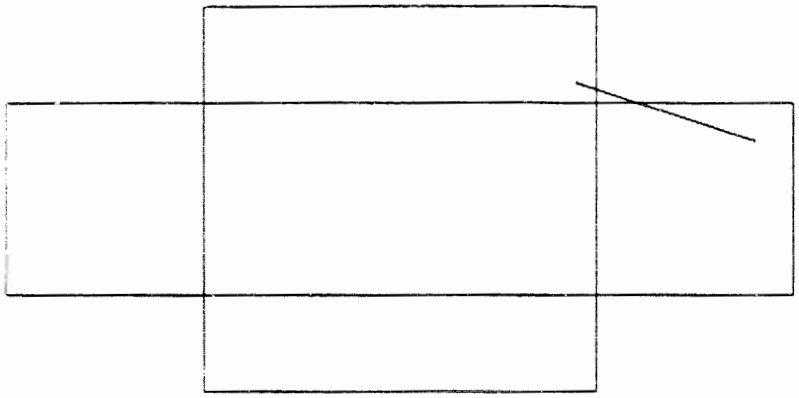

Figure 2.2.2: If the square and rectangle are aligned, no point in the square moves far.

We now show that if we use the above procedure until each region has on the average $\Theta(\log n)$ points, then with high probability each region will contain $\Theta(\log n)$ points. Since after $2i$ steps the regions are transformed squares of a $2^i \times 2^i$ grid, all we need do is show that if a square containing n points is divided into a grid with edge length $\sqrt{\alpha \log n}$, then with high probability each grid square will contain between $\alpha_1 \log n$ and $\alpha_2 \log n$ points, where α_1 and α_2 are some fixed constants. This is shown in the lemma below.

Lemma 2.2.1: If n points are distributed uniformly in the unit square, a region of area $\alpha \log n/n$ will, with probability $1/n^a$ for any a, have between $\alpha_1 \log n$ and $\alpha_2 \log n$ points, where α_1 and α_2 depend on a. Furthermore, by choosing α large enough, we can make α_1/α_2 arbitrarily close to 1.

Proof: This follows from Lemma 3.2.19, proved later in this thesis. In this proof, we will take $\log n$ to be the natural logarithm. Changing the base of the logarithm only affects the constants. Take $\sigma = \sqrt{3a \log n}$. Then with probability $e^{-\sigma^2/3} = n^{-a}$, the region contains $A \pm \sigma\sqrt{A}$, or $(\alpha \pm \sqrt{3a\alpha}) \log n$ points. This holds as long as $2\sigma^2 \leq \alpha$, or $6a \leq \alpha$. By taking α large enough, we ensure that $6a \leq \alpha$ and that $\sqrt{3a\alpha} \ll \alpha$, proving the result. ∎

We must now show that after $\log(n/\log n)$ steps, with high probability the rectangles have aspect ratios which are bounded by some constant. We do this by bounding the change of the aspect ratio of the rectangle at each step, and then multiplying these changes together. These factors form a convergent product dominated by the last term. By the lemma above, this last term is with high probability bounded by a constant.

Lemma 2.2.2: Suppose that rectangles are constructed as described above, by dividing the rectangles at the previous stage in half, and moving the boundaries so the area of a rectangle is proportional to the number of points in it. Suppose further that at the ith stage each rectangle has at least $c_1 n/2^i$ points for some constant c_1. Then with probability at least $1 - 1/n^a$, at the $\log(n/\log n)$ stage, the aspect ratio is less than some constant r_0.

