Research and Tutorial Exposition

Gilbert Strang

Abstract— **My research is concentrated on applications of linear algebra in engineering, including wavelet analysis and structured matrices and (currently) approximation of large dense matrices by a mosaic of low rank blocks.**

Keywords— **Kalman filter, DCT, eigenvalues, distance learning**

I. Introduction

I have selected material from six recent papers that I hope will be of interest.

II. Tridiagonal Matrices and the Kalman Filter

The Kalman filter is an entirely natural algorithm that deserves to be well known in numerical linear algebra. It is often derived via conditional probabilities, but we describe it using ordinary elimination (an alternative is QR from Gram-Schmidt). The most convenient starting point for (block) tridiagonal matrices is their triangular factorization into $T = LDU$.

The reverse factorization $T = U_- D_- L_-$ from eliminating upwards is occasionally useful. The pivots in D and D_- give a neat formula for the inverses of the diagonal entries in T^{-1}, and we show how this formula $(T^{-1})^{-1}_{kk} = D_{kk} + (D_-)_{kk} - T_{kk}$ is applied in one part of the Kalman filter. The output from the filter is the solution $\widehat{x}$ to a block tridiagonal system, and also (most importantly) the diagonal blocks $(T^{-1})^{-1}_{kk}$ are covariance matrices for errors in $\widehat{x}$.

We also mention the corresponding formula when T is not tridiagonal but "tree-diagonal." Thus $T_{ij} = 0$ if the tree has no edge from node i to node j. These matrices share important properties of tridiagonal matrices (which come from straight trees with no branches).

It is an understatement to say that tridiagonal matrices are important in mathematics. We have no hope of describing this enormous subject, and we won't try. Instead we discuss one particular identity that we found intriguing, for the diagonal entries of the inverse matrix. Since the continuous analogue of a tridiagonal matrix is a second-order differential operator, there must be (and there is) a corresponding identity for the diagonal slice $x = y$ of the Green's function $G(x, y)$. It is tempting to include this too, with its application to the continuous Kalman filter. We are in the situation of Oscar Wilde, who could resist anything but temptation. But we did resist. We will stay with matrices.

The identity is equation (1) below. We learned it from a problem posed by Parlett; it was known earlier. Before starting on its applications, may we mention another identity (also for symmetric tridiagonal matrices) that has just now contributed to a remarkable formula of Dumitriu and Edelman [1]. This second identity, found by Paige, connects the first components q_i of the eigenvectors (unit length) to the characteristic polynomials of T and of its last $n-1$ rows and columns:

$$q_i^2 = \frac{P_{n-1}(\lambda_i)}{P_n'(\lambda_i)}$$

See pages—of Parlett's book [7] for proof. The first Dumitriu-Edelman formula gives the Vandermonde determinant of the eigenvalues (thus the discriminant of P_n):

$$\prod_{i<j} |\lambda_i - \lambda_j| = \frac{\prod (T_{i,i+1})^i}{\prod q_i}.$$

From this they are able to compute the Jacobian of a transformation of R^{2n-1} that is crucial for symmetric tridiagonal matrices. The mapping takes the matrix entries $T_{i,i}$ and $T_{i,i+1}$ to the eigenvalues λ_i and the q_i. Its Jacobian is beautifully simple:

$$J = \frac{\prod T_{i,i+1}}{\prod q_i}$$

The application in [1] is to the eigenvalue statistics for new classes of random matrices.

Please forgive one more extraneous remark before this paper begins. It concerns a quite different fact about the inverse of a tridiagonal matrix. *The inverse has low rank* (in fact rank one!) *above and below the main diagonal.* Thus the entries of T^{-1} have the form $u_i v_j$ for $i \le j$. Other band matrices have a corresponding property (low rank triangles). This underlies the success of the Greengard-Rokhlin Fast Multipole Method.

For large dense matrices arising from integral equations we very often see large blocks of (nearly) low rank. It seems probable that approximation by low rank blocks will lead to useful algorithms in numerical linear algebra. Our expository paper [6] is a guide to the original references.

III. The Discrete Cosine Transform

Each Discrete Cosine Transform uses N real basis vectors whose components are cosines. In the DCT-4, for example, the jth component of v_k is $\cos(j+\frac{1}{2})(k+\frac{1}{2})\frac{\pi}{N}$. These basis vectors are orthogonal and the transform is extremely useful in image processing. If the vector x gives the intensities along a row of pixels, its cosine series $\sum c_k v_k$ has the coefficients $c_k = (x, v_k)/N$. They are quickly computed from an FFT of length $2N$. But a direct proof of orthogonality, by calculating inner products, does not reveal how natural these cosine vectors are.

