Algebra and Complexity[1]

V. Strassen, Universität Konstanz

[1]Preprint of an article that appeared in the Proceedings of the first European Congress of Mahtematics, Paris July 1992, Birkhäuser Verlag (1994), 429-446

1 Matrix Multiplication

In this survey we shall discuss the relation between algebra and complexity by looking at a ubiquitous computational problem: Matrix multiplication. Our point of view will be asymptotic with regard to the size of the matrices considered.

For simplicity we will work over the field of complex numbers, although the main results covered in this paper hold over fields of any characteristic, algebraically closed or not. Once the groundfield is fixed we may define the so-called *exponent* ω of matrix multiplication, which controls the computational complexity of multiplying large matrices:

$$\omega := \inf\{\tau : L(\langle h, h, h\rangle) = O(h^\tau)\}. \tag{1}$$

Here $\langle h, h, h\rangle$ stands for the multiplication map of matrices of order h and $L(\langle h, h, h\rangle)$ denotes its complexity, i.e. the minimal number of arithmetical operations sufficient to compute the product of two generic matrices (by straightline algorithms). Thus ω is the smallest number τ, "smallest" in the sense of infimum, such that matrices of sufficiently high order h may be multiplied by an algorithm using only h^τ arithmetic operations.

Why is matrix multiplication so important? There are certainly many theoretical reasons, such as the role of the matrix algebra as a building block in Wedderburn's classification of semisimple associative algebras, or the fact that the structure maps of the classical simple Lie groups and Lie algebras have close ties with matrix multiplication. More generally, any finite dimensional associative algebra and any linear algebraic group may be respectively imbedded into a full matrix algebra and a general linear group.

Perhaps even more significant than these theoretical considerations is the central role that is played by matrix multiplication in numerical linear algebra. (Of course theoretical and practical aspects are not unrelated.) For example it can be shown that matrix inversion, evaluation of the determinant, computation of the coefficients of the characteristic polynomial, the construction of various normal forms, all have the same asymptotic exponent as matrix multiplication [24],[20],[1],[14].

The proofs of these facts are achieved by reduction. In order to show, for instance, that evaluating determinants is not much harder than multiplying matrices, one organizes a standard determinant algorithm as a recursive procedure via block decomposition with blocks of approximately half size. Each level of the recursion is made to consist essentially of a constant number of matrix multiplications of half size together with two calls of the procedure, i.e. two determinant evaluations for matrices of half size. This is easily seen to be sufficient for comparing exponents. In a similar, though sometimes more complicated, way one reduces any of the above mentioned problems of numerical linear algebra to matrix multiplication.

Reductions in the opposite direction are not always as straightforward. For instance, it is not clear how a fast determinant algorithm can be of any help for matrix multiplication, in particular since the latter problem requires the evaluation of many functions (the coefficients of the product matrix) rather than only one as in the case of the determinant.

Nevertheless a reduction is possible with the help of the so-called *derivative inequality* [1] for the complexity of a polynomial $f = f(x_1, \ldots, x_n)$:

$$L(f, \frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n}) \leq 4 \cdot L(f).$$

This inequality had originally been designed to prove meaningful lower complexity bounds for single polynomials via the degree bound [25]. The degree bound estimates the complexity of polynomials $f_1, \ldots, f_r$ in indeterminates $x_1, \ldots, x_n$ from below by the binary logarithm of the algebro-geometric degree of the graph of the polynomial map $\mathbf{C}^n \longrightarrow \mathbf{C}^r$ defined by $(f_1, \ldots, f_r)$. It yields the best results when n and r are large and have the same order of magnitude. On the other hand, for $r = 1$ the degree of the graph is simply the degree of the polynomial $f = f_1$ and the result is trivial. The derivative inequality is used by first replacing the complexity of f by the complexity of f together with its first order derivatives and then applying the degree bound. In this way one obtains lower bounds of order $n \log n$ for the complexity of various interesting polynomials, such as $\sum_{i=1}^n x_i^n$, $\prod_{i \neq j}(x_i - x_j)$ or any elementary symmetric function in the middle range. A nice consequence for interpolation is proved in [23].

Application of the derivative inequality to the determinant function shows that it is not much harder to compute the determinant of a matrix of order h together with all minors of order $h-1$ than it is to compute the determinant alone. Hence by Cramers rule matrix inversion is not much harder than determinant evaluation. An old trick reduces matrix multiplication to matrix inversion:

$$\begin{pmatrix} I & A & \\ & I & B \\ & & I \end{pmatrix}^{-1} = \begin{pmatrix} I & -A & AB \\ & I & -B \\ & & I \end{pmatrix}.$$

Here A, B are matrices of order h and I denotes the corresponding unit matrix. Thus matrices of order h may be multiplied by a single inversion of a matrix of order $3h$. (Actually, the argument is not cogent for the model of straightline algorithm, since it must be allowed that the domain of validity of such an algorithm be smaller than the domain of definition of the set of functions to be computed, in our case $GL(3h, \mathbf{C})$. A fast algorithm for matrix inversion may therefore be undefined on *every* matrix of the form appearing on the left hand side of the above matrix equation. The difficulty can be overcome, at least for comparing exponents, by a suitable perturbation [26].) Combining the three reductions indicated above one sees that a fast determinant evaluation yields a fast matrix multiplication algorithm, as claimed.