Proof: If a rectangle has k points in it, then at the next stage it will be divided into two rectangles having x and $k - x$ points, where x is a random variable having a binomial distribution. The amount the aspect ratio changes is the amount that the rectangle is stretched. The side of a rectangle will be multiplied by $2x/k$. Now, with probability $1 - 1/n^a$, $|x - \frac{k}{2}| = O(\sqrt{k}\sqrt{\log n})$, so the aspect ratio is at worst multiplied by $1 - c_1\sqrt{\log n}/\sqrt{k}$. At the ith stage, $k \geq c_1 n/2^i$, and at the last stage, $k \geq c_1 \alpha \log n$. Thus, at the ith stage from the end, with at least $2^i c_1 \alpha \log n$ points in a rectangle, the aspect ratio is at worst multiplied by $(1 - c/\sqrt{2^i \alpha c_1})$. Let c_2 be $c/\sqrt{\alpha c_1}$. Then, the final aspect ratio is with high probability at most

$$(1 - c_2)^{-1}(1 - c_2/\sqrt{2})^{-1}(1 - c_2/2)^{-1}(1 - c_2/2\sqrt{2})^{-1} \cdots.$$

If $c_2 < 1$, which we can ensure by making α large enough, this infinite product converges, so there is a constant β such that the aspect ratio is bounded by β with probability $1 - 1/n^a$. ∎

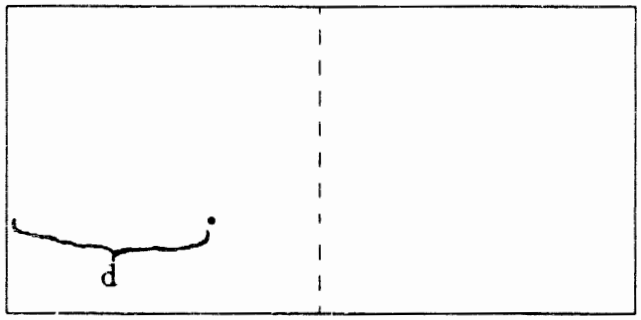

Figure 2.2.3:

We now prove another lemma: this one enables us to complete the matching once we have divided the square into rectangles.

Lemma 2.2.3: Suppose we have two partitions of a unit square into n regions of equal area, $Q_1, Q_2, ..., Q_n$ and $R_1, R_2, ..., R_n$. The R_i's are pairwise disjoint, as are the Q_i's, and $\bigcup Q_i = \bigcup R_i = \mathbf{S}$. Then there is a matching σ between the R_i's and the Q_i's such that $Q_i \cap R_{\sigma(i)} \neq \emptyset$.

Proof: This is a consequence of Hall's theorem. We need to show that any j of the Q_i's can always be matched to j of the R_i's. This is true unless we can find $j-1$ R_i's and j Q_i's such that

$$\bigcup_{i=1}^{j} Q_{\sigma(i)} \subseteq \bigcup_{i=1}^{j-1} R_{\pi(i)}.$$

However, since all the regions have area $\frac{1}{n}$, this would mean that a region of area j/n is contained in a region of area $j - 1/n$, a contradiction. ∎

We must now show that the average distance moved by the center of each square is small. At each step, we move the center of a square with equal probability in one of two directions. The center of the square will be contained in some rectangle (See Figure 2.2.3). Suppose that the center is at distance d from the left edge of the rectangle and that the width of the rectangle is s. Suppose further that the point is in the left half of the rectangle. Let the number of

points in the rectangle be k and the number of points in the left half of the rectangle be x. Then the size of the left half of the rectangle will be changed from $s/2$ to sx/k. This means the point p will be moved from distance d from the left side of the rectangle to distance $2dx/k$ from the left edge of the rectangle. The point thus moves distance $d(k-2x)/k$. Since x is taken from a binomial distribution on k, this distance is distributed symmetrically about 0. Thus the distance moved by any point is a martingale.

In each step the distance moved has mean 0 and variance at most $\frac{1}{n}$. After $\Theta(\log n)$ steps, the mean is 0 and the variance is $O(\log n)/n$, so the average distance moved is $O(\sqrt{\log n}/\sqrt{n})$. This proves the upper bound for average edge length matching. ∎

Chapter 3. Up-right and Maximum Distance Matching

3.1. The Lower Bound

In this section, we obtain the $\Omega(n^{1/2}\log^{3/4} n)$ lower bound for up-right matching.

