Improving computations by using low complexity matrix algebras

FIRST LECTURE, Friday August 29, 2014 (Moscow, LMSU)

Bernoulli polynomials and numbers

Set $B_0(x) = 1$, and let $B_n(x)$ the degree n polynomial uniquely defined by the following two conditions

$$B_n(x+1) - B_n(x) = nx^{n-1}, \quad \int_0^1 B_n(x)\,dx = 0$$

(the proof of the fact that B_n is uniquely defined is left to the reader, it could be a further exercise).

First observe that any B_n is monic, in fact, if $B_n(x) = a_n x^n + a_{n-1}x^{n-1} + \dots$ then

$$\begin{aligned} B_n(x+1) - B_n(x) &= (a_n(x+1)^n + a_{n-1}(x+1)^{n-1} + \dots) - (a_n x^n + a_{n-1}x^{n-1} + \dots) \\ &= [a_n(x^n + nx^{n-1} + \dots) + a_{n-1}(x^{n-1} + (n-1)x^{n-2} + \dots) + \dots] - (a_n x^n + a_{n-1}x^{n-1} + \dots) \\ &= a_n n x^{n-1} + (\cdot)x^{n-2} + \dots \end{aligned}$$

and thus the condition $B_n(x+1) - B_n(x) = nx^{n-1}$ implies $a_n n = n$, i.e. $a_n = 1$.

Let us compute $B_1(x)$ and $B_2(x)$.

$B_1(x)$ is of the form $B_1(x) = x + \beta$, thus $B_1(x+1) - B_1(x) = (x+1+\beta) - (x+\beta) = 1$, i.e. the first condition is satisfied $\forall\,\beta$. Moreover, $\int_0^1 (x+\beta)\,dx = [x^2/2 + \beta x]_0^1 = \frac{1}{2} + \beta$ must be zero, so $B_1(x) = x - \frac{1}{2}$.

$B_2(x)$ is of the form $B_2(x) = x^2 + \alpha x + \beta$, thus $B_2(x+1) - B_2(x) = ((x+1)^2 + \alpha(x+1) + \beta) - (x^2 + \alpha x + \beta) = 2x + 1 + \alpha$, and the first condition is satisfied if $\alpha = -1$. Moreover, $\int_0^1 (x^2 - x + \beta)\,dx = [x^3/3 - x^2/2 + \beta x]_0^1 = 1/3 - 1/2 + \beta$ must be zero, so $B_2(x) = x^2 - x + 1/6$.

The numbers $\{B_{2j}(0)\}_{j=0}^{+\infty}$ are known as Bernoulli numbers. We can say that $B_0(0) = 1$, $B_2(0) = 1/6$. Euler guessed the following "explicit" formula for the generic Bernoulli number:

$$B_{2j}(0) = (-1)^{j+1}\frac{2(2j)!}{(2\pi)^{2j}}\zeta(2j), \quad \zeta(s) = \sum_{k=1}^{+\infty}\frac{1}{k^s}.$$

The value in 0 of degree odd Bernoulli polynomials is less interesting. In fact, $B_1(0) = -\frac{1}{2}$, and $B_{2j+1}(0) = 0$, $j = 1, 2, \dots$ (see below).

Recall that $\sum_{x=1}^n x^j$ is n for $j = 0$, is $n(n+1)/2$ if $j = 1$, is $(n(n+1)/2)^2$ if $j = 3$, and for $j = 2$ and j generic? Note that

$$\sum_{x=1}^n x^j = \sum_{x=1}^n \frac{1}{j+1}(B_{j+1}(x+1) - B_{j+1}(x)) = \frac{1}{j+1}[B_{j+1}(n+1) - B_{j+1}(1)].$$

It follows that $\varphi_j(n) := \sum_{x=1}^n x^j$ can be explicitely written in terms of the values of B_{j+1} in $n+1$ and 1. For example, in order to write $\sum_{x=1}^n x^2$ it is sufficient to know the values of the third Bernoulli polynomial B_3 in $n+1$ and 1.

Let us deduce B_3 by a way different from that used to deduce B_1 and B_2.

Exercise (IMP). Prove that

$$B_n(1-x) = (-1)^n B_n(x), \quad B'_{n+1}(x) = (n+1)B_n(x).$$

(hint: note that also the polynomials $(-1)^n B_n(1-x)$ and $\frac{1}{n+1}B'_{n+1}(x)$ satisfy the conditions that uniquely define $B_n(x)$).

The first of the equalities in the exercise implies that the Bernoulli polynomials are, with respect to $x = \frac{1}{2}$, symmetric if n is even and antisymmetric if n is odd. In particular, we have $B_n(1) = (-1)^n B_n(0)$, $n = 0, 1, 2, \ldots$. But we also know that $B_n(1) = B_n(0)$, $n = 0, 2, 3, 4, \ldots$ (by the first condition defining Bernoulli polynomials). Thus,

$$B_n(1) = B_n(0) = 0, \quad n \text{ odd }, \ n \neq 1.$$

The same equality in the Exercise (IMP) also implies $B_n(\frac{1}{2}) = (-1)^n B_n(\frac{1}{2})$, from which we deduce that

$$B_n(\frac{1}{2}) = 0, \quad n \text{ odd}.$$

It follows that

$$B_1(x) = x - \frac{1}{2}, \quad B_3(x) = x(x - \frac{1}{2})(x - 1), \quad B_{2j+1}(x) = x(x - \frac{1}{2})(x - 1)q_{2j-2}(x), \ j = 2, 3, \ldots,$$

where q_{2j-2} is monic of degree $2j - 2$. Now that we know B_3 we can say that

$$\sum_{x=1}^{n} x^2 = \frac{1}{3}[B_3(n + 1) - B_3(1)] = \frac{1}{3}(n + 1)(n + 1 - \frac{1}{2})(n + 1 - 1) = \frac{1}{6}n(n + 1)(2n + 1)$$

Exercise. Prove that $B_{2j+1}(x)$ $(j \geq 1)$ is zero in $[0, 1]$ if and only if $x = 0, \frac{1}{2}, 1$.

The second equality in Exercise (IMP) let us observe that the zeros of B_n are all the stationary points for B_{n+1}, and viceversa. So, for example, in $[0, 1]$ the only stationary points of the even degree Bernoulli polynomials B_n are $x = 0, \frac{1}{2}, 1$ (for $n = 2$ only $x = \frac{1}{2}$). Moreover, the second equality yields the following integral formula for $B_n(x)$

$$B_n(x) = B_n(0) + n \int_0^x B_{n-1}(t)\, dt.$$

Finally, note that, by the Euler formula, the Bernoulli numbers $B_{2j}(0)$ are rational numbers, and $|B_{2j}(0)| \to +\infty$ if $j \to +\infty$.

Exercise. Prove that in $[0, 1]$ if $j \to +\infty$, then $B_{2j}(x)/B_{2j}(0) \to \cos(2\pi x)$.

Exercise. Prove that for $x \in [0, 1]$ one has $|B_{2j}(x)| \leq |B_{2j}(0)|$, $\forall\, j$. (It would be sufficient to prove that $|B_{2j}(\frac{1}{2})| \leq |B_{2j}(0)|$, why?)

Bernoulli numbers are involved in the well known Euler-Mclaurin summation formula:

$$m, n \in \mathbb{Z}, \ m < n, \quad \sum_{r=m}^{n} f(r) = \frac{1}{2}(f(m) + f(n)) + \int_m^n f(x)\, dx + \sum_{j=1}^{k} \frac{B_{2j}(0)}{(2j)!}[f^{(2j-1)}(n) - f^{(2j-1)}(m)] + u_{k+1},$$

$$u_{k+1} = \frac{1}{(2k + 1)!} \int_m^n f^{(2k+1)}(x) \overline{B}_{2k+1}(x)\, dx$$

where $\overline{B}_n(x)$ is the periodic extension on $\mathbb{R}$ of $B_n(x)|_{[0,1)}$.

Proof. The starting point is to integrate by parts $\int_k^{k+1} f'(x)\overline{B}_1(x)\, dx$. The main steps of the proof are in the Appendix to the first lecture. $\square$

The euler-Mclaurin formula can be used to compute approximations of $\zeta(s)$, $s > 1$, for example of the yet mysterious number $\zeta(3) = \sum_{k=1}^{+\infty} 1/k^3$ (set $f(r) = 1/r^3$ and let n go to infinite), or it can be used to obtain a formula for the error of the trapezoidal rule in approximating $\int_a^b g(x)\, dx$ (set $f(t) = g(a + th)$, $h = (b - a)/n$,

$m = 0$; recall that the trapezoidal rule is $\mathcal{I}_h = h[\frac{g(a)}{2} + \frac{g(b)}{2} + \sum_{i=1}^{n-1} g(a+ih)]$). For some more details, see the Appendix to the first lecture.

Bernoulli numbers solve a lower triangular semi-infinite linear system

Now let us show that Bernoulli numbers solve a semi-infinite lower triangular linear system. It is known that

$$\frac{te^{xt}}{e^t - 1} = \sum_{n=0}^{+\infty} \frac{B_n(x)}{n!} t^n,$$

$$\frac{t}{e^t - 1} = \sum_{n=0}^{+\infty} \frac{B_n(0)}{n!} t^n = -\frac{1}{2}t + \sum_{k=0}^{+\infty} \frac{B_{2k}(0)}{(2k)!} t^{2k}.$$

Multiply the latter identity by $e^t - 1$, expand e^t in terms of powers of t, and set to zero the coefficients of t^j of the right hand side, $j = 2, 3, 4, \ldots$:

$$\begin{aligned} t &= \big(-\frac{1}{2}t + \sum_{k=0}^{+\infty} \frac{B_{2k}(0)}{(2k)!} t^{2k}\big)\big(\sum_{r=0}^{+\infty} \frac{t^{r+1}}{(r+1)!}\big), \\ t &= -\frac{1}{2}\sum_{j=2}^{+\infty} \frac{t^j}{(j-1)!} + \sum_{k,r=0}^{+\infty} \frac{B_{2k}(0) t^{2k+r+1}}{(2k)!(r+1)!}, \\ t &= -\frac{1}{2}\sum_{j=2}^{+\infty} \frac{t^j}{(j-1)!} + \sum_{j=1}^{+\infty} \sum_{k=0}^{[(j-1)/2]} \frac{B_{2k}(0) t^j}{(2k)!(j-2k)!}, \end{aligned}$$

$$-\frac{1}{2}\frac{1}{(j-1)!} + \sum_{k=0}^{[(j-1)/2]} \frac{B_{2k}(0)}{(2k)!(j-2k)!} = 0, \quad j = 2, 3, 4, 5, \ldots.$$

Thus

$$-\frac{1}{2}j + \sum_{k=0}^{[\frac{j-1}{2}]} \binom{j}{2k} B_{2k}(0) = 0, \quad j = 2, 3, 4, 5, \ldots \ .$$

In particular, for $j = 2, 4, 6, 8, \ldots$, we obtain the equations:

$$W\mathbf{b} = \begin{bmatrix} \binom{2}{0} & & & & \\ \binom{4}{0} & \binom{4}{2} & & & \\ \binom{6}{0} & \binom{6}{2} & \binom{6}{4} & & \\ \binom{8}{0} & \binom{8}{2} & \binom{8}{4} & \binom{8}{6} & \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix} \begin{bmatrix} B_0(0) \\ B_2(0) \\ B_4(0) \\ B_6(0) \\ \cdot \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ \cdot \end{bmatrix},$$

i.e. the lower triangular linear system we were looking for.

So, we can for instance easily compute the first Bernoulli numbers:

$$1, \ \frac{1}{6}, \ -\frac{1}{30}, \ \frac{1}{42}, \ -\frac{1}{30}, \ \frac{5}{66}, \ -\frac{691}{2730}, \ \frac{7}{6}, \ B_{16}(0) = -\frac{3617}{510}.$$

In the next lecture it will be shown that the coefficient matrix W of the above linear system turns out to have an analytic representation. In order to prove this fact, it is enough to observe that W is a suitable

submatrix of the Tartaglia-Pascal matrix which can be represented as a power series: $X = \sum_{k=0}^{+\infty} \frac{1}{k!} Y^k$, where

$$Y = Z \cdot \operatorname{diag}\big(i : i = 1, 2, 3, \dots\big), \quad Z = \begin{bmatrix} 0 & & \\ 1 & 0 & \\ & 1 & \cdot \\ & & \cdot \end{bmatrix}.$$

The analytic representation of W will let us introduce a lower triangular Toeplitz linear system solved by Bernoulli numbers.

A FINAL REMARK on the first lecture.

In [Ramanujan, 1919] Ramanujan writes explicitly 11 sparse equations solved by the absolute values of the 11 Bernoulli numbers $B_2(0)$, $B_4(0)$, ..., $B_{22}(0)$. They are the first of an infinite set of sparse equations solved by the absolute values of all the Bernoulli numbers. The Ramanujan equations, written together and directly in terms of the $B_{2i} := B_{2i}(0)$, $i = 1, 2, \dots, 11$, form the linear system displayed here below:

$$\begin{bmatrix} 1 & & & & & & & & & & & \\ 0 & 1 & & & & & & & & & & \\ 0 & 0 & 1 & & & & & & & & & \\ \frac{1}{3} & 0 & 0 & 1 & & & & & & & & \\ 0 & \frac{5}{2} & 0 & 0 & 1 & & & & & & & \\ 0 & 0 & 11 & 0 & 0 & 1 & & & & & & \\ \frac{1}{5} & 0 & 0 & \frac{143}{4} & 0 & 0 & 1 & & & & & \\ 0 & 4 & 0 & 0 & \frac{286}{3} & 0 & 0 & 1 & & & & \\ 0 & 0 & \frac{204}{5} & 0 & 0 & 221 & 0 & 0 & 1 & & & \\ \frac{1}{7} & 0 & 0 & \frac{1938}{7} & 0 & 0 & \frac{3230}{7} & 0 & 0 & 1 & & \\ 0 & \frac{11}{2} & 0 & 0 & \frac{7106}{5} & 0 & 0 & \frac{3553}{4} & 0 & 0 & 1 & \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix} \begin{bmatrix} B_2 \\ B_4 \\ B_6 \\ B_8 \\ B_{10} \\ B_{12} \\ B_{14} \\ B_{16} \\ B_{18} \\ B_{20} \\ B_{22} \\ \cdot \end{bmatrix} = \begin{bmatrix} \frac{1}{6} \\ -\frac{1}{30} \\ \frac{1}{42} \\ \frac{1}{45} \\ -\frac{1}{132} \\ \frac{4}{455} \\ \frac{1}{120} \\ -\frac{1}{306} \\ \frac{3}{665} \\ \frac{1}{231} \\ -\frac{1}{552} \\ \cdot \end{bmatrix}.$$

For example, by using the last but three of Ramanujan equations, from the Bernoulli numbers $B_2(0), \dots, B_{16}(0)$ listed previously, the following further Bernoulli numbers can be easily obtained:

$$B_{18}(0) = \frac{43867}{798}, \quad B_{20}(0) = -\frac{174611}{330}, \quad B_{22}(0) = \frac{854513}{138}.$$

Exercise. Let R be the coefficient matrix of the above Ramanujan linear system. Prove that there exists a diagonal matrix D such that $RD = D\tilde{R}$ where $\tilde{R}$ is lower triangular Toeplitz and sparse (for a more precise statement of this exercise see the Appendix to the first lecture).

APPENDIX TO THE FIRST LECTURE

On the Euler-Mclaurin formula:

Proof. First observe that

$$\int_k^{k+1} f'(x)\overline{B}_1(x)\,dx = \frac{1}{2}(f(k) + f(k+1)) - \int_k^{k+1} f(x)\,dx.$$

Then sum on k from m to $n-1$, call u_1 the integral $\int_m^n f'(x)\overline{B}_1(x)\,dx$, and set

$$u_j = \frac{1}{(2j-1)!} \int_m^n f^{(2j-1)}(x)\overline{B}_{2j-1}(x)\,dx.$$

Find a formula for u_j in terms of u_{j+1} and apply it for $j = 1, 2, \dots, k$. ...

Remark. If $f^{(2k+2)}$ does not change sign in $[m,n]$, then

$$|u_{k+1}| \leq 2\frac{|B_{2k+2}(0)|}{(2k+2)!}|f^{(2k+1)}(n) - f^{(2k+1)}(m)|.$$

This result allows to bound the error in the approximations of $\zeta(s)$ calculated via the Euler-Mclaurin formula.

Remark. Set $\mathcal{I} = \int_a^b g(x)\,dx$, $h = \frac{b-a}{n}$, $\mathcal{I}_h = h[\frac{g(a)}{2} + \sum_{i=1}^{n-1} g(a+ih) + \frac{g(b)}{2}]$. An expression of R in the equality $\mathcal{I} = \mathcal{I}_h + R$ can be found by using Euler-Mclaurin formula for $m = 0$ and $f(t) = g(a+th)$. Note that $R = c_1h^2 + c_2h^4 + c_3h^6 + \ldots$ and this justifies the efficiency of the number

$$\tilde{\mathcal{I}}_h = \frac{2^2\mathcal{I}_h - \mathcal{I}_{2h}}{2^2 - 1}$$

as an approximation of $\mathcal{I}$ ($\mathcal{I} - \tilde{\mathcal{I}}_h = O(h^4)$).

• $\zeta(3) = \sum_{r=1}^{+\infty} \frac{1}{r^3}$ is an irrational number (proved in 1973). It is not known if it is algebraic or trascendental. Show that the Euler-Mclaurin formula can be used to obtain approximations of $\zeta(3)$ as good as possible (hint: choose $m > 1$ and let n go to infinite).

• The Eulero-Mascaroni constant is

$$\gamma = \lim_{n\to+\infty}(\sum_{k=1}^{n}\frac{1}{k} - \log_e n).$$

Find approximations as good as possible of γ by using the Euler-Mclaurin formula.

• Find a diagonal matrix D and a lower triangular Toeplitz matrix $\tilde{R}$ of the following type

$$\tilde{R} = \sum_{k=0}^{+\infty}(\cdot)_k Z^{3k}, \quad Z = \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ & 1 & 0 & \\ & & \cdot & \cdot \end{bmatrix},$$

such that $RD = D\tilde{R}$, where R is the matrix whose 11×11 upper-left part has been shown at the end of the first lecture (find $(\cdot)_k$ explicitely for all k).