Let us return to the exponent ω defined in (1). The standard matrix multiplication algorithm shows that 3 is in the competition set of the infimum in (1), hence $\omega \leq 3$. It is also clear that $\omega \geq 2$, since any multiplication algorithm has to compute h^2 different functions.

The following table displays the progress that has been made in estimating ω from above by the time of writing. No lower bound > 2 is known.

$\omega <$			
2.81	(1969)	[24]	*RANK*
2.79	(1979)	[17]	
2.78	(1979)	[4], [3]	*BORDER RANK*
2.55	(1981)	[21]	τ − *THEOREM*
2.53	(1981)	[18]	
2.52	(1982)	[19]	
2.50	(1982)	[7]	*REFINEMENT OF THE* τ − *THEOREM*
2.48	(1987/8)	[27], [28]	*ASYMPTOTIC SPECTRUM*
2.38	(1990)	[8]	*RANDOMIZATION*

2 Bilinear Maps

More important than the various records of estimates are the general notions and techniques that have been introduced to make these estimates possible and some of which are indicated on the right hand side of the table. These notions and techniques apply or can be generalized to apply to arbitrary bilinear maps. In our survey the main objects of discussion will therefore be bilinear maps

$$f : U \times V \longrightarrow W$$

between finite dimensional complex vector spaces. Multiplication of general (rectangular) matrices provides instances of such maps. We use the Schönhage notation

$$\langle e, h, l \rangle : \mathbf{C}^{e \times h} \times \mathbf{C}^{h \times l} \longrightarrow \mathbf{C}^{e \times l}$$

for the multiplication of $(e \times h)$- with $(h \times l)$-matrices. Many good examples come as structure maps of associative algebras, Lie-algebras or their modules. In these cases we often denote the structure map by the same symbol as the underlying algebra or module. Thus for instance $\mathbf{C}[T]/(F)$, where F is a polynomial in T of degree n, denotes the bilinear map that multiplies polynomials of degree $< n$ and then reduces the product $\mathrm{mod} F$.

$f \simeq g$ means that f and g are *isomorphic*. An isomorphism is of course given by *three* vector space isomorphisms compatible with f and g. However, in the case when f and g are multiplication maps of associative algebras, the bilinear maps are isomorphic exactly when the algebras are.

The *direct sum* $f \oplus g$ is defined in the obvious way: Just execute f and g in parallel. Again, for multiplication maps of algebras the direct sum coincides with the multiplication map of the product algebra.

As a simple illustration of the notions of isomorphism and direct sum we have (by the chinese remainder theorem)

$$\mathbf{C}[T]/(T^n - \gamma) \simeq \mathbf{C} \oplus \ldots \oplus \mathbf{C} = \mathbf{C}^n \tag{2}$$

for $\gamma \neq 0$. $\mathbf{C}^n$ is the *diagonal map* of size n, i.e. the map of coordinatewise multiplication of vectors in $\mathbf{C}^n$.

The *tensor product* $f \otimes g$ of two bilinear maps f and g also generalizes the corresponding notion for associative algebras. $f \otimes g$ is a map that has a block decomposition such that each block looks like g whereas the superstructure is modelled after f. This is best visualized by looking at an example: Matrix multiplication of order $2h$ may be interpreted as multiplication of (2×2)-matrices, whose coefficients are not complex numbers but matrices of order h. This fact can be elegantly expressed in the language of tensor products:

$$\langle 2h, 2h, 2h \rangle \simeq \langle 2, 2, 2 \rangle \otimes \langle h, h, h \rangle.$$

By induction we obtain

$$\langle 2^N, 2^N, 2^N \rangle \simeq \langle 2, 2, 2 \rangle^{\otimes N}. \tag{3}$$

Loosely speaking, this says that large matrix multiplication maps are high tensor powers of (2×2)-matrix multiplication. (3) will be of considerable use later on.

3 Degeneration and Border Rank

In our presentation, the notion of *degeneration* of bilinear maps is central. It is a special case of the notion of degeneration (via orbit closure) in modules over linear algebraic groups (see e.g. [15]), but we prefer to give here a simple direct definition: $f \trianglelefteq g$ (f is a degeneration of g), when f is the limit of isomorphic copies of g. More precisely, we require that there is a sequence $g_1, g_2, \ldots$ of bilinear maps defined on the same vector spaces as f and each isomorphic to g, which converges to f in the classical topology. (The definition is not quite correct: It may happen that the spaces U, V, W on which f is defined are too small to accomodate isomorphic copies of g. In this case U, V, W are allowed to be enlarged, f being extended in a trivial way.) Here is an example:

$$\mathbf{C}[T]/(T^n) \trianglelefteq \mathbf{C}^n, \tag{4}$$

since for any nullsequence (γ_j) of nonzero complex numbers we have

$$\mathbf{C}[T]/(T^n - \gamma_j) \longrightarrow \mathbf{C}[T]/(T^n)$$

as $j \to \infty$, while $\mathbf{C}[T]/(T^n - \gamma_j) \simeq \mathbf{C}^n$ for all j by (2). Thus polynomial multiplication truncated at n (equivalently, truncated power series multiplication) is a degeneration of the diagonal map of size n. Note that for $n \geq 2$ the two maps in (4) are not isomorphic, since otherwise the underlying algebras would be isomorphic, contradicting the fact that only one of them has nilpotent elements.

