Matrices : Theory & Applications
Solutions to the exercises

Denis Serre
École Normale Supérieure
de Lyon

1 Elementary theory

1. Obviously, $G^{\mathbb{C}}$ is an $\mathbb{R}$-vector space. Proving that it is a $\mathbb{C}$-vector space amounts to check that $(\lambda\mu)x = \lambda(\mu x)$ for every $\lambda, \mu \in \mathbb{C}$ and $x \in G^{\mathbb{C}}$. This must be done by direct calculation.

2. The rank of M is the dimension of the space $R(M)$ spanned by the columns

$$M^{(1)}, \dots, M^{(m)}$$

of M in K^n. Each column of MM', being a linear combination (check that!) of the $M^{(j)}$'s, belong to F. Therefore the space spanned by the columns of MM' is a subspace of F and must not have a greater dimension. This proves $\mathrm{rk}(MM') \leq \mathrm{rk}\, M$. Applying that to M'^T and M^T instead, gives

$$\mathrm{rk}(MM') = \mathrm{rk}(MM')^T = \mathrm{rk}(M'^T M^T) \leq \mathrm{rk}\, M'^T = \mathrm{rk}\, M'.$$

3. (a) Let $r := \mathrm{rk}\, B$. Let G be a subspace of K^m such that $K^m = R(B) \oplus G$. Then $\dim G = m - r$. We have $R(A) = A(R(B) \oplus G) = A(R(B)) + A(G) = R(AB) + A(G)$. Since $\dim A(G) \leq \dim G$, there comes

$$\mathrm{rk}\, A = \dim(R(AB) + A(G)) \leq \dim R(AB) + \dim G \leq \mathrm{rk}(AB) + m - \mathrm{rk}\, B.$$

(b) Let assume first that B be onto. Then apply the first question to A and BC ; we obtain $\mathrm{rk}\, A + \mathrm{rk}(BC) \leq m + \mathrm{rk}(ABC)$. This is the desired inequality, since $\mathrm{rk}\, B = m$ and $\mathrm{rk}(AB) = \dim R(AB) = \dim R(A) = \mathrm{rk}\, A$.

We now consider the general case. The previous analysis is obviously valid for homomorphisms. We apply it to the triplet (A', B, C), where A' is the restriction of A to $R(B)$. We obtain

$$\mathrm{rk}(A'B) + \mathrm{rk}(BC) \leq \mathrm{rk}\, B + \mathrm{rk}(A'BC).$$

This is the desired result, since $A'B$ acts exactly the same as AB.

4. (a) The bilinearity is clear. That the range spans $M_{nn'\times mm'}$ comes from the general fact that $M_{p\times q}$ is spanned by all the elementary matrices $E^{k,l}$ defined by

$$E^{k,l}_{i,j} = \delta^k_i \delta^l_j.$$

Last, the tensor product of elementary matrices is an elementary matrix, and one obtains all elementary matrices in $M_{nn'\times mm'}$ as such tensor products. Hence the range of the tensor product spans $M_{nn'\times mm'}$. However, the map is not onto in general, since it requires that all $n' \times m'$ blocks be colinear to a single one C. One proves that it is onto if and only if either $n = m = 1$ or $n' = m' = 1$.

(b) One finds $(BD) \otimes (CE)$.

(c) The uniqueness of L follows from the fact that tensor products span $M_{nn'\times mm'}$. The existence is given by the formula

$$L\left(\sum_{i,j,k,l} a^{kl}_{ij} E^{ij} \otimes E^{kl}\right) := \sum_{i,j,k,l} a^{kl}_{ij} \phi\left(E^{ij}, E^{kl}\right).$$

This defines a linear form. There remains to check that $L(B \otimes C) = \phi(B, C)$, which is done by expanding B and C on the bases of elementary matrices and using the bilinearity of ϕ.

Nota. in this question, one may replace the target space of ϕ by any K-vector space.

2 Square matrices

1. Let A, B be upper triangular matrices : $a_{ij} = b_{ij} = 0$ whenever $i > j$. Then,

$$(AB)_{ij} = \sum_k a_{ik} b_{kj} = \sum_{i \le k \le j} a_{ik} b_{kj},$$

and the last sum is void, hence null, if $i > j$.

2. Let us do it for the case of an upper triangular matrix M. For a product $m_{1\sigma(1)} \cdots m_{n\sigma(n)}$ to be non zero, one needs $j \le \sigma(j)$ for every index j. Summing up, there comes

$$\frac{n(n+1)}{2} = \sum_j j \le \sum_j \sigma(j) = \sum_k k = \frac{n(n+1)}{2}$$

and hence all the inequalities are equalities. Finally, there remains at most one non zero term in the determinant, namely the one corresponding to the identity. The product is that of the diagonal entries and the signature is $+1$.

3.

$$M = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad N = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

4. Let A be orthogonal and upper triangular. Then $A^T = A^{-1}$ must be upper triangular (see exercise 8). Therefore A is lower triangular, hence diagonal. Then $I_n = A^T A = A^2$ shows that every diagonal entry is either $+1$ or -1.

5. (a) Let x be in K^n. Define $y = Px$ and $z = x - y$. Then $Pz = y - P^2x = y - y = 0$ gives $z \in F$. Also, $(I_n - P)y = y - P^2x = 0$. Hence $K^n = E + F$. Last, assume that $x \in E \cap F$. Then $y = 0$ and $z = x$. Since $z \in F$, there holds $z = Pz = Px = 0$. Finally, $x = 0$.

   (b) From $(P - Q)^2 = P + Q - PQ - QP$, we find

$$P(P - Q)^2 = P - PQP = (P - Q)^2 P.$$

   By symmetry, $(P - Q)^2$ commutes with Q too. Also, $(I_n - P - Q)^2 = I_n - P - Q - PQ - QP$ gives the identity.

6. Let A be $p \times q$ and D be $r \times s$. For AD to be meaningful, one needs $q = r$. Alike, for BC to make sense, one needs $s = r$. At last, M being square, one has $p + r = q = s$, hence $p = q = r = s$. Now, if $p \geq 2$ the following is a counter-example to the formula :

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \quad C = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad D = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

   Schur's formula is $\det M = \det A \det(D - BA^{-1}C)$. When $p = q = r = s$ and if A and B commute, then it implies $\det M = \det(AD - ABA^{-1}C) = \det(AD - BC)$.

7. Because $AC = CA$, there holds

$$\begin{pmatrix} I_m & 0_m \\ -C & A \end{pmatrix} M = \begin{pmatrix} A & B \\ 0_m & AD - CB \end{pmatrix},$$

   form which we infer $\det A \det M = \det A \det(AD - CB)$. This is the desired equality if $\det A \neq 0$.

   In the general case, we define a matrix $M(X)$ by replacing A by $A - XI_m$. Let us denote $Q(X) = \det M(X) - \det((A - XI_m)D - CB)$. From the previous case, we know that every scalar $z \in \bar{K}$ is either a root of the characteristic polynomial P_A, or a root of Q. Since $\bar{K}$ is infinite (it would be enough to consider z in a large enough extension of K) and P_A has finitely many roots, Q must have infinitely many root and therefore must be the null polynomial.

8. Prove it either directly or by means of Cayley-Hamilton's Theorem.

9. From exercise 2, the characteristic polynomial of a triangular matrix M equals the product of the monomials $X - m_{jj}$. Its roots are the diagonal entries m_{jj}. The algebraic multiplicity of an eigenvalue λ equals the number of indices j for which $m_{jj} = \lambda$.

10. Let assume that there exists a null matrix extracted from A, of size $k \times l$ with $k + l > n$, corresponding to the row indices in K and column indices in L. For every permutation σ, the sum of cardinals of $\sigma(K)$ and L is $k + l$, and thus these sets must have a common element $j = \sigma(i)$. Form $i \in K$ and $\sigma(i) \in L$, we obtain $a_{i\sigma(i)} = 0$. Hence every diagonal contains a null element.

    We prove the converse by induction on n. Let the result be true at order $n - 1$ and let A be $n \times n$ and such that every diagonal contains a zero element. If the first row is identically zero, we already have an $1 \times n$ null block and $1 + n > n$. Otherwise, there exists a non zero element a_{1j} and we consider the $(n-1) \times (n-1)$ matrix obtained by deleting the first row and the j-th column. By induction, it contains an $r \times s$ null block with $r + s \geq n$. If $r + s > n$, we are done. Otherwise, $r + s = n$ and we may suppose, up to a permutation of rows and columns, that this block is the upper left one. Hence, A reads as

$$\begin{pmatrix} 0_{k\times l} & B \\ C & D \end{pmatrix},$$

    where B, C are square matrices. If each one has a diagonal with non zero elements, then A has one too. Hence the induction hypothesis can be applied to one of both matrices B, C. Without loss of generality, we may assume that B (of size $k \times k$) contains a null block of size $a \times b$, with $a + b > k$, corresponding to row indices in a set I and columns indices in a set J. Then A contains a null block of size $a \times (b + l)$, corresponding to row indices in I and columns indices in $J \cup \{1, \ldots, l\}$. At last, $a + b + l > k + l = n$.

11. By inspection, $\boldsymbol{GL}_2(\mathbb{Z}/2\mathbb{Z})$ has 6 elements. Since matrices

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

    do not commute, this group is not abelian. At last, every non abelian group with six elements is isomorphic to $\boldsymbol{S}_3$.

12. (a) One easily find that $\pi^k = \mathrm{circ}(\ldots, 0, 1, 0, \ldots)$, where the entry 1 is in k position, *modulo* n. In particular, $\pi^n = I_n$. Therefore, the algebra $\mathbb{C}[\pi]$ equals the subspace spanned by $I_n, \pi, \ldots, \pi^{n-1}$, which is $\mathcal{C}_n$. Since the kernel of $X \mapsto \pi$, from $\mathbb{C}[X]$ to $\mathbb{C}[\pi]$, is the ideal spanned by $X^n - 1$, the characteristic polynomial of π, $\mathcal{C}_n$ is isomorphic to the quotient ring $\mathbb{C}[X]/(X^n - 1)$.

    (b) Clearly, the hermitian adjoint of $\mathrm{circ}(a_0, \ldots, a_{n-1})$ is $\mathrm{circ}(\overline{a_0}, \overline{a_{n-1}}, \ldots, \overline{a_1})$. We already know that $P(C)$ is circulant, since $\mathcal{C}_n$ is an algebra. A matrix is circulant if and only if it commutes with π (check that!). Hence C^{-1}, when defined, is circulant.

    (c) Just because they are polynomials in π, and π has distinct eigenvalues (the n-th roots of unity) and thus is diagonalizable.

    (d) Same as in the previous question, because there are n distinct n-th roots of unity under this assumption.

(e) We shall prove that now π is not diagonalizable. Let p denote the characteristic of K, then the algebraic mulitplicities of the eigenvalues of π are greater than 1, since the derivative of the characteristic polynomial $X^n - 1$ vanishes identically. However, every eigenvalue ω is geometrically simple, since the eigenspace is spanned by $(\omega^{n-1}, \ldots, \omega, 1)^T$. Other argument : the rank of $\pi - \omega I_n$ is $n-1$ at least since this matrix contains a invertible block I_{n-1}. Hence its kernel has dimension one at most.

13. Let A be alternate and $D := \mathrm{diag}(\lambda, 1, \ldots, 1)$. Then

$$\lambda \mathrm{Pf}(A) = (\det D)\mathrm{Pf}(A) = \mathrm{Pf}(D^T A D).$$

Since $D^T AD$ is obtained my multiplying the entries in the first row (and therefore those in the first column, to keep the alternate shape) by λ, this shows that Pf is homogeneous of degree 1 in the entries of the first row. Since Pf is a polynomial, this homogeneity means linearity. Using permutations, or an other diagonal matrix D, we conclude that Pf is linear with respect to each row and each column.

Let $\mathrm{Pf} = QR$ be a factorization of the Pfaffian. Obviously, Q and R must be homogeneous. Also, considering the i-th row dependence, one of both, say Q, must be linear and the other must not depend on it. If R is not constant, it must contain some variable X_{jk}, and therefore be linear in the k-th column. Thus Q does not depend on the k-th column. Hence, neither Q nor R depend on the entry X_{ik}, which contradicts the fact that Pf actually depends on every entry.

14. Let a be an eigenvalue of A and $E := \ker(A - aI_n)$. For every $M \in S$, there holds $M(A - aI_n) = (A - aI_n)M$, and thus $M(E) \subset E$. By assumption, one has $E = k^n$, hence $A = aI_n$.

15. (a) Let assume that such sequences exist for every pair (j, k). Let $I \cup J$ be a non-trivial partition of the set of indices. Let $i \in I$ and $j \in J$ be given and let $i = l_1, \ldots, l_r = j$ be such a sequence. Let p be the smallest integer such that $l_{p+1} \notin I$. Then $l_p \in I$, $l_{p+1} \in J$ and $a_{l_p l_{p+1}} \neq 0$. Hence the block $A_{I \times J}$ is non zero, and A is irreducible.
On the contrary, let us assume that there exists a pair (i, j) such that there does not exist a such sequence. Then let I be the set of attainable indices with such a sequence starting from i. Finally, let J be the complement of I. By assumption, $I \cup J$ is a non-trivial partition of the indices, but $A_{I \times J} = 0$. Hence A is reducible.

(b) If $i < j$, one uses the (non zero) entries $a_{k,k+1}$ to make a sequence from i to j. If $i > j$, one uses the entries $a_{k+1,k}$ instead. If $i = j$ and $n \geq 2$, we use two sequences, one from i to some other index k and one from k to i. If $n = 1$, there is nothing to prove.

16. (a) Just remark that $q \in I_x$.

(b) Let Q be the lcm of the r_j's. Since r_j divides q, Q divides q as well. On an other hand, every vector x writes as $\sum_j x_j \vec{e^j}$, and thus $Q(A)x = 0$, since $Q(A)\vec{e^j} = 0$. Hence q divides Q, that is $Q = q$.

(c) We notice that the set of monic divisors of q is finite. If x_m converges to x, then by continuity, p_x will divide every cluster point of the sequence p_{x_m}, that is every polynomial which appears infinitely many times in that sequence. There follows that p_x actually divides p_y for every y in a small neighbourhood. Hence V_p is open.

(d) If p_x is of maximal degree, then let $p := p_x$. Then V_p contains x. If $y \in V_p$, then p divides p_y. The maximality of p implies that actually $p_y = p$. Since V_p is a non-void open set, one can choose a basis of k^n, contained in V_p. By linearity, there follows that $p(A)$ vanishes on every vector, and therefore q divides p. Thus $q = p$.

Conclusion. There exists a vector x such that $p(A)x = 0$ implies that the minimal polynomial divides p.

17. (a) One has

$$\begin{pmatrix} I_n & 0 \\ -B & XI_m \end{pmatrix} M = \begin{pmatrix} I_n & A \\ 0 & X^2 I_m - BA \end{pmatrix}.$$

Taking the determinant in the identity gives the desired identity.

(b) On an other hand, one also has

$$\begin{pmatrix} XI_n & -A \\ 0 & I_m \end{pmatrix} M = \begin{pmatrix} X^2 I_n - AB & 0 \\ B & XI_m \end{pmatrix}.$$

Taking the determinant gives $X^n \det M = X^m \det(X^2 - AB)$. Using both identities, there comes $X^{2n} P_{BA}(X^2) = X^{2m} P_{AB}(X^2)$. In other words, $X^n P_{BA}(X) = X^m P_{AB}(X)$.

(c) We may assume that $m \leq n$. Then $P_{AB} = X^{n-m} P_{BA}$. If $n \neq m$, then the spectrum of AB equals the spectrum of BA, augmented of 0. If $n = m$, both spectra are equal. The multiplicity of a non-zero eigenvalue is always the same for AB and BA. The multiplicities of the null eigenvalue differ by $n - m$.

18. (a) Let W, Z be two vectors such that $Z^T W = 1$, and define $\alpha := \theta(WZ^T)$. Then, for every X, Y, there holds

$$\theta(XY^T) = \theta(XZ^T WY^T) = \theta(WY^T XZ^T) = (Y^T X)\theta(WZ^T) = \alpha Y^T X.$$

(b) Decompose any matrix A as $\sum_{ij} a_{ij} \vec{e_i} \vec{e_j}^T$, where the $\vec{e_i}$'s are the canonical basis vectors. Then apply the previous question.

19. (a) We consider Δ, a polynomial of n variables, as a polynomial in the single variable X_i with coefficients in $A_{n-1} = K[\ldots, X_{i-1}, X_{i+1}, \ldots]$. We may specialize X_i to any value in a field containing K. For instance, X_i may be assigned the value $X_j \in A_{n-1}$ for some $j \neq i$. Then Δ vanishes since the matrix M has two identical columns. Hence $X_i - X_j$ divides Δ in $K[k]$, where k is the fraction field of A_{n-1}. However, since $X_i - X_j$ is monic, the quotient belongs to $K[A_{n-1}] = A_n$.

(b) Since the polynomials $X_i - X_j$ $(1 \leq i < j \leq n)$ are pairwise coprime, their product must also divide Δ. Since both this product and Δ have same degree $n(n-1)/2$, the quotient is a scalar.

(c) By induction on n, the coefficient of this monomial is 1 in both the product and Δ. Hence $a = 1$, that is

$$\Delta = \prod_{i<j}(X_j - X_i).$$

(d) If two distinct exponents p_r are equal, then the determinant vanishes since the matrix has two identical rows. Otherwise, up to a permutation of rows, we may assume that $p_1 < p_2 < \cdots < p_n$. One easily factorizes

$$\left(\prod_j X_j\right)^{p_1} \prod_{i<j}(X_j - X_i).$$

The quotient of Δ by this polynomial is a symmetric polynomial, whose determination is beyond the scope of our book.

20. From the previous exercise, the determinant equals $\prod_{i<j}(a_j - a_i)$. Therefore it vanishes if and only if at least one of the factors does.

21. (a) Just use Cauchy-Binet's formula.

(b) Since the principal minors

$$A\begin{pmatrix} 1 & \cdots & p \\ 1 & \cdots & p \end{pmatrix}$$

are positive, therefore non-zero, A admits a unique LU factorization. Apply first the Cauchy-Binet's formula to the product LU, using the p first lines. Then the sum reduces to only one term because all minors of L are zero but the principal one. Therefore,

$$U\begin{pmatrix} 1 & \cdots & p \\ j_1 & \cdots & j_p \end{pmatrix} = A\begin{pmatrix} 1 & \cdots & p \\ j_1 & \cdots & j_p \end{pmatrix}$$

is positive for every increasing sequence $j_1, \ldots, j_p$ and every length $p \in \{1, \ldots, n\}$. Likewise,

$$L\begin{pmatrix} i_1 & \cdots & i_p \\ 1 & \cdots & p \end{pmatrix}$$

is positive for every increasing sequence $i_1, \ldots, i_p$ and length p.

The end of the proof is more delicate and does not involve A any more. It uses the general fact that, for a triangular matrix, the (strict) positivity of non-trivial minors follows from that of minors of the form above. This result is a consequence of the algebraic relations between minors within a single matrix. We refer to S. Karlin, *Total Positivity*, Stanford University Press (1968).

(c) We have to show that any determinant of the form

$$F(x_1, \dots, x_r) := \det((x_i^{p_j}))_{1 \le i,j \le r}$$

is positive whenever $x_1 < \cdots < x_r$ and $p_1 < \cdots < p_r$. Let us suppose that $F(x) = 0$. Then an eigenvector of the corresponding matrix gives a non-zero polynomial $Q = \sum_l v_l X^{p_l}$, such that $Q(x_1) = \cdots = Q(x_r) = 0$. Now, Descartes' rule of signs (G. Polya, G. Szego, exercize V.36 in *Problems and Theorems in Analysis*, Grundl. der mat. Wiss., band 216, Springer-Verlag) ensures that the number of positive roots of Q does not exceed the number of sign changes in the list of coefficients, here $v_1, \dots, v_r$. Hence, there should not be r positive roots of Q, a contradiction. Thus F does not vanish under the constraints on the vectors x and p. We conclude by induction : given an admissible vector p, F keeps a constant sign ϵ on admissible vectors x. Letting $x_1 \to 0^+$, the limit of $x_1^{-p_1} F(x)$ is the same determinant of order $r-1$, with admissible vectors. From the induction hypothesis, it is positive. Hence F is positive for small values of x_1 and ϵ must be $+1$.

22. Compute the product, then use that $\det A = \det A^T$ and the formula for the determinant of the Vandermonde matrix.

23. (a) That Δ is symmetric is obvious, because the transposition $i \leftrightarrow j$ exchanges factors $(X_i - X_k)^2$ and $(X_j - X_k)^2$ $(k \neq i, j)$ while it leaves unchanged all other factors. Obviously, Δ has coefficients in $\mathbb{Z}$. The fact that a symmetric polynomial expresses as a polynomial in terms of $\sigma_1, \dots, \sigma_n$ is the fundamental Theorem of symmetric polynomials. This theorem also tells that one may keep the same ring of scalars, here $\mathbb{Z}$.

(b) We apply the former identity to the eigenvalues $(\lambda_1, \dots, \lambda_n)$ of the matrix A. Since the $\sigma_j(\lambda_1, \dots, \lambda_n)$'s are the coefficients (up to ± 1) of the characteristic polynomial P_A, they write as polynomials in the entries of A, with coefficients in $\mathbb{Z}$.

(c) A symmetric matrix with real entries has real eigenvalues only. Hence each term $(\lambda_j - \lambda_i)^2$ is real non-negative and the product is non-negative too. If $n = 2$ the discriminant of

$$A = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$$

is $D_S(A) = (a-c)^2 + b^2$ which is not the square of any polynomial. If D_S were the square of some polynomial Q for some $n \ge 2$, then the restriction of Q to the set of block-diagonal matrices where the first block is symmetric 2×2 and the others are scalars $a_1, \dots, a_{n-2}$ would be the square of a polynomial too. But this restriction, being

$$\Big((a-c)^2 + b^2\Big) \prod_{j<k} (a_j - a_k)^2 \prod_j \Big((a - a_j)(c - a_j) - b^2\Big)^2,$$

is the product of such a square by $(a-c)^2 + b^2$. Hence the latter would be a square, a contradiction.

24. Obviously, the λ_i's are the roots of $P = X^n + a_1 X^{n-1} + \cdots + a_n$.

25. (a) Let D denotes the commutator $[A, B]$. It is traceless, hence satisfies, from Cayley-Hamilton's Theorem, $D^2 = dI_2$, where $d = -\det D$. Hence D^2 commutes with every matrix C.

    (b) For every 2×2 matrix, Cayley-Hamilton's Theorem reads $M^2 - (\operatorname{Tr} M)M + \det M = 0$, and $\det M = (\operatorname{Tr} M)^2 - \operatorname{Tr} M^2$. Hence the identity reads $\phi(M, M) = 0$, with

    $$\phi(M, N) = MN + NM - (\operatorname{Tr} M)N - (\operatorname{Tr} N)M + \{(\operatorname{Tr} M)(\operatorname{Tr} N) - \operatorname{Tr}(MN)\}I_2.$$

    Since ϕ is symmetric and bilinear, one deduces that

    $$4\phi(M, N) = \phi(M + N, M + N) - \phi(M - N, M - N) = 0.$$

    (c) Just write that $\operatorname{Tr}(\phi(N_1, N_2)N_3) = 0$.