SECOND LECTURE, Monday September 1, 2014 (Moscow, INM RAS)

Normalized Bernoulli numbers solve a lower triangular semi-infinite Toeplitz linear system

Note that

$$Z_{a_2,a_3,a_4,\ldots} = \begin{bmatrix} 0 & & & & \\ a_2 & 0 & & & \\ & a_3 & 0 & & \\ & & a_4 & \cdot & \\ & & & \cdot & \cdot \end{bmatrix},$$

$$Z^2_{a_2,a_3,a_4,\ldots} = \begin{bmatrix} 0 & & & & \\ 0 & 0 & & & \\ a_2a_3 & 0 & 0 & & \\ & a_3a_4 & 0 & \cdot & \\ & & a_4a_5 & \cdot & \cdot \\ & & & \cdot & \cdot \end{bmatrix},$$

$$Z^3_{a_2,a_3,a_4,\dots}=\begin{bmatrix} 0 & & & & \\ 0 & 0 & & & \\ 0 & 0 & 0 & & \\ a_2a_3a_4 & 0 & 0 & \cdot & \\ & a_3a_4a_5 & 0 & \cdot & \cdot \\ & & a_4a_5a_6 & \cdot & \cdot \\ & & & \cdot & \cdot \end{bmatrix},$$

$$Z^k_{a_2,a_3,a_4,\dots}=\begin{bmatrix} 0 & & & & \\ \cdot & 0 & & & \\ 0 & \cdot & 0 & & \\ a_2a_3\cdot a_{1+k} & 0 & \cdot & \cdot & \\ & a_3a_4\cdot a_{2+k} & 0 & \cdot & \cdot \\ & & a_4a_5\cdot a_{3+k} & \cdot & \cdot \\ & & & \cdot & \cdot \end{bmatrix}.$$

Thus

$$[Z^k_{a_2,a_3,a_4,\dots}]_{ij}=\begin{cases} a_{j+1}a_{j+2}\cdot a_{j+k} & if\ i=k+j\ per\ j=1,2,3,\dots \\ 0 & otherwise \end{cases}$$

Set

$$\phi=Z_{2,12,30,56,\dots}=Z\cdot \operatorname{diag}\big((2i-1)2i:\ i=1,2,3,\dots\big)$$

or, equivalently, $\phi=Z_{a_2,a_3,a_4,\dots}$ where $a_r=(2r-3)(2r-2)$. Then

$$[\phi^k]_{ij}=(2j-1)(2j)(2j+1)(2j+2)\cdots(2j+2k-3)(2j+2k-2)=\frac{(2j+2k-2)!}{(2j-2)!}.$$

Now observe that

$$W=(*),\quad (*)=\operatorname{diag}(2,12,30,56,\dots)\cdot\sum_{k=0}^{+\infty}\frac{1}{(2k+2)!}\phi^k$$

[in fact,

$$[(*)]_{ij}=\frac{(2i)!}{(2i-2)!}\Big(\sum_{k=0}^{+\infty}\frac{1}{(2k+2)!}\phi^k\Big)_{ij}=\frac{(2i)!}{(2i-2)!}\frac{1}{(2i-2j+2)!}[\phi^{i-j}]_{ij}=\frac{(2i)!}{(2i-2)!}\frac{1}{(2i-2j+2)!}\frac{(2i-2)!}{(2j-2)!}=\binom{2i}{2j-2}.$$

]. Thus the lower triangular system solved by Bernoulli numbers can be rewritten as follows:

$$\sum_{k=0}^{+\infty}\frac{2}{(2k+2)!}\phi^k\mathbf{b}=\mathbf{q}^e,\quad \mathbf{b}=\begin{bmatrix} B_0(0) \\ B_2(0) \\ B_4(0) \\ B_6(0) \\ \cdot \end{bmatrix},\ \mathbf{q}^e=\begin{bmatrix} 1 \\ 1/3 \\ 1/5 \\ 1/7 \\ \cdot \end{bmatrix}. \qquad (**)$$

Set $D=\operatorname{diag}(d_1,\,d_2,\,d_3,\,\dots)$, $d_i\neq 0$. By investigating the nonzero entries of the matrix $D\phi D^{-1}$, it is easy to observe that it can be forced to be equal to a matrix of the form xZ; just choose $d_k=x^{k-1}d_1/(2k-2)!$, $k=1,2,3,\dots$ (d_1 can be chosen equal to 1). So, if

$$D=\operatorname{diag}\big(1,\,\frac{x}{2!},\,\frac{x^2}{4!},\,\cdot\,,\,\frac{x^{n-1}}{(2n-2)!},\,\cdot\,\big),$$

then we have the equality $D\phi D^{-1}=xZ$.

Now, since $D\phi^k D^{-1} = (D\phi D^{-1})^k = x^k Z^k$, it is easy to show the equivalence of (**) with the following lower traingular Toeplitz linear system:

$$[\sum_{k=0}^{+\infty} \frac{2x^k}{(2k+2)!} Z^k] D\mathbf{b} = D\mathbf{q}^e,$$

Exercise. By investigating the powers of the matrix Z, verify that the matrix $[\cdot]$ is lower triangular Toeplitz.

The set of all ε-circulant matrices and, in particular, the set of all lower triangular Toeplitz matrices, form a low complexity matrix algebra

Consider the following two $n \times n$ matrices:

$$\Pi_\varepsilon = \begin{bmatrix} 0 & 1 & 0 & \cdot & 0 \\ \cdot & 0 & 1 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & \cdot & \cdot & 0 & 1 \\ \varepsilon & 0 & \cdot & \cdot & 0 \end{bmatrix}, \ \varepsilon \in \mathbb{C}, \quad J = \begin{bmatrix} & & & \cdot & 1 \\ & & & & \\ & & 1 & & \\ & & & & \\ 1 & & & & \end{bmatrix}.$$

We want to study the algebra generated by Π_ε, i.e. $C_\varepsilon = \{p(\Pi_\varepsilon) : p = \text{polynomials}\}$. By writing the powers $I, \Pi_\varepsilon, \Pi_\varepsilon^2, \ldots, \Pi_\varepsilon^{n-1}$, one easily realizes that

$$C_\varepsilon(\mathbf{z}) = \sum_{k=1}^{n} a_k \Pi_\varepsilon^{k-1} = \begin{bmatrix} a_1 & a_2 & a_3 & \cdot & a_n \\ \varepsilon a_n & a_1 & a_2 & \cdot & \cdot \\ \cdot & \varepsilon a_n & a_1 & \cdot & a_3 \\ \varepsilon a_3 & \cdot & \cdot & \cdot & a_2 \\ \varepsilon a_2 & \varepsilon a_3 & \cdot & \varepsilon a_n & a_1 \end{bmatrix}$$

(note that $\Pi_\varepsilon^n = \varepsilon I$).

ASSUME $\varepsilon \neq 0$ (the matrix algebra of all $n \times n$ ε-circulant matrices)

Eigenvalues of Π_ε:

$$\det(\lambda I - \Pi_\varepsilon) = \det \begin{bmatrix} \lambda & -1 & & & \\ 0 & \lambda & -1 & & \\ \cdot & \cdot & \cdot & \cdot & \\ 0 & & & \cdot & -1 \\ -\varepsilon & 0 & \cdot & 0 & \lambda \end{bmatrix} = \lambda\lambda^{n-1} + (-1)^{n+1}(-\varepsilon)(-1)^{n-1} = \lambda^n - \varepsilon$$

(the determinant has been computed with respect to the first column). Thus, the eigenvalues of Π_ε are the roots of the algebraic equation $\lambda^n - \varepsilon = 0$.

Eigenvectors of Π_ε:

If λ is such that $\lambda^n = \varepsilon$, then

$$\begin{bmatrix} 0 & 1 & & \\ \cdot & 0 & \cdot & \\ 0 & & \cdot & 1 \\ \varepsilon & 0 & \cdot & 0 \end{bmatrix} \begin{bmatrix} 1 \\ \lambda \\ \cdot \\ \lambda^{n-1} \end{bmatrix} = \begin{bmatrix} \lambda \\ \cdot \\ \lambda^{n-1} \\ \varepsilon \end{bmatrix} = \lambda \begin{bmatrix} 1 \\ \lambda \\ \cdot \\ \lambda^{n-1} \end{bmatrix}.$$

Let $\lambda_0, \lambda_1, \ldots, \lambda_{n-1}$ denote the eigenvalues of Π_ε (note that they are distinct!). Then

$$\Pi_\varepsilon X = \begin{bmatrix} 0 & 1 & & \\ \cdot & 0 & \cdot & \\ 0 & & \cdot & 1 \\ \varepsilon & 0 & \cdot & 0 \end{bmatrix} \begin{bmatrix} 1 & \cdot & \cdot & 1 \\ \lambda_0 & \cdot & \cdot & \lambda_{n-1} \\ \cdot & \cdot & \cdot & \cdot \\ \lambda_0^{n-1} & \cdot & \cdot & \lambda_{n-1}^{n-1} \end{bmatrix} = \begin{bmatrix} 1 & \cdot & \cdot & 1 \\ \lambda_0 & \cdot & \cdot & \lambda_{n-1} \\ \cdot & \cdot & \cdot & \cdot \\ \lambda_0^{n-1} & \cdot & \cdot & \lambda_{n-1}^{n-1} \end{bmatrix} \begin{bmatrix} \lambda_0 & & & \\ & \lambda_1 & & \\ & & \cdot & \\ & & & \lambda_{n-1} \end{bmatrix} = X\Lambda.$$

Note that the matrix X is invertible since its columns are eigenvectors of Π_ε corresponding to distinct eigenvalues of Π_ε (or since it is a vandermonde matrix). Thus, from the above equality, we obtain the following spectral representations, of Π_ε:

$$\Pi_\varepsilon = X\Lambda X^{-1}$$

and of $C_\varepsilon(\mathbf{a})$, $\mathbf{a} \in \mathbb{C}^n$:

$$C_\varepsilon(\mathbf{a}) = \sum_{k=1}^{n} a_k \Pi_\varepsilon^{k-1} = X\Big(\sum_{k=1}^{n} a_k \Lambda^{k-1}\Big)X^{-1} = X\,\mathrm{diag}\,(\sum_{k=1}^{n} a_k \lambda_i^{k-1},\ i = 0, 1, \ldots, n-1)X^{-1} = Xd(X^T\mathbf{a})X^{-1}.$$

Remark. One can easily verify that effectively the first row of $Xd(X^T\mathbf{a})X^{-1}$ is $\mathbf{a}^T$: $\mathbf{e}_1^T(Xd(X^T\mathbf{a})X^{-1}) = (\mathbf{a}^T X)d(X^T\mathbf{e}_1)X^{-1} = \mathbf{a}^T$.

Remark. Set $\varepsilon = \rho_\varepsilon e^{\mathbf{i}\theta_\varepsilon}$, $\theta_\varepsilon \in [0, 2\pi)$ ($\rho_\varepsilon > 0$). Then note that $[X]_{r,k} = \lambda_k^r = (\cdot)^r \omega_n^{rk}$, $0 \le r, k \le n-1$, that is,

$$X = \begin{bmatrix} 1 & & & \\ & (\cdot) & & \\ & & \ddots & \\ & & & (\cdot)^{n-1} \end{bmatrix} \begin{bmatrix} 1 & 1 & \cdot & 1 \\ 1 & \omega_n & \cdot & \omega_n^{n-1} \\ \cdot & \cdot & \cdot & \cdot \\ 1 & \omega_n^{n-1} & \cdot & \omega_n^{(n-1)(n-1)} \end{bmatrix}$$

where $\omega_n = e^{\mathbf{i}\frac{2\pi}{n}}$ ($\omega_n^n = 1$, $\omega_n^i \neq 1$ if $0 < i < n$) and $(\cdot) = |\rho_\varepsilon^{\frac{1}{n}}|e^{\mathbf{i}\frac{\theta_\varepsilon}{n}}$, so that $\lambda_k = (\cdot)\omega_n^k$, $k = 0, 1, \ldots, n-1$.

Now let F be the $n \times n$ Fourier matrix $F = \frac{1}{\sqrt{n}}(\omega_n^{ij})_{i,j=0}^{n-1}$. Note that F is unitary, $F^H F = I$ (this is an exercise!) and that

$$X = \sqrt{n}DF, \quad D_{ii} = (\cdot)^{i-1},\ i = 1, \ldots, n.$$

Thus

$$XX^H = nDFF^H D^H = n \begin{bmatrix} 1 & & & \\ & |(\cdot)|^2 & & \\ & & \ddots & \\ & & & |(\cdot)|^{2(n-1)} \end{bmatrix}$$

i.e. the rows of X are orthogonal. If $|\varepsilon| = 1$, then $|(\cdot)| = 1$ and $XX^H = nI$, i.e. also the columns of X are orthogonal and the matrix $\frac{1}{\sqrt{n}}X = DF$, $D_{kk} = e^{\mathbf{i}\theta_\varepsilon(k-1)/n}$, $k = 1, \ldots, n$, is unitary (because D is unitary).

In other words, we have shown that if $|\varepsilon| = 1$ then the ε-circulant matrices are simultaneously diagonalized by a unitary transform, i.e. by $U = \frac{1}{\sqrt{n}}X = DF$. Actually, the condition $|\varepsilon| = 1$ is also necessary for C_ε being diagonalized by a unitary transform. To prove this simply observe that Π_ε is a normal matrix if and only if $|\varepsilon| = 1$ (prove this!). Recall that $A \in \mathbb{C}^{n\times n}$ is normal ($AA^H = A^H A$) if and only if there exists U unitary such that $U^H AU$ is diagonal.

In conclusion,

$$\varepsilon \in \mathbb{C} \quad \Rightarrow \quad C_\varepsilon(\mathbf{a}) = DFd(\sqrt{n}FD\mathbf{a})F^H D^{-1},$$

$$|\varepsilon| = 1 \quad \Rightarrow \quad C_\varepsilon(\mathbf{a}) = DFd(\sqrt{n}FD\mathbf{a})F^H\overline{D} = Ud(\sqrt{n}U^T\mathbf{a})U^H,\ U = DF\ \textit{unitary}.$$

From the above spectral representations of the matrices from C_ε it follows that C_ε is a space of low complexity, in the sense that

1) any matrix-vector product $C_\varepsilon(\mathbf{a})\mathbf{v}$,
2) the solution of any linear system $C_\varepsilon(\mathbf{a})\mathbf{x} = \mathbf{b}$ and
3) the computation of the eigenvalues of any $C_\varepsilon(\mathbf{a})$

are all operations that can be done with $O(n \log_2 n)$ arithmetic operations. This is true for any value of $\varepsilon \neq 0$. (Check these assertions!)

Remark. Note however that the problem of the eigenvalues of matrices from C_ε is optimally conditioned if and only if $|\varepsilon| = 1$. This assertion is a consequence of the Bauer-Fike theorem which states that for any eigenvalue $\tilde{\lambda}$ of $C_\varepsilon(\mathbf{a}) + \Delta$, there is an eigenvalue λ of $C_\varepsilon(\mathbf{a})$ such that

$$|\tilde{\lambda} - \lambda| \le \mu \|\Delta\|_2, \quad \mu = \inf_{M: M^{-1} C_\varepsilon(\mathbf{a}) M = diag} \mu_2(M),$$

and of the fact that μ is equal to 1 if and only if $|\varepsilon| = 1$ (see the Exercise here below).

Exercise. Let M be a $n \times n$ matrix. Prove that $\mu_2(M) = \|M\|_2 \|M^{-1}\|_2 = 1$ if and only if $M = cU$ for some matrix U unitary and $c \in \mathbb{C}$.

THIRD LECTURE, Tuesday September 2, 2014 (Moscow, INM RAS)

Exercise. Find a value of x for which the entry $(D\mathbf{b})_n$ of the solution of the linear system $[\sum_{k=0}^{+\infty} 2x^k/(2k+2)! Z^k] D\mathbf{b} = D\mathbf{q}^e$ for $n \to +\infty$ remains bounded.

From yesterday, we have

$$C_\varepsilon(\mathbf{a}) = \sum_{k=1}^{n} a_k \Pi_\varepsilon^{k-1} = DFd(\sqrt{n} FD\mathbf{a}) F^H D^{-1} = DFd(\sqrt{n} FD\mathbf{a})(DF)^H$$

where the last identity holds if and only if $|\varepsilon| = 1$ ($D_{kk} = (\cdot)^{k-1}$, $k = 1, \ldots, n$, $(\cdot) = |\rho_\varepsilon^{\frac{1}{n}}| e^{\mathbf{i}\frac{\theta_\varepsilon}{n}}$, $F_{ij} = \frac{1}{\sqrt{n}} \omega_n^{ij}$, $0 \le i, j \le n-1$, $\omega_n = e^{\mathbf{i}\frac{2\pi}{n}}$, $\varepsilon = \rho_\varepsilon e^{\mathbf{i}\theta_\varepsilon}$).

For $\varepsilon = 1$: $D = I \Rightarrow C_1(\mathbf{a}) = Fd(\sqrt{n} F\mathbf{a}) F^H$.
For $\varepsilon = -1$: $C_{-1}(\mathbf{a}) = DFd(\sqrt{n} FD\mathbf{a})(DF)^H$, $D = \text{diag}\,((e^{\mathbf{i}\frac{\pi}{n}})^{k-1}, k = 1, \ldots, n)$.

1) $C_\varepsilon(\mathbf{a})\mathbf{v}$: two or three FFT, thus $O(n \log_2 n)$ arithmetic operations
2) $C_\varepsilon(\mathbf{a})\mathbf{z} = \mathbf{f}$: $\mathbf{z} = DFd(\sqrt{n} FD\mathbf{a})^{-1} F^H D^{-1} \mathbf{f}$, as above
3) eigenvalues of $C_\varepsilon(\mathbf{a})$: one FFT, as above

ASSUME $\varepsilon = 0$ (the matrix algebra of all $n \times n$ upper triangular matrices)

Now we will prove that the operations in 1), 2), 3) can be done with $O(n \log_2 n)$ arithmetic operations also in the case $\varepsilon = 0$. Note that no representation of $C_0(\mathbf{a})$ of the type $Md(\mathbf{v})M^{-1}$ can hold since, for example, Π_0 is not diagonalizable.

Set

$$L(\mathbf{a}) = C_0(\mathbf{a})^T = \begin{bmatrix} a_0 & & & \\ a_1 & a_0 & & \\ \cdot & a_1 & a_0 & \\ a_{n-1} & \cdot & a_1 & a_0 \end{bmatrix}$$

3) The eigenvalues of $L(\mathbf{a})$: 0 computations
1) $L(\mathbf{a})\mathbf{v}$: $O(n \log_2 n)$ arithmetic operations (see the exercise here below)

Exercise. Prove that any Toeplitz matrix $T = (t_{i-j})_{i,j=1}^n$ can be written as $C_1(\mathbf{u}) + C_{-1}(\mathbf{w})$, for suitable $\mathbf{u}, \mathbf{w} \in \mathbb{C}^n$.

