Finite Analogs of Szemerédi's Theorem

Paul RAFF[1] and *Doron ZEILBERGER*[1]

Szemerédi's Celebrated Theorem

One of the crowning achievements of combinatorics is

Szemerédi's Theorem ([S]): Given an integer $n \geq 1$ and an integer $k \geq 3$, let $r_k(n)$ denote the size of any largest subset S of $[n] := \{1, 2, \ldots, n\}$ for which there are **no** subsets of the form

$$\{i, i+d, i+2d, \ldots, i+(k-1)d\} \quad (i \geq 1 \quad , \quad 1 \leq d < \infty) \quad ,$$

then $r_k(n) = o(n)$.

The depth and mainstreamness of this remarkable theorem is amply supported by the fact that at least four Fields medalists (Klaus Roth, Jean Bourgain, Tim Gowers, and Terry Tao) and at least one Wolf prize winner (Hillel Furstenberg) made significant contributions.

This article is yet another such contribution, and while it may not have the "depth" of the work of the above-mentioned human luminaries, it does have *one* advantage over them. We "cheat" and use a computer. It is true that, so far, we can only talk about *finite* analogs, but we do believe that the present approach could be eventually extended to sharpen the current rather weak bounds.

More specifically, we prove:

Finite version of Szemerédi's Theorem: Given an integer $n \geq 1$ and integers $k \geq 3$, $D \geq 1$, let $R_{k,D}(n)$ denote the size of any largest subset S of $[n] := \{1, 2, \ldots, n\}$ for which there are **no** subsets of the form

$$\{i, i+d, i+2d, \ldots, i+(k-1)d\} \quad (\, i \geq 1 \quad , \quad 1 \leq d \leq D\,) \quad ,$$

then there exists a rational number $\alpha_{k,D} = P_{k,D}/Q_{k,D}$ such that

$$\lim_{n \to \infty} \frac{R_{k,D}(n)}{n} = \alpha_{k,D} \quad .$$

We have (rigorously!) computed $\alpha_{k,D}$ for small k and D in the table below.

---

[1] Department of Mathematics, Rutgers University (New Brunswick), Hill Center-Busch Campus, 110 Frelinghuysen Rd., Piscataway, NJ 08854-8019, USA. `[praff,zeilberg] at math dot rutgers dot edu` ,
`http://www.math.rutgers.edu/~[praff,zeilberg]` . First written: July 13, 2009. Second version: Oct. 20, 2009 (a few typos corrected, thanks to Brian Nakamura). This version: Nov. 11, 2009 (adding a few more details for the obtuse reader). Accompanied by the Maple package `ENDRE`, as well as Mathematica and Java programs downloadable from
`http://www.math.rutgers.edu/~zeilberg/mamarim/mamarimhtml/szemeredi.html` .
The work of DZ was supported in part by the USA National Science Foundation.

	3	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
1	$\frac{2}{3}$	$\frac{3}{4}$	$\frac{4}{5}$	$\frac{5}{6}$	$\frac{6}{7}$	$\frac{7}{8}$	$\frac{8}{9}$	$\frac{9}{10}$	$\frac{10}{11}$	$\frac{11}{12}$	$\frac{12}{13}$	$\frac{13}{14}$
2	$\frac{2}{3}$	$\frac{2}{3}$	$\frac{4}{5}$	$\frac{4}{5}$	$\frac{6}{7}$	$\frac{6}{7}$	$\frac{8}{9}$	$\frac{8}{9}$	$\frac{10}{11}$	$\frac{10}{11}$	$\frac{12}{13}$	$\frac{12}{13}$
3	$\frac{4}{8}$	$\frac{8}{12}$	$\frac{4}{5}$	$\frac{4}{5}$	$\frac{6}{7}$	$\frac{6}{7}$	$\frac{6}{7}$	$\frac{20}{23}$	$\frac{10}{11}$	$\frac{10}{11}$	$\frac{12}{13}$	$\frac{12}{13}$
4	$\frac{4}{9}$	$\frac{3}{5}$	$\frac{4}{5}$	$\frac{4}{5}$	$\frac{6}{7}$	$\frac{6}{7}$	$\frac{6}{7}$	$\frac{26}{30}$	$\frac{10}{11}$	$\frac{10}{11}$	$\frac{12}{13}$	$\frac{12}{13}$
5	$\frac{4}{9}$	$\frac{4}{7}$	$\frac{16}{24}$	$\frac{22}{30}$	$\frac{6}{7}$							
6	$\frac{4}{9}$	$\frac{4}{7}$										
7	$\frac{4}{9}$	$\frac{6}{11}$										
8	$\frac{4}{9}$	$\frac{6}{11}$										
9	$\frac{4}{10}$											
10	$\frac{4}{11}$											
11	$\frac{8}{24}$											
12	$\frac{56}{177}$											
13	$\frac{6}{19}$											
14	$\frac{6}{19}$											
15	$\frac{6}{19}$											
16	$\frac{6}{19}$											
17	$\frac{6}{19}$											