We prove orthogonality in a different way. Each DCT basis comes from the eigenvectors of a symmetric "second difference" matrix. By varying the boundary conditions

Math Department, 77 Massachusetts Avenue, Cambridge, MA 02139-4307

we get the established transforms DCT-1 through DCT-4. Other combinations lead to four additional cosine transforms. The type of boundary condition (Dirichlet or Neumann, centered at a meshpoint or a midpoint) determines the applications that are appropriate for each transform. The centering also determines the period: $N-1$ or N in the established transforms, $N-\frac{1}{2}$ or $N+\frac{1}{2}$ in the other four. The key point is that all these "eigenvectors of cosines" come from simple and familiar matrices.

Just as the Fourier series is the starting point in transforming and analyzing periodic functions, the basic step for vectors is the Discrete Fourier Transform (DFT). It maps the "time domain" to the "frequency domain". A vector with N components is written as a combination of N special basis vectors v_k. Those are constructed from powers of the complex number $w = e^{2\pi i/N}$:

$$v_k = \left(1, w^k, w^{2k}, \ldots, w^{(N-1)k}\right), \quad k = 0, 1, \ldots, N-1.$$

The vectors v_k are the columns of the Fourier matrix $F = F_N$. *Those columns are orthogonal.* So the inverse of F is its conjugate transpose, divided by $\| v_k \|^2 = N$. The Fourier series $x = \sum c_k v_k$ is $x = Fc$. The inverse $c = F^{-1}x$ uses $c_k = (x, v_k)/N$ for the (complex) Fourier coefficients.

Two points to mention, about orthogonality and speed, before we come to the purpose of this note. For these DFT basis vectors, a direct proof of orthogonality is very efficient:

$$(v_k, v_\ell) = \sum_{j=0}^{N-1} (w^k)^j (\bar{w}^\ell)^j = \frac{(w^k \bar{w}^\ell)^N - 1}{w^k \bar{w}^\ell - 1}.$$

The numerator is zero because $w^N = 1$. The denominator is nonzero because $k \neq \ell$. This proof of $(v_k, v_\ell) = 0$ is short but not very revealing. I want to recommend a different proof, which recognizes the v_k as *eigenvectors*. We could work with any circulant matrix, and we will choose below a symmetric A_0. Then linear algebra guarantees that its eigenvectors v_k are orthogonal.

Actually this second proof, verifying that $A_0 v_k = \lambda_k v_k$, brings out a central point of Fourier analysis. The Fourier basis diagonalizes every periodic constant coefficient operator. Each frequency k (or $2\pi k/N$) has its own frequency response λ_k. The complex exponential vectors v_k are important in applied mathematics because they are eigenvectors!

The second key point is speed of calculation. The matrices F and F^{-1} are full, which normally means N^2 multiplications for the transform and the inverse transform: $y = Fx$ and $x = F^{-1}y$. But the special form $F_{jk} = w^{jk}$ of the Fourier matrix allows a factorization into very sparse and simple matrices. This is the Fast Fourier Transform (FFT). It is easiest when N is a power 2^L. The operation count drops from N^2 to $\frac{1}{2}NL$, which is an enormous saving. But the matrix entries (powers of w) are complex.

The purpose of this note is to consider *real transforms that involve cosines.* Each matrix of cosines yields a Discrete Cosine Transform (DCT). There are four established types, DCT-1 through DCT-4, which differ in the boundary conditions at the ends of the interval. (This difference is crucial. The DCT-2 and DCT-4 are constantly applied in image processing; they have an FFT implementation and they are truly useful.) All four types of DCT are orthogonal transforms. The usual proof is a direct calculation of inner products of the N basis vectors, using trigonometric identities.

We want to prove this orthogonality in the second (indirect) *way.* The basis vectors of cosines are actually eigenvectors of symmetric second-difference matrices. This proof seems more attractive, and ultimately more useful. It also leads us, by selecting different boundary conditions, to four less familiar cosine transforms. The complete set of eight DCT's was found in 1985 by Wang [8], and we want to derive them in a simple way. We begin now with the DFT.

IV. Trees with Cantor eigenvalue distribution

We study a family of trees with degree k at all interior nodes and degree 1 at boundary nodes. The eigenvalues of the adjacency matrix have high multiplicities. As the trees grow, the graphs of those eigenvalues approach a piecewise-constant "Cantor function". For each value $\frac{m}{n}$, we will find the fraction of the eigenvalues that are given by $\lambda = 2\sqrt{k-1}\cos\left(\frac{\pi m}{n}\right)$.

V. Localized eigenvectors from widely spaced matrix modifications

This paper is about the eigenvalues and eigenvectors of familiar structured matrices, after changes in a small number of entries. The actual changes need not be small, so we refer to them as modifications rather than perturbations. The number of changes is small relative to the size of the matrix, because the modifications are required to be "widely spaced". They occur in entries that are far apart. They produce new eigenvectors that are localized in and near the components that correspond to changed rows. By knowing the approximate form of the eigenvectors, we also determine a very close (and simple) approximation to the eigenvalues.

