections have been of interest for many years. In 1984, Joan S. Birman and Caroline Series [5] discovered a combinatorial algorithm that determines whether a curve is simple, or does not intersect itself, from its curve word. In 1987, Marshall Cohen and Martin Lustig [6] created an algorithm to count the number of intersections between two curves from their curve words. In 2004, Moira Chas [7] proved a bijection between intersections and linked pairs, which are broadly, pairs of words whose corresponding arcs must intersect.

One type of intersections is the *self intersection*, which is the minimal number of times a curve in a class geometrically intersects itself. Denote by C_n the set of all classes of closed curves of length $n \in \mathbb{N}$. Moira Chas and Steven P. Lalley [8]–[10] have studied the distribution of the number of self intersections of curves in C_n. In 2012, they proved that the distribution approaches a Gaussian distribution as $n \to \infty$ [11].

We investigate instead the distribution of the intersections of two classes of curves, with one fixed and the other varying. Fix a free homotopy class of closed curves ω on a surface S. Let c be a randomly chosen class of curves from C_n, and let $i(\omega, c)$ be the minimal number of intersections between curves in ω and c. We study the distribution of $i(\omega, c)$.

Computer generated data suggests that there exist positive constants $\mu(\omega)$ and $\sigma(\omega)$ such that the random variable $\frac{i(\omega,c)-\mu(\omega)\cdot n}{n^{\frac{1}{2}}} \to \sigma(\omega) \cdot N(0,1)$ as $n \to \infty$.

Our work involves the study of $\mu(\omega)$. In our results Theorem 5.3 and Theorem 5.7, we present a calculation for $\mu(\omega)$ for any curve ω, from which it follows that that $\mu(\omega)$ is rational. We have also developed an algorithm to calculate $\mu(\omega)$ and written a program that calculates $\mu(\omega)$ for any given closed curve ω.

In addition, we model a randomly generated curve of infinite length as the result of a Markov Chain. From the Markov Chain Central Limit Theorem [12], it follows that that a variable that estimates $i(\omega, c)$ approaches a normal distribution as $n \to \infty$. In Theorem 6.7, we use this result to prove that $i(\omega, c)$ approaches a Gaussian distribution as $n \to \infty$.

An outline of this paper is as follows. In Section 2, we present fundamental definitions regarding surfaces and curves. In Section 3, we present the theorem bijecting linked pairs to intersections and give a few examples of linked pairs corresponding to intersections. In Section 4, we give computer generated data supporting our conjecture that $i(\omega, c)$ approaches a Gaussian distribution as $n \to \infty$. Section 5 contains our main results regarding the expected value of $i(\omega, c)$ and its calculation. Then in Section 6, we introduce Markov Chains and their relevance to arbitrarily long curves. We then use the Markov Chain Central Limit Theorem to show that an estimate of $i(\omega, c)$ approaches a normal distribution as $n \to \infty$, which leads us to a proof that $i(\omega, c)$ approaches a Gaussian distribution as $n \to \infty$. Finally, in Section 7, we summarize our work and describe future goals.

2 Preliminaries

2.1 Surfaces

On an orientable surface with at least one boundary component, we consider a maximum number of disjoint arcs with both endpoints on boundary components such that the surface is not rendered disconnected. If there are k arcs, we can number the arcs from 1 to k and consider an *alphabet* $A = \{a_1, a_2, \ldots, a_k, \overline{a_1}, \overline{a_2}, \ldots, \overline{a_k}\}$ with $a_1, a_2, \ldots, a_k$ distinct letters. For each $i \in \{1, 2, \ldots, n\}$, we assign one side of the arc with a_i and the other with $\overline{a_i}$. Upon removing the arcs, the surface is a polygon with $4k$ sides, with alternating edges labeled with letters such that each letter in the alphabet is used exactly once. This polygon formed is called the *planar model*. A *surface word* associated to a surface is the cyclic sequence of letters read from the planar model in a counterclockwise direction. For surface S, we choose its surface word and denote it by $W(S)$. We also define the *length of a surface word* to be the number of letters in a surface word, equal to twice the number of cuts made in forming the planar model. For surface S, we denote the length of its surface word as $|S|$.

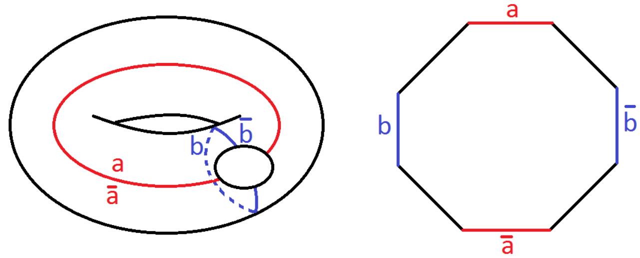

Figure 2.1: Torus with One Boundary

Example 2.1. In Figure 2.1, the torus with one boundary component can be cut along the red and blue arcs which are labeled with a and $\overline{a}$, and b and $\overline{b}$, respectively. Upon removing these cuts, we attain an octogon with alternating sides labeled with $a, b, \overline{a}$, and $\overline{b}$. Hence we read the surface word as $ab\overline{a}\overline{b}$.

2.2 Curves

We consider free homotopy classes of directed curves on a fixed surface S. Consider a curve u on S. Suppose that u intersects a minimal number of the arcs which are removed from S to form the planar model. A class of curves can be labeled by the ordered sequence of letters of the edges it enters. We call this sequence of letters the *curve word*; for curve u, let its curve word be $W(u)$. We define the *curve length* anagolous to the length of a surface word: it is the number of letters in the curve word. We denote the curve length of a curve u as $|u|$.

When u intersects a minimal number of arcs, its curve word is *reduced*, meaning that for a curve word $u_1u_2u_3 \ldots u_{|u|}$, $u_i \neq \overline{u_{i+1}}$ for $1 \leq i \leq k-1$, where $\overline{\overline{x}} = x$. For instance, the curve word $aa\overline{a}ab$ is not reduced, while aab is. For the remainder of this paper, a curve word denotes a reduced curve word.

Our study involves, more specifically, classes of closed curves. Note that for classes of closed curves, the curve word is *cyclic*, or word $u_ju_{j+1} \ldots u_{|u|}u_1 \ldots u_{j-1}$ is equivalent to $u_1u_2 \ldots u_{|u|}$. Hence we consider *cyclically reduced curve words* for closed curves, meaning that $u_i \neq \overline{u_{i+1}}$ for $1 \leq i \leq k-1$ and $u_k \neq \overline{u_1}$. We use $W_j(u)$ to refer to the curve word shifted by $j-1$ places, $u_ju_{j+1} \ldots u_{|u|}u_1 \ldots u_{j-1}$.

We have two types of words: cyclic and *linear*, or non-cyclic. Cyclic words correspond to closed curves while linear ones are associated with open curves.

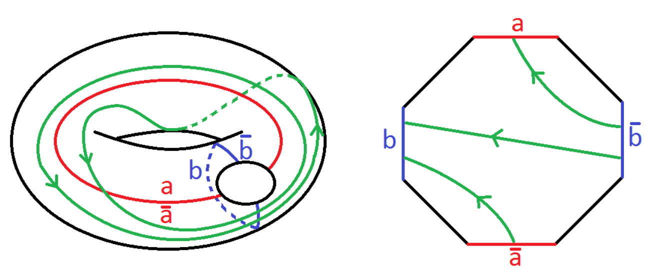

Figure 2.2: Curve abb on Torus with One Boundary

Example 2.2. Figure 2.2 displays a closed curve, drawn in green, on the torus with one boundary. Upon forming the planar model $ab\overline{a}\overline{b}$, the curves become green arcs between labeled edges, as shown. Then the curve word is abb, as the curve enters edges a, b, and b, in that order.

Example 2.3. Figure 2.3 shows open curve $\overline{b}\overline{a}\overline{a}$ on surface $ab\overline{a}\overline{b}$.

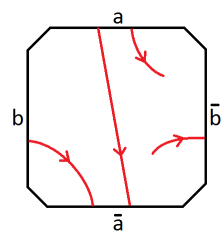

Figure 2.3: Open Curve $\overline{b}\overline{a}\overline{a}$ on $ab\overline{a}\overline{b}$

Definition 2.4. Let the *representative* of a word w be the class of open curves with curve word w if w is linear, or the class of closed curves with curve word w if w is cyclic.

The power of curve words can be seen in the following theorem:

Theorem 2.5. *[13] Fix a planar model of an orientable surface S. There is a bijection between curve words and free homotopy classes of closed curves on S.*

Most of the topological properties of a free homotopy class of curves are recorded in a string of letters. This allows us to study classes of curves by simply studying their corresponding curve words.

If open curve u has curve word $u_1u_2u_3\dots u_{|u|}$, let $\overline{u}$ be the open curve such that $W(\overline{u}) = \overline{u_{|u|}u_{|u|-1}u_{|u|-2}}\dots\overline{u_1}$. We let $\overline{W(u)} = \overline{u_{|u|}u_{|u|-1}u_{|u|-2}}\dots\overline{u_1}$, so that $W(\overline{u}) = \overline{W(u)}$. Essentially, if we reverse the direction of each curve in u, we obtain $\overline{u}$.

3 Intersections of Two Curves

We now consider the interactions of two curves on a surface.

Definition 3.1. For any two free homotopy classes of curves u and v on the same surface S, the *minimal intersection number* of u and v, denoted by $i(u,v)$, is the smallest number of intersection points of u and v, counted with multiplicity.

Properties of curve words that imply intersections have been studied by many people, notably by Moira Chas [7]. In this section, we review her results.