2) $L(\mathbf{a})\mathbf{z} = \mathbf{f}$: $L(\mathbf{a})$ is not diagonalizable since even $L(\mathbf{e}_2)$ is not diagonalizable! However, the cost of solving a lower triangular Toeplitz system is $O(n \log_2 n)$ arithmetic operations. In the following of the present lecture we shall prove this fact.

$L(\mathbf{a})\mathbf{z} = \mathbf{f}$, $\mathbb{N}$, $\mathbf{a} \in \mathbb{C}^{\mathbb{N}}$, $\mathbf{a} = \begin{bmatrix} a_0 \\ a_1 \\ \cdot \end{bmatrix}$ semi-infinite vector.

$L(\mathbf{a}) = \sum_{k=0}^{+\infty} a_k Z^k$, $Z = \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ & 1 & 0 & \\ & & \cdot & \cdot \end{bmatrix}$

The space $\{L(\mathbf{a}) : \mathbf{a} \in \mathbb{C}^n\}$ is closed under matrix multiplication and commutative. Moreover, it is closed under inversion since $\{L(\mathbf{a}) : \mathbf{a} \in \mathbb{C}^n\} = \{A : AZ = ZA\}$ (so $AZ = ZA \Rightarrow A^{-1}Z = ZA^{-1}$).

LEMMA 1. $\mathbf{a}, \mathbf{b} \in \mathbb{C}^{\mathbb{N}}$: $L(\mathbf{a})\mathbf{b} = \mathbf{c} \Leftrightarrow L(\mathbf{a})L(\mathbf{b}) = L(\mathbf{c})$

Proof. $\Leftarrow$: check the first column. $\Rightarrow$: note that $L(\mathbf{a})L(\mathbf{b})$ is a lower triangular Toeplitz matrix and that its first column is $L(\mathbf{a})\mathbf{b}$. $\square$

LEMMA 2. Set $E = \begin{bmatrix} 1 & 0 & & \\ 0 & 0 & & \\ 0 & 1 & 0 & \\ 0 & 0 & 0 & \\ \cdot & \cdot & \cdot & \end{bmatrix}$, $E\mathbf{v} = [v_0\ 0\ v_1\ 0\ v_2\ 0\ \cdots]^T$. Then $E^s L(\mathbf{u})\mathbf{v} = L(E^s\mathbf{u})E^s\mathbf{v}$, $s \in \mathbb{N}$.

Proof. For $s = 0$ trivial. If we prove the thesis for $s = 1$, $EL(\mathbf{u})\mathbf{v} = L(E\mathbf{u})E\mathbf{v}$, then, by multiplying on the left by E we obtain the thesis for all other s. But the fact that $EL(\mathbf{u})\mathbf{v} = L(E\mathbf{u})E\mathbf{v}$ follows from an easy check:

$$EL(\mathbf{u})\mathbf{v} = E \begin{bmatrix} u_0 & & & & \\ u_1 & u_0 & & & \\ u_2 & u_1 & u_0 & & \\ \cdot & \cdot & \cdot & \cdot \end{bmatrix} \begin{bmatrix} v_0 \\ v_1 \\ v_2 \\ \cdot \end{bmatrix} = \begin{bmatrix} u_0v_0 \\ 0 \\ u_1v_0 + u_0v_1 \\ 0 \\ u_2v_0 + u_1v_1 + u_0v_2 \\ 0 \\ \cdot \end{bmatrix},$$

$$L(E\mathbf{u})E\mathbf{v} = \begin{bmatrix} u_0 & & & & & \\ 0 & u_0 & & & & \\ u_1 & 0 & u_0 & & & \\ 0 & u_1 & 0 & u_0 & & \\ u_2 & 0 & u_1 & 0 & u_0 & \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix} \begin{bmatrix} v_0 \\ 0 \\ v_1 \\ 0 \\ v_2 \\ \cdot \end{bmatrix} = \begin{bmatrix} u_0v_0 \\ 0 \\ u_1v_0 + u_0v_1 \\ 0 \\ u_2v_0 + u_1v_1 + u_0v_2 \\ 0 \\ \cdot \end{bmatrix}.$$

$\square$

$L(\mathbf{a})\mathbf{z} = \mathbf{f}$, $\mathbf{a} \in \mathbb{C}^{\mathbb{N}}$, $\mathbf{f} \in \mathbb{C}^{\mathbb{N}}$

STEP 1. Find $\hat{\mathbf{a}} \in \mathbb{C}^{\mathbb{N}}$ such that $L(\mathbf{a})\hat{\mathbf{a}} = E\mathbf{a}^{(1)}$ for some $\mathbf{a}^{(1)} \in \mathbb{C}^{\mathbb{N}}$. Then, by Lemma 1 and commutativity,

$$L(\hat{\mathbf{a}})L(\mathbf{a}) = L(\mathbf{a})L(\hat{\mathbf{a}}) = L(E\mathbf{a}^{(1)}).$$

Multiply the system on the left by $L(\hat{\mathbf{a}})$:

$$L(\hat{\mathbf{a}})L(\mathbf{a})\mathbf{z} = L(\hat{\mathbf{a}})\mathbf{f}, \quad L(E\mathbf{a}^{(1)})\mathbf{z} = L(\hat{\mathbf{a}})\mathbf{f}. \tag{$*$}$$

Note that

$$L(E\mathbf{a}^{(1)}) = \begin{bmatrix} a_0^{(1)} & & \\ 0 & a_0^{(1)} & \\ a_1^{(1)} & 0 & \cdot \\ 0 & a_1^{(1)} & \cdot \\ a_2^{(1)} & 0 & \cdot \\ \cdot & \cdot & \cdot \end{bmatrix}.$$

STEP 2. Find $\hat{\mathbf{a}}^{(1)} \in \mathbb{C}^{\mathbb{N}}$ such that $L(\mathbf{a}^{(1)})\hat{\mathbf{a}}^{(1)} = E\mathbf{a}^{(2)}$ for some $\mathbf{a}^{(2)} \in \mathbb{C}^{\mathbb{N}}$. Then, by Lemma 2 $L(E\mathbf{a}^{(1)})E\hat{\mathbf{a}}^{(1)} = EL(\mathbf{a}^{(1)})\hat{\mathbf{a}}^{(1)} = E^2\mathbf{a}^{(2)}$, and, by Lemma 1 and commutativity,

$$L(E\hat{\mathbf{a}}^{(1)})L(E\mathbf{a}^{(1)}) = L(E\mathbf{a}^{(1)})L(E\hat{\mathbf{a}}^{(1)}) = L(E^2\mathbf{a}^{(2)}).$$

Multiply the system (*) on the left by $L(E\hat{\mathbf{a}}^{(1)})$:

$$L(E\hat{\mathbf{a}}^{(1)})L(E\mathbf{a}^{(1)})\mathbf{z} = L(E\hat{\mathbf{a}}^{(1)})L(\hat{\mathbf{a}})\mathbf{f}, \quad L(E^2\mathbf{a}^{(2)})\mathbf{z} = L(E\hat{\mathbf{a}}^{(1)})L(\hat{\mathbf{a}})\mathbf{f}. \tag{**}$$

Note that

$$L(E^2\mathbf{a}^{(2)}) = \begin{bmatrix} a_0^{(2)} & & \\ 0 & a_0^{(2)} & \\ 0 & 0 & \cdot \\ 0 & 0 & \cdot \\ a_1^{(2)} & 0 & \cdot \\ 0 & a_1^{(2)} & \cdot \\ \cdot & \cdot & \cdot \end{bmatrix}.$$

STEP 3. Find $\hat{\mathbf{a}}^{(2)} \in \mathbb{C}^{\mathbb{N}}$ such that $L(\mathbf{a}^{(2)})\hat{\mathbf{a}}^{(2)} = E\mathbf{a}^{(3)}$ for some $\mathbf{a}^{(3)} \in \mathbb{C}^{\mathbb{N}}$. Then, by Lemma 2 $L(E^2\mathbf{a}^{(2)})E^2\hat{\mathbf{a}}^{(2)} = E^2L(\mathbf{a}^{(2)})\hat{\mathbf{a}}^{(2)} = E^3\mathbf{a}^{(3)}$, and, by Lemma 1 and commutativity,

$$L(E^2\hat{\mathbf{a}}^{(2)})L(E^2\mathbf{a}^{(2)}) = L(E^2\mathbf{a}^{(2)})L(E^2\hat{\mathbf{a}}^{(2)}) = L(E^3\mathbf{a}^{(3)}).$$

Multiply the system (**) on the left by $L(E^2\hat{\mathbf{a}}^{(2)})$:

$$L(E^2\hat{\mathbf{a}}^{(2)})L(E^2\mathbf{a}^{(2)})\mathbf{z} = L(E^2\hat{\mathbf{a}}^{(2)})L(E\hat{\mathbf{a}}^{(1)})L(\hat{\mathbf{a}})\mathbf{f}, \quad L(E^3\mathbf{a}^{(3)})\mathbf{z} = L(E^2\hat{\mathbf{a}}^{(2)})L(E\hat{\mathbf{a}}^{(1)})L(\hat{\mathbf{a}})\mathbf{f}.$$

Note that

$$L(E^3\mathbf{a}^{(3)}) = \begin{bmatrix} a_0^{(3)} & & \\ 0 & a_0^{(3)} & \\ 0 & 0 & \cdot \\ 0 & 0 & \cdot \\ 0 & 0 & \cdot \\ 0 & 0 & \cdot \\ 0 & 0 & \cdot \\ 0 & 0 & \cdot \\ a_1^{(3)} & 0 & \cdot \\ 0 & a_1^{(3)} & \cdot \\ \cdot & \cdot & \cdot \end{bmatrix}.$$

Assume now that the lower triangular Toeplitz system we have to solve is made up with 8 equations. Then we stop the process here, and we note that, by commutativity,

$$L(E^3\mathbf{a}^{(3)})\mathbf{z} = L(\hat{\mathbf{a}})L(E\hat{\mathbf{a}}^{(1)})L(E^2\hat{\mathbf{a}}^{(2)})\mathbf{f}.$$

Assume moreover that $\mathbf{f} = E^2\mathbf{v}$ for some vector $\mathbf{v} \in \mathbb{C}^{\mathbb{N}}$ (important: note that this assumption is satisfied by $\mathbf{f} = \mathbf{e}_1$). Then, by Lemma 2,

$$L(E^3\mathbf{a}^{(3)})\mathbf{z} = L(\hat{\mathbf{a}})L(E\hat{\mathbf{a}}^{(1)})L(E^2\hat{\mathbf{a}}^{(2)})E^2\mathbf{v} = L(\hat{\mathbf{a}})L(E\hat{\mathbf{a}}^{(1)})E^2L(\hat{\mathbf{a}}^{(2)})\mathbf{v} = L(\hat{\mathbf{a}})EL(\hat{\mathbf{a}}^{(1)})EL(\hat{\mathbf{a}}^{(2)})\mathbf{v},$$

and this equality between semi-infinite vectors imply the following equality between 8-dimensional vectors:

$$a_0^{(3)}\{\mathbf{z}\}_8 = \{L(\hat{\mathbf{a}})\}_{8,8}\{E\}_{8,8}\{L(\hat{\mathbf{a}}^{(1)})\}_{8,8}\{E\}_{8,8}\{L(\hat{\mathbf{a}}^{(2)})\}_{8,8}\{\mathbf{v}\}_8 = \{L(\hat{\mathbf{a}})\}_{8,8}\{E\}_{8,4}\{L(\hat{\mathbf{a}}^{(1)})\}_{4,4}\{E\}_{4,2}\{L(\hat{\mathbf{a}}^{(2)})\}_{2,2}\{\mathbf{v}\}_2$$

[The last four columns of $E_{8,8}$ are null; the last two columns of $E_{4,4}$ are null].

The latter result let us state that the first column of the inverse of a 8×8 lower triangular Toeplitz matrix can be computed by performing two matrix-vector products where the matrices are lower triangular Toeplitz 4×4 and 8×8, respectively (set $\mathbf{v} = \mathbf{e}_1$). In the general case in which $n = 2^s$ the lower triangular Toeplitz matrices that have to be multiplied by vectors are respectively of order 4×4, 8×8, ..., $2^s \times 2^s$. So, we have the following

Result: If $O(j2^j)$ is the cost of computing the product of a $2^j \times 2^j$ lower triangular Toeplitz matrix by a vector, then the cost of the computation of the first column of the inverse of a $2^s \times 2^s$ lower triangular Toeplitz matrix by the above described algorithm is: $\sum_{j=2}^{s} O(j2^j) = O(s2^s) = O(n \log_2 n)$.

Finally, once $\{\tilde{\mathbf{z}}\}_{2^s} = \{L(\mathbf{a})\}^{-1}_{2^s,2^s}\{\mathbf{e}_1\}_{2^s}$ is known, in order to solve the system $\{L(\mathbf{a})\}_{2^s,2^s}\{\mathbf{z}\}_{2^s} = \{\mathbf{f}\}_{2^s}$, note that

$$\{\mathbf{z}\}_{2^s} = \{L(\mathbf{a})\}^{-1}_{2^s,2^s}\{\mathbf{f}\}_{2^s} = \{L(\tilde{\mathbf{z}})\}_{2^s,2^s}\{\mathbf{f}\}_{2^s},$$

and that $\{L(\tilde{\mathbf{z}})\}_{2^s,2^s}\{\mathbf{f}\}_{2^s}$ can be computed with $O(s2^s) = O(n \log_2 n)$ arithmetic operations (via the representation of $\{L(\tilde{\mathbf{z}})\}_{2^s,2^s}$ as the sum of a circulant and of a (-1)-circulant matrix).

Important remark (for the proof of the above result). How to find $\hat{\mathbf{a}}$ such that $L(\mathbf{a})\hat{\mathbf{a}} = E\mathbf{a}^{(1)}$ for some $\mathbf{a}^{(1)} \in \mathbb{C}^{\mathbb{N}}$? No computation is required to obtain such a vector $\hat{\mathbf{a}}$, in fact

$$\begin{bmatrix} 1 & & & & \\ a_1 & 1 & & & \\ a_2 & a_1 & 1 & & \\ a_3 & a_2 & a_1 & 1 & \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix} \begin{bmatrix} 1 \\ -a_1 \\ a_2 \\ -a_3 \\ \cdot \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 2a_2 - a_1^2 \\ 0 \\ \neq 0 \\ \cdot \end{bmatrix} \qquad (RES)$$

FINAL REMARK.

Find $\hat{\mathbf{a}}$ such that $L(\mathbf{a})\hat{\mathbf{a}} = E\mathbf{a}^{(1)}$ is equivalent to find $\hat{\mathbf{a}}$ such that $L(\hat{\mathbf{a}})L(\mathbf{a}) = L(E\mathbf{a}^{(1)})$ (by Lemma 1 and by commutativity), i.e. to find $\hat{a}_k$ such that

$$(\sum_{k=0}^{+\infty} \hat{a}_k Z^k)(\sum_{k=0}^{+\infty} a_k Z^k) = (\sum_{k=0}^{+\infty} a_k^{(1)} Z^{2k}).$$

More in general, how to find $\hat{a}_k$ such that

$$(\sum_{k=0}^{+\infty} \hat{a}_k Z^k)(\sum_{k=0}^{+\infty} a_k Z^k) = (\sum_{k=0}^{+\infty} a_k^{(1)} Z^{rk}),$$

or, equivalently, given a polynomial $a(z)$ how to find a polynomial $\hat{a}(z)$ such that $\hat{a}(z)a(z) = a^{(1)}(z^r)$ for some polynomial $a^{(1)}$? An answer is in the following

Exercise. Prove that $\hat{a}(z) = a(zt)a(zt^2)\cdots a(zt^{r-1})$ where $t^r = 1$, $t^j \neq 1$, $0 < j < r$, is such that $\hat{a}(z)a(z) = a^{(1)}(z^r)$.

For example, for $r = 2$ we have $\hat{a}(z) = a(-z)$, and we retrieve the result observed in (RES).

Exercise (not for evaluation). If $E\mathbf{v} = [v_0\ 0\ 0\ v_1\ 0\ 0\ v_2\ 0\ \cdots]^T$, then analogous Lemmas 1 and 2 hold, and analogous algorithm ok for $n = 3^s$ of complexity $O(s3^s)$ can be conceived ...

FOURTH LECTURE, Monday September 8, 2014 (Rome)

Hessenberg algebras and displacement matrix formulas

Exercise 1. Prove the following assertion:
If B is $n \times n$ real symmetric positive definite, and $\mathbf{y}, \mathbf{s} \in \mathbb{R}^n$ are such that $\mathbf{y}^T\mathbf{s} > 0$, then

$$A = B + \frac{1}{\mathbf{y}^T\mathbf{s}}\mathbf{y}\mathbf{y}^T - \frac{1}{\mathbf{s}^TB\mathbf{s}}B\mathbf{s}\mathbf{s}^TB$$

is real symmetric positive definite.

Exercise 2. Prove that eigenvectors $\mathbf{x}$ and $\mathbf{y}$ corresponding to distinct eigenvalues λ and μ of a hermitian (or unitary) matrix A ($A\mathbf{x} = \lambda\mathbf{x}$, $A\mathbf{y} = \mu\mathbf{y}$) are such that $\mathbf{x}^H\mathbf{y} = 0$.

Exercise 3. Find values of α and β such that $\sum_{i=1}^{n} \sin^2 \frac{ij\pi}{n+1} = \alpha n + \beta$ (for all $j = 1, 2, \ldots, n$).