Imagine a large number of nodes around a circle. Edges go only to the two neighbors of every node. Each row of the adjacency matrix A of this cyclic graph has two 1's. The matrix is a circulant with 1's on the first subdiagonal and superdiagonal, coming from the neighbors to the left and right. Now add a few edges going "across" the circle, so that the nodes involved are widely spaced. The modified graph has an adjacency matrix (symmetric if the added edges are undirected, but this is not required) with 1 in the i, j entry when an edge connects node i to node j. A typical example of our work is to find the "new" eigenvalues and eigenvectors of this modified matrix.

The first author mentioned in SIAM News the simplest case of this example. Only one undirected edge crosses the circle, from node i to a distant node j. This added edge modifies A by $a_{ij} = a_{ji} = 1$, in other words by a widely spaced submatrix with entries from $B = [01;10]$.

The modified matrix has two new 1's, far from the main diagonal. The two new eigenvalues are almost exactly $\sqrt{(5)}$ and $-\sqrt{(5)}$. The corresponding eigenvectors show a sum or difference of two spikes, as in Figure 1, centered at the positions i and j connected by the "shortcut edge". The remaining eigenvalues stay in the interval $[-2, 2]$ that contains all eigenvalues of the original A. Their eigenvectors still oscillate like the original eigenvectors, but orthogonality to the new ones produces the pinching at i and j that is illustrated by Figure 2.

This brief mention in SIAM News brought suggested proofs from three friends, Beresford Parlett and Bill Trench and Jackie Shen. All four approaches are different! Shen connected the problem to the theory of perturbed Schrodinger operators, and we believe that our work can be seen as a small contribution (possibly not new) to that established theory. We stay with this example in our first section, and we find the following formula linking the (nearly exact) new eigenvalues λ to the eigenvalues μ of B:

$$\lambda = (\text{sign } \mu)\sqrt{(4 + \mu^2)}.$$

The rank two perturbation from one undirected edge and $B = [0\ 1;\ 1\ 0]$ has $\mu = 1$ and -1, confirming that $\lambda = \sqrt{(5)}$ and $-\sqrt{(5)}$. In the two localized eigenvectors, the heights of the "spikes" are given by the eigenvectors of B. We also determine the ratio t between neighboring entries near positions i and j (a smaller t means a sharper spike and a more localized eigenvector). This pattern extends to any widely spaced modification by a nonsingular B.

Later sections of the paper extend the theory beyond the circle of nodes and its particular adjacency matrix A. We mention that an infinite string of nodes would give the same results, or even a finite string with a tridiagonal matrix A, provided the modifications occur far from the ends of the string (the first and last rows of A). The Laplacian matrix of a graph (a circle or a tree or an N-dimensional grid) is another important source of examples.

A. The Model Problem

We start with a line of nodes (the graph has a node at every integer). Its adjacency matrix A will be infinite, with 1's on the first subdiagonal and first superdiagonal: $a_i, i-1 = a_i, i+1 = 1$ for $-\infty < i < \infty$. The modification of A will be governed by an M by M matrix B, which need not be symmetric. We choose M widely spaced indices $r_1 < \ldots < r_M$; the differences between these indices all exceed a number $L \gg 1$. Then the i, j entry of B is added to the r_i, r_j entry of A. By a terrible abuse of notation, we call the modified matrix $A + B$. Our problem is to estimate the "new" eigenvalues and eigenvectors after the modification:

$$(A + B)x = \lambda x. \tag{1}$$

The key is that we expect each eigenvector x to be a sum of M spikes. For a given eigenvector, suppose the spike centered at the r_k entry of x has height h_k. The "spike ratio" between neighboring entries is denoted by t. Then the jth component of this eigenvector has the form

$$x_j = \text{ sum from } k = 1 \text{ to } M \text{ of } t^{|j - r_k|} h_k. \tag{2}$$

When $j = r_k$, the kth term reduces to h_k as desired. The ratio t will be different for different eigenvectors (t will depend on λ).

Now substitute this form for x into the equation $(A + B)x = \lambda x$. We distinguish the special rows $j = r_1, \ldots, r_M$ from the other rows (ordinary unmodified rows). In those ordinary rows there is no contribution from B. The matrix A has 1's to the left and right of the diagonal, which produce an extra factor of t and $1/t$ in every spike. The eigenvalue equation (1) in the ordinary rows becomes

$$(t + 1/t + 0)x_j = \lambda x_j. \tag{3}$$

In the special row $j = r_k$, where B has an effect, the equation is

$$2th_k + (Bh)_k = \lambda h_k + O(t^L). \tag{4}$$

You see the change. For the kth spike, centered on this special row, both 1's in A produce a factor t (thus $2t$). All the other spikes in x are of order t^L in this entry, because the special rows are far apart. For the same reason x_j on the right side equals $h_k + O(t^L)$.