    (d) Again, write the Cayley-Hamilton's Theorem for an arbitrary matrix M. Let write the coefficients of the characteristic polynomial P_M in terms of the sums

    $$s_k := \sum_j \lambda_j^k,$$

    where the λ_j's are the eigenvalues of M. This is achieved through the Newton's formula (see Section 10.5). Now, Cayley-Hamilton's formula appears as a polynomial identity in M, homogeneous of degree n. It may be written as $\phi(M, \ldots, M)$, where ϕ is symmetric n-linear. There follows the polynomial identity $\phi(N_1, \ldots, N_n) = 0$ for every $n \times n$ matrices $N_1, \ldots, N_n$. Multiplying by N_{n+1}, and taking the trace yields the desired formula.

26. (a) We proceed by induction over p. There is nothing to prove if $p = 0$. If $p \geq 1$ and if the statement is true at order $p - 1$, then we expand the derivative of the characteristic polynomial :

    $$P'_A(X) = \sum_i P_{A(i)}(X), \quad A(i) := A_{J(i)},$$

    where $J(i)$ is the set of all indices but i. By induction, λ is a root of multiplicity p (at least) of all terms in the sum, and therefore of P'_A. Since by assumption it is also a root of P_A, its multiplicity must be $p + 1$ at least.

    (b) Let E_J be the subspace spanned by $\vec{e}_j$ when j runs over J, and let F be the eigenspace associated to λ. Since $\dim F = q$, its intersection with E_J is non-trivial whenever $r := \operatorname{card} J > n - q$. Hence there exists an eigenvector in E_J, which provides an eigenvector of A_J for the eigenvalue λ.

27. If $l < n$ then just take $q_l = X^l$. If $l \geq n$, then write $A^{l-n} P_A(A) = 0$, where P_A is the characteristic polynomial of A. Since $X^{l-n} P_A = X^l + \text{l.o.t.}$, this shows that A^l is a linear combination of $A^{l-n}, \ldots, A^{l-l}$. Then proceed by induction over l. If A is invertible, then $A^{-l} P_A(A) = 0$ shows that A^{-l} is a linear combination of $A^{1-l}, \ldots, A^{n-l}$. For $l = -1$, this gives already the result. Then proceed by induction over $-l$.

28. It is enough to show that the map is one-to-one. So let M be such that $AM = MB$. Then $A^2M = AMB = MB^2$. By induction, there holds $A^kM = MB^k$ for every $k \in \mathbb{N}$ and hence $p(A)M = Mp(B)$ for every polynomial p. Let us take $p = P_B$, the characteristic polynomial of B. From Cayley-Hamilton's Theorem, we obtain $p_B(A)M = 0$. However, the eigenvalues of $p_B(A)$ are the images of those of A under P_B. By assumption, all of them are non-zero and thus $P_B(A)$ is invertible. Hence $M = 0$.

29. (a) Let M_{rs} be a non-zero matrix. Then, from $M_{rj}M_{js} = M_{rs}$, one finds that none of M_{rj} or M_{js} vanish. Repeating the argument, we find tha none of the M_{ij} vanish.

    (b) Trivial : we have $M_{ii}^2 = M_{ii}$.

    (c) Let $x_j \in E_j$. If $\sum_j x_j = 0$, then apply M_{ii} for some index i. Since $M_{ii}x_i = x_i$ and $M_{ii}x_j = 0$ otherwise, there remains $x_i = 0$. Hence $E_1, \ldots, E_n$ are in direct sum. Since M_{ii} is non-zero, E_i is non trivial. From $\sum_j \dim E_j \leq n$ and $\dim E_j \geq 1$, there comes $\dim E_j = 1$.

    (d) Given a generator e_j of E_j, one has $M_{jj}e_j = e_j$. Because of $M_{jk} = M_{jj}M_{jk}$, we know that $R(M_{jk}) \subset E_j$. But since M_{jk} is non zero, we really have $R(M_{jk}) = E_j$. Next, $M_{jk}e_l = M_{jk}M_{ll}e_l = 0$ if $l \neq k$. Hence $R(M_{jk})$ is spanned by $M_{jk}e_k$ and therefore $M_{jk}e_k = \alpha_{jk}e_j$ for some scalar $\alpha_{jk} \neq 0$. At last, writing $M_{ij}M_{jk}e_k = M_{ik}e_k$, we obtain $\alpha_{ij}\alpha_{jk} = \alpha_{ik}$. Let us denote $\beta_j := \alpha_{1j}$. Then there comes (take $i = 1$) $\alpha_{jk} = \beta_k/\beta_j$. Now, renormalizing the generators by $e_j \mapsto e_j/\beta_j$, we obtain $M_{ij}e_j = e_i$ for every indices i, j.

    (e) Let σ be an automorphism of $\boldsymbol{M}_n(k)$. We define $M_{ij} := \sigma(E_{ij})$, where E_{ij}'s are elementary matrices : $(E_{ij})_{kl} = \delta_i^k\delta_j^l$. Then the M_{ij}'s satisfy the assumption. Therefore there exists a basis $\mathcal{B} = \{e_1, \ldots, e_n\}$ of k^n such that $M_{ij}e_k = \delta_j^k e_i$. Let P be the change of basis from the canonical one to $\mathcal{B}$. Then $M_{jk}P = PE_{jk}$. Since the E_{ij}'s form a basis of $\boldsymbol{M}_n(k)$, this extends by linearity, giving $\sigma(M) = PMP^{-1}$.

3 Matrices with real or complex entries

1. Let A be skew-hermitian and X be an eigenvector, $AX = \mu X$. Then $\mu\|X\|^2 = X^*AX$. Hence $\bar{\mu}\|X\|^2 = (X^*AX)^* = X^*A^*X = -X^*AX = -\mu\|X\|^2$. Thus $\bar{\mu} = -\mu$, that is $\mu \in i\mathbb{R}$. Last, a real skew-symmetric matrix is skew-hermitian.

2. The polynomial $f(z) := \det(P + zQ)$ is non trivial since $f(i) \neq 0$. Hence it vanishes only finitely many times. In particular, there are plenty of real numbers b such that $f(b) \neq 0$, that is $P + bQ \in \boldsymbol{GL}_n(\mathbb{R})$.

   Let $M, N \in \boldsymbol{M}_n(\mathbb{R})$ be similar in $\boldsymbol{M}_n(\mathbb{C})$. There exists $P, Q \in \boldsymbol{M}_n(\mathbb{R})$ such that $P + iQ \in \boldsymbol{GL}_n(\mathbb{C})$ and $(P + iQ)M = N(P + iQ)$. Taking real and imaginary parts, there comes $PM = NP$, $QM = NQ$. From above, there exists a real number b such that $P + bQ \in \boldsymbol{GL}_n(\mathbb{R})$. Since $(P + bQ)M = N(P + bQ)$, M, N are similar within $\boldsymbol{GL}_n(\mathbb{R})$.

3. Let T be upper triangular and normal. Computing the diagonal entries of $TT^* = T^*T$, we obtain

$$\sum_{k\leq i} |t_{ki}|^2 = \sum_{k\geq i} |t_{ik}|^2.$$

Making $i = 1$, there follows $t_{1k} = 0$ for every $k > 1$. Hence T is block diagonal, where diagonal blocks are upper triangular and inherit the normality. We conclude by induction over n.

If M is normal and if $M = U^*TU$ is a unitary trigonalization, then T is triangular and normal, hence is diagonal.

4. Writing that the restriction of A to the plane spanned by the i-th and j-th basis vectors is positive definite, we have

$$a_{ij}^2 < a_{ii}a_{jj} \leq (\max_l a_{ll})^2.$$

5. (a) A direct computation shows that $h_M \circ h_N = h_{MN}$. Since h_{I_2} is the identity, this shows that h_M is a bijection, with the inverse h_N, $N := M^{-1}$. Notice that the point at infinity is crucial in this procedure, since $h_M(\infty) = a/c$ and $h_M(-d/c) = \infty$.

   (b) There remains to compute the kernel of this multiplicative homomorphism. If h_M is the identity, then $az + b = z(cz + d)$ for every z. Hence, $c = 0 = b$ and $a = d$. The kernel is thus $\mathbb{R}I_2$, the set of homotheties.

   (c) One finds

$$\Im h_M(z) = \frac{\det M}{|cz + d|^2} \Im z.$$

   In particular, $h_M(\mathcal{H}) \subset \mathcal{H}$ whenever $\det M > 0$. Since then $\det M^{-1} > 0$ and $h_{M^{-1}} = h_M^{-1}$, we have actually $h_M(\mathcal{H}) = \mathcal{H}$. Thus $\boldsymbol{GL}_2^+(\mathbb{R})$ acts on $\mathcal{H}$. Notice that the point at infinity is not needed in this analysis, since $h_M(\infty)$ is real.

   (d) Obviously, the quotient $\boldsymbol{GL}_2^+(\mathbb{R})/\mathbb{R}I_2$ acts on $\mathcal{H}$. But the quotient $\boldsymbol{GL}_2^+(\mathbb{R})/\mathbb{R}^+I_2$ is isomorphic to $\boldsymbol{SL}_2(\mathbb{R})$, through $M \mapsto M/\det M$. Hence the former is isomorphic to $\boldsymbol{PSL}_2(\mathbb{R})$.

   (e) Let z be a fixed point of h_M : $az + b = z(cz + d)$. This is a quadratic equation with real coefficients. Its discriminant is $(d - a)^2 - 4bc = (\text{Tr M})^2 - 4$. If $|\text{Tr M}| \geq 2$, then the roots are real and h_M does not have a fixed point in $\mathcal{H}$. If $|\text{Tr M}| < 2$ there are two complex conjugate roots of which precisely one belongs to $\mathcal{H}$. **Nota** : a matrix in $\boldsymbol{SL}_2(\mathbb{R})$ is called *hyperbolic*, *elliptic* or *parabolic* if the modulus of its trace is larger than, equal to or less than 2, respectively. Hyperbolic matrices fix exactly one point in $\mathcal{H}$. Elliptic matrices fix two points on the boundary $\mathbb{R} \cup \{\infty\}$, while parabolic fix only one point on the boundary.

6. Let $\mathcal{F}$ be a set of convex functions on $\mathbb{R}^N$ and F be the supremum of its elements : $F(x) = \sup_{f\in\mathcal{F}} f(x)$. Given vectors x, y and a number $\theta \in [0, 1]$, we have, for every f in $\mathcal{F}$,

$$f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta)f(y) \leq \theta F(x) + (1 - \theta)F(y).$$

Taking the supremum gives the convexity of F.

We apply this principle to the maximal eigenvalue λ_n of an $n \times n$ hermitian matrix. From Rayleigh's formula, it is given as the supremum of affine (thus convex) functions

$$M \mapsto \frac{x^* M x}{x^* x}.$$

7. If M is normal, then $M = V^* D V$ where $V \in \boldsymbol{U}_n$ and D is diagonal. There exists a unitary diagonal matrix W such that $\bar{D} = DW$. Hence

$$M^* = V^* \bar{D} V = V^* D W V = MU, \quad U = V^* W V \in \boldsymbol{U}_n.$$

Conversely, let us assume that $M^* = MU$ with U unitary. Taking the hermitian adjoint, we have $M = U^* M^* = U^* M U$. Hence $UM = MU$. There follows $M^* M = MUM = M^2 U = MM^*$.

8. Within the set of triangular matrices, those having pairwise distinct diagonal entries form a dense subset. Since these ones have distinct eigenvalues, they are diagonalizable. Hence the set of diagonalizable triangular matrices is dense in the set of triangular matrices. Now, let M be any matrix in $\boldsymbol{M}_n(\mathbb{C})$. From Schur's Theorem 3.1.3, it writes as $U^* T U$ where U is unitary and T is triangular. Let T_k be asequence of diagonalizable triangular matrices converging towards T. Then $U^* T_k U$ is a sequence of diagonalizable matrices converging towards M.

9. We may assume that both sequences are ordered increasingly. We shall prove that the maximum is $a_1 b_1 + \cdots + a_n b_n$ and the minimum is $a_1 b_n + \cdots + a_n b_1$. First of all, both these values are achieved by choosing diagonal matrices A and B. By unitary diagonalisation, we may suppose that $A = \mathrm{diag}(a_1, \ldots, a_n)$. We denote $\epsilon_j := a_{j+1} - a_j \geq 0$. From Schur's Theorem 3.4.1, we have

$$\sum_i^n b_{jj} \leq \sum_i^n b_j, \quad 2 \leq i \leq n,$$

while the sums from 1 to n are equal. Hence,

$$\begin{aligned} \mathrm{Tr}(AB) &= \sum_1^n a_j b_{jj} = a_1 \mathrm{Tr}\, B + \epsilon_1 \sum_2^n b_{jj} + \cdots + \epsilon_{n-1} b_{nn} \\ &\leq a_1 \sum_1^n b_j + \epsilon_1 \sum_2^n b_j + \cdots + \epsilon_{n-1} b_n = \sum_1^n a_j b_j. \end{aligned}$$

Using again Schur's Theorem 3.4.1, we have

$$\sum_i^n b_{jj} \geq \sum_1^{n-i+1} b_j, \quad 2 \leq i \leq n,$$

and we obtain

$$\mathrm{Tr}(AB) \geq a_1 \sum_1^n b_j + \epsilon_1 \sum_1^{n-1} b_j + \cdots + \epsilon_{n-1} b_1 = a_1 b_n + \cdots + a_n b_1.$$

10. (a) We notice that $L+l$ depends only on $u_1+u_n, u_2, \ldots, u_{n-1}$, while $L-l$ depends only on $u_1-u_n, u_2, \ldots, u_{n-1}$. Actually, the restriction of $L+l$ to K_n depends only on $u_2, \ldots, u_{n-1}$:

$$L(u)+l(u) = (a_1+a_n)\left(1-\sum_2^{n-1} u_j\right) + \sum_2^{n-1} u_j\left(a_j+\frac{1}{a_j}\right) =: h(u_2,\ldots,u_{n-1}).$$

Therefore, $L+l$ is maximal on K_n by maximizing h under the constraints

$$u_2,\ldots,u_{n-1} \geq 0 \text{ and } \sum_2^{n-1} u_j \leq 1.$$

Remarking that the coefficient $a_j+1/a_j-a_n-a_1$ of u_j in h is non-positive (because $t \mapsto t+1/t$ is convex), we see that the maximum of $L+l$ is achieved for $u_2 = \cdots = u_{n-1} = 0$, with

$$\max_{K_n}(L(u)+l(u)) = a_1+a_n.$$

On an other hand, $L-l$ vanishes in K_n, for instance at the point $(1/2,0,\ldots,0,1/2)^T$ in which $L+l$ is maximal. This point answers the question.

(b) Since $4lL = (L+l)^2 - |L-l|^2$ and $L+l$ is always non-negative, we have from the previous question

$$4\max_{K_n} l(u)L(u) = \left(\max_{K_n}(L(u)+l(u))\right)^2 - \left(\min_{K_n}(L(u)-l(u))\right)^2 = (a_1+a_n)^2.$$

(c) Let us express x in an orthonormal eigenbasis of A. Then

$$\frac{(x^*Ax)(x^*A^{-1}x)}{\|x\|^2} = l(u)L(u),$$

where $u_j = x_j^2/\|x\|^2$, hence u belongs to K_n. If $a_1a_n = 1$, we have the desired inequality from the previous question. Otherwise, let just renormalize by dividing A by $\sqrt{a_1a_n}$.

11. (a) We use the min-max formula, denoting by $r_A(x)$ the Rayleigh quotient $x^*Ax/\|x\|^2$:

$$\gamma_j = \min_{\dim F=j}\max_{x\in F\setminus\{0\}} r_{A+B}(x) \leq \min_{\dim F=j}\max_{x\in F\setminus\{0\}}(r_A(x)+\beta_n) = \alpha_j+\beta_n.$$

Likewise,

$$\gamma_j \geq \min_{\dim F=j}\max_{x\in F\setminus\{0\}}(r_A(x)+\beta_1) = \alpha_j+\beta_1.$$

(b) Actually, for every unit vector $x \in G\cap H$, there holds $x^*Ax \leq R_A(G)$, $x^*Bx \leq R_B(H)$ and therefore $x^*Cx \leq R_A(G)+R_B(H)$. Then take the supremum.

(c) In the former inequality, let us take the infimum over the subspaces G, H of respective dimensions l, m. We obtain

$$\inf R_C(G \cap H) \leq \alpha_l + \beta_m,$$

where the infimum runs over all subspaces $G \cup H$ of dimension between k and $\min(l, m)$. Hence

$$\gamma_k = \min_{k \leq j \leq \min(l,m)} \gamma_j \leq \alpha_l + \beta_m.$$

(d) Just apply the last result to $-A, -B, -C$.

(e) Let us apply the first question to the matrices $A, D = B - A, B$, where the eigenvalues of D are $\delta_1 \leq \cdots \leq \delta_n$. Then

$$|\beta_k - \alpha_k| \leq \max(\delta_n, -\delta_1) = \rho(B - A).$$

Last, we use the fact that $\rho(D) = \|D\|_2$ for hermitian matrices.

12. Use the previous exercise, or use directly the Rayleigh quotients.

13. Apply the previous exercise.

14. From Proposition 8.1.2 (Schur's formula), we have $\det M = \det A \det(C - B^* A^{-1} B)$ (notice that A, being positive definite, is invertible). The inequality $\det(C - B^* A^{-1} B) \leq \det C$ follows from the previous exercise with the obvious fact that $B^* A^{-1} B$ is positive semidefinite and the fact, that we shall prove now, that $C - B^* A^{-1} B$ is positive semidefinite. Given $y \in \mathbb{C}^q$, let us define $x = (-A^{-1} B y, y)^T$. Then $y^*(C - B^* A^{-1} B) y = x^* M x$ is positive whenever $y \neq 0$.

15. The diagonal entries of M^{-1} are the principal minors of M, divided by $\det M$. Hence their product equals $P_{n-1}(M)(\det M)^{-n}$. Applying Hadamard's inequality to M^{-1}, we obtain $(\det M)^{n-1} \leq P_{n-1}(M)$. Last, $P_n(M) = \det M$.

We prove the general formula by applying the former to every principal minor of size $k+1$. This immediately gives an upper bound of $P_{k+1}(M)$ in terms of a product of principal minors of size k. Obviously, this product is a symmetric function on the set of these minors. Hence it equals a power of $P_k(M)$. The power is found by looking at the homogeneity degrees of P_k and P_{k+1} with respect to the entries of M. These degrees are $k \begin{pmatrix} n \\ k \end{pmatrix}$ and $(k+1) \begin{pmatrix} n \\ k+1 \end{pmatrix}$, respectively. Hence the power is the ratio $(n-k)/k$ of these two numbers.

16. (a) Since $d(0_n)(d(M) - 1) \equiv 0$ and d is non constant, one has $d(0_n) = 0$. Since $d(M)(d(I_n) - 1) \equiv 0_n$ we have likewise $d(I_n) = 1$. If P is non singular, then $d(P)d(P^{-1}) = d(I_n) = 1$, hence $d(P) \neq 0$ and $d(P^{-1}) = 1/d(P)$. Last, $d(P^{-1}MP) = d(P^{-1})d(M)d(P) = d(M)$.

(b) Just choose $D_k = \mathrm{diag}(d_{k+1}, \ldots, d_{k+n})$, where $D = \mathrm{diag}(d_1, \ldots, d_n)$, and the indices are computed modulo n. Then $\delta(\det D)^n = d(\det I_n) = d(DD_1 \cdots D_{n-1}) = d(D)d(D_1) \cdots d(D_{n-1})$. We conclude with the remark that $D, D_1, \ldots, D_n$ are similar to each other. Hence $\delta(\det D)^n = d(D)^n$. Last, both δ and d are non-negative.

(c) Since M is similar to a diagonal matrix D, we have $d(M) = d(D) = \delta(\det D) = \delta(\det M)$.

(d) Obviously, δ is multiplicative too. Since M^T is similar to M, we have $d(M^T) = d(M)$. Hence, using the fact that the symmetric matrix $M^T M$ is diagonalizable, we have

$$d(M)^2 = d(M^T)d(M) = d(M^T M) = \delta(\det M^T M) = \delta((\det M)^2) = \delta(\det M)^2.$$

17. Let $C := B^{-1}B^*$. Then $C^{-1} = B^{-*}B$ and $C^* = BB^{-*}$ are similar.

(a) Conversely, let A be given such that $A^* = PA^{-1}P^{-1}$. Then $A^*PA = P$ and thus $A^*P^*A = P^*$. One chooses $H = \mu P + \bar{\mu}P^*$, where $\mu \in \mathbb{C}$. There remains to show that μ can be chosen in such a way that H be non-singular. This is true because $-\mu/\bar{\mu}$ runs over the unit numbers when μ runs over $\mathbb{C}^*$, and $P^{-1}P^*$ has finitely many eigenvalues.

(b) Since, for every complex number a, there holds $BA = B^*$ for $B := (aI_n + \bar{a}A^*)H$, it is enough to find a such that $\det B \neq 0$. By the same argument as above, there exists an a such that $aI_n + \bar{a}A^*$ is non singular. Then B is invertible and $A = B^{-1}B^*$.

18. Since $\sum_{ij} |a_{ij}|^2 = \mathrm{Tr}(A^*A)$ is invariant under unitary conjugation, as well as $\sum_l |\lambda_l|^2$ and normality, and since every matrix is unitary similar to a triangular matrix, it is enough to prove the result when A is triangular. With this restriction, we have seen in Exercise 3 that A is normal if and only if it is diagonal. On the other hand, $\lambda_l = a_{ll}$, so that the equality between sums amounts to writing $a_{ij} = 0$ for every distinct indices.

19. (a) The eigenvalues of A are pure imaginary numbers $i\mu_j$. Thus $\det(I_n + A)$, being a product of numbers $1 + i\mu_j$ whose moduli are greater than or equal to 1, has modulus greater than or equal to 1. Equality means that the spectrum of A reduces to $\{0\}$; since A is normal, hence diagonalizable, this means $A = 0_n$.

(b) Let K be the square root of H^{-1} (K is hermitian positive definite and $K^2 = H^{-1}$). Then $M - M^*$ is skew-hermitian and $H^{-1}(M - M^*) = K(K^*(M - M^*)K)K^{-1}$ is similar to the skew-hermitian matrix $K^*(M - M^*)K$. Using the previous result, we compute

$$\det H \leq \det H |\det(I_n + \frac{1}{2}H^{-1}(M - M^*))| = |\det(H + \frac{1}{2}(M - M^*))| = |\det M|.$$

20. The eigenvalues of M are non-negative, with a sum equal to n, the trace of M. Since one of them is n, all other eigenvalues vanish. Since a real symmetric matrix is diagonalizable, M is similar to $\mathrm{diag}(0, \ldots, 0, n)$ and has rank 1. Hence there exist non-zero vectors X, Y

such than $M = XY^T$. One of them, say X, can be chosen of norm 1 ($\|X\|_2 = 1$, that is $X^TX = 1$). The symmetry $XY^T = YX^T$ shows that X and Y are colinear. Thus $M = \mu XX^T$. Next, $MX = \mu X$, so that μ equals the only non-zero eigenvalue n. Finally, $M = nXX^T$ and $m_{jj} = nx_j^2$ give $x_j = \pm 1/\sqrt{n}$. Conversely, every matrix of the form nXX^T with $x_j = \pm 1/\sqrt{n}$ satisfies all the assumptions.

