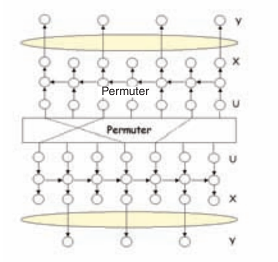
American Mathematical Society

Current Events in Mathematics

An AMS Special Session at the Joint Mathematics Meeting Organized by AMS President *David Eisenbud*

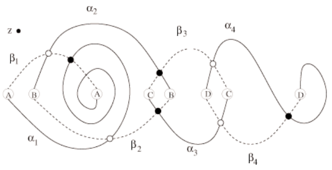
The Green-Tao Theorem on Primes in Arithmetic Progression: A Dynamical Point of View Bryna Kra

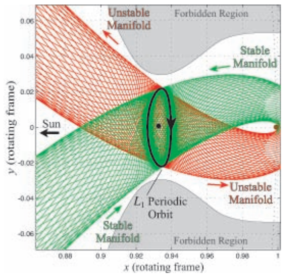
2 3 5 7 11 13 17 19 23 29 31 37 41 43 47 53 59 61 67 71 73 79 83
89 97 101 103 107 109 113 127 131 137 139 149 151 157 163 167
173 179 181 191 193 197 **199** 211 223 227 229 233 239 241 251
257 265 269 271 277 281 283 293 307 311 313 317 331 337 347
349 353 359 367 373 379 383 389 397 401 **409** 419 421 431 433
439 443 449 457 461 463 467 479 487 491 499 503 509 521 523
541 547 557 563 569 571 577 587 593 599 601 607 613 617 **619**
631 641 643 647 653 659 661 673 677 683 691 701 700 719 727
733 739 743 751 757 761 769 773 787 797 809 811 821 823 827
829 839 853 857 859 863 877 881 883 887 907 911 919 929 937
941 947 953 967 971 977 983 991 997 1009 1013 1019 1021 1031
1033 **1039** 10491051 1061 1063 1069 1087 1091 1093 1097 1103
1109 1117 1123 1129 1151 1153 1163 1171 1181 1187 1193 1201
1213 1217 1223 1229 1231 1237 **1249** 1259 1277 1279 1283 1289



Achieving the Shannon Limit: A Progress Report Robert McEliece

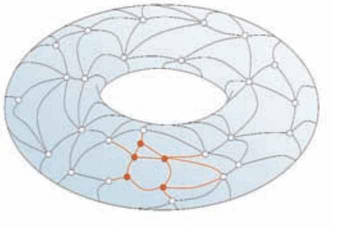
Floer Theory and Low Dimensional Topology Dusa McDuff





New Methods in Celestial Mechanics and Mission Design Jerrold Marsden and Shane Ross

Graph Minors and the Proof of Wagner's Conjecture László Lovász



THE GREEN-TAO THEOREM ON ARITHMETIC PROGRESSIONS IN THE PRIMES: A DYNAMICAL POINT OF VIEW

BRYNA KRA

Abstract. A long standing and almost folkloric conjecture is that the primes contain arbitrarily long arithmetic progressions. Until recently, the only progress on this conjecture was due to van der Corput, who showed in 1939 that there are infinitely many triples of primes in arithmetic progression. In an amazing fusion of methods from analytic number theory and ergodic theory, Ben Green and Terence Tao showed that for any integer k greater than or equal to 3, there exist infinitely many arithmetic progressions of length k consisting only of prime numbers. This is an introduction to some of the ideas in the proof, concentrating on those drawn from ergodic theory.

1. Background

For hundreds of years, mathematicians have made conjectures about patterns in the primes: one of the simplest to state is that the primes contain arbitrarily long arithmetic progressions. It is not clear exactly when this conjecture was first formalized, but as early as 1770 Lagrange and Waring studied the problem of how large the common difference of an arithmetic progression of k primes must be. A natural extension of this question is to ask if the prime numbers contain arbitrarily long arithmetic progressions. Support for a positive answer to this comes from the following simple heuristic. The Prime Number Theorem states that the number of prime numbers less than the integer N is asymptotically $N/\log N$. It follows that the density of primes around a positive large $x \in \mathbb{R}$ is about $1/\log x$. Thus if we model the sequence of primes by choosing integers at random, with the integer n being chosen with probability $1/\log n$, then there ought to be approximately $N^2/\log^k N$ progressions of length k in the prime numbers less than N.

In 1923, Hardy and Littlewood [HL] made a very general conjecture (the *k-tuple conjecture*) about patterns and their distribution in the primes. This conjecture includes as a special case that the number of k-term arithmetic progressions in the primes bounded by N is asymptotically $C_k N^2/\log^k N$ for a certain explicit value of C_k. There are numerous related conjectures about the existence of arithmetic progressions in certain subsets of the integers. For example, the famous conjecture of Erdös and Turán [ET] states that if $A = \{a_1 < a_2 < \ldots\}$ is an infinite sequence of integers with $\sum_i 1/a_i = \infty$, then A contains arbitrarily long arithmetic progressions. In particular, this would imply that the primes contain arbitrarily long arithmetic progressions.

The first major progress on arithmetic progressions in the primes was made by van der Corput [Vdc], who proved in 1939 that the primes contain infinitely many arithmetic progressions of length 3. The next progress was not until 1981, when Heath-Brown [H] showed that there are infinitely many arithmetic progressions of length 4 consisting of three primes and an almost prime, meaning either a prime or a product of two primes. In a slightly different direction are the elegant results of Balog ([Ba1], [Ba2]) on patterns in the primes. For example, he shows that for any positive integer k, there exist infinitely many k-tuples of distinct primes $p_1 < p_2 < \ldots < p_k$ such that $(p_i + p_j)/2$ is prime for all $i, j \in \{1, \ldots, k\}$. For $k = 2$ this implies, in particular, that the primes contain infinitely many arithmetic progressions of length 3.

Computational mathematicians have also given the problem of finding long arithmetic progressions in the primes attention. In 1995, Moran, Pritchard and Thyssen [MPT] found a progression of length 22 in the primes and this record held for almost 10 years. In 2004, Frind, Jobling and Underwood [FJU] found a progression of length 23, starting with the prime 56211383760397 and with common difference 44546738095860.

In 2004, Ben Green and Terence Tao announced a major breakthrough, with a proof of the general case:

Theorem 1.1 (Green and Tao [GT1]). *For every integer $k \geq 1$, the prime numbers contain an arithmetic progression of length k.*

We note that they [GT2] also extract a bound on how far out in the primes one must go in order to guarantee finding an arithmetic progression of length k, showing that there is a k-term arithmetic progression of primes all of whose entries are bounded by

$$2^{2^{2^{2^{2^{2^{2^{2^{2^{(100k)}}}}}}}}} .$$

This bound is considered far from optimal; standard heuristics in number theory, plus a little calculation, leads to the conjecture that there is an arithmetic progression of length k in the primes all of whose entries are bounded by $k! + 7$.

Green and Tao actually obtain the stronger statement that it suffices to have positive density relative to the primes:

Theorem 1.2 (Green and Tao [GT1])**.** *If A is a subset of prime numbers with*

$$\limsup_{N\to\infty} \frac{1}{\pi(N)} |A \cap [1, \dots, N]| > 0 \ ,$$

where $\pi(N)$ is the number of primes less than N, then for every integer $k \geq 1$, A contains an arithmetic progression of length k.

For $k = 3$, this was proved by Green [G].

The theorem of Green and Tao is a beautiful result answering an old conjecture that has attracted much work. Perhaps even more impressive is the fusion of methods and results from number theory, ergodic theory, harmonic analysis, discrete geometry, and combinatorics used in its proof. The starting point for Green and Tao's proof is the celebrated theorem of Szemerédi [S]: a set of integers with positive upper density[1] contains arbitrarily long arithmetic progressions. One of the main ideas is to generalize this, showing that a dense subset of a sufficiently *pseudorandom* collection (see Section 7 for the precise definition) of the integers contains arbitrarily long arithmetic progressions. There are three major ingredients in the proof. The first is Szemerédi's Theorem itself. Since the primes do not have positive upper density, Szemerédi's Theorem can not be directly applied and the second major ingredient in Green and Tao's proof is a certain transference principle that allows one to use Szemerédi's Theorem in a more general setting. The third ingredient is an application of recent work of Goldston and Yildirim [GY] on the distribution of the prime numbers, showing that this generalized Szemerédi Theorem applies to the primes.

It is impossible to give a complete proof of their theorem in this limited space, nor even to do justice to the main ideas. Our goal is to outline the main ingredients and focus on the relation between their work and recent parallel advances in ergodic theory. The interaction between combinatorial number theory and ergodic theory began with Furstenberg's proof of Szemerédi's Theorem (see Section 3) and has led to many new results. Until the present, this interaction has mainly taken the form of using ergodic theory to prove statements in combinatorial number theory, such as Szemerédi's Theorem, its generalizations (including a multidimensional version [FK1] and a polynomial

[1] The upper density $d^*(A)$ of a subset A of the integers is defined to be

$$d^*(A) := \limsup_{N\to\infty} |A \cap [1, \dots, N]|/N \ .$$

version [BL]), or the density Hales-Jewett Theorem [FK2]. Green and Tao's work opens a new chapter in this interaction, with ergodic theoretic *proof* techniques being adapted for use in a number theoretic setting.

Acknowledgments: I appreciate the many helpful comments I received while preparing this manuscript, including those from Nikos Frantzikinakis, Bernard Host, Terence Tao, and Mate Wierdl, and especially those from Andrew Granville.

2. Szemerédi's Theorem

Substituting the set of all integers for the set of primes in Theorem 1.2, one obtains Szemerédi's Theorem. We state an equivalent finite version of this theorem:

Theorem 2.1 (**Finite Szemerédi** [S]). *Let* $0 < \delta \leq 1$ *be a real number and let* $k \geq 1$ *be an integer. Then there exists* $N_0(\delta, k)$ *such that if* $N > N_0(\delta, k)$ *and* $A \subset [1, \ldots, N]$ *with* $|A| \geq \delta N$*, then* A *contains an arithmetic progression of length* k*.*

It is clear that this version implies the first version of Szemerédi's Theorem, and an easy argument gives the converse implication.

Szemerédi's [S] original proof in 1975 was combinatorial in nature. Shortly thereafter, Furstenberg developed the surprising relation between combinatorics and ergodic theory, proving Szemerédi's Theorem via a multiple recurrence theorem (see Section 3). More recently, Gowers [Go] gave a new proof of Szemerédi's Theorem using harmonic analysis, vastly improving the known bounds in the finite version. Although the various proofs (Szemerédi's, Furstenberg's, or Gowers') seem to use very different methods, they all have several features in common. In each, a key idea is the dichotomy in the underlying space (whether it be a subset of the integers, a measure space, or the finite group $\mathbb{Z}/N\mathbb{Z}$) between randomness and structure. One then has to analyze the structured part of the space to understand the intersection of a set with itself along arithmetic progressions. We start by further exploring the connection with ergodic theory.

3. Szemerédi's Theorem and ergodic theory

Furstenberg proved the multiple ergodic theorem:

Theorem 3.1 (**Multiple Recurrence** [F]). *Let* $(X, \mathcal{X}, \mu, T)$ *be a measure preserving probability system*[2] *and let* $k \geq 1$ *be an integer. Then*

[2] By a *measure preserving probability system*, we mean a quadruple $(X, \mathcal{X}, \mu, T)$, where X is a set, $\mathcal{X}$ denotes a σ-algebra on X, μ is a probability measure on $(X, \mathcal{X})$

for any set $E \in \mathcal{X}$ with $\mu(E) > 0$,

$$\liminf_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \mu\left(E \cap T^{-n}E \cap T^{-2n}E \cap \ldots \cap T^{-(k-1)n}E\right) > 0. \tag{3.1}$$

An obvious corollary to Theorem 3.1 is that under the same hypotheses, there exists an integer $n \geq 1$ such that

$$\mu\left(E \cap T^{-n}E \cap T^{-2n}E \cap \ldots \cap T^{-(k-1)n}E\right) > 0 \ .$$

Furstenberg then made the beautiful connection to combinatorics, showing that regularity properties of integers with positive upper density correspond to multiple recurrence results:

Theorem 3.2 (Correspondence Principle [F]**).** *Assume that A is a subset of integers with positive upper density. Then there exist a measure preserving probability system $(X, \mathcal{X}, \mu, T)$ and a measurable set $E \in \mathcal{X}$ with $\mu(E) = d^*(A)$ such that for all integers $k \geq 1$ and all integers $m_0, \ldots, m_{k-1} \geq 1$,*

$$d^*\big(A \cap (A+m_0) \cap \ldots \cap (A+m_{k-1})\big) \geq \mu\big(E \cap T^{-m_0}E \cap \ldots \cap T^{-m_{k-1}}E\big) \ .$$

Taking $m_1 = n, m_2 = 2n, \ldots, m_{k_1} = (k-1)n$, Szemerédi's Theorem follows from the corollary to the Multiple Recurrence Theorem.

Furstenberg's proof relies on a compactness argument, making it difficult to extract any explicit bounds in the finite version of Szemerédi's Theorem. On the other hand, Theorem 3.1 and its proof gave rise to a new area in ergodic theory, called "Ergodic Ramsey Theory", leading to many other results in combinatorics, such as the multidimensional Szemerédi Theorem [FK1] and the polynomial Szemerédi Theorem [BL]. Some of these generalizations are still not attainable by other methods. More recent developments in ergodic Ramsey Theory closely parallel ideas in Green and Tao's work; we return to this in Section 5.

To prove Theorem 3.1, Furstenberg shows that in any measure preserving system, one of two distinct phenomena occurs to make the measure of this intersection positive. The first is weak mixing,[3] when

and $T\colon X \to X$ is a measurable map such that $\mu(A) = \mu(T^{-1}A)$ for all $A \in \mathcal{X}$. Usually, we assume that X is a metrizable compact set and $\mathcal{X}$ is its *Borel σ-algebra* (the σ-algebra generated by the open sets). We always denote the σ-algebra by the calligraphic version of the letter used for the space and when there is no ambiguity, we omit explicit mention of the σ-algebra and instead write (X, μ, T).

[3] The system $(X, \mathcal{X}, \mu, T)$ is *weak mixing* if for all $A, B \in \mathcal{X}$,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \left|\mu\left(T^{-n}A \cap B\right) - \mu(A)\mu(B)\right| = 0 \ .$$

for any set E, $\mu(E \cap T^{-n}E)$ is approximately $\mu(E)^2$ for most choices of the integer n. Then it can be shown that $\mu(E \cap T^{-n}E \cap T^{-2n}E \cap \ldots \cap T^{-(k-1)n}E)$ is approximately $\mu(E)^k$ for most choices of n, which is clearly positive for a set E of positive measure. The opposite situation is rigidity, when for appropriately chosen n, T^n is very close to the identity. Then $T^{jn}E$ is very close to E and $\mu(E \cap T^{-n}E \cap T^{-2n}E \cap \ldots \cap T^{-(k-1)n}E)$ is very close to $\mu(E)$, again giving a positive intersection for a set E of positive measure. One then has to show that the average along arithmetic progressions for any function can be decomposed into a piece which has randomness of the first case (a generalization of weak mixing) and one which has the structure of the second case (a generalization of rigidity). One of the difficulties lies in proving a structure theorem for the latter situation, showing that this portion of the system can be reduced to a finite series of compact extensions of a one point system (a *Furstenberg tower*) and then proving a recurrence statement for this tower.

4. Gowers norms in combinatorics

In his proof of Szemerédi's Theorem, Gowers [Go] shows that averages along arithmetic progressions are controlled by certain norms, now known as *Gowers (uniformity) norms.* We start with a description of this key idea, explaining it in the combinatorial setup in this section and in the ergodic version in the next section. To define these norms, we introduce some notation.

For a positive integer N, let $\mathbb{Z}_N := \mathbb{Z}/N\mathbb{Z}$. If $f : \mathbb{Z}_N \to \mathbb{C}$ is a function, let $\mathbb{E}\left(f(x)|x \in \mathbb{Z}_N\right)$ denote the average value of f on $\mathbb{Z}_N$:

$$\mathbb{E}\left(f(x)|x \in \mathbb{Z}_N\right) = \frac{1}{N}\sum_{x\in\mathbb{Z}_N} f(x) \ .$$

We also use a higher dimensional version of the expectation. For example, by $\mathbb{E}(f(x,y)|x,y \in \mathbb{Z}_N)$, we mean iteration of the one variable expectation:

$$\mathbb{E}\big(\mathbb{E}(f(x,y)|x \in \mathbb{Z}_N)|y \in \mathbb{Z}_N\big) \ .$$

In this terminology, Szemerédi's Theorem becomes:

Theorem 4.1 (Reformulated Szemerédi). *Let $0 < \delta \leq 1$ be a real number and let $k \geq 2$ be an integer. If N is sufficiently large and $f : \mathbb{Z}_N \to \mathbb{R}$ is a function with $0 \leq f(x) \leq 1$ for all $x \in \mathbb{Z}_N$ and $\mathbb{E}\left(f(x)|x \in \mathbb{Z}_N\right) \geq \delta$, then*

$$\mathbb{E}\big(f(x)f(x+r)\ldots f(x+(k-1)r)|x,r \in \mathbb{Z}_N\big) \geq c(k,\delta) \qquad (4.1)$$

for some constant $c(k, \delta) > 0$ which does not depend either on f or on N.

At first glance, this appears to be a stronger version than the original statement of Szemerédi's Theorem, showing not only the existence of a single arithmetic progression but of some positive multiple of N^2 progressions. However, using some combinatorial trickery one can quickly show that the two versions are equivalent.

We now define the norms that control the averages along arithmetic progressions. A variation on the classic van der Corput difference theorem motivates the definition and since we are studying progressions in the context of $\mathbb{Z}_N$, it is particularly easy to state:

Lemma 4.2 (van der Corput Lemma for $\mathbb{Z}_N$). *If $f : \mathbb{Z}_N \to \mathbb{C}$ is a function, then*

$$|\mathbb{E}(f(x)|x \in \mathbb{Z}_N)|^2 = \mathbb{E}(\overline{f(x)}f(x+h)|x, h \in \mathbb{Z}_N) \ .$$

Since each of these expectations is a finite sum, the proof of the lemma is immediate by expanding both sides and using a simple change of variable.

The d^{th}*-Gowers (uniformity) norm* $\|f\|_{U^d}$ of a function $f\colon \mathbb{Z}_N \to \mathbb{C}$ is defined inductively. Set

$$\|f\|_{U^1} := |\mathbb{E}(f(x)|x \in \mathbb{Z}_N)|$$

For $d \geq 2$, we mimic successive uses of the van der Corput Lemma and define

$$\|f\|_{U^d} := \mathbb{E}\left(\|\overline{f}f_h\|_{U^{d-1}}^{2^{d-1}} |h \in \mathbb{Z}_N\right)^{1/2^d} \ , \tag{4.2}$$

where $f_h(x) = f(x+h)$. By definition, $\|f\|_{U^d}$ is non-negative for $d = 1$ and therefore also for all higher d. Furthermore, Equation (4.2) shows that the d^{th}-Gowers norm is shift invariant, meaning that $\|f(x)\|_{U^d} = \|f(x+h)\|_{U^d}$ for any $h \in \mathbb{Z}_N$.

It follows immediately from the definitions and a change of variable that

$$\|f\|_{U^1} = \left(\mathbb{E}(f(x)\overline{f(x+h)}|x, h \in \mathbb{Z}_N)\right)^{1/2} \ . \tag{4.3}$$

Thus for $d=1$, $\|f\|_{U^d}$ is only a seminorm.[4] Using the Fourier expansion of f and computing, we have that

$$\|f\|_{U^2} = \left(\sum_{\xi\in\mathbb{Z}_N} \left|\hat{f}(\xi)\right|^4\right)^{1/4},$$

where $\hat{f}$ denotes the Fourier transform of f. It follows that for $d=2$, $\|f\|_{U^d}$ is nondegenerate and so it is a norm.

To see that $\|f\|_{U^d}$ is a norm for $d\geq 2$, we give an equivalent characterization of the d^{th}-Gowers norm as a certain average over a d-dimensional cube. This also allows us to express the definition of the norm in a closed form. We first need to introduce some more notation.

We consider $\{0,1\}^d$ as the set of vertices of the d-dimensional Euclidean cube, meaning it consists of points $\omega = (\omega_1,\ldots,\omega_d)$ with $\omega_j\in\{0,1\}$ for $j=1,\ldots,d$. For $\omega\in\{0,1\}^d$, define $|\omega| = \omega_1+\ldots+\omega_d$ and if $\omega\in\{0,1\}^d$ and $\mathbf{h}=(h_1,\ldots,h_d)\in\mathbb{Z}_N^d$, we define $\omega\cdot\mathbf{h} := \omega_1h_1+\ldots+\omega_dh_d$. Then if $f\colon\mathbb{Z}_N\to\mathbb{C}$ is a complex valued function, it follows from inductively applying the definition in (4.2) that

$$\|f\|_{U^d} := \mathbb{E}\left(\prod_{\omega\in\{0,1\}^d} C^{|\omega|}f(x+\omega\cdot\mathbf{h})|x\in\mathbb{Z}_N,\mathbf{h}\in\mathbb{Z}_N^d\right)^{1/2^d}, \tag{4.4}$$

where C is the conjugation operator $Cf(x) := \overline{f(x)}$. This presentation allows one to view the Gowers norms as an average over the cube $\{0,1\}^d$.

By repeated applications of the Cauchy-Schwarz Inequality and the definitions of the norms, one obtains the *Gowers Cauchy-Schwarz Inequality* for 2^d functions $f_\omega\colon\mathbb{Z}_N\to\mathbb{C}$:

$$\left|\mathbb{E}\left(\prod_{\omega\in\{0,1\}^d} C^{|\omega|}f_\omega(x+\omega\cdot\mathbf{h})|x\in\mathbb{Z}_N,\mathbf{h}\in\mathbb{Z}_N^d\right)\right| \leq \prod_{\omega\in\{0,1\}^d}\|f_\omega\|_{U^d}. \tag{4.5}$$

From this, one can show that $\|f\|_{U^d}$ is subadditive and so is a seminorm. Furthermore, using the Gowers Cauchy-Schwarz Inequality, one has the chain of inequalities

$$\|f\|_{U^1}\leq\|f\|_{U^2}\leq\ldots\leq\|f\|_{L^\infty}. \tag{4.6}$$

[4] A *seminorm* on a vector space V is a non-negative real valued function such that $\|f+g\|\leq\|f\|+\|g\|$ and $\|cf\|=|c|\cdot\|f\|$ for all $f,g\in V$ and all scalars c. Thus unlike a norm, one may have $\|f\|=0$ for some $f\neq 0$.

Since $\|f\|_{U^d}$ is nondegenerate for $d = 2$, Inequality (4.6) implies that $\|f\|_{U^d}$ is nondegenerate for all higher d, giving that the d^{th}- Gowers norm is actually a norm for $d \geq 2$.

Finally we rewrite the Gowers norms in notation that is closer in spirit to the ergodic theoretic setup. Consider $\mathbb{Z}_N$ endowed with the transformation $T(x) = x + 1 \mod N$ and the uniform measure m assigning weight $1/N$ to each element of $\mathbb{Z}_N$. Then the definition of Equation (4.4) becomes:

(4.7)

$$\|f\|_{U^d} = \left(\int \prod_{\omega \in \{0,1\}^d} C^{|\omega|} f(T^{\omega \cdot \mathbf{h}} x)\, dm(x) dm(h_1) \dots dm(h_d) \right)^{1/2^d} .$$

These norms are used by Gowers (as well as by Host and Kra [HK1] and more recently by Tao [T] and by Green and Tao [GT1]) to control the average along arithmetic progressions, which is the quantity in Equation (4.1). This can be viewed as a generalized version of the von Neumann Ergodic Theorem, which states that the average of a bounded function on a finite measure space converges in mean to its integral. We formalize the statement about this control, as stated by Tao [T]:

Theorem 4.3 (**Generalized von Neumann Theorem** [T]). *Let $k \geq 2$ be an integer, let N be a prime number and assume that $f_0, \dots, f_{k-1} : \mathbb{Z}_N \to \mathbb{C}$ are functions with $|f_0|, \dots, |f_{k-1}| \leq 1$. Then*

$$\left|\mathbb{E}\big(f_0(x) f_1(x+n) \dots f_{k-1}(x + (k-1)n) | x, n \in \mathbb{Z}_N\big)\right| \leq \min_{0 \leq j \leq k-1} \|f_j\|_{U^{k-1}} .$$

The proof of this theorem is an induction argument, using the Cauchy-Schwarz Inequality and the van der Corput Lemma for $\mathbb{Z}_N$ (Lemma 4.2).

To prove Szemerédi's Theorem, Gowers [Go] studies the indicator function $\mathbf{1}_A$ of a set $A \subset \mathbb{Z}_N$. As in Furstenberg's proof, there are two opposite cases. If $\|\mathbf{1}_A - |A|/N\|_{U^{k-1}}$ is small, then one can substitute a constant function for $\mathbf{1}_A$ and get this average is large. If $\|\mathbf{1}_A - |A|/N\|_{U^{k-1}}$ is large, then he shows that the restriction of $\mathbf{1}_A$ to some large subset of $\mathbb{Z}_N$ has many arithmetic properties and so the average in Equation (4.1) is once again large. The difficulty in this proof lies in showing that a usable version of the dichotomy between large and small always occurs. Proving that the second case has the needed structure is easier, since here the structure is a nested sequence of arithmetic progressions.

5. Gowers norms in ergodic theory

Furstenberg's proof of Theorem 3.1 left open the question of the existence of the limit in the left hand side of Equation (3.1). Host and Kra [HK1] show that this lim inf is actually a limit:

Theorem 5.1 (**Multiple Convergence** [HK1]). *Assume that $(X, \mathcal{X}, \mu, T)$ is a measure preserving probability system, $k \geq 1$ is an integer, and $f_1, f_2, \ldots, f_k$ are bounded functions on X. Then the averages*

$$\frac{1}{N}\sum_{n=0}^{N-1} f_1(T^n x) f_2(T^{2n} x) \ldots f_k(T^{kn} x) \tag{5.1}$$

converge in $L^2(\mu)$ as $N \to \infty$.

The first step in proving this theorem is showing that instead of taking the average in the system $(X, \mathcal{X}, \mu, T)$, it suffices to consider the average over some (ostensibly simpler) system $(Y, \mathcal{Y}, \nu, S)$. This amounts to proving a generalized von Neumann Theorem, as in Gowers' proof. This idea is implicit in Furstenberg's [F] proof of Szemerédi's Theorem and made explicit in the proof of Theorem 5.1.

In [HK1], we introduced seminorms that generalize the Gowers norms. We consider a general probability measure preserving space $(X, \mathcal{X}, \mu)$ with an invertible measurable, measure preserving transformation $T : X \to X$ on it. For a function $f \in L^\infty(\mu)$, we define (compare with Equation (4.2))

$$\|f\|_{U^1} := \left|\int f(x)\, d\mu(x)\right|$$

and inductively we define the d^{th}-seminorm by

$$\|f\|_{U^d}^{2^d} := \lim_{N\to\infty} \frac{1}{N}\sum_{n=0}^{N-1} \|f\overline{T^n f}\|_{U^{d-1}}^{2^{d-1}} \ . \tag{5.2}$$

To recover the Gowers norms, we take the space $\mathbb{Z}_N$ with the transformation $x \mapsto x+1 \mod N$ and the uniform measure assigning each element of $\mathbb{Z}_N$ weight $1/N$.

Once again, there is an alternate presentation, analogous to that of Equation (4.7), as the integral with respect to a certain measure and this second presentation makes many properties of $\|f\|_{U^d}$ more transparent. We need some notation to define this measure. (A reader not interested in the technical definition of this measure can omit this alternate presentation.)

Assume that $(X, \mathcal{X}, \mu)$ is a probability space. If $f \in L^1(\mu)$ and $\mathcal{Y} \subset \mathcal{X}$ is a sub-σ-algebra, then the *conditional expectation* $\mathbb{E}(f|\mathcal{Y})$ of

f on $\mathcal{Y}$ is the $\mathcal{Y}$-measurable function such that

$$\int_A f\,d\mu = \int_A \mathbb{E}(f|\mathcal{Y})\,d\mu$$

for all $A \in \mathcal{Y}$.

Fix an ergodic[5] measure preserving probability system $(X, \mathcal{X}, \mu, T)$. Define $X^{[d]} = X^{2^d}$ and write points of $X^{[d]}$ as $\mathbf{x} = (x_\omega : \omega \in \{0,1\}^d)$. Let $T^{[d]} = T \times T \times \ldots \times T$ taken 2^d times. There is a natural identification between $X^{[d+1]}$ and $X^{[d]} \times X^{[d]}$, with a point $\mathbf{x} \in X^{[d+1]}$ being identified with $(\mathbf{x}', \mathbf{x}'') \in X^{[d]} \times X^{[d]}$, where $x'_\omega = x_{\omega,0}$ and $x''_\omega = x_{\omega,1}$ for each $\omega \in \{0,1\}^d$.