3.1 Extended Planar Model

For a surface S, fix a planar model and a surface word. Choose an open curve u on S with curve word $u_1u_2\dots u_{|u|}$. Then an *extended planar model* of S is $|u|-1$ ordered planar models where edge u_i of the ith planar model is connected to edge $\overline{u_i}$ of the $(i+1)$th planar model for $1 \le i \le |u|-1$. Curve u can be drawn on this planar model, starting from edge $\overline{u_1}$ of the first planar model and entering edge u_{i+1} of the ith planar model.

Example 3.2. Figure 3.1 shows the extended planar model of open curve $\overline{b}a\overline{b}\overline{b}$.

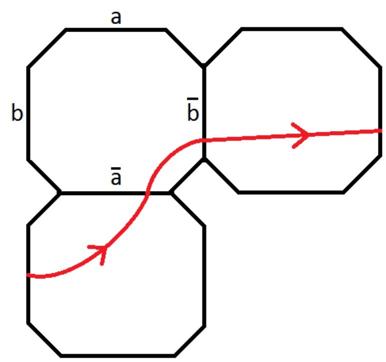

Figure 3.1: Extended Planar Model of Curve $\bar{b}a\bar{b}\bar{b}$ on Torus with One Boundary

3.2 Linked Pairs Theorem

First we introduce a few definitions.

Definition 3.3. A *power* of a word W is the word formed by concatenating W with itself a number of times. In particular, let W^k denote the word formed by concatenating W with itself k times.

For example, the word *abbabbabb* is a power of the word *abb*. Specifically, $abbabbabb = (abb)^3$.

Definition 3.4. We say that word u' is a *subword* of a word W if u' is a substring of W^k for some $k \in \mathbb{N}$.

Note that a subword is a linear word, so the representatives of subwords are directed open curves.

Definition 3.5. A *linked pair* is a pair of linear words (p, q) with $|p| = |q|$ such that one of the following conditions hold:

1. $p = p_1p_2$ and $q = q_1q_2$ for $p_1, p_2, q_1, q_2 \in A$
2. $p = p_1Rp_2$ and $q = q_1Rq_2$ for a linear word R and $p_1, p_2, q_1, q_2 \in A$
3. $p = p_1Rp_2$ and $q = q_1\overline{R}q_2$ for a linear word R and $p_1, p_2, q_1, q_2 \in A$,

and any representatives of p and q must intersect when placed in the same extended planar model.

Figure 3.2 demonstrates a linked pair of the second type. The red edges are letters in R. Note that for linked pair (p, q) of the second type, $(p, \overline{q})$ is a linked pair of the third type.

This leads us to an important theorem:

Theorem 3.6. *(Chas, M. (2004)) [7] For free homotopy classes of curves u and v, consider a pair of representatives of u and v that intersect minimally. There is a bijection between intersections and linked pairs (u', v'), for u' a subword of $W(u)$ and v' a subword of $W(v)$.*

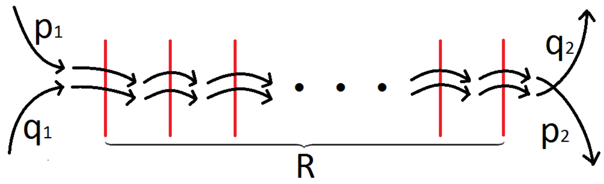

Figure 3.2: Linked Pair

Specifically, the intersection associated with a linked pair (p, q) is the intersection of the representatives of p and q in their shared extended planar model. We demonstrate the linked pairs corresponding to intersections in the following example:

Example 3.7. Displayed in Figure 3.3 are the curves a and abb on surface $ab\bar{a}\bar{b}$. We observe a minimal of two intersections.

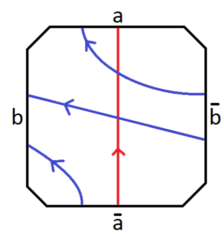

Figure 3.3: Curves a and abb on $ab\bar{a}\bar{b}$

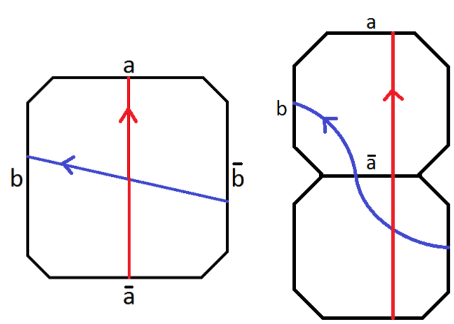

Figure 3.4: Extended Planar Model of Linked Pairs (aa, bb) and (aaa, bab)

Pairs (aa, bb) and (aaa, bab) are linked pairs, as shown in Figure 3.4. Both aa and aaa are subwords of a while bb and bab are both subwords of abb. Then the two intersections of a and abb correspond to the linked pairs (aa, bb) and (aaa, bab).

4 Empirical Distribution

With a computer, for a fixed curve ω and positive integer n, we can study the distribution of $i(\omega, c)$ for $c \in C_n$. We used a program in Java that computes the complete set C_n and determines the number of linked pairs (ω', c') for ω' a subword of $W(\omega)$ and c' a subword of $W(c)$. By running this program and, for each k, determining the number of classes of curves c satisfying $i(\omega, c) = k$, we created histograms.

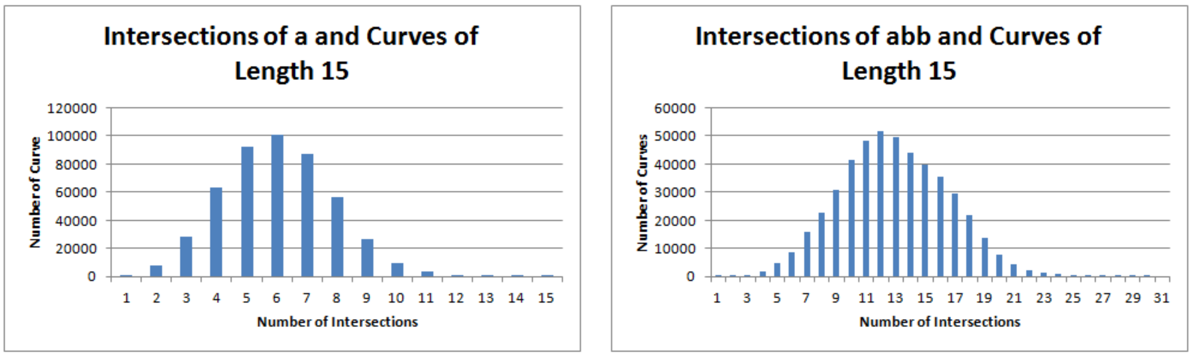

Figure 4.1: Intersections with a and abb on $ab\overline{a}\overline{b}$

Since it takes $O(c^n)$ time to generate C_n, we only ran the program for $n \leq 15$. A clear pattern emerges rather quickly; even at $n = 15$, the distribution appears to approach a normal distribtuion. We see this in Figure 4.1 for $\omega = a$ and $\omega = abb$ on the torus with one boundary component with surface word $ab\overline{a}\overline{b}$.

5 Expected Number of Intersections of a Fixed Curve with Curves of Length n

In this section, we look at the mean $\mathbb{E}_{c \in C_n} i(\omega, c)$ of the distribution of $i(\omega, c)$ for fixed class of curves ω and a random class of curves c in C_n. We show that $\mathbb{E}\frac{i(\omega,c)}{n}$ approaches a constant $\mu(\omega)$ as $n \to \infty$, which we can calculate.

5.1 Formula for the Mean

Recall that a multiset is a set in which repeated elements are considered distinct. Because intersections are implied by linked pairs, we simply need to find the multiset C' of all words c' for which there exists a subword ω' of $W(\omega)$ such that (ω', c') is a linked pair. We also consider a multiset D' of half the words in C' of length at least 3 that will be used in Lemma 5.1 and Theorem 5.7.

Determining all elements in C' can be done by considering individually each subword ω' of our fixed curve word $W(\omega)$. We denote the first character of ω' by ω'_1, the last by ω'_2, and the middle $|\omega'| - 2$ characters by R, such that $\omega' = \omega'_1 R \omega'_2$. To find a word c' such that (ω', c') is a linked pair, we search through all values of c'_1 and c'_2 for ones that give that $(\omega', c'_1 R c'_2)$ is a linked pair. Then we add $c' = c'_1 R c'_2$ to C' and, if $|c'| \geq 3$, to D'. Since $\overline{c'}$ also gives $(\omega', \overline{c'})$ is a linked pair, we must also add $\overline{c'}$ to C'. Define $C'_{[i,j]}$ to be the multiset of all words in C' with length at least i and at most j. Then $C' = C'_{[2,2]} \cup D' \cup (\cup_{c' \in D'} \{\overline{c'}\})$.

Let $E_{c'}$ be the expected number of times a word c' is a subword of a random curve word of length n, divided by n. The process of randomly generating k consecutive letters in a curve word is equivalent to choosing the first at random from $|S|$ possibilities, and each successive letter from $|S|-1$ possibilites. Then, we can give $E_{c'}$ an explicit formula:

$$E_{c'} = \frac{1}{|S|(|S|-1)^{|c'|-1}}.$$

Let $\mu_m(\omega)$ be the sum of $E_{c'}$ for all $|c'| \leq m$. Then $\mu_m(\omega)$ is equal to

$$\begin{aligned}\mu_m(\omega) &= \sum_{c' \in C', |c'| \leq m} E_{c'} \\ &= \sum_{c' \in C', |c'| \leq m} \frac{1}{|S|(|S|-1)^{|c'|-1}}.\end{aligned}$$

We want to show that the sequence $\{\mu_m(\omega)\}_m$ approaches a constant $\mu(\omega)$. To facilitate calculations, we first show that we can partition the curves in $C'_{[3,\infty)}$ into sets $[c']$ for $c' \in C'_{[3,|\omega|+2]}$, such that $[c']$ contains exactly one word of length $|c'| + k|\omega|$ for $k \geq 0$.

