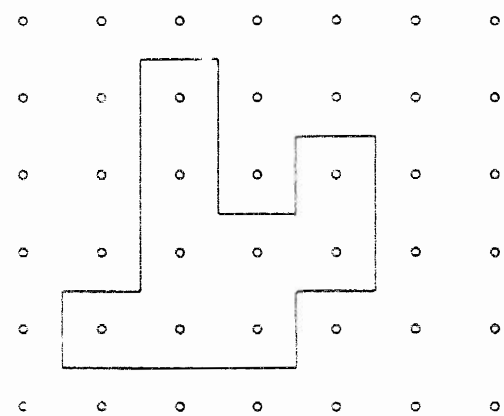

Figure 3.2.3: $Area(R) = |\text{gridpoints} \in R|$

between the expected number of points in the region and the actual number of points in the region. The expected number of points in any region the area of the region. Thus, we get Definition 3.2.3:

Definition 3.2.3: The *discrepancy* $\Delta(R)$ of a region R is the absolute value of the difference between the number of points of X in the region and the area of the region, i.e.

$$\Delta(R) = \Big||X \cap R| - Area(R)\Big| .$$

For rectilinear regions whose boundaries are on a grid between the grid points the area of the region is equal to the number of grid points in the region (See Figure 3.2.3). It follows that for these regions, the discrepancy is the difference between the number of points of X and the number of grid points in the region.

We can show that to prove Theorem 3.2.2 it is sufficient to prove the following: with high probability, given points distributed as above, every simple closed curve R whose boundary is on a grid with grid length $\Theta(\log^{3/4} n)$ has discrepancy at most $cPer(R)\log^{3/4} n$ for some fixed constant c. Here, $Per(R)$ is the length of the perimeter of R. We prove this theorem by approximating the simple closed curve R by a sequence of $O(\log n)$ regions, and showing that the difference of the discrepancies of successive approximations is small.

We divide the proof of the theorem into four sections. First, we prove several lemmas that we will need in the proof. Second, we prove that the dual theorem stated above implies the theorem. Last, we prove this dual theorem. We divide this proof into two parts: a deterministic part and a probabilistic part. The deterministic part defines a series of approximations of a region R. The probabilistic part uses these to show that the discrepancy $\Delta(R)$ is small.

3.2.3. Lemmas and Notation

In this section we will define notation that we will need for the proof of Theorem 3.2.2, and prove several lemmas that we will need in that theorem. Most of these definitions and lemmas are geometric. They deal with figures in the plane, mainly points, lines, grids, regions, paths and curves. The lemmas are mostly fairly easy to prove; some are trivial.

We define the ***symmetric difference*** of two regions in the plane A and B to be the set of points in exactly one of the regions. We use the notation $|A - B|$ for the symmetric difference of two regions. If $B \subseteq A$, then we will use $A - B$ to be the region containing every point in A that is not in B. Otherwise, $A - B$ will denote the ***signed*** difference of A and B; that is, every point in A and not in B is considered positive, and every point in B and not in A is considered negative. We will need the to define the discrepancy of a signed difference. We define the discrepancy of $A - B$ as

$$\Delta(A - B) = \Big| Area(A) - Area(B) - |X \cap A| + |X \cap B| \Big| .$$

We now have

$$|\Delta(A) - \Delta(B)| \leq \Delta(A - B).$$

We now prove certain lemmas that will be necessary to the proof of maximum edge length matching. Most of these are easy. They deal with geometric constructs and grids.

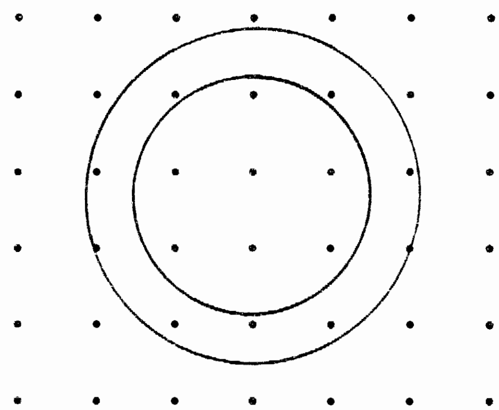

Figure 3.2.4: Expanding the circle to include squares around all the grid points.