Hessenberg algebras.

$$X = \begin{bmatrix} r_{11} & b_1 & & & \\ r_{21} & r_{22} & b_2 & & \\ \cdot & \cdot & \cdot & \cdot & \\ \cdot & & & \cdot & b_{n-1} \\ r_{n1} & \cdot & & \cdot & r_{nn} \end{bmatrix}, \quad b_i \neq 0, \ \forall\, i. \tag{$Hess$}$$

Note that the following matrices

$$J_1 = X^0 = I, \quad J_2 = \frac{1}{b_1}(X - r_{11}I), \quad J_3 = \frac{1}{b_2}(J_2X - r_{21}I - r_{22}J_2), \ldots,$$

$$J_n = \frac{1}{b_{n-1}}(J_{n-1}X - r_{n-1,1}I - r_{n-1,2}J_2 \ldots - r_{n-1,n-1}J_{n-1})$$

are polynomials in X (of degree 0, 1, ..., $n-1$) and have as first row, respectively, $\mathbf{e}_1^T$, $\mathbf{e}_2^T$, $\mathbf{e}_3^T$, ..., $\mathbf{e}_n^T$. Then consider the set H_X of all polymomials in X. Note that its dimension is less than or equal to n, since by Cayley-Hamilton theorem X^n must be a linear combination of the previous powers of X. But there are n matrices in H_X which are linearly independent, the J_k. So, the dimension of H_X is n,

$$H_X = \{p(X) : \, p = \text{polynomials}\} = \text{Span}\,\{J_1, J_2, \ldots, J_n\} = \{A \in \mathbb{C}^{n\times n} : \, AX = XA\},$$

and any matrix of H_X is uniquely determined by its first row, i.e., given any vector $\mathbf{a}$ in $\mathbb{C}^n$ it is well defined the matrix of H_X with first row $\mathbf{a}^T$:

$$H_X(\mathbf{a}) = \sum_{k=1}^{n} a_k J_k, \quad \mathbf{e}_1^T H_X(\mathbf{a}) = \mathbf{a}^T.$$

Of course, the space H_X is closed under matrix multiplication (H_X is a matrix algebra), and is commutative. In particular, we have that $J_kJ_s = J_sJ_k$ (for all k, s), and thus

$$\mathbf{e}_i^T H_X(\mathbf{a}) = \mathbf{e}_i^T \sum_{k=1}^{n} a_k J_k = \sum_{k=1}^{n} a_k \mathbf{e}_k^T J_i = \mathbf{a}^T J_i, \quad \forall\, i$$

and, multiplying the previous identity by the scalar v_i and summing on $i = 1, \ldots, n$, we obtain the equality

$$\mathbf{v}^T H_X(\mathbf{a}) = \mathbf{a}^T H_X(\mathbf{v}), \quad \mathbf{a}, \mathbf{v} \in \mathbb{C}^n.$$

Note that if a matrix $H_X(\mathbf{a})$ in H_X is invertible, then its inverse must be a polynomial in X, i.e. it must exist a vector $\mathbf{z} \in \mathbb{C}^n$ such that $H_X(\mathbf{a})^{-1} = H_X(\mathbf{z})$ (see below). From the equality $H_X(\mathbf{z})H_X(\mathbf{a}) = I$ it follows that $\mathbf{z}$ can be determined by solving the linear system $\mathbf{z}^T H_X(\mathbf{a}) = \mathbf{e}_1^T$.

The space of ε-circulant matrices (for any value of ε) is of course an example of Hessenberg algebra. We will study also another important example of Hessenberg algebra, the τ matrix algebra.

Now we want to state an example of general displacement formula, which involves Hessenberg algebras, i.e. given any matrix A we represent it as a sum of products of types M_iN_i where M_i and N_i are matrices of two general Hessenberg algebras H_X and $H_{X'}$. The number of addenda in such sum is $\alpha+1$ where α is the rank of $AX-XA$. Such formula for A can be extremely useful if α does not depend on the order n of the matrix A.

Lemma. Let A be $n\times n$ with complex entries. If $AX-XA=\sum_{m=1}^{\alpha}\mathbf{x}_m\mathbf{y}_m^T$ $(\mathbf{x}_m,\mathbf{y}_m\in\mathbb{C}^n)$, then $\sum_{m=1}^{\alpha}\mathbf{x}_m^Tp(X)^T\mathbf{y}_m=0$, for any polynomial p.

Proof.

$$\sum_{m=1}^{\alpha}\mathbf{x}_m^Tp(X^T)\mathbf{y}_m=\sum_{m=1}^{\alpha}\sum_{i,j=1}^{n}(\mathbf{x}_m)_i(p(X^T))_{ij}(\mathbf{y}_m)_j=\sum_{i,j=1}^{n}\Big(\sum_{m=1}^{\alpha}(\mathbf{x}_m)_i(\mathbf{y}_m)_j\Big)(p(X^T))_{ij}$$

$$=\sum_{i,j=1}^{n}(AX-XA)_{ij}(p(X^T))_{ij}=\sum_{i,j=1}^{n}\sum_{k}A_{ik}X_{kj}(p(X^T))_{ij}-\sum_{i,j=1}^{n}\sum_{k}X_{ik}A_{kj}(p(X^T))_{ij}$$

$$=\sum_{i,k=1}^{n}A_{ik}\sum_{j}(p(X^T))_{ij}(X^T)_{jk}-\sum_{k,j=1}^{n}A_{kj}\sum_{i}(X^T)_{ki}(p(X^T))_{ij}=\sum_{i,k=1}^{n}A_{ik}(p(X^T)X^T)_{ik}-\sum_{k,j=1}^{n}A_{kj}(X^Tp(X^T))_{kj}=0.$$

□

The displacement formula obtained in the following theorem involves persymmetric Hessenberg algebras. Such formula and the fact that $\operatorname{rank}(AX-XA)=2$ for A Toeplitz and for $X=\Pi_\varepsilon$ will be used to obtain an efficient representation of the inverse of a Toeplitz matrix, due to Ammar and Gader.

Theorem. Let the lower Hessenberg matrix X be persymmetric, i.e. symmetric with respect to the anti-diagonal ($X^T=JXJ$), and define $X'\in\mathbb{C}^{n\times n}$ and $\beta\in\mathbb{C}$ by the following identity

$$X=X'+(r_{n1}-\beta)\mathbf{e}_n\mathbf{e}_1^T.$$

Let A be $n\times n$ with complex entries. If $AX-XA=\sum_{m=1}^{\alpha}\mathbf{x}_m\mathbf{y}_m^T$ $(\mathbf{x}_m,\mathbf{y}_m\in\mathbb{C}^n)$, then

$$(r_{n1}-\beta)A=-\sum_{m=1}^{\alpha}H_X(\hat{\mathbf{x}}_m)H_{X'}(\mathbf{y}_m)+(r_{n1}-\beta)H_X(JA\mathbf{e}_n)$$

(for $\mathbf{v}\in\mathbb{C}^n$ the symbol $\hat{\mathbf{v}}$ means $J\mathbf{v}$).

Exercise (not for evaluation). Prove that if instead X has a Toeplitz structure, then, under the same assumption $AX-XA=\sum_{m=1}^{\alpha}\mathbf{x}_m\mathbf{y}_m^T$, we have $bA=-\sum_{m=1}^{\alpha}L(Z\mathbf{x}_m)H_X(\mathbf{y}_m)+bH_X(A^T\mathbf{e}_1)$ where $b=b_i$ and $L(\mathbf{z})=C_0(\mathbf{z})^T$ (for $X=Z^T$ such result yields the famous Gohberg-Semencul formula).

Proof. Set $(*)=-\sum_{m=1}^{\alpha}H_X(\hat{\mathbf{x}}_m)H_{X'}(\mathbf{y}_m)$. Then

$$(*)X-X(*)=-\sum_{m=1}^{\alpha}H_X(\hat{\mathbf{x}}_m)\Big[H_{X'}(\mathbf{y}_m)X-XH_{X'}(\mathbf{y}_m)\Big]=-(r_{n1}-\beta)\sum_{m=1}^{\alpha}H_X(\hat{\mathbf{x}}_m)\Big[H_{X'}(\mathbf{y}_m)\mathbf{e}_n\mathbf{e}_1^T-\mathbf{e}_n\mathbf{e}_1^TH_{X'}(\mathbf{y}_m)\Big]$$

$$=-(r_{n1}-\beta)\sum_{m=1}^{\alpha}H_X(\hat{\mathbf{x}}_m)[\hat{\mathbf{y}}_m\mathbf{e}_1^T-\mathbf{e}_n\mathbf{y}_m^T]=(r_{n1}-\beta)\sum_{m=1}^{\alpha}\mathbf{x}_m\mathbf{y}_m^T=(r_{n1}-\beta)(AX-XA).$$

Thus, $(r_{n1}-\beta)A-(*)$ must commute with X, i.e. must be a polynomial in X:

$$(r_{n1}-\beta)A-(*)=H_X(\mathbf{z}),\quad \mathbf{z}\in\mathbb{C}^n.$$

By imposing that the last column of the left matrix and the last column of the right matrix, in the previous equality, are equal, one obtains that the vector $\mathbf{z}$ is defined by the identity $J\mathbf{z}=(r_{n1}-\beta)A\mathbf{e}_n$.

Note that in the proof we have used twice the fact that $\sum_{m=1}^{\alpha}H_X(\hat{\mathbf{x}}_m)\hat{\mathbf{y}}_m$ is the null vector, in fact

$$\mathbf{e}_i^T(\sum_{m=1}^{\alpha}H_X(\hat{\mathbf{x}}_m)\hat{\mathbf{y}}_m)=\sum_{m=1}^{\alpha}\hat{\mathbf{x}}_m^TJ_i\hat{\mathbf{y}}_m=\sum_{m=1}^{\alpha}\mathbf{x}_m^TJ_i^T\mathbf{y}_m$$

($J_i=H_X(\mathbf{e}_i)$) and the last quantity is zero by the Lemma since J_i is a polynomial in X. $\square$

We have seen that any matrix $A=C_\varepsilon(\mathbf{z})$ in the space C_ε is of low complexity, in the sense that the computation of A times a vector, the computation of the eigenvalues of A, and the computation of $\mathbf{z}$ such that $A\mathbf{z}=\mathbf{b}$ are all operations that can be performed with no more that $O(n\log_2 n)$ arithmetic operations.

Is the same true if A has the Toeplitz structure, i.e. $A=T$ where $T=(t_{i-j})_{i,j=1}^n$? We already know that the computation of T times a vector can be done with $O(n\log_2 n)$ arithmetic operations, since any Toeplitz matrix can be expressed as the sum of a circulant and of a (-1)-circulant matrix. For what concerns the computation of the eigenvalues of T, I can say only that, under suitable assumptions on the sequence t_k, $k\in\mathbb{Z}$, there are results that indicate areas in $\mathbb{C}$ (defined in terms of the symbol function of t_k) enclosing such eigenvalues. We shall claim a result of this type. So, it remains to consider the problem of solving a Toeplitz linear system $T\mathbf{x}=\mathbf{b}$. We shall prove a simple representation for the inverse of T of the type $T^{-1}=C_1C_{-1}+C_1'C_{-1}'$, involving two circulant and two (-1)-circulant matrices. Such representation allows us to claim that if we do not count the operations involving only the entries of T (such operations can be done in a preprocessing phase), then $T\mathbf{x}=\mathbf{b}$ can be solved with $O(n\log_2 n)$ arithmetic operations.

Note that, for any Toeplitz matrix $T=(t_{i-j})_{i,j=1}^n$, we have

$$T\Pi_\varepsilon-\Pi_\varepsilon T=\begin{bmatrix}\varepsilon t_{-n+1}-t_1 & 0 & \cdot & 0\\ \cdot & \cdot & \cdot & \cdot\\ \varepsilon t_{-1}-t_{n-1} & 0 & \cdot & 0\\ \varepsilon t_0-\varepsilon t_0 & t_{n-1}-\varepsilon t_{-1} & \cdot & t_1-\varepsilon t_{-n+1}\end{bmatrix}$$

(check the case $n=3$ first), or, shortly,

$$T\Pi_\varepsilon-\Pi_\varepsilon T=\mathbf{u}\mathbf{e}_1^T-J\mathbf{e}_1\mathbf{u}^TJ,\quad \mathbf{u}=\varepsilon T\mathbf{e}_n-\Pi_\varepsilon T\mathbf{e}_1.$$

The Ammar-Gader representation of the inverse of a Toeplitz matrix

FIFTH LECTURE, Tuesday September 9, 2014 (Rome)

Assume now T invertible. Then $\Pi_\varepsilon T^{-1}-T^{-1}\Pi_\varepsilon=T^{-1}\mathbf{u}\mathbf{e}_1^TT^{-1}-T^{-1}J\mathbf{e}_1\mathbf{u}^TJT^{-1}$, i.e.

$$T^{-1}\Pi_\varepsilon-\Pi_\varepsilon T^{-1}=J\mathbf{p}(T^{-1}\mathbf{u})^TJ-(T^{-1}\mathbf{u})\mathbf{p}^T,\quad \mathbf{p}=(\mathbf{e}_1^TT^{-1})^T=JT^{-1}\mathbf{e}_n.$$

Remark: $T^{-1}\mathbf{u}=\varepsilon\mathbf{e}_n-T^{-1}\begin{bmatrix}t_1\\ \cdot\\ t_{n-1}\\ \varepsilon t_0\end{bmatrix}$. Set $\delta(\sigma)=T^{-1}\begin{bmatrix}t_1\\ \cdot\\ t_{n-1}\\ \sigma\end{bmatrix}$ and $\mathbf{q}=(T^T)^{-1}\begin{bmatrix}\varepsilon t_0\\ t_{n-1}\\ \cdot\\ t_1\end{bmatrix}=J\delta(\varepsilon t_0)$.

Then $JT^{-1}\mathbf{u}=\varepsilon\mathbf{e}_1-\mathbf{q}$.

Now observe that $\Pi_\varepsilon=\Pi_\beta+(\varepsilon-\beta)\mathbf{e}_n\mathbf{e}_1^T$, thus we can apply the displacement theorem for $X=\Pi_\varepsilon$ and $X'=\Pi_\beta$, and obtain

$$(\varepsilon-\beta)T^{-1}=-C_\varepsilon(\mathbf{p})C_\beta(JT^{-1}\mathbf{u})+C_\varepsilon(JT^{-1}\mathbf{u})C_\beta(\mathbf{p})+(\varepsilon-\beta)C_\varepsilon(\mathbf{p}),$$

and, by the Remark,

$$(\varepsilon-\beta)T^{-1}=C_\varepsilon(\mathbf{p})[-\beta I+C_\beta(\mathbf{q})]+[\varepsilon I-C_\varepsilon(\mathbf{q})]C_\beta(\mathbf{p}).$$

[Further remark. Note that the vector $\delta(\sigma)$ can always be expressed in terms of at most three columns of T^{-1}, and one of these columns is always $T^{-1}\mathbf{e}_n$ (see [Di Fiore, Zellini, 1995]). For example, in the particular case in which $s_{11}:=(T^{-1})_{11}\neq 0$ we have

$$\delta(\sigma)=\delta\Big(-\frac{[0\,t_{n-1}\ \cdot\ t_1]T^{-1}\mathbf{e}_1}{s_{11}}+\Big[\sigma-(-\frac{[0\ t_{n-1}\ \cdot\ t_1]T^{-1}\mathbf{e}_1}{s_{11}})\Big]\Big)$$

$$=\delta\Big(-\frac{[0\,t_{n-1}\ \cdot\ t_1]T^{-1}\mathbf{e}_1}{s_{11}}\Big)+\Big[\sigma-(-\frac{[0\ t_{n-1}\ \cdot\ t_1]T^{-1}\mathbf{e}_1}{s_{11}})\Big]T^{-1}\mathbf{e}_n$$

$$=-\frac{1}{s_{11}}Z^TT^{-1}\mathbf{e}_1+\Big[\sigma-(-\frac{[0\ t_{n-1}\ \cdot\ t_1]T^{-1}\mathbf{e}_1}{s_{11}})\Big]T^{-1}\mathbf{e}_n,$$

i.e. the vector $\delta(\sigma)$ is known if the first and the last columns of T^{-1} are known.

Exercise (not for evaluation). Prove that $T\Big(-\frac{1}{s_{11}}Z^TT^{-1}\mathbf{e}_1\Big)=\begin{bmatrix}t_1\\ \cdot\\ t_{n-1}\\ -\frac{[0\,t_{n-1}\ \cdot\ t_1]T^{-1}\mathbf{e}_1}{s_{11}}\end{bmatrix}$.

]

The choices $\varepsilon=1$ and $\beta=-1$ in the above formula for $(\varepsilon-\beta)T^{-1}$ lead to the following well known Ammar-Gader formula, announced above:

$$2T^{-1}=C_1(\mathbf{p})C_{-1}(\mathbf{q}+\mathbf{e}_1)+C_1(\mathbf{e}_1-\mathbf{q})C_{-1}(\mathbf{p}),\quad T(J\mathbf{p})=\mathbf{e}_n,\ T(J\mathbf{q})=\begin{bmatrix}t_1\\ \cdot\\ t_{n-1}\\ t_0\end{bmatrix}.\qquad (AmGa)$$

So, a procedure to solve a linear system $T\mathbf{x}=\mathbf{b}$ is the following:

1) compute $\mathbf{p}$ and $\mathbf{q}$ (for example, if T is real symmetric positive definite, then the cost of this step is $O(n\log_2 n)$ arithmetic operations (see Ammar and Gragg));

2) observe that, by (AmGa) and by the spectral representations of circulant and (-1)-circulant matrices,

$$2T^{-1}=nF\Big[d(F\mathbf{p})F^HDFd(FD(\mathbf{q}+\mathbf{e}_1))+d(F(\mathbf{e}_1-\mathbf{q}))F^HDFd(FD\mathbf{p})\Big]F^H\overline{D},\qquad (+)$$

and compute the vectors $F\mathbf{p}$, $F\mathbf{q}$, $FD\mathbf{p}$, $FD\mathbf{q}$;

3) compute $T^{-1}\mathbf{b}$ via $(+)$ with six FFT.

The operations in 1) and 2) involve only entries of the coefficient matrix of the system. Thus, if we do not count such operations (which can be done in a preprocessing phase), then $O(n\log_2 n)$ arithmetic operations are sufficient to solve $T\mathbf{x}=\mathbf{b}$.

Whenever $A\Pi_\varepsilon-\Pi_\varepsilon A$ is of constant rank with respect to n (f.i. in the case A is a low rank perturbation of a Toeplitz matrix), the matrix A is called Toeplitz-like; it is clear that also for Toeplitz-like matrices efficient formulas for A^{-1} can be stated, because also A^{-1} has constant rank with respect to n.