Theorem 3.1.1: Suppose there are n points uniformly distributed in a unit square, and each point has an equal probability of being a + or a − point. If these points are matched such that every − point is matched to a + point above and to the right of it, then the expected number of unmatched points in a maximum such matching is $\Omega(n^{1/2}\log^{3/4} n)$.

Proof: In order to make the proof easier, we rotate the square 45°. With this rotation, a − point can only be matched to a + point above it, and the slope of the edge joining them must be larger than 1 or smaller than −1. (See Figure 3.1.1.) To obtain a lower bound of k for the number of unmatched points in an up-right matching in a rotated square, it suffices to split the square into two sections, such that in the lower section there are k more + points than − points, and such that the boundary dividing the sections is a curve that joins the left corner of the square to the right corner, and always has slope between −1 and 1. (See Figure 3.1.2.) With such a boundary, the excess + points in the lower section can never be matched to − points in the upper section, so they

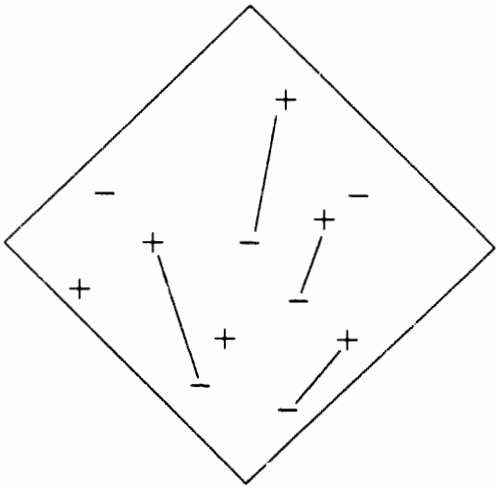

Figure 3.1.1: Rotating the square.

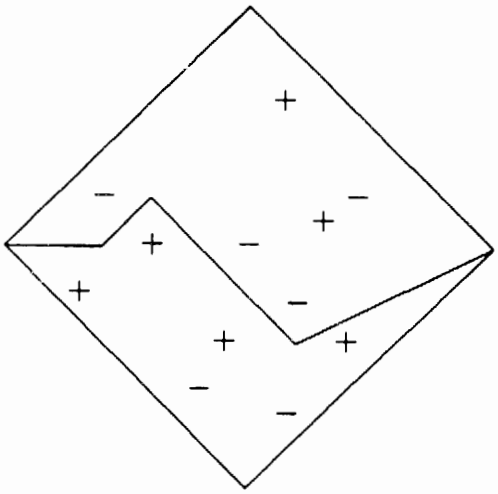

Figure 3.1.2: A lower set.

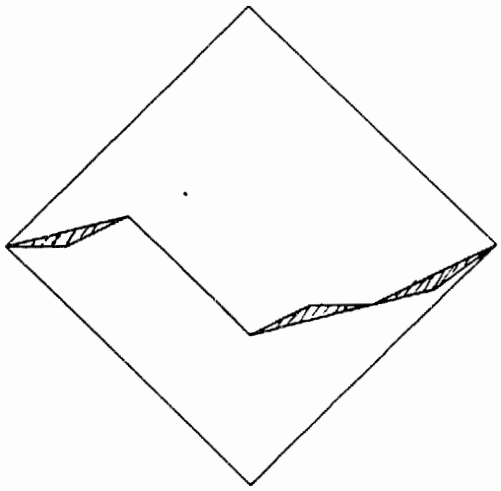

Figure 3.1.3: The boundary at stage 2.

must remain unmatched. We will construct a boundary such that the expected number of extra +'s below it is $\Theta(n^{1/2}\log^{3/4} n)$.

We will produce the boundary in stages. At each stage, the boundary will consist of line segments and triangles. If a triangle is on the boundary, then in any subsequent stage the boundary will pass through two vertices of the triangle, and the portion of the boundary between these vertices will be contained within the triangle. For example, in Figure 3.1.3, the final boundary will lie in the shaded areas. To obtain the boundary at the next stage, we replace every triangle with either a line segment or with two triangles each having a quarter of the area of the old triangles. We finally stop the refinement when the triangles are so small that they contain on the average only one point in each.