Lemma 5.1. *Each word $c' \in C'_{[3,|\omega|+2]}$ determines a set of words $[c'] \in C'_{[3,\infty)}$, with $c'' \in [c']$ if there exists i and j such that $c' = c'_1 R c'_2$, and c'' is either $c'_1 R[W_j(\omega)]^i c'_2$ or $c'_1 R[\overline{W_j(\omega)}]^i c'_2$. Each word $c'' \in C'_{[3,\infty)}$ belongs to exactly one of these sets.*

Proof. Define $D'_{[i,j]}$ to be the multiset of all words in D' with length at least i and at most j. It suffices to prove that each word $c' \in D'_{[3,|\omega|+2]}$ determines an set of words $[c'] \in D'_{[3,\infty)} = D'$ and that each word $c'' \in D'$ belongs to exactly one of these sets.

For any curve $c' \in D'_{[3,|\omega|+2]}$, let $c' = c'_1 R c'_2$ for $c'_1, c'_2 \in A$ and R a subword of $W(\omega)$. If $W(\omega) = \omega_1 \omega_2 \ldots \omega_{|\omega|}$, suppose that $R = \omega_j \omega_{j+1} \ldots \omega_{j+|R|-1}$ for some positive integer j and indices taken modulo $|\omega|$, such that $(\omega_{j-1} R \omega_{j+|R|}, c'_1 R c'_2)$ is a linked pair. Then let any word c'' be in $[c']$ if $c'' = c'_1 R[W_{j+|R|}(\omega)]^k c'_2$ for some integer value of $k \geq 0$. Note that $[c']$ contains exactly one element of length $|c'| + k|\omega|$ for each $k \geq 0$.

It is clear that $c'' \in D'_{[3,\infty)}$ since $(\omega_{j-1} R[W_{j+|R|}(\omega)]^k \omega_{j+|R|}, c'')$ is a linked pair. Also, given $c'' \in D'_{[3,\infty)}$, we can write c'' as $c''_1 R[W_{j+|R|}(\omega)]^k c''_2$ for some values of c''_1, R, j, k, and c''_2, with $|R| \leq |\omega|$. Then $c' = c''_1 R c''_2 \in D'_{[3,|\omega|+2]}$, or $c'' \in [c']$. □

We can now group the words in $C'_{[3,\infty)}$ by sets $[c']$ for $c' \in C'_{[3,|\omega|+2]}$. For a fixed $c' \in C'_{[3,|\omega|+2]}$, let $\mu_m(\omega, c')$ be the sum of all $E_{c''}$ for $c'' \in [c']$ and $|c''| \leq m$. In the following Lemma, we show that the sequence $\{\mu_m(\omega, c')\}_m$ approaches a constant $\mu(\omega, c')$.

Lemma 5.2. *For some $c' \in C'_{[3,|\omega|+2]}$, the sequence $\{\mu_m(\omega, c')\}_m$ approaches a constant $\mu(\omega, c')$ equal to*

$$\frac{(|S|-1)^{|\omega|}}{(|S|-1)^{|\omega|}-1} \cdot E_{c'},$$

with

$$|\mu_m(\omega, c') - \mu(\omega, c')| = \frac{E_{c'}}{(|S|-1)^{|\omega|}-1} \cdot \left[\frac{1}{(|S|-1)^{|\omega|}}\right]^{\left\lfloor \frac{m-|c'|}{|\omega|} \right\rfloor}.$$

Proof. From Lemma 5.1, we have that, for $c' \in C'_{[3,|\omega|+2]}$, $[c']$ contains exactly one word of length $|c'| + k|\omega|$ for each integer $k \geq 0$. Then

$$\begin{aligned}
\mu_m(\omega, c') &= \sum_{c'' \in [c'], |c''| \leq m} E_{c''} \\
&= \sum_{c'' \in [c'], |c''| \leq m} \frac{1}{|S|(|S|-1)^{|c''|-1}} \\
&= \frac{1}{|S|(|S|-1)^{|c'|-1}} \cdot \sum_{k=0}^{\left\lfloor \frac{m-|c'|}{|\omega|} \right\rfloor} \left[\frac{1}{(|S|-1)^{k|\omega|}}\right] \\
&= \frac{1}{|S|(|S|-1)^{|c'|-1}} \cdot \frac{(|S|-1)^{|\omega|} - \left[\frac{1}{(|S|-1)^{|\omega|}}\right]^{\left\lfloor \frac{m-|c'|}{|\omega|} \right\rfloor}}{(|S|-1)^{|\omega|}-1} \\
&= \frac{(|S|-1)^{|\omega|} - \left[\frac{1}{(|S|-1)^{|\omega|}}\right]^{\left\lfloor \frac{m-|c'|}{|\omega|} \right\rfloor}}{(|S|-1)^{|\omega|}-1} \cdot E_{c'}.
\end{aligned}$$

Let

$$\mu(\omega, c') = \frac{(|S|-1)^{|\omega|}}{(|S|-1)^{|\omega|}-1} \cdot E_{c'}.$$

For any $\epsilon > 0$, there is a value of m_0 for which

$$\begin{aligned}
|\mu_m(\omega, c') - \mu(\omega, c')| &= \frac{E_{c'}}{(|S|-1)^{|\omega|}-1} \cdot \left[\frac{1}{(|S|-1)^{|\omega|}}\right]^{\left\lfloor \frac{m-|c'|}{|\omega|} \right\rfloor} \\
&< \epsilon
\end{aligned}$$

for all $m > m_0$. Therefore, $\mu(\omega, c')$ is the limit of the sequence $\{\mu_m(\omega, c')\}_m$. □

Let $g_\omega(k)$ be the number of elements of C' with length equal to k. This leads us to our following result about the mean of the limiting distribution.

Theorem 5.3. *Fix a surface S and its surface word. For fixed class of curves ω, $\{\mu_m(\omega)\}_m$ converges to*

$$\mu(\omega) = \left[\frac{g_\omega(2)}{|S|(|S|-1)}\right] + \left[\frac{(|S|-1)^{|\omega|}}{(|S|-1)^{|\omega|}-1} \cdot \sum_{3 \leq k \leq |\omega|+2} \frac{g_\omega(k)}{|S|(|S|-1)^{k-1}}\right],$$

such that

$$|\mu_m(\omega) - \mu(\omega)| = \sum_{c' \in C'_{[3,|\omega|+2]}} \left(\frac{E_{c'}}{(|S|-1)^{|\omega|} - 1} \cdot \left[\frac{1}{(|S|-1)^{|\omega|}} \right]^{\left\lfloor \frac{m-|c'|}{|\omega|} \right\rfloor} \right).$$

Proof. Let $\epsilon > 0$. Set $\mu(\omega)$ equal to

$$\begin{aligned}
\mu(\omega) &= \sum_{c' \in C'_{[2,2]}} E_{c'} + \sum_{c' \in C'_{[3,|\omega|+2]}} \mu(\omega, c') \\
&= \sum_{c' \in C'_{[2,2]}} E_{c'} + \frac{(|S|-1)^{|\omega|}}{(|S|-1)^{|\omega|} - 1} \sum_{c' \in C'_{[3,|\omega|+2]}} E_{c'} \\
&= \left[\sum_{c' \in C'_{[2,2]}} \frac{1}{|S|(|S|-1)} \right] + \left[\frac{(|S|-1)^{|\omega|}}{(|S|-1)^{|\omega|} - 1} \sum_{c' \in C'_{[3,|\omega|+2]}} \frac{1}{|S|(|S|-1)^{|c'|-1}} \right] \\
&= \left[\frac{g_\omega(2)}{|S|(|S|-1)} \right] + \left[\frac{(|S|-1)^{|\omega|}}{(|S|-1)^{|\omega|} - 1} \cdot \sum_{3 \le k \le |\omega|+2} \frac{g_\omega(k)}{|S|(|S|-1)^{k-1}} \right].
\end{aligned}$$

We have that

$$\mu_m(\omega) = \sum_{c' \in C'_{[2,2]}} E_{c'} + \sum_{c' \in C'_{[3,|\omega|+2]}} \mu_m(\omega, c').$$

Because by Lemma 5.2,

$$|\mu_m(\omega, c') - \mu(\omega, c')| = \frac{E_{c'}}{(|S|-1)^{|\omega|} - 1} \cdot \left[\frac{1}{(|S|-1)^{|\omega|}} \right]^{\left\lfloor \frac{m-|c'|}{|\omega|} \right\rfloor},$$

there exists m_0 such that for all $m > m_0$,

$$\begin{aligned}
|\mu_m(\omega) - \mu(\omega)| &= \left| \left(\sum_{c' \in C'_{[2,2]}} E_{c'} + \sum_{c' \in C'_{[3,|\omega|+2]}} \mu_m(\omega, c') \right) - \left(\sum_{c' \in C'_{[2,2]}} E_{c'} + \sum_{c' \in C'_{[3,|\omega|+2]}} \mu(\omega, c') \right) \right| \\
&= \left| \sum_{c' \in C'_{[3,|\omega|+2]}} \big(\mu_m(\omega, c') - \mu(\omega, c') \big) \right| \\
&= \sum_{c' \in C'_{[3,|\omega|+2]}} \left(\frac{E_{c'}}{(|S|-1)^{|\omega|} - 1} \cdot \left[\frac{1}{(|S|-1)^{|\omega|}} \right]^{\left\lfloor \frac{m-|c'|}{|\omega|} \right\rfloor} \right) \\
&< \epsilon.
\end{aligned}$$

□

Corollary 5.4. $\mathbb{E}_{c \in C_n} \frac{i(\omega,c)}{n}$ *is bounded.*

Proof. This is clear as $g_\omega(k)$ is bounded for all k: Because there are $2 \cdot |\omega|$ choices for ω' for which (ω', c') is a linked pair, and less than $|\omega|$ choices for each of c_1' and c_2', we have that

$$g_\omega(k) < 2 \cdot |\omega|^3.$$

Furthermore, there are finitely many $3 \leq k \leq |\omega| + 2$. □

Corollary 5.5. $\lim_{n\to\infty} \mathbb{E}_{c \in C_n} \frac{i(\omega,c)}{n} \in \mathbb{Q}$.

