ON THE GENERAL POSITION SUBSET SELECTION PROBLEM∗

MICHAEL S. PAYNE† AND DAVID R. WOOD‡

Abstract. Let \( f(n, \ell) \) be the maximum integer such that every set of \( n \) points in the plane with at most \( \ell \) collinear contains a subset of \( f(n, \ell) \) points with no three collinear. First we prove that if \( \ell \leq O(\sqrt{n}) \), then \( f(n, \ell) \geq \Omega(\sqrt{n}/\ln \ell) \). Second we prove that if \( \ell \leq O(n^{1-\epsilon}/2) \), then \( f(n, \ell) \geq \Omega(\sqrt{n \log n}) \), which implies all previously known lower bounds on \( f(n, \ell) \) and improves them when \( \ell \) is not fixed. A more general problem is to consider subsets with at most \( k \) collinear points in a point set with at most \( \ell \) collinear. We also prove analogous results in this setting.

Key words. general position, Erdős problems, discrete geometry

AMS subject classification. 52C10

DOI. 10.1137/120897493

1. Introduction. A set of points in the plane is in general position if it contains no three collinear points. The general position subset selection problem asks, given a finite set of points in the plane with at most \( \ell \) collinear, how big is the largest subset in general position? That is, determine the maximum integer \( f(n, \ell) \) such that every set of \( n \) points in the plane with at most \( \ell \) collinear contains a subset of \( f(n, \ell) \) points in general position. Throughout this paper we assume \( \ell \geq 3 \). Furthermore, as the results in this paper are all asymptotic in \( n \), the expression “fixed \( \ell \)” is shorthand for “\( \ell \) a constant not dependent on \( n \).” Otherwise \( \ell \) is allowed to grow as a function of \( n \).

The problem was originally posed by Erdős, first for the case \( \ell = 3 \) [8], and later in a more general form [9]. Füredi [10] showed that the density version of the Hales–Jewett theorem [11] implies that \( f(n, \ell) \leq o(n) \), and that a result of Phelps and Rödl [20] on independent sets in partial Steiner triple systems implies that

\[
f(n, 3) \geq \Omega(\sqrt{n \ln n}).
\]

Until recently, the best known lower bound for \( \ell \geq 4 \) was \( f(n, \ell) \geq \sqrt{2n/(\ell - 2)} \), proved by a greedy selection algorithm. Lefmann [16] showed that for fixed \( \ell \),

\[
f(n, \ell) \geq \Omega(\sqrt{n \ln n}).
\]

(In fact, his results are more general; see section 3.)

In relation to the general position subset selection problem (and its relatives), Brass, Moser, and Pach [2, p. 318] write, “To make any further progress, one needs to explore the geometric structure of the problem.” We do this by using the Szemerédi–Trotter theorem [25].

We give improved lower bounds on \( f(n, \ell) \) when \( \ell \) is not fixed, with the improvement being most significant for values of \( \ell \) around \( \sqrt{n} \). Our first result (Theorem 2.3)
says that if \( \ell \leq O(\sqrt{n}) \), then \( f(n, \ell) \geq \Omega(\sqrt{n\ln n}) \). Our second result (Theorem 2.5) says that if \( \ell \leq O(n^{(1-c)/2}) \), then \( f(n, \ell) \geq \Omega(\sqrt{n \log \ell n}) \). For fixed \( \ell \), this implies Lefmann’s lower bound on \( f(n, \ell) \) mentioned above.

In section 3 we consider a natural generalization of the general position subset selection problem. Given \( k < \ell \), Erdős [9] asked for the maximum integer \( f(n, \ell, k) \) such that every set of \( n \) points in the plane with at most \( \ell \) collinear contains a subset of \( f(n, \ell, k) \) points with at most \( k \) collinear. Thus \( f(n, \ell) = f(n, \ell, 2) \). We prove results similar to Theorems 2.3 and 2.5 in this setting too.

2. Results. Our main tool is the following lemma.

Lemma 2.1. Let \( P \) be a set of \( n \) points in the plane with at most \( \ell \) collinear. Then the number of collinear triples in \( P \) is at most \( c(n^2 \ln \ell + \ell^2 n) \) for some constant \( c \).

