Green-Tao numbers and SAT

Oliver Kullmann
Computer Science Department
Swansea University

PCV Seminar

http://www.swansea.ac.uk/compsci/research/seminars/pcvseminars.php

Swansea, May 6, 2010

Introduction
Green-Tao numbers
Transversal extensions
The generic boolean translation

Avoiding progressions of size 3 in the primes

Partition $\{2, 3, 5, 7, 11, 13, 17, 19, 23\}$
into two parts
such that no part
contains an arithmetic progression of size 3:

$$2\ 3\ 5\ 7 \mid 11\ 13\ 17\ 19\ 23$$
$$2\ 3\ 5\ 11 \mid 7\ 13\ 17\ 19\ 23$$
$$2\ 3\ 5\ 11\ 13 \mid 7\ 17\ 19\ 23$$

$$\mathbf{grt_2(3, 3)} > 9$$
$$\mathrm{grt}_2(3, 3) = 23$$

The **Green-Tao Theorem** guarantees that
$\mathrm{grt}_m(k_1, \ldots, k_m) \in \mathbb{N}$ always exists.

Introduction

Green-Tao numbers

Transversal extensions

The generic boolean translation

The basic SAT translation

OKlibrary at http://www.ok-sat-library.org

```
(%i1) oklib_load_all();
(%i2) output_greentao2_stdname(3,9);
> more GreenTao_2-3_9.cnf
c Green-Tao problem (diagonal form), created by the OKlibrary:
c 2 parts, arithmetic progressions of size 3, and 9 prime numbers.
c Variables and associated prime numbers:
c 1 : 2
c 2 : 3
c 3 : 5
c 4 : 7
c 5 : 11
c 6 : 13
c 7 : 17
c 8 : 19
c 9 : 23
p cnf 9 14
2 3 4 0
2 4 5 0
3 5 7 0
2 5 8 0
4 6 8 0
2 6 9 0
5 7 9 0
-4 -3 -2 0
-5 -4 -2 0
-7 -5 -3 0
-8 -5 -2 0
-8 -6 -4 0
-9 -6 -2 0
-9 -7 -5 0
```

Overview

In this initial phase (see [5, 6]) of the investigations into

Ramsey theory and SAT

we

- computed “the” basic Green-Tao numbers, and
- determined the “best” available SAT methods.

We put special focus

1. on how to get more problems whose size is not already astronomic
2. how to translate these non-boolean problems into boolean problems.

General methods were found to approach both problems.

In the future we hope to establish Ramsey-type problems as an important class of benchmark problems for SAT solvers.

Outline

1. Introduction
2. Green-Tao numbers
3. Transversal extensions
4. The generic boolean translation

The general definition

By the Green-Tao theorem ([1]) the following definition is justified:

- For a parameter tuple $(k_1, \ldots, k_m)$
- let the **Green-Tao number $\mathbf{grt_m(k_1, \ldots, k_m)}$**
- be defined as the smallest $n_0 \in \mathbb{N}$
- such that for every $n \geq n_0$ and every $f : \{p_1, \ldots, p_n\} \to \{1, \ldots, m\}$,
- where $p_1, \ldots, p_n$ are the first n prime numbers,
- there exists some $i \in \{1, \ldots, m\}$
- such that $f^{-1}(i)$ contains an arithmetic progression of size k_i.

On the history of the Green-Tao theorem

- Arithmetic progressions in the primes have been investigated for more than 200 years.
- The existence of arbitrarily long arithmetic progressions in the primes is a special case of the famous "k-tuple conjecture" of Hardy and Littlewood in 1923 (still wide open).
- The final proof in 2004 by Ben Green and Terence Tao is based on *Additive Number Theory* and *Ramsey Theory*.

On additive number theory

- “Additive number theory” is the offspring of Ramsey Theory (especially van der Waerden’s theorem) into number theory.
- Landau (apparently Lev, the physicist(!), not Edmund, the mathematician) once said

  “Prime numbers are meant to be multiplied, not added.”
- Additive number theory investigates additive number-theoretical structures, and has only been enabled by its interaction with combinatorics.

