Brief notes on linear algebra and its applications

CEP-TEQIP course: 2nd - 6th Jan 2020, held at EE, IIT Bombay

Instructors: Profs. Debasattam Pal, Madhu N. Belur & Harish K. Pillai

Please write about corrections/typos to belur@iitb.ac.in

Corrected and updated notes are being maintained at http://www.ee.iitb.ac.in/%7Ebelur/LAA

Outline: Scalars, vectors, linear independence, spaces, basis, dimension
Examples: span of exponentials, Fourier series, polynomials of a finite degree
n-tuples, row-vector/column-vector
Matrices arising in linear set of equations
Maps between vector spaces, matrices
Rank of a matrix: square/non-square, range/column-span, kernel/null-space
Row/column ranks, full row rank, full column rank, link with range/kernel

Orthogonal matrices, symmetric matrices, invertible matrices, permutation matrices
Methods to solving $Ax = b$: existence, uniqueness of x, inverse of A
SVD, rank, nearest low rank approximation, notion of distance, norms
Ill-conditioned matrices, and link with sensitivity to solution of $Ax = b$
Hands-on session on Octave/Scilab

Different methods for solving $Ax = b$, exposure to accuracy,
Eigenvalues, eigenvectors.
Similarity transformation, Jordan normal form, Schur form: real/complex
Special cases: diagonal, triangular, orthogonal, symmetric (no proofs)

Notions/definition of other decompositions: QR, LU, link with Gauss Elimination
QR, inverse, LU, each being $O(N^3)$
Overview of flop-count, determinant being $O(N^3)$ though definition suggests $O(N!)$
Inner products, inner product spaces. Gram-Schmidt orthogonalization

Orthogonal projection, best approximation,
Application to estimation theory: least-square fit for line/polynomial of a degree
Power method of computing eigenvalues/eigenvectors
Google's page rank algorithm. Perron-Frobenius Theorem
SVD, PCA, POD, data compression
Hands-on session on Octave/Scilab

These notes are meant to be a collection of definitions and facts. Detailed explanation and proofs of the facts can be found in standard books, for example.

References about Linear Algebra:

- Gilbert Strang, Introduction to Linear Algebra
- Kenneth Hoffman & Ray Kunze, Linear Algebra

References about numerical aspects of linear algebra

- Trefethen: Numerical Linear Algebra

- Golub & van Loan: Matrix Computations
- D. Watkins: Fundamentals of Matrix Computations

Research paper:

- J. Kautsky, N.K. Nichols, P. Van Dooren, Robust pole assignment in linear state feedback, *International Journal of Control*, 1985. (For application of how almost perpendicular closed loop eigenvectors help in maximizing robustness to parameter perturbations.)

Notation:
$\mathbb{R}$ stands for set of real numbers and $\mathbb{C}$ stands for set of complex numbers. Both are fields. $\mathbb{R}^n$ stands for n-tuples of real-numbers, and usually stacked as a column. $\mathbb{R}^{m\times n}$ stands for matrices with m-rows and n-columns.

Definition 1 *Consider a vector space $\mathbb{V}$ over a field $\mathbb{F}$. A set of vectors $v_1, v_2, \ldots, v_m \in \mathbb{V}$ are said to linearly dependent if there are scalars $\alpha_1, \alpha_2, \ldots, \alpha_m$, not all zero, such that*

$$\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_m v_m = 0.$$

First understand for $\mathbb{V} = \mathbb{R}^n$ with $\mathbb{F} = \mathbb{R}$, and then the general case. Note that 'not all zero' in above definition means 'at least one α_i is nonzero'.

Definition 2 *If such α_i (with at least one nonzero), as in the previous definition, do not exist, then v_i are said to be linearly independent.*

Fact 3 *Alternative (and equivalent) definition for linear independence:*
For a vector space $\mathbb{V}$ over the field $\mathbb{F}$, a set of vectors $v_1, v_2, \ldots, v_m \in \mathbb{V}$ are said to linearly independent if whenever

$$\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_m v_m = 0.$$

with scalars $\alpha_1, \alpha_2, \ldots, \alpha_m \in \mathbb{F}$, then it implies that each $\alpha_i = 0$.

This means that the only linear combination of $\sum \alpha_i v_i$ that gives the zero vector $0 \in \mathbb{V}$ is the 'trivial' linear combination.
Dimension of a vector space $\mathbb{V}$: the maximum set of independent vectors we can find in $\mathbb{V}$. Note: if vectors v_1, $v_2, \ldots, v_m$ are independent, then $m \leqslant$ dimension of $\mathbb{V}$. Further, $\mathbb{R}^n$ has dimension n (as a vector space over $\mathbb{R}$).
When $A \in \mathbb{R}^{m\times n}$ acts on a vector $v \in \mathbb{R}^n$ and gives vector $Av \in \mathbb{R}^m$, we speak of the kernel and range of A.