The general step of replacing a triangle by two smaller triangles is illustrated in Figure 3.1.4. The points G and H are midpoints of AB and BC, and D, F and E divide AC into quarters. In the next stage, this triangle will be replaced either by triangles ADB and BEC or by triangles AGF and FHC. The *central quadrilateral* $BKFJ$ is defined by the four edges BE, BD, FG and FH; it is shaded in Figure 3.1.4. We put the central quadrilateral below the boundary

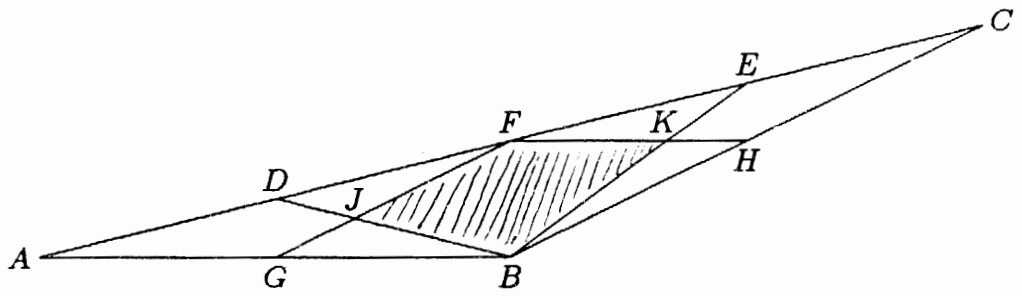

Figure 3.1.4: A typical triangle on the boundary.

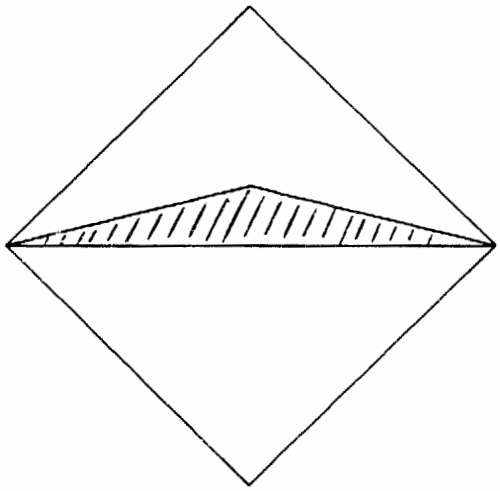

Figure 3.1.5: The boundary at the start.

when it contains more +'s than −'s. In the triangles, the central vertex (B in Figure 3.1.4) can point either up or down. If the triangle's vertex points down, as in the picture, the refinement to AGF and FHC places the central quadrilateral below the boundary. Otherwise, the refinement to AGF and FHC will do this. We begin with one large triangle, as shown in Figure 3.1.5. This is an isosceles triangle which has two vertices at the left and right corners of the square, and which has two sides with slopes of $\pm s = \log^{1/2} n$.

We now list several properties of the triangles we will be using. The letters refer to the example in Figure 3.1.4.

1. The three vertices (A, B, C) have x-coordinates evenly spaced, so the x-coordinate of B is the average of the other two. Thus, the segment BF

is vertical.

2. Each triangle has exactly 1/4 the area of the previous one.

3. The area of the central quadrilateral is 1/3 the area of the triangle.

4. If the slopes of the sides of a triangle are $(k-1)s$, ks and $(k+1)s$, then the two triangles it is refined to will either both be similar to it, in which case they have sides with the same slopes, or one triangle will have sides with slopes $(k-2)s$, $(k-1)s$ and ks, and the other will have sides with slopes ks, $(k+1)s$ and $(k+2)s$.

These properties are easily proved using elementary Euclidean geometry.

Property 1 follows by induction. We must show that the x-coordinates of D and G are halfway between the x-coordinate of A and the x-coordinate of B. This follows from the fact that D and G are the midpoints of AF and AB, respectively.