These numbers can get difficult to compute very quickly, but it can be seen, for example, that $\alpha_{k,1} = \frac{k-1}{k}$. It turns out that even more is true. $R_{k,D}(n)$ is a quasi-linear function of n (i.e. a quasi-polynomial of degree 1) and for $i = 1, \ldots, Q_{k,D}$ there exist integers $a_{k,D,i}$ between 0 and $P_{k,D} - 1$ such that

$$R_{k,D}([Q_{k,D}] \cdot n + i) = [P_{k,D}] \cdot n + a_{k,D,i} \quad .$$

Our proof is algorithmic, and we show how to find these explicit expressions using **rigorous experimental mathematics**.

Note that $\alpha_{k,D}$ is a non-increasing sequence in D, and Szemerédi's theorem is equivalent to the statement that

$$\lim_{D \to \infty} \alpha_{k,D} = 0 \quad .$$

A Wordy Formulation

Every subset S of $[1, n] = \{1, 2, 3, \ldots, n\}$ corresponds to an n-letter word in the alphabet $\{0, 1\}$ defined by $w[i] = 1$ if and only if $i \in S$. S has an arithmetical progression of size k if there is an *Equidistant Letter Sequence* in the sense of the **Bible Codes** of the word 1^k (i.e. 1 repeated k times). Denoting by 2 a place where the occupying letter may be either 0 or 1, we can say that the $r_k(n)$ of Szemerédi's theorem defined above asks to find the maximal number of 1's that an n-letter word in $\{0, 1\}$ may have, that avoids the *infinitely* many patterns

$$(12^d)^{k-1}1 \quad , \quad 0 \le d < \infty.$$

Analogously, the $R_{k,D}(n)$ of the finite-version Szemerédi's theorem defined above asks to find the maximal number of 1's that an n-letter word in $\{0,1\}$ may have, that avoids the *finitely* many patterns

$$(12^d)^{k-1}1 \quad , \quad (0 \leq d \leq D-1) .$$

Define the *weight* of a word w to be $t^{length} z^{\#\ of\ 1s}$. Let $F_{k,D}(z,t)$ be the weight-enumerator of all binary words avoiding the D patterns $(12^d)^{k-1}1 \quad , \quad (0 \leq d \leq D-1)$. We will soon see that $F_{k,D}(z,t)$ is a rational function in (z,t).

Let's treat the more general case of an *arbitrary* set of *generalized patterns.* But let's first define *generalized pattern.*

Definition: A *generalized pattern* is a word in the alphabet $\{0,1,2\}$, where 2 stands for "space".

Now let's say what it means to *contain* a pattern.

Definition: A word $w = w_1 w_2 \dots w_n$ in the alphabet $\{0,1\}$ contains the pattern $p = p_1 p_2 \dots p_m$ if there exists a position i $(1 \leq i \leq n-m+1)$ such that

$$w_{i+j-1} = p_j \quad , \quad if \quad p_j \neq 2 \quad , \quad j = 1, \dots, m \quad .$$

For example, the word 011101101 contains the pattern 12221 (with $i=3$).

A word w avoids a generalized pattern p if it does *not* contain it. A word w avoids a set of generalized patterns P if w avoids all the members of P.

Analogous definitions can be made for an arbitrary finite alphabet, where we can use SPACE (_) instead of 2. We will now digress to that general scenario, and later specialize back to the binary case.

The General Problem

Consider a finite alphabet A together with a symbol SPACE(to be denoted by _) not in A. We are interested in weight-enumerating the set of words that avoid a set of patterns P, according to the weight

$$weight(w_1 w_2 \dots w_n) = x[w_1]x[w_2] \cdots x[w_n] \quad ,$$

where $x[a]$ $(a \in A)$ are *commuting indeterminates.* For example, $weight(PAUL) = x[P]x[A]x[U]x[L] = x[A]x[L]x[P]x[U]$, $weight(DORON) = x[D]x[N]x[O]^2x[R]$.

Let F be the weight-enumerator (sum of weights of its members, a formal power series in the variables $\{x[a], a \in A\}$) of the set of such words (that avoid P), let's call it, for reasons to become clear shortly, $S[P,\emptyset]$. A word belonging to it is either empty, or else starts with one of the letters of our alphabet. If you chop that letter, what remains is a shorter word in $S[P,\emptyset]$, but with *more* conditions, since it can not *start* with a “chopped pattern” obtained by chopping-off the first letter for all those patterns of P that happen to start with that letter or with $\_$.