For each integer $d \geq 0$, we inductively define a $T^{[d]}$-invariant measure $\mu^{[d]}$ on $X^{[d]}$. Define $\mu^{[0]} := \mu$. Assume that $\mu^{[d]}$ is defined for some $d \geq 0$. Let $\mathcal{I}^{[d]}$ denote the $T^{[d]}$-invariant σ-algebra of $(X^{[d]}, \mu^{[d]}, T^{[d]})$. Using the natural identification of $X^{[d+1]}$ with $X^{[d]} \times X^{[d]}$, define the measure preserving (probability) system $(X^{[d+1]}, \mu^{[d+1]}, T^{[d+1]})$ to be the *relatively independent joining* of $(X^{[d]}, \mu^{[d]}, T^{[d]})$ with itself over $\mathcal{I}^{[d]}$: this means that the measure $\mu^{[d+1]}$ is the measure such that for all bounded functions F' and F'' on $X^{[d]}$, we have

$$\int_{X^{[d+1]}} F'(\mathbf{x}')F''(\mathbf{x}'')\,d\mu^{[d+1]}(\mathbf{x}) = \int_{X^{[d]}} \mathbb{E}(F'|\mathcal{I}^{[d]})\mathbb{E}(F''|\mathcal{I}^{[d]})\,d\mu^{[d]}\ .$$

The measure $\mu^{[d+1]}$ is invariant under $T^{[d+1]}$ and the two natural projections on $X^{[d]}$ are each $\mu^{[d]}$. Using induction, this gives that each of the 2^d natural projections of $\mu^{[d]}$ on X is equal to μ. Thus for a bounded function f on X, the integral

$$\int_{X^{[d]}} \prod_{\omega \in \{0,1\}^d} C^{|\omega|} f(x_\omega)\,d\mu^{[d]}(\mathbf{x})$$

is real and nonnegative, where as before $Cf(x) := \overline{f(x)}$. An alternate definition of the seminorms is:

$$\|f\|_{U^d} = \left(\int_{X^{[d]}} \prod_{\omega \in \{0,1\}^d} C^{|\omega|} f(x_\omega)\,d\mu^{[d]}(\mathbf{x})\right)^{1/2^d} . \tag{5.3}$$

To show that these are seminorms, one proceeds in the same manner as in the combinatorial setup, deriving a version of the Cauchy-Schwarz

[5] The system $(X, \mathcal{X}, \mu, T)$ is *ergodic* if the only sets $A \in \mathcal{X}$ with $\mu(T^{-1}A) = \mu(A)$ have measure 0 or 1. Every system can be decomposed as an integral of ergodic systems and so we can assume that the system being studied in Theorem 5.1 is ergodic.

Inequality (analogous to Equation (4.5)) and using it to show subadditivity. Positivity follows immediately from definition (5.2). From the definition of these measures and the Ergodic Theorem, we obtain that this second definition is equivalent to the first definition given in Equation (5.2).

The definition of Equation (5.3) can once again be viewed as an average over the cube $\{0,1\}^d$. A convergence theorem for general averages along cubes is also proved in [HK1].

The first step in proving Theorem 5.1 is showing that the averages along arithmetic progressions are once again controlled by the d-seminorms, meaning an analog of Theorem 4.3:

Theorem 5.2 (Generalized von Neumann, revisited [HK1]**).** *Assume that $(X, \mathcal{X}, \mu, T)$ is a measure preserving probability system. Let $k \geq 2$ be an integer and assume that $f_1, \ldots, f_k$ are bounded functions on X with $\|f_1\|_\infty, \ldots \|f_k\| \leq 1$. Then*

$$\limsup_{N\to\infty} \left\| \frac{1}{N} \sum_{n=0}^{N-1} f_1(T^n x) f_2(T^{2n} x) \ldots f_k(T^{kn} x) \right\|_2 \leq \min_{1\leq j\leq k} \left(j \|f_j\|_{U^k} \right) .$$

The added factor of j which appears on the right hand side of this bound and not in Theorem 4.3 is due to the change in underlying space. In Theorem 4.3, we assumed that N is prime; in this case, for any integer j that is not a multiple of N, the map $n \mapsto jn$ is onto in $\mathbb{Z}_N$, which is not the case in $\mathbb{Z}$. As for the earlier Generalized von Neumann Theorem, Theorem 5.2 is proved using induction, the Cauchy-Schwarz Inequality and a van der Corput lemma. This time we need a Hilbert space variation of this lemma:

Lemma 5.3 (van der Corput Lemma, revisited [B]**).** *Assume that $\mathcal{H}$ is a Hilbert space with inner product $\langle\ ,\ \rangle$ and that ξ_n, $n \geq 0$ is a sequence in $\mathcal{H}$ with $\|\xi_n\| \leq 1$ for all n. Then*

$$\limsup_{N\to\infty} \left\| \frac{1}{N} \sum_{n=0}^{N-1} \xi_n \right\|^2 \leq \limsup_{H\to\infty} \frac{1}{H} \sum_{h=0}^{H-1} \limsup_{N\to\infty} \left| \frac{1}{N} \sum_{n=0}^{N-1} \langle \xi_{n+h}, \xi_n \rangle \right| .$$

Theorem 5.2 allows one to consider an average along arithmetic progressions on an appropriate factor, rather than the whole space. We make this notion more precise.

For a measure preserving system $(X, \mathcal{X}, \mu, T)$, the word *factor* is used with two different but equivalent meanings. First, it is a T-invariant σ-algebra of $\mathcal{X}$. (Strictly speaking, this is a sub-σ-algebra, but throughout we omit the use of the word "sub".) Secondly, if $(Y, \mathcal{Y}, \nu, S)$ is a measure preserving system, a map $\pi : X \to Y$ is a *factor map* if π maps

μ to ν and $S \circ \pi = \pi \circ T$. Then Y is said to be a factor of X and the two definitions coincide up to the identification of $\mathcal{Y}$ with $\pi^{-1}(\mathcal{Y})$. For $f \in L^1(\mu)$, we view $\mathbb{E}(f|\mathcal{Y})$ as a function on $\mathcal{X}$ and let $\mathbb{E}(f|Y)$ denote the function on Y defined by $\mathbb{E}(f|Y) \circ \pi = \mathbb{E}(f|\mathcal{Y})$. It is characterized by

$$\int_Y \mathbb{E}(f|Y)(y) \cdot g(y)\, d\nu(y) = \int_X f(x) \cdot g(\pi(x))\, d\mu(x)$$

for all $g \in L^\infty(\mu)$.

The seminorms are used to define factors of the system $(X, \mathcal{X}, \mu, T)$. One presentation of these factors is by defining their orthogonal complements: for $d \geq 1$, define $\mathcal{Z}_{d-1}$ to be the σ-algebra of $\mathcal{X}$ such that for $f \in L^\infty(\mu)$:

$$\|f\|_{U^d} = 0 \text{ if and only if } \ \mathbb{E}(f|\mathcal{Z}_{d-1}) = 0 \ .$$

Thus a bounded function f is measurable with respect to $\mathcal{Z}_{d-1}$ if and only if $\int fgd\mu = 0$ for all functions $g \in L^\infty(\mu)$ with $\|g\|_{U^{d-1}} = 0$. This motivates an equivalent definition of the factors $\mathcal{Z}_d$ with respect to a dual norm. Namely, defining the dual norm $\|f\|_{(U^d)^*}$ by

$$\|f\|_{(U^d)^*} := \sup_{g \in L^\infty(\mu)} \left\{ \int_X fg\, d\mu \colon \|g\|_{U^d} \leq 1 \right\} \ , \tag{5.4}$$

we have that

$$\|f\|_{(U^d)^*} = 0 \text{ if and only if } \mathbb{E}(f|\mathcal{Z}_d) = 0 \ .$$

Letting Z_j denote the factor associated to the σ-algebra $\mathcal{Z}_j$, we have that Z_0 is the trivial factor and Z_1 is the *Kronecker* factor, meaning the σ-algebra which is spanned by the eigenfunctions of T. Furthermore, the sequence of factors is increasing (compare with Equation (4.6)):

$$Z_0 \leftarrow Z_1 \leftarrow Z_2 \leftarrow \ldots \leftarrow X$$

and if T is weakly mixing, then Z_d is the trivial factor for all d.

Theorem 5.2 says that the factor $\mathcal{Z}_{d-1}$ is a *characteristic factor* for the average of Equation (5.1), meaning that the limit behavior of the averages in $L^2(\mu)$ remains unchanged when each function is replaced by its conditional expectation on this factor. Thus it suffices to prove convergence when one of the factors Z_d is substituted for the original system. For a progression of length k, this amounts to decomposing a bounded function $f = g + h$ with $g = f - \mathbb{E}(f|\mathcal{Z}_{k-1})$. The function g is the uniform component and has zero $k-1$ seminorm and so contributes zero to the average along arithmetic progressions. The second term h is the anti-uniform component and belongs to the algebra of functions measurable with respect to the factor $\mathcal{Z}_{k-1}$ and must be analyzed via a structure theorem for the characteristic factors. This decomposition of an arbitrary bounded function into uniform and anti-uniform components is unique. In the combinatorial setting, a similar decomposition (see Section 6) can only be carried out approximately. Ergodic theory is more precise than combinatorics in describing the second component of this decomposition.

When the description of the factor given by the structure theorem is "simple", one has a better chance of proving convergence in this factor. For the given decomposition, the description of the characteristic factor is as an inverse limit of *nilsystems*, meaning that it can be approximated arbitrarily well by a rotation on a homogeneous space of a nilpotent Lie group.[6]

6. Quantitative ergodic theory

Tao [T] gave a new proof of Szemerédi's Theorem, along the lines of Furstenberg's original proof, but proving it in the finite system $\mathbb{Z}_N$ rather than for an arbitrary measure space. This allows him to extract explicit bounds for $N_0(\delta, k)$ in the finite version (Theorem 2.1), although the bounds are nowhere near as good as those obtained by Gowers [Go].

Once again, a generalized von Neumann Theorem (Theorems 4.3 and 5.2) is used to start the proof. Then, as in the ergodic setup, an arbitrary bounded function f on $\mathbb{Z}_N$ is decomposed into pieces, each of which can be analyzed. This time the decomposition is into a term with small Gowers norm and a structured component, with the wrinkle that one also has to deal with a small error term. The first term corresponds to a uniform component $f-\mathbb{E}(f|\mathcal{Z})$ for a well chosen σ-algebra $\mathcal{Z}$ (similar to the use of a characteristic factor in the ergodic setup) which has small Gowers norm and makes a small contribution to the average in Equation (4.1). Since the space being used is $\mathbb{Z}_N$, the σ-algebra $\mathcal{Z}$ is really a finite partition of $\mathbb{Z}_N$. The second term is the conditional expectation of f relative to $\mathcal{Z}$ and this component is analyzed using a form of recurrence similar to that needed for a Furstenberg tower.

[6] If G is a k-step nilpotent Lie group and Γ is a discrete cocompact subgroup, then $a \in G$ naturally acts on G/Γ by left translation by $T_a(x\Gamma) = (ax)\Gamma$. The Haar measure μ is the unique Borel probability measure on G/Γ that is invariant under this action of G by left translations. For a fixed element $a \in G$, the system $(G/\Gamma, \mathcal{G}/\Gamma, T_a, \mu)$ is a k-step nilsystem. The structure theorem in [HK1] states that the characteristic factor $\mathcal{Z}_{k-1}$ for an average of arithmetic progressions of length k is an inverse limit of such $(k-1)$-step nilsystems.

The second component of the decomposition, called the *anti-uniform* functions by Tao, is essentially dual to the uniform component where the anti-uniform (dual) norm $\|g\|_{(U^{d-1})^*}$ is defined by (compare with Equation (5.4))

$$\|g\|_{(U^{d-1})^*} := \sup_{f\colon \mathbb{Z}_N \to \mathbb{C}} \{|\langle f, g\rangle| : \|f\|_{U^{d-1}} \leq 1\} .$$

The contribution of this term to the average is bounded from below by van der Waerden's Theorem,[7] with the idea being that these functions lie in a sufficiently compact space so that a finite coloring argument can be used. Applying this idea to a function with positive expectation, the average along arithmetic progressions is positive.

This proof follows Furstenberg's proof closely. One advantage is the elimination of the compactness argument, leading to explicit bounds on the size of the set needed to guarantee the existence of a progression of length k. The structure theorem does not need an understanding of the precise structure of the chosen σ-algebra, which corresponds to the tower of compact extensions used by Furstenberg (or to the nilsystems in Host and Kra). However, a more precise understanding of this structure should clarify the apparent link between the anti-uniform functions of level k appearing in Tao's proof and the k-step nilsystems used to prove Theorem 5.1.

7. Arithmetic progressions in the primes

Green and Tao continue in this vein to prove the existence of arithmetic progressions in the primes. The starting point is clear: study the averages of Equation (4.1) for the indicator function of the primes. Roughly speaking, for a large integer N and real number $0 < \delta \leq 1$, they consider the function

$$f(n) = \begin{cases} \log n & \text{if } n \text{ is prime} \\ 0 & \text{otherwise} \end{cases} ,$$

normalized such that its average on $\{0, 1, \ldots, N-1\}$ is δ. Szemerédi's Theorem can not be applied because the function f is not bounded independently of N. They begin by studying the closely related *von*

[7] Van der Waerden's Theorem [Vdw] states that if the integers are partitioned into finitely many pieces, then one of these pieces contains arbitrarily long arithmetic progressions. This theorem motivated Erdös and Turán [ET] to conjecture Szemerédi's Theorem.

Mangoldt function $\Lambda(n)$, where

$$\Lambda(n) = \begin{cases} \log p & \text{if } n = p^m \text{ for some } m \in \mathbb{N} \\ 0 & \text{otherwise} \ , \end{cases}$$

and make use of the fact that this function is more natural analytically than f. Although the von Mangoldt function is supported on the primes and their powers, the powers are sparsely enough distributed so that they only contribute a small error term in the calculations. This function has had many uses in number theory; for example, the unique factorization theorem is equivalent to the statement

$$\log n = \sum_{d|n} \Lambda(d) \text{ for all positive integers } n \ ,$$

and the Prime Number Theorem is equivalent to the statement

$$\frac{1}{N} \sum_{1 \leq n \leq N} \Lambda(n) = 1 + o(1) \ .$$

(Throughout, by $o(1)$, we mean a quantity that tends to 0 as $N \to \infty$, and when this quantity depends on other constants, we include them as subscripts on o.)

The function Λ mostly avoids giving weight to arithmetic progressions $a \mod q$ when a and q are not relatively prime. Such arithmetic progressions are more dense when q has small prime factors. This means that the small primes make a disproportionate contribution to Λ, making it too irregularly distributed for their purposes. Therefore Green and Tao are forced to modify Λ, quotienting out the small primes, so that it becomes *pseudorandom*. The precise definition and modification is given below, but the idea is that the values of a pseudorandom function should be distributed so that using any statistic to measure them, one gets approximately the same measurement as that arising from a random set of the same density.

The goal then becomes to extend Szemerédi's Theorem, showing that not only does a dense subset of the integers contain arbitrarily long arithmetic progressions, but a dense subset of a pseudorandom collection of integers also contains arbitrarily long arithmetic progressions. Green and Tao do this by "transferring" Szemerédi's Theorem to a more general setting: the hypothesis in Theorem 4.1 that $f\colon \mathbb{Z}_N \to \mathbb{R}$ satisfies $0 \leq f(x) \leq 1$ is replaced by f being bounded by a more general function $\nu\colon \mathbb{Z}_N \to \mathbb{R}+$ with certain useful properties.

The function $\nu\colon \mathbb{Z}_N \to \mathbb{R}^+$ is assumed to be a *measure*,[8] meaning that $\mathbb{E}(\nu(x)|x \in \mathbb{Z}_N) = 1 + o(1)$, and ν is also assumed to be pseudorandom. They show:

Theorem 7.1 (**Transference Theorem** [GT1])**.** *Let $0 < \delta \leq 1$ be a real number and let $k \geq 3$ be an integer. If N is sufficiently large, $\nu\colon \mathbb{Z}_N \to \mathbb{R}^+$ is a k-pseudorandom measure, and $f\colon \mathbb{Z}_N \to \mathbb{R}$ is function with $0 \leq f(x) \leq \nu(x)$ for all $x \in \mathbb{Z}_N$ and $\mathbb{E}(f(x)|x \in \mathbb{Z}_N) \geq \delta$, then*

$$\mathbb{E}\big(f(x)f(x+r)\dots f(x+(k-1)r)|x, r \in \mathbb{Z}_N\big) \geq c(k,\delta) - o_{k,\delta}(1) \ , \tag{7.1}$$

where the constant $c(k,\delta)$ is the same as that in Theorem 4.1.

Other than the bounds on f, the only additional modification caused by bounding f by a pseudorandom measure instead of the constant function 1 is the introduction of the error term $o_{k,\delta}(1)$, which tends to 0 as $N \to \infty$. The dependence of this error is only on k and δ.

Before giving an indication of the proof of Theorem 7.1, we make the notion of a pseudorandom measure more precise. (A reader not interested in the technical details can skip the next few paragraphs.) The measure $\nu : \mathbb{Z}_N \to \mathbb{R}^+$ is said to be *k-pseudorandom* if ν satisfies a *k-linear forms condition* and a *k-correlation condition.*

To define the linear forms condition, fix k, the length of the arithmetic progression and assume that N is prime and larger than k. Assume that we have m linear forms ψ_i, $1 \leq i \leq m$, with $m \leq k \cdot 2^{k-1}$ and t variables with $t \leq 3k - 4$. (The exact values of these constants are not important for the proof; the importance lies in showing that a particular choice of pseudorandom function satisfies these conditions. For this, it only matters that the values depend on nothing but k.) Let $L = (L_{ij})$ be an $m \times t$ matrix, whose entries are rational numbers with numerator and denominator bounded in absolute value by k. By choice of N and k, we can view the entries of L as elements of $\mathbb{Z}_N$ (recall that N is prime). Assume further that each of the t columns of L are not identically zero and that the columns are pairwise independent. Let $\psi_i(\mathbf{x}) = b_i + \sum_{j=1}^{t} L_{ij}x_j$ denote the m linear forms, where $\mathbf{x} \in \mathbb{Z}_N^t$ and $b_i \in \mathbb{Z}_N$ for $1 \leq i \leq m$. The measure $\nu : \mathbb{Z}_N \to \mathbb{R}^+$ is said to satisfy the (m, t, L)-linear forms condition if

$$\mathbb{E}\big(\nu\left(\psi_1(\mathbf{x})\right)\dots\nu(\psi_m(\mathbf{x}))\,|\mathbf{x} \in \mathbb{Z}_N^t\big) = 1 + o_{m,t,L}(1) \ ,$$

where the dependence on N is assumed to be uniform in the choice of the b_i. The case $m = 1$ with $\psi(x) = x$ corresponds to the measure ν

[8] As noted by Green and Tao, the name measure is a misnomer, and ν should more accurately be called a probability density.

with $\mathbb{E}(\nu) = 1 + o(1)$, and this is the bound used in the Reformulated Szemerédi Theorem (Theorem 4.1). For higher m, the values of the measure ν evaluated on linear forms up to a certain complexity are, on average, independent. If there were no restriction on the complexity, the measure would be close to the ergodic theoretic notion of weak mixing, meaning its values along any distinct linear forms would be, on average, independent.

We now define the correlation condition. The measure $\nu : \mathbb{Z}_N \to \mathbb{R}^+$ is said to satisfy a 2^{k-1}-correlation condition if for each m with $1 \leq m \leq 2^{k-1}$, there exists a weight function $\tau = \tau_m : \mathbb{Z}_N \to \mathbb{R}^+$ with

$$\mathbb{E}(\tau^q | z \in \mathbb{Z}_N) \leq C(m, q)$$

for a constant $C(m, q)$, for all $1 \leq q < \infty$ and that

$$\mathbb{E}\big(\nu(x + h_1)\nu(x + h_2) \dots \nu(x + h_m) | x \in \mathbb{Z}_N\big) \leq \sum_{1 \leq i < j \leq m} \tau(h_i - h_j)$$

for all $h_1, h_2, \dots, h_m \in \mathbb{Z}_N$.

The correlation condition arises in Goldston and Yildirim's [GY] work and is used for specific estimates applied to the prime numbers. Although the linear forms condition does not arise in their work, their estimates apply to ν satisfying this condition. Most of Green and Tao's computations only need the linear forms condition; the correlation condition is used only in one place where the estimates are highly technical.

In some sense, the Transference Theorem can be thought of as a generalization of Furstenberg's Multiple Recurrence Theorem. In the ergodic set up, a natural choice of measure is the uniform one, assigning each integer in $\{1, \dots, N\}$ the equal weight $1/N$. This measure is invariant with respect to the shift map $x \mapsto x + 1 \mod N$. In Green and Tao's generalization, the measure behaves in a pseudorandom manner with respect to the shift. For a certain choice of R (discussed below), to each number in $\{1, \dots, N\}$ having no prime factors less than R, the new measure assigns the weight $\log R/N$ and in order to make the measure more regular, it assigns a small value to each of the other numbers in $\{1, \dots, N\}$.

Lending credence to the idea that Szemerédi's Theorem should hold for a function bounded by a pseudorandom measure is the fact that a pseudorandom measure is close to the constant function 1 in Gowers norm:

Lemma 7.2 ([GT1]). *Fix an integer $k \geq 1$, let $N > k$ be a prime number, and assume that $\nu : \mathbb{Z}_N \to \mathbb{R}^+$ is a k-pseudorandom measure.*

Then

$$\|\nu - 1\|_{U^d} = o(1)$$

for all $1 \leq d \leq k-1$.

The broad outline of the proof of Theorem 7.1 is similar to that of Tao's proof of Szemerédi's Theorem sketched in the last section, but both the technical details and the combination of ideas from seemingly unrelated areas of mathematics make it a significantly more ambitious undertaking. The innovation is the reduction of Theorem 7.1 to Szemerédi's Theorem. The key argument, again, is a structure theorem, but this time not only is there an error term in the decomposition, but the decomposition is only valid on most of the space. Green and Tao show:

Theorem 7.3 (**Decomposition Theorem** [GT1]). *Let* $k \geq 2$ *be an integer, let* $0 < \epsilon \ll 1$ *be a small parameter, and let* $N = N(\epsilon)$ *be sufficiently large. Assume that* $\nu : \mathbb{Z}_N \to \mathbb{R}^+$ *is a k-pseudorandom measure and that* $f \in L^1(\mathbb{Z}_N)$ *is a function satisfying* $0 \leq f(x) \leq \nu(x)$ *for all* $x \in \mathbb{Z}_N$. *Then there exists a* σ*-algebra* $\mathcal{Z}$ *and an exceptional set* $\Omega \in \mathcal{Z}$ *with* $\mathbb{E}(\nu(x)\mathbf{1}_\Omega(x)|x \in \mathbb{Z}_N) = o_\epsilon(1)$ *such that*

$$\|\mathbf{1}_{\Omega^C}\mathbb{E}(\nu - 1|\mathcal{Z})\|_{L^\infty} = o_\epsilon(1)$$

and

$$\|\mathbf{1}_{\Omega^C}(f - \mathbb{E}(f|\mathcal{Z}))\|_{U^{k-1}} \leq \epsilon^{1/2^k} \ ,$$

where Ω^C *denotes the complement of* Ω.

This means that outside a small subset Ω of $\mathbb{Z}_N$, a function f that is bounded by a pseudorandom measure can be decomposed into a sum of a uniform function g and an anti-uniform function h, plus a small error term. The function g has small Gowers norm and corresponds to $f - \mathbb{E}(f|\mathcal{Z})$ in the ergodic theoretic setup, while the non-negative function h is bounded and corresponds to $\mathbb{E}(f|\mathcal{Z})$. Other than the error terms, this parallels the ergodic theoretic decomposition associated to a characteristic factor described in Section 5 and the decomposition used by Tao described in Section 6.

The next ingredient in the proof of Theorem 7.1 is a way to control the contribution of the Gowers uniform portion in the decomposition, analogous to the generalized von Neumann Theorem. Once again, the bound on the functions changes: instead of being bounded by the constant 1, the functions are now bounded pointwise by 1 plus a pseudorandom measure.

Theorem 7.4 (**Pseudorandom Generalized von Neumann Theorem** [GT1]). *Let* $k \geq 2$ *be an integer, let* N *be a prime number, and* *assume that $\nu : \mathbb{Z}_N \to \mathbb{R}^+$ is a k-pseudorandom measure. Assume that $f_0, \dots, f_{k-1} \in L^1(\mathbb{Z}_N)$ are functions such that*

$$|f_j(x)| \leq \nu(x) + 1 \text{ for all } x \in \mathbb{Z}_N, 0 \leq j \leq k-1 \ .$$

Then

$$\begin{aligned}\bigl|\mathbb{E}\bigl(f_0(x)f_1(x+n)\dots f_{k-1}(x+(k-1)n)|x, n \in \mathbb{Z}_N\bigr)\bigr| \\ = O\left(\inf_{0\leq j\leq k-1} \|f_j\|_{U^{k-1}}\right) + o(1) \ .\end{aligned}$$

We are now ready to outline the proof of Theorem 7.1, still glossing over many technical details. We fix a function f that is bounded by a pseudorandom measure and that has positive expectation on $\mathbb{Z}_N$. Using the Decomposition Theorem, the expectation on the left hand side of Equation (7.1) (which is the average along arithmetic progressions) is larger than the same expectation with $\mathbf{1}_{\Omega^C} f$ substituted for f, where Ω is some small set. Ignoring the error term, we now use the decomposition of this new function into $g+h$, where g is the Gowers uniform portion and h is some bounded function. Much like the idea of a characteristic factor in ergodic theory, we now want to discard the Gowers uniform portion g and replace our function by h. The expectation on the left hand side of Equation (7.1) can be expanded as a sum of 2^k expectations, by making the substitution of $g+h$ for $\mathbf{1}_{\Omega^C} f$. All terms but one contain an occurrence of g in it and each of these terms is small by the Pseudorandom Generalized von Neumann Theorem. We are left only with a single term making a large contribution to the expectation, and this term only contains occurrences of the function h. The good news is that now this function h is bounded and so the usual Szemerédi Theorem applies. Furthermore, f and h have approximately the same expectation, and in particular the expectation on $\mathbb{Z}_N$ of h is also positive. Thus by Szemerédi's theorem, the expectation in Equation (7.1) with f replaced by h is positive. Therefore, the same result holds for f.

Lastly we give an indication of the choice of the function f and measure ν needed to use the Transference Theorem for the primes. The function is a variation on the von Mangoldt function, cut off at a certain point, in order to make a function that is (vaguely speaking) supported on primes of magnitude $\log N$. Unfortunately, it does not suffice to simply use a multiple of this function for ν, since as we noted earlier, the primes, and therefore any multiple of the von Mangoldt function, are not uniformly distributed across all residue classes, whereas a pseudorandom function is. Instead, the measure ν is taken to have its support (again, vaguely speaking) on numbers n such that all the prime factors of $n-1$ are greater than some integer R. One can view this measure as approximately $\log R$ times the characteristic function of such numbers.

The ancient scholar Eratosthenes came up with a simple algorithm for listing all the prime numbers up to a given N, referred to as the sieve of Eratosthenes. Given a list of the numbers between 1 and N, starting with 2, erase all nontrivial multiples of 2 up to N. Call the remaining set P_2. Returning to the beginning, take the first number greater than 2 and erase all of its nontrivial multiples up to N. In general, the *level R almost primes* $P_R(N)$ are defined to be the set of all numbers between 1 and N that contain no nontrivial factors less than or equal to R. Thus if $R = \sqrt{N}$, we have that $P_{\sqrt{N}}(N)$ consists exactly of the prime numbers up to N. Mertens [Me] proved that the size $|P_R(N)|$ is approximately $cN/\log R$ for some positive constant c. Combining this with the estimate from the Prime Number Theorem that the number of primes up to N is approximately $N/\log N$, we have that the density of primes in the almost primes $P_R(N)$ is about a multiple of $\log R/\log N$. Therefore if R is a small power of N, then the primes have positive density in the level R almost primes. This motivates the function and measure Green and Tao use. For completeness, we give the technical definitions.

Let W denote the product of the primes up to $4(k+1)!^2$ and let $R = N^{k^{-1}2^{-k-4}}$. The *truncated von Mangoldt function* is defined to be

$$\Lambda_R(n) = \sum_{d|n, d\leq R} \mu(d)\log(R/d) \ ,$$

where μ is the Möbius function.[9] This is a cut off version of the von Mangoldt function, since if $R > n$ then $\Lambda_R(n) = \Lambda(n)$. This approximation to $\Lambda(n)$ has had wide use in analytic number theory, most recently in the work of Goldston and Yildirim [GY].

Finally we need to define the measure ν that majorizes Λ_R and whose values are more uniformly distributed. The measure $\nu : \mathbb{Z}_N \to \mathbb{R}^+$ is defined for $0 \leq n < N$ to be

$$\nu(n) = \begin{cases} \frac{\phi(W)}{W}\frac{\Lambda_R(Wn+1)^2}{\log R} & \text{for } N/(2^k(k+4)!) \leq n \leq 2N/(2^k(k+4)!) \\ 1 & \text{otherwise ,} \end{cases}$$

[9] The *Möbius function* $\mu(n)$ is defined by $\mu(n) = 0$ if n is not a square free integer and $\mu(n) = (-1)^r$ if n is a square free integer and has r prime factors.

where $\{0, 1, \ldots, N-1\}$ is naturally identified with $\mathbb{Z}_N$ and ϕ denotes the Euler totient function.[10] The function Λ_R is evaluated at $Wn + 1$ to make it well distributed. (This quotienting out by small primes is referred to as the *W-trick.*) The primes bounded by x are not uniformly spread out in arithmetic progressions. (For example, there is only one prime congruent to $0 \mod 2$, while there are approximately $x/\log x$ congruent to $1 \mod 2$.) Furthermore, if a and q are relatively prime integers, the number of primes in the arithmetic progression $a \mod q$ up to x is approximately $\frac{x}{\log x} \cdot \frac{1}{\phi(q)}$. If one considers integers n with $n \equiv a \mod q$ and for which $Wn + 1$ is prime, then there are none only when q and $Wa + 1$ are not relatively prime and this can only happen when q and W are relatively prime. This means that q has no small prime factors and the values of $Wn + 1$ are more uniformly distributed among the arithmetic progressions.