To prove property 2, we must show that

$$Area(AGF) = Area(ABD) = \frac{1}{4}Area(ABC).$$

Since F is the midpoint of AC, $Area(AFB) = \frac{1}{2}Area(ABC)$. Since D is the midpoint of AF, $Area(ADB) = \frac{1}{2}Area(ABF)$. Similarly, G is the midpoint of AB, so $Area(AGF) = \frac{1}{2}Area(ABF)$. This proves property 2.

Property 3 follows from the theorem of elementary geometry that the intersection of the medians of a triangle divides the medians at the $\frac{1}{3}, \frac{2}{3}$ points. This implies that $Area(BFJ) = \frac{1}{3}Area(ABF)$ and $Area(BFK) = \frac{1}{3}Area(BFC)$. Adding areas, we obtain $Area(BKFG) = \frac{1}{3}Area(ABC)$.

To prove property 4, we must show that

$$slope(AD) + slope(DB) = ? \cdot slope(AB).$$

This follows from the facts that BF is vertical and that D is the midpoint of AF.

If we let the first triangle be stage 0, at stage i we are testing 2^i regions each of area $\frac{1}{6}s/4^i$ to decide whether or not to include them. (Recall $0, \pm s$ were the slopes of the sides of the original triangle.) The expected difference between the number of $+$ and $-$ points in a regions of area A is $\Theta(\sqrt{nA})$. This is true since a region of area A contains on the average nA points, and each of these has an equal probability of being a $+$ or a $-$ point. Thus, each stage adds on the average $\Theta(\sqrt{ns})$ extra $+$ points to the lower region. After $\Theta(\log n)$ stages, we have an expected number of $\Theta(\sqrt{ns}\log n) = \Theta(\sqrt{n}\log^{3/4} n)$ extra $+$ points in the lower region.

We still must show that the slope of the boundary stays between -1 and 1. If we keep refining all the triangles, it will not. We must modify the procedure so that any time we would produce an edge of a triangle with slope larger than 1 (or smaller than -1), we stop changing the boundary along this segment. We must then show this will not affect the analysis.

We can consider the triangles to be organized in a binary tree, so the children of any triangle are its two refinements on the next level. By property 4, the slopes of the sides of a triangle differ from the slopes of the sides of its parent by $-s$, 0, or $+s$. If one of the children of a triangle is obtained from its parent by adding $-s$, then by property 4, the other is obtained by adding $+s$. In such a tree, by Lemma 2.1.1, if we stop after $O(\log n)$ levels, at most 1/4 of the nodes exceed a value of $s\sqrt{\log n}$. Since $s = \log^{1/2} n$, we get that the slope exceeds 1 on at most $\frac{1}{4}$ of the triangles. By Lemma 2.1.2, even if we ignore the worst $\frac{1}{4}$ of the triangles, each level will still give $\Theta(\sqrt{ns})$ points. ∎

3.2. The Upper Bound

3.2.1. Intuition For the Upper Bound Proof

To give the intuition behind the proof, we will give the proof of a simpler result with the same $\log^{3/4} n$ bound. This result is an interesting one in its own right, dealing with the decomposition a polygonal regions into triangles. The result does not imply the maximum edge length matching result, but it does give some of the basic ideas behind it. In the full result these ideas are obscured by the technical details needed to prove the result.

The result that we will prove in this section is that for any polygonal region R with n sides and perimeter p, R can be decomposed into the sum and difference of triangles T_i such that the sum of the square roots of the areas of the triangles satisfies $\sum_i \sqrt{\text{Area}(T_i)} = \Theta(p \log^{3/4} n)$. The expected discrepancy of a triangle T is $\Theta(\sqrt{Area(T)})$. If we could assume that the average discrepancy of a triangle in the decomposition of the polygonal region R was this expected discrepancy, then by this result the total discrepancy of the region R would be $O(p \log^{3/4} n)$. Proving this for a general class of regions R is the most difficult part of the proof of maximum edge length matching presented in the next sections, and is done in Theorem 3.2.8.