This motivates the following

Definition: Given a word or pattern $w = w_1 w_2 \dots w_n$, let $BEHEAD(w) := w_2 \dots w_n$.

For example, $BEHEAD(DORON) = ORON$, $BEHEAD(PAUL) = AUL$, $BEHEAD(\_\_L\_OVE) = \_L\_OVE$.

Let P be a set of patterns, and let a be any letter of our alphabet A, then let

$$P/a := \{\ BEHEAD(p)\ |\ p \in P \quad and \quad (\ p_1 = a \quad or \quad p_1 = \_\ )\} \quad .$$

For example, if the alphabet is $\{0, 1\}$, and

$$P = \{000, 0\_0\_0, 0\_\_0\_\_0, 111, 1\_1\_1, 1\_\_1\_\_1, \_\_101\} \quad ,$$

then

$$P/0 = \{00, \_0\_0, \_\_0\_\_0, \_101\} \quad ,$$

$$P/1 = \{11, \_1\_1, \_\_1\_\_1, \_101\} \quad .$$

So if w belongs to our set $S[P, \emptyset]$ and it starts with the letter a, say, then the chopped word obviously also avoids P but in addition avoids P/a at the very beginning. This motivates us to make yet another

Definition: Let P and P' be sets of patterns. The set $S[P, P']$ consists of all words avoiding the patterns in P and in addition avoiding the patterns P' at the very beginning.

Since every word in $S[P, P']$ must be either empty or else begin with one of the letters of our alphabet A, we have the linear equation, for the weight-enumerators $F[P, P'](\{x[a]\})$,

$$F[P, P'] = 1 + \sum_{a \in A} x[a] F[P, P/a \cup P'/a] \quad .$$

If P' contains an empty pattern, then of course we have the **initial condition** $F[P, P'] = 0$, since not even the empty word avoids the empty word as a factor.

Of course, we only care about $F[P, \emptyset]$, but in order to compute it, we need to set up a system of linear equations featuring lots of $F[P, P']$ with many other (unwanted!) P', but nevertheless *finitely* many of them. Since the different values of P' that show up on the right side always contain shorter patterns, and eventually we get P' that contain the empty pattern so that we can use the initial condition, we get *finitely many* (but possibly a very large number) of equations, and *as many equations as unknowns* (because every unknown has its own equation, and we keep going until there are no new unknowns). Also, since we know from the outset that a solution exists (from the combinatorics), it follows that the system of equations is non-singular, and by Cramer's rule that we have a *rational function* in the variables

$$\{x[a] \quad | \quad a \in A \quad \} \quad .$$

Specializing

Going back to the Szemerédi scenario, we have a two-letter alphabet $\{0,1\}$ with weight $x[0] = t, x[1] = zt$. For *any* set of forbidden patterns, in particular, those that avoid arithmetical progression of size k with spacings $\leq D$, the generating function is of the form

$$R(z,t) = \frac{P(z,t)}{Q(z,t)} \quad ,$$

where t keeps track of the length of words and z keeps track of their number of 1s.

Expanding $R(z,t)$ as a power-series of t, we get

$$R(z,t) = \sum_{n=0}^{\infty} r_n(z) t^n \quad ,$$

and $r_n(z)$ is a polynomial whose *degree* (in z) is the largest number 1's in an n-letter word avoiding the set of generalized patterns. By looking at the monomials of the denominator, $Q(z,t)$, and searching for the monomial $z^i t^j$ with *largest* ratio $r := i/j$, we get that the largest number of 1's in an n-letter word in $\{0,1\}$ is asymptotically nr, and more precisely, we have the behavior described above for $R_{k,D}(n)$, as a certain quasi-linear discrete function. (More verbosely: we write $Q(z,t) = 1 - Q_1(z,t)$, where $Q_1(z,t)$ is a sum of non-constant monomials, then we expand $(1-Q_1)^{-1} = \sum_{m=0}^{\infty} Q_1^m$, and use the multinomial theorem for Q_1^m, and look at the coefficient of a typical monomial t^n).