21. (a) Trivial.

    (b) From Proposition 3.2.1, $A = \sum_1^p x_\alpha x_\alpha^*$ and $B = \sum_1^q y_\alpha y_\alpha^*$. Then the desired formula holds with $z_{\alpha\beta}$ defined by $x_\alpha \circ y_\beta$, that is

    $$(z_{\alpha\beta})_i := (x_\alpha)_i (y_\beta)_i.$$

    Therefore, $A \circ B$ is positive semi-definite.

    (c) If A and B are positive semi-definite, then $\{x_1, \ldots, x_n\}$ and $\{y_1, \ldots, y_n\}$ are bases of $\mathbb{C}^n$. Let $p \in \mathbb{C}^n$ be orthogonal to all the $z_{\alpha\beta}$'s. Then, denoting $P := \mathrm{diag}(p_1, \ldots, p_n)$, there holds $x_\alpha \perp P y_\beta$ for every pair (α, β). There follows that $Py_\beta = 0$ and therefore $P = 0_n$, or $p = 0$. This means that the $z_{\alpha\beta}$'s span $\mathbb{C}^n$, so that $A \circ B$ is positive definite.

    (d) Here is an example, with $p = q = 2 < n = 3$:

    $$x_\alpha = y_\alpha = \begin{pmatrix} 1 \\ a_\alpha \\ a_\alpha^2 \end{pmatrix}.$$

    There are three distinct vectors $z_{\alpha\beta}$ and their determinant is of Vandermonde type. It is non zero whenever $ab(b^2 - a^2) \neq 0$.

22. (a) The implication is trivial for $j = 1$ or 2. Let us assume **P3** and let B be a principal sub-matrix of A, with a real eigenvalue μ associated with an eigenvector y. Completing y with null entries, we obtain a non-zero vector $x \in \mathbb{R}^n$, for which $Ax - \mu x$ has zero entries along the lines kept in B. Therefore $(Ax - \mu x)_l x_l = 0$ for every l. Let D be a non-negative diagonal matrix such that $(Ax, Dx) > 0$. Then $(Ax - \mu x, Dx)$ vanishes, being a sum of zeroes. Thus $\mu(x, Dx) > 0$, which implies $\mu > 0$.

    At last, let us assume **P4**. With the notation above, the spectrum of B consists in positive real eigenvalues and non-real eigenvalues which come by complex conjugate pairs. Hence their product $\det B$ is positive.

    (b) The coefficients of the polynomial $D \mapsto \det(A + D)$ are principal minors of A, thus positive numbers. This implies that the polynomial achieves positive values for non-negative data.

    (c) Let $x \neq 0$ be given and $y := Ax$. Let J be the (non-void) set of indices of the non-zero components of x. We denote by $x(J), y(J), A(J)$ the vectors and matrix obtained by preserving only rows and columns of indices in J. In particular $y(J) = A(J)x(J)$. If $y_k x_k$ were non-positive for every k, there would exist a non-negative diagonal matrix

Δ such that $y(J) = -\Delta x(J)$. Hence, there would hold $(A(J) + \Delta)x(J) = 0$, hence $\det(A(J) + \Delta) = 0$. This contradicts the former result, applied to $A(J)$.

Finally, **P5** implies **P1**, so that the five properties are equivalent to each other.

4 Norms

1. From $Mx = (b, x)a$ and Proposition 4.1.2, we have $\|M\|_p = \|a\|_p \|b\|_{p'}$. Let us write that $\|M\|_p = 1$ for $p = 1, 2, \infty$:

$$1 = \|a\|_2^2 \|b\|_2^2 \leq \|a\|_1 \|a\|_\infty \|b\|_1 \|b\|_\infty = 1.$$

   Therefore, $\|b\|_2^2 = \|b\|_1 \|b\|_\infty$ and $\|a\|_2^2 = \|a\|_1 \|a\|_\infty$. This means that, for every index i, $|a_i|$ equals either $\|a\|_\infty$ or zero. Likewise, $|b_i|$ equals either $\|b\|_\infty$ or zero. Then $1 = l\|a\|_\infty \|b\|_\infty = m\|a\|_\infty \|b\|_\infty$, where l, m are the numbers of non-zero components of a, b. Hence $l = m$.

   Conversely, let a, b and l be given such that l components of a have moduli $\|a\|_\infty$ and l components of b have moduli $\|b\|_\infty$. Assume also that $l\|a\|_\infty \|b\|_\infty = 1$. Then $\|M\|_p = 1$ for every p.

2. Let J be the set of indices such that $x_i y_i \neq 0$. Let $x(J)$ and $y(J)$ denote the vector obtained by keeping only the components with index in J. Then $|(x, y)| = |(x(J), y(J))| \leq \|x(J)\|_p \|y(J)\|_{p'} \leq \|x\|_p \|y\|_{p'}$. Let us assume that $|(x, y)| = \|x\|_p \|y\|_{p'}$. Then both inequalities above must be equalities. On a first hand this means $x_i = 0$ (if $p < \infty$) and $y_i = 0$ (if $p > 1$) for every i in the complement of J. On the other hand, if $p < \infty$, there exists a non-zero complex number λ such that $y_i = \lambda x_i |x_i|^{p-2}$ for every $i \in J$. If $p > 1$, we write instead that there is a non-zero μ such that $x_i = \mu y_i |y_i|^{p'-2}$ for every $i \in J$.

   We turn to the equality case in Minkowski inequality. If $p \in (0, \infty)$, this is equivalent to the colinearity of x and y. If $p = 1$, it is equivalent to the fact that y_i/x_i belongs to $\mathbb{R}^+ \cup \{\infty\}$ for every i. If $p = \infty$, it is equivalent to the existence of an index i such that x_i *and* y_i have maximal moduli, and y_i/x_i belongs to $\mathbb{R}^+ \cup \{\infty\}$.

3. Just write

$$\|x\|_\infty \leq \|x\|_p \leq \|x\|_\infty n^{1/p}$$

   and let p tend to $+\infty$.

4. (a) See exercise 2.

   (b) Permutations of coordinates are isometries of $(\mathbb{C}^n, \|\cdot\|_p)$ and therefore the induced norm is invariant under such permutations. Then Corollary 5.5.1 asserts that the non-trivial convex set made of bi-stochastic matrices is contained in the unit sphere for this norm, which is thus not strictly convex.

5. (a) Let $x, y \in \mathbb{C}^n, \lambda \in \mathbb{C}$ be given. For every $\epsilon > 0$, there exists decompositions $x = \sum_l \alpha_l x^l$ and $y = \sum_k \beta_k y^k$ such that

$$\sum_l |\alpha_l| N(x^l) < N_1(x) + \epsilon, \quad \sum_k |\beta_k| N(y^k) < N_1(y) + \epsilon.$$

Taking the union of the x^l's and the y^k's, we may assume that the sets of indices is the same and $x^l = y^l$. Then $x + y = \sum_l (\alpha_l + \beta_l) x^l$ and $\lambda x = \sum_l (\lambda \alpha_l) x^l$, so that

$$N_1(x + y) \leq N_1(x) + N_1(y) + 2\epsilon, \quad N_1(\lambda x) \leq |\lambda|(N_1(x) + \epsilon).$$

Letting $\epsilon \to 0^+$, we have

$$N_1(x + y) \leq N_1(x) + N_1(y), \quad N_1(\lambda x) \leq |\lambda| N_1(x).$$

If $x \in \mathbb{R}^n$, the decomposition $x = x$ gives $N_1(x) \leq N(x)$. On an other hand, for every decomposition of x, we have $x = \sum_l (\Re \alpha_l) x^l$ and, since N is a norm over $\mathbb{R}^n$,

$$N(x) \leq \sum_l |\Re \alpha_l| N(x^l) \leq \sum_l |\alpha_l| N(x^l).$$

Taking the infimum, there follows $N(x) \leq N_1(x)$, that is $N(x) = N_1(x)$. Last, the same computation shows that, for every vector x with real part y, there holds $N_1(y) \leq N_1(x)$. Hence, if $N_1(x) = 0$, there holds $y = 0$; likewise, the imaginary part vanishes and $x = 0$. Therefore, N_1 is a norm.

(b) Each map $x \mapsto [e^{i\theta} x]$ is a norm on $\mathbb{C}^n$, viewed as an $\mathbb{R}$-vector space. Hence N_2 is such one too, as their average. On the other hand, $N_2(e^{i\alpha} x) = N_2(x)$ for every real number α. Hence N_2 is a $\mathbb{C}$-norm. If $x \in \mathbb{R}^n$, then $[e^{i\theta} x] = N(x)\sqrt{\cos^2\theta + \sin^2\theta} = N(x)$ gives $N_2(x) = N(x)$.

(c) For every decomposition of x, we have

$$N_2(x) \leq \sum_l |\alpha_l| N_2(x^l) = \sum_l |\alpha_l| N(x^l).$$

Taking the infimum gives $N_2 \leq N_1$. **Nota** : this fact is true for every norm of $\mathbb{C}^n$ extending N. In other words, N_1 is the largest such norm.

(d) The decomposition on the canonical basis gives immediately $N_1(x) \leq \|x\|_1$. Since $\|\cdot\|_1$ is a norm on $\mathbb{C}^n$ which extends its restriction to $\mathbb{R}^n$, the maximality property mentionned above shows that $\|x\|_1 \leq N_1(x)$. Hence the equality.

Finally, we compute N_2 for $x = (1, i)^T$. The real and imaginary parts of $e^{i\theta} x$ are $(\cos\theta, -\sin\theta)^T$ and $(\sin\theta, \cos\theta)^T$, each one of norm 1. Thus $[e^{i\theta} x] = \sqrt{2}$ and $N_2(x) = \sqrt{2} < N_1(x) = \|x\|_1 = 2$.

6. Since a supremum over $\mathbb{C}^n \setminus \{0\}$ is larger than or equal to the supremum of the same quantity over $\mathbb{R}^n \setminus \{0\}$, and since N_1 and N coincide on $\mathbb{R}^n$, there holds $N_1(M) \geq$

$N(M)$ (this is actually true for every extension of the norm N). Conversely, given a decomposition of a vector $x \in \mathbb{C}^n$, we have

$$N_1(Mx) = N_1\left(\sum_l \alpha_l M x^l\right) \leq \sum_l |\alpha_l| N_1(Mx^l) = \sum_l |\alpha_l| N(Mx^l) \leq N(M) \sum_l |\alpha_l| N(x^l).$$

Taking the infimum over all decompositions of x gives $N_1(Mx) \leq N(M)N_1(x)$. Taking the supremum over the unit ball gives $N_1(M) \leq N(M)$.

Let p, q, θ be as in Theorem 4.3.1, and let M be given in $\mathbf{M}_n(\mathbb{R})$. To distinguish the p-norm of M in $\mathbf{M}_n(\mathbb{R})$ from that in $\mathbf{M}_n(\mathbb{C})$, we denote $\|M\|_p$ and $\|M_{\mathbb{C}}\|_p$ respectively. By the same argument as above, one has $\|M_{\mathbb{C}}\|_p \geq \|M\|_p$. On the other hand, denoting $N_{1,p}$ the N_1-norm constructed from $\|\cdot\|_p$, and recalling its maximality among the norms which extend $\|\cdot\|_p$, we have $\|M_{\mathbb{C}}\|_p \leq N_{1,p}(M)$. Since M has real entries, $\|M\|_p = N_{1,p}(M)$ and hence $\|M_{\mathbb{C}}\|_p = \|M\|_p$, and similarly with q, r instead of p. Applying now Theorem 4.3.1 to $M_{\mathbb{C}}$, we obtain the inequality

$$\|M\|_r \leq \|M\|_p^{\theta} \|M\|_q^{1-\theta}.$$

7. Use Proposition 4.4.1 and the fact that $\|A^n\|^{1/n} \leq \|A\|$.

8. Since D and N commutte,

$$(D+N)^m = \sum_{k=0}^{n-1} \left(\begin{array}{c} m \\ k \end{array} \right) N^k D^{m-k}.$$

Using triangle inequality, and a bound of the norms of matrices $N, \ldots, N^{n-1}$ as well as of $D, \ldots, D^{n-1}$, we obtain

$$\|(D+N)^m\| \leq c_0 m^n \|D^{m-n+1}\|.$$

Taking the m-th root and letting m tend towards $+\infty$, we obtain $\rho(D+N) \leq \rho(D)$. Exchanging the rôles of D and $D+N$, we also have the opposite inequality.

9. If $\lambda = 0$, then $\|B\| = 0$ thus $B = 0$, which proves the desired property. Otherwise, let us assume that X belongs to the range of $B - \lambda$. Then there is a vector Y such that $(B-\lambda)Y = X$. For every scalar $k \in \mathbb{N}$, we have $B^k Y = \lambda^k Y + k\lambda^{k-1} X$, from which there comes

$$\|\lambda^k Y + k\lambda^{k-1} X\| \leq |\lambda|^k \|Y\|.$$

Dividing by $|\lambda|^k$ and letting k tend to $+\infty$, we obtain a contradiction. We conclude that $\ker(B-\lambda)^2 = \ker(B-\lambda)$, meaning that the Jordan component associated to λ is diagonal.

10. By assumption, there exists $P \in \mathbf{GL}_n(\mathbb{C})$ such that $PBP^{-1} = \mathrm{diag}(B_1, B_2) =: D$, where B_1 is diagonal and $\rho(B_2) < \rho(B)$. To this block decomposition corresponds a factorization $\mathbb{C}^n = (\mathbb{C}^p \times \{0\}) + (\{0\} \times \mathbb{C}^{n-p})$. obviously, there is a norm on $\mathbb{C}^p$ for which the induced norm of B_1 equals $\rho(B_1)$, which is nothing but $\rho(B)$. On an other hand, Householder's Theorem gives a norm on $\mathbb{C}^{n-p}$ for which the induced norm of B_2 is less than $\rho(B)$. Summing both norms, we obtain a norm on $\mathbb{C}^n$ which solves our problem.

11. (a) From Proposition 4.4.1, the series is bounded by a convergent series $\sum_k c\kappa^k$ with $(\rho(A)+\epsilon)^{-1}\rho(A) < \kappa < 1$.

    (b) Each term of the sum is a semi-norm, the first one being a norm. Thus the sum is a norm.

    (c) There holds
    $$\|Ax\| = (\rho(A)+\epsilon)(\|x\| - N(x)).$$

12. (a) The induced norm that we have denoted by $\|\cdot\|_2$, and the Frobenius norm, since they are the square roots of the spectral radius and the trace of A^*A. If U, V are unitary, then $(UAV)^*(UAV) = V^*A^*AV$, which is similar to A^*A and thus has the same spectrum.

    (b) Let $\|\cdot\|$ be a unitary invariant norm and let QH be the polar decomposition of A. Then $\|A\| = \|H\|$. Moreover, H is unitarily diagonalizable. Hence $\|A\|$ depends only on the eigenvalues $s_1, \ldots, s_n$ of H, through the formula
    $$\|A\| = \|\mathrm{diag}(s_1, \ldots, s_n)\|.$$

13. (a) The (k,l)-entry of $U^{*j}AU^j$ is $a_{kl}\omega^{j(l-k)}$. We now use the formula
    $$\frac{1}{n}\sum_{j=0}^{n-1} \omega^{j(l-k)} = \delta_k^l.$$
    Since $\|U^{*j}AU^j\| = \|A\|$, the triangle inequality gives $\|D(A)\| \le \|A\|$.

    (b) Let P be the permutation matrix associated to σ. It is unitary and $A^\sigma = D(P^{-1}AP)$. Hence $\|A^\sigma\| \le \|P^{-1}AP\| = \|A\|$.
    By assumption, $\|P\| = \|I_n\|$ for every permutation matrix. Hence Birkhoff's Theorem and the convexity of the norm give $\|M\| \le \|I_n\|$ for every bi-stochastic matrix M.

    (c) Similar to 13a).

    (d) Idem.

    (e) Just because $T_r(A) = D_{-r}(A) + \cdots + D_r(A)$.

    (f)
    $$\|T_r(A)\| \le \frac{1}{2\pi}\int_0^{2\pi} |d_p(\theta)|\,\|U_\theta A U_\theta^*\|\,d\theta = \frac{\|A\|}{2\pi}\int_0^{2\pi} |d_p(\theta)|\,d\theta.$$

    (g) We have $B = T_{2n+1}(C)$, where
    $$C := \begin{pmatrix} 0 & A^* \\ A & 0 \end{pmatrix}.$$
    Hence $\|B\|_2 \le L_n\|C\|_2$ (this would be true for every unitary invariant norm). Last, $C^*C = \mathrm{diag}(A^*A, AA^*)$ has the same spectrum as A^*A and hence $\|C\|_2 = \|A\|_2$. Similarly, $\|B\|_2 = \|\Delta(A)\|_2$. **Nota** : the same argument (but here $\|C\|_S = \sqrt{2}\|A\|_S$) works for the Schur-Frobenius norm.

(h) Same method for $\|\Delta_0(A)\|_2 \leq L_{n-1}\|A\|_2$.

14. (a) We expand

$$(B_F^* B_F)^m = \sum_{i(1),\ldots,i(2m)\in F} A_{i(1)}^* A_{i(2)} \cdots A_{i(2m-1)}^* A_{i(2m)}.$$

From the assumption,

$$\|A_{i(1)}^* A_{i(2)} \cdots A_{i(2m-1)}^* A_{i(2m)}\| \leq \gamma(i(1)-i(2))^2 \gamma(i(3)-i(4))^2 \cdots \gamma(i(2m-1)-i(2m))^2.$$

On an other hand, using $\|A_j\|^2 = \rho(A_j^* A_j) \leq \|A_j^* A_j\| \leq \gamma(0)^2$, we also have

$$\|A_{i(1)}^* A_{i(2)} \cdots A_{i(2m-1)}^* A_{i(2m)}\| \leq \gamma(0)^2 \gamma(i(2)-i(3))^2 \cdots \gamma(i(2m-2)-i(2m-1))^2.$$

Taking the geometric average of both bounds, there comes

$$\|A_{i(1)}^* A_{i(2)} \cdots A_{i(2m-1)}^* A_{i(2m)}\| \leq \gamma(0)\gamma(i(1)-i(2)) \cdots \gamma(i(2m-1)-i(2m)).$$

We apply the triangle inequality, and then sum over $i(1)$, then other $i(2), \ldots$, up to $i(2m-1)$. We obtain

$$\|(B_F^* B_F)^m\| \leq \gamma(0)\|\gamma\|_1^{2m-1} \sum_{i(m)\in F} 1,$$

which gives the desired result.

(b) Again, $\|B_F\|^{2m} = \|B_F^* B_F\|^m = \|(B_F^* B_F)^m\|$ since $B_F^* B_F$ is hermitian. Hence,

$$\|B_F\| \leq (\operatorname{card} F)^{1/2m} \|\gamma\|_1.$$

One concludes by letting m tend towards $+\infty$.

(c) Let x, y be given and $a_j(x,y) := y^T A_j x$. From the previous question, the sum $\left|\sum_{j\in F} a_j\right|$ is bounded independently of F. This ensures that the series $\sum_{\mathbb{Z}} a_j(x,y)$ converges absolutely, thus converges. Obviously, its sum $a(x,y)$ is a bilinear form, thus defines a matrix A through $y^T A x = a(x,y)$. At last, $|a(x,y)| \leq \|x\|\,\|y\|\,\|\gamma\|_1$ gives $\|A\| \leq \|\gamma\|_1$.

(d) Applying the former result to vectors x, y chosen in the canonical basis shows that each entry in the series $\sum_j A_j$ is absolutely summable. This tells that the series $\sum_{\mathbb{Z}} A_j$ is normally convergent. **Nota** : Cotlar's Lemma actually holds when $\mathbb{C}^n$ is replaced by a Hilbert space. Then A_j are bounded operators. The above procedure is a way to define the sum $\sum_j A_j$ as a bounded operator. An important application, in the theory of pseudo-differential operators, is the fact that symbols of order zero give rise to bounded operators on $L^2(\mathbb{R}^m)$. Notice however that, due to the infinite dimension, the sum does not converge normally in general.

15. (a) For ϵ small enough, there holds

$$\|A\|_\infty = \max_i \left(1 - \epsilon b_{ii} + \epsilon \sum_{j\neq i} |b_{ij}|\right).$$

Therefore, the desired inequality holds if and only if

$$\forall i, \quad \omega + \sum_{j\neq i} |b_{ij}| \leq b_{ii}.$$

This tells that B is strictly diagonally dominant. Conversely, let B be strictly diagonally dominant. Then there exist an $\omega > 0$ such that the previous inequality holds for every index i. Then $\|I_n - \epsilon B\|_\infty \leq 1 - \omega\epsilon$ for every $\epsilon > 0$ less then $1/\max_i b_{ii}$.

(b) Since $\|\cdot\|_1$ is the dual norm of $\|\cdot\|_\infty$, $\|I_n - \epsilon B\|_1 \leq 1 - \omega\epsilon$ holds if and only if $\|I_n - \epsilon B^T\|_\infty \leq 1 - \omega\epsilon$. From the previous result this characterizes matrices B such that B^T is strictly diagonally dominant. In other words,

$$\forall j, \quad b_{jj} > \sum_{i\neq j} |b_{ij}|.$$

(c) We have

$$\|I_n - \epsilon B\|_2 = \sup_{X\neq 0} \left|1 - \epsilon \frac{X^T B X}{\|X\|_2}\right|.$$

Also, $X^T B X = \frac{1}{2} X^T (B^T + B) X$. The fraction is bounded by above (by $\|B\|_2$). If $B^T + B$ is positive definite, it is also bounded by below by a positive constant ω, and then the desired inequality holds for $0 < \epsilon < 1/\|B\|_2$. Otherwise, there exists a vector X such that $X^T B X \leq 0$ and thus $\|I_n - \epsilon B\|_2 \geq 1$ for every positive ϵ.

16. (a) Let λ be an eigenvalue. If λ is one of the diagonal entries, then $\lambda \in \mathcal{B}(A)$ trivially. Otherwise, let X be an eigenvector ; it admits at least two non-zero components. Let x_i, x_j be two components of larger moduli. The let us write

$$(\lambda - a_{ii}) x_i = \sum_{k\neq i} a_{ik} x_k, \quad (\lambda - a_{jj}) x_j = \sum_{k\neq j} a_{jk} x_k.$$

Taking the moduli, we obtain

$$|\lambda - a_{ii}|\,|x_i| \leq r_i(A)|x_j|, \quad |\lambda - a_{jj}|\,|x_j| \leq r_j(A)|x_i|.$$

Multiplying both inequalities, and then dividing by $|x_i|\,|x_j|$, we obtain that $\lambda \in \mathcal{B}_{ij}(A)$.