We are now in a position to define a notion that actually appears in the list of records above, namely that of the *border rank* $\underline{R}(f)$ of a bilinear map f:

$$\underline{R}(f) := \min\{r : f \trianglelefteq \mathbf{C}^r\}. \tag{5}$$

Since the diagonal map $\mathbf{C}^r$ embodies r independent multiplications of complex numbers we may interpret $\underline{R}(f)$ as the price of f in the currency of single number-multiplications, paid by degeneration. As an example we have

$$\underline{R}(\mathbf{C}[T]/(T^n)) \leq n,$$

which is equivalent to (4).

The notion of border rank has been introduced by Bini, Capovani, Lotti, Romani [4] as a refinement of the earlier notion of *rank* [26] , which played a similar role and has similar properties as border rank. The definition of Bini, Capovani, Lotti, Romani is formulated in terms of power series and differs considerably from (5). The equivalence of the two definitions rests on a power series approximation lemma of Hilbert that is also crucial for the proof of the Hilbert-Mumford criterion for semistability. (See [11], [16], [27].)

What has the border rank to do with the exponent of matrix multiplication and, more generally, with the complexity of bilinear maps? Let us first clarify what we mean by the complexity of such a map, say of $f : U \times V \longrightarrow W$. The obvious definition is this: Choose bases in U,V and W. Then the complexity $L(f)$ of f with respect to these bases is the minimal number of arithmetic operations sufficient to compute the coordinates of $f(u, v)$ from the coordinates of u and v. Unfortunately, the value of $L(f)$ depends on the choice of bases. Fortunately, however, this dependence becomes irrelevant and in fact complexity may be replaced by border rank in asymptotic considerations. Although it is not clear at this point what "asymptotic" means for bilinear maps in general, it certainly should imply that complexity can be replaced by border rank in the definition (1) of the matrix exponent. This has indeed been shown by Bini [3]:

$$\omega = \inf\{\tau : \underline{R}(\langle h, h, h\rangle) = O(h^\tau)\}. \tag{6}$$

There are several advantages of border rank over complexity: As we have already observed, $\underline{R}(f)$ is an invariant of the map f, while $L(f)$ is not. Moreover, our definition (5) relates $\underline{R}$ to the deformation theory in modules over reductive groups, so the somewhat narrow path of the complexity theory of matrix multiplication opens to a wider landscape of mathematical knowledge and beauty. Of most immediate concern, however, is the fact that border rank has better formal properties than complexity:

$$\underline{R}(f \oplus g) \leq \underline{R}(f) + \underline{R}(g), \tag{7}$$

$$\underline{R}(f \otimes g) \leq \underline{R}(f) \cdot \underline{R}(g). \tag{8}$$

The subadditivity property is also shared by complexity, when bases are chosen in a proper way. Submultiplicativity, however, fails completely for L. Now it is this property that is most useful for estimating ω from above: Suppose that we have a good upper estimate of the border rank of some matrix multiplication, for example of $\langle 2, 2, 2\rangle$. Applying (8) inductively we obtain good border rank estimates for high tensor powers of $\langle 2, 2, 2\rangle$, i.e.

by (3) for large matrix multiplication maps. By (6) such estimates are exactly what is needed for good upper estimates of ω.

This line of reasoning together with a symmetrization argument (Hopcroft, Musinski [12]) can also be applied when we start with a border rank estimate for some *rectangular* matrix multiplication. It yields

$$\underline{R}(\langle e,h,l\rangle) \leq r \Longrightarrow (ehl)^{\omega/3} \leq r, \tag{9}$$

a result that is implicit in the work of Bini, Capovani, Lotti, Romani and enabled them to prove $\omega < 2.78$. The fact that this is only slightly better than previous estimates should not detract from the importance of their contribution of introducing border rank. (We prefer (9) to the equivalent and shorter statement $(ehl)^{\omega/3} \leq \underline{R}(\langle e,h,l\rangle)$, because (9) suggests a recipe for estimating ω.)