On van der Waerden's theorem

1. Van der Waerden's theorem 1927 ([8]) is the basis, showing existence of **van der Waerden numbers** $\mathbf{vdw}_m(k_1, \ldots, k_m)$, defined as GT-numbers, but using $\{1, \ldots, n\}$ instead of $\{p_1, \ldots, p_n\}$.
2. A major strengthening has been conjectured by Erdös and Turán in 1936, and and was finally proved 1975 by Szemerédi in his landmark paper [7].
3. The original proof is combinatorial. A breakthrough was the new proof by Fürstenberg (1977), using his general methods from *ergodic theory*.
4. A further breakthrough was Gower's new proof around 1997, based on *harmonic analysis*.

We remark that there are numerous relations to logic, especially proof theory.

Putting it all together: Pseudo-randomness

Szemerédi's theorem in the infinite setting says:

> Any subset of the integers of *positive density*
> contains progressions of arbitrary length.

The set of prime numbers does not have positive density. Green-Tao show a *transference principle*, which allows to deduce from Szemerédi's theorem that

> Any set of positive *relative* density inside
> of a *sufficiently pseudorandom set*
> contains progressions of arbitrary length.

Then Green-Tao apply that

> A large fraction of the primes can be placed,
> with positive relative density, into a sufficiently
> pseudorandom set of "almost primes".

Simple parameter tuples

Already $\mathrm{grt}_1(k)$ for $k \in \mathbb{N}$ poses a non-trivial mathematical problem, namely

the question is to find the smallest $n \in \mathbb{N}$
such that the first n prime numbers
contain an arithmetic progression of length k.

- Trivially $\mathrm{grt}_1(1) = 1$ and $\mathrm{grt}_1(2) = 2$, while $\mathrm{grt}_1(3) = 4$ (confirmed by the progression $(3, 5, 7)$).
- The computation of $\mathrm{grt}_1(k)$ has nothing to do with SAT solving, so we can't contribute here, but the known values give a first feeling for the growth involved.

The known values of $\mathrm{grt}_1(k)$

k	$\mathrm{grt}_1(k)$
1	1
2	2
3	4
4	9
5	10
6	37
7	155
8	263
9	289
10	316
11	21,966
12	23,060
13	58,464
14	2,253,121
15	9,686,320
16	11,015,837
17	227,225,515
18	755,752,809
19	3,466,256,932
20	22,009,064,470
21	220,525,414,079

We see that for the exploration of non-simple parameter tuples only $k \leq 10$ is feasible.

Core parameter tuples

a \ b	3	4	5	6	7
3	23	79	528	≥ 2072	> 13800
4	-	512	> 4231		
5	-	-	≥ 34309		

a, b \ c	3	4	5
$3, 3$	137	≥ 434	> 1989
$3, 4$	-	> 1662	> 8200

a, b, c \ d	3	4
$3, 3, 3$	> 384	> 1052
$3, 3, 4$	-	> 2750

Regarding the solvers

1. All lower bounds are computed by local-search algorithm.
2. While in the 5 successful cases of computing (core) Green-Tao numbers, for determining unsatisfiability a DPLL-like (complete) SAT solver is used.
3. Already in these cases we see a good variety of algorithms/solvers being best:
   - `OKsolver-2002` (look-ahead) versus `minisat2` (conflict-driven) on the complete side,
   - `adaptnovelty+` for the binary lower bounds,
   - except of $(5, 5)$ where `survey propagation` succeeded,
   - and `rnovelty+` for the non-binary lower bounds.

We need to discuss what to do in the non-binary cases!

Regarding growth

1. Yet nothing has been published on the bounds, except of the cases of simple parameter tuples, that is upper bounds for $\mathrm{grt}_1(k)$: Proven an exponential tower of height 8, conjectured $\mathrm{grt}_1(k) \leq \pi(k! + 1)$.
2. If something can be proven for the general numbers, the bounds might be astronomical.
3. For van-der-Waerden numbers we conjecture that in each row the numbers grow only like $P(k)$, where P is a polynomial depending on the row.
4. Here now we conjecture that the growth in each row is of type $\exp(P(k))$.