Theorem 3.2.1: Any polygonal regions R with n vertices and perimeter p can be decomposed into a sum and difference of triangles T_i such that $\sum \sqrt{Area(T_i)} = O(p \log^{3/4} n)$.

Proof: The algorithm to decompose the region into a sum and difference of triangles is simple. At each step, we will reduce the number of sides of the polygon by one by cutting off a triangle (See Figure 3.2.1) We cut off the triangle formed by a pair of adjacent sides. We pick the adjacent sides $v_{i-1}v_i$ and v_iv_{i+1} minimizing $c = e_i + e_{i+1}$, where e_i is the length of side $v_{i-1}v_i$. We then replace

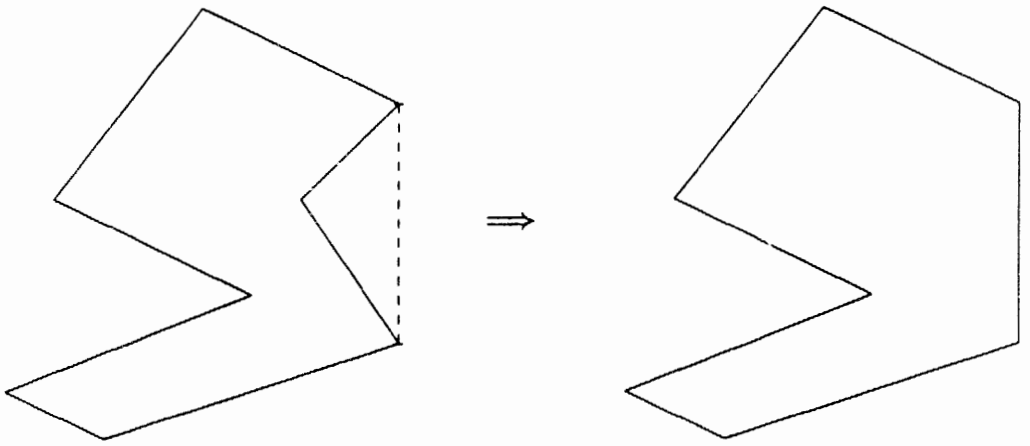

Figure 3.2.1: Removing a triangle.

these two edges by the edge between v_{i-1} and v_{i+1}.

We must show that this algorithm produces a decomposition into triangles such that $\sum_{T_i} \sqrt{Area(T_i)} \leq O(p \log^{3/4} n)$. Let the edge between v_{i-1} and v_{i+1} have length d. We have thus reduced the perimeter by $\Delta = c - d$. The area of the triangle that we cut off is at most $\frac{1}{4} d\sqrt{c^2 - d^2}$ since this area is maximized when $e_i = e_{i+1}$. Now, $c \leq 2p/n$, since we chose the smallest pair of adjacent sides.

We repeat this step until we reduce the number of sides to 2, and thus have no area left. We let c_i be the sum of the two smallest sides at the ith step, d_i be the length of the new edge introduced at the ith step, and $\Delta_i = c_i - d_i$ be the change in perimeter at the ith step. Let T_i be the triangle we cut off at the ith step. Then

$$\begin{aligned}
\sqrt{Area(T_i)} &\leq \tfrac{1}{2} d_i^{1/2} (c_i^2 - d_i^2)^{1/4} \\
&= \tfrac{1}{2} d_i^{1/2} (c_i + d_i)^{1/4} (c_i - d_i)^{1/4} \\
&\leq c_i^{3/4} \Delta_i^{1/4}.
\end{aligned}$$

Since Δ_i was the change in the perimeter at the ith step,

$$\sum_{i=1}^{n-2} \Delta_i \leq p.$$

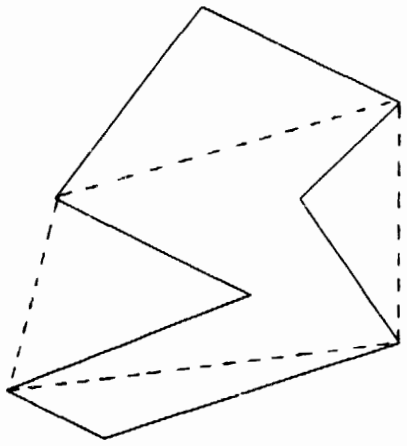

Figure 3.2.2: Removing four triangles.

