ELSEVIER

Computational Geometry 19 (2001) 69–85

Computational Geometry
Theory and Applications

www.elsevier.nl/locate/comgeo

Matching colored points in the plane: Some new results

Adrian Dumitrescu [a], Rick Kaye [b]

[a] *University of Wisconsin, Department of Electrical Engineering Computer Science, EMS Building, PO Box 784, Milwaukee, WI 53201-0784, USA*

[b] *44 Pine Ridge Drive, Edison, NJ 08820, USA*

Communicated by Janos Pach; submitted 17 July 2000; accepted 30 January 2001

Abstract

Let S be a set with $n = w + b$ points in general position in the plane, w of them white, and b of them black. We consider the problem of computing $G(S)$, a largest non-crossing matching of pairs of points of the same color, using straight line segments. We present two new algorithms which compute a large matching, with an improved guarantee in the number of matched points. The first one runs in $O(n^2)$ time and finds a matching of at least 85.71% of the points. The second algorithm runs in $O(n \log n)$ time and achieves a performance guarantee as close as we want to that of the first algorithm. On the other hand, we show that there exist configurations of points such that any matching with the above properties matches fewer than 98.95% of the points. We further extend these results to point sets with a prescribed ratio of the sizes of the two color classes. In the end, we discuss the more general problem when the points are colored with any fixed number of colors.

Keywords: Non-crossing matching; Algorithm; Probabilistic construction

1. Introduction

We call a set S ($|S| \geqslant 3$) of points in the plane independent or in general position, if no three are on a line. Aharoni and Saks [10] considered the following problem: we are given a set S with $n = w + b$ points in general position in the plane, w of them white, and b of them black. Let $G(S)$ be a largest non-crossing matching of pairs of points of the same color, using straight line segments. Define $g(S)$ to be the number of points matched by $G(S)$. It is well known that if S is monochromatic, $G(S)$ can be computed in $O(n \log n)$ time, by sorting the points along a specified direction (e.g., by x-coordinate), and matching the first two points, the next two points and so on. Let $g(n) = \min\{g(S)\colon\ S \subset \mathbb{R}^2$ independent, $|S| = n\}$. They asked if it is always possible to match all but a constant number of points. It was shown in [5] that the answer is negative. In the special case of points in convex position, it is not difficult to show that the answer is affirmative.

E-mail addresses: ad@cs.uwm.edu (A. Dumitrescu), rick@kaye.to (R. Kaye).

0925-7721/01/$ – see front matter
PII: S0925-7721(01)00007-4

Our first result presented in Section 2, is an $\mathrm{O}(n^2)$-time algorithm which computes a matching of at least $\frac{6}{7}n$ of the points. Besides, it gives an improved lower bound on $g(n)$. The algorithm rotates a sweep-line until it finds a suitable direction which allows a reduction in the problem size; then the process is repeated. An asymptotically faster $\mathrm{O}(n\log n)$-time algorithm which comes arbitrarily close to this guarantee, is also provided. This second algorithm employs the divide-and-conquer paradigm, by dividing into subproblems of constant size. In Section 3, we give a more careful analysis of the probabilistic construction in [5], showing the existence of point sets where any non-crossing matching must leave at least $\frac{1}{95}n$ points unmatched. The previous bounds on $g(n)$ given in [5] were $\frac{5}{6}n$ and $\frac{155}{156}n$, respectively.

Theorem 1.1. *For all $n \geqslant 1$,*

$$\frac{6}{7}n - \mathrm{O}(1) \leqslant g(n) \leqslant \frac{94}{95}n + \mathrm{O}(1).$$

In Section 4, the previous results are extended to point sets with a prescribed ratio of the sizes of the two color classes. Let $r \geqslant 1$ be a fixed rational number. Put

$$g^{(r)}(n) = \min\{g(S)\colon\ S \subset \mathbb{R}^2 \text{ independent},\ |S| = n,\ w/b = r\}.$$

Theorem 1.2. *Let $r \in \mathbb{Q}$, $r \geqslant 1$. For all $n \geqslant 1$,*

$$\max\left(\frac{6}{7}, \frac{2r+1}{2r+2}\right)n - \mathrm{O}(1) \leqslant g^{(r)}(n) \leqslant c_r n + \mathrm{O}(1) \quad \textit{for some } c_r < 1.$$

In Section 5, we deal with the case when the points are colored by a fixed number of colors $k \geqslant 3$. We present an $\mathrm{O}(n\log n)$-time algorithm which finds a matching of at least $\frac{12}{6k+1} \cdot n$ of the points. As a byproduct, this gives an improved lower bound on $g_k(n)$, the analogue of $g(n)$ for k colors. The previous bound on $g_k(n)$ given in [5] was $\frac{4}{2k+1} \cdot n$.

Theorem 1.3. *For all $n \geqslant 1$ and each fixed $k \geqslant 3$,*

$$\frac{12}{6k+1} \cdot n - \mathrm{O}(1) \leqslant g_k(n) \leqslant \frac{2}{k} \cdot n + \mathrm{O}(1).$$

Finally, we mention that, to the best of our knowledge, the algorithmic complexity of computing $G(S)$ or $g(S)$ is open at the moment.

Related results

Matching is a well studied problem in graph theory which has attracted much attention in recent years in a geometric setting: matching a set of points in the plane using straight line segments. A usual question is what is a largest size matching under different constraints, such as pairwise non-crossing segments or pairwise crossing segments in the matching.

Since self-crossing in planar configurations is typically an undesirable attribute, the first results addressed non-crossing matchings. One can distinguish between the colored and uncolored version of the problem. For uncolored points, as mentioned in the introduction, if the number of points is even, a non-crossing perfect matching is possible and easy to find. We add here the fact that a minimum length perfect matching is another (efficiently computable) solution to the problem, since the triangle inequality ensures the non-crossing attribute. So the (related) problem of computing a minimum length matching gives for free a non-crossing one. As for the problem of computing a maximum length non-crossing matching, an $O(n \log n)$-time constant factor approximation algorithm is known [2], though the computational complexity of finding an optimal solution is currently unknown. Going back to the maximum size matching, in the two-colored version, the n points are divided into two equally-sized color classes and each segment must join points of different colors. One solution (and algorithm) which gives a perfect matching, if say the two colors are white and black, is to compute a minimum length white–black matching. Again the triangle inequality ensures the non-crossing attribute. Another solution involves the use of so-called *ham-sandwich cuts* [6]. A line which divides the points such that the sizes of the two color classes are equal in both parts is computed, and the argument (and corresponding algorithm) follows by induction (recursion) [11]. In another colored version of this problem (in this paper), the segments must connect points of the same color. A perfect (or almost perfect) matching is not always possible, though a linear size can be guaranteed.