We need more food (for the solvers)!

Cases with $k_i = 2$

We call a parameter-tuple **core**, if

- it has at least two entries
- and every entry is at least 3.

Parameter-values 2 seems not to have been systematically considered until now:

Tuples of the form $(2, \ldots, 2, k_1, \ldots, k_l)$ we call **transversal extensions** of $(k_1, \ldots, k_l)$.

Note that $k_i = 2$ means that only one prime number can get colour i, since every two numbers form an arithmetic progression of size 2.

So finding a solution for m initial 2's means, that we can discard at our will m prime numbers.

Linear growth

Keeping $(k_1, \ldots, k_l)$ and growing the prefix of 2's, we make the observation in the underlying [5, 6] (in a more general framework):

Theorem *If we have the Szemerédi property, then for every tuple* $(k_1, \ldots, k_l)$*, for every* $\varepsilon > 0$ *and for* m *big enough we have:*

$$N_{m+l}(2, \ldots, 2, k_1, \ldots, k_l) \leq (1 + \varepsilon) \cdot m.$$

Since the Green-Tao theorem actually proves the Szemerédi property, we get

$$\mathrm{grt}_{m+l}(2, \ldots, 2, k_1, \ldots, k_l) \leq (1 + \varepsilon) \cdot m.$$

Transversal extension numbers

The numbers $\mathrm{grt}_{m+2}((2,\ldots,2);(3,k))$:

k \ m	0	1	2	3	4	5	6
4	79	117	120	128	136	≥ 142	≥ 151
5	528	581	≥ 582	> 606			

The numbers $\mathrm{grt}_{m+2}((2,\ldots,2);(3,3))$:

m	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
	23	31	39	41	47	53	55	≥ 60	≥ 62	≥ 67	≥ 71	≥ 73	≥ 82	≥ 83	≥ 86

The numbers $\mathrm{grt}_{m+2}((2,\ldots,2);(4,k))$:

k \ m	0	1	2
4	512	≥ 553	> 588

The numbers $\mathrm{grt}_{m+3}((2,\ldots,2);(3,3,k))$:

k \ m	0	1	2
3	137	151	≥ 154
4	≥ 434	≥ 453	> 470

Remarks

- Now the conflict-driven solvers become superior.
- Regarding the lower bounds, the `novelty`-family is no longer dominant, but other solvers often perform best (`saps`, `sapsnr`, `rsaps`, `walksat`, `walksat-tabu`).
- More (good) lower bounds are definitely possible.

Non-boolean clause-sets

The "true" generalisation of boolean CNF to non-boolean CNF seems to be the following:

1. variables v have (finite) domains D_v
2. literals are of the form "$v \neq \varepsilon$" for some $\varepsilon \in D_v$;
3. these clauses are called "no-goods" in constraint solving.

For a systematic investigation see [2, 3, 4].
With these non-boolean clause-sets also the non-boolean Green-Tao problems, for tuples $(k_1, \ldots, k_m)$, now have a canonical representation, using m values.

A general scheme for a boolean translation

Consider a variable v with domain $D_v = \{\varepsilon_1, \ldots, \varepsilon_m\}$.

- So there are m literals, namely $(v, \varepsilon_1), \ldots, (v, \varepsilon_m)$.
- And for assignment $\langle v \longrightarrow \varepsilon_i \rangle$ exactly $m - 1$ of these literals become true, while (v, ε_i) becomes false.
- It wouldn't matter regarding satisfiability if it would be possible to set more than one of them to false.

The idea now is to represent these literals (v, ε_i) by clauses C_i from a clause-set F_v.

- We need to choose m clauses $C_1, \ldots, C_m \in F_v$.
- Since we must not be able to make all literals to true, F_v must be unsatisfiable.
- We demand all clauses C_i to be *necessary* for F_v, that is, removal renders F_v satisfiable — in this way we model that all other literals become true.

That's it!