The cream was skimmed by Schönhage with his τ-Theorem:

Theorem 3.1 (Schönhage [21])

$$\underline{R}(\bigoplus_i \langle e_i,h_i,l_i\rangle) \leq r \Longrightarrow \sum_i (e_i h_i l_i)^{\omega/3} \leq r.$$

Theorem 3.1 differs from (9) only by the accomodation of direct sums. If border rank were additive instead of just subadditive, i.e. if we had equality in (7), the τ -Theorem would be a simple numerical consequence of (9). As things are, Schönhage had to introduce a novel recursion technique to prove Theorem 3.1. In the other direction, the τ -Theorem may be viewed as a kind of asymptotic verification of the additivity of border rank. Ironically, in the same paper [21] Schönhage disproves additivity by showing that

$$\underline{R}(\langle 4,1,4\rangle \oplus \langle 1,9,1\rangle) \leq 17, \tag{10}$$

while it is easy to see that $\underline{R}(\langle 4,1,4\rangle) = 16$ and $\underline{R}(\langle 1,9,1\rangle) = 9$. Note that the inequality (10) has the form of the premise of the τ -Theorem. The conclusion gives $\omega < 2.55$, a great improvement upon previous estimates.

Except for trivial cases the τ -Theorem is never sharp, as has been brought to light by Coppersmith-Winograd [7]. By proving a quantified version of this statement they were able to reach the landmark estimate $\omega < 2.5$. Their general observation is sometimes referred to as "ω is a limit point". This is quite misleading: For all we know complexity and border rank of $\langle h,h,h\rangle$ may be $O(h^\omega)$, regardless of whether $\omega > 2$ or $\omega = 2$.

We still have to explain what we mean by "asymptotics" for bilinear maps in general. There is no problem with matrix multiplication since these maps form an infinite sequence. A natural generalization is suggested by (3): We can consider any bilinear map as a member of the sequence of its tensor powers. This leads to the following definition of *asymptotic rank*, attributable to Gartenberg [9]:

$$\underset{\sim}{R}(f) := \lim_N \underline{R}(f^{\otimes N})^{1/N}. \tag{11}$$

Thus the asymptotic rank of a bilinear map f reflects the border rank of high tensor powers of f. The claim that border rank and complexity are asymptotically equivalent can now be given a precise meaning: in the definition of asymptotic rank we may replace $\underline{R}$ by L (defined via tensor product bases) without changing the result. Moreover, the motivating formula (3) together with (6) yield

$$\underset{\sim}{R}(\langle 2,2,2\rangle) = 2^{\omega}. \tag{12}$$

4 Asymptotic Spectrum

Inspired by K-theory we now shift attention from individual bilinear maps to their isomorphism classes. Actually, for reasons that we have indicated after the definition of degeneration we will work with slightly bigger equivalence classes, which allow the enlargment of the spaces on which a map is defined while extending the map in a trivial way. Taking maps between numerical spaces as representatives, the classes are seen to form a *set* $\mathcal{B}^+$:

$$\mathcal{B}^+ := \{\textit{classes of bilinear maps}\}.$$

Direct sum and tensor product induce operations in $\mathcal{B}^+$, which we write as addition and multiplication. Thus $\mathcal{B}^+$ becomes a commutative semiring with additive cancellation, and in much the same way as the semiring of natural numbers may be imbedded in the ring of integers, $\mathcal{B}^+$ is a positive cone in an essentially unique commutative ring $\mathcal{B}$ such that $\mathcal{B} = \mathcal{B}^+ - \mathcal{B}^+$. (A positive cone in a ring is a subset closed under addition and multiplication.) The additive structure of $\mathcal{B}$ is free abelian, hence $\mathcal{B}$ has characteristic 0. The natural number r interpreted in $\mathcal{B}$ is simply the class of the diagonal map $\mathbf{C}^r$. We call the elements of $\mathcal{B}^+$ *tensor classes*, since bilinear maps may be viewed as tensors, and the elements of $\mathcal{B}$ *generalized tensor classes.*

It is natural to transfer the notions of degeneration, border rank and asymptotic rank to tensor classes. Doing this we obtain a partial order $\trianglelefteq$ in $\mathcal{B}^+$ compatible with addition and multiplication, as well as two numerical functions $\underline{R}$ and $\underset{\sim}{R}$ on $\mathcal{B}^+$, which are related by

$$\underline{R}(b) = \min\{r : b \trianglelefteq r\} \tag{13}$$

and

$$\underset{\sim}{R}(b) = \lim_N \underline{R}(b^N)^{1/N} \tag{14}$$

for $b \in \mathcal{B}^+$. We have not lost ω through this abstraction: (12) remains valid when $\langle 2,2,2\rangle$ is interpreted as the *class* of matrix multiplication of order 2. We emphasize that everything takes place in $\mathcal{B}^+$. It is not possible, for instance, to extend $\trianglelefteq$ to all of $\mathcal{B}$ in a way which is compatible with addition and multiplication (Bürgisser [5]).

(7) and (8) imply that border rank is subadditive and submultiplicative as a function on $\mathcal{B}^+$. These properties are inherited by the asymptotic rank. It turns out that $\underset{\sim}{R}$ has an additional virtue: Take $b \in \mathcal{B}^+$. Then $\underset{\sim}{R}$ restricted to the subcone of $\mathcal{B}^+$ generated by b (which consists of all nonnegative integral polynomials in b), is actually additive and multiplicative. For $b = \langle 2,2,2 \rangle$ this follows from the τ-Theorem, for arbitrary b from a suitable generalization of its proof.