If one insists instead on a pairwise crossing matching, the best results in this direction guarantee only a $\Omega(\sqrt{n})$ size matching, in both uncolored and colored (endpoints of different colors) version [1]. An $O(n \log n)$-time algorithm is provided for this task. On the other hand, no sub-linear upper bounds are known. A related result in the same direction, states that a pairwise crossing perfect matching exists (uncolored version) if and only if the point set S, where $|S| = n$, has precisely $n/2$ halving lines [9]. An $O(n \log n)$-time algorithm solves the decision problem and computes such a matching if it exists.

2. Two colors: algorithms and lower bounds

The next proposition and lemma are from [5].

Proposition 2.1 [5]. *For any integers $k \geqslant 2$, $n \geqslant 1$, $n_1, \ldots, n_k \geqslant 1$, such that $n = \sum_{i=1}^{i=k} n_i$,*

$$g(n) = g\left(\sum_{i=1}^{i=k} n_i\right) \geqslant \sum_{i=1}^{i=k} g(n_i).$$

The above inequality is evident when one thinks about sweeping a vertical line across the point set.

Lemma 2.2 [5]. $g(7) = 4$ *and if* $|S| = 7$, $g(S) = 6$ *unless the set S has a special structure, called a star configuration* (*seen in Fig. 1*): *the convex hull of S has 6 points of alternating colors with the 7th point inside, which we call the center of the star. If the center* (*of the star*) *is white, it is contained in the triangle formed by the 3 black points. If it is black, it is contained in the triangle formed by the 3 white points.*

For a set S of n points in general position in the plane, a k-set is a subset $S' \subseteq S$ such that $S' = S \cap h$ for some half-plane h, and $|S'| = k$ (see [6]). Without loss of generality, we can consider only open half-planes, or half-planes determined by lines that do not contain any of the points in S. For $x, y \in S$, we call the oriented segment $\overline{xy}$ a k-segment of S if its extension to an oriented line has exactly k points of S on its right side.

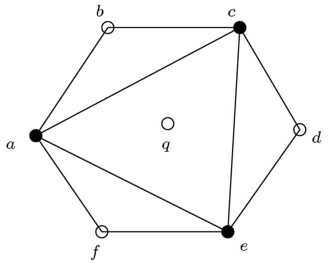

Fig. 1. A star configuration.

The next lemma is the basis for our improved lower bound and corresponding algorithm.

Lemma 2.3. *If $n \geqslant 12$, $g(n) \geqslant 6 + g(n-7)$.*

Proof. Let S be a set of at least 12 points. We can assume that by sweeping S from any direction until we pass 7 points, we get a star configuration, otherwise we are done. In other words, all 7-sets of S are stars. Assume, without loss of generality, that all points have distinct x-coordinates. Denote by C_0 the (7 point) star configuration consisting of the 7 points with the largest x-coordinates, in other words, the one which is obtained by sweeping a vertical line oriented upwards $l_{\theta=0}$ (where θ denotes the angle of the sweep-line) from right to left. Begin to rotate the sweep line clockwise, until we get a different 7-set C_1 when sweeping from that direction. Continue to rotate the sweep line l_θ until we get to sweep from the opposite direction, left to right, ($\theta = 180°$) and obtain along the way the sequence of star configurations $C_0, C_1, \ldots, C_k$, for some $k \geqslant 1$. The last term in the sequence, C_k, corresponds to $\theta = 180°$. For each $i = 0, \ldots, k-1$, C_i and C_{i+1} have 6 points in common. We write $\text{next}(C_i) = C_{i+1}$. By Lemma 2.5 (below), C_i and C_{i+1} have the same center (recall that a star has 6 extreme points and one interior point, the center). Thus $C_0, \ldots, C_k$ all have the same center. If S has at least 12 points, clearly the center points of C_0 and C_k must be different. Hence the assumption that all 7-sets of S are stars must have been false. $\square$

Corollary 2.4. *For every $k \geqslant 0$, $g(7k+12) \geqslant 6k+10$.*

Proof. Immediate from Lemma 2.3 and Proposition 2.1. $\square$

Lemma 2.5. *Let C and C' be two neighboring star configurations, such that $\text{next}(C) = C'$. Then C and C' have the same center.*

Proof. Without loss of generality, let C be C_0 (as in the previous lemma) separated by the vertical line l_0 from the other points. Denote by $p_0, p_1, p_2, p_3, p_4, p_5$, the extreme points of C in clockwise order, and by q the center of C. We can assume that p_0, the leftmost point of C, is black. Let $p = p_i$ be the point that is in C but not in C'. Fig. 2 shows the cases $i = 0$ (a) and $i = 1$ (b). It can be proved (though it is not needed in our arguments) that these two cases are the only possible, under the assumption of the lemma (that C and C' are both stars). Let p' be the "new" point that is in C' but not in C (drawn as a small square in Fig. 2). Note that the sweep line passes through p and p' when C is replaced by C'. Therefore, p could not have been q, the center of C. If $\text{color}(q) = \text{color}(p)$, q is

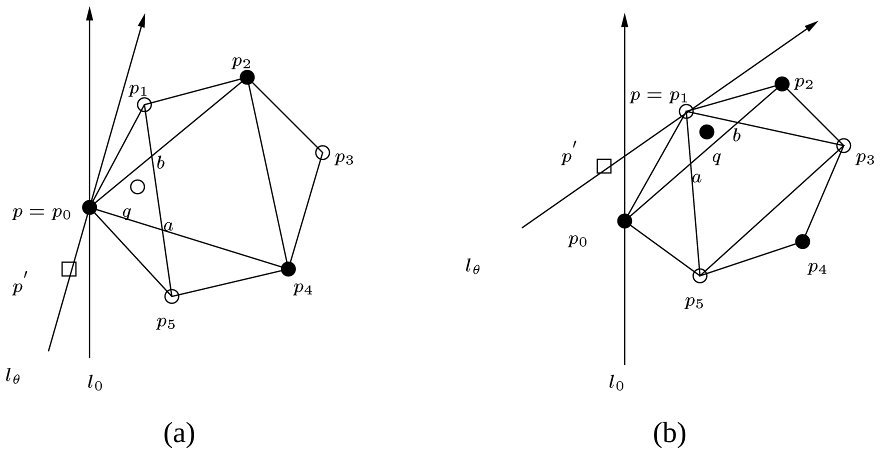

Fig. 2. Rotating sweep-line.

contained in the triangle $\{p_{i-1}, p_{i+1}, p_{i+3}\}$, so it remains interior in C' (all indices are taken modulo 6). So we can assume color$(q) \neq$ color(p). Then the point p' must have the same color as p, otherwise C' has 5 points of one color and 2 points of the other color, contradicting the structure of a star configuration. If color$(q) \neq$ color(p) and q is contained in either of the triangles $\{p_{i+1}, p_{i+2}, p_{i+3}\}$ or $\{p_{i-1}, p_{i-2}, p_{i+3}\}$, it remains interior in C'. The only remaining possibility is that q is in triangle $\{p, a, b\}$ where $a = (p_{i-2}, p_i) \cap (p_{i-1}, p_{i+1})$, $b = (p_i, p_{i+2}) \cap (p_{i-1}, p_{i+1})$. But then the six points in $C \setminus \{p\}$ are in convex position with p_{i-1}, q and p_{i+1} being three consecutive points of the same color, on the convex hull of $C \setminus \{p\}$. When we add point p' to these six points, we get C'. Since C' is a star, it must push one of these three points into the interior. Since the extreme points in stars alternate in color, the point that gets pushed into the middle must be the middle point of the three consecutive points, or in other words q. So q is the center of C'. □

Corollary 2.6. *Let p and p' be the points in which C and C' differ (i.e., $|C \cap C'| = 6$, $C \setminus C' = \{p\}$, $C' \setminus C = \{p'\}$). Then p and p' have the same color.*

Proof. C and C' must have the same center by Lemma 2.5. Therefore, p and p' must be exterior points in C and C' respectively. The exterior points of a star are always split evenly, three black and three white. Since the other five exterior points of C and C' are the same (either 3 black and two white or vice versa), p and p' must have the same color. □

Remark. Lemma 2.3 and its corollary contain the result $g(12) = 10$ from [5] as a special case.