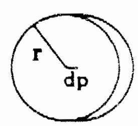

Figure 3.2.5: Moving a circle distance dp.

Lemma 3.2.4: Any circle of radius r on a unit grid can only contain $\pi(r + \sqrt{2}/2)^2$ grid points.

Proof: Expand the circle by $\sqrt{2}/2$ to obtain a new circle. Every grid point in the old circle is contained in a unit square entirely within the new circle. (See Figure 3.2.4) Thus, the number of grid points in the old circle is at most the area of the new circle, or $\pi(r + \sqrt{2}/2)^2$. ∎

Lemma 3.2.5: Let P be a path of length p. Then, if R is the region containing everything within distance r of a point on the path, the area of R is at most $2pr - \pi r^2$.

Proof: We can obtain R by moving a circle with radius r so that its center follows the path P. The area covered by the circle is the region R. If we move the circle a distance dp, then the additional area covered by the circle is $2rdp$. (See Figure 3.2.5) We start with area πr^2, so after moving the circle distance p, we have area at most $\pi r^2 + 2pr$. ∎

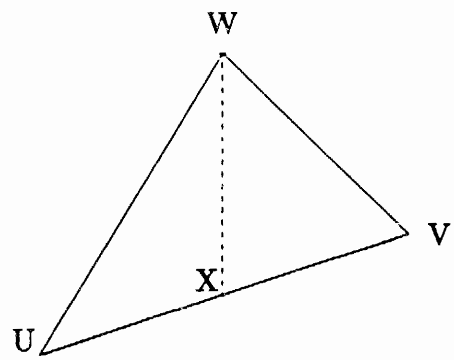

Figure 3.2.6: The triangle UVW.

Lemma 3.2.6: Consider the edge UV in Figure 3.2.6. The locus of points W determined by $k = 2(u^2+v^2)-w^2$ is a circle of radius $\frac{1}{2}\sqrt{k}$ around the midpoint X of UV. Here, $u = |WV|$, $v = |WU|$, and $w = |UV|$,

Proof: Let $U = (x_u, y_u)$, $V = (x_v, y_v)$, and $W = (x, y)$. Then we have the equation

$$k = 2\left((x-x_u)^2 + (y-y_u)^2 + (x-x_v)^2 + (y-y_v)^2\right) - (x_u - x_v)^2 - (y_u - y_v)^2.$$

This equation reduces to

$$k = 4(x^2+y^2) - 4x(x_u+x_v) - 4y(y_u+y_v) + (x_u+x_v)^2 + (y_u+y_v)^2,$$

or

$$\frac{k}{4} = \left(x - \frac{x_u+x_v}{2}\right)^2 + \left(y - \frac{y_u+y_v}{2}\right)^2.$$

This is the equation of a circle with center $(\frac{x_u+x_v}{2}, \frac{y_u+y_v}{2})$ and radius $\sqrt{k/4}$, as claimed. ∎

Lemma 3.2.7: In a triangle with sides of lengths a, b and c,

$$2(a^2+b^2) \geq c^2.$$

Proof: By the triangle inequality, $a+b \geq c$. Squaring both sides, we get $a^2+2ab+b^2 \geq c^2$, Now, adding the inequality $(a-b)^2 \geq 0$, we obtain $2a^2+2b^2 \geq c^2$. ∎

3.2.4. The Dual Problem

In this section, we show that the following theorem implies theorem 3.2.2.

Theorem 3.2.8: Suppose X is a set of n points uniformly distributed in the $\sqrt{n} \times \sqrt{n}$ square. Let G_m be a grid of squares with edge length $\Theta(\log^{3/4} n)$. Then there is a constant c such that with probability at least $1 - n^{-(\log n)^{1/2-\epsilon}}$ for any $\epsilon > 0$ there does not exist a simple closed curve R whose boundary follows G_m and which has discrepancy $\Delta(R) > cPer(R)\log^{3/4} n$.