Suppose we ignore the error $O(t^L)$. Then equation (4) says that the vector h of spike heights is an eigenvector of B. If that eigenvector has an eigenvalue μ, equations (4) and (3) become

$$2t + \mu = \lambda = t + 1/t. \tag{5}$$

Compare the left side with the right side, to find $t + \mu = 1/t$. This quadratic equation yields

$$t = 1/2(-\mu + / - \sqrt{(4 + \mu^2)}). \tag{6}$$

Choose the plus-minus sign to agree with the sign of μ, so that $|t| < 1$. Then substitute t into (5) to find

$$\lambda = 2t + \mu = (\text{ sign of } \mu)\sqrt{(4 + \mu^2)}. \tag{7}$$

This is the (approximate) relation between the new eigenvalue λ of $A + B$ and the eigenvalue μ of B. We want to prove that the error in (7) is of the same order $O(t^L)$ as the terms that were dropped.

Theorem 1: If μ is a non-repeated non-zero eigenvalue of the M by M matrix B, with eigenvector h of norm one, then λ in (7) and x in (2) are within $O(t^L)$ of an exact eigenvalue-eigenvector pair for the (infinite) modified matrix $A + B$.

Example 1: Suppose we change only a single entry on the main diagonal from 0 to 1. The modifying matrix is just $B = [1]$. For the (infinite) modified matrix, the localized eigenvector is exact! The eigenvalue is $\sqrt{(5)}$ and the spike ratio is the golden mean $t = 1/2(-1 + \sqrt{(5)})$. If the single 1 is the $(0, 0)$ entry, the jth component of the eigenvector is $t^{|j|}$.

For a finite matrix, this eigenvalue-eigenvector pair is only approximate. The approximation is good when the modified entry is near the center of the finite matrix and poor (see Figure) as it approaches the ends of the diagonal (where the limiting eigenvalue is ....).

Example 2: Connect three widely spaced nodes i, j, k by three undirected edges. In this case the modifying matrix is

$$B = [0\ 1\ 1;\ 1\ 0\ 1;\ 1\ 1\ 0].$$

Its eigenvalues are $\mu = 2, -1, -1$. The eigenvalues of the large matrix $A + B$ are approximately

$$\lambda = +\sqrt{(4+2^2)} = \sqrt{(8)} \text{ and } \lambda = -\sqrt{(4+(-1)^2)} = -\sqrt{(5)} \text{ (twice)}.$$

The eigenvector $h = (1, 1, 1)$ of the small matrix is correctly reflected in the eigenvector of $A + B$ for $\lambda = \sqrt{(8)}$. It is very nearly a sum of three equal spikes.

The other eigenvalue $\mu = -1$ is repeated. The eigenvalue $\lambda = -\sqrt{(5)}$ is also repeated (very nearly). But Theorem 1 cannot apply as it stands to the eigenvectors, because the small matrix has a plane of eigenvectors for $\mu = -1$. Since $\lambda = -\sqrt{(5)}$ is not exactly repeated, there is no corresponding plane for $A + B$. Experiment shows that its eigenvectors are sums of spikes at i, j, k with heights $h =$ and $h =$.

To extend this example, suppose the modification adds a complete graph on M nodes to the starting graph (which is still an infinite line of nodes connected only to their neighbors). The M by M matrix B has 0's on the diagonal and 1's everywhere else. Its largest eigenvalue $\mu = M - 1$ has eigenvector $h = (1, 1, ..., 1)$. Then the modified matrix $A + B$ has largest eigenvalue

$$\lambda = \sqrt{(4+(M-1)^2)} \text{ with } x = \text{ sum of equal spikes.}$$

The $M - 1$ remaining eigenvalues of the small matrix B all equal -1. Again this produces $\lambda = -\sqrt{(5)}$ as a multiple eigenvalue.

VI. Signal processing for everyone

In the past, signal processing was a topic that stayed almost exclusively in electrical engineering. It was only the specialists who applied lowpass filters to remove high frequencies from digital signals. The experts could cancel unwanted noise. They could compress the signal and then reconstruct. It took two-dimensional experts to do the same for images.

The truth is that everyone now deals with digital signals and images (involving large amounts of data). We all need to understand signal processing—*sampling, transforming, and filtering.* These pages are intended to explain these basic operations, using simple examples. We will reach as far as filter banks (in discrete time) and wavelet expansions (in continuous time).

Most signals start their lives in analog form. They become digital by sampling at equal time intervals. If $x_{\text{analog}}(t)$ is a *continuous time signal*, its samples give a discrete time signal:

$$x_{\text{digital}}(n) = x_{\text{analog}}(nT) \quad n = 0, \pm 1, \pm 2, \ldots \tag{8}$$

The *sampling interval* is T. We often normalize to $T = 1$, by a simple rescaling of the time variable.

A device that actually does this sampling is called an *A-to-D converter.* The input is an analog (A) signal, probably from measurements. The output is a digital (D) signal, probably for computer processing. Usually an A-to-D converter loses high frequency information (or mixes it with low frequencies, which is aliasing). Shannon's Theorem will tell us that when there are no high frequencies in the signal, the analog signal can be recovered at all t from its (digital) samples at the discrete times nT.