Two points on the previous lecture:

1) A Theorem that should be known. For any matrix $X\in\mathbb{C}^{n\times n}$, the set $\{A\in\mathbb{C}^{n\times n}:\ A=p(X)\ p=$ polynomials$\}$ is equal to the set $\{A\in\mathbb{C}^{n\times n}:\ AX=XA\}$ if and only if the minimum and the characteristic

polynomials of X coincide. In general, of course, $\{A \in \mathbb{C}^{n\times n} : A = p(X)\ p = \text{polynomials}\} \subset \{A \in \mathbb{C}^{n\times n} : AX = XA\}$, and moreover

$$\dim\{A \in \mathbb{C}^{n\times n} : A = p(X)\ p = \text{polynomials}\} \le n \le \dim\{A \in \mathbb{C}^{n\times n} : AX = XA\}.$$

Proof. Use Jordan canonical form of X. $\square$

2) If Y $n \times n$ is invertible and p_Y is the characteristic polynomial of Y, then $0 = p_Y(Y) = Y^n - \mathrm{tr}\,(Y)Y^{n-1} + \ldots + gY + (-1)^n \det(Y)I$, $Y(Y^{n-1} - \mathrm{tr}\,(Y)Y^{n-2} + \ldots + gI) = (-1)^{n+1}\det(Y)I$, thus

$$Y[(-1)^{n+1}\frac{1}{\det(Y)}(Y^{n-1} - \mathrm{tr}\,(Y)Y^{n-2} + \ldots + gI) = I,$$

so the inverse of a matrix Y is a finite polynomial in Y.

Extension to Toeplitz-plus-Hankel-like matrices via the τ matrix algebra and displacement formulas involving τ

Consider a generic $n \times n$ Toeplitz-plus-Hankel matrix $T + H$, $(T + H)_{ij} = t_{i-j} + h_{i+j-2}$, $1 \le i, j \le n$. Note that

$$\textbf{Exercise.} \qquad X = \begin{bmatrix} 0 & 1 & & & \\ 1 & 0 & 1 & & \\ & 1 & \cdot & \cdot & \\ & & \cdot & 0 & 1 \\ & & & 1 & 0 \end{bmatrix} \quad \Rightarrow \quad [(T+H)X - X(T+H)]_{ij} = 0,\ 2 \le i, j \le n-2,$$

i.e. the matrix $(T + H)X - X(T + H)$ has rank 4. Thus it would be auspicable to state a displacement formula, that on the basis of an assumption of the type $AX - XA = \sum_{m=1}^{\alpha} \mathbf{x}_m \mathbf{y}_m^T$, represents A as the sum of $\alpha + 1$ matrix products $M_i N_i$ involving matrices which are polynomials in X; such formula would allow us to obtain an efficient representation of the inverse of any Toeplitz-plus-Hankel matrix and thus, thinking to the procedure explained in the Toeplitz case, an efficient procedure for solving Toepitz plus Hankel linear systems.

First let us study the set τ of all polynomials in X. A basis for τ is easily obtained by setting

$$J_1 = I, \quad J_2 = X, \quad J_{i+1} = J_i X - J_{i-1},\ i = 2, \ldots, n,$$

in fact the J_i are degree $i-1$ polynomials linearly independent, since $\mathbf{e}_1^T J_i = \mathbf{e}_i^T$, $i = 1, 2, \ldots, n$ (use induction on i to prove this). Moreover, τ cannot have dimension greater than n (by Cayley-Hamilton theorem applied to X), so τ coincides with the Span of $\{J_1, \ldots, J_n\}$, and, given any vector $\mathbf{a} \in \mathbb{C}^n$, it is well defined the matrix of τ whose first row is $\mathbf{a}^T$:

$$\tau(\mathbf{a}) = \sum_{k=1}^{n} a_k J_k.$$

The eigenvalues of X are the scalars $2\cos\frac{j\pi}{n+1}$, $j = 1, \ldots, n$, in fact, the following vector identities hold

$$\begin{bmatrix} 0 & 1 & & & \\ 1 & 0 & 1 & & \\ & 1 & \cdot & \cdot & \\ & & \cdot & 0 & 1 \\ & & & 1 & 0 \end{bmatrix} \begin{bmatrix} \vdots \\ \sin\frac{ij\pi}{n+1} \\ \vdots \end{bmatrix} = 2\cos\frac{j\pi}{n+1} \begin{bmatrix} \vdots \\ \sin\frac{ij\pi}{n+1} \\ \vdots \end{bmatrix}, \quad j = 1, \ldots, n.$$

The eigenvectors $\mathbf{y}_j$, $(\mathbf{y}_j)_i = \sin\frac{ij\pi}{n+1}$ $i = 1, \ldots, n$, are orthogonal ($\mathbf{y}_j^H \mathbf{y}_k = 0$ if $j \ne k$) since they are eigenvectors corresponding to distinct eigenvalues of a hermitian matrix. So, if $Y = [\mathbf{y}_1\ \mathbf{y}_2\ \cdots\ \mathbf{y}_n]$ and

$\Lambda = \text{ diag}\,(2\cos\frac{j\pi}{n+1},\ j=1,\ldots,n)$, then

$$XY = Y\Lambda,\ X = Y\Lambda Y^{-1},\ \ Y^HY = \frac{n+1}{2}I,\ S = \sqrt{\frac{2}{n+1}}Y,\ S^H = S^T = S,\ S^2 = I,\ X = S\Lambda S.$$

Thus $\tau(\mathbf{a}) = \sum_{k=1}^n a_k p_{k-1}(X) = S\Big(\sum_{k=1}^n a_k p_{k-1}(\Lambda)\Big)S = Sd(S\mathbf{a})d(S\mathbf{e}_1)^{-1}S$.

The last identity must hold since also the matrix $Sd(S\mathbf{a})d(S\mathbf{e}_1)^{-1}S$ is a τ matrix and such that $\mathbf{e}_1^T Sd(S\mathbf{a})d(S\mathbf{e}_1)^{-1}S = \mathbf{a}^T$, so it must coincide with $\tau(\mathbf{a})$.

The fact that A times a vector, the eigenvalues of A, the solution of $A\mathbf{x} = \mathbf{b}$ are all things that can be computed with $O(n\log_2 n)$ arithmetic operations when $A \in \tau$ follows from the identity

$$\mathbf{i}(I - F^2_{2(n+1)})F_{2(n+1)} = \begin{bmatrix} 0 & \mathbf{0}^T & 0 & \mathbf{0}^T \\ \mathbf{0} & S & \mathbf{0} & -SJ \\ 0 & \mathbf{0}^T & 0 & \mathbf{0}^T \\ \mathbf{0} & -JS & \mathbf{0} & JSJ \end{bmatrix},\quad F^2 = \begin{bmatrix} 1 & & & & \\ & & & & 1 \\ & & & 1 & \\ & & \cdot & & \\ & 1 & & & \end{bmatrix},$$

which links the $n\times n$ sine matrix with the $2(n+1)\times 2(n+1)$ Fourier matrix.

Exercise. Prove the latter identity involving sine and Fourier discrete transforms.

Exercise. Prove the following Theorem on a displacement formula involving symmetric Hessenberg algebras: Theorem. Let X be the lower Hessenberg matrix in (Hess). Assume X symmetric. Define X' by the following identity

$$X = \begin{bmatrix} r_{11} & b_1\ 0\ \cdots\ 0 \\ b_1 & \\ 0 & \quad X' \\ \cdot & \\ 0 & \end{bmatrix}.$$

Let A be $n\times n$ with complex entries. If $AX - XA = \sum_{m=1}^\alpha \mathbf{x}_m\mathbf{y}_m^T$ $(\mathbf{x}_m, \mathbf{y}_m \in \mathbb{C}^n)$, then

$$A = \sum_{m=1}^{\alpha} \begin{bmatrix} 0 & 0\ \cdots\ \cdots\ 0 \\ 0 & \\ \cdot & \quad H_{X'}(I_n^2\mathbf{x}_m) \\ 0 & \end{bmatrix} H_X(\mathbf{y}_m) + H_X(A^T\mathbf{e}_1),\quad I_n^2\mathbf{x}_m = \begin{bmatrix} (\mathbf{x}_m)_2 \\ \cdot \\ (\mathbf{x}_m)_n \end{bmatrix}.$$

As a consequence, write a displacement formula involving only matrices from the algebra τ, but of different dimensions (choose X as the tridiagonal $0,1$ matrix which generates τ). Finally, show that one can determine from the latter formula an efficient representation of the inverse of a Toeplitz plus Hankel matrix.

The best approximation of A in $\mathcal{L} = \text{Span}\,\{J_1, J_2, \ldots, J_m\}$

Let $A \in \mathbb{C}^{n\times n}$, and $\mathcal{L} = \text{Span}\,\{J_1, J_2, \ldots, J_m\}$, with J_k linearly independent. Then it is well defined the matrix $\mathcal{L}_A$ in $\mathcal{L}$ such that

$$\|\mathcal{L}_A - A\|_F = \min_{X\in\mathcal{L}} \|X - A\|_F,$$

and the matrix $\mathcal{L}_A$ is also characterized by the orthogonality condition $(X, \mathcal{L}_A - A)_F = 0$, $X \in \mathcal{L}$. In fact, $\mathbb{C}^{n\times n}$ is a Hilbert space with respect to the inner product $(X,Y)_F = \sum_{i,j=1}^n \overline{x}_{ij}y_{ij}$, the norm induced from $(\cdot,\cdot)_F$ is the Frobenius norm, and $\mathcal{L}$ is a closed subspace of $\mathbb{C}^{n\times n}$, thus, by the Hilbert projection theorem, it is well defined the projection in $\mathcal{L}$ of any $A \in \mathbb{C}^{n\times n}$.

We call $\mathcal{L}_A$ the best (least squares) approximation of A in $\mathcal{L}$.

Here below there are some simple results that follow immediately from the definition of $\mathcal{L}_A$:

- If A is real ($\overline{A} = A$), then $\mathcal{L}_A$ is real if and only if $\overline{\mathcal{L}} \subset \mathcal{L}$.

Proof. Assume $\overline{\mathcal{L}} \subset \mathcal{L}$: $\|\mathcal{L}_A - A\|_F = \|\overline{\mathcal{L}_A} - \overline{A}\|_F = \|\overline{\mathcal{L}_A} - A\|_F \Rightarrow \overline{\mathcal{L}_A} = \mathcal{L}_A$ (since $\overline{\mathcal{L}_A} \in \mathcal{L}$).

For example, for $\mathcal{L} = C_\varepsilon$ one has $\overline{\mathcal{L}} \subset \mathcal{L}$ if and only if $\varepsilon \in \mathbb{R}$: C_ε is closed under conjugation ($\overline{C_\varepsilon} \subset C_\varepsilon$) if and only if $\varepsilon \in \mathbb{R}$.

In fact, for $n = 2$ we have

$$J_2 = \begin{bmatrix} 0 & 1 \\ \varepsilon & 0 \end{bmatrix}, \quad \overline{J}_2 = \begin{bmatrix} 0 & 1 \\ \overline{\varepsilon} & 0 \end{bmatrix}$$

and it is clear that in order to have $\overline{J}_2 \in C_\varepsilon$, i.e. $\overline{J}_2$ linear combination of J_1, J_2, the only possibility is that $\overline{J}_2 = J_2$, and that this can happen if and only if $\varepsilon \in \mathbb{R}$. Nothing changes if n is generic.

- If A is hermitian ($A^H = A$), then $\mathcal{L}_A$ is hermitian if and only if $\mathcal{L}^H \subset \mathcal{L}$.

Proof. Assume $\mathcal{L}^H \subset \mathcal{L}$: $\|\mathcal{L}_A - A\|_F = \|\mathcal{L}_A^H - A^H\|_F = \|\mathcal{L}_A^H - A\|_F \Rightarrow \mathcal{L}_A^H = \mathcal{L}_A$ (since $\mathcal{L}_A^H \in \mathcal{L}$).

For example, for $\mathcal{L} = C_\varepsilon$ one has $\mathcal{L}^H \subset \mathcal{L}$ if and only if $|\varepsilon| = 1$: C_ε is closed under conjugate transposition ($C_\varepsilon^H \subset C_\varepsilon$) if and only if $|\varepsilon| = 1$.

In fact, for $n = 3$ we have

$$J_1 = I, \quad J_2 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ \varepsilon & 0 & 0 \end{bmatrix}, \quad (J_2)^H = \begin{bmatrix} 0 & 0 & \overline{\varepsilon} \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad J_3 = \begin{bmatrix} 0 & 0 & 1 \\ \varepsilon & 0 & 0 \\ 0 & \varepsilon & 0 \end{bmatrix},$$

and it is clear that in order to have $J_2^H \in C_\varepsilon$, i.e. J_2^H linear combination of J_1, J_2, J_3, the only possibility is that $J_2^H = \overline{\varepsilon} J_3$, and that this can happen if and only if $|\varepsilon| = 1$. Nothing changes if n is generic.

- If A is real symmetric ($A^H = \overline{A} = A$), then $\mathcal{L}_A$ is real symmetric if and only if $\overline{\mathcal{L}} \subset \mathcal{L}$ and $\mathcal{L}^H \subset \mathcal{L}$.

For example, for $\mathcal{L} = C_\varepsilon$ one has $\overline{\mathcal{L}} \subset \mathcal{L}$ and $\mathcal{L}^H \subset \mathcal{L}$ if and only if $\varepsilon = \pm 1$: C_ε is closed under conjugation and under conjugate transposition ($\overline{C_\varepsilon} \subset C_\varepsilon$ and $C_\varepsilon^H \subset C_\varepsilon$) if and only if $\varepsilon = \pm 1$.

Set

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \mathcal{L} = C_\varepsilon.$$

Prove that $\mathcal{L}_A^H \neq \mathcal{L}_A$ ($\mathcal{L}_A^H \notin \mathcal{L}$) if $|\varepsilon| \neq 1$.
Prove that $\overline{\mathcal{L}_A} \neq \mathcal{L}_A$ ($\mathcal{L}_A \in \mathbb{C}^{2\times 2} \backslash \mathbb{R}^{2\times 2}$) if $\varepsilon \in \mathbb{C} \backslash \mathbb{R}$.

Solving: Consider the real symmetric matrix $A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$. The values of α, β for which

$$\|A - C_\varepsilon(\begin{bmatrix} \alpha \\ \beta \end{bmatrix})\|_F = \left\| \begin{bmatrix} -\alpha & 1-\beta \\ 1-\beta\varepsilon & -\alpha \end{bmatrix} \right\|_F = 2|\alpha|^2 + |1-\beta|^2 + |1-\beta\varepsilon|^2$$

is minimum are $\alpha = 0$ and β some point on the segment connecting 1 to $1/\varepsilon$. Thus $(C_\varepsilon)_A = \beta \begin{bmatrix} 0 & 1 \\ \varepsilon & 0 \end{bmatrix}$. ...

(It is not necessary to compute exactly $\mathcal{L}_A$ to solve this exercise).

Exercise. Let U be a $n \times n$ unitary matrix, and consider the set of matrices $\mathcal{L} = \{Ud(\mathbf{z})U^H : \mathbf{z} \in \mathbb{C}^n\}$ (examples of such sets $\mathcal{L}$ are τ and C_ε, $|\varepsilon| = 1$). The set $\mathcal{L}$ is a closed subspace of $\mathbb{C}^{n\times n}$, closed under matrix multiplication and commutative. Investigate under what conditions on U, the inclusions $\overline{\mathcal{L}} \subset \mathcal{L}$ and $\mathcal{L}^H \subset \mathcal{L}$ hold, separately or together.

Partial solution: $\overline{\mathcal{L}} \subset \mathcal{L}$ if U real, or if $\overline{U} = UPd(\mathbf{x})$, P permutation, $d(\mathbf{x})$ unitary. ...

SIXTH LECTURE, Wednesday September 10, 2014 (Rome)

Question: characterize all spaces $\mathcal{L} = \{J_1, J_2, \ldots, J_m\} \subset \mathbb{C}^{n\times n}$ (J_j linearly independent) for which

$$A \text{ hermitian positive definite implies } \mathcal{L}_A \text{ hermitian positive definite.} \qquad (hpd)$$

Possible way of investigation: Introduce an operator *hpd* such that $hpd(A) = A$ if and only if A is hermitian positive definite. Once such operator is introduced may be the characterization would be $hpd(\mathcal{L}) \subset \mathcal{L}$ (analogously to the conditions $\overline{\mathcal{L}} \subset \mathcal{L}$ and $\mathcal{L}^H \subset \mathcal{L}$ that we have stated above).

The implication (hpd) is true if $\mathcal{L}$ is the set of all matrices diagonalized by a unitary matrix (see below). More general classes of n-dimensional spaces $\mathcal{L}$ for which (hpd) is true are introduced in [Di Fiore, Zellini, 2001]. In particular, (hpd) turns out to be true also if $\mathcal{L}$ is a group algebra, i.e. $\mathcal{L} = \{A \in \mathbb{C}^{n\times n} : a_{ij} = a_{ki,kj},\ i,j,k \in \mathcal{G}\}$, where $\mathcal{G} = \{1, 2, \ldots, n\}$ is a group.

A representation of $\mathcal{L}_A$. The orthogonality condition characterizing $\mathcal{L}_A = \sum_{k=1}^m \alpha_k J_k$ can be rewritten as follows

$$0 = (J_s, A - \sum_{k=1}^m \alpha_k J_k)_F = (J_s, A)_F - \sum_{k=1}^m \alpha_k (J_s, J_k)_F, \quad s = 1, 2, \ldots, m.$$

So, the vector α of the coefficients of $\mathcal{L}_A$ must be such that

$$B\alpha = \mathbf{c}, \quad B_{s,k} = (J_s, J_k)_F,\ c_s = (J_s, A)_F,\ 1 \le s, k \le m,$$

i.e. $\mathcal{L}_A = \sum_{k=1}^m (B^{-1}\mathbf{c})_k J_k$. Note that the Gram matrix B is hermitian positive definite. In [Di Fiore, Zellini, 2001] it is shown that for a class of n-dimensional spaces $\mathcal{L}$ the matrix B is itself a matrix of $\mathcal{L}$.

Exercise. Find the first row of the matrix $(C_\varepsilon)_T$, $\varepsilon \in \mathbb{R}$, where $T = (t_{i-j})_{i,j=1}^n$, i.e. $\mathbf{z} \in \mathbb{C}^n$ such that $(C_\varepsilon)_T = C_\varepsilon(\mathbf{z})$, and note that it can be computed in $O(n)$ arithmetic operations.

In the case $\mathcal{L} = \{Ud(\mathbf{z})U^H : \mathbf{z} \in \mathbb{C}^n\}$, U unitary $n \times n$, we have a further representation for $\mathcal{L}_A$. In fact,

$$\|Ud(\mathbf{z})U^H - A\|_F = \|d(\mathbf{z}) - U^H A U\|_F$$

is clearly minimum if $z_i = (U^H A U)_{ii}$. So, $\mathcal{L}_A = U \operatorname{diag}((U^H A U)_{ii}) U^H$.