Algorithm A1.
Input: a set S of n 2-colored points (white and black).
Output: a matching M.
Sort the points by x-coordinate (assume all distinct). It is known that using the following procedure, which we shall employ in the background with $k = 7$, one can generate all k-sets of S [7,8]. Sweep a vertical line l from right to left, until it passes through a point p_1 and leaves $k - 1$ points to the right of l. Orient l upwards, and let l_0 be this initial position. Rotate l clockwise (we denote by l_θ the current

position of l) around p_1 until it passes through another point p_2. Rotate l clockwise around p_2 until it passes through another point p_3. Continue this process until l rotates with 360° degrees, returning to its initial vertical upward orientation l_0. Let the sequence of (not necessarily distinct) points generated be $p_1, \ldots, p_r$. To find next point p_{i+1}, compute the minimum angle of rotation until a point is hit by l_θ. We distinguish two possibilities:
(a) an “L” hit is when $\overline{p_i p_{i+1}}$ has an opposite orientation with l_θ (i.e., $\overline{p_{i+1} p_i}$ is a $(k-1)$-segment);
(b) an “R” hit is when $\overline{p_i p_{i+1}}$ has the same orientation with l_θ (i.e., $\overline{p_i p_{i+1}}$ is a $(k-2)$-segment).
Update C_θ, the current k-set, when (a) occurs: $C_\theta := C_\theta \cup \{x\} \setminus \{y\}$, if x, y are the endpoints of the corresponding $(k-1)$-segment.

We now return to the description of our algorithm. Start with a vertical line $l_{\theta=0}$ oriented upwards, passing through a point of S, and having exactly 6 points in its open right half-plane. Consider C_0, the configuration of seven points defined in this way. We assume that C_0 is a star, otherwise, the algorithm recurses on a smaller size subset of points, after matching 6 out of 7 points in C_0. Let $L = \{a, b, c, d, e, f\}$ be the circular list containing the 6 extreme points of C_0 (see Fig. 1). Begin to rotate the sweep line l_θ, around the current point p_i (initially, $i = 1$, $p_1 = a$). Since a, b, c, d, e, f are in convex position, with clockwise orientation, l_θ hits them in this order. For each $u \in L$, let v be the element following u in this list. Consider the rotation of l_θ. Let θ_u be the value of the angle θ when l_θ hits point u for the first time. Similarly, we define θ_v. The *interval* $[u, v)$ consists of the points which lie on l_θ for $\theta \in [\theta_u + \varepsilon, \theta_v - \varepsilon]$, for a small $\varepsilon > 0$. Note that $u \in [u, v)$ and $v \notin [u, v)$. For example, the interval $[c, d)$ consists of point c and all the others hit by l_θ after it starts rotating around c, and before it hits point d. The *length* of the interval $[u, v)$ is the number of rotations of l_θ from the moment it hits point u and before it hits point v. The rotation at the end of which l_θ passes through u is counted, but not the one at the end of which l_θ passes through v (for the interval $[a, b)$, we count an extra rotation at $\theta = 0$). The algorithm keeps track of the lengths of the 6 intervals $[a, b), [b, c), [c, d), [d, e), [e, f), [f, a)$, as long as $\theta < 180°$. After at most 43 rotations, one of the following two favorable events E_1 or E_2 must occur (see also Lemma 2.7, to follow). Then the algorithm recurses on a smaller size subset of points, as indicated below.

1. Event E_1: A non-star configuration is found along the way; 6 points are matched out of 7 by 3 monochromatic segments, which are added to M. The algorithm recurses on the remaining set of $(n-7)$ points which are contained in a convex region (half-plane).
2. Event E_2: A convex chain $A = \{a_1, \ldots, a_8\}$ of 8 points of the same color is found, and a set B of 6 points, such that for $i = 1, \ldots, 7$, $\overline{a_i a_{i+1}}$ is a 6-segment, and the corresponding 6-set is B. Moreover, for $i = 1, \ldots, 8$, a_i and B form a star. In this case, for $i = 1, \ldots, 4$, we match a_{2i-1} with a_{2i} and 4 points of B (recall that $g(6) \geqslant 4$ [5]). The algorithm recurses on the remaining set C of $(n-14)$ points which are contained in a convex region (see also Fig. 3).

Lemma 2.7. *Assume $|S| \geqslant 12$. Start with $\theta = 0$. After at most 43 rotations, either*
(i) *a non-star has been found, hence 6 points are matched out of 7, and the remaining set of $(n-7)$ points are contained in a half-plane, or*
(ii) *12 points are matched out of 14, and the remaining set of $(n-14)$ points are contained in a convex region.*

Proof. If after at most 43 rotations, $\theta \geqslant 180°$, a non-star configuration has been found (sweeping across the point set from left and from right, gives star configurations having different centers, as in the proof of Lemma 2.3). So we can assume that for the first 43 rotations, $\theta < 180°$. We make the following

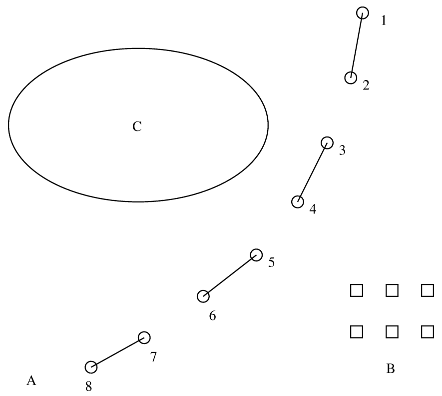

Fig. 3. Illustration of the proof of Lemma 2.7; $k = 8$.

observations. When $\theta = 0$, the $n-1$ points which are not yet hit by l_θ, are contained in two convex regions, $L_{\theta=0}$ and $R_{\theta=0}$, the left and right open half-planes determined by l_0. During rotation of l_θ, this property is maintained in the sense that the points which are not yet hit by l_θ, are contained in two open convex regions, L_θ and R_θ, $L_\theta \subseteq L_0$, $R_\theta \subseteq R_0$. The center q is always in R_θ (otherwise, if l_θ hits q at some point, q is not an interior point of the current star C_θ). If l_θ rotates around a point and the next point hit is an "L"-hit, the two points must have the same color by Corollary 2.6.