Notice that the signal is assumed to be infinitely long, with no start and no finish. The time line is $-\infty < t < \infty$. Then the discrete signal $x(n)$ is defined for all integers $(-\infty < n < \infty)$. Neither of these assumptions is exactly true for real signals. The realistic assumption (this is often well justified) is that the signal is so long that end effects are not significant. By working with the whole line $\mathbf{R}$ and all integers $\mathbf{Z}$, we can use Fourier methods to the utmost.

And those Fourier methods are very powerful. The chief tool in our analysis will be the Discrete Time Fourier Transform, which turns the samples $x(n) = x_{\text{digital}}(n)$ into the coefficients of a 2π-periodic function $X(\omega)$:

$$X(\omega) = \sum_{n=-\infty}^{\infty} x(n) e^{-in\omega}\,. \tag{9}$$

All terms are unchanged when ω is increased by 2π. We refer to ω as the *frequency*, and we graph the transform $X(\omega)$ between $\omega = -\pi$ and $\omega = \pi$. Then "low frequencies" refer to frequencies near zero, and "high frequencies" have $|\omega| \approx \pi$.

Two special signals have the lowest and highest frequencies, $\omega = 0$ and $\omega = \pi$. The pure DC signal $x = (\ldots, 1, 1, 1, 1, 1, \ldots)$ has exactly zero frequency. Its transform $X(\omega)$ has a Dirac delta function at $\omega = 0$. More precisely, $X(\omega)$ is a periodic train of delta functions of magnitude 2π. The pure AC signal $x = (\ldots, 1, -1, 1, -1, \ldots)$ has the highest frequency $\omega = \pi$ (and $\omega = -\pi$). Its transform is a train of delta functions at $\omega = \pm\pi, \pm 3\pi, \ldots$. This alternation between 1 and -1 gives the fastest oscillation of any discrete signal. Between $\omega = 0$ and $\omega = \pm\pi$ is the family of pure sinusoidal signals with frequency $-\pi < \omega < \pi$:

$$x_\omega(n) = e^{in\omega} \qquad \text{for each } n. \tag{10}$$

We are frequently working with systems that respond to these pure inputs with pure outputs. The output has *no change in frequency.* The only change is in amplitude and phase, from multiplication by $H(\omega)$. This is a Linear Time Invariant system:

LTI systems: *The input $x_\omega(n)$ produces the output $H(\omega)x_\omega(n)$.* (11)

The amplifying factor $H(\omega)$, also written $H(e^{i\omega})$, is the *frequency response*. It varies from one frequency to another, but separate frequencies stay separate. $H(\omega)$ is an "eigenvalue" of the system, when the eigenvector is the oscillating signal $x_\omega(n)$. A Linear Time Invariant system is often called a *filter*.

We will study filters in detail. First we look again at these special signals—complex exponentials and real sinusoids. Fourier (and Mozart too) assembled all signals out of these pure harmonics.

VII. Teaching and learning on the Internet

This paper comes from personal experience rather than philosophy. My experience is with MIT's linear algebra course and it is ongoing. You will see that I have recently got myself into some kind of box (a new box instead of the usual one). As a result I don't know exactly what to do in the linear algebra lectures this fall. Writing this paper in the summer of 2000 gives me a small chance to think through this rapidly approaching problem. It involves using the Internet, and videos in particular, in combination with ordinary lectures and homework. I believe that the reader will encounter the same problem, in some form, soon.

In spite of the opening sentence I suppose there is a "philosophy" that underlies my teaching of mathematics. Many of the students are learning engineering and science, and they care first of all about applications. This seems to fit with my approach. I get a lot of pleasure from showing them examples, and connecting with their interests, and convincing them that mathematics is directly useful. It is true that I use the words *beautiful* and *wonderful* to call their attention to ideas that are especially neat. But the beauty is alive and not frozen.

The only theorem that I mention by name is the "Fundamental Theorem of Linear Algebra". I would not want the rest of the faculty to know how seldom I complete a proof in the lectures. An example can be much more memorable anyway. Two examples are totally convincing! (My favorite proof remains the one I found in a book by Ring Lardner: *"Shut up" he explained.* But I only use this in class when desperate.)

Let me come directly to the recent events that present new problems.

1. My linear algebra lectures and review sessions last fall were videotaped live. They are on the web page `http://web.mit.edu/18.06/www`, and they can be viewed with (free) Real Player software. The compression makes my own motion a little jumpy, but the blackboard is surprisingly clear. So all students are going to have the lectures available when they want them (not only MWF at 1).

2. Independently of the videotaping, I joined with David Jerison and Haynes Miller in a proposal to a new funding source within MIT (established by a gift from Microsoft). Our proposal was to introduce "new communication links" in calculus (Jerison) and differential equations (Miller) and linear algebra (Strang). It was nearly successful but in the end was not funded for next year. I hope that some of the ideas might be of interest to readers of this article.