Proof. The expression obtained for $\lim_{n\to\infty} \mathbb{E}_{c \in C_n} \frac{i(\omega,c)}{n}$ in Theorem 5.3 is a finite sum of rational numbers, so the sum is also a rational number. □

But we can calculate $g_\omega(k)$ from the curve word of ω. For fixed surface S, we fix a planar model and surface word. We define a function $d : A^2 \to \mathbb{Z}$. On surface word $z_1 z_2 \dots z_{|S|}$, let

$$d(z_i, z_j) = \begin{cases} j - i - 1 & \text{if } i < j \\ |S| + i - j - 1 & \text{if } i > j \\ 0 & \text{if } i = j. \end{cases}$$

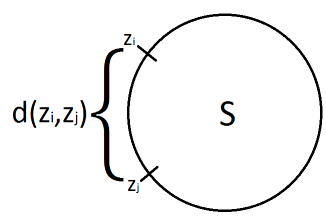

Figure 5.1: $d(z_i, z_j)$

Note that d is essentially counting in a counterclockwise direction the number of labeled edges in the planar model of S between z_i and z_j.

Definition 5.6. Fix a planar model and surface word of a surface S. We say that an edge z_l is *between* two edges z_i and z_j if either $i < l < j$, $j < i < l \leq |S|$, or $l < j < i$.

Then $d(z_i, z_j)$ counts the number of edges z_l that are between z_i and z_j. Now we can find a formula for $g_\omega(k)$.

Theorem 5.7. *For fixed class of curves ω on a fixed surface S with fixed planar model and surface word, let the curve word of ω be $\omega_1 \omega_2 \dots \omega_{|\omega|}$. Then the function $g_\omega(k)$ is given by*

$$g_\omega(k) = \begin{cases} 2 \cdot \sum_{i=1}^{|\omega|} \left[d(\overline{\omega_i}, \omega_{i+1}) \cdot d(\omega_{i+1}, \overline{\omega_i}) \right] & \text{if } k = 2 \\ 2 \cdot \sum_{i=1}^{|\omega|} \left[d(\overline{\omega_i}, \omega_{i+1}) \cdot d(\omega_{i+k-1}, \overline{\omega_{i+k-2}}) + d(\omega_{i+1}, \overline{\omega_i}) \cdot d(\overline{\omega_{i+k-2}}, \omega_{i+k-1}) \right] & \text{if } 3 \leq k \leq |\omega| + 2, \end{cases}$$

where indices are taken modulo $|\omega|$.

Proof. Since for linked pair (ω', c'), we have that $|\omega'| = |c'|$, we consider all subwords ω' of $W(\omega')$ of length k for each $3 \leq k \leq |\omega| + 2$.

First we consider $k = 2$. Each ω' of length 2 is of the form $\omega_i \omega_{i+1}$ for $1 \leq i \leq |\omega|$ and indices taken modulo $|\omega|$. Then either c_1' is between $\overline{\omega_i}$ and ω_{i+1} and c_2' is between ω_{i+1} and $\overline{\omega_i}$, or c_1' is between ω_{i+1} and $\overline{\omega_i}$ and c_2' is between $\overline{\omega_i}$ and ω_{i+1}. The first case gives $d(\overline{\omega_i}, \omega_{i+1}) \cdot d(\omega_{i+1}, \overline{\omega_i})$ possibilities, and the second gives $d(\omega_{i+1}, \overline{\omega_i}) \cdot d(\overline{\omega_i}, \omega_{i+1})$ possiblilities, for a total of $2 \cdot [d(\overline{\omega_i}, \omega_{i+1}) \cdot d(\omega_{i+1}, \overline{\omega_i})]$ possible values of (c_1', c_2'). Summing across all values of i gives

$$g_\omega(2) = 2 \cdot \sum_{i=1}^{|\omega|} \left[d(\overline{\omega_i}, \omega_{i+1}) \cdot d(\omega_{i+1}, \overline{\omega_i}) \right].$$

Now consider $3 \leq k \leq |\omega| + 2$. For $\omega' = \omega_i \omega_{i+1} \ldots \omega_{i+k-2} \omega_{i+k-1}$, with indices taken modulo $|\omega|$, let $R = \omega_{i+1} \omega_{i+2} \ldots \omega_{i+k-2}$. Since there is a bijection between words $c' = c_1' R c_2' \in D'$ and $\overline{c_2'} \overline{R} \overline{c_1'}$, $g_\omega(k)$ counts twice as many words as there are words of length k in D'. Therefore we can count the number of linked pairs (ω', c') with $c' = c_1' R c_2' \in D'$ for $c_1', c_2' \in A$ and multiply by 2 to account for words $c' = \overline{c_2'} \overline{R} \overline{c_1'}$ not in D'.

Then we are looking for the number of pairs (c_1', c_2') such that $(\omega', c_1' R c_2')$ is a linked pair. There are two cases. The first is if c_1' is between $\overline{\omega_i}$ and ω_{i+1} and c_2' is between ω_{i+k-1} and $\overline{\omega_{i+k-2}}$, which gives a total of $d(\overline{\omega_i}, \omega_{i+1}) \cdot d(\omega_{i+k-1}, \overline{\omega_{i+k-2}})$ possible pairs (c_1', c_2'). The other case is where c_1' is between ω_{i+1} and $\overline{\omega_i}$, and c_2' is between $\overline{\omega_{i+k-2}}$ and ω_{i+k-1}, which gives $d(\omega_{i+1}, \overline{\omega_i}) \cdot d(\overline{\omega_{i+k-2}}, \omega_{i+k-1})$ possibilites. If we sum these two cases across all values of i and multiply by 2, we get that

$$g_\omega(k) = 2 \cdot \sum_{i=1}^{|\omega|} \left[d(\overline{\omega_i}, \omega_{i+1}) \cdot d(\omega_{i+k-1}, \overline{\omega_{i+k-2}}) + d(\omega_{i+1}, \overline{\omega_i}) \cdot d(\overline{\omega_{i+k-2}}, \omega_{i+k-1}) \right].$$

□

Once $g_\omega(k)$ has been found for $2 \leq k \leq |\omega| + 2$, $\lim_{n \to \infty} \mathbb{E} \frac{i(\omega, c)}{n}$ can be calculated. A computer can both determine $g_\omega(k)$ as well as perform the calculations. We have written such a program that outputs the mean of the limiting distribution of $\frac{i(\omega, c)}{n}$, using the process described in this section. The output of the program is two relatively prime positive integers k_1 and k_2 such that $\mu(\omega) = \frac{k_1}{k_2}$. The value of $\frac{k_1}{k_2}$ matches the experimental mean calculated by another program that considers $i(\omega, c)$ for each $c \in C_n$ individually.

5.2 Example Calculation

We now calculate $\mu(\omega)$ for $\omega = abb$ to demonstrate the process of calculation concretely.

Example 5.8. Let $\omega = abb$ on surface $ab\overline{a}\overline{b}$. We calculate $\mu(abb)$.

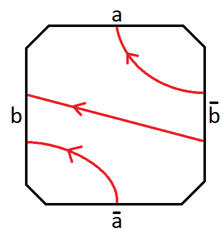

Figure 5.2: Curve abb on surface $ab\bar{a}\bar{b}$

First we find $g_{abb}(2)$. All subwords ω' of abb of length 2 are $\{ab, bb, ba\}$. Since $d(\bar{b}, a) = 0$, there are no words c' of length 2 such that (ab, c') or (ba, c') is a linked pair. Hence the only linked pairs with $|\omega'| = 2$ are (bb, aa) and $(bb, \overline{aa})$, or $g_{abb}(2) = 2$.

Now we need to find $g_{abb}(k)$ for $3 \leq k \leq |abb| + 2 = 5$. In the following table, we consider each value of k and find each subword ω' of abb of length k. Then if $\omega' = \omega'_1\omega'_2 \ldots \omega'_k$, we find $d(\overline{\omega'_1}, \omega'_2)$, $d(\omega'_k, \overline{\omega'_{k-1}})$, $d(\omega'_2, \overline{\omega'_1})$, and $d(\overline{\omega'_{k-1}}, \omega'_k)$. Then the total possible values of c' such that (ω', c') is a linked pair is $2 \cdot [d(\overline{\omega'_1}, \omega'_2) \cdot d(\omega'_k, \overline{\omega'_{k-1}}) + d(\omega'_2, \overline{\omega'_1}) \cdot d(\overline{\omega'_{k-1}}, \omega'_k)]$, which we also record, under the "Total" column.