(b) Obviously, $\mathcal{B}_{ij}(A)$ is contained in the union of Geschgöring discs $D_i(A) \cup D_j(A)$. Hence $\mathcal{B}(A)$ is a subset of the Geschgöring domain. In general, the inclusion is strict. For let assume that both sets are equal and let z be a boundary point. Therefore, z

is on the boundary of discs $D_i(A)$ for $i \in I$, and is exterior to the others. Since there exists a pair i, j of distinct indices for which $z \in \mathcal{B}_{ij}(A)$, one immediately finds that $i, j \in I$. Therefore, each point of $\partial\mathcal{G}$ belongs to at least two Gerschgöring discs. In other words, the Gerschgöring domain is covered by discs which occur twice in the list of Gerschgöring discs. A very rare event.

(c) There is only one pair of distinct indices i, j and we have $|\sum_{k\neq i} a_{ik}x_k| = r_i(A)|x_j|$, $|\sum_{k\neq j} a_{jk}x_k| = r_j(A)|x_i|$. Hence $(|\lambda - a_{11}|\,|\lambda - a_{22}| - r_1(A)r_2(A))|x_1|\,|x_2| = 0$. Notice that if $\lambda = a_{ii}$, then $r_i = 0$ and $\mathcal{B}(A)$ is just the union of two points.

17. (a) We only have to remark that, if $\|\cdot\|$ is a hermitian norm, then $x \mapsto \|Px\|$ is another one, for every invertible matrix P.

(b) Let us assume that $\rho(B) < 1$. We just have shown that there exists a hermitian norm, say $x \mapsto x^*Ax$ with $A \in \boldsymbol{HDP}_n$, such that $\|B\| < 1$. In other words, $x^*B^*ABx < x^*Ax$ for every non-zero vector, that is $A - B^*AB \in \boldsymbol{HDP}_n$. Conversely, let $A \in \boldsymbol{HDP}_n$ be such that $A - B^*AB \in \boldsymbol{HDP}_n$. Then $\|B\| < 1$ for the norm induced by $x \mapsto x^*Ax$.

18. (a) The assumptions imply that the discs D_j are pairwise disjoint. Then Theorem 4.5.1 implies that each disc contains exactly one eigenvalue. In particular, the eigenvalues are simple. We shall name hereafter λ_i the eigenvalue belonging in D_i.

(b) Just remark that A^ρ is obtained from A by conjugating by $P = \mathrm{diag}(\ldots, 1, \delta, 1, \ldots)$.

(c) La borne finale est $2n\epsilon^2/\delta$. Obviously, A^ρ has the same diagonal entries as A, including a_{ii}. But $r_i(A^\rho) = \rho r_i(A) \leq \rho\epsilon(n-1)$, while $r_j(A^\rho) \leq \epsilon(n - 2 + 1/\rho)$ otherwise. Since $\rho = 2\epsilon/\delta < 1$, we have for every index $j \neq i$,

$$r_i(A^\rho) + r_j(A^\rho) \leq n\epsilon(\rho + 1) + \frac{\delta}{2} \leq 2n\epsilon + \frac{\delta}{2} < \delta \leq |a_{ii} - a_{jj}|.$$

Therefore the i-th Gerschgöring disc of A^ρ is disjoint from the others. By Theorem 4.5.1, it contains a unique eigenvalue of A^ρ, that is of A. Obviously it is λ_i. We deduce that

$$|\lambda_i - a_{ii}| \leq r_i(A^\rho) \leq n\epsilon\rho = \frac{n\epsilon^2}{\delta}.$$

Nota : this result is interesting in its own, since it tells that an $\mathcal{O}(\epsilon)$ perturbation of the off-diagonal entries of a diagonal matrix is responsible for an $\mathcal{O}(\epsilon^2)$ perturbation of the eigenvalues, when the diagonal entries are pairwise distinct.

19. We begin with the diagonal case ($S = I_n$, $A = D$). Given an eigenvalue μ of $D+E$, either it is one of the d_j's, or $I_n - (D-\mu)^{-1}E$ is singular. The latter case implies (Proposition 4.1.5) $\|(D-\mu)^{-1}E\| \geq 1$ and therefore, in both cases, $\min_j |\mu - d_j| \leq \|E\|$.

In the general case, the spectrum of $A + E$ equals the one of $D + S^{-1}ES$ and we just have to apply the previous result.

20. Writing $|a_{ij}x_j| = |a_{ij}|^{1/2}|a_{ij}|^{1/2}|x_j|$ and using Cauchy-Schwarz inequality, we have

$$\begin{aligned}\|Ax\|_2^2 &= \sum_i \left|\sum_j a_{ij}x_j\right|^2 \leq \sum_i \left(\sum_j |a_{ij}|\right)\left(\sum_j |x_j|^2|a_{ij}|\right) \leq \|A\|_\infty \sum_{i,j} |x_j|^2|a_{ij}| \\ &\leq \|A\|_\infty \|x\|^2 \sum_i \max_j |a_{ij}| = \|A\|_1\|A\|_\infty\|x\|^2.\end{aligned}$$

21. The equality is true for every diagonal matrix. Also, both ρ and $\|\cdot\|_2$ are unitary invariant. Since normal matrices are unitary similar to diagonal matrices, the equality holds true for every normal matrix.

22. (a) The numbers R, S are achieved ; for instance, R is the supremum of N_1 (continuous) on the sphere (a compact set) of N_2. Let thus x, y be such that

$$N_1(y) = N_2(x) = 1, \quad N_1(x) = R, \quad N_2(y) = S.$$

From Hahn-Banach Theorem, there exists a linear form ℓ on $\mathbb{C}^n$, such that $|\ell(w)| \leq N_2(w)$ for every vector, while $\ell(y) = S$. This form may be represented as $\ell(w) = z^*w$ for some vector z. Let B denote the matrix xz^*. Then $N_2(Bw) \leq N_2(w)$ for every w and $N_2(By) = S = N_2(x)$. Thus $\mathcal{N}_2(B) = 1$. On the other hand, $N_1(By) = SN_1(x) = RS$ gives $\mathcal{N}_1(B) \geq RS$. Hence (notice that, as above, the supremum is achieved)

$$\max_{A\neq 0} \frac{\mathcal{N}_1(A)}{\mathcal{N}_2(A)} \geq RS.$$

Next, for every vector w and every matrix A, we have

$$N_1(Aw) \leq RN_2(Aw) \leq R\mathcal{N}_2(A)N_2(w) \leq RSN_1(w)\mathcal{N}_2(A)$$

and therefore

$$\max_{A\neq 0} \frac{\mathcal{N}_1(A)}{\mathcal{N}_2(A)} \leq RS.$$

This shows the first equality. Exchanging the rôles of N_1 and N_2, we also have the second one.

(b) If $\mathcal{N}_1 = \mathcal{N}_2$, the previous result gives $RS = 1$. Applying the definitions of R, S, we immediately obtain that N_2/N_1 is constant.

(c) If $\mathcal{N}_1 \leq \mathcal{N}_2$, we have $RS \leq 1$. But, RS being larger than or equal to 1 by definition, we again have $RS = 1$, with the same conclusion.

23. (a) Obvious.

(b) From

$$\|Ax\|_y = \|Axy^*\| \leq \|A\|\,\|xy^*\| = \|A\|\,\|x\|_y.$$

(c) If $\|\cdot\|$ is induced and if larger than or equal to an algebra norm $\|\cdot\|_1$, our last result gives an induced norm $\mathcal{N}$, less than or equal to both. Thanks to exercise 22, $\mathcal{N} = \|\cdot\|$ and thus $\|\cdot\|_1 = \|\cdot\|$. Hence $\|\cdot\|$ is minimal.

If $\|\cdot\|$ is a minimal algebra norm, the minimality and the inequality $\mathcal{N}_y \leq \|\cdot\|$ show that $\mathcal{N}_y = \|\cdot\|$.

At last, we assume that $\mathcal{N}_y = \|\cdot\|$ holds true for every $y \neq 0$. Then $\|\cdot\|$ is an induced norm since $\mathcal{N}_y$ is such one.

24. (a) From the previous exercise, we have $\mathcal{N}_y = \|\cdot\| = \mathcal{N}_z$. Hence, Exercise 22 shows that $\|\cdot\|_y / \|\cdot\|_z$ is constant.

(b) Let y, t be given. From the previous question, there exists a non-zero constant such that, for every x, z, there holds $\|xy^*\| = c\|xt^*\|$. The desired equality follows immediately.

25. This matrix norm is unitarily invariant. Since hermitian matrices are unitary diagonalizable, we see that, up to a unitarily conjugation of M, it is enough to prove the inequality when $H = D$ is a diagonal matrix with positive real diagonal entries. Next, replacing DMD by M, we are led to proving the inequality

$$\|M\|_2 \leq \left\| \frac{1}{2}(DMD^{-1} + D^{-1}MD) \right\|_2 .$$

Last, factorizing D as a product of diagonal matrices where all but one diagonal entries equal 1, it is enough to prove the inequality when $D = \mathrm{diag}(d, 1, \ldots, 1)$. The matrix $M(d)$ in the right hand side depends only on $t := (d + 1/d)/2 \geq 1$, in the following way :

$$M(d) = \begin{pmatrix} m & tx^* \\ ty & M' \end{pmatrix} =: N(t).$$

Since $t \mapsto N(t)$ is affine, the function $\phi(t) := \|N(t)\|_2$ is convex on $\mathbb{R}^+$. But

$$N(t) \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} m \\ ty \end{pmatrix}, \quad N(t) \begin{pmatrix} 0 \\ Y \end{pmatrix} = \begin{pmatrix} x^*Y \\ M'Y \end{pmatrix},$$

show that $\phi(t) \geq \max(|m|, \|M'\|_2) = \|N(0)\|_2 = \phi(0)$. Hence the convex function ϕ is non-decreasing on $\mathbb{R}^+$. In particular, when $t = (d + 1/d)/2$ (which is not less than 1), $\|M\|_2 = \phi(1) \leq \phi(t)$ gives the desired inequality.

26. (a) Let M be an extremal point of B. From Theorem 7.7.1 (singular value decomposition), there exist two orthogonal matrices U, V and a non-negative diagonal matrix D such that $M = UDV$. Since B is invariant under the left and right multiplication by orthogonal matrices, which are linear transformations, D is an extreme point of B too. Since $\mathrm{diag}(a_1, a_2) \in B$ if and only if $|a_1|, |a_2| \leq 1$, the extremality gives $D = I_n$, which means that M is orthogonal.

Conversely, the set of extreme points of B is non-void (because B is compact, Krein-Milman Theorem) and thus contains at least one orthogonal matrix. On another hand, it is invariant under the (say, left) action of $\boldsymbol{O}_2(\mathbb{R})$, since B is so. Hence it contains $\boldsymbol{O}_2(\mathbb{R})$. We have thus proven that it equals $\boldsymbol{O}_2(\mathbb{R})$.

(b) Again, Σ is invariant under the left and right action of $\boldsymbol{O}_2(\mathbb{R})$. Hence, every point M in Σ reads PDQ where $D \in \Sigma$ is diagonal. Since $\|D\|_2 = \max_i |d_{ii}|$, we may always assume a form $D = \mathrm{diag}(a, 1)$ with $a \in [0, 1]$.

(c) A diagonal matrix $D = \mathrm{diag}(a, 1)$ belonging to Σ is such that $a \in [-1, 1]$. It is thus an element of the segment $[I_2, S]$, where S is the symmetry $\mathrm{diag}(-1, 1)$. The image of this segment through $D \mapsto PDQ$, where P, Q are given in $\boldsymbol{O}_2(\mathbb{R})$, is a segment whose one vertex is the rotation $\pm PQ$ and the other one is the symmetry $\mp PSQ$, where $\pm 1 = \det P \det Q$.

Conversely, let $[r, s]$ be a segment with $r \in \mathcal{R}$ and $s \in \mathcal{S}$. The left multiplication by r^{-1} sends it onto $[I_2, s']$, where s' is an other symmetry. Let $x \neq 0$ be such that $s'x = x$. Then every point in the segment fixes x, and therefore has norm larger than or equal to 1. The converse inequality follows from convexity. Hence $[I_2, s'] \subset \Sigma$ and similarly $[r, s] \subset \Sigma$.

(d) Since $[r', s']$ can be sent, through left and right orthogonal multiplications, onto $[I_2, S]$, we may assume that $r' = I_2$ and $s' = S$. So let $[r, s]$ contain some point of the form $D = \mathrm{diag}(a, 1)$. Then the $(2, 2)$-entry of either r or s is not less than 1. Since the rows and columns of the corresponding matrix have norm 1, this entry is actually 1 and the off-diagonal entries are zeroes. Hence it has the same diagonal form, which shows that either $r = I_2$ or $s = S$. Hence, two distinct segments have at most a vertex in common. In other words, distinct open segments do not intersect.

(e) An orthogonal symmetry has the form $s = I_2 - 2xx^T$ for some unitary vector x. Since $\|xx^T\|_2 = 1$, we deduce that $d(I_2, s) = 2$. Next, given the rotation r of angle θ, $I_2 - r$ equals $2\sin(\theta/2)$ times an other rotation. Hence $d(I_2, r) = 2|\sin(\theta/2)|$ is less than 2 except in the case $r = -I_2$. Last $S - s$ is also $2\sin(\theta/2)$ times an other symmetry, for a suitable θ. We again have $d(S, s) \leq 2$ with equality if and only if $s = -S$. Using orthogonal invariance of the distance, we see that the equality in $d(r, r') \leq 2$ or $d(s, s') \leq 2$ occurs if and only if $r' = -r$ or $s' = -s$, respectively ; likewise, there always holds $d(r, s) = 2$.

Let $M, N \in \Sigma$ be such that $d(M, N) = 2$. From question c), there exist two segments such that $M \in [r, s]$ and $N \in [r', s']$. Since we already treated the case of rotations and symmetries, we may assume that $M \in (r, s)$ and $N \in (r', s')$. Since $d(N, r), d(N, s) \leq 2$, convexity of the $d(N, \cdot)$ implies $d(N, r) = d(N, s) = 2$. Therefore, convexity of the $d(r, \cdot)$ implies $d(r, r') = d(r, s') = 2$; we obtain likewise $d(s, s') = 2$. Hence $r' = -r$ and $s' = -s$.

5 Non-negative matrices

1. (a) By assumption, $\sum_i m_{ij} = 1$ for every j, and $\max_j \sum_i |m_{ij}| \leq 1$. Obviously, this implies $m_{ij} \geq 0$.

   (b) By assumption, $\sum_i m_{ij} = 1$ for every j. Since $m_{ij} \geq 0$, this implies $\sum_i |m_{ij}| = 1$ and in particular $\|M\|_1 = 1$.

   (c) No, since 0_n satisfies **P1, P3** but not **P2**. Even the stronger assumption $\|M\|_1 = 1$, together with **P1**, do not imply **P2** ; see for instance $M := \mathrm{diag}(1, 0)$.

2. (a) In fact, K is the intersection of $\ker(A - \rho(A)I_n)$ with the simplex defined by $y \geq 0$ and $\sum_j y_j = 1$. Thus it is a compact convex set.

   (b) If $\dim \ker(A - \rho(A)I_n) \geq 2$, then let consider an affine eigenline L passing through x. Then $K \cap L$ is a segment, where x is an interior point as well as every positive points. Let y be a boundary point of $K \cap L$. Then $y \geq 0$ without being positive : it has at least a zero component. But this contradicts Lemma 5.3.2.

   (c) If the multiplicity of $\rho(A)$ were larger than 1, there would exist (from the previous result) a vector z such that $Az - \rho(A)z = x$. For sufficiently large $a \in \mathbb{R}$, the vector $y := z + ax$ is positive and satisfies$Ay - \rho(A)y = x$ too. In particular, $Ay - \rho(A)y > 0$. This contradicts the maximality of $\rho(A)$, in view of Lemma 5.3.2.

3. We already know that M is invertible (Corollary 4.5.1). Let $b \geq 0$ be given and x be the solution of $Mx = b$. Let x_i be its minimal component. Then

$$m_{ii}x_i = b_i - \sum_{j \neq i} m_{ij}x_j \geq \sum_{j \neq i} |m_{ij}|x_j \geq x_i \sum_{j \neq i} |m_{ij}|.$$

   If M is strictly diagonally dominant, we immediately obtain $x_i \geq 0$. If M is strongly diagonally dominant and irreducible, and if $x_i < 0$, then $m_{ii} = \sum_{j \neq i} |m_{ij}|$, and $m_{ij} = 0$ whenever $x_j > x_i$. Then the set I of indices i of minimal components, together with its complement J, make a partition, such that $m_{ij} = 0$ for every $(i, j) \in I \times J$. Irreducibility implies then $J = \emptyset$. Thus $m_{ii} = \sum_{j \neq i} |m_{ij}|$ for every i, which contradicts the assumption. Hence $x \geq 0$. Proposition 5.1.1 then tells that $M^{-1} \geq 0$.

4. (a) By assumption, $B + \epsilon$ admits a positive eigenvector x^ϵ associated to the spectral radius. One may assume that $\|x^\epsilon\|_1 = 1$, that is $\sum_j x_j^\epsilon = 1$. Since the eigenvalues are continuous functions of the entries, the spectral radius is, too. We may extract a converging sequence of x^ϵ as ϵ tends to zero. Its limit is a non-negative non-zero eigenvector of B, associated to the limit $\rho(B)$.

   (b) The series $\sum_k (\lambda^{-1}B)^k$ converges whenever $\rho(\lambda^{-1}B) < 1$, that is $|\lambda| > \rho(B)$, and its sum is $(I_n - \lambda^{-1}B)^{-1}$. Hence the formula. Thus the function h reads as

$$h(\lambda) = \lambda - a - \sum_0^\infty \lambda^{-k} \xi^T B^{k-1} \eta.$$

Since ξ, B and η are non-negative, each term in this expression is non-decreasing on $(\rho(B), +\infty)$. The first one being increasing, h is so. More precisely, $h' > 0$. Likewise, $x(\lambda) > 0$ (the term with $k = 1$ is positive).

(c) This is Schur's formula.

(d) Zeroes of P_A in $(\rho(B), +\infty)$ are those of h. Since h is strictly increasing and even more, its derivative is strictly positive, we only have to prove that it vanishes on $(\rho(B), +\infty)$. Also, the result does not depend on a, so that we must prove that the range of h is the whole line. Since we have (easy) $h(+\infty) = +\infty$, there remains to show that the decreasing function

$$g(\lambda) := \xi^T(\lambda I - B)^{-1}\eta.$$

tends to $+\infty$ as λ decays towards $\rho(B)$. This is the difficult part, that we are going to prove now.

First of all, up to a conjugation by a permutation matrix, we may assume that B is lower block-triangular, with at most two diagonal blocks B_0, B_1, and where $\rho(B_0) \leq \rho(B)$, while B_1 is irreducible and $\rho(B_1) = \rho(B)$. The lower off-diagonal block is denoted by B_{10}. In this decomposition, B_1 is always present, but B_0 may be absent (it is provided B itself is irreducible). We decompose η and ξ accordingly. The irreducibility of A implies the following properties :

$$\eta_0 \neq 0, \quad (\eta_1, B_{10}) \neq 0, \quad (\xi_0, B_{10}) \neq 0, \quad \xi_1 \neq 0.$$

Let now define $R(\lambda) := (\lambda I - B)^{-1}$. Using $R_i(\lambda) := (\lambda I - B_i)^{-1}$, we have

$$R(\lambda) = \begin{pmatrix} R_0(\lambda) & 0 \\ R_1(\lambda)B_{10}R_0(\lambda) & R_0(\lambda) \end{pmatrix}.$$

Then there comes

$$g(\lambda) \geq \sum_{i=0}^{1} \xi_i^T R_{ii}(\lambda)\eta_i + \xi_1^T R_1(\lambda)B_{10}R_0(\lambda)\eta_0. \tag{1}$$

We remark that, as $\lambda \to \rho(B) + 0$, there holds

$$\xi_1^T R_1(\lambda) \sim \frac{1}{\lambda - \rho(B)} Y_1^T,$$

where Y_1 is a positive eigenvector of B_1^T.

We now argue by absurdum. Assuming that g is bounded by above near $\rho(B)$, we obtain that each terms, in the sum of the right-hand of (1) are bounded. From the asymptotics above, we derive on the one hand that $\eta_1 = 0$. On the other hand, the monotonicity of $B_{10}R_0(\lambda)\eta_0$ implies that the function$\lambda \mapsto \eta_1^T R_1(\lambda)B_{10}R_0(\sigma)\eta_0$ is

bounded on $(\rho(B), +\infty)$ for every parameter $\sigma > \rho(B)$. Using again the asymptotics above, we derive the identity $B_{10}R_0(\lambda)\eta_0 = 0$ for all $\lambda > \rho(B)$. In other words,

$$B_{10}B_0^k\eta_0 = 0, \quad \forall k \geq 0. \tag{2}$$

We now see that, thanks to (2), and because of $\eta_1 = 0$, the lower left block in the block decomposition of every power of A vanishes. In particular, the lower left block of $(I + A)^{n-1}$ vanishes, contradicting Proposition 5.1.2 and therefore the irreducibility of A. Hence g is unbounded near $\rho(B)$, meaning that $H(\rho(B)) = -\infty$. Now, applying the intermediate value Theorem, we conclude that h vanishes at some point in $(\rho(B), +\infty)$.

(e) Since A^T is non-negative and irreducible and has λ_0 in its spectrum, and since $\rho(B^T) = \rho(B)$, this eigenvalue is the only one of A^T in the interval $(\rho(B), +\infty)$, and is associated to a positive eigenvector ℓ. We therefore have $l^T(A - \lambda_0 I_n) = 0$.

(f) From $AX = \mu X$ and triangle inequality, we obtain $(A - |\mu|I_n)|X| \geq 0$. Multiplying by ℓ^T gives $(\lambda_0 - |\mu|)\ell^T|X| \geq 0$. Since X is non-zero and $\ell > 0$, we have $\ell^T|X| > 0$, hence $\lambda_0 \geq |\mu|$, which means that $\lambda_0 = \rho(A)$.

5. (a) For $0 < h \max_i a_{ii} < 1$, the matrix $I_n - hA$ is strictly diagonally dominant and its off-diagonal entries are non-positive. From Exercise 3, it is non-singular and its inverse is non-negative.

(b) Let $t > 0$ be given. For m large enough, $R(t/m; A) \geq 0$ and thus $R(t/m; A)^m \geq 0$. Passing to the limit, Trotter's formula gives $\exp tA \geq 0$.

(c) Obvious from the last result.

(d) Since the spectrum of $\exp tA$ is the image of that of A under the map $\mu \mapsto e^{t\mu}$, the spectral radius of $\exp tA$ is the number $e^{t\sigma}$. Applying Perron-Frobenius to $\exp tA$, we obtain that $t\sigma + 2ik\pi$ is an eigenvalue of tA for a suitable integer k. In other words, $\sigma + 2ik\pi/t \in \mathrm{Sp}(A)$. Letting t varying, and using the finiteness of $\mathrm{Sp}(A)$, we have easily $\sigma \in \mathrm{Sp}(A)$.