An Experimental-Yet-Rigorous Shortcut

Solving a huge system of linear equations with *symbolic* coefficients is very time- and memory-consuming. Restricting attention to the alphabet $\{0,1\}$, and letting $f(P,P')(n)$ be the maximum number of 1's in an n-letter word that avoids the patterns in P and in addition, at the beginning, the patterns in P', we get, for $n > 0$,

$$f(P,P')(n) \; = \; max(\quad f(P, P/0 \cup P'/0) \; (n-1) \quad , \quad f(P, P/1 \cup P'/1) \; (n-1) \, + \, 1 \quad) \quad .$$

(Remember that *any* word in $\{0,1\}^n$, not just the one with the largest number of ones avoiding P and P', must start with either a 0 or a 1!). We ask the computer to *first* find the **scheme**, in terms of a binary tree where the left-child of P' is $P/0 \cup P'/0$ and its right-child is $P/1 \cup P'/1$. Then we ask the computer to *crank-out* lots of data, say, the first 500,000 terms (or whatever is needed), and then the computer automatically *guesses* explicit expressions of the form

$$R_{k,D}([Q_{k,D}] \cdot n + i) = [P_{k,D}] \cdot n + a_{k,D,i} \quad , i = 1 \ldots Q_{k,D} \quad ,$$

for certain integers $P_{k,D}$, $Q_{k,D}$, and $a_{k,D,i}$. Once guessed, the computer *automatically* gives a fully rigorous proof, *a posteriori*, by checking all the above equations, this time *symbolically*. See the sample output of ENDRE at the webpage of this article for an example.

Supporting Software

All this is implemented in the Maple package ENDRE. A Mathematica program is also provided, but only for the problems in the context of Szemerédi's Theorem. For efficiency's sake, a Java program is also available. See the webpage http://www.math.rutgers.edu/~zeilberg/mamarim/mamarimhtml/szemeredi.html for these packages, as well as sample input and output.

Exact Enumeration

From Sloane's point of view, it is interesting to crank-out as many terms as possible of $R_{k,D}(n)$, both for their own sake, and also because they offer upper bounds for $r_k(n)$. The interesting and efficient methods of the recent paper [GGK], that treats $r_3(n)$, may be useful to output more terms of $R_{k,D}(n)$ for larger D, but of course our focus is completely different. We do *symbol-crunching* rather than *number-crunching*.

The entries from the above table for $\alpha_{k,D}$, imply upper bounds for $r_4(n), r_5(n), \ldots$.

The Maple package ENDRE also contains programs for the *straight enumeration* of words of length n avoiding a set of generalized patterns, and for computing generating functions, from which the exact asymptotics of the enumerating sequence can be easily determined.

Finite Version of van der Waerden

van der Waerden's theorem (for two colors) tells you that $w_k(n)$, the number of n-letter words in the alphabet $\{0,1\}$, that avoids the generalized patterns

$$(12^d)^{k-1}1 \quad , \quad (02^d)^{k-1}0 \quad , \quad (\ 0 \le d < \infty\ )$$

is eventually 0. It is still of interest to investigate the finite version, $W_{k,D}(n)$, the number of n-letter words in the alphabet $\{0,1\}$, that avoids the generalized patterns

$$(12^d)^{k-1}1 \quad , \quad (02^d)^{k-1}0 \quad , \quad (\ 0 \le d \le D-1\ ) \quad .$$

The Maple package ENDRE can handle these problems as well.

Pipe dreams

For a fixed k, $\alpha_{k,D}$ gets harder and harder to compute as D gets larger and larger, **but** we believe that a clever analysis of the max equations, might lead, one day, to a *quantitative* understanding of how $\alpha_{k,D}$ decreases with D, that may (who knows?) lead to an easier proof of Szemerédi's theorem, and more importantly, improved lower bounds on $r_k(n)$.

What we are essentially doing is solving a system of recurrences of the form

$$f_i(n) = max\,(\,f_{a(i)}(n-1)+1\,,\,f_{b(i)}(n-1)\,) \quad ,$$

for N sequences $\{f_i(n)\}$, $i = 1 \dots N$. Here $a(i)$ $b(i)$ are some functions from $[1, N]$ to $[1, N]$. It may be worthwhile to study such recurrences *for their own sake*, abstractly, and come up with a study of the asymptotic density as they depend on $a(i), b(i)$. It is not hard to show that $f_i(n)$ can be modeled as

$$R(Qn+i) = Pn + c_i \quad ,$$

however, it is not necessarily true that $0 \leq c_i \leq P$. Regardless, hopefully we can get some general theorems, and since $a(i)$ and $b(i)$ are *arbitrary*, there is lots of elbow-room for induction.

Finally, we would check that the *particular* $a(i)$, $b(i)$ that show up satisfy some general conditions that would enable us to get upper bounds on $\alpha_{k,D}$ as a function of D.

References

[GGK] W. Gasarch, J. Glenn, C.P. Kruskal, *Finding large 3-free sets I: The small n case*, Journal of Computer and System Sciences **74** (2008), 628-655. `http://www.cs.loyola.edu/ jglenn/Papers/3apI.pdf`

[S] E. Szemerédi, *On sets of integers containing no k elements in arithmetic progression*, Acta Arith. **27**(1975), 199-245.