$\|\omega'\|$	ω'	$d(\overline{\omega'_1}, \omega'_2)$	$d(\omega'_k, \overline{\omega'_{k-1}})$	$d(\omega'_2, \overline{\omega'_1})$	$d(\overline{\omega'_{k-1}}, \omega'_k)$	Total
3	abb	2	1	0	1	4
3	bba	1	2	1	0	4
3	bab	0	0	2	2	8
4	$abba$	2	2	0	0	8
4	$bbab$	1	0	1	2	4
4	$babb$	0	1	2	1	4
5	$abbab$	0	2	2	0	0
5	$bbabb$	1	1	1	1	4
5	$babba$	0	2	2	0	0

Then $g_{abb}(3) = 16$, $g_{abb}(4) = 16$, and $g_{abb}(5) = 2$. Now we calculate $\mu(abb) = \sum_{c' \in C'} E_{c'}$:

$$\begin{aligned}
\sum_{c' \in C'} E_{c'} &= \sum_{c' \in C'[2,2]} E_{c'} + \sum_{c' \in C'_{[3,|\omega|+2]}} \mu(\omega, c') \\
&= \left[\frac{g_{abb}(2)}{|S|(|S|-1)}\right] + \left[\frac{27}{26} \cdot \sum_{3 \leq k \leq |abb|+2} \frac{g_{abb}(k)}{|S|(|S|-1)^{k-1}}\right] \\
&= \frac{2}{4 \cdot 3} + \frac{27}{26} \cdot \left(\frac{16}{4 \cdot 3^2} + \frac{16}{4 \cdot 3^3} + \frac{4}{4 \cdot 3^4}\right) \\
&= \frac{31}{39}.
\end{aligned}$$

6 Gaussian Distribution

We have studied the mean of the distribution, and now we look at its shape. In this section, we prove that $i(\omega, c)$ approaches a Gaussian distribution as $n \to \infty$. First, we must introduce Markov Chains, which are instrumental to this proof.

6.1 Markov Chains

A *Markov Chain* is a sequence of random variables $(X_i : i \geq 0)$ that take on values in a state space $\mathbb{S}$ and satisfy

$$P(X_{n+1} = x | X_n = X_n, X_{n-1} = X_{n-1}, \ldots, X_0 = x_0) = P(X_{n+1} = x | X_n = x_n).$$

In other words, in a Markov Chain, the value of the next variable X_{n+1} is only dependent on the value of X_n.

For our problem, we study only *finite state Markov Chains*, or where $\mathbb{S}$ is a finite space.

The probabilities at time t associated with transitioning from one state to another are called *transition probabilities.* For a finite state Markov Chain whose transition probabilities between states are independent, we can assign a *transition matrix* P such that $P_{i,j}$ is the probability of transitioning from state i to state j.

A finite state Markov Chain is *irreducible* if it is possible to get from any state to another. A state r is *recurrent* if $P(X_i = r \text{ for infinitely many } i) = 1$. For a recurrent state r, let the *first return time of* r be equal to $\tau(r) = \inf\{i \geq 1 : X_i = r\}$. A Markov Chain is *recurrent* if all its states are recurrent. For a recurrent Markov Chain, $P(X_n = x)$ as $n \to \infty$ approaches a fixed constant for each value of $x \in \mathbb{S}$. For finite state Markov Chains that are both irreducible and recurrent, there is a unique *stationary distribution* $\pi(x)$ of the Markov Chain, which is the probability that $X_n = x$ for $n \to \infty$, or $\pi(x) = P(X_n = x | n \to \infty)$.

We can also define a *reward function* $f : \mathbb{S} \to \mathbb{R}$ to be a function on the states of a Markov Chain to the real numbers. Finally, to state a formula for the variance, we let τ_0 and τ_1 be independently and identically distributed positive random variables. Then we have the following Central Limit Theorem for finite state Markov Chains:

Theorem 6.1. *(Markov Chain Central Limit Theorem [12]) Let $X = (X_n : n \geq 0)$ be an irreducible, positive, and recurrent Markov chain on a discrete state space $\mathbb{S}$. Then,*

$$\frac{\sum_{i=0}^{n-1} f(X_i) - n\pi f}{n^{\frac{1}{2}}} \to \sigma N(0,1)$$

as $n \to \infty$ for $\sigma^2 = \frac{(\mathbb{E}[\sum_{j=\tau_0}^{\tau_0+\tau_1}(f - \pi f)(X_j)])^2}{\mathbb{E}[\tau_1]}$.

6.2 Markov Chain Estimate of $i(\omega, c)$

As we are looking at the limiting distribution of $i(\omega, c)$, we are looking at curves of arbitrarily large length. Hence we can view the process of generating a random curve $c \in C_\infty$ as generating an infinite sequence $(y_i : i \geq 0)$ of random characters $y_i \in A$, where $P(y_{i+1} = \overline{y_i}) = 0$ and $P(y_{i+1} = x) = \frac{1}{|S|-1}$ for $x \neq \overline{y_i}$.

The process of selecting each character one by one can be seen as a Markov Chain with state space equal to the alphabet A. However, we can instead let the state space $\mathbb{S}_m$ of a Markov chain be all reduced words of length m for some m. Then the transition matrix P has $P_{i,j} = \frac{1}{|S|-1}$ if the word formed by the last $m-1$ characters of i is equal to the word formed by the first $m-1$ characters of j, and $P_{i,j} = 0$ otherwise. For instance, for surface $S = ab\bar{a}\bar{b}$, $P_{abb,bb\bar{a}} = \frac{1}{3}$ and $P_{b\overline{aa}\bar{b},\overline{aa}b} = 0$. Then generating a random curve $c \in C_\infty$ can be viewed instead as a sequence of random variables $(Y_i | i \geq 0)$, where Y_i corresponds to the word $y_{i+1}y_{i+2}\ldots y_{i+m}$.

The stationary distribution $\pi(x)$ of this Markov Chain has $\pi(x) = \frac{1}{|S|(|S|-1)^{m-1}}$ for all x, since Y_i is equally likely to equal each word of length m.

We say that a word w *begins* another word w' if the string formed from the first $|w'|$ letters of w is equal to w'. Then we can define a reward function $f(x) : \mathbb{S}_m \to \mathbb{R}$ to be the number of words $c' \in C'$ that x begins. Then $f(x) < \infty$ for all x because x has finite length and hence can only begin finitely many elements of C'. Essentially, $f(x)$ accounts for all words in C' with length less than or equal to $|x|$, so that we count occurrences of words $c' \in C'$ with $|c'| \leq m$ in the Markov Chain.

In random curve $c = y_1y_2\ldots y_n$, let $N(c, [i, j])$ be the random variable that counts the number of times a word $c' \in C'$ with length between i and j inclusive appears as a substring $y_iy_{i+1}\ldots y_{i+|c'|-1}$ of c. Then $N(c, [2, m]) = \sum_{i=0}^{n-m} f(Y_i)$. For large m, $N(c, [2, m])$ offers an approximation of $i(\omega, c)$. Furthermore, $N(c, [2, m])$ approaches a normal distribution as $n \to \infty$, which we state in the following lemma:

Lemma 6.2. *For any $m \in \mathbb{N}$,*

$$\frac{N(c, [2, m]) - n\pi f}{n^{\frac{1}{2}}} \to \sigma(\omega)N(0, 1)$$

for $c \in C_n$ as $n \to \infty$ for some constant $\sigma(\omega)$.

Proof. Applying Theorem 6.1 to $N(c, [2, m])$, π, and f as defined gives the stated result. □

In other words, for any $m \in \mathbb{N}$, the Markov Chain Central Limit Theorem is sufficient to prove normality for the number of occurences of the elements in $C'_{[2,m]}$.

Define $G_m = \sigma(\omega) \cdot N(0, 1)$, or the limiting distribution of $\frac{N(c,[2,m])}{n}$ as $n \to \infty$, and let h_m be the random variable with distribution G_m.

It remains to generalize this normality to when all elements of C' are considered. But as the length of a word $c' \in C'$ increases, the expected number of times it appears as a subword of the curve word of a curve $c \in C_n$ for fixed n decreases, so intuitively, the effect of longer words $c' \in C'$ on the value of $i(\omega, c) = N(c, [2, \infty])$ should be small. Furthermore, as m increases, $N(c, [2, m])$ yields a sharper and sharper estimate of $i(\omega, c)$ for $c \in C_n$. The fact that $\frac{N(c,[2,m])-n\pi f}{n^{\frac{1}{2}}} \to \sigma N(0, 1)$ as $n \to \infty$ for any arbitrarily large m suggests that $N(c, [2, \infty]) = i(\omega, c)$ approaches a Gaussian distribution as $n \to \infty$ as well. We prove this in the following section.

6.3 Proof of Gaussian Distribution

Consider the sequence $\{G_m\}_m$. If it has a limit G, then let h be the random variable with distribution G. We want to show that the distribution of $\frac{i(\omega,c)}{n}$ approaches G. This bounding can be split into considering three differences: $\frac{i(\omega,c)}{n}$ and $\frac{N(c,[2,m])}{n}$, $\frac{N(c,[2,m])}{n}$ and h_m, and h_m and h.

In Theorem 6.7, we assert the existence of the limit G of $\{G_m\}$ as a result of Theorem 5.3 and Lemma 6.6. We also consider the difference in the third difference between h_m and h.

The second difference, between $\frac{N(c,[2,m])}{n}$ and h_m, is a result of Lemma 6.2. Hence we have left to consider the first difference between $\frac{i(\omega,c)}{n}$ and $\frac{N(c,[2,m])}{n} = \frac{N(c,[m+1,\infty])}{n}$.