Proof. For \( 2 \leq i \leq \ell \), let \( s_i \) be the number of lines containing exactly \( i \) points in \( P \). A well-known corollary of the Szemerédi–Trotter theorem [25] states that for some constant \( c \geq 1 \), for all \( i \geq 2 \),

\[
\sum_{j \geq i} s_j \leq c \left( \frac{n^2}{i^3} + \frac{n}{i} \right).
\]

Thus the number of collinear triples is

\[
\sum_{i=2}^{\ell} \binom{i}{3} s_i \leq \sum_{i=2}^{\ell} i^2 \sum_{j=i}^{\ell} s_j \leq \sum_{i=2}^{\ell} ci^2 \left( \frac{n^2}{i^3} + \frac{n}{i} \right)
\]

\[
\leq c \sum_{i=2}^{\ell} \left( \frac{n^2}{i} + in \right) \leq c(n^2 \ln \ell + \ell^2 n). \]

Note that Lefmann [15] proved Lemma 2.1 for the case of the \( \sqrt{n} \times \sqrt{n} \) grid via a direct counting argument. A statement similar to Lemma 2.1 with \( \ell = \sqrt{n} \) also appears in the book by Tao and Vu [26, Corollary 8.8].

To apply Lemma 2.1 it is useful to consider the 3-uniform hypergraph \( H(P) \) determined by a set of points \( P \), with vertex set \( P \), and an edge for each collinear triple in \( P \). A subset of \( P \) is in general position if and only if it is an independent set in \( H(P) \). The size of the largest independent set in a hypergraph \( H \) is denoted \( \alpha(H) \). Spencer [23] proved the following lower bound on \( \alpha(H) \).

Lemma 2.2 (Spencer [23]). Let \( H \) be an \( r \)-uniform hypergraph with \( n \) vertices and \( m \) edges. If \( m < n/r \), then \( \alpha(H) > n/2 \). If \( m \geq n/r \), then

\[
\alpha(H) > \frac{r - 1}{r^r(r - 1)} \frac{n}{(m/n)^{1/(r - 1)}}.
\]

Lemmas 2.1 and 2.2 imply our first result.

Theorem 2.3. Let \( P \) be a set of \( n \) points with at most \( \ell \) collinear. Then \( P \) contains a set of \( \Omega(n/\sqrt{n \ln \ell + \ell^2}) \) points in general position. In particular, if \( \ell \leq O(\sqrt{n}) \), then \( P \) contains a set of \( \Omega(\sqrt{n \ln \ell}) \) points in general position.

Proof. Let \( m \) be the number of edges in \( H(P) \). By Lemma 2.1, \( m/n \leq c(n \ln \ell + \ell^2) \) for some constant \( c \). Now apply Lemma 2.2 with \( r = 3 \). If \( m < n/3 \), then \( \alpha(H(P)) > n/2 \), as required. Otherwise,

\[
\alpha(H(P)) > \frac{2n}{3^3/2(m/n)^{1/2}} \geq \frac{2n}{3^{3/2} \sqrt{c(n \ln \ell + \ell^2)}} = \frac{2n}{3 \sqrt{3c} \sqrt{n \ln \ell + \ell^2}}.
\]
Note that Theorem 2.3 also shows that if \( \ell^2/\ln \ell \geq n \), then \( f(n, \ell) \geq \Omega(n/\ell) \). This improves upon the greedy bound mentioned in the introduction, and is within a constant factor of optimal, since there are point sets with at most \( \ell \) collinear that can be covered by \( n/\ell \) lines.

Theorem 2.3 answers, up to a logarithmic factor, a symmetric Ramsey-style version of the general position subset selection problem posed by Gowers [13]. He asked for the minimum integer \( \text{GP}(q) \) such that every set of at least \( \text{GP}(q) \) points in the plane contains \( q \) collinear points or \( q \) points in general position. Gowers noted that \( \Omega(q^2) \leq \text{GP}(q) \leq O(q^3) \). Theorem 2.3 with \( \ell = q - 1 \) and \( n = \text{GP}(q) \) implies that \( \Omega(\sqrt{\text{GP}(q)/\ln(q-1)}) \leq q \) and so \( \text{GP}(q) \leq O(q^2 \ln q) \).