To be specific, what we will do is show that if for an arrangement of the points such that the discrepancy of every simple closed curve R is $O(\log^{3/4} n)Per(R)$, the optimal matching has edges of length $O(\log^{3/4} n)$. We will need Hall's Theorem.

Hall's Theorem: In a bipartite graph G between two sets of points $\mathcal{P}^+$ and $\mathcal{P}^-$, the number of unmatched + points in a maximal matching is

$$\max_{A \subseteq \mathcal{P}^+} |A| - |R(A)|,$$

where $R(A)$ is the set of vertices of $\mathcal{P}^-$ that are adjacent to the vertices of A.

We use this to prove the following:

Lemma 3.2.9: Suppose that for a set X of n points in the $\sqrt{n} \times \sqrt{n}$ square all regions R made of squares from some grid of size c satisfy $\Delta(R) \leq cPer(R)$. Then there is a matching between points of X and grid points which has maximum edge length d, where $d = 16c$.

By Hall's Theorem, if for every set A of x points of X, there are x grid points within distance d of them, then there is a matching with all edge lengths less than d.

We first construct in the square a grid of smaller squares with sides of size $\frac{d}{4} = 4c$, so that there are $4\sqrt{n}/d$ grid squares on a side of the $\sqrt{n} \times \sqrt{n}$ square.

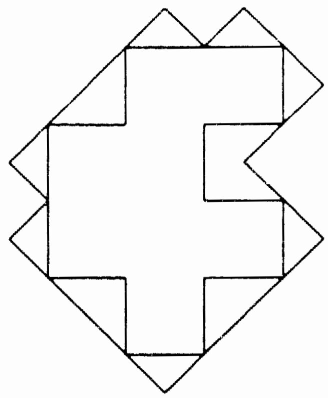

Figure 3.2.7: Forming R' by adding triangles.

Now, consider any subset $A \subseteq X$. Let the region R consist of all grid squares containing a point from A. Form a slightly larger region R' by adding an isosceles right triangle with hypotenuse $4c$ on each of the sides of the squares (See Figure 3.2.7). All the points in this larger region R' are within distance d of a point of A. If we can prove that the number of grid points in R' is larger than the number of points of X in the region R, then we are done.

We show that the number of grid points in R' is larger than the number of points of X in R. This follows from the inequalities

$$\begin{aligned} |\text{gridpoints in } R'| - |X \cap R| &\geq Area(R') - |X \cap R| \\ &\geq Area(R' - R) - \Delta(R) \\ &\geq \tfrac{d}{16} Per(R) - cPer(R) \\ &\geq 0. \end{aligned}$$

The number of grid points in R' will be larger than the area of R' since R' is a region which is a union of right isosceles triangles with grid points at their right angle, and this is true for any such region (See figure 3.2.8). The next inequality holds since $\Delta(R) = |Area(R) - |X \cap R||$. For every grid edge of length $d/4$ on the perimeter of our region, we have an isosceles right triangle in $R' - R$ with area $d^2/64$, so $Area(R' - R) = \frac{d}{16} Per(R)$. Finally, from the hypothesis of

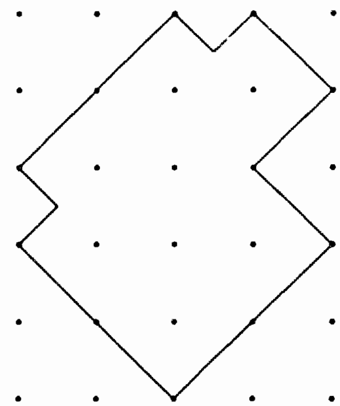

Figure 3.2.8: $Area(R') \leq |\text{gridpoints} \in R'|$.