Those ideas are "speculative" and very much in flux. A second proposal, more directly involved with the structure of lectures and recitations, was awarded a planning grant from a different fund. Haynes and David are testing new possibilities in the calculus lectures. Eric Mazur's website `http://mazur-www.harvard.edu/education/educationmenu.html` has been a source of inspiration. They have very properly suggested that this paper should concentrate on my own part of the original proposal, which was to create an "online encyclopedia" of short and specific pieces of undergraduate mathematics. These will be quite different from complete lectures, but a camera will still be involved.

Note added in proof: I have now seen a device that records as you write on a whiteboard, without needing a cameraman (it transmits the writing but not the speaker). Combined with audio, this may become useful in communicating mathematics in real time.

THE VIDEOTAPES IN LINEAR ALGEBRA

I can explain first about the videotapes in 18.06. The year before, when Gian-Carlo Rota died so suddenly, I expressed to the class my regret that we had no permanent record of his lectures. They were exceptional in every way. The conversation in class moved toward more ordinary things, but several students emailed me afterward. They suggested that I contact the Center for Advanced Engineering Studies. I discovered that the Center was embarking on a large-scale videotaping project in physics (with Walter Lewin). Eventually we realized for a small additional cost, the cameramen could stay in the lecture room and tape the 18.06 lectures.

The original tapes were digitized and compressed (and saved) by David Mycue, in between his work on engineering classes that are running jointly with Singapore. This was all a part of MIT that I had never seen.

I insisted on only one point, that the lectures must be freely available to everyone. Modulo congestion on the web (which depends on the viewer's modem), this is now the case. I had no idea what use might be made of the tapes, it just seemed a good thing to try. I still have no idea! Readers of this paper are very welcome to make suggestions. I can mention two developments within MIT, and I hope there will be more outside:

1. In the semester of videotaping, there were a few classes that I had to miss. Those tapes were made in advance without an audience. I asked the students whether they would prefer to have a substitute teacher, but they firmly chose the tape. Apparently they did come to class and watch quietly.

You will realize the implications. One is that I can be away more and more—leaving a shadow of myself behind. On the other hand (and more seriously), the students can be away more and more. Let me come back to this new freedom, which is partly desirable and partly alarming.

2. The MIT Lincoln Laboratory learned about the tapes, and decided to offer a linear algebra course this summer.

The volunteer students are mature scientists and engineers, who watch two tapes each Thursday afternoon. My best teaching assistant, Peter Clifford, is there to answer questions. I went three times, to be part of the group and ask for their reactions. I frankly thought it would be a horrible experience to watch the tapes with the class, but it wasn't. They are seeing the uncompressed form, not so different from a live lecture. The volunteers at Lincoln Lab continue to attend and their comments are very positive.

I now realize more clearly and urgently that students will have an alternative to attending lectures this fall—or possibly the tapes will be more a supplement or a complement than an alternative. My question now is what to do in class when they can watch the lectures at their own convenience. I could vary the examples, and I certainly will. But I can't vary the mathematics...

18.06 splits into ten or more recitation sections, one hour per week, to discuss homework problems and unanswered questions. The lecture hours could now be more interactive (subject to the limitations of a large class). I am very much in favor of active learning, and I mix in questions as I go. I don't always wait for the answers! Students are hesitant to stand out in a large anonymous group, but I am learning about the successful use of flash cards and class votes.

In general it will not be possible to assume that students have watched the lectures in advance. Do I want to assign specific lectures as part of their homework (and risk developing a habit that will lead them to skip class)? The new situation offers more freedom, but with it comes change and uncertainty. Every innovation implies an altered set of rules. Most definitely, students have learned to deal successfully with the old rules. After long acquaintance, those rules are more or less accepted as fair. Any alteration implies that somehow or somewhere, an extra effort is required. This is likely to be unwelcome.

In the present case, a more active lecture hour might succeed. I have no means of compelling students to attend, and don't want any. I do already try to make the hour more productive for the class than an hour spent reading the textbook (which unfortunately I wrote). I was already competing with myself, and now even more so! Where I previously offered a focus on the more important points, and used the medium of speech to bring home those points, now the video lectures offer speech too, at all times of the day.

Will a live lecture three days a week be preferred to a videotape available at all hours, seven days a week? In the long run I really don't know. Perhaps in the short run, inertia (and maybe the lack of anything better to do) will bring most of them to the classroom.

And there is another question. Could the videotapes affect linear algebra classes around the country? The text is widely used. I am hopeful that instructors will welcome the availability of lectures on the web, as a supplement to their own courses. This is really a key question that will surely arise throughout our teaching—how to make the Internet into a "TA".