We monitor the number of hits of l_θ, as θ varies from $0°$ to its final value $\theta_{43} < 180°$ after the first 43 rotations. Since a, b, c, d, e, f are in convex position, l_θ hits them in this order. Consider the 6 half-closed intervals $[a,b), [b,c), [c,d), [d,e), [e,f), [f,a)$. Call any such interval *short* if its length is at most 7, and *long* otherwise. We notice that at least one of the above 6 intervals must be long, otherwise, after at most $6 \cdot 7 + 1 = 43$ rotations, $\theta \geqslant 180°$. Let $[u,v)$ be such a long interval, of length at least 8. Recall that $\mathrm{color}(u) \neq \mathrm{color}(v)$. When l_θ rotates around $u = f_1$, the first point hit, f_2, is an "L"-hit (an "R"-hit would be v, which cannot occur, since the interval $[u,v)$ is long). By an earlier observation, $\mathrm{color}(f_2) = \mathrm{color}(f_1)$. For $2 \leqslant k \leqslant 8$, let $f_2, \ldots, f_k$ be a maximum length consecutive subsequence of length $k-1 \leqslant 7$ of "L"-hits as it occurs during the next rotations of l_θ. The k distinct points $f_1, \ldots, f_k$ have the same color and form a convex chain.

If $k = 8$, $f_1, \ldots, f_8$ are matched by 4 segments along this chain, together with 4 out of the six points of C_θ, all together 12 points out of 14. The remaining $(n-14)$ points of S lie in the convex region $L_{\theta'}$, as noted earlier (θ' is given by the line which passes through f_{k-1} and f_k), thus they lie above the convex chain $f_1, \ldots, f_8$ extended to infinity at both ends. This is illustrated in Fig. 4. The points are numbered as they are hit by the rotating line (from 1 to 8 in this example).

If $k < 8$, as l_θ rotates around f_k, the next point hit r, is an "R"-hit. Note that $r \neq v$, since $[u,v)$ is a long interval. If $\mathrm{color}(f_k) = \mathrm{color}(r)$, we get a non-star configuration, having two adjacent points f_k, r, of the same color, hence 6 points can be matched out of 7. The remaining $(n-7)$ points of S lie in a convex region (half-plane), as noted earlier. If $\mathrm{color}(f_k) \neq \mathrm{color}(r)$, we get again a non-star configuration, having two adjacent points r, v, of the same color (here we have used the fact that point q is always interior in C_θ). So also in this case, 6 points are matched out of 7, and the remaining $(n-7)$ points of S lie in a convex region (half-plane).

An example is shown in Fig. 4, with $u = b$, $v = c$, $k = 4$. The points are numbered as they are hit by the rotating line (from 1 to 8 in this example). This concludes the proof of our lemma. $\square$

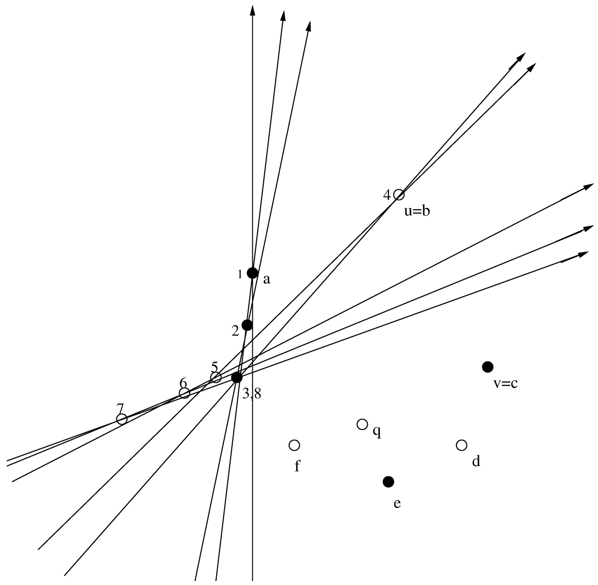

Fig. 4. Illustration of the proof of Lemma 2.7; $k < 8$.

In $\mathrm{O}(n)$ time, the algorithm achieves conditions (i) or (ii) in the lemma: there is a constant bound on the number of rotations, and each can be implemented in $\mathrm{O}(n)$ time, used to select the minimum value of a set of $(n-1)$ angles. Adding up the cost of the recursion, the total complexity of our algorithm is $\mathrm{O}(n^2)$. It is easy to see that it uses $\mathrm{O}(n)$ space. The algorithm guarantee is a matching of $\frac{6}{7}n - \mathrm{O}(1)$ of the points.

Algorithm A2. Given a positive $\varepsilon > 0$, choose a positive integer k, such that

$$\frac{6k+10}{7k+12} > \frac{6}{7} - \varepsilon.$$

After sorting the n points according to their x-coordinate and dividing them into groups of $7k+12$, $6k+10$ are matched in each group. An $\mathrm{O}(n \log n)$-time algorithm with a guarantee of $(\frac{6}{7} - \varepsilon)n - \mathrm{O}(1)$ of the points is obtained. The constant hidden in the O notation depends on ε. For example, to obtain a guarantee of 85%, the algorithm divides the points into groups of 40, and matches 34 in each group, (e.g., using Algorithm A1 on a constant size input).

3. Two colors: upper bounds

In all our constructions, all possible collinearities allowed by the description of the point set are avoided by small perturbations of the points.

In [5], the authors asked whether $g(14) = 10$ or 12 and whether $g(16) = 12$ or 14. We show here that in both cases the smaller value is correct. It is easy to see that $g(14) \geqslant 10$ and $g(16) \geqslant 12$.

The 14 point configuration in Fig. 5 has 5 white and 9 black points. To match 12 points out of 14, one has to leave only one point unmatched from each color. It is easy to see that neither of the two

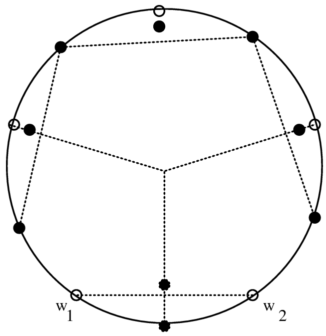

Fig. 5. A 14-point configuration.