Let $p(c, c'')$ be the number of times that c'' is a subword of c, for word c'' and curve word c of length n. For $n \in \mathbb{N}$ and $c' \in C'_{[3,|\omega|+2]}$, define $X(c', n, m)$ to be the maximal possible number of occurences of words $c'' \in [c']$ with $|c''| > m$ in c, divided by n, over all $c \in C_n$, or:

$$X(c', n, m) = \max_{c \in C_n} \sum_{c'' \in [c'], |c''| > m} \frac{p(c, c'')}{n}.$$

Then $\frac{N(c,[m+1,\infty])}{n} \leq \sum_{c' \in C'_{[3,|\omega|+2]}} X(c', n, m)$.

We say that a word c' is *located* at a position i in a curve word $W(c)$ for $c \in C_n$ if $c = y_1 y_2 \dots y_n$ and $c' = y_i y_{i+1} \dots y_{i+|c'|-1}$. This leads us to the following lemma limiting the number of words c'' in fixed $[c']$ that are located at a certain position.

Lemma 6.3. *At most one word $c'' \in [c']$ can be located at any position i in a curve word $W(c)$, for a fixed $c' \in C'_{[3,|\omega|+2]}$.*

Proof. We first prove that for any two distinct words c'' and c''' both belonging to the same set $[c']$ for $c' \in C'_{[3,|\omega|+2]}$, and for any $1 \leq i \leq n - \max(|c''|, |c'''|) + 1$, at most one of c'' and c''' can be located at position i of curve word $W(c)$. Suppose that $c' \in D'_{[3,|\omega|+2]}$ (the other case follows similarly). Also, suppose without loss of generalization that $|c''| < |c'''|$. Then we can write $c'' = c'_1 R c'_2$ and $c''' = c'_3 R [W_j(\omega)]^k c'_4$ for some $k \geq 1$.

Suppose for the sake of contradiction that c'' and c''' are both located at posititition i. Then c'' equals the first $|c''|$ letters of c''', or $c'_1 R c'_2 = c'_3 R \omega_j$, so $c'_2 = \omega_j$. As $c'' \in C'$, the pair $(c'', \omega_{j-|R|-1} R \omega_j) = (c'_1 R c'_2, \omega_{j-|R|-1} R \omega_j)$ is a linked pair, but this is a contradiction as $c'_2 = \omega_j$ contradicts the definition of a linked pair.

Now suppose there are K words $c'' \in [c']$ located at a position i. If $K = 0$, then the result clearly follows. Else, consider c'' located at position i. Then no other $c''' \in [c']$ can also be located at i, or $K = 1$. □

The following lemma is another bound on, for fixed c', the number of $c'' \in [c']$ that can exist in a curve word.

Lemma 6.4. *For $m \geq |\omega| + 2$ and two words c'' and c''', with $|c''|, |c'''| > m$, both belonging to the same set $[c']$ for $c' \in C'_{[3,|\omega|+2]}$, suppose c'' is located in c at position i_1 and c''' is located at position i_2 with $i_1 < i_2$. Then $i_2 - i_1 \geq |c''| - |\omega| - 2$.*

Proof. Suppose that $c' \in D'_{[3,|\omega|+2]}$ as the other case is similar. Let $c'' = c''_1 R_1 c''_2$ and $c''' = c'''_1 R_2 c'''_2$. For the sake of contradiction, suppose that $i_2 - i_1 < |c''| - |\omega| - 2$. Then the first $|\omega|$ characters of R_2, equal to $W_j(\omega)$ for some j, are contained within R_1. Furthermore, since $i_2 > i_1$, $c'''_1 W_j(\omega)$ is contained within R_1, implying that $c'''_1 = \omega_{j-1}$, contradiction. □

Finally, we bound $X(c', n, m)$ for $c' \in C'_{[3,|\omega|+2]}$ with a value that depends on m and is independent of n.

Lemma 6.5. *For $m \geq |\omega| + 2$, $c' \in C'_{[3,|\omega|+2]}$, $0 \leq X(c', n, m) < \frac{1}{m-|\omega|-1}$.*

Proof. First recall that $0 \leq \frac{N(c,[m+1,\infty])}{n} \leq X(c', n, m)$, hence $X(c', n, m) \geq 0$.

Let K be the number of occurences of words $c'' \in [c']$ with $|c''| \geq m + 1$ in $c \in C_n$. If $c = y_1 y_2 \dots y_n$, let the words $c'' \in [c']$ of length greater than m contained in c be $c''_1, c''_2, \dots, c''_K$. Furthermore, suppose that c''_j is located at position i_j, and that $i_1 \leq i_2 \leq \cdots \leq i_K$. By Lemma 6.3, $i_j \neq i_{j+1}$, so we have $i_1 < i_2 < \cdots < i_K$. By Lemma 6.4, $i_{j+1} - i_j \geq |c''_j| - |\omega| - 2$. Then

$$\begin{aligned}
K \cdot (m - |\omega| - 1) &= K \cdot [(m+1) - |\omega| - 2] \\
&= (K-1) \cdot [(m+1) - |\omega| - 2] + [(m+1) - |\omega| - 2] \\
&< \sum_{j=1}^{K-1} \left[|c''_j| - |\omega| - 2 \right] + |c''_K| \\
&\leq n,
\end{aligned}$$

or

$$K < \frac{n}{m - |\omega| - 1}.$$

Since $X(c', n, m) = \frac{K}{n}$, we have that

$$\begin{aligned}
X(c', n, m) &< \frac{\frac{n}{m-|\omega|-1}}{n} \\
&= \frac{1}{m - |\omega| - 1}.
\end{aligned}$$

□

We now have an upper bound for $X(c', n, m)$ that decreases as m increases.

Before we prove that $i(\omega, c)$ approaches a normal distribution as $n \to \infty$, we must first consider its variance. Let $\sigma^2_{n,m}(\omega)$ be the variance of the distribution of $\frac{N(c,[2,m])}{n}$. By Lemma 6.2, we have that $\{\sigma^2_{n,m}(\omega)\}_n$ is a sequence that converges to a constant $\sigma^2_m(\omega)$. We consider the sequence $\{\sigma^2_m(\omega)\}_m$ and show that this also converges to a constant $\sigma^2(\omega)$.

Lemma 6.6. *There exists a value of $\sigma^2(\omega)$ such that, for $m \geq |\omega| + 2$,*

$$|\sigma_m(\omega) - \sigma(\omega)| < \frac{2 \cdot |C'_{[3,|\omega|+2]}|}{m - |\omega| - 1}$$

and

$$|\sigma^2_m(\omega) - \sigma^2(\omega)| < \frac{4 \cdot |C'_{[3,|\omega|+2]}|}{m - |\omega| - 1} \cdot \sigma(\omega) + \left(\frac{2 \cdot |C'_{[3,|\omega|+2]}|}{m - |\omega| - 1}\right)^2.$$

Then $\{\sigma^2_m(\omega)\}_m$ converges to $\sigma^2(\omega)$.

Proof. For sufficiently large $m_1 < m_2$, note that

$$\begin{aligned} |\sigma_{m_1}(\omega) - \sigma_{m_2}(\omega)| &< \sum_{c' \in C'_{[3,|\omega|+2]}} X(c', \infty, m_1) \\ &< \frac{|C'_{[3,|\omega|+2]}|}{m_1 - |\omega| - 1}. \end{aligned}$$

Since for any $\epsilon > 0$, we can choose m_1 arbitrarily large so that $\frac{|C'_{[3,|\omega|+2]}|}{m_1 - |\omega| - 1} < \epsilon$, $\{\sigma_m(\omega)\}_m$ is a Cauchy sequence that converges to a constant $\sigma(\omega)$, with

$$|\sigma_m(\omega) - \sigma(\omega)| < \frac{2 \cdot |C'_{[3,|\omega|+2]}|}{m - |\omega| - 1}.$$

Then,

$$\sigma(\omega) - \frac{2 \cdot |C'_{[3,|\omega|+2]}|}{m - |\omega| - 1} < \sigma_m(\omega) < \sigma(\omega) + \frac{2 \cdot |C'_{[3,|\omega|+2]}|}{m - |\omega| - 1},$$

or

$$\begin{aligned} \sigma^2(\omega) - \frac{4 \cdot |C'_{[3,|\omega|+2]}|}{m - |\omega| - 1}\sigma(\omega) + \left(\frac{2 \cdot |C'_{[3,|\omega|+2]}|}{m - |\omega| - 1}\right)^2 &< \sigma^2_m(\omega) \\ &< \sigma^2(\omega) + \frac{4 \cdot |C'_{[3,|\omega|+2]}|}{m - |\omega| - 1} \cdot \sigma(\omega) + \left(\frac{2 \cdot |C'_{[3,|\omega|+2]}|}{m - |\omega| - 1}\right)^2, \end{aligned}$$

so

$$|\sigma_m^2(\omega) - \sigma^2(\omega)| < \frac{4 \cdot |C'_{[3,|\omega|+2]}|}{m - |\omega| - 1} \cdot \sigma(\omega) + \left(\frac{2 \cdot |C'_{[3,|\omega|+2]}|}{m - |\omega| - 1}\right)^2.$$

For any $\epsilon > 0$, there exists m such that $\frac{4 \cdot |C'_{[3,|\omega|+2]}|}{m-|\omega|-1} \cdot \sigma(\omega) + \left(\frac{2 \cdot |C'_{[3,|\omega|+2]}|}{m-|\omega|-1}\right)^2 < \epsilon$, hence $\sigma_m^2(\omega)$ converges to $\sigma^2(\omega)$. □

We can now prove that the distribution of $i(\omega, c)$ approaches a Gaussian distribution as $n \to \infty$.