Thus $\underset{\sim}{R}$ is subadditive and submultiplicative on all of $\mathcal{B}^+$ and additive and multiplicative on any subcone generated by one element. Perhaps the reader is reminded of the maximum functional on the cone, say, of nonnegative continuous real functions on some compact space Δ. (When b is such a function, it assumes its maximum at some point $\delta \in \Delta$. It is clear that all nonnegative integral polynomials in b are also maximal at δ. This means that the maximum functional on the cone generated by b is simply "evaluation at δ" and is therefore additive and multiplicative.) The two situations are of course quite different: On the one hand tensor classes and a functional based on the degeneration of bilinear maps, on the other hand continuous functions and the maximum functional. Nevertheless there is a deeper connection, which rests in an asymptotic variant of degeneration order, called *asymptotic degeneration order*:

$$a \lesssim b :\Longleftrightarrow a^N \trianglelefteq b^{N+o(N)}. \tag{15}$$

Representing the classes a, b by bilinear maps f, g we can paraphrase (15) as follows: f is an asymptotic degeneration of g when high tensor powers of f are degenerations of slightly higher tensor powers of g.

It is easy to see that $\lesssim$ is a preorder in $\mathcal{B}^+$ compatible with addition and multiplication. Surprisingly $\lesssim$ (unlike $\trianglelefteq$) can be extended to all of $\mathcal{B}$ as a ring preorder. Moreover it satisfies a technical condition that turns $\mathcal{B}$ into a so-called Stone ring. The amount of work that goes into proving these facts is well rewarded: We may bring in the beautiful structure theory for Stone rings due to Stone, Kadison and Dubois [22], [13], [2].

We formulate our main result [28] relative to a subset X of $\mathcal{B}^+$, which serves to focus attention on those bilinear maps that we are interested in at a particular time.

Theorem 4.1 *Given $X \subset \mathcal{B}^+$, there is a compact Hausdorff space $\Delta(X)$ together with a ring homomorphism*

$$\varphi : \mathbf{Z}[X] \longrightarrow C(\Delta(X))$$

from the subring of $\mathcal{B}$ generated by X to the ring of continuous real functions on $\Delta(X)$, such that $\varphi(X)$ separates points and such that

$$\forall a, b \in \mathbf{Z}[X] \qquad a \lesssim b \Longleftrightarrow \varphi(a) \leq \varphi(b). \tag{16}$$

$\Delta(X)$ (together with φ) is uniquely determined up to canonical isomorphism and is called *asymptotic spectrum* of X. Its crucial property (16) implies that the asymptotic degeneration order between all classes of bilinear maps under discussion is faithfully reflected by the pointwise order between the associated functions on the asymptotic spectrum.

We can now explain the analogy between asymptotic rank and maximum functional:

Corollary 4.2
$$\forall b \in \mathbf{N}[X] \qquad \underset{\sim}{R}(b) = \max \varphi(b).$$

(In order to relate this to our previous discussion take $X = \mathcal{B}^+$.)

In connection with its definition (5) we have interpreted $\underline{R}(f)$ as the price of f in the sense of the minimal number of multiplications that can degenerate into f. Analogously $\underset{\sim}{R}(f)$ may be interpreted as the *asymptotic price* of f, i.e. the approximate normalized price of high tensor powers of f. When we blur the distinction between a tensor class b and a map f that represents it, we may rephrase Corollary 4.2 as

$$\max \varphi(b) = \text{asymptotic price of } b.$$

Similarly we have

$$\min \varphi(b) = \text{asymptotic value of } b,$$

where the *asymptotic value* of a map f is the approximate normalized *maximal* number of multiplications that can be degenerated *out of* high tensor powers of f. Clearly, value is always $\leq$ price, and it is easy to construct examples, e.g. by Theorem 5.2 below, where their ratio is arbitrarily small. We live in a real world.

5 Concrete Spectra

Theorem 4.1 and Corollary 4.2 show that in view of both the deformation and the complexity theory of bilinear maps it is a worthwhile task to compute asymptotic spectra. Let us now indicate how this can be done in general, though unfortunately only in principle.

According to the theory of Stone, Kadison and Dubois $\Delta(X)$ is the set (suitably topologized) of morphisms $\xi : \mathbf{Z}[X] \to \mathbf{R}$ of preordered rings, and φ is the evaluation map given by $\varphi(b)(\xi) := \xi(b)$. We explain this more concretely in the case of a *finite* set $X \subset \mathcal{B}^+$, say $X = \{x_1, \ldots, x_n\}$. Here the asymptotic spectrum of X can be canonically embedded as a compact subset in $\mathbf{R}^n$, such that the ring homomorphism φ is given as follows: By definition any $a \in \mathbf{Z}[X]$ may be written as an integral polynomial in $x_1, \ldots, x_n$, perhaps in several ways. Now if $a = A(x_1, \ldots, x_n)$ is one such representation, $\varphi(a)$ is simply the polynomial function A on $\mathbf{R}^n$ restricted to $\Delta(X)$.