The last major step is verifying that this choice of ν is k-pseudorandom. This relies on techniques from analytic number theory, using and extending recent results of Goldston and Yildirim [GY] on finding small gaps between primes.

8. Further directions

At this time, Green and Tao's Theorem seems out of the reach of ergodic theory. All theorems of combinatorial number theory that have been proved using ergodic theory rely in some way or another on the Correspondence Principle, which only applies to sets of integers with positive upper density. However, the many similarities between Green and Tao's approach and proofs in ergodic theory make it clear that the exact connection has yet to be understood. Perhaps a first step in further understanding this connection would be to translate the notion of a pseudorandom function (or sequence) to ergodic theory, somehow replacing the given definition on the finite space $\mathbb{Z}_N$ with a definition on some infinite space. The ultimate goal would be to use translations of the proof techniques of Green and Tao to obtain new convergence results in ergodic theory.

Although Tao's proof [T] of Szemerédi's Theorem removes the compactness argument needed in Furstenberg's proof, it still requires a lengthy induction to replace compactness. This induction only needs finitely many steps, but the number of steps is not explicitly known. A better understanding of the structure theorem used would probably

[10] The *Euler totient function* $\phi(n)$ is defined to be the number of positive integers less than or equal to n that are relatively prime to n, with 1 being counted as relatively prime to all numbers.

improve the bounds that Tao extracts with this method. It seems that finding the exact link between the anti-uniform functions of level k and the k-step nilsystems introduced in the work of Host and Kra [HK1] would clarify the connections between the two fields and probably lead to new and interesting developments.

A natural question arises from these considerations. Bergelson and Leibman [BL] used ergodic theory to establish a polynomial Szemerédi type theorem[11] and perhaps it is possible to carry out a similar program to that of Green and Tao for this situation. Namely, transfer the polynomial Szemerédi Theorem subsets of the integers with positive upper density contain polynomial patterns and show that dense subsets of pseudorandom sets also contain polynomial patterns. This would prove, for example, that there exist infinitely many triples (p, k, n) of positive integers such that $p, p+n, p+n^2, \dots, p+n^k$ consists only of prime numbers.

References

[Ba1] A. Balog. The prime k-tuplets conjecture on average. *Analytic number theory (Allerton Parl, IL., 1989)*, 47–75, *Progr. Math.*, **85**, Birkhäuser Boston, 1990.

[Ba2] A. Balog. Linear equations in primes. *Mathematika*, **39** (1992), 367–378.

[B] V. Bergelson. Weakly mixing PET. *Erg. Th. & Dyn. Sys.*, **7** (1987), 337–349.

[BL] V. Bergelson and A. Leibman. Polynomial extensions of van der Waerden's and Szemerédi's theorems. *J. Amer. Math. Soc.*, **9** (1996), 725–753.

[ET] P. Erdös and P. Turán. On some sequences of integers. *J. Lond. Math. Soc.*, **11** (1936), 261-264.

[FJU] M. Frind, P. Jobling and P. Underwood. 23 primes in arithmetic progression. Available at http://primes.plentyoffish.com/

[11] More precisely, Bergelson and Leibman's Theorem [BL] generalizes Furstenberg's Multiple Ergodic Theorem (Theorem 3.1). They show that if $(X, \mathcal{X}, \mu, T)$ is an invertible measure preserving probability system, $k \geq 1$ is an integer, $p_1(n), \dots, p_k(n)$ are polynomials taking integer values on the integers with $p_1(0) = \dots = p_k(0) = 0$, and $A \in \mathcal{X}$ with $\mu(A) > 0$, then

$$\liminf_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \mu\big(T^{p_1(n)}A \cap \ldots \cap T^{p_k(n)}A\big) > 0 \ .$$

As for arithmetic progressions, the limits of the related averages

$$\frac{1}{N} \sum_{n=0}^{N-1} f_1(T^{p_1(n)}x) \ldots f_k(T^{p_k(n)}x)$$

for bounded functions $f_1, \dots, f_k$ are known to exist in $L^2(\mu)$ (see [HK2] and [L]).

[F] H. Furstenberg. Ergodic behavior of diagonal measures and a theorem of Szemerédi on arithmetic progressions. *J. d'Analyse Math.*, **31** (1977), 204–256.

[FK1] H. Furstenberg and Y. Katznelson. An ergodic Szemerédi theorem for commuting transformations. *J. d'Analyse Math.*, **34** (1979), 275–291.

[FK2] H. Furstenberg and Y. Katznelson. A density version of the Hales-Jewett Theorem. *J. d'Analyse Math.*, **57** (1991), 64–119.

[Go] T. Gowers. A new proof of Szemerédi's Theorem. *GAFA*, **11** (2001), 465–588.

[G] B. Green. Roth's Theorem in the primes. To appear, *Ann.Math.*

[GT1] B. Green and T. Tao. The primes contain arbitrarily long arithmetic progressions. Preprint.

[GT2] B. Green and T. Tao. A bound for progressions of length k in the primes. Preprint.

[GY] D. Goldston and C. Y. Yildirim. Small gaps between primes, I. Preprint.

[HL] G. H. Hardy and J. E. Littlewood. Some problems of "partitio numerorum" III: on the expression of a number as a sum of primes. *Acta Math.*, **44** (1923), 1–70.

[H] D. R. Heath-Brown. Three primes and an almost prime in arithmetic progression. *J. Lond. Math. Soc. (2)*, **23** (1981), 396–414.

[HK1] B. Host and B. Kra. Nonconventional ergodic averages and nilmanifolds. To appear, *Ann. Math.*

[HK2] B. Host and B. Kra. Convergence of polynomial ergodic averages. To appear, *Isr. J. Math.*

[L] A. Leibman. Convergence of multiple ergodic averages along polynomials of several variables. To appear, *Isr. J. Math.*

[Me] F. Mertens. Ein Beitrag zur analytischen Zahlentheorie. *Journal für Math.*, **78** (1874), 46–62.

[MPT] A. Moran, P. Pritchard and A. Thyssen. Twenty-two primes in arithmetic progression. *Math. Comp.*, **64** (1995), 1337–1339.

[S] E. Szemerédi. On sets of integers containing no k elements in arithmetic progression. *Acta Arith.*, **27** (1975), 299-345.

[T] T. Tao. A quantitative ergodic theory proof of Szemerédi's theorem. Preprint.

[Vdc] J. G. van der Corput. Über Summen von Primzahlen und Primzahlquadraten. *Math. Ann.*, **116** (1939), 1–50.

[Vdw] B. L. van der Waerden. Beweis einer Baudetschen Vermutung. *Nieuw Arch. Wisk.*, **15** (1927), 212–216.

Department of Mathematics, Northwestern University, 2033 Sheridan Road, Evanston, IL 60208-2730

E-mail address: kra@math.northwestern.edu

Draft of 8 Dec 2004 12:05 p.m.

Approaching the Shannon Limit: A Progress Report*

R. J. McEliece

1. Introduction.

In 1948 Claude Shannon [26] published an historic monograph entitled *A Mathematical Theory of Communication* which contained a series of 23 theorems that now form the cornerstone of modern telecommunications and data storage systems. Shannon's style of proof was informal and intuitive, which briefly provoked controversy, but every single one of Shannon's assertions has stood the test of time. This article is about Shannon's Theorem 11, which is the intellectual zenith of *A Mathematical Theory of Communication*. It deals with the problem of communicating reliably over unreliable channels, using a communication paradigm of the general type depicted in Fig. 1.

Theorem 11 (Shannon, 1948). *For [almost]* any channel, there exists a positive number C, the channel capacity, such that for any desired data rate $R < C$, and any desired decoded bit error probability $p > 0$, there exists an encoder-decooder pair that permits transmission of data over the channel at rate R and decoded error probability $< p$. For $R > C$, arbiitrarly small p is not attainable.*

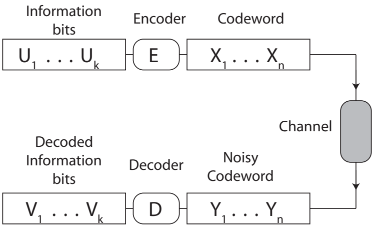


Figure 1 . A General Communication system. Here the rate is $R = \frac{k}{n}$ bits per channel input, and the decoded error probability is $P_b = \frac{1}{k}\sum_{i=1}^{k} \Pr\{V_i \neq U_i\}$. .

* This research was supported by NSF, NASA, Qualcomm, Sony, and the Lee Center for Advanced Networking

* Shannon proved his theorem for a relatively restricted class of channel models. Later researchers have extended the proof to cover a vast array of channels.

Shannon's Theorem 11 and a bit more is illustrated in Figure 2, which shows the smallest attainable decoded bit error probability as a function of the data rate R, where R is measured in multiples of capacity. Note the phase transition at $R = C$. For $R < C$, the minimum attainable p is $0+$, whereas for $R > C$ $p_{\min} = H_2^{-1}(1 - C/R)$ where $H_2(x) = -x \log_2(x) - (1-x)\log_2(1-x)$ is the binary entropy function.

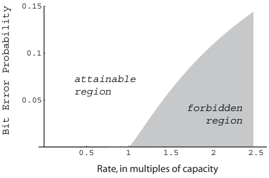


Figure 2 . Shannon's Theorem.

Shannon's Theorem is an existence theorem (note the phrase "there exists an encoder-decoder pair") and we naturally ask "How hard is it to communicate at rate R and decoded error probability p?" We will restrict ourselves to the case $R < C$, for if $R > C$ we would need to consider data compression, which is beyond the scope of this article. We are especially interested in communicating reliably at rates very near capacity, s let us assume in fact that

$$R = (1 - \epsilon)C,$$

where ϵ is a small positive number. Now let's define $\chi_e(\epsilon, p)$ to be the minimum possible *encoding* complexity and $\chi_D(\epsilon, p)$ the minimum possible *decoding* complexity , both mmeasured in ariithmetic operation per information bit, for an encoder-decoder pair that operates at rate $R = (1 - \epsilon)C$ and a decoded bit error probability p.

It is impoortant to know the behavior of $\chi_e(Dta, p)$ and $\chi_D(\epsilon, p)$ for fixed p as $\epsilon \to 0$. naturally we expect a singularity at $\epsilon = 0$, but how severe is it? The classical results (i.e., prior to 1993) in this direction are not encouranging.

1.1 Theorem. *On a discrete memoryless channel of capacity C, for any fixed $0 < p > 1/2$ as $\epsilon \to 0$,*

$$\chi_E(p, \epsilon) = O(1/\epsilon^2)$$
$$\chi_D(p, \epsilon 0 = O(1/e^8).$$

Proof: (Sketch) Use linear codes with (per-bit) encoding complexity $O(n)$ and concatenated codes with decoding complexity $O(n^4)$ [7]. The blocklength n is related to ϵ by $n = O(1/\epsilon^2)$, because of the random coding error exponent, which says that the average

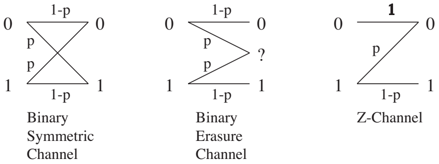


Figure 4.. *Three binary-input Discrete Memoryless Channels.*

error probability for the ensemble of linear codes of rate R satisfiies

$$p \leq \exp(-nE_r(R),$$

and

$$E_r(C(1-\epsilon)) \approx K\epsilon^2 \quad \text{as } \epsilon \to 0.$$

Theorem 1.1 tells us that the encoding problem is not especially difficult, but it suggests that decoding will be a bottleneck. There is one special case, however, for which the decoding problem is, even classically, much less complex.

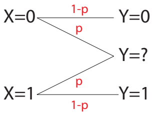


Figure 3 . The Binary Erasure Channel.

1.2 Theorem. *For the binary erasure channel, $\overline{\chi}_D$ can be improved to*

$$\overline{\chi}_D(\epsilon, \pi) = O(1/\epsilon^4),$$

for fixed π, as $\epsilon \to 0$.

Proof: Decode with (per-bit) complexity $O(n^2)$ by solving linear equations for the erased positions. ▪

In the next section we will see how the "turbo revolution" has changed all this: the complexity of communicating near the Shannon Limit now appears to be $O(\epsilon^{-1})$!

2. Turbo Codes.

The complexity estimates developed in the previous section were completely overthrown in 1993 when Claude Berrou and Alain Glavieux introduced turbo-codes [3]. Turbocodes are arguably the most important single invention in the history of coding, but since they have been largely superceded by LDPC codes (which we discuss in the next section), our coverage here will be scanty.

The key idea of the turbocode revolution is that of *suboptimal iterative decoding.* With hindsight it is clear that pre-1993 coding theory and practice was hopelessly mired in a maximum-likelihood paradigm. The justly celebrated turbo decoding algoritm is a low-complexity iterative approximation to maximum a posteriori probability decoding, whose performance, while demonstrably suboptimal, has nevertheless proved to be nearly optimal in an impressive array of experiments around the world.

The turbo idea is illustrated in Figures 4–9. The data is encoded twice, once directly as presented to the encoder and again after being scrambled. or "interleaved." Correspondingly, the decoder hahs two distinct modules, one for each encoding. These subdecoders then exchange their estimates, and iteratively update their estimates until a consensus is reached.

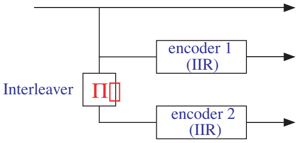


Figure 4 . A classical "parallel" turbo code.

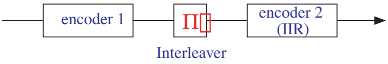


Figure 5 . A "serial" turbo code.

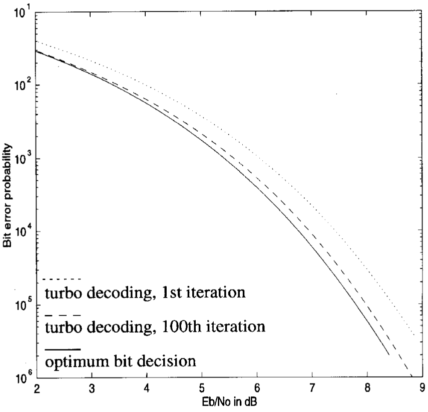


Figure 6 . Typical Performance of Turbo Decoding vs. maximum a posteriori decoding.

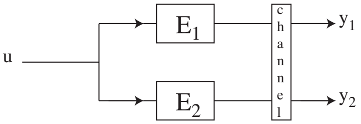


Figure 7 . The Turbo Decoding Problem (Simplified):Infer $\boldsymbol{u}$ from $\{\boldsymbol{y}_1, \boldsymbol{y}_2\}$.

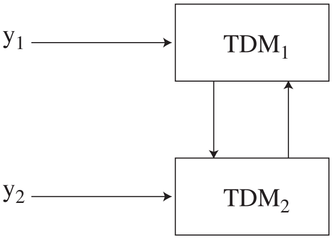


Figure 8 . The “Turbo” Decoder Structure: D_1 and D_2 communicate theiir results to each other, updating their soft estimates of $\boldsymbol{u}$ as they go, until a consennsus is reached

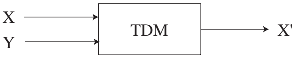


Figure 9 . The Turbo Decoder Module: $\boldsymbol{X} = (X_1, \ldots, X_k)$, the priors in the form of *likelihood* ratios; $\boldsymbol{Y}$ = the channel evidence; $\boldsymbol{X}' = (X'_1, \ldots, X'_k)$, the “extrinsic” likelihoods:

$$X'_i = \frac{\Pr\{U_i = 1 | \boldsymbol{Y}\}}{\Pr\{U_i = 0 | \boldsymbol{Y}\}} \Bigg/ X_i$$

(evidence is not counted twice.)

3. LDPC Codes.

Low-density parity-check (LDPC) codes now appear to be the "final solution" to the 1948 Shannon challenge. They are replacing turbo-codes in many applications, so that it is reasonable to predict that in a few years turbo-codes will be obsolete. This is certainly a strange development, since LDPC codoes were invented by Robert Gallager in 1962 [9]! However, LDPC codes were largely forgotten until their rediscovery by Mackay [18], who not only rediscovered them but used high-performance computers (which were not available to Gallager) to simulate their performance and thereby demonstrate their astonishing power.

Figure 10 shows the parity-check matrix and corresponding Tanner graph [28]. The valid codewords, whcih are transmited over the noisy channel, are required to satisify the parity-checks: $\boldsymbol{H}\boldsymbol{x}^T = 0$. These equations are represented by the bipartite Tanner graph with a variable node (circle) for each column and a check node (box) for each row. Thus every nonzero entry in the H-matrix corresponds to an edge in the Tanner graph.

As will be seen in Figure 11, the noisy version of each transmitted code bit is entered as "evidence" at the appropriate variable node. The evidence is ideally represented as a "likelihoood ratio" of the form

$$m = \frac{\Pr\{X_v = 1|\mathcal{E}\}}{\Pr\{X_v = 0|\mathcal{E}\}},$$

where X_v is the random variable corresponding to the variable node v.

This channel evidence is then modified and circulated around the Tanner graph until a decision about the values of the codeword components can be reached. It remains to describe the nature of the messages and the rules by which they are computed and updated. (See Figures 14 and 13.)

These rules are explained in the captions to Figures 11–16. Briefly, the check-to-variable messages (the λ-messages) are initialized to 1 and the variable-to-check messages (the μ-messages) are initialized to 0. The update rule is

$$\text{Mout} = B\left(\prod_i \text{Min}_i\right).$$

where $B(x) = (1-x)/(1+x)$ is the so-called bilinear transformation. The message from a variable node v to a check node c is

$$\begin{aligned} m(v \to c) &= \Pr\{X_v = 0\} - \Pr\{X_v = 1\} \\ &= \mu(X_v). \end{aligned}$$

where X_v is the random variable associated with the node v. Similarly, the message from a check node c to a variable node v is

$$\begin{aligned} m(v \to c) &= \frac{\Pr\{X_v = 1\}}{\Pr\{X_v = 0\}} \\ &= \lambda(X_v). \end{aligned}$$

Given a message (λ or μ) we can use it to make a decision about the corresponding message bit:

$$[\lambda] = \begin{cases} 0 & \text{if } \lambda \leq 1 \\ 1 & \text{if } \lambda \geq 1. \end{cases}$$

$$[\mu] = \begin{cases} 0 & \text{if } \mu \geq 0 \\ 1 & \text{if } \mu \leq 0. \end{cases}$$

We say that a message is correct if the corresponding decision matches the underlying message bit. The hope, of course, is that after sufficiently many rounds of message passing, the messages will be correct.

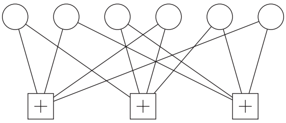


Figure 10 . A Small Tanner Graph, corresponding to the parity-check matrix

$$H = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 \end{pmatrix}.$$

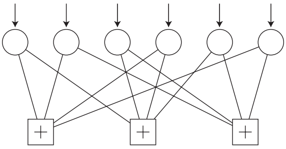


Figure 11 . Evidence from the Channel enters the graph at the variable nodes.

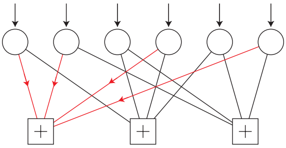


Figure 12 . Messages to a Check Node

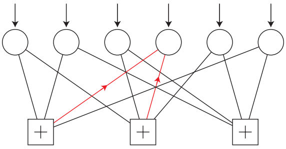


Figure 13 . Messages to a Variable Node

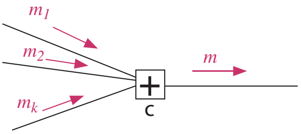


Figure 14 . The μ-messages. Given the incoming messages $m_1, \ldots, m_k$, what should the outgoing message m be? Answer: $m = m_1 \cdots m_k$, provided the m's are defined properly:

$$m(v \to c) = \mu(X_v) = \Pr\{X_v = 0\} - \Pr\{X_v = 1\}.$$

These message rules are based on the following simple observations.

3.1 Theorem. *Let $X_1, \ldots, X_n$ be independent $GF(2)$–valued random variables, and let $S_n = X_1 + \cdots + X_n$. Then*

$$\Pr\{S_n = 0\} - \Pr\{S_n = 1\} = \prod_{i=1}^{n} (\Pr\{X_i = 0\} - \Pr\{X_i = 1\})$$

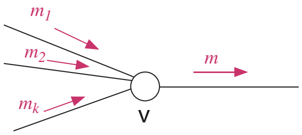


Figure 15 . The λ-messages. Given the incoming messages to a variable node, what should the outgoing messages be? Answer: $m = m_1 \cdots m_k$, provided the m's are defined properly:

$$m(c \to v) = \lambda(X_v) = \frac{\Pr\{X_v = 0\}}{\Pr\{X_v = 1\}}.$$

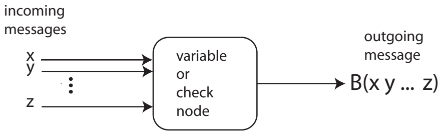


Figure 16 . The complete message update rule.

3.2 Theorem. *Let the a priori probability of the binary random variable X be*

$$\Pr\{X = 1\} = \Pr\{X = 0\} = 1/2.$$

Let $y_1, \ldots, y_m$ be independent (noisy) observations of x.

$$\frac{\Pr\{X = 1|y_1, \ldots, y_m\}}{\Pr\{X = 0|y_1, \ldots, y_m\}} = \prod_{i=1}^{m} \frac{\Pr\{X = 1|y_i\}}{\Pr\{X = 0|y_i\}}.$$

3.3 Lemma. *Let $B(x) = \frac{1-x}{1+x}$. Then*

$$\mu(X) = B(\lambda(X))$$
$$\lambda(X) = B(\mu(X)).$$

4. Introducton to Density Evolution.

The message-passing algorithm described in the previous section is easy to implement, but if the underlying graph (or ensemble of graphs) has cycles, it will not converge to the exact a posteriori probabilities. Nevertheless, experimentally it works extraordinarily well. In this section we will present an introduction to a deep general theory that at least partially explains this remarkable performance. The key idea is that of density evolution.

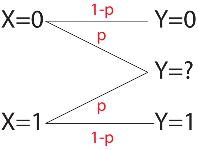


Figure 17 . Special Case: The Binary Erasure Channel. Here $m(v \to c) \in \{+1, -1, 0\}$ and $m(c \to v) \in \{0, \infty, 1\}$.

In density evolution, the idea is to treat the messages sent as random variables over the ensemble of code selection and noise sample, and to track the probability density function of the messages, thereby obtaininng estimates of the probaility that the message is correct.

As an example, consider the BEC (Fig. 17). Then the μ-messages ca assume the values $\{\pm 1, 0\}$, and the λ-messages are $\{\infty, 0, 1\}$. The bilinear transformation is described as follows.

$$\begin{aligned} \mu &\overset{B}{\leftrightarrow} \lambda \\ 1 &\overset{B}{\leftrightarrow} 0 \quad (X = 0) \\ -1 &\overset{B}{\leftrightarrow} \infty \quad (X = 1) \\ 0 &\overset{B}{\leftrightarrow} 1 \quad (X = ?) \end{aligned}$$

Note that a λ-message is guaranteed to be correct unless it equals 1, and a μ-message is correct unless it equals zero. Thus one can simply track the probability that a given message is an "erasure."

A typical LDPC code, or rather code ensemble (i.e. collection of codes) is shown in Fig. 18. In this particular ensemble there are 6 degree 3 variable nodes and three degree 6 check nodes. Thus there are $6 \times 3 = 18$ edges connected to varible nodes and 18 edges connected to thhe three check nodes. These edges are connected to each other via the "interleaver," Π so that the ensemble depicted in Fig. 18 represents 18! different Tanner graphs.

Following the calculations in Figs. 19–26, we see that the $(3, 6)$ ensemble threshold

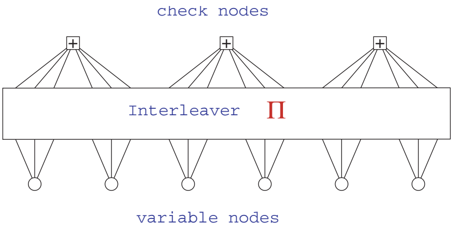


Figure 18 . Tanner Graph for a (regular) $(3,6)$ LDPC Code Ensemble.

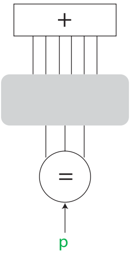


Figure 19 . The $(3,6)$ Ensemble, showing just one degree-3 variable node and one degree-6 check node. Evidence from the channel arrives from below. Thisi evidence, which “seeds” the decoder, is absent (erased) with probability p.

for the BEC is $p = 0.425$, as compared to the Shannon limit for codes of rate 1/2, viz. $p = 0.50$.

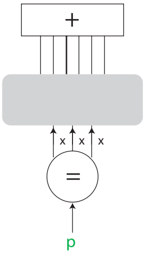


Figure 20 . The First Step. x = probability that the indicated message is “erasure.” (On the first iteration, $x = p$.)

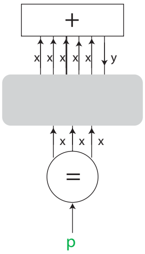


Figure 21 . The Second Step. y is an erasure iff at least one of the x’s is an erasure. Thus $y = 1 - (1 - x)^5$.

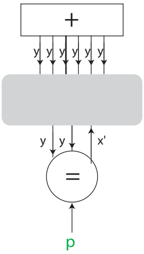


Figure 22 . The Third Step. x' is an erasure iff both y's and the channel input are: $x' = py^2 = p(1-(1-x)^5)^2$.

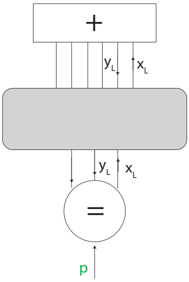


Figure 23 . To summarize: if the probability of an erased message is x_L on the Lth iiteration, then

$$x_{L+1} = f(x_L),$$

where

$$f(x) = p(1-(1-x)^5)^2.$$

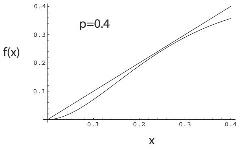


Figure 24 . With $p = 0.4$, the only solution to the equation $f(x) = x$ is $x = 0$.

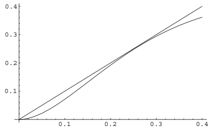


Figure 25 . With $p = 0.425$, the curves are just tangent. Therefore the noise threshold for the $(3, 6)$ ensemble is 0.425, which should be compared to the Shannon Lmit for codes of ate 12, viz., $p = 0.5$.

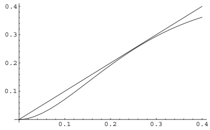


Figure 26 . With $p = 0.45$, there are two nonzero solutions.

References and Recommended Reading.

Note: A superb source of information about iterative message-passing algorithms is the Special Issue of the IEEE Trans. Inform. Theory vol. IT-47 (Feb. 2001). Many of the individual papers listed below are in the special issue.

1. S. Aji, *Graphical Models and Iterative Decoding.* Caltech Ph.D. thesis, 2000.

2. S. M. Aji and R. J. McEliece, "The generalized distributive law." *IEEE Trans. Inform. Theory*, vol. IT-46, no. 2 (March 2000), pp. 325–343.

3. C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon limit error-correcting coding: Turbo codes." *Proc. 1993 IEEE International Conference on Communications* (Geneva, Switzerland, May 1993), pp. 1064–1070.

4. S.-Y. Chung, G. D. Forney, Jr., T. J. Richardson and R. Urbanke, "On the design of low-density parity-check codes within 0.0045 dB from the Shannon limit," IEEE Commun. Letters, vol. 5, pp. 58-60, Feb. 2001.

5. D. Divsalar, S. Dolinar, and F. Pollara, "Low complexity turbo-like codes," Proc. 2nd International Symp. Turbo Codes & Related Topics, Brest, France, Sept. 2000, pp. 73–80.

6. Dariush Divsalar, Hui Jin, and R. McEliece, "Coding Theorems for 'Turbo-Like' Codes," Proc. 1998 Allerton Conference, pp. 201-210.

7. G. D. Forney, Jr., *Concatenated Codes.* Cambridge, Mass.: MIT Press, 1966.

8. B. J. Frey and D. J. C. MacKay, "Irregular turbo-like codes," Proc. 2nd International Symp. Turbo Codes & Related Topics, Brest, France, Sept. 2000, pp. 67–72.

9. R. G. Gallager, *Low density parity check codes*, Cambridge, Mass.: MIT Press, 1963.

10. Hui Jin, A. Khandekar, and R. J. McEliece, "Irregular Repeat-Accumulate Codes," Proc. 2nd International Symp. Turbo Codes & Related Topics, Brest, France, Sept. 2000, pp. 1-8.

11. Hui Jin and R. J. McEliece, "Coding Theorems for Turbo Code Ensembles.' Submittted to *IEEE Trans. Inform. Theory*, January 2001.

12. I. Kanter and D. Saad, "Error-correcting codes that nearly saturate Shannon's bound," *Phys. Rev. Letters.* vol. 83, no. 12 (27 Sept. 1999), pp. 2660–2663.