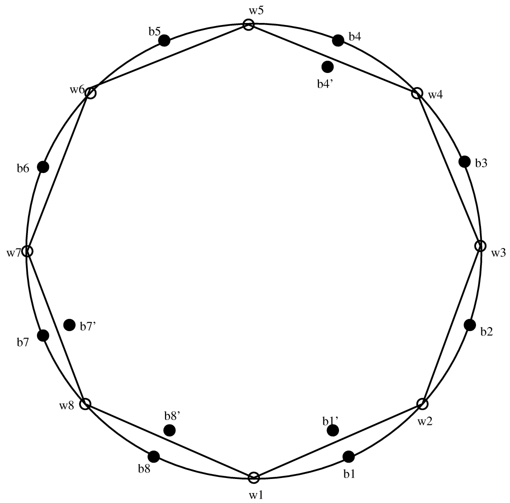

Fig. 6. A 20-point configuration describing a random construction.

white points w_1 and w_2 can be matched in any way while satisfying this requirement. Any matching of w_1 (or w_2) will force at least two black points to remain unmatched. This shows $g(14) \leqslant 10$. A 16 point configuration where at most 12 points can be matched, can be easily obtained from the 14 point configuration by adding 2 more black points near any one of the interior black points and following the same analysis. In conjunction with Corollary 2.4, this fills the gaps in the list of exact values of $g(n)$, for small n, that we previously had (see [5]): $g(1) = g(2) = 0$; $g(3) = g(4) = 2$; $g(5) = g(6) = g(7) = 4$; $g(8) = g(9) = 6$; $g(10) = g(11) = 8$; $g(12) = g(13) = g(14) = 10$; $g(15) = g(16) = 12$; $g(17) = g(18) = 14$; $g(19) = g(20) = 16$.

Next, we prove the upper bound in Theorem 1.1 through a more careful analysis of the following random construction (see [5]). For a given n, place n white and n black points alternately on a circle as a regular convex $2n$-gon, $w_1, b_1, \ldots, w_n, b_n$, say in counterclockwise order. For each $i = 1, \ldots, n$ we randomly place b_i', a twin of b_i on the other side of the segment $w_i w_{i+1}$ and close to the middle of this segment as in Fig. 6. The twin point is added with probability $\frac{1}{2}$ and independently for each i. This random configuration S has n white points and $n + Y$ black points, Y being the number of successes in n Bernoulli trials (with parameter $\frac{1}{2}$), and a total of $|S| = N = 2n + Y$ points in all. Clearly $2n \leqslant N \leqslant 3n$.

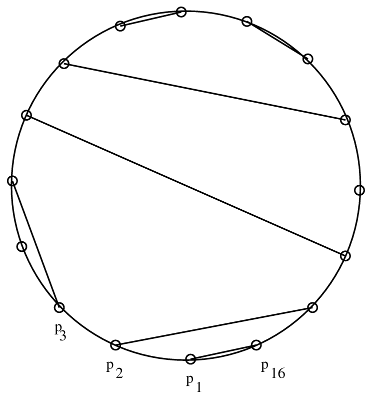

Fig. 7. A non-crossing matching of 16 points with 2 unmatched points and 3 sides.

We say that a matching has a certain color (white/black) if it matches only points of that particular color. For $q < \frac{1}{2}$, $q =$ constant, we study the events

$$A_1 = \{g(S) \geqslant N - qn\}, \qquad A_2 = \{|N - 2.5n| > n^{2/3}\}, \qquad A = A_1 \cup A_2. \tag{1}$$

A_1 is the event that our random set S admits a non-crossing matching with at most qn unmatched points; in other words, that there exist a white matching M and a black matching B such that at most qn points are unmatched using $M + B$ (this is a shorthand for $M \cup B$), a non-crossing matching of S. We will show that a sufficiently small q, (e.g., $q = \frac{1}{38}$) allows $\mathrm{Prob}(A) < 1$.

It is well known [3] that for $\alpha > 0$,

$$\mathrm{Prob}\left(\left|Y - \frac{n}{2}\right| \geqslant \alpha\right) \leqslant 2\mathrm{e}^{-2\alpha^2/n},$$

thus $\mathrm{Prob}(A_2) = \mathrm{o}(1)$, and in order to obtain $\mathrm{Prob}(A) < 1$ we will ensure that $\mathrm{Prob}(A_1) < 1$.

Fix M, a non-crossing white (imperfect in general) matching of S. If m is the number of matched points, M partitions the circle into $m' = m/2 + 1$ convex regions. We say that a region is odd if the number of black points inside it is odd and even otherwise. A region R is bounded by elements which could be either arcs of the circle or straight line segments in M. The segments could be either short chords, joining two adjacent white points of S, or long chords otherwise. When the region is bounded only by an arc and a short chord, we call it a *singleton* region since it contains exactly one black point. We call a *side* of M a matched pair of adjacent white points of S (see Fig. 7). A somewhat different meaning was attributed in [5].

We only consider white matchings, for which the number of (white) unmatched points plus the number of singleton regions is at most qn (otherwise the total number of unmatched points exceeds qn). Denote by $H(q) = -q \log q - (1 - q) \log(1 - q)$ the binary entropy of q (here log stands for the logarithm base 2). The next claim is easy to prove [5].

Claim 3.1 [5]. *Let R be a region determined by a white matching M. Then*

$$\mathrm{Prob}(R \text{ is odd}) \geqslant \frac{1}{2}.$$

More precisely, if R is a singleton region, $\mathrm{Prob}(R \text{ is odd}) = 1$; *otherwise* $\mathrm{Prob}(R \text{ is odd}) = \frac{1}{2}$.

Denote by $\mathrm{Bin}(n, p)$ the binomial random variable with parameters n, p (the number of successes in n Bernoulli trials with success probability p). The number of odd regions which are not singletons is distributed as a binomial random variable: $Z \sim \mathrm{Bin}(m' - s, \frac{1}{2})$, where s stands for the number of singleton regions determined by M out of m'. Put x_2 for the number of unmatched points in M.

Claim 3.2. *Let q' be a constant, $0 < q' \leqslant q < \frac{1}{2}$. For M and Z defined earlier,*

$$\mathrm{Prob}\big(Z \leqslant q'n\big) \leqslant D = \frac{2^{H(2q')n/2}}{2^{(1-2q)n/2}}.$$

Proof. Since $m + x_2 = n$ and $x_2 + s \leqslant qn$,

$$m' - s = \frac{m}{2} + 1 - s \geqslant \frac{n - x_2 - 2s}{2} \geqslant \frac{n - 2x_2 - 2s}{2} \geqslant \frac{n(1-2q)}{2}$$

and

$$m' - s \leqslant \frac{n}{2},$$

from which we get

$$\mathrm{Prob}\big(Z \leqslant q'n\big) \leqslant \sum_{0 \leqslant k \leqslant q'n} \binom{m'-s}{k} \frac{1}{2^{m'-s}} \leqslant \sum_{0 \leqslant k \leqslant q'n} \binom{n/2}{k} \frac{1}{2^{(1-2q)n/2}} \leqslant \frac{2^{H(2q')n/2}}{2^{(1-2q)n/2}}.$$

In the last line we used the following known bound on the sum of binomial coefficients (see [4] for a proof, see also [3] for a similar inequality): if $0 < q \leqslant \frac{1}{2}$ is a constant,

$$\sum_{0 \leqslant m \leqslant nq} \binom{n}{m} \leqslant 2^{H(q)n}. \tag{2}$$

□

We bound the probability of A_1 in (1). Put $l = qn$, and let $a, b < \frac{1}{2}$ be two positive constants to be specified later. Denote the white points by $p_1, \ldots, p_n$ along the circle in say clockwise order. Encode each matching by a $\{0, 1, 2\}$ string of length n as follows: scan the points in increasing order (from p_1 to p_n) and for each matched point, write a 0 if the endpoint belongs to a segment which is seen for the first time and a 1 otherwise (if it is the second time). Write a 2 for each unmatched point. If the number of matched points is m, we obtain a string containing $m/2$ 0's and $m/2$ 1's. For our example in Fig. 7, the encoding is 0002100010112111.