Theorem 6.7. *Fix a free homotopy closed class of curves ω on a surface S. Then for $c \in C_n$ for $n \in \mathbb{N}$,*

$$\frac{i(\omega, c) - \mu(\omega) \cdot n}{n^{\frac{1}{2}}} \to \sigma(\omega) N(0,1)$$

as $n \to \infty$ for positive constant $\sigma(\omega)$.

Proof. Recall that h_m is the random variable with distribution G_m, the limiting distribution of $\frac{N(c,[2,m])}{n}$, which by Lemma 6.2 is a Gaussian distribution. Since the sequence $\{G_m\}_m$ has controlled mean by Theorem 5.3 and controlled variance by Lemma 6.6, it approaches a limit Gaussian distribution, G. Let h be the random variable with distribution G.

Let $\mathbb{P}_{(a,b)}(x)$ denote the probability that a variable x falls in the range (a, b). Consider $\epsilon > 0$. It suffices to show that there exists n_0 such that for all $n > n_0$,

$$\left|\mathbb{P}_{(a,b)}\left(\frac{i(\omega, c)}{n}\right) - \mathbb{P}_{(a,b)}(h)\right| < \epsilon.$$

By the triangle inequality,

$$\begin{aligned}\left|\mathbb{P}_{(a,b)}\left(\frac{i(\omega, c)}{n}\right) - \mathbb{P}_{(a,b)}(h)\right| &\leq \left|\mathbb{P}_{(a,b)}\left(\frac{i(\omega, c)}{n}\right) - \mathbb{P}_{(a,b)}\left(\frac{N(c,[2,m])}{n}\right)\right| \\ &+ \left|\mathbb{P}_{(a,b)}\left(\frac{N(c,[2,m])}{n}\right) - \mathbb{P}_{(a,b)}(h_m)\right| \\ &+ \left|\mathbb{P}_{(a,b)}(h_m) - \mathbb{P}_{(a,b)}(h)\right|.\end{aligned}$$

We proceed to show that there exists m_0 and n_0 so that for $m > m_0$ and $n > n_0$, each summand is less than $\frac{\epsilon}{3}$.

First, set $\epsilon_1 = \frac{|C'_{[3,|\omega|+2]}|}{m-|\omega|-1}$, $\epsilon_2 = \frac{2 \cdot |C'_{[3,|\omega|+2]}|}{m-|\omega|-1}$, and $\epsilon_3 = \sum_{c' \in C'_{[3,|\omega|+2]}} \left(\frac{E_{c'}}{(|S|-1)^{|\omega|-1}} \cdot \left[\frac{1}{(|S|-1)^{|\omega|}}\right]^{\left\lfloor \frac{m-|c'|}{|\omega|} \right\rfloor}\right)$.

There exists $m_0 \geq |\omega|+2$ such that for all $m > m_0$, the following three inequalities are all satisfied:

$$\frac{|C'_{[3,|\omega|+2]}|}{m-|\omega|-1} < \frac{\epsilon}{3} \tag{6.1}$$

$$\left|\frac{\epsilon_3 \cdot \sigma(\omega) + \epsilon_2 \cdot a - \epsilon_2 \cdot \mu(\omega)}{\sigma(\omega) \cdot (\sigma(\omega) - \epsilon_2)}\right| < \frac{\epsilon}{6} \tag{6.2}$$

$$\left|\frac{\epsilon_3 \cdot \sigma(\omega) + \epsilon_2 \cdot b - \epsilon_2 \cdot \mu(\omega)}{\sigma(\omega) \cdot (\sigma(\omega) - \epsilon_2)}\right| < \frac{\epsilon}{6}. \tag{6.3}$$

We first consider the first summand, $\left|\mathbb{P}_{(a,b)}\left(\frac{i(\omega,c)}{n}\right) - \mathbb{P}_{(a,b)}\left(\frac{N(c,[2,m])}{n}\right)\right|$, which we want to show is less than $\frac{\epsilon}{3}$ for sufficiently large m ($m > m_0$), independent of n. By Lemma 6.5, we have that

$$\begin{aligned}\left|\frac{i(\omega,c)}{n} - \frac{N(c,[2,m])}{n}\right| &\leq \sum_{c' \in C'_{[3,|\omega|+2]}} X(c',n,m) \\ &< \frac{|C'_{[3,|\omega|+2]}|}{m-|\omega|-1}.\end{aligned}$$

Recall that $\epsilon_1 = \frac{|C'_{[3,|\omega|+2]}|}{m-|\omega|-1}$. Then from Equation 6.1, for $m > m_0$, $\epsilon_1 = \frac{|C'_{[3,|\omega|+2]}|}{m-|\omega|-1} < \frac{\epsilon}{3}$. We have the following two equations:

$$\begin{aligned}\mathbb{P}_{(a,b)}\left(\frac{i(\omega,c)}{n}\right) &\geq \mathbb{P}_{(a,b-\epsilon_1)}\left(\frac{N(c,[2,m])}{n}\right) \\ &= \mathbb{P}_{(a,b)}\left(\frac{N(c,[2,m])}{n}\right) - \mathbb{P}_{(b-\epsilon_1,b)}\left(\frac{N(c,[2,m])}{n}\right),\end{aligned}$$

and

$$\begin{aligned}\mathbb{P}_{(a,b)}\left(\frac{i(\omega,c)}{n}\right) &\leq \mathbb{P}_{(a-\epsilon_1,b)}\left(\frac{N(c,[2,m])}{n}\right) \\ &= \mathbb{P}_{(a-\epsilon_1,a)}\left(\frac{N(c,[2,m])}{n}\right) + \mathbb{P}_{(a,b)}\left(\frac{N(c,[2,m])}{n}\right).\end{aligned}$$

These imply that

$$\begin{aligned}\left|\mathbb{P}_{(a,b)}\left(\frac{i(\omega,c)}{n}\right) - \mathbb{P}_{(a,b)}\left(\frac{N(c,[2,m])}{n}\right)\right| &\leq \max\left(\mathbb{P}_{(b-\epsilon_1,b)}\left(\frac{N(c,[2,m])}{n}\right), \mathbb{P}_{(a-\epsilon_1,a)}\left(\frac{N(c,[2,m])}{n}\right)\right) \\ &\leq \epsilon_1 \\ &< \frac{\epsilon}{3}.\end{aligned}$$

The next summand is $\left|\mathbb{P}_{(a,b)}\left(\frac{N(c,[2,m])}{n}\right) - \mathbb{P}_{(a,b)}(h_m)\right|$. By Lemma 6.2, we have that there exists n_0 such that for $n > n_0$, $\left|\mathbb{P}_{(a,b)}\left(\frac{N(c,[2,m])}{n}\right) - \mathbb{P}_{(a,b)}(h_m)\right| < \frac{\epsilon}{3}$.

Finally, we have the summand $\left|\mathbb{P}_{(a,b)}(h_m) - \mathbb{P}_{(a,b)}(h)\right|$. Let h_N be the random variable with distribution equal to $N(0,1)$. Then

$$\mathbb{P}_{(a,b)}(h_m) = \mathbb{P}_{\left(\frac{a-\mu_m(\omega)}{\sigma_m(\omega)}, \frac{b-\mu_m(\omega)}{\sigma_m(\omega)}\right)}(h_N)$$

and

$$\mathbb{P}_{(a,b)}(h) = \mathbb{P}_{\left(\frac{a-\mu(\omega)}{\sigma(\omega)}, \frac{b-\mu(\omega)}{\sigma(\omega)}\right)}(h_N).$$

We then have that

$$\begin{aligned}
\left|\mathbb{P}_{(a,b)}(h_m) - \mathbb{P}_{(a,b)}(h)\right| &= \left|\mathbb{P}_{\left(\frac{a-\mu_m(\omega)}{\sigma_m(\omega)}, \frac{b-\mu_m(\omega)}{\sigma_m(\omega)}\right)}(h_N) - \mathbb{P}_{\left(\frac{a-\mu(\omega)}{\sigma(\omega)}, \frac{b-\mu(\omega)}{\sigma(\omega)}\right)}(h_N)\right| \\
&\leq \mathbb{P}_{\left(\min\left(\frac{a-\mu_m(\omega)}{\sigma_m(\omega)}, \frac{a-\mu(\omega)}{\sigma(\omega)}\right), \max\left(\frac{a-\mu_m(\omega)}{\sigma_m(\omega)}, \frac{a-\mu(\omega)}{\sigma(\omega)}\right)\right)}(h_N) \\
&+ \mathbb{P}_{\left(\min\left(\frac{b-\mu_m(\omega)}{\sigma_m(\omega)}, \frac{b-\mu(\omega)}{\sigma(\omega)}\right), \max\left(\frac{b-\mu_m(\omega)}{\sigma_m(\omega)}, \frac{b-\mu(\omega)}{\sigma(\omega)}\right)\right)}(h_N) \\
&\leq \left|\frac{a-\mu_m(\omega)}{\sigma_m(\omega)} - \frac{a-\mu(\omega)}{\sigma(\omega)}\right| + \left|\frac{b-\mu_m(\omega)}{\sigma_m(\omega)} - \frac{b-\mu(\omega)}{\sigma(\omega)}\right|.
\end{aligned}$$