We obtain a translation $F \rightsquigarrow T(F)$ for a non-boolean clause-set F as follows:

- Choose variable-disjoint such F_v for each variable — except of variable-disjointness the choices are completely independent.
- Literals (v, ε_i) are replaced by the clauses C_i.
- The "remainder clauses" in $R_v := F_v \setminus \{C_1, \ldots, C_m\}$ are all added to the translation.

Note that

$$n(T(F)) = \sum_{v \in \mathrm{var}(F)} n(F_v)$$

$$c(T(F)) = c(F) + \sum_{v \in \mathrm{var}(F)} c(R_v).$$

Example: The direct translations

Here we choose

$$F_v = \{ \{v_1\}, \ldots, \{v_m\}, \{\overline{v_1}, \ldots, \overline{v_m}\} \},$$

and we choose the unit-clauses to correspond to the values.

1. For the *weak form* (using only ALO-clauses) that's it (so we have one remainder clause).
2. For the *strong form* we add all positive binary clauses to F_v (so obtaining the AMO-clauses).

Example: The simple logarithmic translation

- If $m = 2^p$, then choose the (minimally) unsatisfiable clause-set F_v with p variables and 2^p clauses (which are all the full clauses using all variables).
- So here are no remainder clauses.
- If m is not a power of two, then for the simple case just use the smallest p with $m < 2^p$, use the same F_v, and choose m of these clauses (the remaining clauses become remainder clauses).

The weak nested translation

Here we use $p := m - 1$ (boolean) variables $v_1, \ldots, v_p$ and

$$F_v = \left\{ \{v_1\}, \{\overline{v_1}, v_2\}, \ldots, \{\overline{v_1}, \ldots, \overline{v_{p-1}}, v_p\}, \{\overline{v_1}, \ldots, \overline{v_p}\} \right\}.$$

There are no remainder clauses.
Yet we tested these (and other, related) translations only on the Green-Tao instances, but this we did rather extensively.

Big surprise:

> For "large" m the logarithmic translation was best, and for all other m the weak nested translation — for all solver types.

"Best" means by orders of magnitudes.

Summary

I. This concludes for us the initial phase.

II. Several other forms of problems from Ramsey theory have also been considered.

III. We see a rich diverse behaviour, where every solver can be best on some class for some parameters.

IV. Now the task is to understand what's going on!

V. Especially the unsatisfiable cases need to be improved.

VI. So amongst others we will investigate tree-resolution and full-resolution complexity in detail, experimentally as well as theoretically.

VII. As an aside, the resolution proofs found for the Green-Tao instances seem to have surprising (number-theoretical) regularities.

References

Ben Green and Terence Tao.

The primes contain arbitrarily long arithmetic progressions.
Annals of Mathematics, 167(2):481–547, 2008.

Oliver Kullmann.

Constraint satisfaction problems in clausal form: Autarkies and minimal unsatisfiability.
Technical Report TR 07-055, version 02, Electronic Colloquium on Computational Complexity (ECCC), January 2009.

Oliver Kullmann.

Constraint satisfaction problems in clausal form I: Autarkies and deficiency.
Fundamenta Informaticae, 2010. To appear.

Oliver Kullmann.

Constraint satisfaction problems in clausal form II: Minimal unsatisfiability and conflict structure.
Fundamenta Informaticae, 2010. To appear.

Oliver Kullmann.

Exact Ramsey theory: Green-Tao numbers and SAT.
Technical Report arXiv:1004.0653v2 [cs.DM], arXiv, April 2010.

Oliver Kullmann.

Green-Tao numbers and SAT.
In Ofer Strichman and Stefan Szeider, editors, *Theory and Applications of Satisfiability Testing - SAT 2010*, volume 6175 of *Lecture Notes in Computer Science*. Springer, 2010.

E. Szemerédi.

On sets of integers containing no k elements in arithmetic progression.
Acta Arithmetica, 27:299–345, 1975.

B.L. van der Waerden.

Beweis einer Baudetschen Vermutung.
Nieuw Archief voor Wiskunde, 15:212–216, 1927.

End