6. (a) From

$$|\lambda + \tau|^2 = \tau^2 + 2\tau\Re\lambda + |\lambda|^2,$$

we see that $|\lambda+\tau|$ is the largest, for large values of τ, when λ is one of the eigenvalues of maximal real part, and when $|\lambda|$ is maximal among these. Hence $\rho(A + \tau I_n) = |\mu + \tau|$ for large $\tau > 0$.

(b) We weak form of Perron-Frobenius' Theorem (Theorem 5.2.1) tells us that $|\mu + \tau|$ is an eigenvalue of $A + \tau I_n$, for $\tau > 0$ large enough. Hence $|\mu + \tau| - \tau \in \mathrm{Sp}A$. Since

$$|\mu + \tau| - \tau = \Re\mu + \frac{1}{2\tau}(\Im\mu)^2 + \mathcal{O}\left(\frac{1}{\tau^2}\right),$$

we deduce that $|\mu + \tau| = \tau + \Re\mu$ for τ large enough and thus $\Im\mu = 0$. Therefore $\mu = \sigma$. In other words, $\sigma \in \mathrm{Sp}A$. From its definition, σ is not less than $\rho(A)$. Since $\rho(A)$ is the largest real eigenvalue (from Perron-Frobenius' Theorem), and σ is itself an eigenvalue of A, we deduce that $\sigma = \rho(A)$.

7. For $\tau > 0$ large enough, we have $B + \tau I_n > 0$. By Perron-Frobenius, we deduce that B admits a real eigenvalue μ, associated to a positive eigenvector X. Let us define $D := \mathrm{diag}(x_1, \ldots, x_n)$ and $B' := D^{-1}BD$. Then $B'e = \mu e$, where $e = (1, \ldots, 1)^T$. By assumption, $\mu < 0$. Now, there holds

$$\sum_j b'_{ij} = \mu < 0, \qquad \forall i.$$

8. (a) Let $(X, Y)^T$ be a positive eigenvector, associated to λ. There holds $(B - \lambda)X = 0$ and $(B - \lambda)Y = -X$. For small positive a, the vector $Z := X - aY$ is positive, while $(B - \lambda)Z = aX > 0$. Thus there is a μ, larger than λ, such that $BZ \geq \mu Z$. Following the proof of Theorem 5.2.1, we obtain $\rho(B) \geq \mu$, that is $\rho(B) > \lambda$.

   (b) Perron-Frobenius' Theorem tells that $\rho(A)$ (which is nothing but $\rho(B)$), is an eigenvalue of A^T, associated to a non-negative eigenvector ℓ. Then

$$0 = \ell^T(A - \lambda I_n)\begin{pmatrix} X \\ Y \end{pmatrix} = (\rho(B) - \lambda)\ell^T \begin{pmatrix} X \\ Y \end{pmatrix}$$

   yields a contradiction, since both factors are positive.

9. (a) Let $\mu_1, \ldots, \mu_p$ these simple eigenvalues of modulus 1, and $\mathbb{C}x^j$ the corresponding eigenspaces. Denoting by ℓ^j the corresponding eigenvectors of B^*, the orthogonal F to $\{\ell^1, \ldots, \ell^p\}$ is invariant under B, and $\mathbb{C}^n$ is the direct sum of F and $G := \oplus_j \mathbb{C}x^j$. The restriction of B^m to G is defined by

$$B^m \sum_j a_j x^j = \sum_j \mu_j^m a_j x^j,$$

   and is therefore bounded. Since the spectral radius of the restriction of B to F is less than one, Householder's Theorem implies that its m-th power tends to zero as $m \to +\infty$. Hence B^m remains bounded as $m \to +\infty$.

   (b) i. We split $\mathbb{R}^n = \mathbb{R}x \oplus y^\perp$, where both factors are invariant subspaces for M. Since they are also invariant for L, they are for B as well. On $\mathbb{R}x$, B vanishes, while it equals to M on $y^\perp$. Since 1 is a simple eigenvalue of M, it is not an eigenvalue of its restriction to $y^\perp$. Thus $B - I_n$ is invertible.

   ii. From Theorem 5.4.1, eigenvalues of M of modulus 1, hence those of B, are simple. Question a) then tells us that B^m remains bounded. Last, one verifies easily that $BL = LB = 0$ and $L^2 = L$. Therefore, the binomial formula holds true and gives $M^m = B^m + L$.

   iii. We have

$$\frac{1}{N}\sum_{m=0}^{N-1} M^m = L + \frac{1}{N}(B - I_n)^{-1}(B^N - I_n).$$

   The right-hand side tends to L since B^N remains bounded.

iv. The property is equivalent to $B^N \to 0_n$, which amounts to saying that $\rho(B) < 1$. In other words, it holds true if and only if 1 is the only eigenvalue of M of that modulus.

10. (a) This is a direct application of Lemma 5.3.3.

    (b) Notice that X_t belongs to a compact set and therefore has at least a cluster point X, a unitary non-negative vector. Passing to the limit in $BX_t \leq (B + tC)X_t = r_t X_t$, we obtain $BX \leq rX$. If r is finite, this implies $B'Y \leq 0$, where B' is the lower-left block of B. Since $B \geq 0$ and $Y > 0$, this gives $B' = 0$.

    (c) In other words, B is block-triangular. Since B is irreducible, this block-triangular form is trivial, that is $X = Y$, or equivalently $X > 0$.

    (d) Likewise, passing to the limit in $CX_t \leq (r_t/t)X_t$ gives $CX \leq 0$. Since $C \geq 0$ and $X > 0$, we derive $C = 0$. This contradicts the assumption. Therefore $r = +\infty$.

    (e) Apply the intermediate value Theorem.

11. The stability of Δ under multiplication is obvious. Therefore, if M is bi-stochastic, its powers belong to the compact set Δ, and thus remain bounded.

12. Form Perron-Frobenius' Theorem, $\rho(M)$ is the only eigenvalue associated to a positive eigenvector. Since $Me = 1e$, we conclude that $\rho(M) = 1$. With the notations of Exercise 9, there hold $x = e$, $\ell = n^{-1}x^T$ and the desired equality. The case

$$M = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

    shows that M^m does not always converge.

13. Since $J_n = n^{-1}e^T e$, Hölder inequality gives $\|J_n\|_p \leq n^{-1}\|e\|_p\|e\|_{p'}$. Applying J_n to e, we see that both numbers are actually equal. Since $\|e\|_q = n^{1/q}$ for every q, we obtain $\|J_n\|_p = n^{1/p+1/p'-1} = 1$.

14. Since both P and P^{-1} send the simplex K_n into itself, we see that $P(K_n) = K_n$, and P is a bijection. Let i be an index. Then there exists x in K_n such that $Px = \vec{e^i}$. From Proposition 5.5.1, we have $o(x) \geq n-1$, which means that x is itself a $\vec{e^j}$. We act similarly with P^{-1} and conclude that P is a permutation matrix.

15. Let $i_1, j_1, i_2, \ldots$ be such a chain (we shall say an *admissible* chain), and let I_l be the set in the canonical decomposition of M, to which i_1 belongs. Then $j_1 \in J_l$, which in turns implies $i_2 \in I_l$. By induction, all i_k's belong to I_l. Thus the equivalence class I of some $i \in I_l$ is contained in I_l, which turns out to be a union of such classes. On the other hand, let J be the set of indices j appearing in such chains which originate from some i' in I. If $j \in J$ and $i'' \in I^c$, there exists an i' in I and an admissible chain $i', \ldots, j$. However, the chain $i', \ldots, j, i''$ is not admissible and therefore we have $m_{i''j} = 0$. Likewise, we have $m_{ij} = 0$ whenever $i \in I$ and $j \in J^c$. The minimality of the canonical decomposition then implies that $I = I_l$.

We apply this result to the matrix

$$M := \begin{pmatrix} 1/2 & 1/2 & 0 & \cdots & 0 \\ 1/2 & 0 & 1/2 & \ddots & \vdots \\ 0 & 1/2 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 0 & 1/2 \\ 0 & \cdots & 0 & 1/2 & 1/2 \end{pmatrix}.$$

The chain $i_1 = 1, j_1 = 2, 3, 4, \ldots$ is admissible and therefore $(2k+1)\mathcal{R}1$ for every k. The chain $1, 1, 2, 3, 4, \ldots$ is also admissible, so that $(2k)\mathcal{R}1$ for every k. Hence the coset of 1 is $\{1, \ldots, n\}$. Therefore the canonical decomposition is trivial and $M \in \boldsymbol{S\Delta}_n$.

16. If $x \in \partial K_n$, then $o(M'Mx) \leq o(Mx) < o(x)$. Hence $M'M \in \boldsymbol{S\Delta}_n$. If $o(M'x) < o(x)$, then $o(MM'x) \leq o(M'x) < o(x)$. Otherwise, $M'x \in \partial K_n$ and we have $o(MM'x) < o(M'x) = o(x)$. In both cases, there holds $o(MM'x) < o(x)$, thus $MM' \in \boldsymbol{S\Delta}_n$.

17. Let K be the image of K_n under M^{n-1}. Since $x \in \partial K_n$ implies $o(Mx) < o(x)$, K is contained in the interior (namely those points with $o(x) = 0$) of K_n. Under the translation $x \mapsto x - e$, we obtain two convex compact subsets K_1, K_2 of the hyperplane $H := \{x\,;\, \sum_i x_i = 0\}$, the latter being in the interior of the former. Hence there exists a number $\mu < 1$ such that $K_2 \subset \mu K_1$. Since K_1 is the unit ball of H (for the norm $\|\cdot\|_1$), we conclude that the norm of the restriction of M^{n-1} to H is strictly less than 1, and the sequences of its powers tends to 0. This implies that the N-th power of the restriction of M to H tends to 0 as well. At last, $\mathbb{R}^n = H \oplus \mathbb{R}e$ and $M^N e = e$ for every N. Thus M^N converges towards J_n, defined in exercise 12.

18. (a) Obviously, $\|M\|_1 = \|M\|_\infty = 1$.

    (b) Use $\|M\| \geq \rho(M)$ and $Me = e$.

    (c) At first sight, it differs from Corollary 5.5.1 in that this one applies to norms induced on $\boldsymbol{M}_n(\mathbb{R})$ by norms of $\mathbb{R}^n$, while here the norm is the restriction to $\boldsymbol{M}_n(\mathbb{R})$ of an induced norm of $\boldsymbol{M}_n(\mathbb{C})$. However, in view of Exercise 6, Section 4, where we showed that both induced norms $\|\cdot\|_p$ are equal, the present result is a consequence of Corollary 5.5.1.

19. (a) Let us differentiate the identity $(A + tB - \lambda_j I_n)X_j = 0$:

$$(B - \lambda_j' I_n)X_j + (A + tB - \lambda_j I_n)X_j' = 0.$$

    Multiplying by X_j^T kills the second term. Since X_j is unitary, there comes the desired identity.

    (b) We have $\alpha_j = \lambda_j(0)$, $\gamma_j = \lambda_j(1)$. Therefore,

$$\gamma_j - \alpha_j = \int_0^1 \lambda_j'(t)dt = \int_0^1 (BX_j, X_j)dt.$$

(c) For every pair of orthonormal bases (W, Z), the matrix M defined by

$$m_{ij} = |(W_i, Z_j)|^2$$

is bi-stochastic (actually, it is orthostochastic). We apply this remark to the bases $W = X(t)$ and $Z = Y$, and then integrate from 0 to 1 using the convexity of $\mathbf{\Delta}_n$.

(d) Since $(BX_j(t), X_j(t)) = \sum_k \beta_k |(X_j(t), Y_k)|^2$, we immediately obtain $\gamma - \alpha = \Sigma\beta$. Thanks to Kirchhoff's Theorem, Σ may be written as a barycenter of permutation matrices. Hence $\Sigma\beta$ is a barycenter of the vectors obtained from β by permutations of its coordinates.

20. (a) Since $C(a)$ is defined by large linear inequalities, it is closed and convex. Since every b in $C(a)$ satisfies $a_1 \leq \min_j b_j$ and $\max_j b_j \leq a_1$, this set is also bounded, thus compact.

Let $b \in C(a)$ be non-decreasing, with $b_j < b_{j+1}$ for some index j. Then $b^{\pm} \in C(a)$, where $b^{\pm}$ is obtained from b via the two perturbations $b_j^{\pm} = b_j \pm \epsilon$, $b_{j+1}^{\pm} = b_{j+1} \mp \epsilon$ (notice that $j < n$), with $\epsilon > 0$ small enough. Hence b is not extremal. We deduce that if b is extremal in $C(a)$, then $s_j(a) = s_j(b)$ for every j. In other words, b is a permutation a^σ of a : $b_j = a_{\sigma(j)}$.

Conversely, let $b \in C(a)$ be such that $s_j(a) = s_j(b)$ for every j. Assume that $b = (b' + b'')/2$ with $b', b'' \in C(a)$. Clearly, each s_j is a concave function, and we have

$$s_j(b'), s_j(b'') \geq s_j(a) = s_j(b),$$

hence $s_j(b') = s_j(b'') = s_j(a)$. We may assume that b is non-decreasing. Then $b_1 = a_1 = s_1(b') \leq b'_1$ and likewise $b_1 \leq b''_1$; this implies $b'_1 = b''_1 = b_1$. Working by induction over j, we find $b' = b'' = b$. Therefore b is extremal in $C(a)$.

(b) We use a formula that is proved in the same way as Theorem 3.3.2. If M is hermitian, with spectrum $\mu_1 \leq \cdots \leq \mu_n$, then, for every k,

$$\phi_k(M) := \mu_1 + \cdots + \mu_k = \min_{\dim F = k} \mathrm{Tr}\,(M|_F)\,.$$

Hence, ϕ_k is concave, from which we immediately have

$$(\theta \in [0,1],\ \phi_k(M), \phi_k(N) \geq s_k(a)) \Rightarrow \phi_k((1-\theta)M + \theta N) \geq s_k(a).$$

Since moreover ϕ_n is linear, we deduce that $Y(a)$ is convex. Obviously, it is closed. Last, $M \in Y(a)$ implies $\mu_1 \geq a_1$ and $\mu_n \leq a_n$ and thus $\|M\|_2 = \rho(M)$ (always true for hermitian matrices) is bounded by $\max(-a_1, a_n)$. Hence, $Y(a)$ is compact.

We notice that $Y(a)$ is invariant under orthogonal conjugation. Therefore, its extremal subset $\mathrm{ext}(Y(a))$ has the same invariance property. Let M be an extreme point of $Y(a)$. It is orthogonaly similar to a diagonal matrix D, which is therefore extremal. And conversely, extremality of D implies extremality of all $M \in \boldsymbol{Sym}_n(\mathbb{R})$

with $\mathrm{Sp}M = \{d_1, \ldots, d_n\}$. Since the subset of diagonal matrices in $Y(a)$ is isomorphic to $C(a)$, extremality of $D = \mathrm{diag}(d_1, \ldots, d_n)$ implies that of $d = (d_1, \ldots, d_n)$ in $C(a)$, that is $d = a^\sigma$, for some permutation σ. Since $\mathrm{ext}(Y(a))$ is non-void, by Krein-Milman's Theorem, we conclude that there exists at least one extremal matrix of the form $\mathrm{diag}(a^\sigma)$. By orthogonal invariance, every such diagonal matrix is extremal. Finally, the set of extremal points of $Y(a)$ consists in symmetric matrices whose spectrum equals $\{a_1, \ldots, a_n\}$, with consistent multiplicities. This set is denoted by $X(a)$ in the next question.

(c) Just apply Krein-Milman's Theorem.

(d) We may assume that a be non-decreasing. Then $ks_n(a) - ns_k(a) = k(a_{k+1} + \cdots + a_n) - (n-k)(a_1 + \cdots + a_k) \geq k(n-k)(a_{k+1} - a_k) \geq 0$. Hence $a' \in C(a)$. Let $b \in C(a)$ be given. The same inequality, applied to b, gives

$$s_k(b) \leq \frac{k}{n} s_n(b) = \frac{k}{n} s_n(a) = s_k(a').$$

Since also $s_n(b) = s_n(a) = s_n(a')$, we conclude that $b \prec a'$.

(e) If $\mathrm{Sp}M \prec a'$, then $\mathrm{Tr}M = s_n(a)$. Conversely, let $M \in \boldsymbol{Sym}_n(\mathbb{R})$ satisfy $\mathrm{Tr}M = s_n(a)$. Then let $b := \mathrm{Sp}M$; we have $b \prec b' = a'$. Hence the set consists precisely in those symmetric matrices whose trace equals $s_n(a)$.

6 Matrices with entries in a principal domain ; Jordan's reduction

1. Let A be a principal domain. We say that a is *prime* if $a = bc$ implies that either b or c is invertible, and a is not itself invertible (this definition is slightly different from the usual one, but makes our proof easier). It amounts to saying that (a) is a maximal ideal (because A is principal) and $(a) \neq A$. If p is prime, then $p = ab$ implies that p divides either a or b, for otherwise Bézout identities for the pairs (p, a) and (p, b) give immediately the false conclusion $1 \in (p)$. More generaly, if the prime p divides $ab \cdots z$, then p divides one of the factors.

   Let $x \neq 0$ belong to A. If x is not prime, then it admits a strict divisor x_1, that is a divisor neither invertible, nor associated to x. Because A is Nœtherian, maximal sequences $x_0 = x, x_1, \ldots$, where each term is a strict divisor of the previous one, are finite ; obviously, the last term of such a sequence is prime. Therefore x admits a prime divisor p. Let now y_1 denote x/p (recall that our rings are integral). By induction, we find at least one maximal sequence $y_0 = x, y_1, \ldots$, where each term is a strict divisor of the previous one, and the quotient is prime. Again, this sequence must be finite, and the last term y_r must be prime. Denoting by q_j the quotient y_j/y_{j+1}, and $q_r = y_r$, we obtain $x = \prod_{j=1}^r q_r$: x is a product of prime elements.

   Let now

$$x = p_1 \cdots p_s = q_1 \cdots q_r$$

be two factorizations in prime elements. Then p_1 divides one of the q_j's, say the first one. But since q_1 is prime and p_1 is not invertible, the quotient q_1/p_1 must be invertible. Hence $p_2 \cdots p_r = uq_2 \cdots q_s$. Similarly, p_2 is associated to one of the remaining q_j's, say q_2 (it cannot divide the unit u). By induction, we obtain that $r = s$ and each p_j's is associated to $q_{\phi(j)}$, for some bijection ϕ.

2. Let us write in a unique way $P(X) = XQ(X) + a_n$. Let expand $\det(XI_n - B_P)$ with respect to the first column. One obtains $X \det(XI_{n-1} - B_Q) + \Delta$, where Δ is a determinant of size $n-1$ whose first line is $(0, \ldots, 0, a_n)$. We expand Δ with respect to this line and obtain $(-1)^n a_n \delta$. Now, δ is the determinant of an $(n-2) \times (n-2)$ triangular matrix whose diagonal is $(-1, \ldots, -1)$. Hence $\delta = (-1)^{n-2}$. Finally, $\det(XI_n - B_P) = X \det(XI_{n-1} - B_Q) + a_n$. An induction over n gives the result.

3. Since M and M^T are similar, they have same rank. Also, we already now that the rank of AB is not larger than that of A or that of B. Hence the rank of $M^T M$ is less than or equal to that of M. If $k = \mathbb{R}$, $M^T Mx = 0$ implies $x^T M^T Mx = 0$, that is $\|Mx\|_2 = 0$, or $Mx = 0$. Therefore $\ker M = \ker M^T M$, which shows that M and $M^T M$ have same rank $n - \dim \ker M$. This property is false for $k = \mathbb{C}$, as shown by the following example

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

4. Well, that's difficult to do by hand. You should use your favorite linear algebra software package.

5. Let M be any invertible matrix whose first row is X^T, and N be any invertible matrix whose first column is Y. Then

$$MAN = \begin{pmatrix} 1 & v^T \\ u & A' \end{pmatrix}.$$

Multiplying at left and right by

$$\begin{pmatrix} 1 & 0^T \\ -u & I_{n-1} \end{pmatrix}, \quad \begin{pmatrix} 1 & -v^T \\ 0 & I_{n-1} \end{pmatrix}$$

respectively, we obtain $M_1 A N_1 = \mathrm{diag}(1, A_1)$, where M_1, N_1 are invertible, and the first row of M_1 is X^T, the first column of N_1 is Y. Since the reduction of $\mathrm{diag}(1, A_1)$ to the canonical diagonal form involves only invertible matrices of the form $\mathrm{diag}(1, R)$, we find that P may be taken as $\mathrm{diag}(1, M_2)M_1$, whose first row is X^T. And likewise for Q.

We deduce that $PAY = (1, 0, \ldots, 0)^T$ and $X^T AQ = (1, 0, \ldots, 0)$. There follows $PBQ = \mathrm{diag}(0, I_{r-1}, 0_{n-r})$, thus the rank of B is $r - 1$.

Likewise, if $X^T AY = I_m$, we may choose P and Q in such a way that

$$PAQ = \mathrm{diag}(I_r, 0_{n-r}) \text{ and } PAY = \mathrm{diag}(I_m, 0_{m \times (n-m)})^T,$$

$X^TAQ = \mathrm{diag}(I_m, 0_{(n-m)\times m})$. Hence, defining $B := A - (AY)(X^TA)$, we have $PBQ = \mathrm{diag}(0_m, I_{r-m}, 0_{n-m})$, so that the rank of B is $r - m$.

If X^TAY is only invertible, we apply the last result to the pair (X, Y'), where $Y' := Y(X^TAY)^{-1}$.

6. (a)
$$\frac{dy}{dt} = PAP^{-1}y.$$

   (b) There exists an invertible P such that $PAP^{-1} =: B$ has a canonical Jordan form. To the elementary divisor $(X-a)^m$, there corresponds a block $J(a;m)$ in B. We may assume that $B = \mathrm{diag}(J(b;m), \ldots)$. Then (6.1) admits solutions of the form $x = (z(t), 0)^T$, where z is any solution of
$$\frac{dz}{dt} = J(a;m)z.$$
   Writing $z = q\exp(at)$, this amounts to solving
$$\frac{dq}{dt} = J(0;m)q,$$
   whose general solution is $q(t) = (Q(t), Q'(t), \ldots, Q^{m-1}(t))^T$, where Q is any polynomial of degree less than m. Choosing $Q(t) = t^k$ solves the question.

7. (a) The characteristic polynomial of M is P. From Section 6.3.1, the similarity invariants of M are $1, \ldots, 1, P$. Therefore, its elementary divisors are the monomials $(X-a)^{n_a}$, each one with multiplicity 1. According to Theorem 6.3.7, the Jordan form of M is $\mathrm{diag}(J(a_1;n_1), \ldots, J(a_r;n_r))$, where $a_1, \ldots, a_r$ are the distinct roots of P and n_j are their respective multiplicities.

   (b) Define $X := (x, x', \ldots, x^{(n-1)})$. Then X solves the differential equation $X' = M^TX$. We conclude with the previous exercise.