The bound \( \text{GP}(q) \geq \Omega(q^2) \) comes from the \( q \times q \) grid, which contains no \( q + 1 \) collinear points, and no more than \( 2q + 1 \) in general position, since each row can have at most \( 2 \) points. Determining the maximum number of points in general position in the \( q \times q \) grid is known as the \textit{no-three-in-line problem}, first posed by Dudeney in 1917 [4]. See [14] for the best known bound and for more on its history.

As an aside, note that Pach and Sharir [18] proved a result somewhat similar to Lemma 2.1 for the number of triples in \( P \) determining a fixed angle \( \alpha \in (0, \pi) \). Their work did not consider the parameter \( \ell \), without which the case \( \alpha = 0 \) is exceptional since \( P \) could be entirely collinear, and all triples could determine the same angle.

The following lemma of Sudakov [24, Proposition 2.3] is a corollary of a result by Duke, Lefmann, and Rödl [5].

**Lemma 2.4 (Sudakov [24]).** Let \( H \) be a 3-uniform hypergraph on \( n \) vertices with \( m \) edges. Let \( t \geq \sqrt{m/n} \) and suppose there exists a constant \( \epsilon > 0 \) such that the number of edges containing any fixed pair of vertices of \( H \) is at most \( t^{1-\epsilon} \). Then \( \alpha(H) \geq \Omega\left(\frac{n}{t^{2}} \sqrt{\ln t}\right) \).

Lemmas 2.1 and 2.4 can be used to prove our second result.

**Theorem 2.5.** Fix constants \( \epsilon > 0 \) and \( d > 0 \). Let \( P \) be a set of \( n \) points in the plane with at most \( \ell \) collinear points, where \( \ell \leq (dn)^{(1-\epsilon)/2} \). Then \( P \) contains a set of \( \Omega(\sqrt{n \log n}) \) points in general position.

**Proof.** Let \( m \) be the number of edges in \( H(P) \). By Lemma 2.1, for some constant \( c \geq 1 \),

\[
m \leq c\ell^2 n + cn^2 \ln \ell < cdn^2 + cn^2 \ln \ell \leq (d + 1)cn^2 \ln \ell.
\]

Define \( t := \sqrt{(d + 1)cn \ln \ell} \). Thus \( t \geq \sqrt{m/n} \). Each pair of vertices in \( H \) is in less than \( \ell \) edges of \( H \), and

\[
\ell \leq (dn)^{(1-\epsilon)/2} < ((d + 1)cn \ln \ell)^{(1-\epsilon)/2} = t^{1-\epsilon}.
\]

Thus the assumptions in Lemma 2.4 are satisfied. So \( H \) contains an independent set of size \( \Omega\left(\frac{n}{t^{2}} \sqrt{\ln \ell}\right) \). Moreover,

\[
\frac{n}{t} \sqrt{\ln \ell} = \sqrt{\frac{n}{(d + 1)cn \ln \ell}} \sqrt{\ln(\sqrt{(d + 1)cn \ln \ell})} \geq \sqrt{\frac{n}{(d + 1)cn \ln \ell}} \sqrt{\frac{1}{2} \ln n}.
\]
Thus $P$ contains a subset of $\Omega(\sqrt{n \log\ell} n)$ points in general position.

3. Generalizations. In this section we consider the function $f(n, \ell, k)$ defined to be the maximum integer such that every set of $n$ points in the plane with at most $\ell$ collinear contains a subset of $f(n, \ell, k)$ points with at most $k$ collinear, where $k < \ell$.

Brass [1] considered this question for fixed $\ell = k + 1$ and showed that

\[
o(n) \geq f(n, k + 1, k) \geq \Omega(n^{(k-1)/k}(\ln n)^{1/k}).\]

This can be seen as a generalization of the results of Füredi [10] for $f(n, 3, 2)$. As in Füredi’s work, the lower bound comes from a result on partial Steiner systems [22], and the upper bound comes from the density Hales–Jewett theorem [12]. Lefmann [16] further generalized these results for fixed $\ell$ and $k$ by showing that

\[
f(n, \ell, k) \geq \Omega(n^{(k-1)/k}(\ln n)^{1/k}).\]

The density Hales–Jewett theorem also implies the general bound $f(n, \ell, k) \leq o(n)$.