Which points of $\mathbf{R}^n$ lie in $\Delta(X)$? Take bilinear maps $f_1, \ldots, f_n$ that represent the tensor classes $x_1, \ldots, x_n$. What we need are real valued functionals on the closure of $\{f_1, \ldots, f_n\}$ under $\oplus$ and $\otimes$, which are additive with respect to $\oplus$, multiplicative with respect to $\otimes$ and monotone with respect to $\trianglelefteq$. Whenever ζ is such a functional we have $(\zeta(f_1), \ldots, \zeta(f_n)) \in \Delta(X)$, and every point of the asymptotic spectrum can be obtained in this way. For example, let $\zeta(f) := \dim U$ for any bilinear map $f : U \times V \to W$. Then ζ is a functional with the above properties and therefore $(\zeta(f_1), \ldots, \zeta(f_n))$ is a spectral point. It is called the *first gauge point* of $\Delta(X)$. Of course there are two more gauge points, but they do not all have to be different.

The first asymptotic spectrum we should like to discuss is that of the set of square matrix multiplications:

$$\Delta_c := \Delta(\{\langle h,h,h\rangle : h \geq 1\}).$$

Since by (3) large matrix multiplications may be approximated by powers of $\langle 2,2,2\rangle$ it is not hard to see that Δ_c is canonically isomorphic to the asymptotic spectrum of $\langle 2,2,2\rangle$ alone. Hence it can be realized as a compact subset of the real line:

$$\Delta_c = \Delta(\langle 2,2,2\rangle) \subset \mathbf{R}^1.$$

Taking into account the identification made above one shows that

$$\forall h \geq 1 \qquad \varphi(\langle h,h,h\rangle) = x^{\log h}|\Delta_c, \tag{17}$$

where log always stands for the binary logarithm.

How does $\Delta_c = \Delta(\langle 2,2,2\rangle)$ look like? The first gauge point is 4, hence

$$4 \in \Delta_c.$$

The other gauge points yield nothing new. We can identify, however, the *largest* point of Δ_c: By formula (17), Corollary 4.2 and formula(12) we have

$$\max \Delta_c = \max_{\Delta_c} x = \max \varphi(\langle 2,2,2\rangle) = \underset{\sim}{R}(\langle 2,2,2\rangle) = 2^\omega, \tag{18}$$

in particular

$$2^\omega \in \Delta_c.$$

This simple fact allows for a 3-line proof of the τ-Theorem, here for square matrix multiplication: The premise is a border rank estimate, equivalent to a degeneration inequality. This implies the corresponding asymptotic inequality, which in turn is equivalent to an inequality of functions on the asymptotic spectrum Δ_c. Evaluation at 2^ω yields the conclusion.

It is unlikely that we could determine Δ_c completely at this time, since we would have to know the precise value of ω. The next theorem [28] falls just short of this:

Theorem 5.1

$$\Delta_c = [4, 2^\omega].$$

If we knew ω, the theorem would tell us everything about asymptotic degeneration between direct sums of square matrix multiplications. But also without this assumption it contains concrete and interesting information, for instance that the asymptotic value of $\langle 2,2,2\rangle$ is 4. (Use an argument analogous to (18).) Thus, unlike the cynic, who according to Oscar Wilde knows the price of everything and the value of nothing, we know the asymptotic value of $\langle 2,2,2\rangle$ precisely, while our knowledge of its asymptotic price is limited by our ignorance about ω. Surprisingly, the exact knowledge of the left end of Δ_c can be used to obtain an improved estimate for its right end, namely $\omega < 2.48$. The method employed is called *laser method* [27], since it is reminiscent of the generation of coherent light.

The laser method has been perfected by Coppersmith-Winograd [8]: Using a very clever probabilistic construction they were able to prove $\omega < 2.38$, the best estimate known at present. (See also [6].)

Let us say a few words about the asymptotic spectrum of the set of general (not necessarily square) matrix multiplications: It is canonically isomorphic to the spectrum of $\{\langle 2,1,1\rangle, \langle 1,2,1\rangle, \langle 1,1,2\rangle\}$, hence it can be canonically imbedded in $\mathbf{R}^3$. The gauge points are $(2,2,1), (1,2,2)$ and $(2,1,2)$. Their logarithmic convex hull belongs to the spectrum and is in fact the set of its minimal points with respect to the product order of $\mathbf{R}^3$. The whole spectrum is logarithmically star shaped relative to this subset and coincides with it exactly when $\omega = 2$.

Estimating asymptotic spectra from outside is usually achieved by constructing concrete degenerations. Estimating them from inside means exhibiting spectral points, which by a previous remark is equivalent to constructing additive, multiplicative and monotone functionals of bilinear maps. We now describe a method [29] for the design of such functionals, which is applicable in many situations of interest and leads to complete results in a few. (For $\langle 2,2,2\rangle$ it gives only the gauge point 4, for general matrix multiplication the logarithmic convex hull of the gauge points.)