Since $c_i \leq \frac{2}{n+1-i}p$,

$$\sum_{i=1}^{n-2} c_i \leq 2p \log n.$$

Thus, we have

$$\begin{aligned} \textstyle\sum_{i=1}^{n-2} \sqrt{Area(T_i)} &\leq \textstyle\sum_{i=1}^{n-2} c_i^{3/4} \Delta_i^{1/4} \\ &\leq \left(\textstyle\sum_{i=1}^{n-2} c_i\right)^{3/4} \left(\textstyle\sum_{i=1}^{n-2} \Delta_i\right)^{1/4} \\ &\leq 2p \log^{3/4} n. \end{aligned}$$

The second step is a special case of Hölder's inequality. ∎

If we calculate the sum of the square roots of the areas of the triangles using just the bounds on the lengths of the sides of the triangles, we get $\sqrt{Area(T_i)} \leq p/n$, since the sides of triangle T_i were at most p/n. This gives $\sum \sqrt{Area(T_i)} \leq p \log n$. The extra factor of $\log^{1/4} n$ comes either because the triangles have smaller angles or shorter sides than one would naively expect. This fact is also what gives us the extra $\log^{1/4} n$ factor in Theorem 3.2.2.

The proof of Theorem 3.2.1 will also work if we take the decomposition where we cut off the odd vertices of the polygon at each step. (See Figure 3.2.2) Thus, each step halves the number of vertices and produces half this number

of triangles. We have that $\sqrt{Area(T_i)} \leq c_i^{3/4}\Delta_i^{1/4}$ and $\sum_{i=1}^{n-2} \Delta_i \leq p$ for the same reasons as before. The only thing we need to show is $\sum c_i \leq p \log n$. This is true since there are $\log n$ stages, and at every stage, $\sum c_i \leq p$. This decomposition is the one that we will be using in the proof of maximum edge length matching.

3.2.2. Outline of Proof

The problem of maximum edge length matching is as follows: given a set X of n points uniformly distributed in the $\sqrt{n} \times \sqrt{n}$ square, and the regular $\sqrt{n} \times \sqrt{n}$ grid in this square, what is the expected maximum edge length of the optimal matching between the grid points and the points of X (i.e., the matching that minimizes this length). Notice that in this section, we have rescaled the unit square of the previous section to a $\sqrt{n} \times \sqrt{n}$ square. This will simplify several of the expressions involved in the proof.

In this section we will show an upper bound of $O(\log^{3/4} n)$ for the maximum distance in an optimal matching that holds with probability at least $1 - n^{-(\log n)^{1/2-\epsilon}}$ for any $\epsilon > 0$.

Theorem 3.2.2: If a set X of n points are uniformly and independently distributed in a $\sqrt{n} \times \sqrt{n}$ square, then with probability $1 - n^{-(\log n)^{1/2-\epsilon}}$ for any $\epsilon \geq 0$, there is a matching between the points of X and the regular $\sqrt{n} \times \sqrt{n}$ grid with unit edge length in the square such that no point is matched farther than $O(\log^{3/4} n)$.

To prove this theorem, we go to the dual problem. We prove that with high probability, in any region R with boundary on a grid G_m with edge length $\Theta(\log^{3/4} n)$, the number of points of X in R is $Area(R) \pm O(\log^{3/4} n \ Per(R))$. This is not true for regions where the perimeter is significantly smaller than $\log^{3/4} n$. This result will prove Theorem 3.2.2 above.

We will need to define the *discrepancy* of a region. This is the difference