8. With the notations of the previous exercise, and $U_m := (u_m, u_{m+1}, \ldots, u_{m+n-1})$, we have $U_{m+1} = MU_m$. If PMP^{-1} is a Jordan form of M, then $V_m := PU_m$ solves $V_{m+1} = PMP^{-1}V_m$, which decomposes as decoupled inductions of the form $W_{m+1} = J(a;n_a)W_m$. Hence $W_m = (aI_{n_a} + N)^mW_0$, where N is nilpotent. Therefore
$$(aI_{n_a} + N)^m = \sum_{k=0}^{n_a-1} \binom{m}{k} a^{m-k}N^k.$$
   This is of the form $a^mQ(m)$, where Q is a polynomial of degree $n_a - 1$. Coming back to u, we obtain that every solution is a linear combination of such $a^mR(m)$. Since the set of such linear combinations is a vector space of dimension $\sum_a n_a = n$, and contains the set of solutions of the linear induction, itself of dimension n, both are equal.

9. (a) One may choose (other choices are possible) $x_j = j$, $y_j = j-1$. Since $M = XX^T - YY^T$, the rank of M is at most 2. Since X and Y are not colinear (or as well, M admits non null 2×2 minors), the rank is exactly 2.

(b) Because the rank of M is 2, the list of invariant factors is $(p, q, 0, \dots, 0)$, where $q \neq 0$ and $p|q$. Substracting the first row to the others give the equivalent matrix

$$M_1 = \begin{pmatrix} 1 & 2 & \cdots & n \\ 1 & 1 & \cdots & 1 \\ \vdots & & & \vdots \\ n-1 & n-1 & \cdots & n-1 \end{pmatrix}.$$

Substracting $k-1$ times the second row to the k-th one (if $k \geq 3$) and once to the first one yields

$$M_2 = \begin{pmatrix} 0 & 1 & 2 & \cdots & n-1 \\ 1 & 1 & 1 & \cdots & n-1 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & & \cdots & & \vdots \end{pmatrix}.$$

Substracting the second column to the over ones yields

$$M_3 = \begin{pmatrix} -1 & 1 & 1 & \cdots & n-2 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & & \cdots & & \vdots \end{pmatrix}.$$

Adding $k-2$ times the first column to the k-th one (if $k \geq 3$), and once to the second column gives at last $M_4 = \mathrm{diag}(-1, 1, 0, \dots, 0)$. Hence, four step suffice, and $p = q = 1$.

10. (a)

$$NB = \begin{pmatrix} 0 & \cdots & 0 & 1 & 0 \\ \vdots & \ddots & \ddots & & \\ 0 & \ddots & \ddots & & \vdots \\ 1 & & & & \\ 0 & & \cdots & & 0 \end{pmatrix}, \quad BN = \begin{pmatrix} 0 & & \cdots & & 0 \\ & & & & 1 \\ \vdots & & \ddots & \ddots & 0 \\ & & \ddots & \ddots & \vdots \\ 0 & 1 & 0 & \cdots & 0 \end{pmatrix}, \quad BNB = N^T.$$

Since B is real symmetric, $S^*S = (I + B^2)/2 = I$. Thus S is unitary.

(b) The displayed matrix is $\frac{1}{2}(N + BNB + i(BN - NB)) = SNS^* = SNS^{-1}$, which is similar to N.

(c) We have shown that every Jordan block is similar to a complex symmetric matrix. Since every complex matrix is similar to a blockwise diagonal matrix with Jordan diagonal blocks, we deduce that every complex matrix is similar to a symmetric matrix. Obviously, this differs from the real case, where only the diagonalizable matrices (with real eigenvalues) are similar to symmetric matrices.

7 Exponential of a matrix, polar decomposition and classical groups

1. The same argument as for other proofs of similar results : $H \mapsto H^2$ is continuous on $\boldsymbol{HPD}_n$, bijective and the preimage of bounded sets are bounded sets. The latter property comes from $\|H\|_2 = \rho(H)$ and $\rho(H^2) = \rho(H)^2$.

2. The set of matrices of the form $Q(M)$, where Q runs over $k(X)$, is a subspace, hence a closed set. Since it contains all finite sums

$$\sum_0^m \frac{1}{k!} M^k,$$

it contains their limit $\exp M$.

If P were the same for every matrix, it would not be constant. Then, choosing a matrix such that $P(M) = 0_n$ (that is always possible), we should have $\exp M = 0_n$, which contradicts the fact that $\exp M$ is invertible.

3. One has $p'_{ij} = p_{i,j+1}$, so that $P' = LP$ with $L = J(0;\infty)$ the generalized Jordan matrix, $l_{ij} = \delta_i^{j-1}$.

4. (a) The rank of P is an integer-valued continuous function, hence is constant.

   (b) We differentiate and obtain $PP' + P'P = P'$. Multiplying by P at right, for instance, there comes $PP'P = P'P - P'P^2$, thus $PP'P = 0$.

   (c) We have $[Q, P] = P'P - 2PP'P + PP' = P'P + PP' = P'$.

   (d) Since Q is continuous, the Cauchy problem for U admits a unique solution. Let us compute

$$(PU)' = P'U + PU' = (P' + PQ)U = QPU.$$

   Thus $PU - UP(t_0)$ solves the Cauchy problem $Y' = QY$, with the initial data 0_n. Hence it vanishes identically : $P(t) = U(t)P(t_0)U(t)^{-1}$. This shows that, given a $\mathcal{C}^1$ family $E(t)$ of subspaces of $\mathbb{R}^n$ (or $\mathbb{C}^n$ as well), it is possible to choose $\mathcal{C}^1$ vector fields $X_j(t)$ such that $(X_1(t), \ldots, X_n(t))$ be a basis of $E(t)$. Actually, the regularity of the vector fields is the same as that of $E(t)$.

5. Let $J_p = \mathrm{diag}(1, \ldots, 1, 0, \ldots, 0)$ denote the standard projector of rank p. Since every projector is diagonalisable, a projector P of rank p has the form $P = QJ_pQ^{-1}$ where $Q \in \boldsymbol{GL}_n(\mathbb{R})$. Up to the change $Q \mapsto Q$, one may assume $Q \in \boldsymbol{GL}_n(\mathbb{R})^+$. Use now the connectedness of $\boldsymbol{GL}_n(\mathbb{R})^+$.

6. (a) Let H denote $\sqrt{A}$. Then $AB = H^2B = H(HBH)H^{-1}$ is similar to, and thus has same spectrum as, HBH. Since HBH is hermitian, it is diagonalizable with real eigenvalues ; hence AB has the same properties. Also, the eigenvalues of AB, namely those of HBH, have same signs as those of B, since HBH and B represent the same hermitian forms in different bases.

(b) Up to a transposition, we may assume that C is positive definite. Then let us define $E := C^{-1}$ and $D = ABC$. We have $AB = DE$, where all matrices are hermitians and three of them are positive definite. Without loss of generality, we assume that B, D, E are positive definite. From the previous question, we now that eigenvalues of DE (therefore also those of AB) are real and positive. But the signs of the eigenvalues of A must be the same as those of the eigenvalues of AB. Hence the eigenvalues of A are positive : A is positive definite.

7. If $HQ = QH$, then $M^* = Q^{-1}H = HQ^{-1}$, from which it follows $M^*M = H^2 = MM^*$. Conversely, let us assume that M be normal. Then M is unitary diagonalizable : $M = U^*DU$. Denoting $|D| := \mathrm{diag}(|d_1|, \ldots, |d_n|)$ and $V := D|D|^{-1} = |D|^{-1}D$, we have $H = U^*|D|U$ and $Q = U^*VU$ which commute obviously.

8. (a) We have $W(F) = W(H)$, hence w is the restriction of W to $\boldsymbol{SPD}_n(\mathbb{R})$.

   (b) From the assumption, we have $w(Q^THQ) = W(Q^THQ) = W(H) = w(H)$. Therefore w is constant on the classes of orthogonal conjugation in $\boldsymbol{SPD}_n$. Thus is depends only on the eigenvalues of H (the singular values of F), that is on the coefficients of its characteristic polynomial.

9. (a) The map $U \mapsto \|A - U\|$ is continous on the compact set $\boldsymbol{U}_n$, hence achieves its lower bound at some matrix Q. Unitary invariance of Schur's norm shows that I_n minimizes $\|S - U\|$ over $\boldsymbol{U}_n$.

   (b) The hermitian adjoint of $U(t) := \exp(itH)$ is the exponential of $(itH)^* = -itH$, hence it is $U(t)^{-1}$. Thus $U(t)$ is unitary. Form the previous question, and since the square of Schur's norm is $\mathcal{C}^1$, we must have

$$\left.\frac{d\|S - U(t)\|^2}{dt}\right|_{t=0} = 0.$$

   This reads $\Re\mathrm{Tr}((S - I_n)(iH) = 0$. In other words, $S - I_n$ is orthogonal to every skew-hermitian matrix. Therefore it is hermitian, that is $S \in \boldsymbol{H}_n$.

   (c) If $S = V^*DV$ with $V \in \boldsymbol{U}_n$, then

$$\|D - I_n\| = \|S - I_n\| \leq \|S - U\| = \|D - VUV^*\| = \|DU' - I_n\|$$

   with $U' = VU^*V^*$, which runs over $\boldsymbol{U}_n$ with U. Chosing $U' = D^{-1}|D|$ ($|D|$ the "module" od D as in Chapter 5), we obtain $\|D - I_n\| \leq \| \, |D| - I_n\|$, which immediately gives $D = |D|$. Hence the eigenvalues of S are real non-negative numbers : S is positive semi-definite.

   (d) If A is invertible, then so is S, which turns out to be positive definite. Since $A = QS$, we have the (left) polar decomposition.

   (e) If $H \in \boldsymbol{HPD}_n$, the polar decomposition writes $H = I_nH$. Therefore $Q = I_n$ minimizes $\|H - U\|$ over $\boldsymbol{U}_n$: there holds $\|H - I_n\| \leq \|H - U\|$ for every unitary U. The equality holds only if $U = I_n$, since the polar decomposition is unique.

(f) This is obtained by passing to the limit in the previous question, since H belongs to the closure of $\boldsymbol{HPD}_n$.

10. (a) The map $h \mapsto R(h;A) - \exp(hA)$ is analytic in the disc $D(0;1/\rho(A))$ and its Taylor series begins with $h^2A^2/2$. Hence the map $h \mapsto h^{-2}(R(h;A) - \exp(hA))$ is analytic, thus bounded on the compact set $\overline{D(0;r)}$.

(b) Each term B^lC^{m-1}, but the first and the last ones, appears exactly twice, once with coefficient $+1$, once with coefficient -1. Thus they cancel. There remains C^m, with coefficient $+1$, and B^m, with coefficient -1.

We apply the formula to $C = R(h;A)$ and $B = e^{hA}$. There comes

$$\|C^m - B^m\| \leq \|C - B\|(\|C\|^{m-1} + \cdots + \|B\|^{m-1}) \leq c_0h^2m(\max(\|B\|, \|C\|))^m.$$

Since $\max(\|B\|, \|C\|) \leq 1 + c_2|h|$ in the disc, we obtain

$$\|R(h;A) - e^{hA}\| \leq c_0m|h|^2(1 + c_2|h|)^m \leq c_0m|h|^2e^{c_2m|h|}.$$

(c) When $h = t/m$ with fixed and $m \to +\infty$, we may apply the previous inequality when m is large enough. The right-hand side is a constant times $|h|$ and tends to zero. Hence $R(t/m;A)^m$ tends to $\exp(tA)$.

11. (a) Multiplying by e^{-b}, we may assume that $b = 0$, thus $a = 1$. We look for N nilpotent, such that $\exp N = I_r + K$, where given K is itself nilpotent. We use the $\log(1+x)$ series to define

$$N = K - \frac{1}{2}K^2 + \cdots + (-1)^rK^{r-1}.$$

Computing $\exp N - I_r - K$, we find a convergent series in K, whose term of lower degree has degree r, thus is zero (because $K^r = 0$). Hence $\exp N = I_r + K$.

(b) The range of exp is stable under conjugation, thus it is enough to determine the Jordan forms which belong to it. From the previous question, every invertible Jordan form is an exponential. Hence every invertible matrix is an exponential. Conversely, we know that every exponential is invertible. However, exp is not one-to-one, since $\exp(2i\pi I_n) = I_n = \exp 0_n$.

If $Y \in \boldsymbol{GL_n(C)}$, we have shown that there exists an $A \in \boldsymbol{M}_n(\boldsymbol{C})$ such that $Y = \exp A$. Then $M := \exp(A/2)$ satisfies $M^2 = Y$ and $M \in \boldsymbol{GL}_n(\boldsymbol{C})$.

We remind that the minimal polynomial of $Y := J(0;n)$ is X^n. If $M^2 = Y$, then $M^{2n} = 0_n$. Thus M is nilpotent, and since its size is $n \times n$, it must satisfy $M^n = 0_n$. This shows that $Y^m = 0_n$, where m is the integral part of $(n+1)/2$. Therefore $(n+1)/2 \geq n$, that is $n \leq 1$. Hence $M \mapsto M^2$ is onto $\boldsymbol{M}_n(\boldsymbol{C})$ only when $n = 1$.

12. (a) We have $(J_2 + I_2)^2 = 0_2$. If $M^2 = J_2$, we deduce that the characteristic polynomial of M divides $(X^2+1)^2$. Since it is real of degree two, it must be $X^2 + 1$. Hence $M^2 + I_2 = 0_2$, which is impossible since $J_2 \neq -I_2$.

(b) Let $Y \in \boldsymbol{M}_2(\mathbb{R})$ be such that $Y^2 = -I_2$, for instance

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

Then

$$M := \begin{pmatrix} Y & I_2 \\ 0_2 & -Y \end{pmatrix}$$

satisfies $M^2 = J_4$.

Let $M \in \boldsymbol{M}_4(\mathbb{R})$ be such that $M^2 = J_3$. Then its eigenvalues are $\pm i$. Since they come by complex conjugate pairs, each one is double and the characteristic polynomial of M is $(X^2+1)^2$. Applying Cayley-Hamilton, we obtain $(I_4 + J_3)^2 = 0_4$, which is obviously false (compute the entry at first row and last column).

(c) Let us assume that $\exp M = J_2$. From Exercise 2, $J_2 = aI_2 + bM$ for some real numbers a, b. Obviously, $b \neq 0$ and therefore M is triangular. Thus its eigenvalues are real and those of J_2, being their exponentials, must be real positive. An obvious contradiction.

We just have seen that neither the exponential nor the square are onto $\boldsymbol{GL}_2(\mathbb{R})$. A more involved result is that, as in the complex case, both maps have the same range, which means that every invertible matrix of the form M^2 (with $M \in \boldsymbol{GL}_n(\mathbb{R})$) is the exponential of some matrix with real entries. The next question supports this comment, as well as the fact that $\exp M$ is the square of $\exp(M/2)$. This could have been used above, to show that J_2 is not an exponential.

(d) Let us take $Z \in \boldsymbol{M}_2(\mathbb{R})$ such that $\exp Z = -I_2$, for instance (see Exercise 14)

$$Z = \begin{pmatrix} 0 & \pi \\ -\pi & 0 \end{pmatrix}.$$

Then

$$M := \begin{pmatrix} Z & -I_2 \\ 0 & Z \end{pmatrix}$$

satisfies

$$\exp M := \begin{pmatrix} -I_2 & I_2 \\ 0_2 & -I_2 \end{pmatrix}.$$

Conjugating by the matrix of the permutation $(2, 3)$, we obtain $J_4 = \exp P$, where

$$P = \begin{pmatrix} 0 & -1 & \pi & 0 \\ 0 & 0 & 0 & \pi \\ -\pi & 0 & 0 & -1 \\ 0 & -\pi & 0 & 0 \end{pmatrix}.$$

As mentioned above, the fact that J_3 is not a square implies that it is not an exponential. We may also make a direct proof :

Let us assume that some matrix satisfies $\exp M = J_3$. Then its eigenvalues are of the form $\pm\mu_k$, $\mu_k := (2k+1)i\pi$. They come by complex conjugate pairs. Since J_3 is not diagonalizable (in $\boldsymbol{M}_4(\mathbb{C})$), M is not, which implies that M has at least one non-simple eigenvalue. Thus its spectrum is $\{\pm\mu_k\}$, where each eigenvalue has multiplicity 2 and is non semi-simple . This means that the similarity invariants of M are $(1, 1, 1, (X^2 - \mu_k^2)^2 =: P)$. Hence M is similar (in $\boldsymbol{M}_4(\mathbb{R})$), to

$$N = \begin{pmatrix} S_k & I_2 \\ 0_2 & S_k \end{pmatrix}, \quad S_k := (2k+1)\begin{pmatrix} 0 & \pi \\ -\pi & 0 \end{pmatrix},$$

because N has the same characteristic polynomial P which turns out to be the minimal one, and thus has the same list of similarity invariants as M. We have

$$\exp N = \begin{pmatrix} -I_2 & W \\ 0_2 & -I_2 \end{pmatrix}$$

for some W, from which it follows $(I_4 + \exp N)^2 = 0_4$. This implies again $(I_4 + J_3)^2 = 0_4$, which is false.

13. If D is unitary and diagonal, its diagonal entries are numbers of unit modulus, thus are of the form e^{ia_j} for real numbers a_j. Such a D is the exponential of $D' := \mathrm{diag}(ia_1, \ldots, ia_n)$, a skew-hermitian matrix. If U is unitary, it is normal and therefore unitary diagonalisable : $U = V^*DV$, where D is unitary and diagonal. Then $U = \exp(V^*D'V)$, where $V^*D'V$ is skew-hermitian.

14. (a) We have $B^{2k} = (-1)^k\theta^{2k}I_2$ and thus $B^{2k+1} = (-1)^k\theta^{2k+1}B_1$, where B_1 is the B-matrix for $\theta = 1$. Hence

$$\exp B = (\cos\theta)I_2 + (\sin\theta)B_1 = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$$

is the rotation of angle θ.

(b) We recall that $M \in \boldsymbol{SO}_n(\mathbb{R})$ is orthogonally similar to a block-diagonal matrix M', whose diagonal blocks are either 2×2 rotations or 1's (see the proof of Proposition 7.5.1). From the previous question, M' is the exponential of a block-diagonal matrix A', whose diagonal blocks are either B-matrices or 0's. The matrix A' is skew-symmetric and M is the exponential of A, orthogonally similar to A', thus skew-symmetric.

15. (a) Since $\phi(M^2) = \phi(M)^2$ is non-negative, it equals $|\phi(M^2)|$, thus $\delta(\det M^2)$. Since the range of exp is contained in that of $M \mapsto M^2$, there follows that $\phi(A) = \delta(\det A)$ holds true on the range of exp.

(b) From Exercise 14, we know that the range of exp contains $\boldsymbol{SO}_n$. On the other hand, it contains also $\boldsymbol{SPD}_n$ (real version of Proposition 7.2.3). Hence $\phi = \delta \circ \det$ on these sets.

(c) Let $M \in \boldsymbol{GL}_n^+(\mathbb{R})$ and $M = SO$ be its polar decomposition. Then $\phi(M) = \phi(S)\phi(O) = \delta(\det S)\delta(1)$ from the previous question, thus is non-negative. Hence $\phi(M) = |\phi(M)| = \delta(\det M)$.

Let now $P \in \boldsymbol{GL}_n(\mathbb{R})$ be some matrix with $\det P < 0$. If $\det M < 0$, let N denote $P^{-1}M$. The equality $\phi(M) = \phi(PN) = \phi(P)\phi(N)$ shows that the sign of $\phi(M)$ is the same as the one of $\phi(P)$, since $\det N > 0$. Therefore ϕ takes a constant sign on matrices of negative determinant. If it is non-negative, then $\phi \equiv \delta \circ \det$. If it is non-positive, then $\phi \equiv \operatorname{sgn}(\det)\delta \circ \det$.

16. (a) Clear, since the series of norms is dominated by the exponential series of $\|x\|$. We have $\|\exp x\| \leq \exp \|x\|$.

(b) If $[x, y] = 0$, we may apply the binomial formula for $(x + y)^m$. This immediately gives $\exp(x+y) = (\exp x)(\exp y)$. Moreover, the partial sums of $\exp y$ commute with x, hence the sum of the series does : $[\exp y, x]$=0.

(c) The series of derivatives converges normally too, uniformly on bounded sets. Therefore $t \mapsto \exp tx$ is differentiable, and its derivative is the sum of the series of derivatives. We immediately obtain the formula. By induction, the map is $\mathcal{C}^\infty$.

(d) i. Differentiating twice this expression, we obtain

$$\exp(-tx)(x^3y - 2x^2yx + xyx^2)\exp(tx),$$

that is $\exp(-tx)x[x,[x,y]]\exp(tx)$, which is zero by assumption. Hence the expression is of the form $at + b$. Making $t = 0$ gives $b = xy$. Differentiating at $t = 0$ gives $a = x[y, x]$, that is $a = [y, x]x$ from the assumption.

ii. The derivative of the right-hand side is $[y, x](e-tx)\exp(-tx)$. Using the previous question, we find the same expression for the derivative of the left-hand side. Therefore both expression differ only from a constant. This constant is zero, by evaluation at $t = 0$.

iii. Using our last results, this derivative equals

$$e^{-ty}\left[e^{-tx}, y\right]e^{t(x+y)} = te^{-ty}[y, x]e^{-tx}e^{t(x+y)}.$$

Since $[y, x]$ commutes with y, it commutes with e^{-ty} too. The expression, denoted by $Y(t)$, verifies therefore $Y' = t[y, x]Y$. Since $Y(0) = I_n$, this Cauchy problem gives

$$Y(t) = \exp\left(\frac{t^2}{2}[y, x]\right).$$

Making $t = 1$ gives the Campbell-Hausdorff formula.

(e) The following matrices work

$$x = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad y = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}.$$

17. The map is well-defined : the spectrum of H is real, thus $iI_n - H$ is invertible. Since both factors commute to each other, we have

$$f(H)^* f(H) = (-iI_n - H)^{-1}(-iI_n + H)(iI_n + H)(iI_n - H)^{-1} = I_n,$$

thus $f(H) \in \boldsymbol{U}_n$. Also, the spectrum of $f(H)$ is the image of that of H, which is real, by the rational map $t \mapsto (it+t)/(it-t)$. It cannot contain -1. Let E be the set of unitary matrices whose spectrum does not contain -1. Then $f : \boldsymbol{U}_n \to E$ is bijective, since it has an inverse

$$f^{-1}(U) = i(U - I_n)(U + I_n)^{-1}$$

(check that this matrix is hermitian for every $U \in E$). Last, both f and f^{-1} are continous. Using the formula $(I_n - X)^{-1} = \sum_{k \geq 0} X^k$ for small matrices X, we obtain

$$f(tH) - \exp(-2itH) \sim \frac{t^2}{2} H^2.$$

18. (a) There holds $[M,N]^* J + J[M,N] = N^*(M^*J + JM) + (M^*J + JM)N + M^*(N^*J + JN) + (N^*J + JN)M$. Therefore $M, N \in \mathcal{G}$ implies $[M,N] \in \mathcal{G}$.