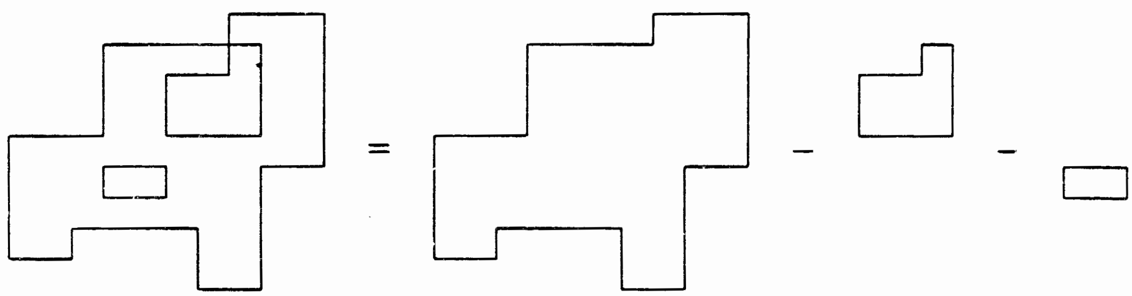

Figure 3.2.9: Decomposing an arbitrary region into simply connected regions.

our theorem, we have $\Delta(R) \leq cPer(R)$. This gives the last inequality.

Lemma 3.2.10: Suppose that every simply connected region R bounded by grid lines satisfies $\Delta(R) \leq cPer(R)$, where c is some constant. Then every region R bounded by grid lines satisfies $\Delta(R) \leq cPer(R)$.

Proof: We decompose an arbitrary region R composed of grid squares into a sum and difference of simply connected regions. (See Figure 3.2.9) We do this in the obvious way: the perimeter of the region decomposes into simple closed curves, each of which determines the simply connected region inside it. By adding and subtracting these regions, we obtain our original region. Since when we add all the regions, we add their perimeter, we have that the perimeter is the sum of the perimeters of the regions. The discrepancy of the total region is less than or equal to the sum of the discrepancies of the component regions. Since each of the component regions has discrepancy bounded by $c \cdot Per(R_i)$, the original region will also have discrepancy bounded by $c \cdot Per(R)$. ∎

It is easy to see that Lemma 3.2.10, Lemma 3.2.9 and Theorem 3.2.8 imply Theorem 3.2.2. By Theorem 3.2.2, with high probability every simply connected region R with boundary on a grid G_m with edge length $\Theta(\log^{3/4} n)$ has discrepancy at most $c \log^{3/4} n Per(R)$ for some constant c. By Lemma 3.2.10, this implies that for *all* regions R whose boundaries follow G_m, $\Delta(R) \leq c \log^{3/4} n Per(R)$. By Lemma 3.2.9, this implies that there is a matching with maximum edge length $\Theta(\log^{3/4} n)$. Thus, to prove Theorem 3.2.2, we must now prove Theorem 3.2.8

3.2.5. Deterministic Part of Proof

We now prove the theorem:

Theorem 3.2.8: Suppose X is a set of n points uniformly distributed in the $\sqrt{n} \times \sqrt{n}$ square. Let G_m be a grid of squares with edge length $\Theta(\log^{3/4} n)$. Then there is a constant c such that with probability at least $1 - n^{-(\log n)^{1/2-\epsilon}}$ for any $\epsilon > 0$ there does not exist a simple closed curve R whose boundary follows G_m and which has discrepancy $\Delta(R) > cPer(R)\log^{3/4} n$.

We will divide the proof of this theorem into two sections, a deterministic section and a probabilistic section. In the deterministic section, we show that we can produce a sequence of approximations R_i to the region R that satisfy certain conditions. In the probabilistic section, we use these approximations to bound the discrepancy of R_i by bounding the difference in the discrepancies of R_i and R_{i+1}. Showing that we can find these approximations R_i is the hardest part of the proof. The conditions on these approximations R_i are somewhat technical. They are stated in the following lemma.

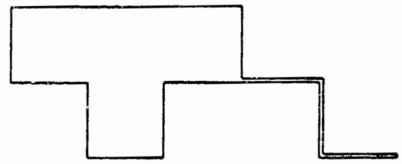

Figure 3.2.10: Making the perimeter a power of 2.