This is an injective function from the set of matchings to the set of $\{0, 1, 2\}$ strings of length n (it is implied by the non-crossing condition). Each side corresponds to a 01 transition, with at most one exception, if it matches p_1 with p_n. If a matching has s' sides, there are at most s' 01 transitions and at most $(s' - 1)$ 10 transitions overall. One can specify such an encoding by first choosing the positions of the 2's in the string, then specifying the $0/1$ character after each maximal consecutive substring of 2's and finally specifying the positions of the 01 and 10 transitions in the remaining free spots (there are less than $n - x_2$ such positions from which to select). The string should start with a 0 or 2 and end with a 1 or 2; it may in general not correspond to a valid matching since it may have, for example, a different number of 0's and 1's.

Let M have x_2 unmatched points and s singleton regions. Then in the corresponding string, the number of 2's is x_2, and the number of 01 transitions is $y_{01} \leqslant s$. The total number of 01 and 10 transitions is $y_{01} + y_{10} \leqslant 2y_{01} \leqslant 2s$. Making repeated use of Claim 3.2, we prove that $q = \frac{1}{38}$ implies $\mathrm{Prob}(A_1) < 1$.

$$\mathrm{Prob}(A_1) \leqslant \sum_{M: z = x_2 + y_{01} \leqslant l} \mathrm{Prob}(Z \leqslant l - s - x_2) \leqslant P_1 + P_2,$$

where

$$P_1 = \sum_{M: x_2 \leqslant a \cdot l, z \leqslant l} \mathrm{Prob}(Z \leqslant l - s - x_2), \qquad P_2 = \sum_{M: a \cdot l \leqslant x_2 \leqslant l, z \leqslant l} \mathrm{Prob}(Z \leqslant l - s - x_2).$$

$$\begin{aligned} P_1 &\leqslant \sum_{0 \leqslant x_2 \leqslant a \cdot l} \binom{n}{x_2} 2^{x_2} \Biggl\{ \sum_{0 \leqslant p \leqslant 2b \cdot l} \binom{n - x_2}{p} \mathrm{Prob}(Z \leqslant l) + \sum_{2b \cdot l \leqslant p \leqslant 2l} \binom{n - x_2}{p} \mathrm{Prob}(Z \leqslant l - b \cdot l) \Biggr\} \\ &\leqslant 2^{H(a \cdot q)n} \cdot 2^{a \cdot q \cdot n} \bigl\{ 2^{H(2b \cdot q)n} \cdot 2^{H(2q)n/2} + 2^{H(2q)n} \cdot 2^{H(2(1-b)q)n/2} \bigr\} \frac{1}{2^{(1-2q)n/2}} \\ &\leqslant 2^{(E_1(a,b,q)-1)n/2} \quad \text{for } n \geqslant n_1, \end{aligned}$$

where

$$E_1(a,b,q) = 2H(a \cdot q) + 2a \cdot q + \max\bigl\{2H(2b \cdot q) + H(2q), 2H(2q) + H\bigl(2(1-b)q\bigr)\bigr\} + 2q.$$

The first sum in the expression of P_1 bounds the contribution to $\mathrm{Prob}(A_1)$ of the white matchings with a small number of unmatched points ($\leqslant a \cdot l$) and a small number of singleton regions ($\leqslant b \cdot l$). The second sum in the expression of P_1 bounds the contribution to $\mathrm{Prob}(A_1)$ of the white matchings with a small number of unmatched points ($\leqslant a \cdot l$) and a large number of singleton regions ($\geqslant b \cdot l$). Similarly, we can upper bound P_2, the contribution to $\mathrm{Prob}(A_1)$ of the white matchings with a large number of unmatched points ($\geqslant a \cdot l$).

$$\begin{aligned} P_2 &\leqslant \sum_{a \cdot l \leqslant x_2 \leqslant l} \binom{n}{x_2} \cdot 2^{x_2} \Biggl\{ \sum_{0 \leqslant p \leqslant 2b \cdot l} \binom{n - x_2}{p} \mathrm{Prob}\bigl(Z \leqslant (l - a \cdot l)\bigr) \\ &\qquad\qquad + \sum_{2b \cdot l \leqslant p \leqslant 2l} \binom{n - x_2}{p} \mathrm{Prob}(Z \leqslant l - a \cdot l - b \cdot l) \Biggr\} \\ &\leqslant 2^{H(q)n} \cdot 2^{q \cdot n} \bigl\{ 2^{H(2b \cdot q)n} \cdot 2^{H(2(1-a)q)n/2} + 2^{H(2(1-a)q)n} \cdot 2^{H(2(1-a-b)q)n/2} \bigr\} \frac{1}{2^{(1-2q)n/2}} \\ &\leqslant 2^{(E_2(a,b,q)-1)n/2} \quad \text{for } n \geqslant n_2, \end{aligned}$$

where

$$\begin{aligned} E_2(a,b,q) = {} & 2H(q) + 2q + \max\bigl\{2H(2b \cdot q) + H\bigl(2(1-a)q\bigr), 2H\bigl(2(1-a)q\bigr) \\ & + H\bigl(2(1-a-b)q\bigr)\bigr\} + 2q. \end{aligned}$$

We would like to determine some values for a, b and a value for q as large as possible, while satisfying $E_1(a,b,q) < 1$, $E_2(a,b,q) < 1$. It can be checked that $E_1(0.32, 0.44, 0.02633) < 1$, $E_2(0.32, 0.44, 0.02633) < 1$. Thus for $n \geqslant n_0$, $P_1 < \frac{1}{2}$, $P_2 < \frac{1}{2}$, giving $\mathrm{Prob}(A_1) < 1$. It is easy to see that for fixed a, b, $\lim_{q \to 0} E_1(a,b,q) = 0$ and $\lim_{q \to 0} E_2(a,b,q) = 0$ and that $E_1(a,b,\cdot)$ and $E_2(a,b,\cdot)$ are increasing functions on the interval $(0, \frac{1}{4})$. In the end, we choose $q = \frac{1}{38} < 0.02633$.

Since $N \approx 2.5n$ with high probability, we conclude that at least $\frac{1}{(2.5)\cdot 38} = \frac{1}{95}$ of the total number of points are unmatched with positive probability, so there exists a configuration with this property as claimed. To prove the upper bound in Theorem 1.1 for all n, we use the following easy statement.