From Lemma 6.6, we have that $|\sigma_m(\omega) - \sigma(\omega)| < \frac{2 \cdot |C'_{[3,|\omega|+2]}|}{m-|\omega|-1} = \epsilon_2$, and from Theorem 5.3,

$$|\mu_m(\omega) - \mu(\omega)| = \sum_{c' \in C'_{[3,|\omega|+2]}} \left(\frac{E_{c'}}{(|S|-1)^{|\omega|} - 1} \cdot \left[\frac{1}{(|S|-1)^{|\omega|}} \right]^{\left\lfloor \frac{m-|c'|}{|\omega|} \right\rfloor} \right) = \epsilon_3.$$

Then

$$\begin{aligned}
\frac{a-\mu_m(\omega)}{\sigma_m(\omega)} - \frac{a-\mu(\omega)}{\sigma(\omega)} &> \frac{a-\mu(\omega)-\epsilon_3}{\sigma(\omega)+\epsilon_2} - \frac{a-\mu(\omega)}{\sigma(\omega)} \\
&= \frac{-\epsilon_3 \cdot \sigma(\omega) - \epsilon_2 \cdot a + \epsilon_2 \cdot \mu(\omega)}{\sigma(\omega) \cdot (\sigma(\omega)+\epsilon_2)},
\end{aligned}$$

and

$$\begin{aligned}
\frac{a-\mu_m(\omega)}{\sigma_m(\omega)} - \frac{a-\mu(\omega)}{\sigma(\omega)} &< \frac{a-\mu(\omega)+\epsilon_3}{\sigma(\omega)-\epsilon_2} - \frac{a-\mu(\omega)}{\sigma(\omega)} \\
&= \frac{\epsilon_3 \cdot \sigma(\omega) + \epsilon_2 \cdot a - \epsilon_2 \cdot \mu(\omega)}{\sigma(\omega) \cdot (\sigma(\omega)-\epsilon_2)}.
\end{aligned}$$

But $\left|\frac{-\epsilon_3 \cdot \sigma(\omega) - \epsilon_2 \cdot a + \epsilon_2 \cdot \mu(\omega)}{\sigma(\omega) \cdot (\sigma(\omega)+\epsilon_2)}\right| < \left|\frac{\epsilon_3 \cdot \sigma(\omega) + \epsilon_2 \cdot a - \epsilon_2 \cdot \mu(\omega)}{\sigma(\omega) \cdot (\sigma(\omega)-\epsilon_2)}\right| < \frac{\epsilon}{6}$ by Equation 6.2, so

$$\left|\frac{a-\mu_m(\omega)}{\sigma_m(\omega)} - \frac{a-\mu(\omega)}{\sigma(\omega)}\right| < \frac{\epsilon}{6}.$$

Similarly,

$$\begin{aligned}\frac{b-\mu_m(\omega)}{\sigma_m(\omega)} - \frac{b-\mu(\omega)}{\sigma(\omega)} &> \frac{b-\mu(\omega)-\epsilon_3}{\sigma(\omega)+\epsilon_2} - \frac{b-\mu(\omega)}{\sigma(\omega)} \\ &= \frac{-\epsilon_3\cdot\sigma(\omega)-\epsilon_2\cdot b+\epsilon_2\cdot\mu(\omega)}{\sigma(\omega)\cdot(\sigma(\omega)+\epsilon_2)}\end{aligned}$$

and

$$\begin{aligned}\frac{b-\mu_m(\omega)}{\sigma_m(\omega)} - \frac{b-\mu(\omega)}{\sigma(\omega)} &< \frac{b-\mu(\omega)+\epsilon_3}{\sigma(\omega)-\epsilon_2} - \frac{b-\mu(\omega)}{\sigma(\omega)} \\ &= \frac{\epsilon_3\cdot\sigma(\omega)+\epsilon_2\cdot b-\epsilon_2\cdot\mu(\omega)}{\sigma(\omega)\cdot(\sigma(\omega)-\epsilon_2)},\end{aligned}$$

so since $\left|\frac{-\epsilon_3\cdot\sigma(\omega)-\epsilon_2\cdot b+\epsilon_2\cdot\mu(\omega)}{\sigma(\omega)\cdot(\sigma(\omega)+\epsilon_2)}\right| < \left|\frac{\epsilon_3\cdot\sigma(\omega)+\epsilon_2\cdot b-\epsilon_2\cdot\mu(\omega)}{\sigma(\omega)\cdot(\sigma(\omega)-\epsilon_2)}\right| < \frac{\epsilon}{6}$ by Lemma 6.3, we have that

$$\left|\frac{b-\mu_m(\omega)}{\sigma_m(\omega)} - \frac{b-\mu(\omega)}{\sigma(\omega)}\right| < \frac{\epsilon}{6}$$

as well.

Therefore,

$$\begin{aligned}\left|\mathbb{P}_{(a,b)}(h_m) - \mathbb{P}_{(a,b)}(h)\right| &\leq \left|\frac{a-\mu_m(\omega)}{\sigma_m(\omega)} - \frac{a-\mu(\omega)}{\sigma(\omega)}\right| + \left|\frac{b-\mu_m(\omega)}{\sigma_m(\omega)} - \frac{b-\mu(\omega)}{\sigma(\omega)}\right| \\ &< \frac{\epsilon}{6} + \frac{\epsilon}{6} \\ &= \frac{\epsilon}{3}.\end{aligned}$$

Now we return to our original equation, and see that for $m > m_0$ and $n > n_0$,

$$\begin{aligned}\left|\mathbb{P}_{(a,b)}\left(\frac{i(\omega,c)}{n}\right) - \mathbb{P}_{(a,b)}(h)\right| &\leq \left|\mathbb{P}_{(a,b)}\left(\frac{i(\omega,c)}{n}\right) - \mathbb{P}_{(a,b)}\left(\frac{N(c,[2,m])}{n}\right)\right| \\ &\quad + \left|\mathbb{P}_{(a,b)}\left(\frac{N(c,[2,m])}{n}\right) - \mathbb{P}_{(a,b)}(h_m)\right| \\ &\quad + \left|\mathbb{P}_{(a,b)}(h_m) - \mathbb{P}_{(a,b)}(h)\right| \\ &< \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} \\ &= \epsilon.\end{aligned}$$

□

7 Conclusion

In this paper, we studied the distribution of $i(\omega, c)$ for fixed curve ω and $c \in C_n$ for some $n \in \mathbb{N}$. We have found an explicit calculation for the mean of the distribution of $i(\omega, c)$, and have written a program in Java that computes $\lim_{n\to\infty} \mathbb{E}\frac{i(\omega,c)}{n}$ for any fixed curve ω. We have also proved that the distribution of $i(\omega, c)$ approaches a Gaussian distribution as $n \to \infty$.

In the future, it would be useful to investigate the standard deviation of this distribution in order to fully quantify it. While there is a way to readily calculate the mean, it remains to be shown whether the standard deviation can be likewise calculated.

8 Acknowledgements

First of all, I'd like to acknowledge my mentor Dr. Moira Chas (SUNY), who not only proposed my project but also guided me throughout the entire process. I'm very thankful for her valuable help and dedication. I would also like to thank all of the staff of PRIMES-USA for giving me this research opportunity. In particular, I appreciate all the work of Dr. Tanya Khovanova (MIT), Dr. Slava Gerovitch (MIT), and Dr. Jonathon Harmon (MIT), who handled the logistics of the program.

References

[1] W. Goldman, “Invariant functions on lie groups and hamiltonian flows of surface group representations”, *Inventiones mathematicae*, vol. 85, no. 2, pp. 263–302, 1986.

[2] M. Mirzakhani, “Growth of the number of simple closed geodesics on hyperbolic surfaces”, *Ann. of Math. (2)*, vol. 168, no. 1, pp. 97–125, 2008, ISSN: 0003-486X.

[3] S. Katok and I. Ugarcovici, “Arithmetic coding of geodesics on the modular surface via continued fractions”, *CWI Tracts 135*, pp. 59–77, 2004.

[4] G. Margulis, “Applications of ergodic theory to the investigation of manifolds of negative curvature”, *Functional Analysis and Its Applications*, vol. 3, no. 4, pp. 335–336, 1970.

[5] J. S. Birman and C. Series, “An algorithm for simple curves on surfaces”, *J. London Math. Soc. (2)*, vol. 29, no. 2, pp. 331–342, 1984, ISSN: 0024-6107.

[6] M. Cohen and M. Lustig, “Paths of geodesics and geometric intersection numbers. I”, in *Combinatorial group theory and topology (Alta, Utah, 1984)*, ser. Ann. of Math. Stud. Vol. 111, Princeton, NJ: Princeton Univ. Press, 1987, pp. 479–500.

[7] M. Chas, “Combinatorial Lie bialgebras of curves on surfaces”, *Topology*, vol. 43, no. 3, pp. 543–568, 2004. [Online]. Available: `http://arxiv.org/abs/math/0105178v2`.

[8] ——, “Relations between word length, hyperbolic length and self-intersection number of curves on surfaces.”, *Recent Advances in Mathematics RMS-Lecture Notes*, vol. 21, pp. 45–75, 2015.

[9] ——, “Self-intersections are empirically gaussian”, *ArXiv preprint 1011.6085*, 2010. arXiv: `1011.6085v1 [math.GT]`.

[10] S. Lalley, “Statistical regularities of self-intersection counts for geodesics on negatively curved surfaces”, *ArXiv:1111.2060v2 [math.DS]*, 2011.

[11] M. Chas and S. P. Lalley, “Self-intersections in combinatorial topology: statistical structure”, *Inventiones Mathematicae*, vol. 188, no. 2, pp. 429–463, 2012.

[12] P. W. Glynn, *Markov chains*, 2008.

[13] A. Hatcher, *Algebraic Topology*. Cambridge University Press, 2002.