Lemma 3.2.11: Let G be a grid with edge length $\Theta(\log^{3/4} n)$. Then there is a constant C and a scheme for approximating all connected regions R with boundary on G satisfying the following: Any connected region R with boundary on G with perimeter p in the $\sqrt{n} \times \sqrt{n}$ square is approximated by regions $R_1, R_2, \cdots, R_m$, where $m \leq \log p$, such that there are numbers $s_1, s_2, \cdots, s_m$ with $\sum_{i=1}^{m} s_i \leq C$ satisfying

1. The area of the difference between successive approximations satisfies $Area(R_{i+1} - R_i) \leq 2^{-i} p^2$.
2. The number of possible sequences $R_1, R_2, \cdots, R_{i+1}$, given a bound s_i' on s_i, and *not* given R, is at most

$$2^{2^{i+1} \log(s_i' \log n)}.$$

Let R be a simple closed curve with boundary following G and let p be the perimeter of R. We can assume without loss of generality that p is a power of 2. Given a region R, we can add to its boundary with extensions of zero area to increase its perimeter to a power of 2. (See Figure 3.2.10.) This addition changes the boundary of the region, but leaves the region itself unchanged. The approximations we derive from the new boundary will thus approximate the same region R.

To produce the sequence of approximations R_i we will use two different sequences of polygons. The first sequence will have vertices lying on the boundary of R. We will call the ith polygon of this kind A_i. These polygons A_i will then

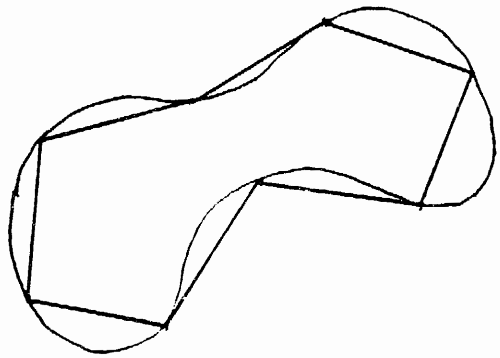

Figure 3.2.11: Obtaining A_i from the boundary of R.

be approximated further by polygons B_i with vertices on a grid G_i. If B_i is a simple closed curve, then R_i is its interior. Although we begin with a simple closed curve, neither the A_i nor the B_i approximations will necessarily be simple closed curves, which will cause further problems, forcing us to define a region "enclosed" by the polygon B_i. This "enclosed" region will be R_i. These regions R_i will not necessarily be connected, even though R is connected.

To obtain the approximation A_i, we mark 2^i points at equal distances along the perimeter of the curve. (See Figure 3.2.11) We call these points a_{i0}, a_{i1}, ..., $a_{i,2^i-1}$. The starting point $a_{i0} = a_0$ will be the same for all A_i. We then join these points in order by edges. Half the vertices of A_{i+1} are also vertices of A_i, specifically, $a_{ij} = a_{i+1,2j}$. We let the length of the edge between $a_{i,j-1}$ and a_{ij} be e_{ij}.

The polygon B_i is obtained by approximating A_i using points on a grid G_i. We let the jth vertex b_{ij} of B_i be the nearest grid point to the jth vertex a_{ij} of A_i (See Figure 3.2.12). If several grid points are equidistant to some vertex, we need to break the tie. Any consistent rule for breaking ties can be used. Alternatively, we can perturb the points so that there are no ties. We will assume that there are no ties.

The grid G_i will have points spaced evenly at distance $\Theta(p/(2^i\sqrt{\log n}))$. The

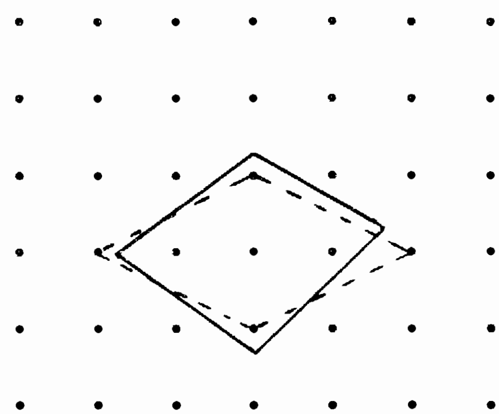

Figure 3.2.12: Obtaining B_i from A_i.