    (b) Use the series defining the exponential to derive the asymptotic formula. If $A, B \in \mathcal{G}$, then $\exp tA, \exp tB$ and their inverses $\exp(-tA), \exp(-tB)$ belong to G. Thus the product of the four exponentials belongs to G. Writing

$$(I_n + t^2[A,B] + \mathcal{O}(t^3))^* J (I_n + t^2[A,B] + \mathcal{O}(t^3)) = J,$$

    and identifying the powers of t^2, we obtain $[A,B]^* J + J[A,B] = 0_n$.

    (c) Just develop the expressions.

19. By definition, $G_{++} \cup G_{+-}$ is the subset of $\boldsymbol{O}(1,q)$ defined by $m_{11} > 0$. This is the set of matrices $M \in \boldsymbol{O}(1,q)$ such that the image of $\tau := (1,0,\ldots,0)^T$ by M belongs to the upper half-plane defined by $x_1 > 0$. Since $\tau_1^2 - \tau_2^2 - \cdots - \tau_n^2 > 0$, its image $x = M\tau$ must satisfy the same inequality. With $x_1 > 0$, we obtain $x_1 > \sqrt{x_2^2 + \cdots + x_n^2}$. Conversely, if $x = M\tau$ satisfies this inequality, then $x_1 > 0$.

20. (a) Obviously, $\sigma_H(M^{-1}N) = \sigma_H(M)^{-1}\sigma_H(N)$: σ_H is a homomorphism of $GL_n(\mathbb{R})$. Also it leaves invariant $\boldsymbol{O}(p,q)$. It is an automorphism of $\boldsymbol{O}(p,q)$, since it has the inverse $\sigma_{H^{-1}}$. Hence it sends G_0, which is a connected component of $\boldsymbol{O}(p,q)$, onto some connected component X. But since $I_n = \sigma_H(I_n) \in X$, we obtain $X = G_0$.

    (b) If $N \in \mathcal{G}$, then $\exp tN \in G$ for every t. Since it is I_n for $t = 0$, connectedness implies that it belongs to G_0 for every t. Hence $H \exp tN = (\exp tN)H$. Differentiating at $t = 0$, we obtain $HN = NH$.

    Let us write blockwise

$$H = \begin{pmatrix} X & Y \\ W & Z \end{pmatrix}.$$

Taking $N = \mathrm{diag}(0_p, A)$, with A skew-symmetric, there holds $N \in \mathcal{G}$ and thus $HN = NH$, which gives $YA = 0_{p,q}$, $AW = 0_{q,p}$ and $AZ = ZA$. This holds for every skew A and thus immediately gives $Y = 0_{p,q}$, $W = 0_{q,p}$ and $Z = zI_q$ for some $z \in \mathbb{R}$. Likewise, taking $N = \mathrm{diag}(A, 0_q)$, with A skew-symmetric, we obtain $X = xI_p$, for some $x \in \mathbb{R}$. Last, taking

$$N = \begin{pmatrix} 0_p & B \\ B^T & 0_q \end{pmatrix},$$

which also belongs to $\mathcal{G}$, we obtain $xB = zB$, that is $x = z$. Hence H is a homothety.

(c) We remark that $H \mapsto \sigma_H$ is a group homomorphism from G to $\mathrm{Aut}(G_0)$, the latter denoting the set of automorphisms of G_0. Its kernel is made of those $H \in G$ which commute with every element of G_0. We just have shown that such an H must be of the form hI_n for some $h \in \mathbb{R}$. Then, $J = H^T JH$ gives $h^2 = 1$, that is $H = \pm I_n$. Therefore, if $H, K \in G$, $\sigma_H = \sigma_K$ is equivalent to $K = \pm H$. When $K \in G_0$, this gives $H \in G_0 \cup G_{\mu,\beta}$. Conversely, if $H \in G_0$, we choose $K = H$, while if $H \in G_{\mu,\beta}$, we choose $K = -H$.

21. Let G_0 denote the connected component of the unit element e. The image of the connected set $G_0 \times G_0$ by the continous map $(g, h) \mapsto g^{-1}h$ is connected and contains $e = ee$, hence is contained in G_0. This shows that G_0 is a subgroup. Last, its image under a conjugation by any $g \in G$ is still connected and contains $e = geg^{-1}$, thus is contained in G_0. Hence G_0 is normal.

Let H be an open subgroup. Then, for every $g \in G$, gH is open, as the preimage of H under the left multiplication by g^{-1}. Let $E \subset G$ be a set of representatives of the cosets gH. Then G is the disjoint union of the gH's when G runs over E. Therefore, the complement of H is such a union and, as a union of open sets, is open. Hence H is closed. We warn the reader that in general, closed subgroups need not be open ! For instance, $\mathbb{R}$, $\mathbb{R}/\mathbb{Z}$ are topological groups, which contain discrete subgroups like $\mathbb{Z}$ in the first case, or $\frac{1}{N}\mathbb{Z}/\mathbb{Z}$ in the second one. These subgroups are closed but not open.

22. (a) We have, under that identification, $\tilde{M}(x + iy) = Ax + By + i(Cx + Dy)$. If $\tilde{M}$ is $\mathbb{C}$-linear, it must be $\tilde{M}(x + iy) = (A + iC)(x + iy)$. This happens if and only if $D = A$ and $B + C = 0$.

(b) We remind that $\boldsymbol{Sp}_n \cap \boldsymbol{O}_{2n}$ is made of matrices of the form

$$M = \begin{pmatrix} A & B \\ -B & A \end{pmatrix},$$

where $A + iB$ is unitary. Form the previous question, $\tilde{M}$ is $\mathbb{C}$-linear and equals precisely $A - iB$, which is unitary as well. One checks easily that the composition laws aggree : $M \mapsto \tilde{M}$ is an isomorphism from $\boldsymbol{Sp}_n \cap \boldsymbol{O}_{2n}$ onto $\boldsymbol{U}_m$.

8 Matrix factorizations

1. We use this algorithm in the non abelian ring $M_n(k)$, to derive a method for multiplying within $M_{Nn}(k)$. We view an $Nn \times Nn$ matrix as an $N \times N$ matrix with entries in $M_n(k)$. Thus multiplication of $Nn \times Nn$ matrices needs K multiplications and L additions of $n \times n$ matrices. In terms of complexity, we obtain

$$P_{Nn} \leq KP_n + Ln^2.$$

We argued that we shall not find a better result than $P_n = \mathcal{O}(n^2)$. Actually, noone reasonable even expects this optimal result. We shall estimate P_n using the induction

$$\pi_{j+1} \leq K\pi_j + L\,N^{2j}, \quad \pi_j := P_{N^j}.$$

We distinguish three cases, according to the position of α with respect to 2. Only the first one is reasonable. We begin by noting that the induction gives

$$\pi_j \leq L\,K^j \sum_{k=0}^{j-1} (N^2K^{-1})^k.$$

- If $\alpha > 2$, then $N^2K^{-1} < 1$ and there comes $\pi_j \leq cK^j$. Arguing as in the case of the Strassen algorithm, we obtain $P_n \leq c'n^\alpha$.
- If $\alpha = 2$, we have $\pi_j \leq cjN^{2j}$, which gives $P_n \leq c'n^2 \log n$.
- If $\alpha < 2$, we have $\pi_j \leq cN^{2j}$, that is $P_n \leq c'n^2$.

2. Using the second method, we pass from the size $n-1$ to the size n by solving a triangular system ($(n-1)^2$ operations) and extract a square root. Summing up, the complexity is $n^3/3 + \mathcal{O}(n^2)$.

3. By induction, using the second method, one easily shows that L is bidiagonal, that is $l_{ij} = 0$ whenever $i \geq j+2$. More generally, if $m_{ij} = 0$ whenever $|j-i| \geq r$, then $l_{ij} = 0$ for $i - j \geq r$.

4. Use again the second method, and remark that if L is lower triangular and non-singular, and if the k first components of R vanish, then the corresponding components of $L^{-1}R$ vanish too.

5. Same argument as usual. Let $\mathcal{T}_n$ denote the set of complex upper triangular matrices with positive real diagonal entries. The map $(Q, R) \mapsto QR$ is continuous and bijective from $\boldsymbol{U}_n \times \mathcal{T}_n$ to $\boldsymbol{GL}_n(\mathbb{C})$. It will be sufficient to prove that the preimage of converging sequences are relatively compact. Thus let $M_k = Q_kR_k$ converge towards M in $\boldsymbol{GL}_n$. Since $\boldsymbol{U}_n$ is compact, we may extract a subsequence such that Q_k converges. Then $R_k = Q_k^* M_k$ converges too.

6. The missing block is $-BA^{-1}$. Then $THT^* = \mathrm{diag}(A, C - BA^{-1}B^*)$. The second diagonal block is the Schur complement of A.

   Let H' denote $H - \mathrm{diag}(0_{n-k}, W) =: H - D$. Then H' is positive (semi-)definite if and only if $TH'T^*$ is so, since both matrices represent the same hermitian form in different bases. However, $TDT^* = D$. Therefore, $TH'T^* = \mathrm{diag}(A, S - W)$, which is positive (semi-)definite if and only if $S - W$ is so.

7. From the previous exercise, we now that

$$H - \begin{pmatrix} 0 & 0 \\ 0 & S(H) \end{pmatrix} \geq 0, \quad H' - \begin{pmatrix} 0 & 0 \\ 0 & S(H') \end{pmatrix} \geq 0.$$

   Summing up, there comes

$$H + H' - \begin{pmatrix} 0 & 0 \\ 0 & S(H) + S(H') \end{pmatrix} \geq 0.$$

   Using again the previous exercise, we obtain $S(H + H') - S(H) - S(H') \geq 0$.

   Let us first consider the case $H - H' > 0$. Then, defining $H'' := H - H'$ we apply our last result :

$$S(H) = S(H' + H'') \geq S(H') + S(H'') > S(H').$$

   Since the Schur complement S is a continuous function, we may pass to the limit as H'' tends to a positive semi-definite matrix. There comes $S(H) \geq S(H')$.

8. Since Q is unitary, its spectrum is made of complex numbers of unit modulus, and it is diagonalizable. Since its is triangular with real positive diagonal entries, the spectrum is made of such numbers. Hence $\mathrm{Sp}Q = \{1\}$. Since Q is diagonalizable, there follows $Q = I_n$, which means uniqueness.

9. In both cases, $x^\dagger = \|x\|_2^{-2} x^*$.

10. For a orthogonal projector P, one has $P^\dagger = P$. Every other orthogonal projector Q whose range is a subspace of that of P satisfies $PQ = QP = Q$. Hence every such Q satisfies the property that QP and PQ be orthogonal projectors, though only P is the generalized inverse of P.

11. (a) We first consider the case $c = 0$, that is $a = Bd$, or equivalently $B^\dagger Bd = d$. One obtains

$$AA^\dagger = BB^\dagger, \quad A^\dagger A = \begin{pmatrix} B^\dagger B - (1 + |d|^2)^{-1} dd^* & (1 + |d|^2)^{-1} d \\ (1 + |d|^2)^{-1} d^* & (1 + |d|^2)^{-1} |d|^2 \end{pmatrix},$$

    which are hermitian. Then $AA^\dagger A = (AA^\dagger)A = BB^\dagger(B, a) = (BB^\dagger B, BB^\dagger a) = (B, a) = A$, and

$$A^\dagger AA^\dagger = A^\dagger BB^\dagger = \begin{pmatrix} B^\dagger BB^\dagger - dbBB^\dagger \\ bBB^\dagger \end{pmatrix} = A^\dagger,$$

because of $bBB^\dagger = b$.

We now examine the case $c \neq 0$. Then $AA^\dagger = BB^\dagger - Bdc^\dagger + ac^\dagger = BB^\dagger + cc^\dagger$ is hermitian. Next, noticing that $c^\dagger$ is parallel to $c^* = d^*(I_q - BB^\dagger)$ (see Exercise 9), we have $c^\dagger B = 0$, from which there follows

$$A^\dagger A = \begin{pmatrix} B^\dagger B & (1 - c^\dagger a)d \\ 0 & c^\dagger a \end{pmatrix}.$$

From Exercise 9, $c^\dagger = \|c\|^{-2}c^*$. Since $c = \pi a$, where $\pi := (I_p - BB^\dagger)$ is an orthogonal projection, we have $c^\dagger a = \|\pi a\|^{-2}a^*\pi a = 1$. Finally, $A^\dagger A = \mathrm{diag}(B^\dagger B, 1)$ is hermitian.

From the latter result, we have $AA^\dagger A = (BB^\dagger B, a) = A$ and

$$A^\dagger AA^\dagger = \begin{pmatrix} B^\dagger BB^\dagger - B^\dagger Bdc^\dagger \\ c^\dagger \end{pmatrix} = A^\dagger,$$

since $B^\dagger Bd = d$. This shows that the suggested formula gives actually the generalized inverse of A.

(b) We proceed by induction over q, keeping p constant. From Proposition 8.4.1, we may assume that $q \leq p$. When $q = 1$, we use the formula $a^\dagger = \|a\|^{-2}a^*$ if $a \neq 0$ and $a^\dagger = 0$ otherwise (see Exercise 9). We now examine the step from $q - 1$ to q. The computation of d requires about $2pq$ operations, like that of c. The computation of $b = c^\dagger$, in the case $c \neq 0$, is negligible. If $c = 0$, it is not, requiring about $2pq$ operations ; however, this case means exactly $a \in R(B)$, a fact which is unlikely since B is $p \times (q-1)$ and $q - 1 < p$. Thus we shall not take this case in account in the study of complexity. Last, the computation of $B^\dagger - db$ requires also $2pq$ operations. Finally, the whole computation of $A^\dagger$ is done in about $3pq^2$ operations.

9 Iterative methods for linear problems

1. The Jacobi matrix has the property that $J_{ij} \neq 0$ implies $|j - i| = 1$. Thus $-J = PJP^{-1}$ where $P := \mathrm{diag}(1, -1, 1, -1, \ldots)$. There follows that $\mathrm{Sp}(-J) = \mathrm{Sp}(J)$, which is point 4) of Proposition 9.4.1.

2. Let $\lambda = (\lambda_1, \ldots, \lambda_n)$ be any sequence of real numbers such that $\lambda \prec 0 := (0, \ldots, 0)$. This means that, ordering λ in the non-decreasing way, there holds $\lambda_1 + \cdots + \lambda_k \leq 0$ for every k and $\lambda_1 + \cdots + \lambda_n \leq 0$. By Theorem 3.4.2, there exists a symmetric matrix J whose diagonal is $(0, \ldots, 0)$ and spectrum is λ. This matrix is the Jacobi matrix for $A := I_n - J \in \boldsymbol{Sym}_n$. The matrix A is positive definite whenever $1 - \lambda_n > 0$. For instance, $\lambda = (-(n-1)\mu, \mu, \ldots, \mu)$ works for every $\mu \in [0, 1)$. For $\mu > 1/(n-1)$, there holds $\rho(J) > 1$ and the Jacobi method for A does not converge. When μ tends to 1, $\rho(J)$ tends to $n - 1$.

Conversely, if $D = I_n$ and $\gamma := \mathrm{Sp}(A)$, then $\lambda := 1 - \gamma$ is the spectrum of J. From Theorem 3.4.1, there holds $\gamma \prec 1$, that is $\lambda \prec 0$ (notice that $a \prec b$ is equivalent to $-a \prec -b$, a counter-intuitive fact in view of the notation). Also, since $A \in \boldsymbol{SPD}_n$, we have $\lambda_j < 1$ for every j. Hence $\lambda_1 = -\sum_2^n \lambda_j > -(n-1)$. This proves that $\rho(J) < n-1$. Finally,

$$\sup\{\rho(J)\,;\, A \in \boldsymbol{SPD}_n, D = I_n\} = n - 1.$$

3. (a) The spectrum of J is even since A is tridiagonal. It is real since A and thus J, are hermitian.

   (b) The diagonal D is positive, so let us define $K := D^{-1/2}(E+F)D^{-1/2}$, which is hermitian and similar to J. If $Y \in \mathbb{C}^n$, then $Y^*KY = X^*(E+F)X$ where $X = D^{-1/2}Y$. Since A is positive definite, we have $X^*(E+F)X < X^*DX$ whenever $X \neq 0$. Hence $Y \neq 0$ implies $Y^*KY < \|Y\|_2^2$. This exactly means $\rho(K) < 1$, that is $\rho(J) < 1$.

   (c) Since $\rho(J) < 1$, the Jacobi method is convergent.

4. (a) The idea used in Exercise 3 gives the inequality $|((E+E^*)v, v)| \leq \rho(J)\|D^{1/2}v\|^2$. Then let μ_1 and μ_n denote the smallest and largest eigenvalues of hermitian matrices. From $(Av, v) \leq (Dv, v) + |((E+E^*)v, v)| \leq (1+\rho(J))\|D^{1/2}v\|^2$, we deduce $\mu_n(A) \leq (1+\rho(J))\mu_n(D)$. Similarly, $\mu_1(A) \geq (1-\rho(J))\mu_1(D)$. Finally, we obtain

$$K(A) \leq \frac{1+\rho(J)}{1-\rho(J)} K(D).$$

   (b) Actually, there holds

$$g\left(\frac{1+y}{1-y}\right) = \frac{1-\sqrt{1-y^2}}{y}.$$

   (c) If $D = I_n$, we remark that the inequalities above are equalities, so that $K(A) \leq (1+\rho(J))/(1-\rho(J))$. Since g is an increasing function, we have

$$\theta = -\log g\left(\frac{1+\rho(J)}{1-\rho(J)}\right) = -\log \frac{1-\sqrt{1-y^2}}{y} = -\frac{1}{2}\tau,$$

   where τ is the convergence ratio of SOR with the optimal parameter.

5. (a) i. Let X be an eigenvector associated to λ. Then

$$(1-\omega-\lambda)DX + \omega(E^* + e^{2i\theta}E)X = 0.$$

   We multiply this identity by $e^{-i\theta}$, so that the last parenthesis becomes hermitian. We then multiply at left by X^*. Since X^*DX (non nul) and $X^*(e^{-i\theta}E^* + e^{i\theta}E)X$ are real, we obtain that $(1-\omega-\lambda)e^{-i\theta}$ is real.

   ii. Taking the imaginary part, there comes $(\omega - 2)\sin\theta = 0$, that is $\sin\theta = 0$. Equivalently, $\lambda = 1$. However, this implies $AX = 0$, which is impossible since A is positive definite.

iii. Since P_ω does not vanish on the unit circle S^1, we have the Cauchy integral

$$m(\omega) = \frac{1}{2i\pi}\int_{S^1} \frac{P'_\omega(z)}{P_\omega(z)}dz.$$

This shows that m is a continuous function on $(0,2)$. Thus m is constant.

(b) i. There holds $\mathcal{L}_\omega - I_n = -\omega(D-\omega E)^{-1}A$. Hence the limit is $-D^{-1}A$.

ii. The matrix $D^{-1}A$ is similar to $D^{-1/2}AD^{-1/2}$, which is hermitian positive definite. Hence its eigenvalues $\mu_1, \dots, \mu_n$ are real and positive. For small ω, the eigenvalues of $\mathcal{L}_\omega$ obey to the asymtpotics $1-\omega\mu_j + \mathcal{O}(\omega^2)$, which turns out to be less than 1 for positive values. Hence $\rho(\mathcal{L}_\omega) < 1$ for small positive ω, that is $m(\omega) = n$. Since m is constant on $(0,2)$, we deduce $m \equiv n$, which means that $\rho(\mathcal{L}_\omega) < 1$ for every $\omega \in (0,2)$. Hence the relaxation method converges for all ω in $(0,2)$.

6. (a) Since $\rho(J) < 1$, one has $\lambda_a < 1$. For $\omega = 1$, $\Delta(\lambda_a) = \lambda_a^2 > 0$, while for $\omega = 2$, $\Delta(\lambda_a) = 4(\lambda_a^2 - 1) < 0$. Since $\Delta(\lambda_a)$ is a polynomial of degree two in ω, this implies that its two roots are real, with exactly one being in $(1,2)$ and one larger than 2.

(b) If $\omega \in D$, we have $\omega = 1 + e^{2i\gamma}r^2$ with $r \in [0,1)$. Let us assume that this polynomial has two roots x, y of equal moduli. This modulus will be (see the product of roots) $|\omega - 1|^{1/2}$, say

$$x = e^{i\alpha}|\omega-1|^{1/2}, \quad y = e^{i\beta}|\omega-1|^{1/2}.$$

The product gives $\alpha + \beta = 2\gamma$. Then the sum gives

$$(e^{-i\gamma} + e^{i\gamma}r^2)\lambda_a = 2r\cos(\beta - \gamma).$$

Taking the imaginary part yields $(r^2-1)\lambda_a \sin\gamma = 0$, that is $\omega = 1 + r^2 \in [1,2)$. At last, for $\omega \in [1, \omega_a)$, the discriminant $\Delta(\lambda_a)$ is positive, so that the polynomial has two distinct real roots. These turn out to have equal signs (their product is positive), thus are of distinct moduli.

Finally, whenever the polynomial has roots of distinct moduli, whose product is $\omega - 1$, exactly one of both must have a modulus larger than $|\omega - 1|$.

(c) The implicit function Theorem, in its holomorphic version, implies the holomorphy of $\omega \mapsto \mu_a$ on its domain.

When $\omega = 1 + e^{2i\gamma}$, the roots of the polynomial are $e^{i(\gamma \pm \alpha)}$ where $\cos\alpha = \lambda_a \cos\gamma$. The continuity of the roots of a polynomial in terms of its coefficients allows for the conclusion

$$\lim_{|\omega-1|\to 1} |\mu_a(\omega)|^2 = 1.$$

Also, when $\omega \in [\omega_a, 2)$ both roots have moduli $\omega - 1$. Hence the continuity argument yields

$$\lim_{\omega\to\gamma} |\mu_a(\omega)|^2 = \gamma - 1, \quad \gamma \in [\omega_a, 2).$$

(d) The maximum principle for holomorphic functions says that the modulus of μ_a cannot achieve a local maximum on the open set $D \setminus [\omega_a, 2)$. Since this modulus has a continuous extension up to the boundary, the latter being not greater than 1, and since its value is less than 1 at some interior point (take ω close to ω_a), we conclude that this modulus is less than 1 everywhere in $D \setminus [\omega_a, 2)$.

This implies that $\rho(\mathcal{L}_\omega) < 1$ everywhere in D, namely that the relaxation method converges for every such parameters ω.

(e) On a other hand, $|\mu_r|$ does not vanish in $D \setminus [\omega_a, 2)$, since it is always larger than $|\omega - 1|^{1/2}$. Hence the maximum principle, applied to the holomorphic function $1/\mu_a$, tells that the lower bound of the $|\mu_r(\omega)|$, which is continuous up to the boundary, is achieved only at interior points. From the previous question, this lower bound is achieved only at ω_r. Since $\rho(\mathcal{L}_\omega) \geq |\mu_r(\omega)|$, we deduce that

$$\rho(\mathcal{L}_\omega) > |\mu_r(\omega_r)| = \rho(\mathcal{L}_{\omega_r})$$

for every $\omega \in D \setminus \{\omega_r\}$.