In the same way as a linear map may be represented by a matrix, when bases have been chosen, a bilinear map f can be described by a spatial matrix, the so-called coordinate tensor (f_{ijk}) of f. By definition, the *support* of f (with respect to the given bases) is the set of triples $(i,j,k) \in \mathbf{N}^3$ such that $f_{ijk} \neq 0$. f is called *oblique*, when bases exist so that the support of f is an antichain with respect to the product order of $\mathbf{N}^3$. This means that there do not exist two different points (i,j,k) and (p,q,r) in the support of f such that $(i,j,k) \leq (p,q,r)$ coordinatewise. For example, when $f = \mathbf{C}[T]/(T^n)$ and when we choose the bases $(1, T, \ldots, T^{n-1})$ (twice) and $(T^{n-1}, \ldots, 1)$, the support of f is $\{(i,j,k) \in \mathbf{N}^3 : i+j+k = n-1\}$, which is an antichain. Hence f is oblique.

Now let $\theta \in \mathbf{R}^3$ be a probability vector, i.e. an element of the standard 2-simplex, and let f be any oblique map. Choose bases such that the support of f is an antichain and define

$$\zeta_\theta(f) := \max_P 2^{\sum_{j=1}^{3} \theta_j H(P_j)}, \tag{19}$$

where P ranges over probability measures on the support of f (a subset of $\mathbf{N}^3$), the P_j are the three marginal distributions of P and H denotes the entropy function. It turns out that $\zeta_\theta(f)$ does not depend on the choice of bases and that ζ_θ is an additive, multiplicative and monotone functional on the collection of oblique maps. (This collection is closed under $\oplus$ and $\otimes$.) Thus, whenever $x_1, \ldots, x_n$ are the classes of oblique maps $f_1, \ldots, f_n$, the point $(\zeta_\theta(f_1), \ldots, \zeta_\theta(f_n)) \in \mathbf{R}^n$ belongs to the asymptotic spectrum of $X = \{x_1, \ldots, x_n\}$. Since this point is easily seen to depend on θ in a continuous way we actually get a singular

2-simplex ζ of the compact space $\Delta(X)$. It is called the *support simplex* of $\Delta(X)$. By methods of convex optimization ζ may be computed very efficiently.

Although "oblique" is not a generic property, many prominent bilinear maps are oblique. For collections of such maps the image of the support simplex can be regarded as an obstruction to asymptotic degeneration. Even in simple cases this image is topologically nontrivial. For example, take all null-algebras (products are zero as far as is compatible with the existence of a unit element) and their rotations (rotate their coordinate tensor). The image of the support simplex turns out to be homeomorphic, but not diffeomorphic, to a 2-sphere.

At least when "oblique" is replaced by the somewhat stronger condition "*tight*" [29], the image of the support simplex can be shown to contain all minimal points of the asymptotic spectrum. Sometimes this suffices to determine the spectrum completely. Let us discuss one such case. Define a function z from nonnegative integers to positive reals by

$$z(m) := \frac{\mu^m - 1}{\mu - 1}\mu^{-2(m-1)/3},$$

where μ is the unique positive solution of

$$\frac{1}{\mu - 1} - \frac{m}{\mu^m - 1} = \frac{m-1}{3}.$$

z may be computed efficiently by Newton iteration. The following table gives a few values:

m	$z(m) \approx$
2	1.89
3	2.76
4	3.61
5	4.46
6	5.31
7	6.16
8	7.00
9	7.85
10	8.69
100	84.4
1000	842

The table indicates that $z(m)$ is always smaller than m. This should increase confidence in the following result [29]:

Theorem 5.2 *Given a polynomial $F \in \mathbf{C}[T]$ of positive degree, write*

$$F = \beta \prod_1^r (T - \alpha_i)^{m_i},$$

where $\alpha_i \neq \alpha_j$ for $i \neq j$. Then

$$\Delta(\mathbf{C}[T]/(F)) = [\sum_1^r z(m_i), \sum_1^r m_i].$$

Note that the right endpoint of the spectral interval is simply the degree of f and that neither endpoint involves a fuzzy number such as ω in Theorem 5.1. In order to decide whether one direct sum of tensor powers of $\mathbf{C}[T]/(F)$ degenerates into another, just compare the corresponding polynomials on the interval given in Theorem 5.2. This can be done for instance by Sturm's algorithm. When the second polynomial is not dominated by the first, degeneration is impossible, even asymptotically. Otherwise we have at least asymptotic degeneration.

We seem to have digressed far from our theme, the complexity of matrix multiplication. Actually this is not the case: In the same way as the left endpoint of the interval in Theorem 5.1 is the key to the estimate $\omega < 2.48$, the left endpoint of the interval in Theorem 5.2 in the simplest nontrivial case $F = T^2$ is the key to Coppersmith-Winograd's first estimate $\omega < 2.41$ in [8] (Their final estimate $\omega < 2.38$ involves a bilinear map that is not of the form $\mathbf{C}[T]/(F)$.)