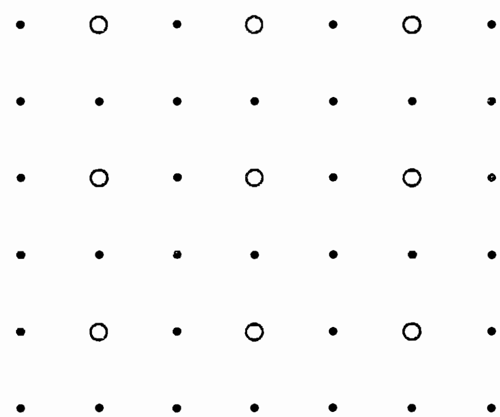

Figure 3.2.13: The grids G_{i+1} and G_i.

edge length of G_{i+1} will be half that of G_i. The grid G_{i+1} is a refinement of G_i, so a fourth of the points of G_{i+1} are also points of G_i. (See Figure 3.2.13.) We denote the edge length of G_i by $g_i = g_1/2^{i-1}$. We will want g_i to be a power of 2, so we choose g_i to be the smallest power of 2 larger than $p/(2^i\sqrt{\log n})$.

Thus, to obtain A_i, we do the following:

1. From some fixed point a_0 along the perimeter, mark every point at distance $p/2^i$ along the curve.
2. Connect these points with straight segments to produce a polygon.

To produce B_i, we add step $1\frac{1}{2}$ between steps 1 and 2:

1. From some fixed point a_0 along the perimeter, mark every point at dis-

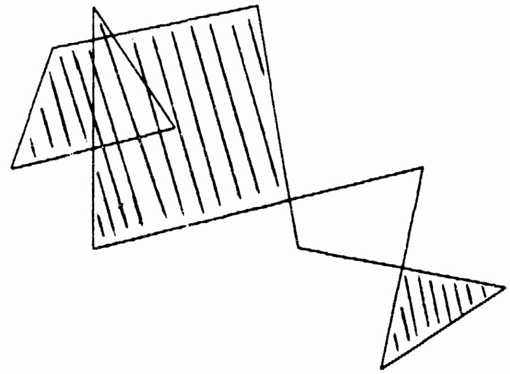

Figure 3.2.14: Obtaining R_i from B_i: R_i is shaded.

tance $p/2^i$ along the curve.

$1\frac{1}{2}$. Approximate these 2^i points by points of the grid G_i.

2. Connect these points with straight segments to produce a polygon.

We have now produced a polygon B_i. If this polygon is a simple closed curve, then the area inside it will be our ith approximation R_i to the region we wish to approximate. If it is not, we must do some more work. We wish to avoid double counting areas. If an area is enclosed by the polygon twice or more (i.e., has winding number ≥ 2), we still wish to count each point inside it at most once when calculating the discrepancy. We also do not want to count any point with winding number 0. We can do this in the following manner: If the winding number of a point is positive with respect to B_i, we include it in our region. If the winding number is zero or negative, we do not include it. This gives the region R_i which we will use as an approximation to the region R. (See Figure 3.2.14.)

We must show that the area of the differences of two successive approximations, $Area(|R_{i+1} - R_i|)$, is small. If we look at the approximations A_i instead of R_i, we see that the difference between A_{i+1} and A_i is a region formed by 2^i triangles around the border of A_i (See Figure 3.2.15.) We will let T_{ij} be the

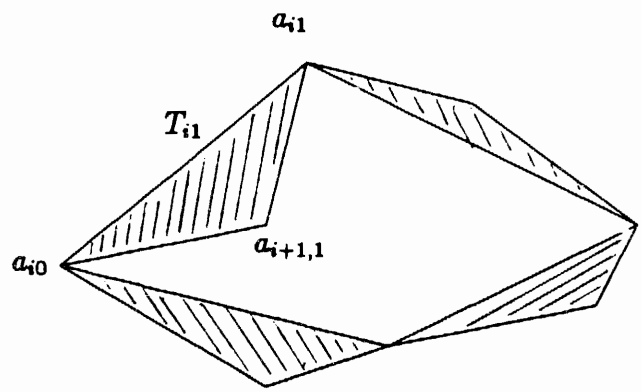

Figure 3.2.15: The triangles T_i.