13. F. Kschischang and B. Frey, "Iterative decoding of compound codes by probability propagation in graphical models," *IEEE J. Sel. Areas Comm.*, vol. 16, no. 2 (February 1998), pp. 219–230.

14. F. Kschischang, B. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," IEEE Trans. Inform. Theory vol.IT-47 (Feb. 2001), pp. 498–519.

15. M. Luby, M. Mitzenmacher, A. Shokrollahi, D. Spielman, and V. Stemann, "Practical

loss-resilient codes," Proc. 29th ACM Symp. on the Theory of Computing (1997), pp. 150-159.

16. M. Luby, M. Mitzenmacher, A. Shokrollahi, and D. Spielman, "Analysis of low-density codes and improved designs using irregular graphs," Proc. 30th ACM Symp. on the Theory of Computing (1998), pp. 249-258.

17. M. Luby, M. Mitzenmacher, A. Shokrollahi, and D. Spielman, "Efficient Erasure Correcting Codes,"IEEE Trans. Inform. Theory vol.IT-47 (Feb. 2001), pp. 569–584.

18. D. J. C. MacKay, "Good error-correcting codes based on very sparse matrices." *IEEE Trans. Inform. Theory*, vol. 45 (March 1999), pp. 399–431.

19. D. J. C. MacKay and R. M. Neal, "Near Shannon limit performance of low density parity check codes." *Electronic Letters*, vol. 32 (Aug. 1996), pp. 1645–1646. Reprinted in *Electronic Letters*, vol. 33 (March 1997), pp. 457–458.

20. R. J. McEliece, *The Theory of Information and Coding.* Reading, Mass.: Addison-Wesley, 1977.

21. R. J. McEliece, D. J. C. MacKay, and J. -F. Cheng, "Turbo decoding as an instance of Pearl's 'belief propagation' algorithm," *IEEE J. Sel. Areas Comm.*, vol. 16, no. 2 (Feb. 1998), pp. 140–152.

22. J. Pearl, *Probabilitic Reasoning in Intelligent Systems.* San Francisco: Morgan-Kaufmann, 1988.

23. L. Ping and K. Y. Wu, "Concatenated tree codes," Proc. 2nd International Symp. Turbo Codes & Related Topics, Brest, France, Sept. 2000, pp. 161–164.

24. T. J. Richardson and R. Urbanke, "The capacity of low-density parity-check codes under message passing decoding," IEEE Trans. Inform. Theory vol.IT-47 (Feb. 2001), pp. 599–619498–519.

25. T. J. Richardson, A. Shokrollahi,, and R. Urbanke, "Design of capacity-approaching irregular LDPC codes, IEEE Trans. Inform. Theory vol.IT-47 (Feb. 2001), pp. 619–637.

26. C. E. Shannon, *The Mathematical Theory of Communication.* Reprinted by the University of Illinois Press, 1998.

27. M. A. Shokrollahi, "New sequences of linear time erasure codes approaching channel capacity," Proc. 1999 ISITA (Honolulu, Hawaii, November 1999) pp. 65–76.

28. R. M. Tanner, "A recursive approach to low complexity codes," *IEEE Trans. Inform. Theory*, vol. IT-27 (Sept. 1981). pp. 533-547.

29. N. Wiberg, *Codes and Decoding on General Graphs.* Linköping Studies in Science and Technology, Dissertations no. 440. Linköping, Sweden, 1996.

FLOER THEORY AND LOW DIMENSIONAL TOPOLOGY

DUSA MCDUFF

Abstract. My lecture will aim to give a pictorial introduction to the new 3- and 4-manifold invariants recently constructed by Ozsvath and Szabo. These are based on a Floer theory associated with Heegaard diagrams. The following notes try to give somewhat more of the background than would be possible in a lecture. Readers wanting to know more should consult Ozsvath and Szabo's recent survey article [8].

1. The Floer complex

This section begins by outlining traditional Morse theory, using the Heegaard diagram of a 3-manifold as an example. It then describes Witten's approach to Morse theory, a finite dimensional version of Floer theory. Finally, it discusses Lagrangian Floer homology. This is fundamental to Ozsvath and Szabo's work; their Heegaard–Floer theory is a special case of this general construction.

1.1. **Classical Morse theory.** Morse theory attempts to understand the topology of a space X by using the information provided by real valued functions $f : X \to \mathbb{R}$. In the simplest case, X is a smooth m-dimensional manifold, compact and without boundary, and we assume that f is generic and smooth. This means that its critical points p are isolated and there is a local normal form: in suitable local coordinates $x_1, \dots, x_m$ near the critical point $p = 0$ the function f may be written as

$$f(x) = -x_1^2 - \cdots - x_i^2 + x_{i+1}^2 + \cdots + x_m^2.$$

The number of negative squares occurring here is independent of the choice of local coordinates and is called the **Morse index** $\mathrm{ind}(p)$ of the critical point.

Functions $f : X \to \mathbb{R}$ that satisfy these conditions are called **Morse functions**. One analyses the structure of X by considering the family of sublevel sets

$$X^c := f^{-1}(-\infty, c].$$

Date: 30 November, 2004.

Key words and phrases. Floer complex, Morse complex, Heegaard diagram, Ozsvath–Szabo invariant, low dimensional topology.

Partly supported by the NSF grant DMS 0305939.

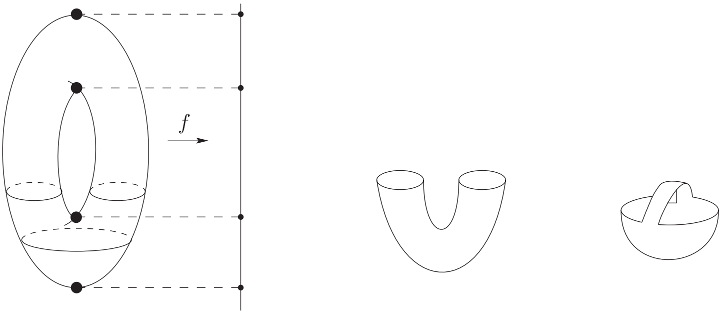


Figure 1: Morse decompostion of the torus

These spaces are diffeomorphic as c varies in each interval of regular (i.e. noncritical) values, and their topology changes in a predictable way as c passes a critical level.

One way to prove this is to consider the **negative gradient flow** of f. Choose a generic metric μ on X. Then the gradient vector field ∇f is perpendicular to the level sets $f^{-1}(c)$ at regular values and vanishes only at the critical points. Therefore one can push a regular level $f^{-1}(c)$ down to $f^{-1}(c-\varepsilon)$ by following the flow of ∇f. Moreover one can understand what happens to the sublevel sets as one passes a critical level by looking at the set of downward gradient trajectories emanating from the critical point p. The points on this set of trajectories form the **unstable manifold**

$$W^u_f(p) := \{p\} \cup \big\{u(s) : s \in \mathbb{R},\ \dot{u}(s) = -\nabla f(u(s)),\ \lim_{s\to-\infty} u(s) = p\big\}.$$

It is easy to see that $W^u_f(p)$ is diffeomorphic to $\mathbb{R}^d$ where $d = \operatorname{ind}(p)$. Similarly, each critical point has a stable manifold $W^s_f(p)$ consisting of trajectories that converge towards p as $s \to \infty$.

For example, if c is close to $\min f$ (and we assume that f has a unique minimum) then the sublevel set X^c is diffeomorphic to the closed ball $D^m := \{x \in \mathbb{R}^m : \|x\| \le 1\}$ of dimension m. When c passes a critical point p of index 1 a one handle (homeomorphic to $[0,1] \times D^{m-1}$) is added. One should think of this handle as a neighborhood of the unstable manifold $W^u_f(p) \cong \mathbb{R}$. Similarly, when one passes a critical point of index 2 one adds a 2-handle: see Milnor [4]. When $m = 2$, a 2-handle is just a 2-disc, as one can see in the well known decomposition for the 2-torus $\mathbb{T} = S^1 \times S^1$ given by the height function: cf. Fig. 1.

The next example shows how one can use a Morse function to give a special kind of decomposition of a 3-manifold Y that is known as a Heegaard splitting. This description of Y lies at the heart of Ozsvath and Szabo's theory.

Example 1.1. Heegaard diagram of a 3-manifold. Choose the Morse function $f : Y \to \mathbb{R}$ to be *self-indexing*, i.e. so that all the critical points of

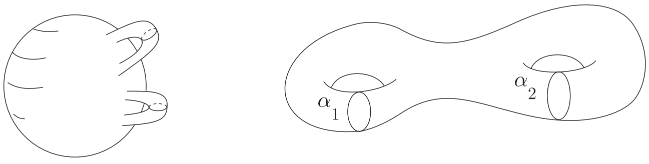


Figure 2: Genus two handlebody

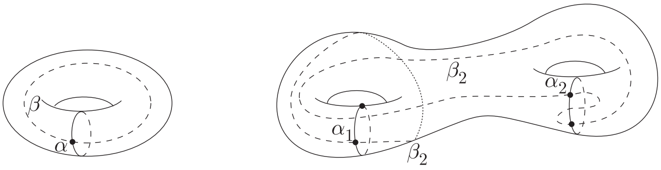


Figure 3: Genus 1 and 2 Heegaard diagrams for S^3

index i lie on the level $f^{-1}(i)$. Then the cut $f^{-1}(3/2)$ at the half way point is a Riemann surface Σ_g of genus g equal to the number of index 1 critical points of f, and the sublevel set $Y^{3/2}$ is a *handlebody of genus* g, i.e. the union of a 3-ball D^3 with g 1-handles: see Fig. 2. By symmetry, the other half $f^{-1}[3/2, 3]$ of Y is another handlebody of genus g. Thus Y is built from a single copy of the surface $\Sigma = f^{-1}(3/2)$ by attaching handlebodies U_α, U_β to its two sides.

The attaching map of U_α is determined by the loops in Σ that bound discs in U_α: if Σ has genus g, there is an essentially unique collection of g disjoint embedded circles $\alpha_1, \dots, \alpha_g$ in Σ that bound discs $D_1, \dots, D_g$ in U_α. These discs are chosen so that when they are cut out the remainder $U_\alpha \setminus \{\alpha_1, \dots, \alpha_g\}$ of U_α is still connected. Therefore Y can be described by two collections $\alpha := \{\alpha_1, \dots, \alpha_g\}$ and $\beta := \{\beta_1, \dots, \beta_g\}$ of disjoint circles on the Riemann surface Σ_g. See Fig. 3. This description (known as a **Heegaard diagram**) is unique modulo some basic moves.[1]

As an example, there is a well known decomposition of the 3-sphere $\{(z_1, z_2) : |z_1|^2 + |z_2|^2 = 1\}$ into two solid tori (handlebodies of genus 1), $U_1 := \{|z_1| \le |z_2|\}$ and $U_2 := \{|z_1| \ge |z_2|\}$, and the corresponding circles in the 2-torus $\Sigma_1 = \{|z_1| = |z_2|\}$ are

$$\alpha_1 = \Big\{\frac{1}{\sqrt{2}}(e^{i\theta}, 1) : \theta \in [0, 2\pi]\Big\}, \qquad \beta_1 = \Big\{(\frac{1}{\sqrt{2}}(1, e^{i\theta}) : \theta \in [0, 2\pi]\Big\},$$

with a single intersection point $(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$. Section 2 in [5] contains a nice description of the properties of Heegaard diagrams.

The Oszvath–Szabo invariants capture information about the intersection points $\alpha_j \cap \beta_k$ of these two families. Note that each α_j is the intersection

[1] These are; isotoping the loops in α, β, changing these loops by "handleslides" and finally stabilizing Σ_g by increasing its genus in a standard way.

$W^s(p_j) \cap \Sigma$ of the upward gradient trajectories from some index 1 critical point p_j of f with the level set Σ. Similarly, the β_k are the intersections with Σ of the downward gradient trajectories from the index 2 critical points q_k. Hence, each intersection point $\alpha_j \cap \beta_k$ corresponds to a gradient trajectory from q_k to p_j.

This traditional version of Morse theory is useful in some infinite dimensional cases as well, especially in the study of closed geodesics. Here one looks at the length (or energy) functional $\mathcal{F}$ on the space $\mathcal{X}$ of smooth loops in X. Its critical points are closed geodesics. They may not be isolated but they have finite index. For further discussion see Bott's wonderful survey article [1].

1.2. **The Morse–Witten complex.** Witten observed that the sublevel sets $f^{-1}(-\infty, c]$ have little physical meaning. More relevant are the gradient trajectories between critical points, which occur as "tunnelling effects" in which one state (regime at a critical point) affects another. His influential paper [10] pointed out that one could use these trajectories to build a complex $C_*(X; f)$ that calculates the homology of a manifold X as follows. The k-chains are finite sums of critical points of index k:

$$C_k(X; f) = \Big\{ \sum_{x \in \mathrm{Crit}_k(f)} a_x \langle x \rangle \, : \, a_x \in \mathbb{Z} \Big\}, \tag{1.1}$$

and the boundary operator $\partial : C_k(X; f) \to C_{k-1}(X; f)$ has the form

$$\partial \langle x \rangle = \sum_{y \in \mathrm{Crit}_{k-1}(f)} n(x, y) \, \langle y \rangle, \tag{1.2}$$

where $n(x, y)$ is the number of gradient trajectories of f from x to y. (Here one either counts mod 2 or counts using appropriate signs that come from suitably defined orientations of the trajectory spaces.) Note that the chain groups depend only on f but the boundary operator depends on the choice of a generic auxiliary metric μ.

We claim that $C_*(X; f)$ is a **chain complex**, i.e. that $\partial^2 = 0$. To see this, note that

$$\begin{aligned} \partial^2 \langle y \rangle &= \sum_{y \in \mathrm{Crit}_{k-1}(f)} n(x, y) \, \partial \langle y \rangle \\ &= \sum_{y \in \mathrm{Crit}_{k-1}(f)} \; \sum_{z \in \mathrm{Crit}_{k-2}(f)} n(x, y) n(y, z) \langle z \rangle. \end{aligned}$$

The coefficient $\sum_y n(x, y) n(y, z)$ of $\langle z \rangle$ in this expression is the number of once-broken gradient trajectories from x to z and vanishes because these occur in cancelling pairs; the space

$$\widehat{\mathcal{M}}(x, z) := \mathcal{M}(x, z)/\mathbb{R} = \big(W^u_f(x) \cap W^s_f(y) \big)/\mathbb{R}$$

of (unparametrized)[2] trajectories from x to y is a union of circles and open intervals whose ends may be identified with the set of once-broken gradient trajectories from x to z.

Therefore the homology $H_*(X; f) := \ker\partial/\mathrm{im}\partial$ of this complex is defined. It turns out to be isomorphic to the usual homology $H_*(X)$ of X. In particular, it is independent of the choice of metric μ and function f.

Remark 1.2. Morse–Novikov theory. There is a variant of this construction whose initial data is a closed 1-form ν on X instead of a Morse function. If ν is integral, it has the form $\nu = df$ for some circle valued function $f : X \to S^1$, and there is a cover $\mathbb{Z} \to \widetilde{X} \to X$ of X on which f lifts to a real valued function $\widetilde{f}$. Each critical point of f lifts to an infinite number of critical points of $\widetilde{f}$. The Morse–Novikov complex of f is essentially just the the Morse complex of $\widetilde{f}$. It supports an action of the group ring $\mathbb{Z}[U, U^{-1}]$ of the group $\{U^n : n \in \mathbb{Z}\}$ of deck transformations of the cover $\widetilde{X} \to X$ and is finitely generated over this ring. One of the Heegaard–Floer complexes is precisely of this kind.

Remark 1.3. Operations on the Morse complex. This point of view has proved very fruitful, not only for the applications we discuss later, but also for the understanding of the topology of manifolds and their loop spaces, a topic of central importance in so-called "string topology". Here the aim is to understand various homological operations (e.g. products) at the chain level, and it is very important to have a versatile chain complex to work with. The Morse–Witten complex fits into such theories very well. For example, given three generic Morse functions $f_k,\ k = 1, 2, 3$, one can model the homology intersection product $H_i \otimes H_i \to H_{i+j-m}$ on an m-dimensional manifold by defining a chain level homomorphism

$$\phi : C_i(X; f_1) \times C_j(X; f_2) \to C_{i+j-m}(X; f_3)$$

by counting Y-shaped trajectories from a pair (x_1, x_2) of critical points in $\mathrm{Crit}(f_1) \times \mathrm{Crit}(f_2)$ to a third critical point $x_3 \in \mathrm{Crit}(f_3)$ whose two arms are gradient trajectories for f_1 and f_2 and whose leg is a trajectory for f_3. Thus

$$\phi(x_1, x_2) = \sum n(x_1, x_2, x_3)\langle x_3\rangle,$$

where $n(x_1, x_2, x_3)$ is the number of such trajectories, counted with signs. If the functions f_k and metric μ are generic, then this number is finite and agrees with the number of triple intersection points of the three cycles $W^s_{f_1}(x_1), W^s_{f_2}(x_2)$ and $W^u_{f_3}(x_3)$ which have dimensions i_1, i_2 and $m - i_3$ respectively, where $i_3 = i_1 + i_2 - m$. In fact, there is a bijection between the set of Y-images and the set of such triple intersection points.

This is just the beginning. One thinks of Y as a tree graph with two inputs at the top and one output at the bottom. The nonassociativity of

[2] The elements $a \in \mathbb{R}$ act on the trajectories $u : \mathbb{R} \to X$ in $\mathcal{M}(x, z)$ by reparametrization: $a * u(s) := u(s + a)$.

the intersection product at the chain level gives rise to a new operation that counts maps of trees in X with three inputs and one output. Continuing this way, one may construct the full Morse–Witten A^∞-algebra as well as many other homology operations such as the Steenrod squares: see for example Cohen [2].

The fact that the chain complexes of Lagrangian Floer theory support similar maps is an essential ingredient of Ozsvath and Szabo's work.

1.3. **Floer theory.** Inspired partly by Witten's point of view but also by work of Conley and Gromov, Floer realised that there are interesting infinite dimensional situations in which a similar approach makes sense. In these cases, the ambient manifold $\mathcal{X}$ is infinite dimensional and the critical points of the function $\mathcal{F} : X \to \mathbb{R}$ have infinite index and coindex. Therefore one usually cannot get much information from the sublevel sets $\mathcal{F}^{-1}(-\infty, c]$ of $\mathcal{F}$. Also, one may not be able to choose a metric on $\mathcal{X}$ such that the gradient flow of $\mathcal{F}$ is everywhere defined. However, Floer realised that in some important cases one can choose a metric so that the spaces $\mathcal{M}(x, y)$ of gradient trajectories between distinct critical points x, y of $\mathcal{F}$ have properties analogous to those in the finite dimensional case. Hence one can define the **Floer chain complex** using the recipe described in equations (1.1) and (1.2) above.

We now describe the version of Floer theory used by Ozsvath–Szabo. In their situation both the critical points of $\mathcal{F}$ and its gradient flow trajectories have natural geometric interpretations.

Example 1.4. Lagrangian Floer homology. Let M be a $2n$-dimensional manifold with symplectic form ω (i.e. a closed, nondegenerate 2-form) and choose two Lagrangian submanifolds L_0, L_1. These are smooth submanifolds of dimension n on which the symplectic form vanishes identically. (Physicists call them *branes.*) We assume that they intersect transversally and also that their intersection is nonempty, since otherwise the complex we aim to define is trivial.

Denote by $\mathcal{P} := \mathcal{P}(L_0, L_1)$ the space of paths x from L_0 to L_1:

$$x : [0, 1] \to M, \qquad x(0) \in L_0,\ x(1) \in L_1.$$

Pick a base point $x_0 \in L_0 \cap L_1$ considered as a constant path in $\mathcal{P}$ and consider the universal cover $\widetilde{\mathcal{P}}$ based at x_0. Thus elements in $\widetilde{\mathcal{P}}$ are pairs, $(x, \hat{x})$ where $\hat{x}$ is an equivalence class of maps $\hat{x} : [0, 1]\times[0, 1] \to M$ satisfying the boundary conditions

$$\hat{x}(0, t) = x_0, \quad \hat{x}(s, i) \in L_i, \quad \hat{x}(1, t) = x(t).$$

The function $\mathcal{F}$ is the action functional $\mathcal{A} : \widetilde{\mathcal{P}} \to \mathbb{R}$ given by

$$\mathcal{A}(x, \hat{x}) = \int_0^1 \int_0^1 \hat{x}^*(\omega),$$

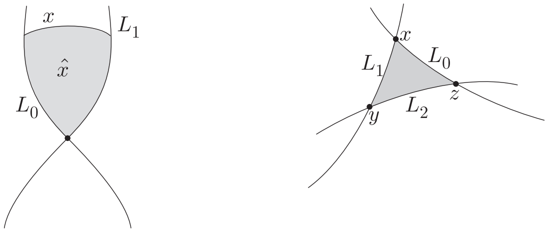


Figure 4: The Path space $\widetilde{P}$ and a holomorphic triangle

and its critical points are the lifts to $\widetilde{\mathcal{P}}$ of the points of the intersection $L_0 \cap L_1$. (See Fig. 4. $\mathcal{A}$ does not depend on the homotopy class of the map $\hat{x}$ because ω is closed and vanishes on the L_i.)

Now, let us consider the $\mathcal{A}$-gradient trajectories between the critical points. Since $\mathcal{P}$ is infinite dimensional these depend significantly on the choice of metric. We use a metric on $\mathcal{P}$ that is determined by a particular kind of Riemannian metric on M, namely a metric g_J given in the form $g_J(v, w) = \omega(v, Jw)$ where $J : TM \to TM$ is an almost complex structure on M that is compatible with ω in the sense that[3]

$$\omega(Jv, Jw) = \omega(v, w), \quad \omega(v, Jv) > 0, \text{ for } v, w \in T_pM \setminus \{0\}.$$

It then turns out[4] that the gradient trajectories $\hat{u} : \mathbb{R} \to \widetilde{\mathcal{P}}$ of $\mathcal{A}$ are given by J-holomorphic strips

$$u : \mathbb{R} \times [0, 1] \to M, \qquad u(s, t) := \hat{u}(s)(t),$$

in M with boundary on L_0 and L_1:

$$\partial_s u + J(u)\partial_t u = 0, \quad u(s, 0) \in L_0, \quad u(s, 1) \in L_1. \tag{1.3}$$

One cannot always define a Floer complex in this setup because ∂^2 may not always vanish. The basic problem is that it may be impossible to define a good compactification of the 1-dimensional trajectory spaces $\widehat{\mathcal{M}}(x, z)$ simply by adding once-broken trajectories. (There is recent work by Fukaya–Oh–Ohta–Ono that sets up a framework in which to measure the obstructions to the existence of the Floer complex.) However, Ozsvath–Szabo consider a

[3] These equations generalise the well known relations between the Kähler metric g_J and Kähler form ω on a Kähler manifold M. The only difference is that the almost complex structure J need not be integrable, i.e. need not come from an underlying complex structure on the manifold M.

[4] The associated L_2-inner product on the tangent bundle of the path space is defined as follows. Given a path $x : [0, 1] \to M$ the tangent space $T_x(\mathcal{P})$ consists of all (smooth) sections ξ of the pullback bundle $x^*(TM)$, i.e. $\xi(t) \in T_{x(t)}M$ for all $t \in [0, 1]$ and satisfies the boundary conditions $\xi(i) \in T_{x(i)}L_i$ for $i = 0, 1$. Given two such sections ξ, η, we set

$$\langle \xi, \eta \rangle := \int_0^1 g_J(\xi(t), \eta(t))\, dt.$$

Then the g_J-gradient of $\mathcal{A}$ is the vector field $\nabla \mathcal{A}$ defined by setting the inner product $\langle \nabla \mathcal{A}, \xi \rangle$ equal to $d\mathcal{A}(\xi)$, (the differential $d\mathcal{A}$ evaluated on the tangent vector ξ.)

very special case of this construction in which the Lagrangian submanifolds arise from the geometry of the Heegaard diagram. In their case, $\partial^2 = 0$ and so the Floer homology groups $HF_*(L_0, L_1)$ are defined. Moreover they are independent of the choice of almost complex structure J on M and of any perturbations used in their construction.

Just as in the case of the Morse complex where one can define various products on the chain level by counting images of Ys and other trees, one can define topologically interesting chain maps between the complexes $CF_*(L_i, L_j)$ for different Lagrangian pairs by counting holomorphic triangles (think of these as fattened up Ys) or other polygons, with each boundary component mapping to a different Lagrangian submanifold L_i. For short we refer to the collection of such maps as the **naturality properties** of Lagrangian Floer theory. These properties lie at the heart of the proof that the Heegaard–Floer groups depend only on the manifold Y rather than on the chosen Heegaard splitting. They can also be used to establish various interesting long exact sequences in the theory. Similar structures appear in Seidel's work [9] on the **Fukaya category** of a symplectic manifold, the basis of one side of the homological mirror symmetry conjecture.

2. Heegaard–Floer theory

In this section we first define the Heegaard–Floer complexes. Then we briefly describe some applications.

2.1. Definition of the invariants.

We saw in Example 1.1 that a 3-manifold Y is completely determined by a triple (Σ, α, β) where Σ is a Riemann surface of genus g and α, β are sets of disjoint embedded circles

$$\alpha = \{\alpha_1, \dots, \alpha_g\}, \quad \beta = \{\beta_1, \dots, \beta_g\}.$$

Ozsvath and Szabo's idea is to use this data to construct a symplectic manifold (M, ω) together with a pair of Lagrangian submanifolds $\mathbb{T}_\alpha, \mathbb{T}_\beta$ and then to consider the corresponding Floer complex. This is a very rough version of their idea: in fact the manifold is not quite symplectic, the submanifolds are not quite Lagrangian and they also put some extra structure on the Floer complex. The most amazing thing about their construction is that it does give interesting 3-manifold invariants.

For simplicity we shall assume throughout the following discussion that Y is a **rational homology sphere**, i.e. that $H_*(Y; \mathbb{Q}) \cong H_*(S^3 : \mathbb{Q})$. This means that the abelianization $H_1(Y; \mathbb{Z})$ of the fundamental group $\pi_1(Y)$ is finite, that $H_2(Y; \mathbb{Z}) = 0$, and that Y is orientable. However, the invariants may be defined for all Y.

The manifold M: This is the g-fold symmetric product $M_g := \mathrm{Sym}^g \Sigma_g$ of Σ_g, i.e. the quotient

$$M_g := \prod_g \Sigma / S_g,$$

of the g-fold product $\prod_g \Sigma := \Sigma \times \cdots \times \Sigma$ by the obvious action of the symmetric group S_g on g letters. M_g is smooth: if $\mathbb{C}$ is a local chart in Σ then the points in $\mathrm{Sym}^g\mathbb{C}$ are unordered sets of g points in $\mathbb{C}$ and hence are the roots of a unique monic polynomial whose coefficients give a local chart on $\mathrm{Sym}^g\mathbb{C}$. However, M_g has no *natural* smooth structure; it inherits a complex structure J_M from the choice of a complex structure j on Σ, but different choices of j give rise to different[5] smooth structures on M_g. Similarly, although (M_g, J_M) is a Kähler manifold and so has symplectic structures, there is no natural choice of symplectic structure on M_g.

The manifold M_g has rather simple homotopy and cohomology. For example, in genus two $M_2 := \mathrm{Sym}^2\Sigma_2$ is a 1-point blow up of the standard 4-torus $\mathbb{T}^4$, i.e. topologically it is the connected sum of $\mathbb{T}^4$ with a negatively oriented copy of the complex projective plane. In general, $\pi_1(M_g) \cong H_1(M_g;\mathbb{Z})$ is abelian of rank $2g$. In fact the inclusion $\Sigma \times \mathrm{pt} \times \cdots \times \mathrm{pt}$ induces an isomorphism $H_1(\Sigma_g) \cong \pi_1(M_g)$. When $g > 1$ the cohomology ring of M_g has one other generator in $H^2(M_g)$ that is Poincaré dual to the submanifold

$$\{z\} \times \mathrm{Sym}^{g-1}(\Sigma) \subset M_g,$$

where z is any fixed point in Σ. Correspondingly $\pi_2(M_g) = \mathbb{Z}$, with generator

$$S^2 \equiv \Sigma/\rho \overset{\iota}{\hookrightarrow} \mathrm{Sym}^2(\Sigma) \to \mathrm{Sym}^g(\Sigma),$$

where we think of the 2-sphere S^2 as the quotient of Σ by a suitable involution ρ (e.g. the hyperelliptic involution) and set

$$\iota(z) := [z, \rho(z)] \in \mathrm{Sym}^2\Sigma, \quad z \in \Sigma.$$

The fact that $\pi_2(M_g)$ has rank 1 and is generated by a holomorphic 2-sphere with trivial normal bundle is one of the reasons why Ozsvath–Szabo's boundary operator ∂ has $\partial^2 = 0$. (Technically, this is in the **monotone** case.)

The tori $\mathbb{T}_\alpha, \mathbb{T}_\beta$: Because the circles α_i are mutually disjoint, the product

$$\alpha_1 \times \cdots \times \alpha_g \subset \prod_g \Sigma$$

maps bijectively onto a torus $\mathbb{T}_\alpha$ in M_g. This torus is clearly totally real, i.e. its tangent bundle $T\mathbb{T}_\alpha$ intersects $J_M(T\mathbb{T}_\alpha)$ transversally. There is no natural smooth symplectic structure on M_g that makes it Lagrangian, but this does not really matter since its inverse image in the product is Lagrangian for product symplectic forms.