The result of Lefmann may be generalized to include the dependence of $f(n, \ell, k)$ on $\ell$ for fixed $k \geq 3$, analogously to Theorems 2.3 and 2.5 for $k = 2$. The first result we need is a generalization of Lemma 2.1. It is proved in the same way.

**Lemma 3.1.** Let $P$ be a set of $n$ points in the plane with at most $\ell$ collinear. Then, for $k \geq 4$, the number of collinear $k$-tuples in $P$ is at most $c(\ell^{k-3}n^2 + k^{-1}n)$ for some absolute constant $c$.

Lemmas 2.2 and 3.1 imply the following theorem, which is proved in the same way as Theorem 2.3.

**Theorem 3.2.** If $k \geq 3$ is fixed and $\ell \leq O(\sqrt{n})$, then $f(n, \ell, k) \geq \Omega\left(n^{(k-1)/k}\right)$.

For $\ell = \sqrt{n}$ and fixed $k \geq 3$, Theorem 3.2 implies $f(n, \sqrt{n}, k) \geq \Omega\left(n^{(k-1)/k}\right) = \Omega\left(n^{(2k-2-k+2)/2k}\right) = \Omega(\sqrt{n})$. This answers completely a generalized version of Gowers’ question [13], namely, to determine the minimum integer $\text{GP}_k(q)$ such that every set of at least $\text{GP}_k(q)$ points in the plane contains $q$ collinear points or $q$ points with at most $k$ collinear, for $k \geq 3$. Thus $\text{GP}_k(q) \leq O(q^2)$. The bound $\text{GP}_k(q) \geq \Omega(q^2)$ comes from the following construction. Let $m := \lceil (q-1)/k \rceil$ and let $P$ be the $m \times m$ grid. Then $P$ has at most $m$ points collinear, and $m < q$. If $S$ is a subset of $P$ with at most $k$ collinear, then $S$ has at most $k$ points in each row. So $|S| \leq km \leq q - 1$.

Theorem 2.5 can be generalized using Lemma 3.1 and a theorem of Duke, Lefmann, and Rödl [5] (the one that implies Lemma 2.4).

**Theorem 3.3 (Duke, Lefmann, and Rödl [5]).** Let $H$ be a $k$-uniform hypergraph with maximum degree $\Delta(H) \leq t^{k-1}$ where $t \gg k$. Let $p_t(H)$ be the number of pairs of edges of $H$ sharing exactly $j$ vertices. If $p_j(H) \leq n!^{2k-j-1} \gamma$ for $j = 2, \ldots, k-1$ and some $\gamma > 0$, then $\alpha(H) \geq C(k, \gamma)^{1/2}(\ln n)^{1/(k-1)}$ for some constant $C(k, \gamma) > 0$.

**Theorem 3.4.** Fix constants $d > 0$ and $\epsilon \in (0, 1)$. If $k \geq 3$ is fixed and $4 \leq \ell \leq dn^{1-\epsilon}/2$, then

\[
f(n, \ell, k) \geq \Omega\left(\frac{n^{(k-1)/k}}{\ell^{(k-2)/k}}(\ln n)^{1/k}\right)\]

**Proof.** Given a set $P$ of $n$ points with at most $\ell$ collinear, a subset with at most $k$ collinear points corresponds to an independent set in the $(k + 1)$-uniform hypergraph
$H_{k+1}(P)$ of collinear $(k+1)$-tuples in $P$. By Lemma 3.1, the number of edges in $H_{k+1}(P)$ is $m \leq c(n^2 k^{k-2} + n^k)$ for some constant $c$. The first term dominates since $\ell \leq o(\sqrt{n})$. For $n$ large enough, $m/n \leq 2cn^{k-2}$.