Let us tie together the open ends of this survey with a conjecture. [Oder lieber nicht? Die Vermutung könnte falsch sein. Vermutlich *ist* sie falsch... Aber auch eine falsche ist besser als gar keine:]

Conjecture 5.3 *For sets of tight tensor classes the image of the support simplex coincides with the asymptotic spectrum.*

The conjecture implies $\omega = 2$, and this at least is believed by most experts.

As to *algebra and complexity*, in the previous discussion it would be difficult to even separate the two aspects from each other. While complexity theory perhaps profits most from this symbiosis, algebraic projects such as the deformation theory of tensors also benefit: The asymptotic point of view and the notion of asymptotic spectrum offer a new simplicity as well as a whole range of fascinating new problems.

References

[1] *W. Baur* and *V. Strassen,* The complexity of partial derivatives, Theor. Computer Science **22**(1982), 317–330.

[2] *E. Becker* and *N. Schwartz,* Zum Darstellungssatz von Kadison-Dubois, Archiv Math. **40** (1983), 421–428.

[3] *D. Bini,* Relations between EC-algorithms and APA-algorithms, applications, Nota interna B **79/8**, Pisa 1979.

[4] *D. Bini, M. Capovani, G. Lotti* and *F. Romani,* $O(n^{2.7799})$ complexity for matrix multiplication, Inf. Proc. Letters **8** (1979), 234–235.

[5] *P. Bürgisser,* Doctoral Thesis, Universität Konstanz 1990.

[6] *D. Coppersmith,* How not to multiply matrices, Preprint, IBM Research, T. J. Watson Research Center, Yorktown Heights, NY 10598, 1988.

[7] *D. Coppersmith* and *S. Winograd,* On the asymptotic complexity of matrix multiplication, SIAM J. Comp. **11** (1982), 472–492.

[8] *D. Coppersmith* and *S. Winograd,* Matrix Multiplication via arithmetic progressions, J. Symbolic Comput. 9/3, pp. 251-280 (1990).

[9] *P. A. Gartenberg,* Fast rectangular matrix multiplication, PhD thesis, Los Angeles 1985.

[10] *H. F. de Groote,* Lectures on the complexity of bilinear problems, Lect. Notes in Comp. Science **245**, Berlin-Heidelberg-New York 1987.

[11] *D. Hilbert,* Über die vollen Invariantensysteme, Math. Ann. **42** (1893), 313–373.

[12] *J. E. Hopcroft* and *J. Musinski,* Duality applied to the complexity of matrix multiplications and other bilinear forms, SIAM J. Comp. **2** (1973), 159-173.

[13] *R. V. Kadison,* A representation theory for commutative topological algebra, Mem. Amer. Math. Soc. **7** (1951).

[14] *W. Keller-Gehrig,* Fast algorithms for the characteristic polynomial, Theor. Comp. Sci. **36** (1985), 309–317.

[15] *H. Kraft,* Geometric methods in representation theory, in: Representations of Algebras, Workshop Proc., Puebla, Mexico 1980, LNM **944**, Berlin-Heidelberg-New York 1982.

[16] *H. Kraft,* Geometrische Methoden in der Invariantentheorie, Aspekte der Mathematik, Braunschweig 1984.

[17] *V. Ya. Pan,* Strassen's algorithm is not optimal, Proc. 19th IEEE Symposium on Foundations of Computer Science, Ann. Arbor, Mich. 1978, 166–176.

[18] *V. Ya. Pan,* New fast algorithms for matrix operations, SIAM J. Comput. **9** (1980), 321–342.

[19] *F. Romani,* Some properties of disjoint sums of tensors related to matrix multiplication, SIAM J. Comput. **11** (1982), 263–267.

[20] *A. Schönhage,* Unitäre Transformationen großer Matrizen, Num. Math. **20** (1973), 409–417.

[21] *A. Schönhage,* Partial and total matrix multiplication, SIAM J. Comp. **10** (1981), 434–455.

[22] *M. H. Stone,* A general theory of spectra I, Proc. N.A.S. **26** (1940), 280–283.

[23] *J. Stoss,* The complexity of evaluating interpolation polynomials, Theoret. Comput. Sci. **41** (1985), 319–323.

[24] *V. Strassen,* Gaussian Elimination is not optimal, Num. Math. **13** (1969), 354–356.

[25] *V. Strassen,* Die Berechnungskomplexität von elementarsymmetrischen Funktionen und von Interpolationskoeffizienten, Num. Math. **20/3** (1973), 238–251.

[26] *V. Strassen,* Vermeidung von Divisionen, J. reine angew. Math. **264** (1973), 184–202.

[27] *V. Strassen,* Relative bilinear complexity and matrix multiplication, J. reine angew. Math. **375/376** (1987), 406–443.

[28] *V. Strassen,* The asymptotic spectrum of tensors, J. reine angew. Math. **384** (1988), 102–152.

[29] *V. Strassen,* Degeneration and complexity of bilinear maps: Some asymptotic spectra, J. reine angew. Math. **413** (1991), 127–180.