[5] These smooth structures s_j are diffeomorphic. They are different in the sense that the identity map $(M, s_j) \to (M, s_{j'})$ is not smooth. Readers familiar with complex geometry might note that $\mathrm{Sym}^g\Sigma_g$ is a rather special complex manifold. It is birationally equivalent to the Picard variety $\mathrm{Pic}^g(\Sigma) \cong \mathbb{T}^{2g}$ of Σ: to get a map $\mathrm{Sym}^g\Sigma \to \mathrm{Pic}^g(\Sigma)$ think of the set of g points as a divisor and map it to the point in $\mathrm{Pic}^g(\Sigma)$ given by the corresponding degree g line bundle.

If the α_j and β_k intersect transversally then the two tori $\mathbb{T}_\alpha, \mathbb{T}_\beta$ also intersect transversally. Each intersection point can be written as

$$\mathbf{x} := (x_1, \dots, x_g), \qquad x_k \in \alpha_k \cap \beta_{\pi(k)}, \ k = 1, \dots, g, \ \pi \in S_g.$$

The trajectory spaces $\mathcal{M}(\mathbf{x}, \mathbf{y})$: Fix a complex structure j on Σ and consider the corresponding complex structure $J = J_M$ on the symmmetric product M_g. Given two intersection points $\mathbf{x}, \mathbf{y} \in \mathbb{T}_\alpha \cap \mathbb{T}_\beta$ the elements in $\mathcal{M}(\mathbf{x}, \mathbf{y})$ are the J-holomorphic strips $u : \mathbb{R} \times S^1 \to M_g$ from $\mathbf{x}$ to $\mathbf{y}$ satisfying the conditions of (1.3). The domain $\mathbb{R} \times S^1$ is conformally equivalent to the closed unit disc $\mathbb{D}$ in $\mathbb{C}$ with the two boundary points $\pm i$ removed. Thus Ozsvath–Szabo think of the strips as continuous maps

$$u : \mathbb{D} \to M_g,$$

that are holomorphic in the interior $\operatorname{int} \mathbb{D}$ and take the left boundary $\partial \mathbb{D} \cap \{\Re z < 0\}$ to $\mathbb{T}_\alpha$ and the right boundary $\partial \mathbb{D} \cap \{\Re z > 0\}$ to $\mathbb{T}_\beta$. Continuous maps $\phi : \mathbb{D} \to M_g$ that satisfy these boundary conditions but are not necessarily holomorphic are called **Whitney discs** from $\mathbf{x}$ to $\mathbf{y}$.

One can define a complex whose vertices are the intersection points $\mathbb{T}_\alpha \cap \mathbb{T}_\beta$ and whose boundary map is defined as in (1.2) by counting the number of elements in the 0-dimensional components of $\mathcal{M}(\mathbf{x}, \mathbf{y})/\mathbb{R}$. However, this complex contains no interesting information: its homology depends just on $H_*(Y)$. Therefore Ozsvath and Szabo add two pieces of extra structure. Firstly, they observed that this complex decomposes into a direct sum of subcomplexes that are indexed by the Spin^c-structures[6] $\mathfrak{s}$ on Y. Secondly, they work in a suitable covering $\widetilde{\mathcal{P}}$ of the path space $\mathcal{P}(\mathbb{T}_\alpha, \mathbb{T}_\beta)$ with deck transformation group $\mathbb{Z}$. By taking the action of the generator U of this group into account as in Remark 1.2, they define various different, but related, chain complexes

$$CF^\infty(Y, \mathfrak{s}), \quad \widehat{CF}(Y, \mathfrak{s}), \quad CF^+(Y, \mathfrak{s}), \quad CF^-(Y, \mathfrak{s}), \quad CF_{\text{red}}(Y, \mathfrak{s}).$$

Whitney discs and Spin^c-structures: Given $\mathbf{x}, \mathbf{y} \in \mathbb{T}_\alpha \cap \mathbb{T}_\beta$ we denote by $\pi_2(\mathbf{x}, \mathbf{y})$ the set of homotopy classes of Whitney discs from $\mathbf{x}$ to $\mathbf{y}$. Recall from Example 1.1 that each intersection point $\alpha_j \cap \beta_k$ lies on a unique f-gradient trajectory in Y that connects an index 2 critical point q_k to an index 1-critical point p_j. Thus the point $\mathbf{x} \in \mathbb{T}_\alpha$ can be thought of as a g-tuple of such gradient flow lines connecting each p_j to some q_k. The corresponding 1-chain $\gamma_{\mathbf{x}}$ in Y is called a **simultaneous trajectory**.

When $g > 1$ there is a Whitney disc $\phi : \mathbb{D} \to M_g$ from $\mathbf{x}$ to $\mathbf{y}$ only if the 1-cycle $\gamma_{\mathbf{x}} - \gamma_{\mathbf{y}}$ is null homologous. To see this, consider the commutative

[6] This gives a point of contact with the Seiberg–Witten invariants, which depend for their very definition on the choice of a Spin^c-structure.

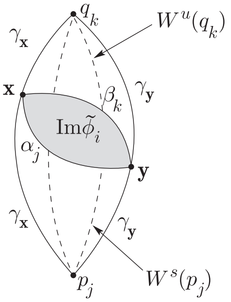


FIGURE 5: WHITNEY DISC GIVES CHAIN IN Y WITH BOUNDARY $\gamma_{\mathbf{x}} - \gamma_{\mathbf{y}}$

diagram

$$\begin{tikzcd} F \arrow[r, "\widetilde{\phi}"] \arrow[d] & \prod \Sigma \arrow[d, "\pi"'] \\ \mathbb{D} \arrow[r, "\phi"] & \mathrm{Sym}^g\Sigma, \end{tikzcd} \tag{2.1}$$

where $F \to \mathbb{D}$ is a suitable (possibly disconnected) branched g-fold cover (the pullback of π by ϕ.) Denote the component functions of $\widetilde{\phi}$ by $\widetilde{\phi}_\ell : F \to \Sigma$. The inverse images of the points $\pm i$ divide the boundary of F into arcs that that are taken by the $\widetilde{\phi}_\ell$ alternately into subarcs of the α and β curves joining the intersections in $\mathbf{x}$ to those in $\mathbf{y}$. Each such subarc in an α_j-curve extends to a triangle in Y consisting of f-gradient flow lines in the stable manifold $W^s(p_j)$. Similarly the subarcs in β_k extend to triangles in the unstable manifolds $W^u(q_k)$, and it is not hard to see that the union of these triangles with the surfaces $\widetilde{\phi}_\ell(F)$ form a 2-chain with boundary $\gamma_{\mathbf{x}} - \gamma_{\mathbf{y}}$: see Fig. 5.

We say that two intersection points $\mathbf{x}, \mathbf{y}$ are equivalent if $\pi_2(\mathbf{x}, \mathbf{y})$ is nonempty. Using the Mayer–Vietoris sequence for the decomposition $Y = Y_1 \cup Y_2$ one can check that the differences $\gamma_{\mathbf{x}} - \gamma_{\mathbf{y}}$ generate $H_1(Y; \mathbb{Z})$. Hence these equivalence classes form an affine space modelled on the finite abelian group $H_1(Y; \mathbb{Z})$. The set of Spin^c structures on Y is also an affine space modelled on $H_1(Y; \mathbb{Z}) \cong H^2(Y; \mathbb{Z})$.

We now explain how the choice of a point $z \in \Sigma$ that does not lie on any α_j or β_k curve determines a natural map

$$s_z : \mathbb{T}_\alpha \cap \mathbb{T}_\beta \to \mathrm{Spin}^c(Y)$$

such that $s_z(\mathbf{x}) = s_z(\mathbf{y})$ iff $\gamma_{\mathbf{x}} - \gamma_{\mathbf{y}} = 0$.

A Spin^c-structure on Y may be thought of as a decomposition of the (trivial) tangent bundle TY into the sum $L \oplus \mathbb{R}$ of a complex line bundle L with a trivial real line bundle,[7] and so corresponds to a nonvanishing vector

[7] A Spin^c-structure on Y is a lift of the structural group SO(3) of the tangent bundle TY to the group $\mathrm{Spin}^c(3) := \mathrm{Spin}(3) \times_{\mathbb{Z}/2\mathbb{Z}} S^1 = \mathrm{SU}(2) \times_{\mathbb{Z}/2\mathbb{Z}} S^1$.

field ξ on Y (a section of $\mathbb{R}$) that is well defined up to *homology*.[8] Therefore to define $s_z(\mathbf{x})$ we just need to associate a nonvanishing vector field $\sigma_{\mathbf{x}}$ to $\mathbf{x}$ that is well defined modulo homology. But z lies on unique f-gradient trajectory γ_z from $\max f$ to $\min f$. This, together with the simultaneous trajectory $\gamma_{\mathbf{x}}$, pairs up the set of critical points of f. Since each pair has index sum 3, the gradient vector field ∇f of f can be modified near these trajectories to a nonvanishing vector field $\sigma_{\mathbf{x}}$. Then $\sigma_{\mathbf{x}} = \nabla f$ outside a union of 3-balls and so is well defined up to homology. We therefore set

$$s_z(\mathbf{x}) = [\sigma_{\mathbf{x}}] \in \mathrm{Spin}^c(Y).$$

Definition of $CF^\infty(Y, \mathfrak{s})$: Given a Spin^c-structure $\mathfrak{s}$, denote by $\mathcal{S} \subset \mathbb{T}_\alpha \cap \mathbb{T}_\beta$ the corresponding set of intersection points. We define $CF^\infty(Y, \mathfrak{s})$ to be the free abelian group with generators $[\mathbf{x}, i] \in \mathcal{S} \times \mathbb{Z}$ and with relative grading

$$\mathrm{gr}\big([\mathbf{x}, i], [\mathbf{y}, j]\big) := \mu(\phi) - 2(i - j + n_z(\phi)).$$

Here ϕ is any Whitney disc from $\mathbf{x}$ to $\mathbf{y}$, $n_z(\phi)$ is its intersection number with the generator $\{z\} \times \mathrm{Sym}^{g-1}(\Sigma)$ of $H_{2n-2}(M_g)$ and $\mu(\phi)$ is its Maslov index, that is, the expected dimension of the set $\mathcal{M}(\mathbf{x}, \mathbf{y}; \phi)$ of all components of the trajectory space $\mathcal{M}(\mathbf{x}, \mathbf{y})$ that contain elements homotopic to ϕ. One can show that the number $\mu(\phi) - 2n_z(\phi)$ is independent of the choice of ϕ. We then define the boundary operator δ^∞ by:

$$\delta^\infty([\mathbf{x}, i]) = \sum_{\mathbf{y} \in \mathcal{S}} \sum_{\phi \in \pi_2(\mathbf{x}, \mathbf{y}) : \mu(\phi) = 1} n(\mathbf{x}, \mathbf{y}; \phi)\ [\mathbf{y}, i - n_z(\phi)],$$

where $n(\mathbf{x}, \mathbf{y}; \phi)$ denotes the (signed) number of elements in

$$\widehat{\mathcal{M}}(\mathbf{x}, \mathbf{y}; \phi) := \mathcal{M}(\mathbf{x}, \mathbf{y}; \phi)/\mathbb{R}.$$

For the reasons outlined in Example 1.4, $(\delta^\infty)^2 = 0$. Hence $CF^\infty(Y, \mathfrak{s})$ is a chain complex.

Definition of $CF^\pm(Y, \mathfrak{s})$ and $\widehat{CF}(Y, \mathfrak{s})$: Since the submanifold $\{z\} \times \mathrm{Sym}^{g-1}(\Sigma)$ is a complex hypersurface, any holomorphic trajectory meets it positively. In other words, $n_z(\phi) \geq 0$ whenever $\mathcal{M}(\mathbf{x}, \mathbf{y}; \phi)$ is nonempty. Therefore the subset $CF^-(Y, \mathfrak{s})$ generated by the elements $[\mathbf{x}, i]$ with $i < 0$ forms a subcomplex of $CF^\infty(Y, \mathfrak{s})$. We define $CF^+(Y, \mathfrak{s})$ to be the quotient $CF^\infty(Y, \mathfrak{s})/CF^-(Y, \mathfrak{s})$, i.e. the complex generated by $[\mathbf{x}, i]$, $i \geq 0$. All three complexes are $Z[U]$-modules where U acts by

$$U \cdot [\mathbf{x}, i] = [\mathbf{x}, i - 1],$$

reducing grading by 2. Finally we define $\widehat{CF}(Y, \mathfrak{s})$ to be the complex generated by the kernel of the U-action on $CF^+(Y, \mathfrak{s})$. Thus we may think of

[8] Two nonvanishing vector fields are called homologous if one can be homotoped through nonvanishing vector fields to agree with the other except on a finite union of 3-balls.

$\widehat{CF}(Y, \mathfrak{s})$ as generated by the elements $\langle \mathbf{x} \rangle, \mathbf{x} \in \mathcal{S}$, with differential

$$\widehat{\partial} \langle \mathbf{x} \rangle = \sum_{\mathbf{y}} \sum_{\phi \in \pi_2(\mathbf{x}, \mathbf{y}) : \mu(\phi) = 1, n_z(\phi) = 0} n(\mathbf{x}, \mathbf{y}; \phi) \langle \mathbf{y} \rangle,$$

i.e. we count only those trajectories that do not meet $\{z\} \times \mathrm{Sym}^{g-1}(\Sigma)$. The corresponding homology groups are related by exact sequences
(2.2)

$$\begin{array}{ccccccc} \ldots \longrightarrow & HF^-(Y, \mathfrak{s}) & \xrightarrow{i} & HF^\infty(Y, \mathfrak{s}) & \xrightarrow{\pi} & HF^+(Y, \mathfrak{s}) & \xrightarrow{\delta} \ldots \\ \ldots \longrightarrow & \widehat{HF}(Y, \mathfrak{s}) & \xrightarrow{j} & HF^+(Y, \mathfrak{s}) & \xrightarrow{U} & HF^+(Y, \mathfrak{s}) & \longrightarrow \ldots . \end{array}$$

There is yet another interesting group, namely $HF_{\mathrm{red}}(Y, \mathfrak{s})$, the cokernel of the above map π. This vanishes for the 3-sphere and for lens spaces. Later, we will use the fact that there is a pairing $HF^+ \otimes HF^- \to \mathbb{Z}$, that induces a pairing

$$\langle \cdot, \cdot \rangle : HF_{\mathrm{red}} \otimes HF_{\mathrm{red}} \to \mathbb{Z}. \tag{2.3}$$

The following result is proved in [5].

Theorem 2.1. *Each of these relatively $\mathbb{Z}$-graded $\mathbb{Z}[U]$-modules is a topological invariant of the pair $(Y, \mathfrak{s})$.*

The proof that these homology groups are independent of the choice of almost complex structure j on Σ, of isotopy class of the loops α_i, β_j and of basepoint z, uses fairly standard arguments from Gromov–Witten–Floer theory. To see that they remain unchanged under handle slides of the curves in α, β one uses the naturality properties of Lagrangian Floer homology, defining a chain map by counting suitable holomorphic triangles. Finally the fact that they are invariant under the stabilization of the Heegaard splitting uses a "stretch the neck" argument.

At first glance it is not at all clear why one needs such a variety of homology groups. However, if we ignore the action of U and consider only HF^∞ we get very little information. Thus, for example, it is shown in [5] that when Y is a homology 3-sphere

$$HF^\infty(Y, \mathfrak{s}) \cong \mathbb{Z}[U, U^{-1}],$$

for all choices of $\mathfrak{s}$.[9] In fact the different complexes HF are just ways of encoding the subtle information given by the basepoint z. For they may all be defined in terms of the chain complex $CF^-(Y, \mathfrak{s})$ of $\mathbb{Z}[U]$-modules:

- $CF^\infty(Y, \mathfrak{s})$ is the "localization" $CF^-(Y, \mathfrak{s}) \otimes \mathbb{Z}[U, U^{-1}]$,
- $CF^+(Y, \mathfrak{s})$ is the cokernel of the localization map, and
- $\widehat{CF}(Y, \mathfrak{s})$ is the quotient $CF^-(Y, \mathfrak{s}) / U \cdot CF^-(Y, \mathfrak{s})$.

[9] A similar phenomenon occurs in the Hamiltonian Floer theory of the loop space of a symplectic manifold M. The resulting homology groups $FH_*(M; H, J)$ are always (additively) isomorphic to the homology of M, but one gets interesting information by filtering by the values of the action functional.

This terminology is not merely fanciful. In the conjectured equivalence between this and the Seiberg–Witten–Floer theory of Y, which is an S^1-equivariant theory, the element U corresponds to the generator of $H_2(BS^1)$, although the underlying geometric reason for this is not yet understood: see Lee[3].

Example 2.2. Consider the case $Y = S^3$. We saw in Example 1.1 that this has a Heegaard splitting consisting of a torus $\mathbb{T}^2$, with a single α and a single β curve intersecting once transversally at $\mathbf{x}$. Denote by $\mathfrak{s}$ the unique Spin^c-structure on S^3. Then the complex $CF^-(S^3, \mathfrak{s})$ has generators $[\mathbf{x}, i], i < 0$, and trivial boundary map (this has to vanish since the relative gradings are even). This determines all the other groups; for example, $\widehat{HF}(S^3) \cong \mathbb{Z}$. There are many other Heegaard splittings for S^3. Ozsvath–Szabo give an example in [8, §2.2] of a genus 2 splitting where the differential ∂ *depends on the choice of complex structure* j on Σ_2. This might seem paradoxical. The point is that the differential is given by counting holomorphic discs in $\mathrm{Sym}^g(\Sigma)$, but as in diagram (2.1) these correspond to counting images in Σ of some branched cover F of the disc, and these images can have nontrivial moduli. This shows that Heegaard–Floer theory is not entirely combinatorial: the next big advance might be the construction of combinatorial invariants, possibly similar to Khovanov's new knot invariants.

One can make various additional refinements to the theory. For example, when Y is a rational homology sphere it is possible to lift the relative $\mathbb{Z}$-grading to an absolute $\mathbb{Q}$-grading that is respected by the naturality maps we discuss below: see [§3.2][8]. Ozsvath–Szabo also define knot invariants in [7] and use them to give a new obstruction for a knot to have unknotting number one.

2.2. **Properties and Applications of the invariants.**

The power of Heegaard–Floer theory comes from the fact that it is well adapted to certain natural geometric constructions in 3-manifold theory, such as adding a handle or performing a Dehn surgery on a knot, because these have simple descriptions in terms of Heegaard diagrams. Here is the basic geometric construction.

Suppose given three sets α, β, γ of g disjoint curves on the Riemann surface Σ_g that are the attaching circles for the handlebodies $U_\alpha, U_\beta, U_\gamma$. Then there are three associated manifolds

$$Y_{\alpha,\beta} = U_\alpha \cup U_\beta, \quad Y_{\alpha,\gamma} = U_\alpha \cup U_\gamma, \quad Y_{\beta,\gamma} = U_\beta \cup U_\gamma.$$

We now construct a 4-manifold $X = X_{\alpha\beta\gamma}$ with these three manifolds as boundary components. Let Δ be a triangle (or 2-simplex) with vertices $v_\alpha, v_\beta, v_\gamma$ and edges $e_\alpha, e_\beta, e_\gamma$ (where e_α lies opposite v_α), and form X from the four pieces

$$\big(\Delta \times \Sigma\big) \sqcup \big(e_\alpha \times U_\alpha\big) \sqcup \big(e_\beta \times U_\beta\big) \sqcup \big(e_\gamma \times U_\gamma\big)$$

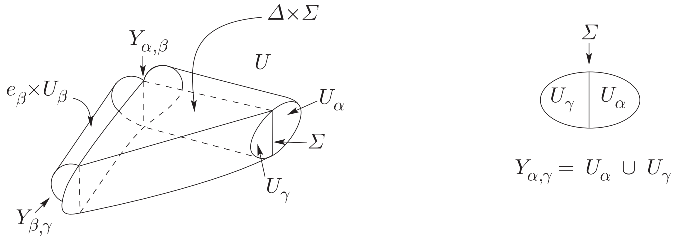


Figure 6: The 4-manifold $X_{\alpha\beta\gamma}$ (with Σ represented as an interval)

by making the obvious identifications along $\partial\Delta\times\Sigma$ and then smoothing. For example, the part $e_\alpha\times\Sigma$ of $\partial\Delta\times\Sigma$ is identified with $e_\alpha\times\partial U_\alpha$: see Fig. 6. The resulting manifold has three boundary components, one corresponding to each vertex, with $Y_{\alpha,\beta}$ lying over $v_\gamma = e_\alpha\cap e_\beta$ for example. One can orient X so that

$$\partial X = -Y_{\alpha,\beta} - Y_{\beta,\gamma} + Y_{\alpha,\gamma}.$$

This elementary cobordism is called a **pair of pants** cobordism. Counting holomorphic triangles in M_g with boundaries on the three tori $\mathbb{T}_\alpha, \mathbb{T}_\beta, \mathbb{T}_\gamma$ gives rise under good circumstances to a map

$$(2.4)\qquad f^\infty : CF^\infty(Y_{\alpha,\beta}, \mathfrak{s}_{\alpha,\beta}) \otimes CF^\infty(Y_{\beta,\gamma}, \mathfrak{s}_{\beta,\gamma}) \to CF^\infty(Y_{\alpha,\gamma}, \mathfrak{s}_{\alpha,\gamma}).$$

(Here the Spin^c structures are assumed to extend to a common Spin^c structure $\mathfrak{s}$ on X.)

There are some interesting special cases of this construction. For example, it can be used to obtain a **long exact surgery sequence** which is very useful in analysing the effect of rational Dehn surgeries on Y. Here we shall concentrate on explaining some of the corresponding naturality properties of the theory.

Maps induced by cobordisms: Suppose that Y_2 is obtained from Y_1 by doing a 0**-surgery along a framed knot** K. This means that we choose an identification[10] of a neighborhood $N(K)$ with $S^1\times D^2$, attach one part $\partial D^2\times D^2$ of the boundary of the 4-ball $D^2\times D^2$ to Y_1 via the obvious map

$$\psi : \partial D^2\times D^2 \to S^1\times D^2 \equiv N(K) \subset Y_1,$$

and then define Y_2 to be a smoothed out version of the union

$$Y_2 = \Big(Y_1 \setminus \operatorname{int} N(K)\Big) \sqcup_\psi (D^2\times S^1),$$

where ψ identifies the boundary torus $S^1\times S^1$ in $D^2\times S^1$ to $\partial N(K)$. Note that the 4-manifold $W = \big([0,1]\times Y_1\big) \sqcup_\psi (D^2\times D^2)$ is a cobordism from Y_1 to Y_2 obtained by adding a 2-handle to $[0,1]\times Y_1$ along $N(K)\subset\{1\}\times Y_1$.

[10] This is called a **framing** of the knot. It corresponds to choosing a pair of linearly independent vector fields along K that trivialize its normal bundle. Note that any knot in S^3 has a canonical framing: because $H_2(S^3) = 0$, K bounds an embedded surface S in S^3 and one can choose the first vector field to be tangent to S.

Given a knot K in Y_1, one can always choose a Heegaard diagram (Σ, α, β) for Y_1 so that K lies in the surface $\Sigma - \beta_2 - \cdots - \beta_g$ (and is given the obvious framing) and intersects β_1 once transversally. Pushing K into U_β, one sees that this is equivalent to requiring that K is disjoint from the discs D_j with boundary β_j for $j > 1$ and meets D_1 transversally in a single point. Hence one can construct a suitable diagram by starting with a neighborhood $N(K)$ of the knot and then adding 1-handles to obtain U_β. Since doing 0-surgery along K adds a disc with boundary K, it is easy to check that

$$Y_2 = Y_{\alpha,\gamma}, \quad \gamma := \{K, \beta_2, \dots, \beta_g\}.$$

Further $Y_{\beta,\gamma}$ is a connected sum $\#(S^2 \times S^1)$ of copies of $S^2 \times S^1$, and so is standard. Pairing the map f^∞ in equation (2.4) with a canonical element in $HF^\infty(Y_{\beta,\gamma})$ one obtains a map

$$f^\infty : HF^\infty(Y_1; \mathfrak{s}_1) \to HF^\infty(Y_2; \mathfrak{s}_2)$$

for suitable $\mathfrak{s}_i$.

This construction can be extended to any cobordism.

Lemma 2.3. *Suppose that X is an oriented connected cobordism from Y_1 to Y_2, where each Y_i is an oriented connected 3-manifold. Then, for each* Spin^c *structure $\mathfrak{s}$ on X there is a natural induced map*

$$F^\infty_{X,\mathfrak{s}} : HF^\infty(Y_1, \mathfrak{s}_1) \to HF^\infty(Y_2, \mathfrak{s}_2),$$

where $\mathfrak{s}_i$ is the restriction of $\mathfrak{s}$ to Y_i.

There are corresponding maps for the other groups $HF^\pm, \widehat{HF}$ and so on. All of them have the obvious functorial properties, behaving well for example under compositions of cobordisms. Another important property is that the image of the induced map

$$F^-_{X,\mathfrak{s}} : HF^-(Y_1, \mathfrak{s}_1) \to HF^-(Y_2, \mathfrak{s}_2)$$

is contained in HF_{red} if $b_2^+(X) \ge 1$. (This is the first appearance so far of this condition[11] on b_2^+ which is so ubiquitous in Seiberg–Witten theory.) In other words, $\operatorname{im} F^-_{X,\mathfrak{s}}$ is contained in the image of the boundary map $\delta : HF^+ \to HF^-$ in the long exact sequence (2.2).

A 4-manifold invariant: We now define an invariant $\Phi_{X,\mathfrak{s}}$ of a closed connected 4-manifold X with Spin^c-structure $\mathfrak{s}$. Conjecturally it agrees with the Seiberg–Witten invariant. Its construction illustrates the use of the different groups HF.

Suppose that X is a closed connected 4-manifold with $b_2^+(X) > 1$ and Spin^c-structure $\mathfrak{s}$. (For example, any symplectic 4-manifold has a canonical Spin^c-structure.) An **admissible cut** of X is a decomposition of X into

[11] Given a connected, oriented 4-manifold X, $b_2^+(X)$ is the number of positive squares in the diagonalization of the cup product pairing on $H^2(X, \partial X)$. The relevant fact here is that when $b_2^+(X) \ge 1$ there is a closed surface C in X with self-intersection $C \cdot C \ge 0$.

two pieces X_1, X_2, each with $b_2^+(X_i) \geq 1$, along a 3-manifold $Y := X_1 \cap X_2$. We assume also that the restriction map

$$H^2(X) \to H^2(X_1) \oplus H^2(X_2)$$

is injective. Delete small 4-balls from the interior of each piece X_i and consider them as giving cobordisms from S^3 to Y. Then, for a certain canonical element $\theta \in HF^-(S^3)$, consider

$$\Phi_{X,\mathfrak{s}} := \Big\langle \delta^{-1} \circ F^-_{X_1,\mathfrak{s}_1}\,\theta, F^-_{X_2,\mathfrak{s}_2}\theta \Big\rangle,$$

where we use the pairing (2.3) on $HF_{\mathrm{red}}(Y, \mathfrak{s})$. This element turns out to be independent of choices and nonzero for symplectic manifolds. (In this case it can be calculated using a decomposition of X coming from a Donaldson–Lefschetz pencil.) Hence in any admissible cut of a symplectic manifold, $HF_{\mathrm{red}}(Y)$ must be nonzero. Ozsvath–Szabo conclude in [6] that:

Proposition 2.4. *A connected closed symplectic* 4*-manifold X has no admissible cut $X = X_1 \cup X_2$ such that $Y := X_1 \cap X_2$ has $HF_{\mathrm{red}}(Y, \mathfrak{s}) = 0$ for all $\mathfrak{s}$.*

The first proof of this was in the case $Y = S^3$ and is due to Taubes. He combined the well known fact that gauge theoretic invariants vanish on connected sums together with his proof that the Seiberg–Witten invariants do not vanish on symplectic 4-manifolds.

Rational homology 3-spheres Y for which HF^+ has no torsion and where $HF_{\mathrm{red}}(Y, \mathfrak{s}) = 0$ for all $\mathfrak{s}$ are called **L-spaces** in [8, §3.4]. All lens spaces are L-spaces, but not all Brieskorn homology spheres are: $\Sigma(2, 3, 5)$ is an L-space, but $\Sigma(2, 3, 7)$ is not. The class of L-spaces is not yet fully understood, but it has interesting geometric properties. For example it follows from the above proposition that L-spaces do not support any taut foliations, i.e. foliations in which the leaves are minimal surfaces for some Riemannian metric on Y; for if Y supports such a foliation then results of Thurston, Eliashberg, Giroux and Etnyre about contact structures allow one to construct a symplectic manifold X that has an admissible cut with $Y = X_1 \cap X_2$.

As a final corollary, we point out that similar arguments imply that Heegaard–Floer theory can **detect the unknot** in S^3. This means the following. Suppose that K is a knot in S^3 and denote by $S^3_0(K)$ the result of doing 0-surgery along K with the canonical framing described above.

Corollary 2.5. *If $HF(S^3_0(K)) = HF(S^3_0(\text{unknot}))$ then K is the unknot.*

Sketch of proof. Let $Y = S^3_0(K)$. Suppose that $Y \neq S^3_0(\text{unknot}) = S^2 \times S^1$. By a deep result of Gabai, Y admits a taut foliation. As above, this implies that $HF_{\mathrm{red}}(Y, \mathfrak{s}) \neq 0$ for some $\mathfrak{s}$. But $HF_{\mathrm{red}}(S^2 \times S^1, \mathfrak{s})$ is always 0. □

References

[1] R. Bott, Morse theory indomitable, *Publi. I.H.E.S.*

[2] R. T. Cohen, Morse theory, graphs and string topology, GT/0411272.