To limit the maximum degree of $H_{k+1}(P)$, discard vertices of degree greater than $2(k+1)m/n$. Let $\bar{n}$ be the number of such vertices. Considering the sum of degrees, $(k+1)m \geq \bar{n}2(k+1)m/n$, and so $\bar{n} \leq n/2$. Thus discarding these vertices yields a new point set $P'$ such that $|P'| \geq n/2$ and $\Delta(H_{k+1}(P')) \leq 4(k+1)cn^{k-2}$. Note that an independent set in $H_{k+1}(P')$ is also independent in $H_{k+1}(P)$.

Set $t := (4(k+1)cn^{k-2})^{1/k}$, so $m \leq 2^{1/(k+1)}nt^k$ and $\Delta(H_{k+1}(P')) \leq t^k$, as required for Theorem 3.3. By assumption, $\ell \leq dn^{(1-\epsilon)/2}$. Thus

$$\ell \leq d\left(\frac{t^k\ell^{2-k}}{4(k+1)c}\right)^{\frac{1}{\epsilon'}}.$$

Hence $\frac{\ell}{\ell^{k-2}+k-2} \leq \frac{d^{2/(1-\epsilon)}t^k}{4(k+1)c}$, implying $\ell \leq C_1(k)\ell^{1-\epsilon}\times t^{1-\epsilon} = C_1(k)\ell^{1-\epsilon} \times t^{1-\epsilon}$ for some constant $C_1(k)$. Define $\epsilon' := 1 - \frac{1}{1+\epsilon\epsilon'}$, so $\epsilon' > 0$ (since $\epsilon < 1$) and $\ell \leq C_1(k)\ell^{1-\epsilon'}$. To bound $p_j(H_{k+1}(P'))$ for $j = 2, \ldots, k$, first choose one edge (which determines a line), then choose the subset to be shared, then choose points from the line to complete the second edge of the pair. Thus for $\gamma := \epsilon'/2$ and sufficiently large $n$,

$$p_j(H_{k+1}(P')) \leq m\left(\begin{array}{c} k+1 \\ j \end{array}\right)\left(\frac{\ell - k - 1}{k + 1 - j}\right) \leq C_2(k)\epsilon t^{k+1-\gamma} \leq C_2(k)(C_1(k))^{k+1-j}nt^k(1-\epsilon')(k+1-j) \leq nt^{2k+1-j-1-\gamma}.$$

Hence the second requirement of Theorem 3.3 is satisfied. Thus

$$\alpha(H_{k+1}(P')) \geq \Omega\left(\frac{n}{t}(\ln t)^{1/k}\right) \geq \Omega\left(\frac{n^{(k-1)/k}}{t^k(k-2)^{1/k}}\left(\ln((nt^{k-2})^{1/k})\right)^{1/k}\right) \geq \Omega\left(\frac{n^{(k-1)/k}}{t^k(k-2)^{1/k}}(\ln n)^{1/k}\right).$$

4. Conjectures. Theorem 3.2 suggests the following conjecture, which would completely answer Gowers’ question [13], showing that $\text{GP}(q) = \Theta(q^2)$. It is true for the $\sqrt{n} \times \sqrt{n}$ grid [14], [7, Appendix].

**Conjecture 4.1.** $f(n, \sqrt{n}) \geq \Omega(\sqrt{n})$.

A natural variation of the general position subset selection problem is to color the points of $P$ with as few colors as possible, such that each color class is in general position. An easy application of the Lovász local lemma shows that under this requirement, $n$ points with at most $\ell$ collinear are colorable with $O(\sqrt{n})$ colors. The following conjecture would imply Conjecture 4.1. It is also true for the $\sqrt{n} \times \sqrt{n}$ grid [27].

**Conjecture 4.2.** Every set $P$ of $n$ points in the plane with at most $\sqrt{n}$ collinear can be colored with $O(\sqrt{n})$ colors such that each color class is in general position.

The following proposition is somewhat weaker than Conjecture 4.2.
Proposition 4.3. Every set \( P \) of \( n \) points in the plane with at most \( \sqrt{n} \) collinear can be colored with \( O(\sqrt{n}\ln^{3/2} n) \) colors such that each color class is in general position.