Exercise. If $\mathcal{L} = \{Ud(\mathbf{z})U^H : \mathbf{z} \in \mathbb{C}^n\}$ with U unitary, then $\mathcal{L}_A$ is hermitian or hermitian positive definite whenever A is hermitian or hermitian positive definite.

As an example, let us calculate $\mathcal{L}_{\mathbf{y}\mathbf{y}^T}$, $\mathbf{y} \in \mathbb{C}^n$. We have

$$\mathcal{L}_{\mathbf{y}\mathbf{y}^T} = U \operatorname{diag}((U^H \mathbf{y}\mathbf{y}^T U)_{ii}) U^H = U \operatorname{diag}((U^H\mathbf{y})_i (U^T\mathbf{y})_i) U^H.$$

In particular, if $\mathbf{y} \in \mathbb{R}^n$, then $\mathcal{L}_{\mathbf{y}\mathbf{y}^T} = Ud(|U^H\mathbf{y}|^2)U^H$, where for a vector $\mathbf{v} \in \mathbb{C}^n$ by the symbol $|\mathbf{v}|^2$ we mean the vector $[|v_1|^2\ |v_2|^2\ \cdots\ |v_n|^2]^T$.

Exercise. Let X be the $n \times n$ $0,1$ tridiagonal matrix that generates the algebra τ, and assume n even. Show that there exists $\mathbf{z}$ such that $\mathbf{z}^T X = \mathbf{e}_1^T$, and thus X is invertible and $X^{-1} = \tau(\mathbf{z})$. Write the matrix $\tau(\mathbf{z})$ for $n = 6$ by exploiting the fact that the entries a_{ij} of any τ matrix A satisfy the cross-sum rule with zero border conditions, i.e.

$$a_{i,j-1} + a_{i,j+1} = a_{i+1,j} + a_{i-1,j}, \quad a_{0,j} = a_{n+1,j} = a_{i,0} = a_{i,n+1} = 0, \quad 1 \le i, j \le n$$

(this cross-sum rule follows from the equalities $[AX]_{ij} = [XA]_{ij}$).

Low complexity matrix algebras $\mathcal{L}$ and $\mathcal{L}_A$ in iterative methods for solving real symmetric positive definite Toeplitz linear systems

Let $T = (t_{|i-j|})_{i,j=1}^n$ be a real symmetric positive definite Toeplitz matrix.

Exercise. Well known examples of such matrices are the two corresponding, respectively, to the choices $t_0 = 2$, $t_1 = -1$, $t_k = 0$, $k > 1$, and $t_k = p^k$ with $p \in \mathbb{R}$ such that $|p| < 1$. Why they are positive definite?

Assume we have to solve a linear system of type $T\mathbf{x} = \mathbf{b}$, $\mathbf{b} \in \mathbb{R}^n$. First note that finding $T^{-1}\mathbf{b}$ is equivalent to finding the minimum of the following function from $\mathbb{R}^n$ to $\mathbb{R}$:

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T T\mathbf{x} - \mathbf{x}^T\mathbf{b},$$

in fact $\nabla f(\mathbf{x}) = T\mathbf{x} - \mathbf{b}$ and $\nabla^2 f(\mathbf{x}) = T$, thus f has only one stationary point, $T^{-1}\mathbf{b}$, which is a global minimum since f is strictly convex in $\mathbb{R}^n$ (the Hessian is strictly definite positive everywhere).

Let $\mathbf{x}_k \in \mathbb{R}^n$ be an approximation of $T^{-1}\mathbf{b}$. In order to see if it is a good approximation, one can evaluate a norm of the residual $\mathbf{r}_k = \mathbf{b} - T\mathbf{x}_k$, which turns out to be equal to $-\nabla f(\mathbf{x}_k)$. If it is not a good approximation of $T^{-1}\mathbf{b}$, how to generate a better approximation?

First introduce a descent search direction in $\mathbf{x}_k$ for f, i.e. a vector $\mathbf{d}_k$ such that $\mathbf{d}_k^T \nabla f(\mathbf{x}_k) < 0$ (recall that $\nabla f(\mathbf{x}_k)$ is the direction of max increasing of f in a neighborhood of $\mathbf{x}_k$ and is orthogonal to the level ipersurface $\{\mathbf{x} \in \mathbb{R}^n : f(\mathbf{x}) = f(\mathbf{x}_k)\}$ (which is a neighborhood of $T^{-1}\mathbf{b}$ in the energy metric $\|\mathbf{u} - \mathbf{v}\|_T \ldots$).

Then find $\lambda_k > 0$ such that $f(\mathbf{x}_k + \lambda_k \mathbf{d}_k) = \min_{\lambda \in \mathbb{R}} f(\mathbf{x}_k + \lambda \mathbf{d}_k)$ and set $\mathbf{x}_{k+1} = \mathbf{x}_k + \lambda_k \mathbf{d}_k$. Note that such λ_k is uniquely defined, it is the abscissa of the vertex of the convex parabola $f(\mathbf{x}_k + \lambda \mathbf{d}_k)$

Exercise. Find the explicit formula of such λ_k.

A suitable choice at each step k of the descent search direction $\mathbf{d}_k$, allows to generate a sequence $\{\mathbf{x}_k\}$ of approximations of $T^{-1}\mathbf{b}$ convergent to $T^{-1}\mathbf{b}$.

The CG method

a) In particular, in the CG (Conjugate Gradient) method at each step $\mathbf{d}_k$ is defined as follows:

$$\mathbf{d}_k = -\nabla f(\mathbf{x}_k) + \beta \mathbf{d}_{k-1}, \quad \beta \text{ such that } \mathbf{d}_k^T T \mathbf{d}_{k-1} = 0$$

($\mathbf{d}_0 = -\nabla f(\mathbf{x}_0)$). It can be shown that if this choice of $\mathbf{d}_k$ is applied, for $k = 0, 1, 2, \ldots$, then in no more than n steps one obtains $T^{-1}\mathbf{b}$, i.e. there exists $s \leq n$ such that $\mathbf{x}_s = T^{-1}\mathbf{b}$. This result can be easily understood geometrically in the case $n = 2$.

Actually, a stronger result holds:

Let m be the number of distinct eigenvalues of T. Then there exists $s \leq m$ such that $\mathbf{x}_s = T^{-1}\mathbf{b}$.

b) Each step of the CG method require the computation of a matrix-vector product where the matrix is T (besides some scalar products whose cost is $O(n)$). We know that such computation can be done with $O(n \log_2 n)$ arithmetic operations.

c) When the eigenvalues of the coefficient matrix are "clustered" the rate of convergence is much greater, in the sense that we obtain a good approximation of $T^{-1}\mathbf{b}$ after a relatively small number of iterations.

For example, it can be shown the following bound for the error after k step of CG method applied to $T\mathbf{x} = \mathbf{b}$:

$$\|\mathbf{x}_k - T^{-1}\mathbf{b}\|_T \leq c(k) 2 \left(\frac{\sqrt{\frac{1+\varepsilon}{1-\varepsilon}} - 1}{\sqrt{\frac{1+\varepsilon}{1-\varepsilon}} + 1}\right)^{k - r_\varepsilon} \|\mathbf{x}_0 - T^{-1}\mathbf{b}\|_T, \quad k \geq r_\varepsilon,$$

where ε is arbitrary in (0,1), r_ε is the number of the eigenvalues of T which are not in $[1-\varepsilon, 1+\varepsilon]$, $c(k) = 1$ if no eigenvalue of T is less than $1-\varepsilon$, and $c(k) > 1$ and grows with k if some eigenvalues of T are near zero.

The above bound says that even in case $\mu_2(T) = \max \lambda_i(T) / \min \lambda_i(T)$ is great, the CG method can produce a good approximation of $T^{-1}\mathbf{b}$ after a small number of iterations if most of the eigenvalues of T are

in $[1-\varepsilon, 1+\varepsilon]$ for a small ε. And if this distribution of the eigenvalues, good for the rate of convergence of CG, does not hold for T, we can try to introduce a real symmetric positive definite linear system $\tilde{T}\mathbf{y} = \tilde{\mathbf{b}}$ equivalent to $T\mathbf{x} = \mathbf{b}$, but where $\tilde{T}$ has a better distribution of eigenvalues, and apply CG to this second system. This is equivalent to minimize the function $\tilde{f}(\mathbf{y}) = \frac{1}{2}\mathbf{y}\tilde{T}\mathbf{y} - \mathbf{y}^T\tilde{\mathbf{b}}$.

d) Let $\mathcal{L}$ be a subspace of $\mathbb{C}^{n\times n}$ such that $\mathcal{L}_T$, the projection of T on $\mathcal{L}$, is real symmetric positive definite (for example $\mathcal{L} = C_{\pm 1}, \tau$ or $\mathcal{L} = \{Ud(\mathbf{z})U^H : \mathbf{z} \in \mathbb{C}^n\}$ with U real unitary, or $\mathcal{L}$ =a group matrix algebra,...).

Then $\mathcal{L}_T = EE^T$, for some E non singular, and we can consider the linear system

$$\tilde{T}\mathbf{y} = (E^{-1}TE^{-T})(E^T\mathbf{x}) = E^{-1}\mathbf{b} = \tilde{\mathbf{b}}$$

equivalent to $T\mathbf{x} = \mathbf{b}$, but with a coefficient matrix which is similar to $\mathcal{L}_T^{-1}T$ (prove it!), and thus has its same eigenvalues. Now, for some classes of real symmetric positive definite Toeplitz matrices T there are suitable spaces $\mathcal{L}$ for which the eigenvalues of $\mathcal{L}_T^{-1}T$ *cluster around* 1, briefly $\sigma(\mathcal{L}_T^{-1}T) \approx 1$. Here below we claim a theorem with a result of this type.

Theorem. Let $\{t_k\}_{k=0}^{+\infty}$, $t_k \in \mathbb{R}$, be such that $\sum_{k=0}^{+\infty} |t_k| < +\infty$. Set $t(\theta) = \sum_{k\in\mathbb{Z}} t_{|k|}e^{\mathbf{i}k\theta} = t_0 + 2\sum_{k=1}^{+\infty} t_k \cos(k\theta)$, $t_m = \min t(\theta)$, and $t_M = \max t(\theta)$. Finally, set $T^{(n)} = (t_{|i-j|})_{i,j=1}^n$,

Then $\sigma(T^{(n)}) \subset [t_m, t_M]$ for all n.

Let $\mathcal{L} = C_{\pm 1}$ or $\mathcal{L} = \tau$. If $t_m > 0$ (note that in this case the $T^{(n)}$ and the $\mathcal{L}_{T^{(n)}}$ are all real symmetric positive definite matrices), then $\sigma(\mathcal{L}_{T^{(n)}}^{-1}T^{(n)}) \approx 1$, or, more precisely, $\forall\,\varepsilon > 0$, there exist $\nu_\varepsilon, k_\varepsilon \in \mathbb{N}$ such that for all $n > \nu_\varepsilon$ in $[1-\varepsilon, 1+\varepsilon]^c$ there are at most k_ε eigenvalues of $\mathcal{L}_{T^{(n)}}^{-1}T^{(n)}$.

Exercise. Prove the result $\sigma(T^{(n)}) \subset [t_m, t_M]$ stated in the above theorem.

In [Di Fiore, Zellini, 2001] it is shown that the clustering result stated in the theorem can be extended to matrix algebras $\mathcal{L} = \{Ud(\mathbf{z})U^H : \mathbf{z} \in \mathbb{C}^n\}$ where U are Hartley-type transforms. Recall here only the Hartley transform: $U_{ij} = \frac{1}{\sqrt{n}}(\cos\frac{2ij\pi}{n} + \sin\frac{2ij\pi}{n})_{i,j=0}^{n-1}$.

As a consequence of the result stated in the theorem, by applying the CG method to the system $(E^{-1}TE^{-T})(E^T\mathbf{x}) = E^{-1}\mathbf{b}$, or, equivalently, by mimimizing the function

$$\frac{1}{2}\mathbf{y}^T(E^{-1}TE^{-T})\mathbf{y} - \mathbf{y}^T(E^{-1}\mathbf{b}),$$

by the CG method, one obtains a sequence of approximations $\mathbf{y}_k$ which converge superlinearly to $\mathbf{y} = E^T\mathbf{x}$.
...

Low complexity matrix algebras $\mathcal{L}$ in unconstrained minimization methods

Assume we have to minimize a generic function $f : \mathbb{R}^n \to \mathbb{R}$. Let $\mathbf{x}_k \in \mathbb{R}^n$ be an approximation of $\mathbf{x}_*$, a (local) minimum of f. In order to see if it is a good approximation, one can evaluate a norm of $\nabla f(\mathbf{x}_k)$. If $\mathbf{x}_k$ is not a good approximation of $\mathbf{x}_*$, how to generate a better approximation?

First introduce a descent search direction in $\mathbf{x}_k$ for f, i.e. a vector $\mathbf{d}_k$ such that $\mathbf{d}_k^T\nabla f(\mathbf{x}_k) < 0$ (recall that $\nabla f(\mathbf{x}_k)$ is the direction of max increasing of f in a neighbourhood of $\mathbf{x}_k$ and is orthogonal to the level ipersurface $\{\mathbf{x} \in \mathbb{R}^n : f(\mathbf{x}) = f(\mathbf{x}_k)\}$).

Then find $\lambda_k > 0$ suitable such that $f(\mathbf{x}_k + \lambda_k\mathbf{d}_k)$ is enough smaller than $f(\mathbf{x}_k)$, and $\mathbf{x}_k + \lambda_k\mathbf{d}_k$ is enough far from $\mathbf{x}_k$, and set $\mathbf{x}_{k+1} = \mathbf{x}_k + \lambda_k\mathbf{d}_k$. Procedures defining a step-length λ_k with such properties are due to Armijo, Goldstein and Wolfe. They have to be preferred to the procedures that try to compute λ_k for which $f(\mathbf{x}_k + \lambda\mathbf{d}_k)$, $\lambda \in \mathbb{R}$, is (at least locally) minimum in λ_k, because the latter are too expensive. Note that the most known Armijo-Goldstein-Wolfe procedures (line searches) for defining λ_k yield a new guess $\mathbf{x}_{k+1}$ such that $\mathbf{s}_k^T\mathbf{y}_k > 0$, where $\mathbf{s}_k = \mathbf{x}_{k+1} - \mathbf{x}_k$, $\mathbf{y}_k = \nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)$ (see for instance [Dennis, Schnabel, 1983], [Di Fiore et al, 2003] or [Cipolla et al, 2014]).

The Newton method

In Newton method $\mathbf{d}_k = -\nabla^2 f(\mathbf{x}_k)^{-1}\nabla f(\mathbf{x}_k)$.

This choice of $\mathbf{d}_k$ is very good from a mathematically point of view if $\mathbf{x}_k$ is near the minimum we are approaching. In fact, in this case $\mathbf{d}_k^T\nabla f(\mathbf{x}_k)$ is of course negative (since $\nabla^2 f(\mathbf{x}_k)$ is r.s.p.d.) and thus $\mathbf{d}_k$ is a descent direction for f in $\mathbf{x}_k$, and, moreover, $\mathbf{x}_{k+1}$ turns out to approximate $\mathbf{x}_*$ much better than $\mathbf{x}_k$.

But from other points of view, the Newton direction is not so good. First it fails to be a descent direction in $\mathbf{x}_k$ for f if $\mathbf{x}_k$ is "far" from $\mathbf{x}_*$ ($\nabla^2 f(\mathbf{x}_k)$ could be not positive definite). Second, the computation of it (at each step) can be too expensive, since it requires the evaluation of $n(n+1)/2$ second derivatives (the entries of $\nabla^2 f(\mathbf{x}_k)$) and the solution of a linear system ($\nabla^2 f(\mathbf{x}_k)\mathbf{d}_k = -\nabla f(\mathbf{x}_k)$), which in general requires $O(n^3)$ arithmetic operations.

The Quasi-Newton Secant methods and BFGS

In quasi-Newton methods $\mathbf{d}_k = -B_k^{-1}\nabla f(\mathbf{x}_k)$, where B_k is chosen at each step real symmetric positive definite. [Note the generality of quasi-Newton methods; in fact, any descent direction in $\mathbf{x}_k$ for f, say $\mathbf{d}_k$, must be of the type $\mathbf{d}_k = -B_k^{-1}\nabla f(\mathbf{x}_k)$ for some r.s.p.d. matrix B_k (see [Di Fiore, Fanelli, Zellini, 2007]).]

An important requirement made on B_k is usually that

$$B_k\mathbf{s}_{k-1} := B_k(\mathbf{x}_k - \mathbf{x}_{k-1}) = \nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}_{k-1}) =: \mathbf{y}_{k-1}. \qquad (sec)$$

This condition is the vectorial analogous of the scalar condition $b_k(x_k - x_{k-1}) = f'(x_k) - f'(x_{k-1})$ which defines uniquely the well known secant method for finding the stationary points of a function $f : \mathbb{R} \to \mathbb{R}$:

$$x_{k+1} = x_k - \lambda_k b_k^{-1} f'(x_k) = x_k - \lambda_k \frac{f'(x_k)}{\frac{f'(x_k)-f'(x_{k-1})}{x_k - x_{k-1}}}.$$

If B_k solves the *secant equation* (sec) then the quasi-Newton direction $\mathbf{d}_k = -B_k^{-1}\nabla f(\mathbf{x}_k)$ is said secant.

Among the many possible real symmetric positive matrices B_k solving the secant equation (there is an infinite number of such matrices if $\mathbf{s}_{k-1}^T\mathbf{y}_{k-1} > 0$), the most effective is the one proposed simultaneously by Broyden, Fletcher, Goldfarb, Shanno (BFGS). In the following we illustrate the corresponding BFGS minimization method (which can be implemented with $O(n^2)$ per step and require no computation of second derivatives) and some much cheaper but efficient versions of such method, which involve low complexity matrix algebras $\mathcal{L}$ and best least squares fit to B_k in $\mathcal{L}$ (see the previous sections) and thus allow the application of the very effective BFGS-type scheme to large scale minimization problems.