[3] Yi-Jen Lee, Heegaard splittings and Seiberg–Witten monopoles, GT/0409536.

[4] J. Milnor. *Morse Theory*, Annals of Math. Studies #51, Princeton Univ Press.

[5] P. Ozsvath and Z. Szabo, Holomorphic discs and topological invariants for rational homology three-spheres, SG/0101206, *Ann. Math.*

[6] P. Ozsvath and Z. Szabo, Holomorphic disc invariants for symplectic four manifolds, SG/0210127.

[7] P. Ozsvath and Z. Szabo, Knots with unknotting number one and Heegaard Floer homology, GT/0401426.

[8] P. Ozsvath and Z. Szabo, Heegaard diagrams and holomorphic discs, GT/0403029.

[9] P. Seidel, Fukaya categories and deformations, SG/0206155.

[10] E. Witten, Supersymmetry and Morse theory, *J. Diff. Geo.* **17** (1982) 661–692.

Department of Mathematics, Stony Brook University, Stony Brook, NY 11794-3651, USA

E-mail address: dusa@math.sunysb.edu

URL: http://www.math.sunysb.edu/ dusa

New Methods in Celestial Mechanics and Mission Design

Jerrold E. Marsden
Control and Dynamical Systems
Caltech 107-81, Pasadena, CA 91125
marsden@cds.caltech.edu

Shane D. Ross
Department of Aerospace and Mechanical Engineering
University of Southern California, RRB 217
Los Angeles, CA 90089-1191
shane@cds.caltech.edu

13 December 2004

Abstract

The title of this paper comes from Poincaré, who introduced many key dynamical systems methods through his study of celestial mechanics and especially the three body problem. Since then, many researchers have contributed to his legacy by developing and applying these methods to problems in celestial mechanics and, more recently, with the design of real space missions. This paper will give a survey of some of these exciting ideas. In an upcoming monograph Koon, Lo, Marsden, and Ross [2005], this approach and its application to real missions is discussed in detail.

One of the key ideas is that the competing gravitational pull between celestial bodies creates a vast array of passageways that wind around the sun, planets and moons. These passageways are realized geometrically as invariant manifolds attached to equilibrium points and periodic orbits in interlinked three body problems. In particular, tube-like structures form an interplanetary transport network which will facilitate the exploration of Mercury, the Moon, the asteroids, and the outer solar system, including a mission to assess the possibility of life on Jupiter's icy moons.

1 Astrodynamics and Dynamical Astronomy

Astrodynamics and dynamical astronomy apply the principles of mechanics, including the law of universal gravitation to the determination of the motion of objects in space. Orbits of astronomical bodies, such as planets, asteroids, and comets are calculated, as are spacecraft trajectories, from launch through atmospheric re-entry, including all the needed orbital maneuvers.

While there are no sharp boundaries, astrodynamics has come to denote primarily the design and control of spacecraft trajectories, while dynamical astronomy is concerned with the motion of other bodies in the solar system (origin of the moon, Kuiper belt objects, etc). From a dynamical systems perspective of interest to us, it is quite useful to mix these subjects. There is one obvious commonality: the model used for studying either a spacecraft or, say, the motion of an asteroid is the restricted $N+1$ body problem, where N celestial bodies move under the influence of one another and the spacecraft or asteroid moves in the field of these bodies, but has a mass too small to influence their motion.

The Ephemeris and Its Approximations. In the case of motion within the solar system, the motion of the N bodies (planets, moons, etc) can be measured and predicted to great accuracy, producing an ephemeris. An *ephemeris* is simply a listing of positions and velocities of celestial bodies as a function of time with respect to some coordinate system. An ephemeris can be considered as the solution of the N-body gravitational problem, and forms the gravitational field which determines a spacecraft or asteroid's motion.

While the final trajectory design phase of a space mission or the long term trajectory of an asteroid will involve a solution considering the most accurate ephemeris, insight can be achieved by considering simpler, approximate ephemerides (the plural of ephemeris). An example of such an ephemeris is a simplified solution of the N-body problem, where N is small, for example, the motion of the Earth and Moon under their mutual gravitation, a two-body solution. The simplest two-body solution of massive bodies which gives rise to interesting motion for a spacecraft is the circular motion of two bodies around their common center of mass. The problem of the spacecraft's motion is then known as the *circular restricted three-body problem*, or the CR3BP.

Introduction to the Trajectory Design Problem. The set of possible spacecraft trajectories in the three-body problem can be used as building blocks for the design of spacecraft trajectories in the presence of an arbitrary number of bodies. Consider the situation shown in Figure 1.1, where we have a spacecraft, approximated as a particle, P, in the gravitational field of N massive bodies. We assume P has a small enough mass that it does not influence the motion of the N massive bodies, which move in prescribed orbits under their mutual gravitational attraction. In the solar system, one can think of a moon, M_2, in orbit around a planet, M_1, which is in orbit around the Sun, M_0.

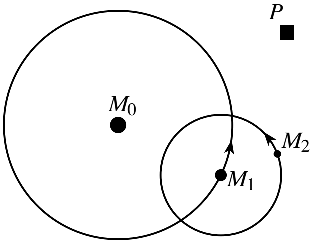


Figure 1.1: A spacecraft P in the gravitational field of N massive bodies which move in prescribed orbits.

The goal of trajectory design is to find a transfer trajectory, such as the one shown in Figure 1.2(a), which takes the spacecraft from a

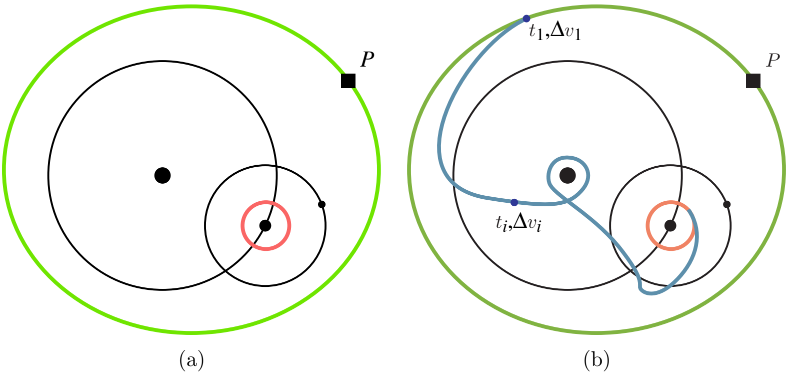


Figure 1.2: (a) The goal is to find a transfer trajectory which takes the spacecraft from an initial orbit to a final orbit using controls. (b) Assuming impulsive controls, i.e., several instantaneous changes in the spacecraft's velocity, with norm Δv_i at time t_i, we can effect such a transfer.

prescribed initial orbit to a prescribed final orbit using controls. The initial orbit may be an orbit around the Earth and the final orbit an orbit around one of the moons of Jupiter, for instance. To effect this transfer, we could use high thrust or low thrust propulsion systems. In the low thrust case, we have a small continuous control which can operate at all times. In the high thrust case, we assume that the control is discretized into several instantaneous changes in the spacecraft's velocity. These instantaneous changes have a magnitude at time t_i that is traditionally denoted Δv_i. Under a high thrust assumption, the Δv's are proportional to the fuel consumption:

$$\Delta v = -v_e \frac{\Delta m}{m}$$

where m is the mass of the rocket and Δm is the mass of propellant ejected at an exhaust velocity v_e (Roy [1988]). As spacecraft are limited in the amount of fuel that they can carry on-board for their journey, we often want to consider an *optimal control problem*: minimize the fuel consumed (equivalently, energy). In other words, we want to find the maneuver times t_i and sizes Δv_i to minimize

$$\sum_i \Delta v_i,$$

the total change in velocity, or "ΔV" as it is called.

It is typical in space missions to use the magnitude of the required ΔV as a measure of the spacecraft fuel performance. The propellant mass is a much less stable quantity as a measure of spacecraft performance, since it is dependent on the spacecraft mass and various other parameters which change frequently as the spacecraft is being built. The ΔV comes from astrodynamics considerations only and is independent of the mass and type of spacecraft. Thus, for a given mission objective, one generally wants to *minimize* ΔV.

2 The Patched Three-Body Approximation

To get a spacecraft from, say, Earth to other parts of the solar system, it is necessary to find solutions for the motion of the spacecraft under the influence of N bodies, a notoriously difficult problem. Furthermore, one needs to find solutions with a desired behavior, e.g., flying by the giant outer planets as Voyagers 1 and 2 did, while satisfying engineering constraints, e.g., low fuel consumption, short time of flight, low radiation dose, etc.

Patched-Conic Approach and the Voyager Trajectory. For many purposes it is satisfactory to simplify the general trajectory problem by considering the gravitational force between the spacecraft and only one other body at a time. Even for the case of interplanetary transfer, this simplification will suffice for many calculations. That is, one may consider escape from or capture by a planet to be an interaction between the spacecraft and that particular planet alone, whereas the transfer process is considered an interaction between the spacecraft and the Sun alone. NASA's spectacular multiple flyby missions such as Voyager and Galileo are based on this Keplerian decomposition of the solar system, known as the *patched-conic approximation* (or *patched two-body approximation*), discussed in Bate, Mueller, and White [1971].

The strategy of the designers of the Voyager missions was to initially approximate the full N-body solution of the spacecraft's motion as a linkage of several two-body solutions, the well known conic solutions discovered by Kepler. The spacecraft's trajectory as it coasted between two planets was considered as a heliocentric hyperbolic trajectory. The heliocentric trajectory was cleverly chosen to come close to the destination planet, in order to fly by it. When the spacecraft came within the "sphere of influence"[1] of a planet, it was considered as a hyperbolic conic section trajectory centered on the planet. This patched-conic solution could be used as an initial guess for a numerical procedure which produced a fully integrated N-body solution.

High vs. Low Relative Velocities. For missions such as Voyager and Galileo, the speed of the spacecraft relative to the bodies is high and therefore the time during which the acceleration on the spacecraft due to two bodies is comparable is very short, and results in a minor perturbation away from a conic solution. But when one needs to deal with the unpropelled, or ballistic, capture regime of motion,[2] where the relative speed is low, a three-body decomposition of the solar system is necessary.

Some Missions Cannot Be Approximated by the Patched-Conic Approach. For Voyager and Galileo, the patched-conic approach worked very well. But as space missions have become more demanding, other approaches have become necessary. For example, the Genesis, L_1 Gateway, and multi-moon orbiter trajectories discussed below resemble solutions of the restricted three- and four-body problems much more than two-body problems. In fact, methods based on a patched-conic approximation would have a very difficult time finding these complicated trajectories, as they are fundamentally non-Keplerian, restricted N-body solutions.

[1] The sphere of influence of a planet is the radius at which the acceleration on a spacecraft due to the planet and the Sun are approximatedly equal (Roy [1988]).

[2] *Ballistic capture* means that no propulsion is necessary (i.e., no ΔV) to achieve a capture orbit at the destination body. In general, this "capture" is temporary.

Taking Better Advantage of N-Body Dynamics. It is possible to satisfy mission constraints using spacecraft solutions which do not take advantage of the N-body dynamics of a system. But this may require using more fuel than is necessary.[3] Worse yet, because of the fuel restrictions on interplanetary spacecraft, some missions may not be possible if only a patched-conic approach is used. An interesting example in this category, which also served as motivation for much of our group's work, is the "rescue" of a malfunctioned Japanese space mission to the moon by Belbruno and Miller of JPL in June, 1990. The mission originally had two spacecraft, MUSES-A and MUSES-B; B was to go into orbit around the moon, with A remaining in earth orbit as a communications relay. But B failed and A did not have sufficient fuel to make the journey. However, by utilizing a trajectory concept originally discovered by Belbruno in 1986, which is more energy-efficient than the one planned for B, MUSES-A (renamed Hiten) left Earth orbit in April, 1991 and reached the moon that October. As a result, Japan became the third nation to send a spacecraft to the moon. After a series of scientific experiments, Hiten was purposely crashed into the Moon in April, 1993. See Belbruno [2004] for additional details of this fascinating story.

An ESA (European Space Agency) mission currently underway, SMART-1, which is a mission from the Earth to the Moon, also uses some of these same ideas; see `http://sci.esa.int/science-e/www/area/index.cfm?fareaid=10`.

A Hierarchy of Models. We want to make use of the natural dynamics in the solar system as much as possible; that is, we wish to take advantage of the phase space geometry, integrals of motion, and lanes of fast unpropelled travel. We envision generating a trajectory via a hierarchy of models. One starts with simple models which capture essential features of natural dynamics. One then uses simple model solutions as initial guess solutions in more realistic models. The approach described above does this conceptually, using the patched-conic approximation to generate the first guess solution. But there are regimes of motion where conics are simply not a good approximation to the motion of the spacecraft. There is much to be gained by starting with not two-body solutions, but three-body solutions to the spacecraft's motion.

The Patched Three-Body Approximation. Motivated by the Belbruno and Miller work, we consider a restricted four-body problem

[3] For example, Dunn [1962] proposed to use a satellite for lunar far side communications by placing it in a position where it would requre approximately 1500 m/s per year for stationkeeping. A few years later, Farquhar [1966] proposed a trajectory for the same mission which used only 100 m/s per year by taking advantage of three-body dynamics.

wherein a spacecraft moves under the influence of three massive bodies whose motion is prescribed, as shown schematically in Figure 1.1. For Belbruno and Miller, these four bodies were the Sun, the Earth, the Moon and the spacecraft.

To begin with, we restrict the motion of all the bodies to a common plane, so the phase space is only four-dimensional. As in the patched-conic approach, the patched three-body approach uses solutions obtained from two three-body problems as an initial guess for a numerical procedure which converges to a full four-body solution.

As an example of such a problem where there is no control, consider the four-body problem where two adjacent giant planets compete for control of the same comet (e.g., Sun-Jupiter-comet and Sun-Saturn-comet). When close to one of the planets, the comet's motion is dominated by the corresponding planet's 3-body dynamics. Between the two planets, the comet's motion is mostly heliocentric and Keplerian, but is precariously poised between two competing three-body dynamics, leading to complicated transfer dynamics between the two adjacent planets.

When we consider a spacecraft with control instead of a comet, we can intelligently exploit the transfer dynamics to construct low energy trajectories with prescribed behaviors, such as transfers between adjacent moons in the Jovian and Saturnian systems (Lo and Ross [1998]). For example, by approximating a spacecraft's motion in the $N+1$ body gravitational field of Jupiter and N of its planet-sized moons into several segments of purely three body motion—involving Jupiter, the ith moon, and the spacecraft—we can design a trajectory for the spacecraft which follows a prescribed itinerary in visiting the N moons. In an earlier study of a transfer from Ganymede to Europa, we found our fuel consumption for impulsive burns, as measured by the total norm of velocity displacements, ΔV, to be less than half the Hohmann transfer value (Koon, Lo, Marsden, and Ross [1999]). We found this to be the case for the following example multi-moon orbiter tour shown schematically in Figure 2.1: starting beyond Ganymede's orbit, the spacecraft is ballistically captured by Ganymede, orbits it once, escapes in the direction of Europa, and ends in a ballistic capture at Europa.

One advantage of this multi-moon orbiter approach as compared with the Voyager-type flybys is the "leap-frogging" strategy. In this new approach to mission design, the spacecraft can orbit a moon for a desired number of circuits, escape the moon, and then perform a transfer ΔV to become ballistically captured by a nearby moon for some number of orbits about that moon, etc. Instead of brief flybys lasting only seconds, a scientific spacecraft can orbit several different moons for any desired duration. Furthermore, the total ΔV necessary is *much less than that necessary using purely two-body motion segments.* One can also systematically construct low energy transfers from the Earth to

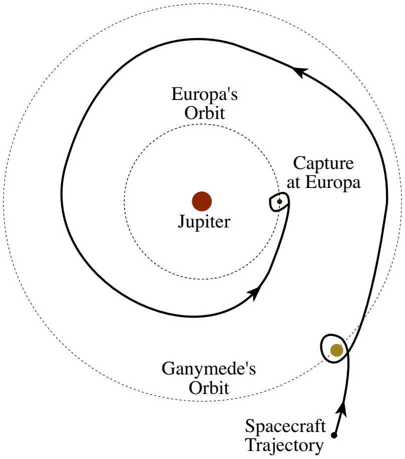


Figure 2.1: Leap-frogging mission concept: a multi-moon orbiter tour of Jupiter's moons Ganymede and Europa.

the Moon using the Sun's perturbation, and from lunar libration point orbits to Earth libration point orbits.

Three-Body Dynamics. To patch three-body solutions (the spacecraft's motion in the presence of two bodies), one needs a good handle on what those solutions are. Studying the CR3BP solutions in detail is an interesting topic in its own right. This is a topic that goes back to the basic work of Poincaré in the late 1800s and provided the context in which he developed modern dynamical systems theory and the notion of chaos.

In the CR3BP, we thus have two primaries that move in circles; the smaller third body moves in the gravitational field of the primaries (without affecting them). We typically view the motion in a rotating frame so that the primaries appear stationary. It is important to consider both the planar and the spatial problems, but we shall focus on the planar problem for the moment.

One may derive the equations of motion using a little elementary mechanics as follows. Let the masses of the two primaries be denoted m_1 and m_2 and set $\mu = m_2/(m_1 + m_2)$. We can normalize the distance between the primaries to be unity and then in the rotating frame, normalized to rotate with unit angular velocity, the two bodies may be located on the x-axis at the points $(-\mu, 0)$ and $(1 - \mu, 0)$. Let the position of the third body be denoted (x, y) in the rotating frame. The kinetic energy of this third body (whose mass we take to be unity) with

with respect to an inertial frame but written in a frame rotating with unit angular velocity is the usual $\frac{1}{2}mv^2$ expression:

$$K(x, y, \dot{x}, \dot{y}) = \frac{1}{2}\left[(\dot{x} - y)^2 + (\dot{y} + x)^2\right]$$

Let r_1 be the distance from the third body to the first primary; that is, $r_1 = \sqrt{(x+\mu)^2 + y^2}$ and let r_2 be the distance to the second primary, that is, $r_2 = \sqrt{(x - 1 + \mu)^2 + y^2)}$. Then the gravitational potential energy of the third body is, again in normalized units,

$$V(x, y) = -\frac{1-\mu}{r_1} - \frac{\mu}{r_2}.$$

The Lagrangian of the third body is its kinetic minus potential energies, namely

$$L(x, y, \dot{x}, \dot{y}) = K(x, y, \dot{x}, \dot{y}) - V(x, y);$$

Now one gets the equations of motion simply by writing down the corresponding Euler-Lagrange equations:

$$\ddot{x} - 2\dot{y} = -\frac{\partial \overline{V}}{\partial x}, \qquad \ddot{y} + 2\dot{x} = -\frac{\partial \overline{V}}{\partial y} \tag{2.1}$$

where the *effective potential* is

$$\overline{V} = V - \frac{x^2 + y^2}{2}$$

Being Euler–Lagrange equations, there is a conserved energy that one computes via the Legendre transformation to be

$$E = \frac{1}{2}\left(\dot{x}^2 + \dot{y}^2\right) + \overline{V}(x, y).$$

Equilibria. These occur when the the third body moves in a circular orbit with the same frequency as the primaries, so that it is stationary in the rotating frame. We find these points by finding the equilibrium points, in the standard sense of ode's, of the equations (2.1). It is clear that this task is equivalent to finding the critical points of the effective potential, an analysis that is found in every book on celestial mechanics. The result is that there are five such points. There are three collinear points on the x-axis that were discovered by Euler around 1750 and are denoted L_1, L_2, L_3 and there are two equilateral points discovered by Lagrange around 1760 and are denoted L_4, L_5. They are indicated in Figure 2.2.

Equations (2.1) may be interpreted as those of a particle moving in an effective potential plus a magnetic field. Its graph is shown in Figure 2.3. This figure also shows the region one gets by imposing conservation

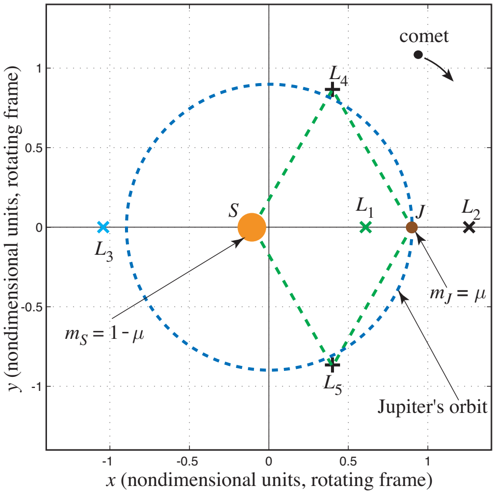
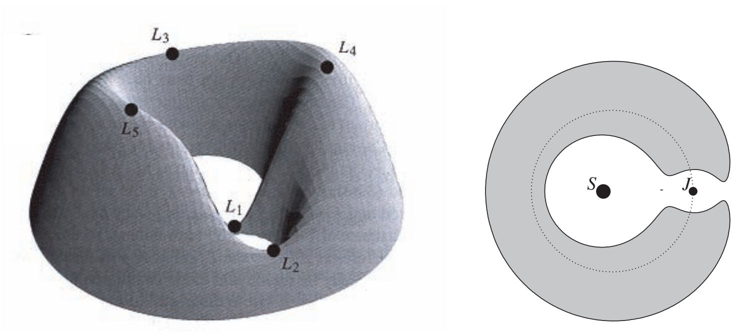


Figure 2.2: Equilibrium points for the three body problem.

of energy and the simple inequality that the kinetic energy is positive. Thus, at a given energy level E, the third body can only move in the region given by the inequality $E - \overline{V} \geq 0$; this is called the *Hill's region* and is obtained by intersecting the graph of the effective potential with a horizontal plane. An example is shown in the right hand side of Figure 2.3 for the Sun-Jupiter-third body system. In this figure, one can see three *realms*, namely the *Sun realm*, the *Jupiter realm* and the *exterior realm* that are connected by the *neck regions*, the left hand neck containing L_1 and the right hand neck containing L_2. For other values of the energy, one or more of these realms may be prohibited due to conservation of energy; that is, the necks may close off.

Of special interest are the two points L_1 and L_2 closest to the secondary body, which a linearized analysis shows are center-saddle points. The famous Liapunov theorem says that there is a family of periodic orbits surrounding each of these points; one can think of this as meaning that one can "go into orbit about these points". These planar periodic orbits are called *Liapunov orbits*, while their counterparts in the 3D problem are called *halo and Lissajous orbits* (which, by the way involves an interesting bifurcation analysis).

Tubes. In the 3 body problem, a key role is played by the invariant manifolds of these periodic orbits, which we call the *Conley-McGehee tubes*. Also key is a network of homoclinic and heteroclinic orbits con-



Effective Potential

Level set shows the Hill region

Figure 2.3: The graph of the effective potential in the 3-body problem. Its critical points are the equilibria.

necting these periodic orbits, also discovered in a preliminary way in work of Conley and McGehee and was extended and thoroughly investigated in Koon, Lo, Marsden, and Ross [2000]. Some of the reasons that these tubes are important can be seen in the context of specific space missions described below.

In fact, the invariant manifold structures of L_1 and L_2 provide the framework for understanding and categorizing the motions of spacecraft as well as, for example, comets that undergo resonance hopping. Moreover, the stable and unstable invariant manifold tubes associated to periodic orbits around L_1 and L_2 are the phase space conduits transporting material between different realms in a single three body system as well as between primary bodies for separate three-body systems. These tubes can be used to construct new spacecraft trajectories as we will indicate below. It is remarkable that the connecting orbits as well as the associated Conley-McGehee tubes are critical for understanding transport in the solar system as well as in molecular systems. It is quite interesting that some of the same techniques used in the celestial context can also be used in the molecular context, and conversely, techniques from Chemistry can be used in celestial problems, as was done in Jaffé, Ross, Lo, Marsden, Farrelly, and Uzer [2002].

Figure 2.4 shows some tubes (projected from phase space to configuration space) associated with periodic orbits about L_1, L_2 for the Earth-Moon system. As this figure indicates, it is the tubes that control the capture and escape properties as well as transit and non-transit orbits.

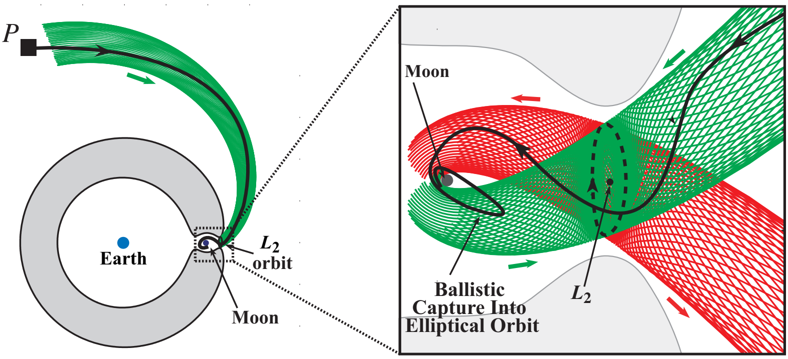


Figure 2.4: Tube leading to ballistic capture around the Moon (seen in rotating frame).

Some Specific Missions. For the complex space missions planned for the near future, greater demands are placed on the trajectory design. In many instances, standard trajectories and classical methods such as the patched two-body approximation are inadequate to support the new mission concepts. Without appropriate and economical trajectories, these missions cannot be achieved. For nearly half a century, space mission planners have depended on trajectory concepts and tools developed in the 1950s and 1960s, based largely on a two-body decomposition of the solar system, the "patched conics" approach. While that approach remains very valuable for some missions, new trajectory paradigms must be developed to meet today's challenges.

A detailed understanding of the three-body problem, and in particular the dynamics associated with libration points, is absolutely necessary to continue the exploration and development of space.

Figure 2.5 shows in metro map format connections between hubs in Earth's neighborhood and beyond. NASA desires to develop a robust and flexible capability to visit several potential destinations. As shown in the figure, NASA has recognized that libration points L_1 and L_2 in the Sun-Earth and Earth-Moon system are important hubs and/or destinations. The fortuitous arrangement of low energy passageways in near-Earth space implies that lunar L_1 and L_2 orbits are connected to orbits around Earth's L_1 or L_2 via low energy pathways.[4] Therefore, a Lunar Gateway Station at the lunar L_1 would be a natural transportation hub to get humanity beyond low-Earth orbit, a stepping stone to the moon, Earth's neighborhood, Mars, the asteroids, and beyond. We

[4] We will sometimes refer to the Sun-Earth L_1 and L_2 as the Earth's L_1 and L_2, since they are much closer to the Earth than the Sun. Similarly, we will occasionally refer to the Earth-Moon L_1 and L_2 as the lunar or the Moon's L_1 and L_2.

will discuss the Lunar L_1 Gateway Station further below.

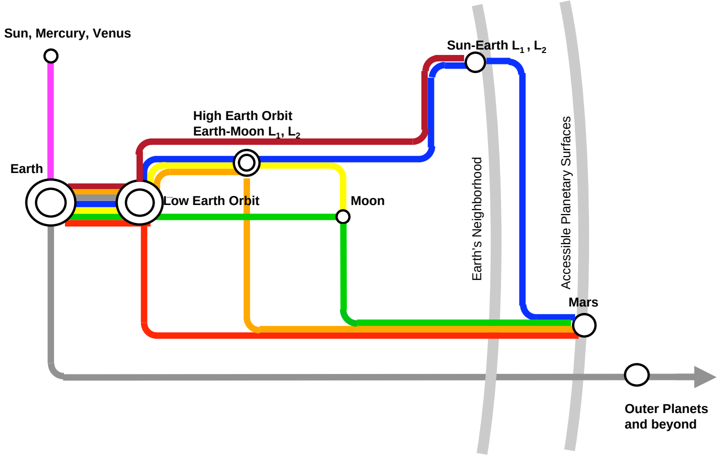


Figure 2.5: A metro map representation showing hubs connected by low energy passageways in the near-Earth neighborhood and beyond (source: Gary L. Martin, NASA Space Architect).

Because of its unobstructed view of the sun, the Sun-Earth L_1 is a good place to put instruments for doing solar science. NASA's Genesis Discovery Mission has been there, the first space mission designed completely using invariant manifolds and other tools from dynamical systems theory (Howell, Barden, and Lo [1997]). The Solar and Heliospheric Observatory (SOHO),[5] a joint project of the European Space Agency and NASA, and NASAs WIND and the Advanced Composition Explorer (ACE) are also there.

Genesis Discovery Mission. Launched in August 2001, the Genesis Discovery Mission spacecraft swept up specks of the sun—individual atoms of the solar wind—on five collector arrays the size of bicycle tires and in an ion concentrator. The goal was to collect solar wind samples and return them safely to the Earth for study into the origins of the solar system. Genesis returned its solar wind cargo to Earth via a sample-return capsule which returned to Earth in September 2004 (see

[5] SOHO is a spacecraft mission designed to study the internal structure of the Sun, its extensive outer atmosphere and the origin of the solar wind, the stream of highly ionized gas that blows continuously outward through the solar system. It is a joint project of the European Space Agency (ESA) and NASA. See `http://soho.estec.esa.nl` for more information.

Lo, Williams, Bollman, Han, Hahn, Bell, Hirst, Corwin, Hong, Howell, Barden, and Wilson [2001]).[6] The sample was the only extraterrestrial material brought back to Earth from deep space since the last of the Apollo landings in 1972, and the first to be collected from beyond the moon's orbit.