triangle formed by the points $a_{i,j-1}$, a_{ij} and $a_{i+1,2j-1}$. This triangle has edges of lengths e_{ij}, $e_{i+1,2j-1}$ and $e_{i+1,2j}$. We will be able to show that on the average two angles of these triangles are small (the angles at the vertices in A_i). This shows the average area of a triangle is small, so the area of the region between A_i and A_{i+1} is small. The region $|R_{i+1} - R_i|$ is an approximation of this, so it also has a small area.

The intuitive reason that the average side angle of a triangle T_{ij} is small is that a large angle adds a lot to the perimeter. For example, if all the angles of triangles on the ith level are 45°, the perimeter of the $(i+1)$st level will be $\sqrt{2}$ times the perimeter of the ith level. If the perimeter goes up by a large factor at each step, and ends at the value p, it must start out very small. Otherwise, there must be a lot of steps where the perimeter does not increase much. In either case, the area of $|R_{i+1} - R_i|$ is small on the average.

Merely showing that the area of $|R_{i+1} - R_i|$ is small does not show that the discrepancy is small. We must also show that there are a relatively small number of choices at each stage for R_{i+1}. Intuitively, there are only a small number of choices because the points in B_{i+1} which are added between two vertices of B_i usually fall near the midpoint of the edge between the two adjacent vertices. There are only a small number of grid points near this midpoint, so the number of choices for B_{i+1} is limited. We show this by showing that the triangles T_{ij} not only have two small angles, but also two nearly equal sides.

We must quantify all the intuitive notions presented above. It turns out that the best quantity to look at is not the perimeter (the sum of the lengths of the edges) but the sum of the squares of the edge lengths. Recall e_{ij}, $1 \leq j \leq 2^i$ were the lengths of the edges of A_i. We will look at $\sum_{j=1}^{2^i} e_{ij}^2$. To obtain a quantity that always increases, we must normalize this sum by 2^i. We will also normalize it by p^2 so as to get a dimensionless number. What we actually use is thus a normalized generalized perimeter $q_i = \frac{2^i}{p^2} \sum_{j=1}^{2^i} e_{ij}^2$.

In the rest of this proof, we will be looking closely at what happens in triangle T_{ij}. It will thus help to have generic names for the points in this situation. When we are talking about a generic edge of A_i, it will have endpoints U and V, and the two edges it is replaced by will be UW and WV. We define $k = 2(|UW|^2 + |VW|^2) - |UV|^2$, as we will be using this quantity often. We define k_{ij} to be this quantity for triangle T_{ij}, so $k_{ij} = 2(e_{i+1,2j-1}^2 + e_{i+1,2j}^2) - e_{ij}^2$.

We now prove some necessary facts about the sum of the squares of the edges, so we can later bound the number of choices for B_i and the area of $R_{i+1} - R_i$.

Let $q_i = \frac{2^i}{p^2} \sum_{j=1}^{2^i} e_{ij}^2$. Here e_{ij} is the length of the jth edge of the ith approximation A_i of R, and p is the perimeter of R. We will show the following claim.

Claim 3.2.12:

$$1 \geq q_{i+1} \geq q_i \text{ for all } i.$$

Proof of Claim:

A) $q_i \leq 1$.

At the ith step, all the edges have length $\leq \frac{p}{2^i}$, and there are 2^i of them. Thus, $\frac{2^i}{p^2} \sum e_i^2 \leq \frac{2^i}{p^2} \sum \left(\frac{p}{2^i}\right)^2 = \frac{2^i}{p^2} 2^i \left(\frac{p}{2^i}\right)^2 = 1$.