Assume we already have B_k (from the previous step); then in BFGS the matrix B_{k+1} is defined from B_k by the following identity:

$$B_{k+1} = \Phi(B_k, \mathbf{s}_k, \mathbf{y}_k), \quad \mathbf{s}_k = \mathbf{x}_{k+1} - \mathbf{x}_k, \ \mathbf{y}_k = \nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k),$$

$$\Phi(B, \mathbf{s}, \mathbf{y}) = B + \frac{1}{\mathbf{y}^T\mathbf{s}}\mathbf{y}\mathbf{y}^T - \frac{1}{\mathbf{s}^T B\mathbf{s}} B\mathbf{s}\mathbf{s}^T B.$$

Note that $B_{k+1}\mathbf{s}_k = \mathbf{y}_k$, so BFGS is a quasi-Newton secant method. Let us resume here below the main instructions of the BFGS algorithm:

$\mathbf{x}_0 \in \mathbb{R}^n$, $B_0 \in \mathbb{R}^{n\times n}$ *r.s.p.d.*
For $k = 0, 1, 2, \ldots$ {
$\mathbf{d}_k = -B_k^{-1}\nabla f(\mathbf{x}_k)$
$\mathbf{x}_{k+1} = \mathbf{x}_k + \lambda_k\mathbf{d}_k$, *with* λ_k *suitable so that (it is possible!)* $f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k) - \eta_k$, $\eta_k > 0$, *and...*
$\mathbf{s}_k = \mathbf{x}_{k+1} - \mathbf{x}_k$, $\mathbf{y}_k = \nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)$ ... $\mathbf{s}_k^T\mathbf{y}_k > 0$. *But this implies...*
$B_{k+1} = \Phi(B_k, \mathbf{s}_k, \mathbf{y}_k)$ ...B_{k+1} *r.s.p.d., and thus* $\mathbf{d}_{k+1}$ *desc direction in* $\mathbf{x}_{k+1}$ }

Each step of the above algorithm is well defined, and a strictly decreasing sequence $\{f(\mathbf{x}_k)\}$ is produced. Under suitable assumptions on f, the method has a local superlinear rate of convergence, i.e. if $\mathbf{x}_0 \approx \mathbf{x}_*$,

then for all k we have $\|\mathbf{x}_k - \mathbf{x}_*\| \leq \eta_k \|\mathbf{x}_{k-1} - \mathbf{x}_*\|$, with $\eta_k \to 0$ (the proof of this not obvious result is due to Dennis-More'). The method can be implemented with $O(n^2)$ arithmetic operations per step. In fact, by Sherman-Morrison formula it is possible to obtain the following expression of B_{k+1}^{-1} in terms of B_k^{-1},

$$B_{k+1}^{-1} = \Big(I - \frac{\mathbf{y}_k \mathbf{s}_k^T}{\mathbf{y}_k^T \mathbf{s}_k}\Big)^T B_k^{-1} \Big(I - \frac{\mathbf{y}_k \mathbf{s}_k^T}{\mathbf{y}_k^T \mathbf{s}_k}\Big) + \frac{\mathbf{s}_k \mathbf{s}_k^T}{\mathbf{y}_k^T \mathbf{s}_k}.$$

Then, thanks to this expression, $B_{k+1}^{-1} \nabla f(\mathbf{x}_{k+1})$ can be computed by performing a matrix-vector product involving B_k^{-1} and cheaper computations, such as scalar products or multiplications of vectors by scalars. ...

BFGS-type and $\mathcal{L}^{(k)}$ QN methods

With the aim to reduce the space and time-per-step complexity of BFGS method, one can modify the above BFGS procedure by defining, at each step, in corrispondence with the current Hessian approximation B_k, a lower complexity r.s.p.d. matrix $\tilde{B}_k$ which maintains as more as possible information on the structure and on the spectrum of B_k, and by updating, via the BFGS-type iterative scheme, in place of B_k, such matrix $\tilde{B}_k$.

This idea produces the following BFGS-type secant algorithm:

SECANT BFGS-type:
$\mathbf{x}_0 \in \mathbb{R}^n$, $B_0 \in \mathbb{R}^{n \times n}$ r.s.p.d.
For $k = 0, 1, 2, \ldots$:
$\mathbf{d}_k = -B_k^{-1} \nabla f(\mathbf{x}_k)$
$\mathbf{x}_{k+1} = \mathbf{x}_k + \lambda_k \mathbf{d}_k$, ... λ_k ...
$\mathbf{s}_k = \mathbf{x}_{k+1} - \mathbf{x}_k$, $\mathbf{y}_k = \nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)$ ... $\mathbf{s}_k^T \mathbf{y}_k > 0$ ...
define $\tilde{B}_k$ r.s.p.d. with $\tilde{B}_k \approx B_k$ (in some sense)
$B_{k+1} = \Phi(\tilde{B}_k, \mathbf{s}_k, \mathbf{y}_k)$ ... B_{k+1} r.s.p.d. ...

Note that yet $B_{k+1}\mathbf{s}_k = \mathbf{y}_k$ and a well defined decreasing sequence $f(\mathbf{x}_k)$ is produced. How to define the $\tilde{B}_k$? Assume that at each step k a space $\mathcal{L}^{(k)} \subset \mathbb{C}^{n \times n}$ is introduced with the property

$$\mathcal{L}^{(k)} \text{ such that } A \text{ r.s.p.d. implies } \mathcal{L}_A^{(k)} \text{ r.s.p.d..} \qquad (updde fpos)$$

(note that such $\mathcal{L}^{(k)}$ must necessarily satisfy the inclusions $\overline{\mathcal{L}^{(k)}} \subset \mathcal{L}^{(k)}$ and $(\mathcal{L}^{(k)})^H \subset \mathcal{L}^{(k)}$). Then one can choose, in the above algorithm, $\tilde{B}_k = \mathcal{L}_{B_k}^{(k)}$:

SECANT $\mathcal{L}^{(k)}$QN:
$\mathbf{x}_0 \in \mathbb{R}^n$, $B_0 \in \mathbb{R}^{n \times n}$ r.s.p.d.
For $k = 0, 1, 2, \ldots$:
$\mathbf{d}_k = -B_k^{-1} \nabla f(\mathbf{x}_k)$
$\mathbf{x}_{k+1} = \mathbf{x}_k + \lambda_k \mathbf{d}_k$, ... λ_k ...
$\mathbf{s}_k = \mathbf{x}_{k+1} - \mathbf{x}_k$, $\mathbf{y}_k = \nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)$ ... $\mathbf{s}_k^T \mathbf{y}_k > 0$ ...
$B_{k+1} = \Phi(\mathcal{L}_{B_k}^{(k)}, \mathbf{s}_k, \mathbf{y}_k)$ ... B_{k+1} r.s.p.d. ...

If $\mathcal{L}^{(k)} = \mathcal{L}$ for all k and $\mathcal{L} = \{Ud(\mathbf{z})U^H : \mathbf{z} \in \mathbb{C}^n\}$ with U fast discrete transform (f.i. $\mathcal{L} = C_{\pm 1}$, $\mathcal{L} = \tau$, $\mathcal{L}$ =Hartley algebra), or if $\mathcal{L}^{(k)} = \{U_k d(\mathbf{z}) U_k^H : \mathbf{z} \in \mathbb{C}^n\}$ with U_k suitable fast discrete transforms (f.i. $\mathcal{L}$ generated by a Householder or by two Householder matrices), then *the eigenvalues of $\mathcal{L}_{B_{k+1}}^{(k+1)}$ can be obtained from those of $\mathcal{L}_{B_k}^{(k)}$ at the cost of transforms involving U_k and U_{k+1}*, and this result, together with the formula

$$B_{k+1}^{-1} = \Big(I - \frac{\mathbf{y}_k \mathbf{s}_k^T}{\mathbf{y}_k^T \mathbf{s}_k}\Big)^T (\mathcal{L}_{B_k}^{(k)})^{-1} \Big(I - \frac{\mathbf{y}_k \mathbf{s}_k^T}{\mathbf{y}_k^T \mathbf{s}_k}\Big) + \frac{\mathbf{s}_k \mathbf{s}_k^T}{\mathbf{y}_k^T \mathbf{s}_k},$$

let us conclude that *the cost of each step of the above $\mathcal{L}^{(k)}$QN secant method is of the order of the cost of transforms involving U_k and U_{k+1}.*

Let us prove this in the case $\mathcal{L}^{(k)} = \mathcal{L}$, $\forall\, k$: by projecting on $\mathcal{L}$ the equation $B_{k+1} = \phi(\mathcal{L}_{B_k}, \mathbf{s}_k, \mathbf{y}_k) = \mathcal{L}_{B_k} + \frac{1}{\mathbf{y}_k^T\mathbf{s}_k}\mathbf{y}_k\mathbf{y}_k^T - \frac{1}{\mathbf{s}_k^T\mathcal{L}_{B_k}\mathbf{s}_k}\mathcal{L}_{B_k}\mathbf{s}_k\mathbf{s}_k^T\mathcal{L}_{B_k}$, we obtain

$$\mathcal{L}_{B_{k+1}} = \mathcal{L}_{B_k} + \frac{1}{\mathbf{y}_k^T\mathbf{s}_k}\mathcal{L}_{\mathbf{y}_k\mathbf{y}_k^T} - \frac{1}{\mathbf{s}_k^T\mathcal{L}_{B_k}\mathbf{s}_k}\mathcal{L}_{(\mathcal{L}_{B_k}\mathbf{s}_k)(\mathcal{L}_{B_k}\mathbf{s}_k)^T}.$$

If we call $\mathbf{z}_k$ the vector of the eigenvalues of $\mathcal{L}_{B_k}$, then

$$Ud(\mathbf{z}_{k+1})U^H = Ud(\mathbf{z}_k)U^H + \frac{1}{\mathbf{y}_k^T\mathbf{s}_k}Ud(|U^H\mathbf{y}_k|^2)U^H - \frac{1}{\mathbf{z}_k^T|U^H\mathbf{s}_k|^2}Ud(|d(\mathbf{z}_k)U^H\mathbf{s}_k|^2)U^H,$$

$$d(\mathbf{z}_{k+1}) = d(\mathbf{z}_k) + \frac{1}{\mathbf{y}_k^T\mathbf{s}_k}d(|U^H\mathbf{y}_k|^2) - \frac{1}{\mathbf{z}_k^T|U^H\mathbf{s}_k|^2}d(|d(\mathbf{z}_k)U^H\mathbf{s}_k|^2),$$

$$\mathbf{z}_{k+1} = \mathbf{z}_k + \frac{1}{\mathbf{y}_k^T\mathbf{s}_k}|U^H\mathbf{y}_k|^2 - \frac{1}{\mathbf{z}_k^T|U^H\mathbf{s}_k|^2}d(\mathbf{z}_k)^2|U^H\mathbf{s}_k|^2. \qquad (updEig)$$

From the latter formula it is clear that the vector $\mathbf{z}_{k+1}$ of the eigenvalues of $\mathcal{L}_{B_{k+1}}$ can be computed from $\mathbf{z}_k$ at a cost equal to the cost of the transforms $U^H\mathbf{s}_k$ and $U^H\mathbf{y}_k$. For example, if U is the Hartley matrix as in [Bortoletti et al, 2003], then such updating of the eigenvalues can be implemented in $O(n\log_2 n)$ arithmetic operations. If U is a Householder matrix or the product of two Householder matrices, then the cost reduces to $O(n)$.

Exercise. Find an eigenvalues updating formula of the type (updEig) in case $\mathcal{L}^{(k)}$ changes at each step and is equal to $\mathcal{L}^{(k)} = \{U_k d(\mathbf{z})U_k^H : \mathbf{z} \in \mathbb{C}^n\}$ with $U_k = H(\mathbf{u}_k) := I - \mathbf{u}_k\mathbf{u}_k^H$, $\mathbf{u}_k \in \mathbb{R}^n$, $\|\mathbf{u}_k\|_2 = \sqrt{2}$. Note that the cost of its implementation is $O(n)$.

In [Bortoletti et al, 2003] it is shown that the SECANT $\mathcal{L}^{(k)}$QN method has a very good performance on experiments even in case $\mathcal{L}^{(k)} = \mathcal{L}$ for all k (in that paper $\mathcal{L}$ is chosen equal to the Hartley matrix algebra). SECANT $\mathcal{L}^{(k)}$QN is competitive with the best known minimization algorithms suitable for large scale problems, such as Limited memory BFGS.

But no convergence result has been proved for SECANT $\mathcal{L}^{(k)}$QN.

Instead, in [Di Fiore et al, 2003] it is shown that a subsequence of the sequence of gradients $\nabla f(\mathbf{x}_k)$ generated by the following NON SECANT BFGS-type algorithm

NON SECANT BFGS-type:
$\mathbf{x}_0 \in \mathbb{R}^n$, $B_0 \in \mathbb{R}^{n\times n}$ r.s.p.d.
For $k = 0, 1, 2, \ldots$:
define $\tilde{B}_k$ r.s.p.d. with $\tilde{B}_k \approx B_k$ (in some sense)
$\mathbf{d}_k = -\tilde{B}_k^{-1}\nabla f(\mathbf{x}_k)$
$\mathbf{x}_{k+1} = \mathbf{x}_k + \lambda_k\mathbf{d}_k$, $\ldots$ λ_k $\ldots$
$\mathbf{s}_k = \mathbf{x}_{k+1} - \mathbf{x}_k$, $\mathbf{y}_k = \nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)$ $\ldots$ $\mathbf{s}_k^T\mathbf{y}_k > 0$ $\ldots$
$B_{k+1} = \Phi(\tilde{B}_k, \mathbf{s}_k, \mathbf{y}_k)$ $\ldots$ B_{k+1} r.s.p.d. $\ldots$

converges to the zero vector if the matrices $\tilde{B}_k$ are chosen such that

$$\det(B_k) \le \det(\tilde{B}_k), \quad \mathrm{tr}\,(B_k) \ge \mathrm{tr}\,(\tilde{B}_k). \qquad (convNS)$$

These conditions are satisfied for $\tilde{B}_k = \mathcal{L}_{B_k}^{(k)}$ if $\mathcal{L}^{(k)} = \{U_k d(\mathbf{z})U_k^H : \mathbf{z} \in \mathbb{C}^n\}$ with U_k unitary such that (upddefpos) holds.

Exercise. Prove the latter assertion by using the formula $\mathcal{L}^{(k)}_{B_k} = U_k \,\mathrm{diag}\,((U_k^H B_k U_k)_{ii}) U_k^H$ and Hadamard inequality for r.s.p.d. matrices.

However, the behaviour of the Non Secant BFGS-type algorithm, in particular in the case $\tilde{B}_k = \mathcal{L}^{(k)}_{B_k}$:

NON SECANT $\mathcal{L}^{(k)}$QN:
$\mathbf{x}_0 \in \mathbb{R}^n$, $B_0 \in \mathbb{R}^{n\times n}$ r.s.p.d.
For $k = 0, 1, 2, \ldots$:
$\mathbf{d}_k = -(\mathcal{L}^{(k)}_{B_k})^{-1}\nabla f(\mathbf{x}_k)$
$\mathbf{x}_{k+1} = \mathbf{x}_k + \lambda_k \mathbf{d}_k$, $\ldots$ λ_k $\ldots$
$\mathbf{s}_k = \mathbf{x}_{k+1} - \mathbf{x}_k$, $\mathbf{y}_k = \nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)$ $\ldots$ $\mathbf{s}_k^T \mathbf{y}_k > 0$ $\ldots$
$B_{k+1} = \Phi(\mathcal{L}^{(k)}_{B_k}, \mathbf{s}_k, \mathbf{y}_k)$ $\ldots$ B_{k+1} r.s.p.d. $\ldots$

is far from being good, in numerical experiences.

Recently, in [Cipolla et al, 2014] it has been proposed to choose $\tilde{B}_k$ such that secant and non secant BFGS-type algorithms essentially coincide, so that one achieves simultaneously convergence and good experimental behaviour. More precisely, it is required that the non secant direction $-\tilde{B}_k^{-1}\nabla f(\mathbf{x}_k)$ is a multiple of the secant one $-B_k^{-1}\nabla f(\mathbf{x}_k)$.

So, in [Cipolla et al, 2014] the equality

$$\tilde{B}_k^{-1}\nabla f(\mathbf{x}_k) = \sigma B_k^{-1}\nabla f(\mathbf{x}_k) \qquad (SECequNONSEC)$$

is investigated for $\tilde{B}_k \in \mathcal{L}^{(k)} = \{U_k d(\mathbf{z}) U_k^H : \mathbf{z} \in \mathbb{C}^n\}$, U_k unitary. First (SECequNONSEC) is studied for $\tilde{B}_k = \mathcal{L}^{(k)}_{B_k} = U_k \,\mathrm{diag}\,((U_k^H B_k U_k)_{ii}) U_k^H$. But, in this case, even the question of existence of U_k satisfying (SECequNONSEC) is not clear; thus the question of existence of a *simple* such U_k (U_k =Householder or product of two Householder) is difficult ...

Then the (SECequNONSEC) condition is investigated for $\tilde{B}_k = U_k \,\mathrm{diag}\,((V_k^H B_k V_k)_{ii}) U_k^H$, with U_k and V_k unitary. In this case, if V_k is such that the numbers $\max_i (V_k^H B_k V_k)_{ii}$ and $\min_i (V_k^H B_k V_k)_{ii}$ satisfy a certain condition (such matrix V_k always exist), then we have the existence of U_k for which $\tilde{B}_k$ satisfies (SECequNONSEC), and such U_k is simple since it is the product of two Householder matrices. One can check the condition first for $V_k = U_{k-1}$: if it is satisfied then also V_k turns out to be simple (i.e. the product of two Householder matrices); otherwise, V_k must be chosen different from U_{k-1} and the procedure for finding a right V_k, even if exists, could be expensive ...

For more details, see [Cipolla et al, 2014].