A reason Genesis was feasible as a mission is that it was designed using low energy passageways. Figure 2.6 shows a three-dimensional view of the Genesis trajectory (kindly supplied by Roby Wilson). The spacecraft was launched to a *halo orbit* in the vicinity of the Sun-Earth L_1 and uses a "heteroclinic-like return" in the three-body dynamics to return to Earth.[7]

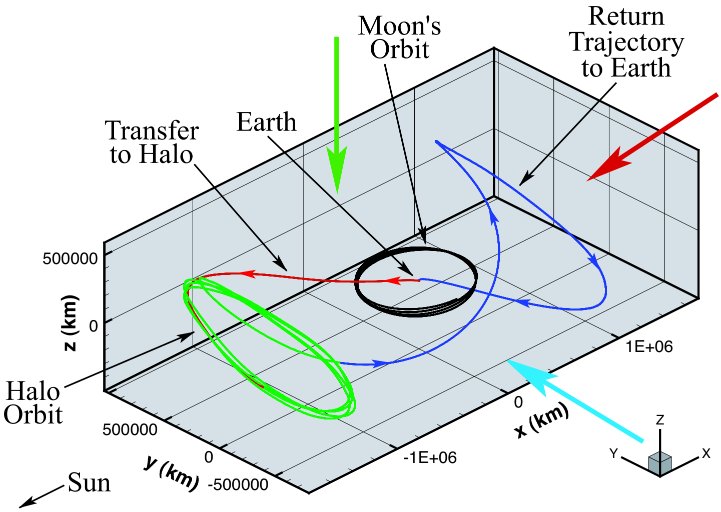


Figure 2.6: The Genesis Discovery Mission trajectory. The three arrows correspond to the three projections shown in Figure 2.7.

As noted above, L_1 is the unstable equilibrium point between the Sun and the Earth at roughly 1.5 million km from the Earth in the direction of the Sun. Genesis took a low energy path to its halo orbit, stayed there collecting samples for about 2 years, and returned home on another low energy path.

[6] See `http://genesismission.jpl.nasa.gov/` for further information.

[7] The orbit is called a "halo orbit" because, as seen from Earth, the flight path follows a halo around the sun. Such orbits were originally named for lunar halo orbits by Farquhar [1968]. By the way, setting a spacecraft exactly to the L_1 point is not a good idea, as the spacecraft's radio signals would be lost in the Suns glare.

Figure 2.7 shows three orthographic projections of the Genesis trajectory. These figures, plotted in a rotating frame, show the key parts of the trajectory: the transfer to the halo, the halo orbit itself, and the return to Earth. The rotating frame is defined by fixing the x-axis along the Sun-Earth line, the z-axis in the direction normal to the ecliptic, and with the y-axis completing a right-handed coordinate system. The y-amplitude of the Genesis orbit, which extends from the x-axis to the maximum y-value of the orbit, is about 780,000 km (see Figures 2.6 and 2.7). Note that this is bigger than the radius of the orbit of the Moon, which is about 380,000 km.

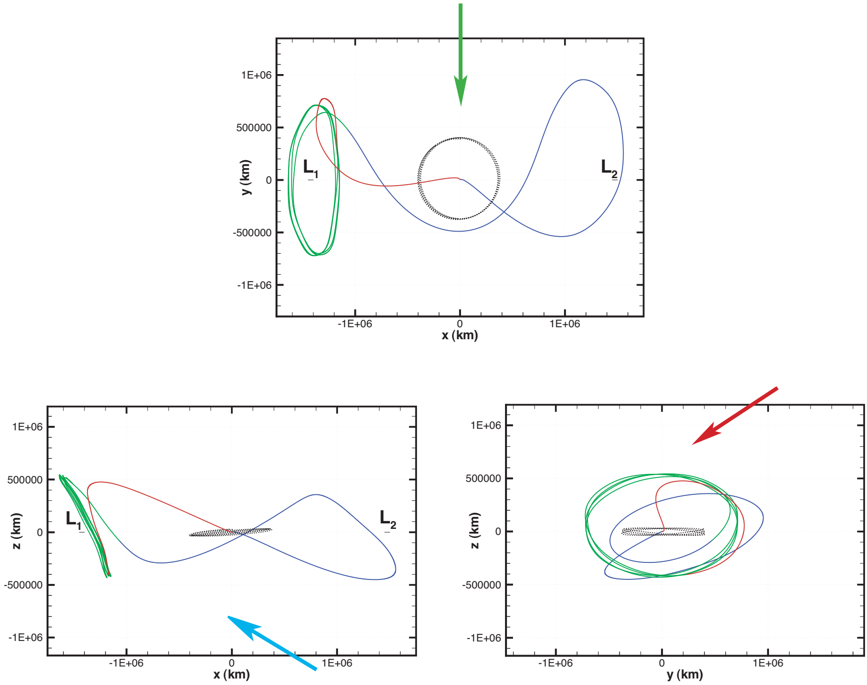


Figure 2.7: The xy, xz, and yz projections of the three dimensional Genesis trajectory shown in the preceding figure.

As Figures 2.6 and 2.7 show, the trajectory travels between neighborhoods of L_1 and L_2; L_2 is roughly 1.5 million km on the opposite side of the Earth from the Sun. In dynamical systems theory, this is closely related to the existence of a *heteroclinic connection* between the L_1 and L_2 regions.

The deeper dynamical significance of the heteroclinic connection for the planar three-body problem is that it allows a classification and a construction of orbits using symbolic dynamics, as was shown in Koon, Lo, Marsden, and Ross [2000], and similar phenomena are seen when the third degree of freedom is included, as discussed in Gomez, Koon, Lo, Marsden, Masdemont, and Ross [2004].

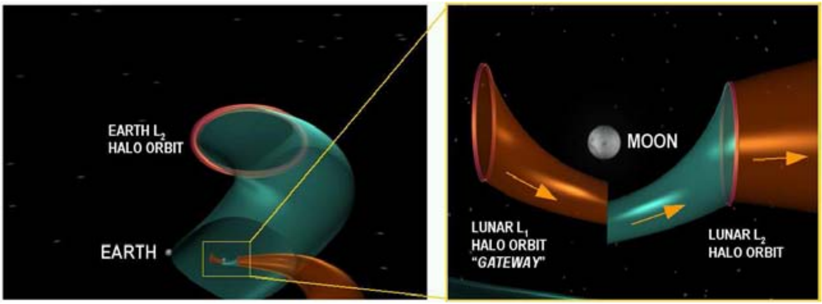
One of the attractive and interesting features of the Genesis trajectory design is that the three year mission, from launch all the way back to Earth return, requires no deterministic maneuver whatsoever and automatically injects into the halo orbit.

It is difficult to use traditional classical algorithms[8] to find a near-optimal solution like that of Genesis, so the design of such a low energy trajectory is facilitated by using dynamical systems methods. This is achieved by using the stable and unstable manifolds as guides in determining the end-to-end trajectory. That Genesis performs its huge exotic trajectory using a deterministic ΔV of *zero* (i.e., no fuel) has created a great deal of interest in both the astronautical and mathematical communities.

Lunar L_1 Gateway Station. The work on Genesis has inspired deeper exploration of the dynamics in Earth's neighborhood (see Lo and Ross [2001]). NASA desires to develop a robust and flexible capability to visit several potential destinations, as suggested by the metro map, Figure 2.5. A Lunar Gateway Station in the vicinity of the lunar L_1 libration point (between the Earth and the Moon) was proposed as a way station for transfers into the solar system and into the Earth-Sun halo orbits. This is enabled by an historical accident: the energy levels of the Sun-Earth L_1 and L_2 points differ from those of the Earth-Moon system by only 50 m/s (as measured by maneuver velocity). The significance of this coincidence to the development of space cannot be overstated. For example, this implies that the lunar L_1 halo orbits are connected to halo orbits around Earth's L_1 and L_2 via low energy pathways, as illustrated in Figure 2.8.

Many of NASA's future space observatories located around the Earth's L_1 or L_2 may be built in a lunar L_1 orbit and conveyed to the final destination with minimal propulsion requirements. When the spacecraft or instruments require servicing, they may be returned from Earth libration orbits to the lunar L_1 orbit where human servicing may be performed, which was shown to be of vital importance for keeping the Hubble Space Telescope operable. Since a lunar L_1 orbit may be reached from Earth in only three days, the infrastructure and complexity of long-term space travel is greatly mitigated. The same orbit could reach

[8] See, for example, Farquhar and Dunham [1981], Farquhar, Muhonen, Newman, and Heuberger [1980], and Farquhar, Muhonen, and Richardson [1977].



(a) (b)

Figure 2.8: (a) The fortuitous arrangement of low energy passageways in near-Earth space implies that lunar L_1 and L_2 halo orbits are connected to halo orbits around Earth's L_1 or L_2 via low energy pathways. Many of NASA's future space telescopes located around the Earth's L_1 or L_2 may be built in a lunar L_1 orbit and conveyed to the final destination with minimal fuel requirements. (b) Shown in this close-up are two halo orbits at the lunar L_1 and L_2, respectively, and the set of invariant manifolds that provide the low energy departures from the lunar L_1 orbit.

any point on the surface of the Moon within hours, making it a perfect location for the return of humans to the Moon. A lunar L_1 orbit is also an excellent point of departure and arrival for interplanetary flights to Mars, the asteroids, and the outer solar system. Several lunar and Earth encounters may be added to further reduce the launch cost and open up the launch period. A lunar L_1 is therefore a versatile hub for a space transportation system.

Multi-Moon Orbiters. Using low energy passageways is in no way limited to the inner solar system. For example, consider a spacecraft in the gravity field of Jupiter and its planet-sized moons. A possible new class of missions to the outer planet moon systems has been proposed by (Koon, Lo, Marsden, and Ross [1999]; Ross, Koon, Lo, and Marsden [2003]). These are missions in which a single spacecraft orbits several moons of Jupiter (or any of the outer planets), allowing long duration observations. Using this *multi-moon orbiter* approach, a single scientific spacecraft orbits several moons of Jupiter (or any of the outer planets) for any desired duration, allowing long duration observations instead of flybys lasting only seconds. For example, a multi-moon orbiter could orbit each of the galilean moons—Callisto, Ganymede, Europa, and Io—one after the other, using a technologically feasible amount of fuel. This approach should work well with existing techniques, enhancing trajectory design capabilities for missions such as NASA's proposed Jupiter Icy Moons Orbiter. Figure 2.9 shows a low energy transfer trajectory

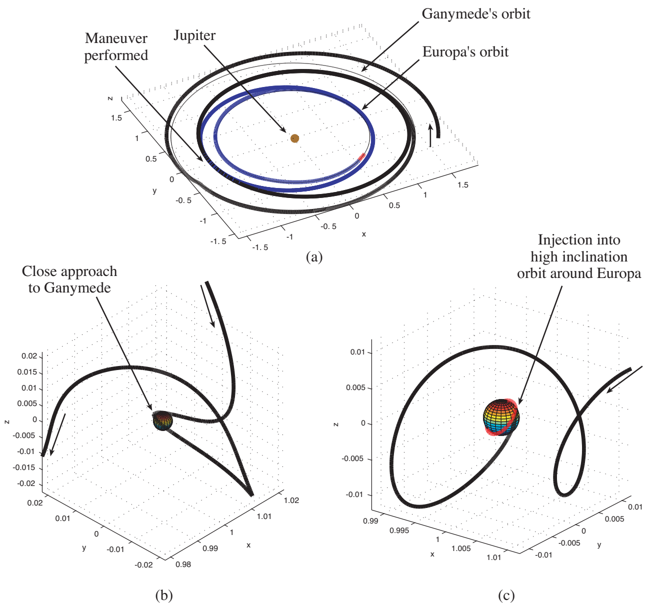


Figure 2.9: A multi-moon orbiter space mission concept for the Jovian moons. (a) We show a spacecraft trajectory coming into the Jupiter system and transferring from Ganymede to Europa using a single impulsive maneuver, shown in a Jupiter-centered inertial frame. (b) The spacecraft performs one loop around Ganymede, using no propulsion at all, as shown here in the Jupiter-Ganymede rotating frame. (c) The spacecraft arrives in Europa's vicinity at the end of its journey and performs a final propulsion maneuver to get into a high inclination circular orbit around Europa, as shown here in the Jupiter-Europa rotating frame.

from an initial Jovian insertion trajectory to Ganymede. After one orbit around Ganymede including a close approach, the spacecraft heads onward to Europa, ending in a high inclination orbit around the icy moon.

3 Transport in the Solar System

As we have indicated, there are many phenomena in the solar system that involve interesting transport processes. Examples include the transport of Mars rocks to Earth (the rocks could be thrown into Mars orbit by a meteor impact, for instance) and the transport of asteroids and comets from outside of Jupiter's orbit to inside of Jupiter's orbit through the two Jupiter necks shown in Figure 2.3. Several comets, such as Oterma did just that (this is described in Koon, Lo, Marsden, and Ross [2000]).

For such processes, one can ask "what is the transport rate?" More specifically, we might wish to compute what percentage of a random distribution on an appropriate energy shell after 1000 years will go from outside of Jupiter's orbit (the exterior realm) to inside (the Sun, or interior, realm)? Similarly, what is the probability of transport of Kuiper belt objects from outside of Neptune's orbit to inside?

To study such questions, we need a few more tools from theoretical and computational dynamical systems.

Poincaré Sections. For the planar 3-body problem, the energy surfaces are three dimensional. Thus, using a Poincaré surface of section at fixed energy E, represents the system as a 2-dimensional area preserving map. For example, in the Sun-Jupiter-third body system, we might choose a section in the interior realm as shown in Figure 3.1.

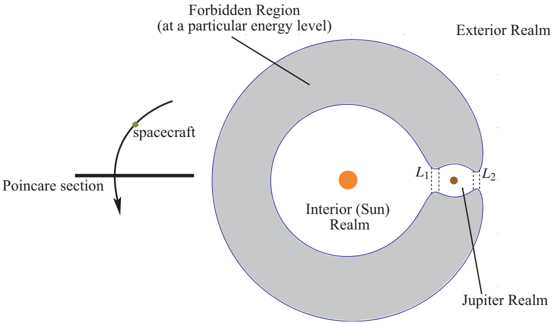


Figure 3.1: A Poincaré surface of section in the exterior realm of the Sun-Jupiter-third body system.

Such a procedure then produces a standard Poincaré map picture, as shown in Figure 3.2.

As we have indicated, and it is also important for transport in both the celestial as well as the molecular context, is that these different Poincaré sections are linked by the Conley-McGehee tubes.

MANGEN. To carry out the needed computations, software is of course required. While there are lots of packages available, we shall concentrate on two of them. First of all, *MANGEN* (Manifold Generation) computes, amongst many other things, invariant manifolds and transport rates between different resonant regions using dynamical systems

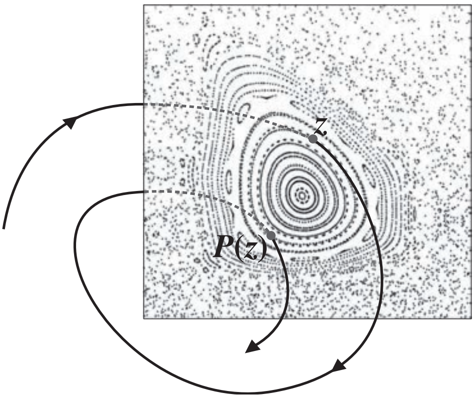


Figure 3.2: A Poincaré map produced by intersecting orbits with a Poincaré section.

methods, such as lobe dynamics (see Rom-Kedar and Wiggins [1990]; Meiss [1992]; Wiggins [1992]; Rom-Kedar [1999]).

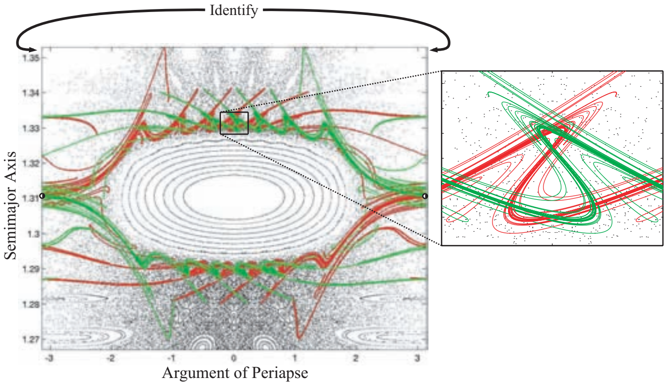


Figure 3.3: Unstable (red) and stable (green) manifolds comprising a homoclinic tangle that bounds a resonant region.

While this software was originally developed for the study of fluid systems (see Lekien [2003]), it has proved to be useful for astrodynamics as well as in molecular systems! Mathematics of course provides the common structures on the two areas that enables this. A sample computation of invariant manifolds in a Poincaré section using MANGEN in the astrodynamics area is shown in Figure 3.3.

One of the interesting questions in dynamical astronomy is to ask about the transport between various regions. Some resonant regions, such at the one shown in Figure 3.4 for a region outside the orbit of Jupiter have “leaky” boundaries due to homoclinic tangles. Using lobe dynamics, MANGEN can compute the transport between such regions and neighboring regions.

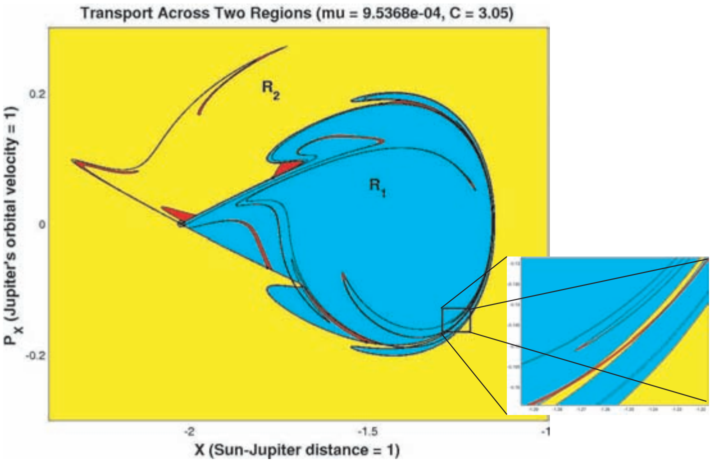


Figure 3.4: MANGEN can compute the transport rate between these two resonant regions R_1 and R_2 in the Sun-Jupiter-third body system.

GAIO. A second piece of software that is very useful is GAIO (Global Analysis of Invariant Objects); see Dellnitz, Froyland, and Junge [2001] and Dellnitz and Junge [2002]. It uses a *set-oriented* methodology, taking a global point of view to compute sets of dynamical interest rather than focusing on individual trajectories. The idea is to cover the region of phase space with boxes and to refine using box subdivisions. A given dynamical system is discretized using a map and then an associated *graph* is constructed. The nodes are the box centers and edges connect those nodes that are dynamically related. This set up allows one to make use of techniques from graph theory, such as graph partitioning.

One of the key concepts in this area is the notion of an *almost invariant set (AIS)*, which corresponds to a set containing relatively long-lived dynamical trajectories. An example of an (AIS) are the above resonance regions. This notion is also important in, for example, biomolecules, where it corresponds to molecular conformations.

There are two related ways to compute almost invariant sets. The first is to use graph partitioning software such as PARTY (Monien, Preis and Diekmann [2000]). The idea is to find efficient ways to cut the graph so that the traffic flow (that is, the transport rate) across the cut is minimized. A second method is the use of eigenfunctions of the associated Perron Frobenius operator (the induced map on measures). We refer to Dellnitz and Junge [2002] and to Dellnitz, Junge, Koon, Lekien, Lo, Marsden, Padberg, Preis, Ross, and Thiere [2004] for a survey of these methods and for further references. In either case, the computation of transport rates between two AIS is naturally computed within these set oriented methods.

Figure 3.5 shows the same resonance region as above, but computed using GAIO.

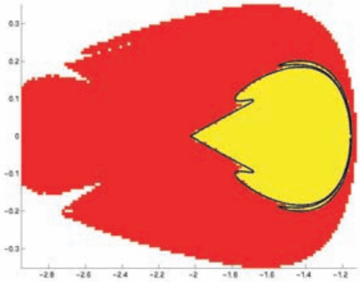


Figure 3.5: Resonance region for the three body Sun Jupiter system computed using GAIO.

In some circumstances, such as shown in Figure 3.6, GAIO and MANGEN can work together to produce more efficient adaptive algorithms.

Using these techniques, one gets very concrete answers for transport rates. For example, it is shown in Dellnitz, Junge, Koon, Lekien, Lo, Marsden, Padberg, Preis, Ross, and Thiere [2004] that there is a 5% probability that a randomly chosen particle will go from one of these regions to the other in 1,800 Earth years.

Mars Crossers. Figure 3.7 shows the data from the Hilda group of asteroids and comets that lie between Jupiter and Mars in a belt relatively close to Jupiter.

Figure 3.8 shows a Poincaré map for a cut chosen at an energy level appropriate to the Hilda group and cutting across it. The coordinates are x along the cut and $\dot{x}$ the corresponding velocity. Also shown is the Mars crosser line; i.e., the coordinates $(x, \dot{x})$ of points whose Keplerian orbit with that initial condition will just graze the orbit of Mars. Hildas

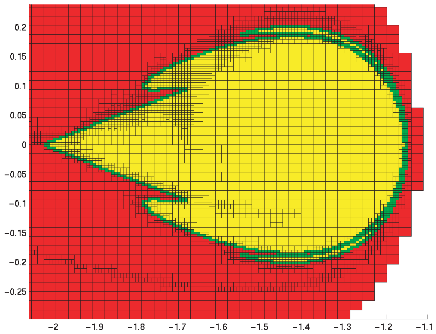


Figure 3.6: GAIO and MANGEN working together.

become Mars-crossers by going from the left of this curve, where they usually reside, to the right of it.

As shown in Dellnitz, Junge, Koon, Lekien, Lo, Marsden, Padberg, Preis, Ross, and Thiere [2004a], GAIO can locate the quasi-Hilda region as one of the AIS in this three body problem. Drawing in the set corresponding to the Mars crosser line, as shown in Figure 3.9, we can then ask "what is the probability that, in a certain period of time, an object in the quasi-Hilda region becomes a Mars crosser?" GAIO gives rather concrete answers. It shows that the probability for a typical particle to leave the quasi-Hilda region is around 6% after 200 iterates of the map, which corresponds to a transit time between 2000 and 6000 Earth years, depending on the location of the particle within the quasi-Hilda region.

But there is much more to the story, as shown in Figure 3.10. Look now at all of the inner planet crosser curves and notice how they lie in the middle of various of the Sun-Jupiter-third body almost invariant sets (as computed by GAIO)! But the almost invariant sets are computed just with the Sun-Jupiter-third body system, which in principle is independent of the knowledge of the other planets. However, this shows that these are not independent at all and it suggests that the Jupiter system in fact drove the formation of the whole solar system. Of course additional analysis and simulation are needed to make this definitive. This does, however highlight one example of the many miracles hidden in the solar system and the relations between the planets! We are also hopeful that such ideas might be useful in the search for Earth-like planets in other solar systems.

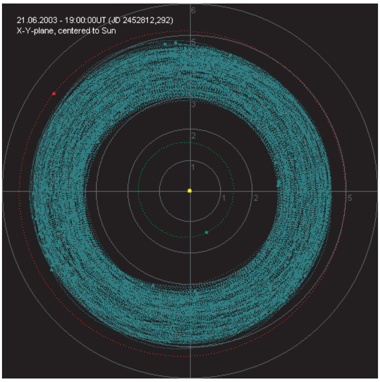


Figure 3.7: Orbits of asteroids in the Hilda group. Jupiter is the large dot on the outer dotted circle and Mars is the large dot on the inner dotted circle

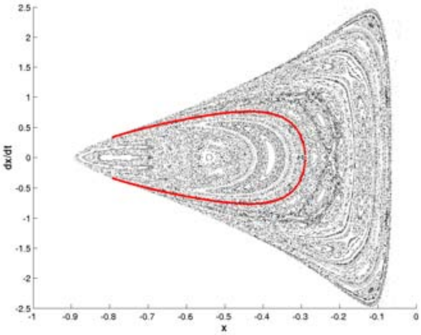


Figure 3.8: A Poincaré map corresponding to a section cutting across the Hilda group.

Acknowledgements. The authors research was partially supported by National Science Foundation grants DMS 0402842, DMS-0204474 and a Max-Planck Research Award.

References

Bate, R. R., D. D. Mueller, and J. E. White [1971], *Fundamentals of Astrodynamics*. Dover, New York.

Belbruno, E. [2004], *Capture Dynamics and Chaotic Motions in Celes-*

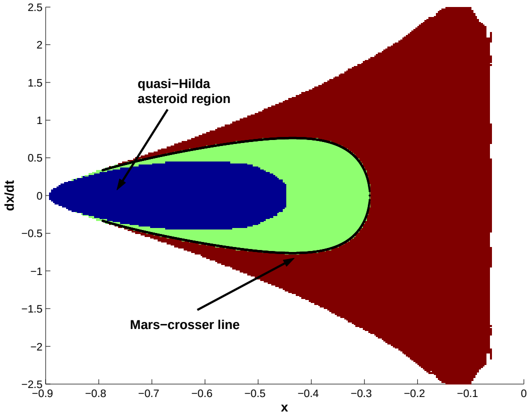


Figure 3.9: The quasi Hilda set (an AIS) and the Mars crosser set.

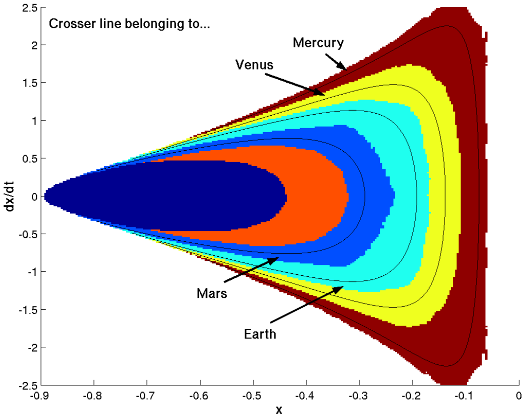


Figure 3.10: Crosser lines of all the inner planets superimposed on almost invariant sets for the Sun-Jupiter-third body system.

tial Mechanics: With Applications to the Construction of Low Energy Transfers, Princeton University Press.

Dellnitz, M., G. Froyland, and O. Junge [2001], The algorithms behind GAIO-set oriented numerical methods for dynamical systems. In *Er-*

godic theory, analysis, and efficient simulation of dynamical systems, number MR1850305 (2002k:65217), pages 145–807. Springer, Berlin.

Dellnitz, M. and O. Junge [2002], Set oriented numerical methods for dynamical systems. In *Handbook of dynamical systems, Vol. 2*, number 1 900 656, pages 221–264. North-Holland, Amsterdam.

Dellnitz, M., O. Junge, W. S. Koon, F. Lekien, M. W. Lo, J. E. Marsden, K. Padberg, R. Preis, S. Ross, and B. Thiere [2004], Transport in dynamical astronomy and multibody problems, *Intern. J. of Bifurcation and Chaos (to appear).*

Dellnitz, M., O. Junge, W. S. Koon, F. Lekien, M. W. Lo, J. E. Marsden, K. Padberg, R. Preis, S. Ross, and B. Thiere [2004a], Transport of Mars-crossers from the quasi-Hilda region, (submitted for publication).

Dunn, G. L. [1962]. A high-speed data link for farside lunar communications, General Electric Co. Report 62 SPC-5, March 1962.

Farquhar, R. W. [1966], Station-Keeping in the Vicinity of Collinear Libration Points with an Application to a Lunar Communications Problem. In *Space Flight Mechanics, Science and Technology Series*, volume 11, pages 519–535. American Astronautical Society, New York.

Farquhar, R. W. and D. W. Dunham [1981], A new trajectory concept for exploring the Earth's geomagnetic tail, *Journal of Guidance and Control* **4**, 192–196.

Farquhar, R. W., D. P. Muhonen, C. Newman, and H. Heuberger [1980], Trajectories and Orbital Maneuvers for the First Libration-Point Satellite, *Journal of Guidance and Control* **3**, 549–554.

Farquhar, R. W., D. P. Muhonen, and D. L. Richardson [1977], Mission Design for a Halo Orbiter of the Earth, *Journal of Spacecraft and Rockets* **14**, 170–177.

Farquhar, R. [1968], *The Control and Use of Libration-Point Satellites*, PhD thesis, Stanford University.

Gomez, G., W. S. Koon, M. W. Lo, J. E. Marsden, J. Masdemont, and S. D. Ross [2004], Connecting orbits and invariant manifolds in the spatial restricted three-body problem, *Nonlinearity* **17**, 1571–1606.

Howell, K., B. Barden, and M. Lo [1997], Application of dynamical systems theory to trajectory design for a libration point mission, *The Journal of the Astronautical Sciences* **45**, 161–178.

Jaffé, C., S. D. Ross, M. W. Lo, J. E. Marsden, D. Farrelly, and T. Uzer [2002], Statistical Theory of Asteroid Escape Rates, *Phys. Rev. Lett.* **89**, 011101–1.

Koon, W. S., M. W. Lo, J. E. Marsden, and S. D. Ross [1999], Constructing a Low Energy Transfer between Jovian Moons. In *Celestial Mechanics : an international conference on celestial mechanics*, Evanston, Illinois.