In short, I took this videotaping step in the belief that it could only be useful—not knowing exactly how, but certainly knowing that lectures on the web are sure to come. They will come in different forms, from different sources. If it becomes clearly helpful to add summaries of the lectures, or answers to frequently asked questions, or additional examples (all probably to be prepared on transparencies), I will try to do that. First I hope to learn how these videos can be used. I will be extremely grateful if readers of this article send thoughts and suggestions by email (to `gs@math.mit.edu`).

The Internal Proposal Within MIT

The second part of this paper will describe some aspects of the proposal that Haynes Miller, David Jerison, and I made in December 1999. We offered to create new on-line possibilities for the students, without a radical change in the existing lecture system. We were unwilling to destroy something that is pretty good (certainly not perfect). The new ideas would definitely need testing and adjustment, and a lot of work.

Our overall goal is to make the experience of freshmen and sophomores more active and positive. In every society, whether on the scale of a nation or a university or a family, there is tremendous constructive energy. Very often this is potential energy, and it is never released. To convert that stored potential into kinetic energy is a central goal of teachers (and of leaders wherever they are).

One tool we proposed to use is the Internet. We know that experiments are going forward in this direction in many mathematics departments. Undoubtedly the results are mixed—the same would surely be true for us. I can reproduce here a substantially edited summary of two ideas, from the proposal that the three of us prepared:

A. We will create an on-line encyclopedia to offer quick help in our basic mathematics courses. This material could extend beyond freshman courses to help all students who have access to the Internet.

The advantage of video and multimedia presentations over textbooks is that the added dimension of time can convey complicated (and also simple!) mathematical ideas more effectively. A critical advantage over lectures is that the information can be delivered in packets at the moment when it is needed. That moment can be the point at which a student gets stuck on a homework problem. It can also be the time in the next month or the next year when a specific piece of information is needed again.

B. We will establish Chat Rooms in which students will be able to discuss ideas and problems. They can manipulate the graphical tools that we plan to develop, and they can have have fun working together. We want MIT students to appreciate the *active* and *cooperative* elements of mathematics. Mathematics depends on communication.

The chat rooms could develop into a new feature of MIT education. Students already form study groups to do problem sets and to review for exams. We do not wish to lose the value of these interactive groups as we move to exploit the power of computer education for every individual. On-line chat rooms allow students to interact

even if they cannot do so in person, or if they do not have access to a compatible group. A group can use graphical and computational tools jointly, to go further with discovery than separate individuals. The transcripts of the chat rooms will be useful in monitoring the whole process and understanding where students get stuck.

Like most universities, MIT relies on lectures by experienced teachers. It would not be wise to overturn this framework tomorrow—the quality of those lecturers (and their dedication) is too valuable. But when we consider the whole experience from the viewpoints of the students, we do see new ways to communicate with them individually—*and new ways in which students can communicate with each other.*

It is changes in the student experience, made possible by the revolution in faculty-student and student-student communication links, that motivate this proposal. The first project is the most straightforward—a direct way for students to access (on line) essential concepts of each course. The second is the most novel—a way for students to talk to each other. We want them to work together, because in some unexpected way that promotes individual learning. Thus our presentation concentrates on these elements:

- Direct on-line access to quick help with the fundamental ideas of each course and their *applications* (with many examples).
- Student-student interaction on homework problems as well as central concepts. The magical moment of understanding generally happens outside the classroom.

May I return from the proposal to this MAA paper, for several comments. One is to repeat that in the actual development, changes in these ideas would be absolutely certain. I will add some details to the descriptions given above, but those changes are already coming. And there was another key part of our proposal that is not reflected in this paper—the idea of modules in differential equations, where the biologists and physical scientists and engineers are often interested in totally different examples.

Student-student on-line interaction.

Student learning often comes in bursts—frequently while working on problem sets. These are crucial times outside the classroom, and we can supply extra help. They are the moments we must concentrate on, because the students are concentrating. We will provide video clips to the students. But we are convinced that their communication *with each other* is a powerful force in learning. We will also provide graphical and computational tools that form integral parts of the problem set. The next paragraphs discuss three types of on-line support: Videos and graphical tools and student chat rooms.

1. Videos. The advantage over textbooks is that video, with the added dimension of time, can convey complicated (and also simple!) mathematical ideas more effectively. The critical advantage over lectures is that the information can be delivered in packets at the moment it is needed. Homework problems will in some cases be explicitly linked to video clips. Successful animations will also enliven our lectures.

2. Graphical Tools. These make a basic concept visual, and they also teach methods of discovery. In using Newton's method, calculus students can discover rates of convergence and periodic orbits and especially basins of attraction. At the same time they will reinforce their basic understanding of linear and quadratic approximations. By varying the coefficients of a quadratic function of two variables, they will see saddle points and maxima and minima. Graphical tools will be heavily used in the "technology-enabled" classrooms of the future.

3. Chat Rooms. Students already form study groups to do problem sets. But this includes only some students—it is good but not fully inclusive. On-line chat rooms allow students to interact even if they cannot do so in person, or if they do not have access to a compatible group. A group can use graphical and computational tools jointly, to go much further with discovery than separate individuals.