[J. E. Dennis, Jr., R. B. Schnabel, 1983] Numerical Methods for Unconstrained Optimization and Nonlinear Equations Prentice-Hall, Englewood Cliffs, New Jersey, 1983

[C. Di Fiore, P. Zellini, 1995] Matrix decompositions using displacement rank and classes of commutative matrix algebras, Linear Algebra Appl., 229 (1995), pp.49-99

[C.Di Fiore, P. Zellini, 2001] Matrix algebras in optimal preconditioning, Linear Algebra Appl., 335 (2001), pp.1-54

[C. Di Fiore, S. Fanelli, F. Lepore, P. Zellini, 2003] Matrix algebras in quasi-Newton methods for unconstrained minimization, Numerische Mathematik, 94 (2003), pp.479-500

[A. Bortoletti, C. Di Fiore, S. Fanelli, P. Zellini, 2003] A new class of quasi-Newtonian methods for optimal learning in MLP-networks, IEEE Transactions on Neural Networks, 14 (Marzo 2003), pp.263-273

[C. Di Fiore, S. Fanelli, P. Zellini, 2007] Low Complexity secant quasi-Newton minimization algorithms for nonconvex functions, Journal of Computational and Applied Mathematics, 210 (2007), pp.167-174

[C. Di Fiore, F. Tudisco, P. Zellini, 2012] Bernoulli, Ramanujan, Toeplitz and the triangular matrices, submitted, December 2012, April 2013

[S. Cipolla, C. Di Fiore, F. Tudisco, P. Zellini, 2014] Adaptive Matrix Algebras in Unconstrained Minimization, submitted, July 1, 2014

Participants

Ivan, Rufina, Nadia, Anastasia, Elena, Pavel, Dmitry, Aleksei, Rotislav, Nikita, Valeriia, Andrei, Sergey, Ilya
Giulia, Virginia, Elena, Isabella, Arianna, Alessandra, Fabio, Pierpaolo, Michela, Dario, Mauro, Vincent, Antonio,
Gianluca, Stefano, Francesco, Tigran, Tatiana, Daniil.
(Rome-Moscow school of Matrix Methods and Applied Linear Algebra, 2014, Wednesday August 20, S.Petersburg; Saturday August 23-Saturday September 6, Moscow; Sunday September 7-Sunday September 21, Rome)
Roberto Peirone, Sergei Goreinov, Eugene Tyrtyshnikov, Carmine Di Fiore, Dario Fasino, Olga Lebedeva, Mikhail Botchev, Dmitry Vetrov, Massimo Picardello, Daniele Bertaccini, Salvatore Filippone, Ivan Oseledets

APPENDIX (the discrete Fast Fourier Transform)

The Fourier matrix, circulants, and fast discrete transforms

Consider the following $n \times n$ matrix

$$\Pi_1 = \begin{bmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 1 & 1 \\ 1 & & & & 0 \end{bmatrix}.$$

Let $\omega \in \mathbb{C}$. Note that

$$\Pi_1 \begin{bmatrix} 1 \\ 1 \\ . \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ . \\ 1 \end{bmatrix} = 1 \begin{bmatrix} 1 \\ 1 \\ . \\ 1 \end{bmatrix}, \ \Pi_1 \begin{bmatrix} 1 \\ \omega \\ . \\ \omega^{n-1} \end{bmatrix} = \begin{bmatrix} \omega \\ \omega^2 \\ . \\ \omega^{n-1} \\ 1 \end{bmatrix} = \omega \begin{bmatrix} 1 \\ \omega \\ . \\ \omega^{n-1} \end{bmatrix},$$

where the latter identity holds if $\omega^n = 1$. More in general, if $\omega^n = 1$, we have the following vectorial identities

$$\Pi_1 \begin{bmatrix} 1 \\ \omega^j \\ . \\ \omega^{(n-1)j} \end{bmatrix} = \begin{bmatrix} \omega^j \\ . \\ \omega^{(n-1)j} \\ 1 \end{bmatrix} = \omega^j \begin{bmatrix} 1 \\ \omega^j \\ . \\ \omega^{(n-1)j} \end{bmatrix}, \ j = 0, 1, \ldots, n-1,$$

or, equivalently, the following matrix identity

$$\Pi_1 W = W D_{1\omega^{n-1}},$$

$$D_{1\omega^{n-1}} = \begin{bmatrix} 1 & & & \\ & \omega & & \\ & & . & \\ & & & \omega^{n-1} \end{bmatrix}, \ W = \begin{bmatrix} 1 & 1 & \cdot & 1 & \cdot & 1 \\ 1 & \omega & & \omega^j & & \omega^{n-1} \\ . & . & & . & & . \\ 1 & \omega^{n-1} & . & \omega^{(n-1)j} & . & \omega^{(n-1)(n-1)} \end{bmatrix}.$$

Proposition. If $\omega^n = 1$ and if $\omega^j \neq 1$ for $0 < j < n$, then $W^*W = nI$.

proof: since $|\omega| = 1$, $\overline{\omega} = \omega^{-1}$, we have

$$\begin{aligned} [W^*W]_{ij} &= [\overline{W}W]_{ij} = \textstyle\sum_{k=1}^n [\overline{W}]_{ik}[W]_{kj} = \sum_{k=1}^n \overline{\omega}^{(i-1)(k-1)} \omega^{(k-1)(j-1)} \\ &= \textstyle\sum_{k=1}^n \omega^{(k-1)(j-i)} = \sum_{k=1}^n (\omega^{j-i})^{k-1}. \end{aligned}$$

Thus $[W^*W]_{ij} = n$ if $i = j$, and $[W^*W]_{ij} = \frac{1-(\omega^{j-i})^n}{1-\omega^{j-i}} = 0$ if $i \neq j$ (note that the assumption $\omega^j \neq 1$ for $0 < j < n$ is essential in order to make $1 - \omega^{j-i} \neq 0$).

By the result of the above Proposition, we can say that the following (symmetric) Fourier matrix

$$F = \frac{1}{\sqrt{n}} W$$

is unitary, i.e. $F^*F = I$.

Exercise. Prove that $F^2 = J\Pi_1$ where J is the permutation matrix $J\mathbf{e}_k = \mathbf{e}_{n+1-k}$, $k = 1, \ldots, n$ (J is usually called anti-identity).

The matrix identity satisfied by Π_1 and W can be of course rewritten in terms of F, $\Pi_1 F = F D_{1\omega^{n-1}}$, thus we obtain the equality

$$\Pi_1 = F D_{1\omega^{n-1}} F^*$$

which states that the Fourier matrix diagonalizes the matrix Π_1, or, more precisely, that *the columns of the Fourier matrix form a system of n unitarily orthonormal eigenvectors for the matrix* Π_1 *with corresponding eigenvalues* $1, \omega, \ldots, \omega^{n-1}$.

But if F diagonalizes Π_1, then it diagonalizes all polynomials in Π_1:

$$\begin{aligned}
\Pi_1^{k-1} &= F D_{1\omega^{n-1}}^{k-1} F^*, \\
\textstyle\sum_{k=1}^n a_k \Pi_1^{k-1} &= F \textstyle\sum_{k=1}^n a_k D_{1\omega^{n-1}}^{k-1} F^* \\
&= F \begin{bmatrix} \sum_{k=1}^n a_k & & & \\ & \sum_{k=1}^n a_k \omega^{k-1} & & \\ & & \ddots & \\ & & & \sum_{k=1}^n a_k \omega^{(n-1)(k-1)} \end{bmatrix} F^* \\
&= F d(W\mathbf{a}) F^* = \sqrt{n} F d(F\mathbf{a}) F^*
\end{aligned}$$

where by $d(\mathbf{z})$ we mean the diagonal matrix whose diagonal entries are $z_1, z_2, \ldots, z_n$.

Let us investigate the matrices Π_1^{k-1}, $k = 1, \ldots, n$, and the matrix $\sum_{k=1}^n a_k \Pi_1^{k-1}$ in the case $n = 4$:

$$\Pi_1^0 = I,\ \Pi_1^1 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix},\ \Pi_1^2 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix},\ \Pi_1^3 = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix},$$

$$\Pi_1^4 = \Pi_1^3 \Pi_1 = \Pi_1^T \Pi_1 = I = \Pi_1^0,$$

$$\sum_{k=1}^4 a_k \Pi_1^{k-1} = \begin{bmatrix} a_1 & a_2 & a_3 & a_4 \\ a_4 & a_1 & a_2 & a_3 \\ a_3 & a_4 & a_1 & a_2 \\ a_2 & a_3 & a_4 & a_1 \end{bmatrix} = \sqrt{4} F d(F\mathbf{a}) F^*,\ F = \frac{1}{\sqrt{4}} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & \omega & \omega^2 & \omega^3 \\ 1 & \omega^2 & \omega^4 & \omega^6 \\ 1 & \omega^3 & \omega^6 & \omega^9 \end{bmatrix},$$

$$\omega^4 = 1,\ \omega^j \neq 1,\ 0 < j < 4\ (\omega = e^{\pm \mathbf{i} 2\pi/4}).$$

Note that, for n generic, we have the identities $\mathbf{e}_1^T \Pi_1^{k-1} = \mathbf{e}_k^T$, $k = 1, \ldots, n$, and $\Pi_1^n = I$ (prove them!). So, the set $C = \{p(\Pi_1)\}$ of all polynomials in Π_1 is spanned by the matrices $J_k = \Pi_1^{k-1}$; the particular polynomial $\sum_{k=1}^n a_k J_k$ is simply denoted by $C(\mathbf{a})$. Note that $C(\mathbf{a})$ is the matrix of C with first row $\mathbf{a}^T$:

$$C(\mathbf{a}) = \sum_{k=1}^n a_k J_k = \begin{bmatrix} a_1 & a_2 & \cdot & a_{n-1} & a_n \\ a_n & a_1 & \cdot & & a_{n-1} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ a_3 & & \cdot & \cdot & a_2 \\ a_2 & a_3 & \cdot & a_n & a_1 \end{bmatrix} = F d(F^T \mathbf{a}) d(F^T \mathbf{e}_1)^{-1} F^{-1}.$$

C is known as the space of circulant matrices.

Exercise. (i) Repeat all, starting from the $n \times n$ matrix

$$\Pi_{-1} = \begin{bmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ -1 & & & & 0 \end{bmatrix}$$

and arriving to the (-1)-circulant matrix whose first row is $\mathbf{a}^T$, $\mathbf{a} \in \mathbb{C}^n$:

$$C_{-1}(\mathbf{a}) = \begin{bmatrix} a_1 & a_2 & \cdot & a_{n-1} & a_n \\ -a_n & a_1 & \cdot & & a_{n-1} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ -a_3 & & \cdot & \cdot & a_2 \\ -a_2 & -a_3 & \cdot & -a_n & a_1 \end{bmatrix}.$$

(ii) Let T be a Toeplitz $n \times n$ matrix, i.e. $T = (t_{i-j})_{i,j=1}^n$, for some $t_k \in \mathbb{C}$. Show that T can be written as the sum of a circulant and of a (-1)-circulant, that is, $T = C(\mathbf{a}) + C_{-1}(\mathbf{b})$, $\mathbf{a}, \mathbf{b} \in \mathbb{C}^n$.

Why circulant matrices can be interesting in the applications of linear algebra? The main reason is in the fact that the matrix-vector product $C(\mathbf{a})\mathbf{z}$ can be computed in at most $O(n \log_2 n)$ arithmetic operations (whereas, usually, a matrix-vector product requires n^2 multiplications).

Proposition FFT. Given $\mathbf{z} \in \mathbb{C}^n$, the complexity of the matrix-vector product $F\mathbf{z}$ is at most $O(n \log_2 n)$. Such operation is called discrete Fourier transform (DFT) of $\mathbf{z}$. As a consequence, the matrix-vector product $C(\mathbf{a})\mathbf{z}$ is computable by two DFTs (after the preprocessing DFT $F\mathbf{a}$).

proof: since $\omega^{(i-1)(k-1)}$ is the (i,k) entry of W and z_k is the k entry of $\mathbf{z} \in \mathbb{C}^n$, we have

$$\begin{array}{rcl} (W\mathbf{z})_i & = & \sum_{k=1}^n \omega^{(i-1)(k-1)} z_k = \sum_{j=1}^{n/2} \omega^{(i-1)(2j-2)} z_{2j-1} + \sum_{j=1}^{n/2} \omega^{(i-1)(2j-1)} z_{2j} \\ & = & \sum_{j=1}^{n/2} (\omega^2)^{(i-1)(j-1)} z_{2j-1} + \sum_{j=1}^{n/2} \omega^{(i-1)(2(j-1)+1)} z_{2j} \\ & = & \sum_{j=1}^{n/2} (\omega^2)^{(i-1)(j-1)} z_{2j-1} + \omega^{i-1} \sum_{j=1}^{n/2} (\omega^2)^{(i-1)(j-1)} z_{2j}. \end{array}$$

Note that ω is in fact a function of n, i.e. the right notation for ω should be ω_n. Then $\omega^2 = \omega_n^2$ is such that $(\omega_n^2)^{n/2} = 1$ and $(\omega_n^2)^i \neq 1$ $0 < i < n/2$; in other words $\omega_n^2 = \omega_{n/2}$. So, we have the identities

$$(W_n\mathbf{z})_i = \sum_{j=1}^{n/2} \omega_{n/2}^{(i-1)(j-1)} z_{2j-1} + \omega_n^{i-1} \sum_{j=1}^{n/2} \omega_{n/2}^{(i-1)(j-1)} z_{2j}, \ i = 1, 2, \ldots, n. \qquad (?)$$

It follows that, for $i = 1, \ldots, \frac{n}{2}$,

$$(W_n\mathbf{z})_i = (W_{n/2} \begin{bmatrix} z_1 \\ z_3 \\ \cdot \\ z_{n-1} \end{bmatrix})_i + \omega_n^{i-1} (W_{n/2} \begin{bmatrix} z_2 \\ z_4 \\ \cdot \\ z_n \end{bmatrix})_i.$$

Moreover, by setting $i = \frac{n}{2} + k$, $k = 1, \ldots, \frac{n}{2}$, in (?), we obtain

$$\begin{array}{rcl} (W_n\mathbf{z})_{\frac{n}{2}+k} & = & \sum_{j=1}^{n/2} \omega_{n/2}^{\frac{n}{2}(j-1)} \omega_{n/2}^{(k-1)(j-1)} z_{2j-1} + \omega_n^{\frac{n}{2}} \omega_n^{k-1} \sum_{j=1}^{n/2} \omega_{n/2}^{\frac{n}{2}(j-1)} \omega_{n/2}^{(k-1)(j-1)} z_{2j} \\ & = & \sum_{j=1}^{n/2} \omega_{n/2}^{(k-1)(j-1)} z_{2j-1} - \omega_n^{k-1} \sum_{j=1}^{n/2} \omega_{n/2}^{(k-1)(j-1)} z_{2j} \\ & = & (W_{n/2} \begin{bmatrix} z_1 \\ z_3 \\ \cdot \\ z_{n-1} \end{bmatrix})_k - \omega_n^{k-1} (W_{n/2} \begin{bmatrix} z_2 \\ z_4 \\ \cdot \\ z_n \end{bmatrix})_k, \ k = 1, \ldots, \frac{n}{2}. \end{array}$$

($\omega_n^{\frac{n}{2}} = -1$; think $\omega = e^{\pm \mathbf{i} 2\pi/n}$). Thus

$$W_n \mathbf{z} = \begin{bmatrix} I & D_{1\omega_n^{\frac{n}{2}-1}} \\ I & -D_{1\omega_n^{\frac{n}{2}-1}} \end{bmatrix} \begin{bmatrix} W_{n/2} & 0 \\ 0 & W_{n/2} \end{bmatrix} Q\mathbf{z},$$

$$D_{1\omega_n^{\frac{n}{2}-1}} = \begin{bmatrix} 1 & & & \\ & \omega_n & & \\ & & . & \\ & & & \omega_n^{\frac{n}{2}-1} \end{bmatrix}, Q = \begin{bmatrix} 1 & & & & & & & \\ 0 & 0 & 1 & & & & & \\ & & . & . & & & & \\ & & & & & & 1 & 0 \\ 0 & 1 & & & & & & \\ 0 & 0 & 0 & 1 & & & & \\ & & & . & . & & & \\ & & & & & & 0 & 1 \end{bmatrix}. \qquad (??)$$

If c_n denotes the complexity of the matrix-vector product $F_n\mathbf{z}$, then, by the previous formula,

$$c_n \leq 2c_{n/2} + rn, \ r \text{ constant}.$$

But this implies $c_n = O(n \log_2 n)$. The proof of the last assertion is left to the reader.

Of course, any time a $n \times n$ matrix U, well defined for all n, satisfies for n even an identity of the type

$$U_n = \Big[\text{ sparse matrix } \Big] \begin{bmatrix} U_{n/2} & 0 \\ 0 & U_{n/2} \end{bmatrix} \Big[\text{ permutation matrix } \Big],$$

the matrix-vector product $U_n\mathbf{z}$ can be computed in at most $O(n \log_2 n)$ arithmetic operations. The above identity is verified for at least 10 matrices U, the Fourier transform and its (-1)-version, and the eight Hartley-type transforms. Note, however, that there are also other 16 discrete transforms of complexity $O(n \log_2 n)$, sine-type and the cosine-type transforms. See [],[].

Exercise. Let $U_n = U$ be the Hartley matrix, i.e. $U = \frac{1}{\sqrt{n}}(\cos\frac{2ij\pi}{n} + \sin\frac{2ij\pi}{n})_{i,j=0}^{n-1}$. Then $U^T = U = U^{-1}$, and

$$U_{2n} = \frac{1}{\sqrt{2}} \begin{bmatrix} I & R \\ I & -R \end{bmatrix} \begin{bmatrix} U_n & O \\ O & U_n \end{bmatrix} Q_{n,2},$$

$R = d(\mathbf{c}) + d(\mathbf{s})J\Pi_1$, $c_h = \cos\varphi_h$, $s_h = \sin\varphi_h$, $\varphi_h = h\pi/n$, $h = 0, 1, \ldots, n-1$, $Q_{n,2}\mathbf{z} = [z_0\ z_2\ \cdots\ z_{2n-2}\ z_1\ z_3\ \cdots\ z_{2n-1}]^T$.

Exercise G. Prove that the $n \times n$ matrix $G = G_n$ defined by

$$G_{ij} = \frac{1}{\sqrt{n}}(\cos\frac{(2i+1)(2j+1)\pi}{2n} + \sin\frac{(2i+1)(2j+1)\pi}{2n}), \ i, j = 0, \ldots, n-1,$$

is symmetric, persymmetric, real, unitary, and satisfies the identity:

$$G_n = \frac{1}{\sqrt{2}} \begin{bmatrix} R_+ & R_- \\ -R_-J & R_+J \end{bmatrix} \begin{bmatrix} G_{n/2} & 0 \\ 0 & G_{n/2} \end{bmatrix} Q, \ R_\pm = D_c \pm D_s J,$$

(J $\frac{n}{2} \times \frac{n}{2}$ anti-identity) for some suitable $\frac{n}{2} \times \frac{n}{2}$ diagonal matrices D_c, D_s. Prove, moreover, that each row of G_n has at least a zero entry when $n = 2+4s$. This is like to say, that a matrix in the space $\{Gd(\mathbf{z})G : \mathbf{z} \in \mathbb{C}^n\}$ is not uniquely defined from any of its rows.