Proof. Color \( P \) by iteratively selecting a largest subset in general position and giving it a new color. Let \( P_0 := P \). Let \( C_i \) be a largest subset of \( P_i \) in general position and let \( P_{i+1} := P_i \setminus C_i \). Define \( n_i := |P_i| \). Applying Lemma 2.1 to \( P_i \) shows that \( H(P_i) \) has \( O(n_i^2 \ln \ell + \ell^2 n_i) \) edges. Thus the average degree of \( H(P_i) \) is at most \( O(n_i \ln \ell + \ell^2) \), which is \( O(n \ln n) \) since \( n_i \leq n \) and \( \ell \leq \sqrt{n} \).

Applying Lemma 2.2 gives \(|C_i| = \alpha(H(P_i)) > cn_i/\sqrt{n\ln n} \) for some constant \( c > 0 \). Thus \( n_i \leq n(1 - c/\sqrt{n\ln n}) \). It is well known (and not difficult to show) that if a sequence of numbers \( m_i \) satisfies \( m_i \leq m(1 - 1/x)^i \) for some \( x > 1 \) and if \( j > x \ln m \), then \( m_j \leq 1 \). Hence if \( k \geq \sqrt{n\ln n} \ln n/c \), then \( n_k \leq 1 \), so the number of colors used is \( O(\sqrt{n}\ln^{3/2} n) \).

The problem of determining the correct asymptotics of \( f(n, \ell) \) for fixed \( \ell \) remains wide open. The Szemerédi–Trotter theorem is essentially tight for the \( \sqrt{n} \times \sqrt{n} \) grid [19], but says nothing for point sets with bounded collinearities. For this reason, the lower bounds on \( f(n, \ell) \) for fixed \( \ell \) remain essentially combinatorial. Finding a way to bring geometric information to bear in this situation is an interesting challenge.

Conjecture 4.4. If \( \ell \) is fixed, then \( f(n, \ell) \geq \Omega(n/\text{polylog}(n)) \).

The point set that gives the upper bound \( f(n, \ell) \leq o(n) \) (from the density Hales–Jewett theorem) is the generic projection to the plane of the \([\log n] \)-dimensional \( \ell \times \ell \times \cdots \times \ell \) integer lattice (henceforth \( [\ell]^d \), where \( d := \lceil \log(n) \rceil \)). The problem of finding large general position subsets in this point set for \( \ell = 3 \) is known as Moser’s cube problem [17, 21], and the best known asymptotic lower bound is \( \Omega(n/\sqrt{n\ln n}) \) [3, 21].

In the coloring setting, the following conjecture is equivalent to Conjecture 4.4 by an argument similar to that of Proposition 4.3.

Conjecture 4.5. For all fixed \( \ell \geq 3 \), every set of \( n \) points in the plane with at most \( \ell \) collinear can be colored with \( O(\text{polylog}(n)) \) colors such that each color class is in general position.

Conjecture 4.5 is true for \( [\ell]^d \), which can be colored with \( O(d^{\ell-1}) \) colors as follows. For each \( x \in [\ell]^d \), define a signature vector in \( \mathbb{Z}^\ell \) whose entries are the number of entries in \( x \) equal to 1, 2, \ldots, \ell. The number of such signatures is the number of partitions of \( d \) into at most \( \ell \) parts, which is \( O(d^{\ell-1}) \). Give each set of points with the same signature its own color. To see that this is a proper coloring, suppose that \( \{a, b, c\} \subseteq [\ell]^d \) is a monochromatic collinear triple, with \( b \) between \( a \) and \( c \). Permute the coordinates so that the entries of \( b \) are nondecreasing. Consider the first coordinate \( i \) in which \( a_i \), \( b_i \), and \( c_i \) are not all equal. Then without loss of generality \( a_i < b_i \). But this implies that \( a \) has more entries equal to \( a_i \) than \( b \) does, contradicting the assumption that the signatures are equal.

Acknowledgments. Thanks to Timothy Gowers for posting his motivating problem on MathOverflow [13]; to Moritz Schmitt and Louis Theran for interesting discussions, and for making us aware of reference [13]; to János Pach for pointing out references [16] and [18]; and to the referee for helpful feedback and for pointing out reference [26].

References