The transcripts of the chat rooms would be recorded. This makes it possible to see where students got stuck. To make better use of the transcripts, we propose to require students to cite their sources of help, with no penalty for accepting help. This will make it possible to track down abuses, such as wholesale copying of homework solutions, and (more important) to track down successful interactions.

Chat rooms should heat up before exams. Students will have access to support at late night crunch times. And especially at exam times, teaching staff could participate. In some cases the chat room can become an on-line recitation. Transcripts of these pre-exam interactions will be particularly helpful in course design. The TA could clip pieces and discuss the interaction with the lecturer.

An interesting problem is how students will communicate with each other in chat rooms if the keyboard is the only link. Mathematicians generally use an informal version of Donald Knuth's typesetting language TeX, where $\int_a^b f(x)dx$ means the definite integral of $f(x)$. There may be a better language for on-line mathematical conversations. And keyboards are presently inadequate for conveying pictures. The Holy Grail is the functional equivalent of a blackboard—which would support distance learning everywhere.

One can object that we are proposing to encourage last minute learning. But students learn when they are receptive, and not before. Any mechanism that makes learning easier must have its place. We may also include problems that specifically develop reading and listening skills, because lectures and a textbook are still central to the course. We could make the chat rooms off limits for certain homework exercises.

Furthermore we will use technology to help with learning *after* the last minute, in the form of post-testing. Most students do not go over their homework to see what was wrong. Even worse, many of them don't review their old tests. Teachers are guilty of encouraging this behavior, by racing to the next topic. This is a crucial learning opportunity and it is frequently lost. No experienced teacher expects everyone to learn subjects the first time. We aim to develop the habit of going back to learn and to reinforce

learning.
We can teach our students more successfully if we deliver information to them when they need it and are ready for it. Students can work cooperatively on line or in recitations. With their friends also on line, students will find it easier to learn to use these tools, to make progress on problems, and to have fun doing it—without the hesitation that we all recognize in a lecture room.
Our main goals are to engage the students more fully; to make better use of their time; to provide more inspired teaching; to encourage them to manipulate graphics in their own mathematical experiments; and to offer easy access to background information at the moment when it is needed. We hope for, and we expect, changes in style and substance.

That concludes the part of this paper which is drawn from our joint proposal. I owe thanks in so many ways to Haynes Miller and David Jerison. And over a very long period, too long to contemplate, I have learned from an army of students. To be truthful, I have seldom thought deeply or carefully about theories of education—it has always been more instinct in a classroom, a feeling for what students might understand and enjoy. And I read aloud (quietly) everything I write, so the same instincts are in control there too.

The arrival of the Internet has opened tremendous new possibilities. Just in time.

This article was included in a recent book published by the Math Association of America.

References

[1] I. Dumitriu and A. Edelman, Matrix models for all beta ensembles, *to appear*, 2001.

[2] Gilbert Strang, Xiangwei Liu and Susan Ott Localized eigenvectors from widely spaced matrix modifications, *in preparation*, 2001.

[3] Gilbert Strang, eds V. Capasso, H. Engl, and J. Periaux. Signal processing for everyone, Computational Mathematics Driven by Industrial Problems, *Springer Lecture Notes in Mathematics* **1739**, 2000.

[4] Gilbert Strang, X.Q. Gao and T. Nguyen Two-channel complex-valued filter banks and wavelets with orthogonality and symmetry properties, *submitted to IEEE Transactions on Signal Processing*, (2000).

[5] Gilbert Strang, G. Boyd, C. Micchelli, and D.X. Zhou Binomial matrices and discrete Taylor series, *in preparation*, 2001.

[6] Gilbert Strang Tridiagonal matrices and low rank inverses, *for submission to SIAM Review*, 2001.

[7] B.N. Parlett *The Symmetric Eigenvalue Problem*, Prentice-Hall (1980).

[8] Z. Wang and B. Hunt, The discrete W-transform, Appl. Math. Comput. **16** (1985) 19–48.

Biography

Gilbert Strang was an undergraduate at MIT and a Rhodes Scholar at Balliol College, Oxford. His doctorate was from UCLA and since then he has taught at MIT. He has been a Sloan Fellow and a Fairchild Scholar and is a Fellow of the American Academy of Arts and Sciences. He is a Professor of Mathematics at MIT and an Honorary Fellow of Balliol College.

Professor Strang has published a monograph with George Fix, "An Analysis of the Finite Element Method", and six textbooks:

Introduction to Linear Algebra (1993,1998)
Linear Algebra and Its Applications (1976,1980,1988)
Introduction to Applied Mathematics (1986)
Calculus (1991)
Linear Algebra, Geodesy, and GPS, with Kai Borre (1997)

He served as President of SIAM during 1999 and 2000. His home page is `http://www-math.mit.edu/~ gs`