7. From

$$(D - E)^{-1} = \begin{pmatrix} I & 0 & 0 \\ -M_2 & I & 0 \\ M_3 M_2 & -M_3 & I \end{pmatrix},$$

we have

$$G = \begin{pmatrix} 0 & 0 & -M_1 \\ 0 & 0 & M_2 M_1 \\ 0 & 0 & -M_3 M_2 M_1 \end{pmatrix}.$$

The spectrum of G is the union of 0 and that of $-M_3 M_2 M_1$. Thus $\rho(G) = \rho(M_3 M_2 M_1)$. Since $B^3 = \mathrm{diag}(M_1 M_3 M_2, M_2 M_1 M_3, M_3 M_2 M_1)$, where the diagonal blocks have the same spectra, except perhaps for the eigenvalue 0, we find also $\rho(B^3) = \rho(M_3 M_2 M_1)$. Finally, there holds $\rho(G) = \rho(B)^3 = \rho(J)^3$, since $J = B$. This tells that if one of both methods converges, then the other one does too, and Gauss-Seidel converges three times faster than Jacobi.

If instead

$$J = B = \begin{pmatrix} 0 & M_3 & 0 \\ 0 & 0 & M_1 \\ M_2 & 0 & 0 \end{pmatrix},$$

then

$$(D - E)^{-1} = \begin{pmatrix} I & 0 & 0 \\ 0 & I & 0 \\ -M_2 & 0 & I \end{pmatrix}, \quad G = \begin{pmatrix} 0 & -M_3 & 0 \\ 0 & 0 & -M_1 \\ 0 & M_2 M_3 & 0 \end{pmatrix}.$$

One still has $\rho(J)^3 = \rho(M_3 M_1 M_2)$. But $\rho(G)$ equals the spectral radius of $\rho(H)$, where

$$H := \begin{pmatrix} 0 & -M_1 \\ M_2 M_3 & 0 \end{pmatrix}.$$

Since H^2 is block-diagonal, we find $\rho(G)^2 = \rho(M_3 M_1 M_2)$ and therefore $\rho(G) = \rho(J)^{3/2}$. This tells that one method converges if and only if the other one does, and then Gauss-Seidel is one and a half faster than Jacobi.

For a cyclic matrix of order p, the best acceleration factor is obtained when the structure of B is such that $B_{ij} = 0$ whenever $i \neq j+1$ modulo p. In case of convergence, Gauss-Seidel is p times faster because $\rho(G) = \rho(J)^p = \rho(M_p \cdots M_1)$. The worst case is when $B_{ij} = 0$ whenever $j \neq i+1$. Then Gauss-Seidel is only $p/(p-1)$ faster than Jacobi. Notice that in this case, G is block triangular, where the diagonal blocks are 0 in one hand and a $(p-1)$-cyclic matrix on the other hand.

10 Approximation of eigenvalues

1. (a) The sequence is obviously well-defined since one divides by non-zero polynomials. Since $\deg P_j$ decays strictly, the sequence is finite. By definition, the last polynomial P_r is a constant. If it is zero, then P_{r-1} divides P_{r-2} and there follows by induction that P_{r-1} divides every other P_j, including P and P'. This contradicts the fact that P has only simple roots. Hence P_r is a non-zero constant. If $P_j(x) = 0$, then $P_{j-1}(x)P_{j+1}(x) = -P_{j-1}(x)^2 \leq 0$. Actually, if it vanished then x would be a common root of P_{j-1} and P_j, and therefore of every P_l, including P_r, a contradiction. Hence $P_{j-1}P_{j+1} < 0$ and we have a Sturm sequence.

   (b) For $P = X^2 + aX + b$, we have $P_1 = 2X + a$ and $P_2 = -b + a^2/4 =: \delta$. We evaluate the total number of real roots by computing $V(-\infty) - V(+\infty)$. Since $V(\pm\infty)$ is the number of sign changes in the sequences $(+\infty, \pm\infty, \delta)$, we have the following cases. If $\delta < 0$, then $V(\pm\infty) = 1$, thus the number of real roots is $1 - 1 = 0$. If $\delta > 0$, then $V(-\infty) = 2$, $V(+\infty) = 0$ and the number of real roots is $2 - 0 = 0$.

   Let now turn to the case of $P = X^3 + pX + q$. Then $P_1 = 3X^2 + p$, $P_2 = -(2p/3)X - q$ and $P_3 = -p - 27q^2/(4p^2) =: -\delta$. Here, we count the sign changes in the sequences $(\pm\infty, +\infty, \mp(\operatorname{sgn} p)\infty, -\delta)$. If $\delta > 0$, we have $V(-\infty) = 2$ and $V(+\infty) = 1$, for $2 - 1 = 1$ real root. If $\delta < 0$, then $p < 0$, and we have $V(-\infty) = 3$ and $V(+\infty) = 0$, for $3 - 0$ roots.

2. (a) If $|j| \leq p-1$, there holds $(W_n x)_{p+j} = (|j|+1)x_{p+j} + x_{p+j-1} + x_{p+j+1}$. Besides, $(W_n x)_1 = px_1 + x_2$ and $(W_n x)_n = px_n + x_{n-1}$. Hence $x \in E'$ (respectively $x \in E''$) implies $W_n x \in E'$ (respectively $W_n x \in E''$).

   (b) Let $\{e^1, \ldots, e^n\}$ be the canonical basis of $\mathbb{R}^n$. We build a basis $\{f^1, \ldots, f^p\}$ of E' by $f^{p-j} := e^{p-j} + e^{p+j}$ when $j = 0, \ldots, p-1$. Likewise, $g^{p-j} := e^{p-j} - e^{p+j}$ $(j = 1, \ldots, p-1)$ defines a basis of E''. The matrix W_n' (respectively W_n'') is the one of the restriction to E' (respectively to E'') of W_n in this basis.

(c) Let us define

$$W_n''' := \begin{pmatrix} p & 1 & & \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & 3 \end{pmatrix} \in M_{p-2}(\mathbb{R}).$$

From Proposition 10.1.2, and since both W_n'' and W_n''' are symmetric, their eigenvalues are simple and those of W_n''' strictly separate those of W_n''. Denoting by P, Q, R the characteristic polynomials of W_n', W_n'', W_n''', we have (expand with respect to last row and last column) $P(X) = (X-1)Q(X) - 2R(X)$. Then proceed as in the proof of Proposition 10.1.2.

3. (a) The eigenvectors are $r^j := (1, \dots, 1, 1-j, 0, \dots, 0)^T$ (with $j-1$ times a 1) for $j = 2, \dots, n$ and $r^{n+1} := (1, \dots, 1)^T$. The corresponding eigenvalues are given by $\lambda_j = \sum_k r_k^j a_k$.

(b) That each sum of a row or of a column equals 1 is clear. Also, $b_j - a_j \geq (j-1)(a_{j-1} - a_j) \geq 0$. Thus $M(a)$ is bi-stochastic. Next, $b_{j+1} - a_{j+1} - (b_j - a_j) = j(a_j - a_{j+1}) \geq 0$. At last, $b_n - a_n = 1 - na_n \leq 1$.

Let us remark that $\lambda_j = b_j - a_j$ for $j = 2, \dots, n$.

(c) We may assume that $\mu_1 \leq \cdots \leq \mu_n$. We need to solve the linear system

$$\begin{aligned} a_1 - a_2 &= \mu_1, \\ a_1 + a_2 - 2a_3 &= \mu_2, \\ &\vdots \\ a_1 + \cdots + a_{n-1} - (n-1)a_n &= \mu_{n-1}, \\ a_1 + \cdots + a_n &= 1. \end{aligned}$$

Elimination between the two last equations gives the value of a_n (through $1 - na_n = \mu_{n-1}$). Then a backward induction gives a unique value of a_j for $j \geq 2$. Last, knowing $a_2, \dots, a_n$, the last equation gives a unique value for a_1. Thus this $n \times n$ system is uniquely solvable.

There remains to prove that $a \in S$. First of all, we have $b_j - a_j = \mu_{j-1}$ and therefore $b_{j+1} - a_{j+1} - (b_j - a_j) \geq 0$, which exactly means that a is non-increasing. Last, $\mu_{n-1} \leq 1$ gives $a_n \geq 0$.

(d) Corollary 5.5.1 tells that Δ is included in Σ. Recall that Δ is the polytope of dimension $(n-1)^2$, made of bi-stochastic matrices. Now, let $P \in \Sigma$. We begin with the case where $P \in \boldsymbol{Sym}_n$ is positive semi-definite. Then $\rho(P) = \|P\|_2 = 1$, which means that its eigenvalues μ_j satisfy $0 < \mu_1 \leq \cdots \leq \mu_n = 1$. Hence there exists $a \in S$ such that $\mu_1, \dots, \mu_n$ be the eigenvalues of $M(a)$. Since both P and $M(a)$ are symmetric, thus orthogonally diagonalizable, we equal spectra, we deduce that they are orthogonally similar : $P = U^T M(a) U$ with $U \in \boldsymbol{O}_n$. Then $U^T \Delta U$ is a polytope

of dimension $(n-1)^2$, containing P and contained in Σ, by unitary invariance of the norm.

In the general case, we use the polar decomposition of P, $P = QH$. It is well-defined if P is non-singular and then $Q \in \boldsymbol{O}_n$, H is symmetric positive definite. If P is singular, compactness of $\boldsymbol{O}_n$ and density of $\boldsymbol{GL}_n(\mathbb{R})$ in $\boldsymbol{M}_n(\mathbb{R})$ show that a polar decompsition does exist, though possibly non-unique. From unitary invariance of the norm, we have $H \in \Sigma$. We may apply our first analysis to H : there exists a polytope K of dimension $(n-1)^2$, contained in Σ and containing H. Then QK is a polytope K of dimension $(n-1)^2$, contained in Σ and containing P.

4. We use the method of Householder for the QR factorization of a matrix M. It consists in multiplying at left by a matrix of rotation in the plane of coordinates x_p, x_q, in order to replace the (p,q) entry by zero. If the choice of the planes is made in the right order, one obtains a triangular matrix after $n(n-1)/2$ such multiplications. In the case of Hessenberg matrices, only $n-1$ sub-diagonal entries have to be replaced by zeroes, and it is done after $n-1$ such multiplications. Thus $R = P_1 \cdots P_{n-1} M$ and $Q = P_{n-1}^T \cdots P_1^T$, where P_j denote the rotation matrices.

   The QR iteration yields the Hessenberg matrix

$$M^{(1)} = RQ = P_1 \cdots P_{n-1} M P_{n-1}^T \cdots P_1^T,$$

   which can be computing by the successive conjugation by P_j.

   In the tridiagonal case, a conjugation is very cheep. It modifies only eight entries. But, the hermitian property being preserved, we really need to actualize only five entries. Each actualization needs at most 10 operations, and often less. Thus the complexity of a QR iteration is an $\mathcal{O}(n)$. The factor 20 takes in account not only the actualization, but also the computation of the cosine and sine of each rotation.

5. In the proof of Theorem 10.1.1, the matrix M_{n-r} is hermitian. Therefore H is already tridiagonal and there holds

$$B = \begin{pmatrix} 0_{n-r-1,r} \\ Z^T \end{pmatrix}.$$

   Also, N is hermitian. When computing $V^{-1} M_{n-r} V$, BS is already known, since it is the hermitian adjoint of SZ. Thus there remains only the computation of SNS, for which we only have to calculate the lower half. In all, it needs only about $7r^2/2$ operations. Summing up from $r = 1$ to $n-2$, the whole procedure needs $7n^3/6$ operations.

6. (a) There holds $KT = t_{nn} K$, $TK = t_{11} K$.

   (b) Expanding with respect to the first row, we find that $\mu \mapsto \det(M - \lambda I_n - \mu K)$ is affine, say $a\mu + b$. Making $\mu = 0$, we have $b = \det(M - \lambda I_n)$. Last, a is the $(n,1)$-cofactor, which is precisely $(-1)^n \det(M - \lambda I_n)_1$.

   We remark that, in this identity, I_n may be replaced by any other matrix in $\boldsymbol{M}_n(\mathbb{R})$.

(c) We use the previous formula :

$$\begin{aligned}(-1)^n\mu\det(A-\lambda I_n)_1+\det(A-\lambda I_n) &= \det(QR-\lambda I_n-\mu K)\\ &= \det R\det(Q-\lambda R^{-1}-\mu K R^{-1})\\ &= \det R\det\left(Q-\lambda R^{-1}-\frac{\mu}{r_{nn}}K\right)\\ &= \det R\left((-1)^n\frac{\mu}{r_{nn}}\det(Q-\lambda R^{-1})_1\right.\\ &\quad \left.+\det(Q-\lambda R^{-1})\right).\end{aligned}$$

Equating the powers of μ gives the desired identity.

(d) We perform a similar computation with $A'=RQ$:

$$\begin{aligned}(-1)^n\mu\det(A'-\lambda I_n)_1+\det(A'-\lambda I_n) &= \det(RQ-\lambda I_n-\mu K)\\ &= \det R\det(Q-\lambda R^{-1}-\mu R^{-1}K)\\ &= \det R\det\left(Q-\lambda R^{-1}-\frac{\mu}{r_{11}}K\right)\\ &= \det R\left((-1)^n\frac{\mu}{r_{11}}\det(Q-\lambda R^{-1})_1\right.\\ &\quad \left.+\det(Q-\lambda R^{-1})\right).\end{aligned}$$

Equating the powers of μ gives

$$\det(A'-\lambda I_n)_1=\frac{\det R}{r_{11}}\det(Q-\lambda R^{-1})_1.$$

Using this identity and that of the previous question, there comes

$$r_{nn}\det(A'-\lambda I_n)_1=r_{11}\det(A-\lambda I_n)_1.$$

(e) If $\ell\geq 1$ is an integer, we consider the matrix

$$K=\begin{pmatrix}0 & L_\ell\\ 0 & 0\end{pmatrix}.$$

If T is upper triangular (block-triangular is sufficient), then

$$KT=\begin{pmatrix}0 & T_+\\ 0 & 0\end{pmatrix},\quad TK=\begin{pmatrix}0 & T_-\\ 0 & 0\end{pmatrix},$$

where T_- and T_+ are the first and last diagonal blocks.
On an other hand, if

$$N=\begin{pmatrix}0 & L\\ 0 & 0\end{pmatrix},$$

then $\det(M-\lambda P-\mu N)$ is a polynomial in μ of degree ℓ, whose leading order term is $\epsilon(n,\ell)\mu^\ell \det L \det(M-\lambda P)_\ell$. Applying this formula to both A and A', with $M=Q$, $P=R^{-1}$ and either $N=KR^{-1}$ or $N=R^{-1}K$, we obtain the identity (r_j denoting r_{jj})

$$r_n \cdots r_{n-\ell+1} \det(A'-\lambda I_n)_\ell = r_1 \cdots r_\ell \det(A-\lambda I_n)_\ell.$$

Since none of the r_j's vanish, we deduce that the polynomial $\det(A-\lambda I_n)_\ell$ keeps the same roots and same multiplicities along a step of the QR algorithm. Let us remark that all this analysis applies also to $\ell=0$, where it just gives the invariance of the eigenvalues, which the QR algorithm is designed for.

For a general matrix, $\det(A-\lambda I_n)_\ell$ is a polynomial of degree $n-2\ell$ whenever $\ell \leq n/2$, and is a constant otherwise. Thus we expect the invariance of $n-2\ell$ roots for every such ℓ. Summing up over ℓ, we obtain that there are in general $p(p+1)$ (if $n=2p$) or $(p+1)^2$ (if $n=2p+1$) invariant roots.

For a Hessenberg matrix, these polynomials are constants for every $\ell \geq 1$. Therefore, we do not find other invariants than the eigenvalues.

7. (a) Since the determinant is invariant under transposition, we have $P_M(h;z) = P_M(1-h;z)$. In other words, the polynomial $P_M(X+1/2;z)$ is even with respect to X and thus contains only powers of X^2. This means that it also reads as $R_M(X^2;z)$ where R_M is a polynomial. Then $Q_M(Y;z) := R_M(-Y+1/4;z)$ works.

   (b) Since

$$(1-h)N + hN^T - zI_n = Q^T((1-h)M + hM^T - zI_n)Q,$$

   there holds $P_N = P_M$ and therefore $Q_N = Q_M$. Hence Q_M remains constant along the QR algorithm.

   (c) In particular, the coefficients $J_{rk}(M)$ of the monomials $Y^k z^{n-r}$ in Q_M remain constants along the QR algorithm. Since P_M is also polynomial in the entries of M, each J_{rk} is a polynomial function. For $k=0$, these are the coefficients of $Q_M(0;\cdot)$, that is of the characteristic polynomial of M. This just confirms the fact that the algorithm was designed in such a way that the spectrum be preserved.

   (d) When $r=n$, we only need to compute $\det((1-h)M+hM^T)$. We find here $J_{21} = -(m_{12}-m_{21})^2$. When Theorem 10.2.1 applies, the entries m_{11}, m_{22} and m_{21} converge towards λ_2, λ_1 and 0. From the invariance of J_{21}, there follows that m_{12}^2 converges towards $-J_{21}$. Hence the sequence has at most two cluster points of the form

$$\begin{pmatrix} \lambda_2 & \pm c \\ 0 & \lambda_1 \end{pmatrix}.$$

   From the continuity of the QR factorization, we know that

$$Q_k \to \mathrm{diag}(\mathrm{sgn}\lambda_2, \mathrm{sgn}\lambda_1).$$

   Since $\det A > 0$, this limit is $\pm I_2$, so that $A_{k+1} - A_k$ tends to zero. Having at most two cluster points, the sequence A_k must therefore converge.

(e) Again, J_{21} is the coefficient of $h(1-h)z^{n-2}$ in P_M. The coefficient of z^{n-2} is the sum of the principal minors of size 2 in $(1-h)M + hM^T$. Thus we are gone back to the 2×2 case done above. We find

$$J_{21}(M) = -\sum_{i<j}(m_{ij} - m_{ji})^2 = -\frac{1}{2}\mathrm{Tr}\left((M - M^T)^2\right).$$

If A_k converges to a diagonal matrix D, then the invariance of J_{21} gives $J_{21}(A) = J_{21}(D)$, that is $J_{21}(A) = 0$, which means that A is symmetric.

8. The convergence of $A^{(k)}$, together with the formula (10.4) and the fact that the angle θ_k belongs to $[-\pi/4, \pi/4)$, show that θ_k tends to zero, with $\theta_k = \mathcal{O}(\|E_k\|)$. From Theorem 10.3.1, we deduce $\theta_k = \mathcal{O}(\rho^k)$. Therefore $R^{(k)} - I_n = \mathcal{O}(\rho^k)$. Thus the series $\sum_k (R^{(k)} - I_n)$ is convergent, and this implies that the product $\prod_k R^{(k)}$ converges.

9. Such a matrix is unitary if and only if $|z_1|^2 + |z_3|^2 = 1$, $|z_2|^2 + |z_4|^2 = 1$ and $\bar{z}_1 z_2 + \bar{z}_3 z_4 = 0$. Using it in the same way as in the real Jacobi method, one vanishes the (p, q)-entry provided

$$\bar{z}_1(z_2 h_{pp} + z_4 h_{pq}) + \bar{z}_3(z_2 h_{qp} + z_4 h_{qq}) = 0.$$

One may specialize to matrices where $z_4 = z_1$ is real and $z_3 = -\bar{z}_2$. Then the sole constraint is $z_1^2 + |z_2|^2 = 1$, which means that there are angles θ, ϕ such that $z_1 = \cos\theta$ and $z_2 = e^{i\phi}\sin\theta$. To vanish the (p, q)-entry, we need $\Im(e^{-i\phi}h_{pq}) = 0$ and $sc(h_{pp} - h_{qq}) + (c^2 - s^2)\Re(e^{-i\phi}h_{pq}) = 0$. The correct choice is $e^{i\phi} = \bar{h}_{pq}/|h_{pq}|$ and

$$\cot 2\theta = \frac{h_{pp} - h_{qq}}{2|h_{pq}|}$$

with $\theta \in [-\pi/4, \pi/4)$. Remark that the other choice $(\phi + \pi, -\theta)$ gives exactly the same unitary conjugation.

10. (a) The decay of $\|E_k\|$ was proved in Theorem 10.3.1 without assuming a specific choice of θ_k. Hence it holds true as well in the present case. Since the matrices $A^{(k)}$ are unitary similar to A, they form a relatively compact sequences, whose cluster values are unitary similar to A and are actually the cluster values of D_k. Hence they are diagonal, with the eigenvalues of A as diagonal entries.

(b) Since the eigenvalues are simple, Formula (10.4) shows that $\tan 2\theta_k$ tends to zero. Since $\pi/4 \leq |\theta_k| \leq \pi/2$, we deduce that θ_k tends towards $\pi/2$. Thus

$$a_{pp}^{k+1} - a_{qq}^k = c^2(a_{pp} - a_{qq})^{(k)} - 2csa_{pq}^{(k)} \to 0,$$

and similarly $a_{qq}^{k+1} - a_{pp}^k \to 0$.

11. Without loss of generality, we may assume that $a_0 = 1$. Then let us form a companion matrix for the polynomial :

$$M = \begin{pmatrix} 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 \\ -a_n & \cdots & \cdots & -a_1 \end{pmatrix}.$$

The vector $x_k := (z_k, \ldots, z_{k+n-1})^T$ satisfies the induction $x_{k+1} = Mx_k$. Up to the normalization, this is the power method. Hence, if the polynomial has only one root of maximal modulus, and if it is simple, the ratio z_{k+1}/z_k converges to that one.

12. (a) Since $\|Mx^k\| = \|M^{k+1}x^0\|/\|M^k x^0\|$, the expression equals

$$\frac{1}{m} \log \frac{\|M^m x^0\|}{\|x^0\|}.$$

(b) The upper bound comes from the Householder Theorem. Applying Householder Theorem to $(M|_E)^{-1}$ gives the lower bound for $\|M^k y^0\|$. Without loss of generality, we may assume that $\mu > \rho(M|_F)$. Then a third use of the same Theorem gives $\|M^k z^0\| = o(\mu^k)$, hence the result.

(c) The last result tells that

$$\frac{1}{m} \log \|M^m x^0\| \to \log \rho(M).$$

Therefore

$$\frac{1}{m} \sum_{k=0}^{m-1} \log \|M^k x^0\| \to \log \rho(M),$$

which exactly tells that $\log \|M^k x^0\|$ converges in the mean (in the Cesaro sense) towards $\log \rho(M)$.

13. Let us recall (Theorem 4.5.1) that M admits exactly one, simple, eigenvalue λ in the disc D_l.

Without loss of generality, we may assume $m_{ll} = 0$. Let r be the radius od D_l. We use $\mu = 0$ as a coarse approximation of λ. Computing $\det M$, we may check whether $\lambda = 0$. If so, then we are done. If not, then the inverse power method consists in applying the power method to M^{-1} ; at each step, one has to solve a linear system $Mx = b$. Since $1/\lambda$ is the unique, simple, eigenvalue of maximal modulus of M^{-1} by assumption, the method converges and gives the eigenvalue, hence it inverse λ.