Koon, W. S., M. Lo, J. E. Marsden, and S. Ross [2000], Heteroclinic connections between periodic orbits and resonance transitions in celestial mechanics, *Chaos* **10**, 427–469.

Koon, W. S., M. Lo, J. E. Marsden, and S. Ross [2005], *Dynamical Systems, the Three-Body Problem and Space Mission Design* (to be published).

Lekien, F. [2003], *Time-Dependent Dynamical Systems and Geophysical Flows*, PhD thesis, California Institute of Technology.

Lo, M., B. G. Williams, W. E. Bollman, D. Han, Y. Hahn, J. L. Bell, E. A. Hirst, R. A. Corwin, P. E. Hong, K. C. Howell, B. Barden, and R. Wilson [2001], Genesis Mission Design, *The Journal of the Astronautical Sciences* **49**, 169–184.

Lo, M. W. and S. D. Ross [1998], Low energy interplanetary transfers using invariant manifolds of L1, L2 and halo orbits. In *AAS/AIAA Space Flight Mechanics Meeting*, Monterey, California.

Lo, M. W. and S. D. Ross [2001], The Lunar L1 Gateway: Portal to the stars and beyond. In *AIAA Space 2001 Conference*, Albequerque, New Mexico.

Meyer, K. R. and R. Hall [1992], *Hamiltonian Mechanics and the N-Body Problem.* Texts in Applied Mathematics Science. Springer-Verlag, Berlin.

Meiss, J. D. [1992], Symplectic maps, variational principles, and transport. *Rev. Mod. Phys.* **64**(3), 795–848.

Monien, B., R. Preis, and R. Diekmann [2000], Quality matching and local improvement for multilevel graph-partitioning. *Parallel Computing*, **26**(12), 1609–1634.

Rom-Kedar, V. [1999], Transport in a class of n-d.o.f. systems, in *Hamiltonian systems with three or more degrees of freedom (S'Agaró, 1995)*, Vol. **533** of *NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci.*, Kluwer Acad. Publ., Dordrecht, 538–543.

Rom-Kedar, V. and S. Wiggins [1990], Transport in two-dimensional maps. *Arch. Rat. Mech. Anal.* **109**, 239–298.

Ross, S. D., W. S. Koon, M. W. Lo, and J. E. Marsden [2003], Design of a Multi-Moon Orbiter. In *13th AAS/AIAA Space Flight Mechanics Meeting*, Ponce, Puerto Rico. Paper No. AAS 03-143.

Roy, A. E. [1988], *Orbital Motion.* Adam Hilger, Bristol, 3rd edition.

Wiggins, S. [1992], *Chaotic transport in dynamical systems.* Interdisciplinary Appl. Math. **2**. Springer, Berlin-Heidelberg-New York.

Graph Minor Theory

László Lovász

December 1, 2004

Abstract

A monumental project in graph theory was recently completed. The project, started by Robertson and Seymour, and later joined by Thomas, led to entirely new concepts and a new way of looking at graph theory.

The motivating problem was Kuratowski's characterization of planar graphs, and a far-reaching generalization of this, conjectured by Wagner: If a class of graphs is minor-closed (i.e., it is closed under deleting and contracting edges), then it can be characterized by a *finite* number of excluded minors. The proof of this conjecture is based on a very general theorem about the structure of large graphs: If a minor-closed class of graphs does not contain all graphs, then every graph in it is glued together in a tree-like fashion from graphs that can almost be embedded in a fixed surface.

We describe the precise formulation of the main results, and survey some of its applications to algorithmic and structural problems in graph theory.

1 Introduction

Let us start with recalling Kuratowski's Theorem [10]:

Theorem 1 *A graph G is embedable in the plane if and only if it does not contain a subgraph homeomorphic to the complete graph K_5 or the complete bipartite graph $K_{3,3}$.*

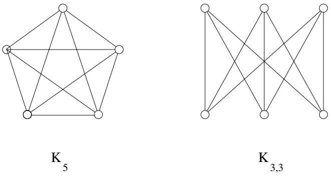


Figure 1: Excluded minors for planar graphs.

It is an immediate and natural question to ask if a similar result holds for other surfaces: can one characterize graphs embedable in a fixed surface Σ by excluding subgraphs homeomorphic to graphs in a finite list? Studies concerning specific surfaces are somewhat discouraging: it seems that the only surface besides the plane (or sphere) for which such a list (of 35 graphs) is known is the projective plane. Nevertheless, the existence of a finite list was proved by Robertson and Seymour [16].

Wagner formulated a fundamental conjecture (apparently only published in 1970 in a textbook [31]), which extends this finite basis result to a much more general setting, namely characterizing *minor-closed classes of graphs.* This conjecture was proved by Robertson and Seymour in series of papers; the final version of the paper in which the proof is completed has just been finished. This gives us the excuse to survey this monumental work.

A crucial element of the proof is a theorem about the structure of graphs not containing a certain minor. Roughly speaking, it says that if a graph does not contain a certain minor, then it is basically 2-dimensional. The exact statement of the theorem (section 4) will be more complicated.

2 Minors and embeddings

Given a graph G, we consider the following three ways of reducing it:

(a) delete an edge;

(b) contract an edge;

(c) delete an isolated node.

Any graph G' that can be produced from G by successive application of these reductions is called a *minor* of G. (In particular, G is a minor of itself.) Every graph that is isomorphic to a minor of G is also called a minor of G. A minor that is not isomorphic to G is called a *proper* minor.

This notion fits well with many notions and problems graph theory studies. In fact, if a graph theorist learns about property that is inherited minors, he or she knows that this property is interesting from a graph theoretical point of view. Planarity of graphs is an example. We can generalize this: the property of being embedable in any other fixed surface is inherited by minors.

There are many simple but important graph properties minor closed (inherited by minors). For example, being *series-parallel*: these are graphs that can be obtained from a single edge by a sequence of parallel extensions (adding an edge parallel to an edge that already exists) and series extensions (subdividing an edge by a new node).

Various topological properties of graphs are also often minor-closed. Every graph is embedable in $\mathbb{R}^3$, but we may impose additional conditions on such embeddings. For example, the graph is *linklessly embedable*, if it has an embedding in which no two disjoint cycles of the graph are linked. A similar notion is *knotlessly embedable*: these graphs have an embedding in 3-space in which no cycle is knotted. Both these topological properties are minor-closed.

3 Wagner's conjecture

3.1 Excluded minor characterizations

Many important and deep theorems characterize minor-closed graph properties by "excluded minors". Kuratowski's Theorem 1 is not directly of this form, but Wagner [30] reformulated it in this way: he showed that instead of excluding K_5 and $K_{3,3}$ as subgraphs up to homeomorphism, it is equivalent to exclude them as minors. Let us quote two further theorems, characterizing minor-closed properties mentioned above. The first is due to Dirac [9]:

Theorem 2 *A graph is series-parallel if and only if it has no K_4 minor.*

The second, much more difficult theorem was conjectured by Sachs and proved by Robertson, Seymour and Thomas [24]:

Theorem 3 *A graph is linklessly embedable if and only if it does not contain any of the seven graphs in Figure 2 as a minor.*

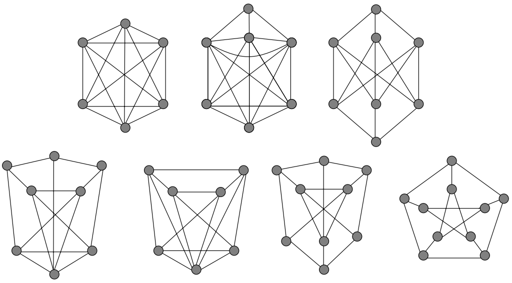


Figure 2: Excluded minors for linklessly embedable graphs (the "Petersen family").

No such theorem is known for knotlessly embedable graphs, even though the main result to be discussed implies that a finite family characterizing them does exist.

3.2 Statements of the theorem

We say that a class $\mathcal{K}$ of graphs is *minor-closed*, if for every $G \in \mathcal{K}$ every minor of G also belongs to $\mathcal{K}$.

Given a family of graphs $\{G_1, G_2, \dots\}$, we can consider the class $\mathcal{K}$ of graphs that do not contain any of $G_1, G_2, \dots$ as a minor. Trivially, this class is minor-closed; we'll say that the graphs $G_1, G_2, \dots$ *characterize $\mathcal{K}$ as excluded minors.* It is also trivial that every minor-closed family can be characterized by excluded minors: just list all graphs not in the family. Wagner's conjecture (now the Robertson–Seymour Theorem) asserts that we can always achieve this by a finite list:

Theorem 4 *Every minor-closed class of graphs can be characterized by a finite family of excluded minors.*

Clearly this theorem is a far-reaching generalization of Kuratowski's Theorem.

For every minor-closed class $\mathcal{K}$ there is a unique *minimal* list of excluded minors characterizing it: this consists of those graphs not in $\mathcal{K}$ for which every proper minor is in $\mathcal{K}$. Theorem 4 asserts that the set of minor-minimal graphs not in $\mathcal{K}$ is finite.

Yet another formulation of this result is that *in every infinite set* $\{G_1, G_2, \dots\}$ *of finite graphs there are two graphs such that one is a minor of the other.* This form puts it in the context of well-quasi-ordering. A partially ordered set $(P, \leq)$ is *well-quasi-ordered*, if every infinite sequence $(x_1, x_2, \dots)$ of its elements has two elements x_i and x_j such that $i < j$ and $x_i \leq x_j$. Theorem 4 says that the set of (isomorphism classes of) finite graphs, with the "minor" relation as partial order, is well-quasi-ordered.

Perhaps the most important special case is embadability in a surface, which was proved earlier [16]:

Corollary 5 *For every closed compact surface there is a finite list of graphs such that a graph* G *is embedable in this surface if and only if it does not contain any of these as a minor.*

It is not hard to see that an analogous theorem holds where a finite list of graphs are excluded as homeomorphic subgraphs (rather than minors).

3.3 Linkages

There is an important graph theoretic problem that plays a central role in the theory. Given a graph and $2k$ nodes $s_1, \dots, s_k, t_1, \dots, t_k$, we may want to know whether there are k disjoint paths $P_1, \dots, P_k$ so that P_i connects u_i to v_i. If such paths exist, we say that the ordered sets $(s_1, \dots, s_k)$ and $(t_1, \dots, t_k)$ are *linked*. If every two ordered k-sets are linked, we say that the graph is *k-linked*.

The linkage problem sounds very similar to Menger's Theorem, which asserts that *for two k-element sets S and T, we have k disjoint paths, each connecting a node in S to a node in T if and only if S and T cannot be separated by $k - 1$ nodes.* The additional condition that we prescribe which element of S should be connected to which element of T makes this problem much more difficult. A complete characterization only exists for $k = 2$ (Thomassen [27], Seymour [25]). Let us assume, to exclude some not-so-interesting complications, that the graph is 4-connected (i.e., it cannot be separated by 3 or fewer nodes).

Theorem 6 *Let G be a 4-connected graph and s_1, s_2, t_1, t_2 four nodes of G. Then (s_1, s_2) and (t_1, t_2) are linked unless G is planar and s_1, s_2, t_1, t_2 are on the boundary of the same face, in this order.*

It is interesting that the answer to a purely graph-theoretic question involves such a strong topological property of the graph.

The linkage problem is very important in many applications: it plays a crucial role in VLSI design, and is closely related to the the Multicommodity Flow Problem.

Linkages are "linked" to graph minors in a number of ways. To illustrate the idea, let us first consider a graph G that is k-linked. Let H be a graph with k edges. then we can find a homeomorphic copy of H in G, by first mapping the nodes of H arbitrarily, specify the edges through which the connecting paths should start, and then solve a linkage problem to map the edges of H onto paths in G.

In the other direction, Robertson and Seymour [19] proved that if a graph is $2k$-connected and has a K_{3k} minor, then the graph is k-linked. Extending these ideas, Bollobás and Thomason [4] proved that every $(22k)$-connected graph is k-linked. For more on this connection, see [5].

3.4 About the proof

We have seen several examples showing that the notion of a minor-closed class is substantially more general than the notion of graphs embedable in a given surface. Still, the proof of Theorem 7 goes through the proof for this special case, namely Corollary 5. Robertson and Seymour show that every graph in a minor-closed class $\mathcal{K}$ (not containing all graphs) can be approximated by a composition of graphs that are embedable in a surface which only depends on $\mathcal{K}$. Graphs that are embedable on this surface can be characterized by a finite number of excluded minors, and from these, the finiteness of the minimal excluded minors for $\mathcal{K}$ can be proved.

The details are very difficult, however. The approximation of graphs in $\mathcal{K}$ by the class of graphs embedable in some surface with bounded genus leads to an exciting and deep structure theorem, which we'll discuss in the next section. The finiteness of the list of excluded minors for a given surface (Corollary 5) is needed in a stronger form, not only for graphs, but also for hypergraphs.

How does topology come in at all? Theorem 2-LINK above may provide a hint. If we find disjoint paths between certain pairs of nodes in a certain part of the graph, then we can use these to construct appropriate minors. If not, then we know that this part of the graph is planar, which could be the beginning of an embedding of the whole graph.

4 Structure theory

4.1 Constructive characterizations

Let us fix a graph H and consider the class $\mathcal{K}_H$ of graphs not containing H as a minor. It is clear that this class is in NP (to certify that a graph G is not in $\mathcal{K}_H$, just exhibit the way H is produced from G as a minor). It follows from Graph Minor Theory that this class is in P, and so also in co-NP. How can we certify that $G \in \mathcal{K}_H$, i.e., that G does not contain H as a minor?

As an illustration, let us quote Wagner's characterization of graphs not containing the complete graph K_5 as a minor [29]. We need some definitions. Let G_1 and G_2 be two graphs, and let $S_i \subseteq V(G_i)$ be a k-clique (a set of k mutually adjacent nodes). Let G be obtained by identifying S_1 with S_2, and deleting some (possibly none, possibly all) edges between the nodes in $S_1 = S_2$. We say that G is a *k-clique-sum* of G_1 and G_2.

We denote by V_8 the graph obtained from a cycle of length 8 by connecting opposite nodes.

Theorem 7 *A graph G has no K_5 minor if and only if it can be obtained by* 0-*,* 1-*,* 2- *and* 3-*clique-sum operations from planar graphs and V_8.*

This theorem can serve as a paradigm for answering such a question: we find a class that does not contain a K_5 minor for topological reasons (planar graphs), throw in some exceptions, and describe gluing rules that preserve the property that there is no K_5 minor. But this theorem also warns us that such a certificate can become quite complicated, and in general it is probably hopeless to explicitly describe the basic classes and gluing rules that would produce $\mathcal{K}_H$.

4.2 Approximate characterizations

The main idea behind a successful structure theory is to prove such a result in an approximate sense. We start with an early result of this kind from [15]. We say that the graph G has *tree-width* at most k is we can write G as the union of subgraphs G_i, which are indexed by the nodes of a tree T, with the following properties.

(i) each G_i has at most $k+1$ nodes;

(ii) if $i, j, k \in V(T)$ and j lies on the path between i and k, then $V(G_i) \cap V(G_k) \subseteq V(G_j)$.

Equivalently, G can be obtained by repeatedly taking clique-sum with graphs with at most $k+1$ nodes (Figure 3.

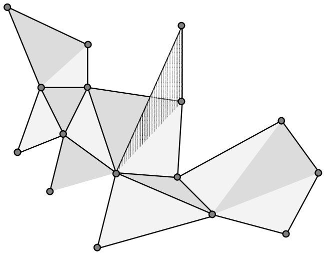


Figure 3: A graph with treewidth 2. The colored triangles indicate the subgraphs G_i.

Theorem 8 (a) *For every planar graph H there is an integer k such that if a graph does not contain H as a minor, then its tree-width is at most k.*

(b) *For every $k > 0$ there is a planar graph H such that no graph with tree-width at most k contains H as a minor.*

The main assertion here is that if a graph does not contain a given planar graph H as a minor, then it has bounded tree-width, and therefore it can be constructed from bounded size graphs, by gluing them together in a tree-like structure.

In an analogous (but much more difficult) fashion, the following construction does not characterize $\mathcal{K}_H$; instead, it describes the approximate structure of every graph in $\mathcal{K}_H$.

We need a definition. Let C be a cycle. Select a family of arcs on C, so that each node is contained in at most k of these arcs. For each arc A, create a new node v_A. Connect v_A to some nodes on the arc A. Also connect any number of pairs (v_A, v_B) for which A and B have a common node. We call this *adding a fringe of width* k to C.

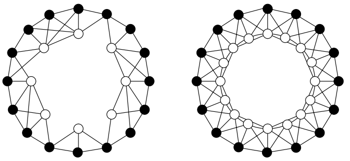


Figure 4: A fringe of width 2 and a fringe of width 3.

For a positive integer k, construct the following class $\mathcal{L}_k$ of graphs:

(i) We start with a graph G embedded in a connected closed surface Σ with genus at most k so that each face is homeomorphic with an open disc.

(ii) We select at most k faces of G and add a fringe of width at most k to each of them.

(iii) We create at most k new nodes and connect them to the other nodes arbitrarily.

(iv) We repeatedly construct the k-clique-sum of the graph we have with another graph constructed using steps (i)–(iii) above.

It is clear that the class $\mathcal{L}_a$ is in NP: to certify that a graph is in $\mathcal{L}_a$, just follow the construction. The assertion that this construction provides an "approximate good characterization" of classes characterized by excluded minors is made precise by the following fundamental theorem [20]:

Theorem 9 *(a) For every graph H there exists an integer $a > 0$ such that $\mathcal{K}_H \subseteq \mathcal{L}_a$;*

(b) For every integer $a > 0$ there exists a graph H such that $\mathcal{L}_a \subseteq \mathcal{K}_H$.

The second assertion is not hard, and it is included here just for completeness. The hard, and useful, part is (a). We can strengthen it in various ways—for example, in (i) we may start with a surface on which H does not embed.

5 Algorithmic consequences

Graph minor theory has an algorithmic consequence that is unprecedented in its generality [19].

Theorem 10 *Every minor-closed property of graphs can be tested in polynomial time.*

The algorithm that follows from the Graph Minor Theory is of complexity $O(n^3)$. The devil is hidden in the big-O; first, the constants are huge and second, they depend on the list of excluded minors. While the finiteness of this list is guaranteed by Theorem 7, it is in general not easy to find and, as we have remarked, it can be very large. So (unless the property that we want to test is given by an explicit list of excluded minors), Theorem 10 only tells us the existence of a polynomial time algorithm; it is a very unusual "pure existence theorem" for an algorithm.

The notion of treewidth introduced above has turned out to be quite useful in algorithm design: there are many graph parameters that are difficult (NP-hard) to compute in general, but which can be computed in polynomial time if the graph in question has bounded treewidth. In view of Theorem 8, this means that if we restrict our attention to graphs not containing a given planar graph as a minor, then we can solve many problems in polynomial time that are NP-hard in general (see e.g. [2]).

Let the chromatic number serve as an example: this is defined as the minimum number of colors needed to color the nodes of a graph G, so that adjacent nodes get different colors. This fundamental parameter is difficult (NP-hard) to compute; but for graphs with bounded treewidth we can use the following method.

Let G be a graph with treewidth at most k; we want to decide whether it is colorable with r colors. We know that G can be glued together from pieces G_i with at most $k+1$ nodes, which are indexed by the nodes of a tree T, satisfying (TW1) above. Let us designate one of these G_i, say G_1, as the *root region.* Our algorithm will be recursive, and in fact do more: it will decide for every r-coloring of the root region, whether or not the coloring can be extended to a (legitimate) r-coloring of the whole graph.

Now the algorithm consists of two easy recursive steps:

(1) If the root region, as a node of T, has degree $d > 1$, we decompose the tree into its "branches" relative to the root. These branches correspond to subgraphs of G, for which the extension problem can be solved recursively. A coloring of the root can be extended to G if and only if it can be extended to every branch.

(2) If the root region has degree $d = 1$ in T, we delete this node from T, and designate its neighbor as the new root region. We solve the extension problem for this smaller graph, and it is easy to check which colorings of the old root can be extended to the new root, and which of these can be extended to the rest.

6 The decomposition paradigm

The excluded minor characterizations and the structure theorems discussed above can serve as prototypical examples of a paradigm that leads to very difficult but important results.

Perhaps most dramatic of these is the recent resolution of the Strong Perfect Graph Conjecture by Chudnovsky, Robertson, Seymour and Thomas [6]. Here again, the key to the proof is a structure theory, which describes how every perfect graph can be glued together from certain basic types. The minor-producing operation in this case is deleting a node.

The paradigm goes way beyond graph theory. A beautiful and rather early example is a pair of difficult theorems on regular matroids. These are matroids that can be coordinatized by a totally unimodular matrix. Interest in them comes from the fact that two standard matroids derived from graphs, the cycle matroid and the cocycle matroid of a graph, are totally unimodular.

The question of characterizing regular matroids is closely related to (but not quite equivalent with) characterizing totally unimodular matrices. Tutte [28] gave a characterization in terms of excluded minors. On the other hand, Seymour [26] gave a constructive description of regular matroids: he showed that they can be glued together (in a way analogous to 1- and 2- and 3-clique-sums) from cycle matroids and cocycle matroids of graphs, and one particular 10 element matroid. Tutte's result tells you why a matroid is *not* regular: it is because it contains, as a minor, one of three particular matroids. Seymour's result tells you why a matroid is regular: because it can be built up in a specific way.

We should also mention the characterization of balanced matrices using the same paradigm by Conforti, Cornuejols, Kapoor, Rao and Vuskovic [7, 8].

7 Research directions

7.1 Simpler proofs

It would be quite important to have simpler proofs with more explicit bounds. Warning: many of us have tried, but only a few successes can be reported. For the generalization of Kuratowski's Theorem to other surfaces (Corollary 5 such proofs are known: Archdeacon and Huneke [1] proved it (before the Robertson–Seymour proof of the general result) for nonorientable case, and Mohar [11] gave a constructive proof for the orientable case.

7.2 Exact structural descriptions

If a class of graphs is defined in terms of excluded minors, then it is in co-NP (it is easy to certify that a graph contains one of these). We also know that it is in P, and hence, also in NP; but is there a direct way to certify that a graph is in this class? A structure theorem could serve this purpose (an example would be Wagner's Theorem 7), but such structure theorems are only known in special cases, and in the general case, we only have the approximate structure theorem 9. (This should also warn us that such a result could be very difficult.)

7.3 Properties of minor-closed classes

It seems that there are many interesting known nd unknown general properties of minor-closed classes; some follow from Theorem 9, others need (or should need) different techniques. To give an example: an old result of Babai [3] asserts that *if $\mathcal{K}$ is a minor-closed class of graphs that does not contain all graphs, then graphs in $\mathcal{K}$ cannot have arbitrary automorphism groups.*

We have seen above that if we restrict our attention to graphs that do not contain a given planar graph as a minor, then many hard algorithmic problems become polynomially solvable. There are also several examples of hard algorithmic problems (for example, a version of the linkage problem in Section 3.3) that are polynomially solvable for planar graphs. In some cases, Theorem 9 allows us to extend these to any minor-closed class of graphs; but there are other such problems, where this extension does not seem to work.

7.4 3-dimensional graphs

One way of interpreting Theorem 9 is that graphs that don't have all minors are essentially 2-dimensional, and vice versa. Is there a similar description of "3-dimensional" graphs? Is there a general notion of "minor", that would correspond to graphs whose structure we feel is 3-dimensional?

References

[1] D. Archdeacon and P. Huneke: A Kuratowski theorem for nonorientable surfaces, *J. Combin. Theory* Ser. B **46** (1989), 173–231.

[2] S. Arnborg and A. Proskurowski: Linear time algorithms for NP-hard problems restricted to partial k-trees, *Discrete Appl. Math.* **23** (1989), 11–24.

[3] L. Babai: Authomorphism groups of graphs and edge contraction, *Discrete Math.* **8** (1974), 13–22.

[4] B. Bollobás and A. Thomason: Highly linked graphs, *Combinatorica* **16** (1996), 313–320.

[5] G. Chen, R.J. Gould, K. Kawarabayashi, F. Pfender and B. Wei: Graph Minors and Linkages (preprint).

[6] M. Chudnovsky, N. Robertson, P.D. Seymour and R. Thomas: The Strong Perfect Graph Theorem (to appear).

[7] M. Conforti, G. Cornuejols and M.R. Rao: Decomposition of Balanced Matrices, *J. Combin. Theory* Ser. B **77** (1999), 292–406.

[8] M. Conforti, G. Cornuejols, A. Kapoor and K. Vuskovic: Balanced 0,+1,-1 matrices, Part I: Decomposition. *J. Combin. Theory* Ser. B **81** (2001), 243–274.

[9] G.A. Dirac: A property of 4-chromatic graphs and some remarks on critical graphs *J. London Math. Soc.* **27** (1952), 85–92.

[10] K. Kuratowski: Sur le probleme des courbes gauches en topologie, *Fund. Math.***16** (1930), 271–283.

[11] B. Mohar: A linear time algorithm for embedding graphs in an arbitrary surface, *SIAM J. Discrete Math.* **12** (1999), 6–26.

[12] B. Mohar: Graph minors and graphs on surfaces, in: *Surveys in Combinatorics*, Proc. 18th British Comb. Conf. (ed. J.W.P. Hirschfeld), London Math. Soc. Lecture Note Ser. *288* (2001), Cambridge Univ. Press, 145–163.

[13] N. Robertson, P.D. Seymour: Graph minors III. Planar tree-width, *J. Combin. Theory* Ser. B **36** (1984), 49–64.

[14] N. Robertson, P.D. Seymour: Graph minors IV. Tree-width and well-quasi-ordering, *J. Combin. Theory* Ser. B **48** (1990), 227–254.

[15] N. Robertson, P.D. Seymour: Graph minors V. Excluding a planar graph, *J. Combin. Theory* Ser. B **41** (1986), 92–114.

[16] N. Robertson, P.D. Seymour: Graph Minors VIII. A Kuratowski Theorem for General Surfaces, *J. Combin. Theory* Ser. B **48** (1990), 255–288.

[17] N. Robertson, P.D. Seymour: Graph Minors IX. Disjoint Crossed Paths, *J. Combin. Theory* Ser. B **49** (1990), 40–77.

[18] N. Robertson, P.D. Seymour: Graph minors X. Obstructions to tree-decomposition, *J. Combin. Theory* Ser. B **52** (1991), 153–190.

[19] N. Robertson, P.D. Seymour: Graph Minors XIII. The disjoint paths problem, *J. Combin. Theory* Ser. B **63** (1995), 65–110.

[20] N. Robertson, P.D. Seymour: Graph minors XVII. Taming a vortex, *J. Combin. Theory* Ser. B **77** (1999), 162–210.

[21] N. Robertson, P.D. Seymour: Graph minors XVIII. Tree-decompositions and well-quasi-ordering, *J. Combin. Theory* Ser. B **89** (2003), 77–108.

[22] N. Robertson, P.D. Seymour: Graph minors XIX. Well-quasi-ordering on a surface, *J. Combin. Theory* Ser. B **90** (2004), 325–385.

[23] N. Robertson, P.D. Seymour: Graph Minors XX. Wagner's Conjecture *J. Combin. Theory* Ser. B (to appear).

[24] N. Robertson, P.D. Seymour and R. Thomas: Sach's linkless embedding conjecture, *Journal of Comb. Theory* Series B, **64** (1995), 185–227.

[25] P. Seymour: Disjoint paths in graphs, *Discrete Math.* **29** (1980), 293–309.

[26] P. Seymour: Decomposition of regular matroids, *Journal of Comb. Theory* Series B, **28** (1980), 305–359.

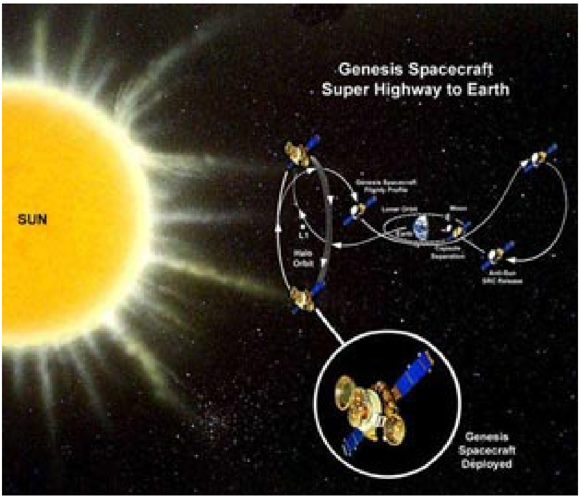
[27] C. Thomassen: 2-linked graphs, *Europ. J. Combin.* **1** (1980), 371–378.

[28] W.T. Tutte: Lectures on Matroids, *J. Res. Nat. Bur. Standards* **69B** (1965), 1–47.

[29] K. Wagner: Über eine Eigenschaft der ebenen Komplexe, *Mathematische Annalen* **114** (1937), 570–590.

[30] K. Wagner: Über eine Erweiterung des Satzes von Kuratowski, *Deutsche Mathematik* **2** (1937), 280–285.

[31] K. Wagner: *Graphentheorie*, B.J. Hochschultaschenbucher **248/248a**, Mannheim (1970), 61.



2005 CURRENT EVENTS COMMITTEE

David Eisenbud, Chair
Alejandro Adem
Dan Freed
Susan Friedlander
Edward Frenkel
Stuart Geman
Andrew Granville
Matthew Miller
John Morgan
Hugo Rossi
Michael Singer
Richard Stanley
Alan Weinstein

Produced by Raquel Storti
American Mathematical Society