Claim 3.3. *Let $U \subseteq V$ be two point sets. Then $|g(V) - g(U)| \leqslant 2|V \setminus U|$.*

In the last step of our point set construction, add (or delete) arbitrarily $|N - 2.5n|$ points to (respectively from) S to get the final point configuration S', where $|S'| = 2.5n$ (we have assumed n is even). We note that the number of added (or deleted) points is $o(n)$, and since the above inequality on q is strict, the multiplicative constant in our upper bound is not affected. This completes the proof of the upper bound in Theorem 1.1.

4. Two colors: point sets with a prescribed color ratio

Let $r = w/b$ be the *color ratio* of a point set having w white and b black points, where $w \geqslant b$. Motivated by the fact that our upper bound construction is not balanced (its color ratio is ≈ 1.5) the following question arises: what happens for balanced point sets (for which $r = 1$), or for highly unbalanced ones (say with $r = 1000$)?

Next, we prove Theorem 1.2. A lower bound of $\frac{6}{7}n - O(1)$ holds by Theorem 1.1; we show a better lower bound for $r > 2.5$.

Lemma 4.1. *Given a family S of n pairwise disjoint segments in the plane, whose endpoints are in general position, and an arbitrary ordering of the segments $s_1, s_2, \ldots, s_n$, extend (in the given order) each segment in both directions until it hits another segment, or a segment extension, or to infinity. Then when the process is complete, the plane will be partitioned into $n + 1$ convex regions.*

Proof. The statement is an easy consequence of Euler's formula for planar graphs. The details are left to the reader. □

Construct a non-crossing matching B of $2\lfloor b/2\rfloor$ of the black points using $\lfloor b/2\rfloor$ segments (at most one point remains unmatched). Use Lemma 4.1 to obtain a convex partitioning (of the plane) by segment extension. Then match the white points in each convex region (at most one per region remains unmatched). In this way we have obtained a matching of $n - b/2 - O(1)$ points; since $r = w/b$, this means $\frac{2r+1}{2r+2}n - O(1)$ points.

Now we describe the upper bound construction, which is a modification of the one used earlier in the unrestricted case: first select n odd; then follow the same steps as for the construction in Fig. 6. Recall that we now have n white points and $n + Y$ black points, where Y stands for the number of black twin points obtained after n coin flips. Let $r = s/t$. Place a cluster of $(t\lceil (n+Y)/t\rceil (s/t) - n)$ white points very close to the center of the circle, but otherwise arbitrarily (see Fig. 8).

We refer to this group of points as the (*center*) *cluster*. Place $(t\lceil (n+Y)/t\rceil - (n+Y))$ black points in a small cluster somewhere outside the circle. Note that $w = s\lceil (n+Y)/t\rceil$ and $b = t\lceil (n+Y)/t\rceil$, so the resulting set S has the prescribed color ratio. Assume for simplicity that $t \mid (n+Y)$, so there are no black points outside the circle (in general, the size of the black cluster is bounded by a constant, t, and its influence can be ignored for large n). In this case, the size of the center cluster is $(n+Y)r - n$.

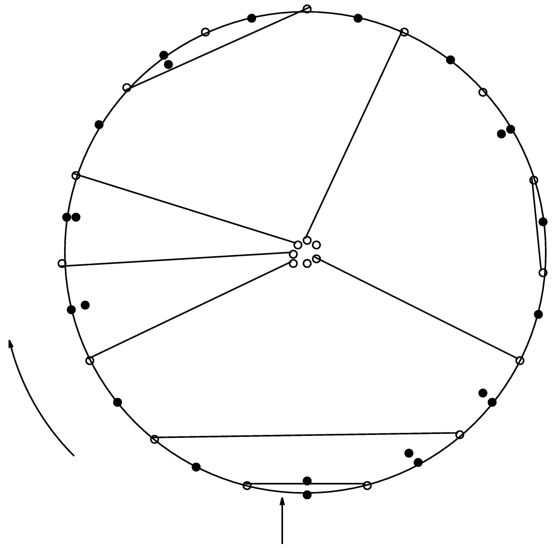

Fig. 8. A white matching in the modified construction.

Let $M + B$ be a matching of points in S. As in Section 3, we restrict our attention to white matchings M for which the number of (white) unmatched points plus the number of singleton regions is at most qn (otherwise the total number of unmatched points exceeds qn); we call such a matching *q-good*. A white matching M consists of three types of segments: (*circle*) *chords*, (*circle*) *rays* and *cluster segments*. Chords match two points on the circle, rays match a point in the cluster with a point on the circle and cluster segments match two points in the cluster. Since n is odd, any chord leaves the center cluster on one side of it. We assume M contains at least two rays, the other case is easy. We distinguish two types of regions: a *chord region* is a (convex) region whose boundary consists of circle arcs and chords; a *sector region* is a region which has two ray segments on its boundary. At most one non-convex sector region can exist. Two adjacent rays can be viewed as having a common endpoint in the center cluster – in the sense that the set of endpoints of rays in the cluster is irrelevant, see below.

Two white matchings are said to be equivalent if they determine the same pattern on the circle, i.e., if the set of points on the circle which are matched by rays and the set of chords are the same in both. So two equivalent matchings will create the same sets of sector (respectively chord) regions. For two equivalent matchings, we will generously assume that the points in the center cluster which are unmatched by rays, are perfectly matched inside the cluster without creating any crossings. From now on, when referring to a white matching, we will not make any distinction between two equivalent ones.

Analogously to the procedure described in the previous section, one can encode a set of equivalent white matchings by a $\{0, 1, 2, 3\}$ string of length n. We write a 3 for a point on the circle which is matched to a point in the cluster (it is irrelevant to which one); the other encoding conventions are the same. For the example in Fig. 8 (with $r = 1$, $Y = 7$), the encoding is 003330213201311. Let x_i be the number of symbols i in the string. We have $x_0 + x_1 + x_2 + x_3 = n$, where $x_0 = x_1$. The number of (sector and chord) regions is $x_0 + x_3$.

Claim 4.2. *The number of q-good white matchings is at most $2^{E(q)n}$, for some function $E(\cdot)$ such that $\lim_{q\to 0} E(q) = 0$ and $E(\cdot)$ is increasing on the interval $(0, \frac{1}{4})$.*

Proof. For a sector region R, define its size to be the number of white points on the circle in between the two rays, and denote it by $\mathrm{size}(R)$. If $\mathrm{size}(R) \geqslant 1$, we call R a *large sector region*, otherwise, a *small sector region*. Observe that each large sector region determines at least one unmatched point: if $\mathrm{size}(R)$ is odd, there will be an unmatched white point in between its extreme rays; if $\mathrm{size}(R)$ is even, the white points in between its extreme rays must form a perfect (non-crossing) matching, so there will exist an adjacent pair connected by a segment (a side of M), leaving an unmatched black point.

To get an upper bound on the number of *q-good* white matchings, we bound the number of encodings. To specify an encoding we choose

(1) the positions of 2 in the string (recall $x_2 \leqslant qn$) and the next symbol after each maximal string of 2's,
(2) the starting positions of large sector regions (their number is also at most qn by the above observation),
(3) the ending positions of large sector regions,
(4) the starting positions of maximal sequences of consecutive small sector regions (their number is bounded by the number of large sector regions, if this number is positive, otherwise is one),
(5) the ending positions of maximal sequences of consecutive small sector regions,
(6) the positions of 01 (respectively 10) transitions in the string.

The number of choices for each item i above is bounded by $2^{e_i(q)n}$ (using (2)) for some function $e_i(\cdot)$ such that $\lim_{q\to 0} e_i(q) = 0$ and $e_i(\cdot)$ is increasing on the interval $(0, \frac{1}{4})$. The claim follows from the fact that the family of these functions is closed with respect to finite sums. □

Claim 4.3. *For each* (*sector or chord*) *region R,*

$$\mathrm{Prob}(R \text{ is odd}) \geqslant \frac{1}{2}.$$

Proof. Clear from construction; see also Claim 3.1, and [5] for a formal proof. □

Claim 4.4. *The number of* (*sector and chord*) *regions is $\geqslant (n - x_2)/2 \geqslant (n - qn)/2$.*

Proof.

$$2(x_0 + x_3) + x_2 \geqslant x_0 + x_1 + x_2 + x_3 = n.$$

The claim is readily implied. □

We want to bound from above $\mathrm{Prob}(A_1)$ as in Section 3: a calculation similar to the one made in [5] (or along the lines of the proof in Section 3) goes through (i.e., $\mathrm{Prob}(A_1) < 1$) based on the fact that the total number of white matchings has been reduced to the number of inequivalent ones. Putting together all of the above we obtain the upper bound in Theorem 1.2. A similar argument to the one given in the proof of Theorem 1.1 shows that the statement holds for any n. The details are left to the reader.

5. Three or more colors

Let $k \geqslant 3$ be the number of colors (fixed). The upper bound in Theorem 1.3 was proved in [5]. Here we prove the lower bound.

Lemma 5.1. *For $k \geqslant 3$, $g_k(6k+1) = 12$.*

Proof. First we prove the statement for $k = 3$. $g_3(19) \leqslant g_3(21) \leqslant (\frac{2}{3})21 - 2 = 12$ (see [5]). To prove the opposite inequality, consider a set S, with $|S| = 19$ points colored by 3 colors $\{1, 2, 3\}$. Write the sizes of the three color classes in nondecreasing order: $n_1 \leqslant n_2 \leqslant n_3$, where $n_1 + n_2 + n_3 = 19$.

If $n_1 \leqslant 4$, $n_2 + n_3 \geqslant 15$ thus $g_3(S) \geqslant g_2(15) \geqslant 3g(5) = 12$.

If $n_1 = 5$ and $n_2 \leqslant 6$, then $n_3 \geqslant 8$. Since $n_1 + n_2 \geqslant 10$ and $g_2(10) = 8$, 8 points of colors 1 and 2 can be matched by 4 disjoint segments. Using Lemma 4.1, by extending these segments, the plane is divided into 5 convex regions. Then either there exists a region containing 4 points of color 3 (from at least 8 of this color), or there exist 2 regions each containing at least 2 points of color 3. In either case, 4 more points can be matched, giving a total of 12 matched points.

If $n_1 = 5$ and $n_2 = 7$, or $n_1 = 6$ and $n_2 = 6$, we have $n_1 + n_2 = 12$ and $n_3 = 7$. Since $g_2(12) = 10$, 10 points of colors 1 and 2 can be matched using 5 disjoint segments. Again using Lemma 4.1, by extending these segments, the plane is divided into 6 convex regions. There must be 2 points (out of 7) of color 3 in one of these regions, which can be matched. Again the total is 12 matched points.

Next we prove the statement for any $k \geqslant 3$. $g_k(6k+1) \leqslant g_k(7k) \leqslant \frac{2}{k}(7k) - 2 = 12$ (see [5]). To show the opposite inequality, consider a set S, with $|S| = n = 6k + 1$ points colored by $k \geqslant 3$ colors. Write the sizes of the color classes in nondecreasing order: $n_1 \leqslant n_2 \leqslant \cdots \leqslant n_k$, where $\sum_{i=1}^{i=k} n_i = n$. Then $n_{k-2} + n_{k-1} + n_k \geqslant 19$ (otherwise $n_i \leqslant 6, \forall i \leqslant k-3$, which implies $\sum_{i=1}^{i=k} n_i = \sum_{i=1}^{i=k-3} n_i + \sum_{i=k-2}^{i=k} n_i \leqslant 6(k-3) + 18 = 6k < n$, a contradiction). Using the result for $k = 3$, $g_3(19) = 12$, we get $g_k(6k+1) \geqslant g_3(19) = 12$. $\square$

Algorithm. The n points are sorted according to their x-coordinate and divided into groups of $6k+1$; then 12 are matched in each group. This is done by matching 12 out of 19 points in the largest three color classes (by Lemma 5.1, with $k = 3$). The time to process a group is $O(k)$, so the total time is $O(n \log n)$. The number of matched points is bounded as in the theorem.

Acknowledgements

We would like to thank János Pach for his suggestions which improved our presentation.

References

[1] B. Aronov, P. Erdős, W. Goddard, D. Kleitman, M. Klugerman, J. Pach, L. Schulman, Crossing families, in: Proceedings of the 7th Annual Symposium on Computational Geometry, 1991, pp. 351–356; also Combinatorica 14 (1994) 127–134.

[2] N. Alon, S. Rajagopalan, S. Suri, Long non-crossing configurations in the plane, in: Proceedings of the 9th Annual Symposium on Computational Geometry, 1993, pp. 257–263.

[3] N. Alon, J. Spencer, The Probabilistic Method, Wiley, New York, 1992.
[4] P. Beame, M. Saks, J. Thathachar, Time-space tradeoffs for branching programs, in: 39th Annual Symposium on Foundations of Computer Science, 1998, pp. 254–263.
[5] A. Dumitrescu, W. Steiger, On a matching problem in the plane, in: Workshop on Algorithms and Data Structures (WADS '99), 1999; also Discrete Mathematics 211 (1–3) (2000) 183–195.
[6] H. Edelsbrunner, Algorithms in Combinatorial Geometry, Springer, New York, 1987.
[7] P. Erdős, L. Lovász, A. Simmons, E.G. Straus, Dissection graphs of planar point sets, in: J.N. Srivastava et al. (Eds.), A Survey of Combinatorial Theory, North-Holland, Amsterdam, 1973, pp. 139–149.
[8] L. Lovász, On the number of halving lines, Ann. Univ. Sci. Budapest Eötvös Sect. Math. 14 (1971) 107–108.
[9] J. Pach, J. Solymosi, Halving lines and perfect cross-matchings, Contemporary Mathematics 223 (1999) 245–249.
[10] M. Saks, Personal communication, 1998.
[11] W. Steiger, Personal communication, 1998.