Definition 4 *For $A \in \mathbb{R}^{m\times n}$, consider the map $A : \mathbb{R}^n \to \mathbb{R}^m$, with $v \in \mathbb{R}^n$ mapped to Av.*

$$\textit{kernel } A := \{x \in \mathbb{R}^n | Ax = 0\},$$
$$\textit{range } A := \{y \in \mathbb{R}^m | y = Ax \textit{ for some } x \in \mathbb{R}^n\}$$

Note: kernel here also is called nullspace. Further, range is also called column-span and image. The zero element is always in these sets for any matrix A. Whether the kernel of A has nonzero elements also or not is very important, see next fact. Whether image of A is the full space $\mathbb{R}^m$ or strictly smaller than $\mathbb{R}^m$ is also important.

Fact 5 *If column vectors $v_i \in \mathbb{R}^n$ are stacked as columns of a matrix, suppose we get matrix $A \in \mathbb{R}^{n\times m}$: first column v_1, and last column as v_m. Then, the following are equivalent.*

1. *The set of vectors v_i are independent.*
2. *The matrix A has rank m.*
3. *There is a $m \times m$ submatrix in A which is nonsingular.*
4. *Nullspace/kernel of A contains only the zero vector in $\mathbb{R}^m$.*
5. *$Ax = 0 \implies x = 0$.*
6. *Number of independent vectors in image of A = m, i.e. dimension of image$(A) = m$.*

We call A defined above as having 'full column rank', i.e. rank (A) = number of columns of A. For this to happen, A has to be 'tall', i.e. number of rows is more than number of columns.
More generally, rank of a matrix cannot exceed the minimum of the number of rows and columns. When the rank of matrix A equals the minimum (of the number of rows/columns), we tell A is full rank. In general, the number of independent columns equals the number of independent rows, which is equal to the rank of matrix A. Rank of a matrix A is also the dimension of the image of A, i.e. number of independent vectors in the range-space of A. The rank-nullity theorem relates rank and dimension of nullspace.

Theorem 6 ***(*Rank nullity theorem) Consider a matrix $A \in \mathbb{R}^{m\times n}$. Then,***

$$n = \dim(\textbf{image } (A)) + \dim(\textbf{kernel } (A)).$$

Definition 7 *A square matrix $P \in \mathbb{R}^{n\times n}$ is called singular if the determinant of P (denoted by* det (P)*) equals* 0*. Otherwise, P is called nonsingular.*

Notions of singular-matrix and nonsingular-matrix are defined only for square matrices.

Definition 8 *A square matrix $U \in \mathbb{R}^{n\times n}$ is called orthogonal if $U^{-1} = U^T$.*

We use the word 'orthogonal matrix' only for square matrices. However, one or more vectors are also defined as mutually 'orthogonal vectors'. Notion of 'orthonormal' is related to these.
Singular value decomposition (SVD) is amongst the most widely used concepts in linear algebra.

Theorem 9 SVD: *For a possibly non-square matrix $A \in \mathbb{R}^{m\times n}$,*
there exist matrices $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$, and a 'diagonal' matrix $\Sigma \in \mathbb{R}^{m\times n}$ such that $A = U\Sigma V^T$,
with Σ as follows:

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \sigma_2 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \vdots & \ddots & \cdots & 0 & \cdots & 0 \\ 0 & 0 & 0 & \sigma_r & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \vdots & \ddots & \cdots & 0 & \ddots & 0 \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \end{bmatrix}, \text{ with } 0 < \sigma_r \leqslant \sigma_{r-1} \leqslant \cdots \leqslant \sigma_2 \leqslant \sigma_1 \,.$$

Note that Σ is the same size of A while U and V have to each be square, but possibly different sizes.

Fact 10 *In the above SVD, $r = \text{rank}\,(A)$, the rank of the matrix A. Further, the σ_i (called singular values of A) are unique, meaning they depend only on A and not on the orthogonal matrices U and V. However, U and V are possibly non-unique.*

Amongst the many different applications of SVD are for following purposes.

- Numerically 'stable': singular values are continuous in entries of A
- There are fast algorithms to compute SVD
- Used for determining 'rank' of a matrix: the number of nonzero singular values
- Zero or nonzero decision: needs a tolerance to be specified (for σ_r)
- Rank := number of singular values LARGER than specified tolerance
- Approximating a matrix by a low-rank one, thus: data-compression
- Proper Orthogonal Decomposition (POD): used in model order reduction
- Principal Component Analysis (PCA): used in unsupervised learning

Fact 11 *Suppose $A \in \mathbb{R}^{m\times n}$ and A has rank r with SVD of $A = U\Sigma V^T$, with $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$. Suppose the columns of U are $u_1, u_2, \ldots, u_m$, while the columns of V are $v_1, v_2, \ldots, v_n$. Then we can express the SVD also as*

$$A = \sum_{i=1}^{r} u_i \sigma_i v_i^T.$$

Thus A can be expressed as a (scaled) sum of rank-one matrices $u_i v_i^T$, and each rank one-matrix $u_i v_i^T$ is scaled by σ_i. This helps in obtaining a lower rank approximation of A: when the last few nonzero singular values are negligible.
The SVD can be used to construct the 'pseudo-inverse' of A.
Quite closely linked to singular values are the notions of eigenvalue and eigenvector of a square matrix.
Note: SVD is defined for non-square matrices also, while eigenvalues and eigenvectors are defined for square matrices only.

Definition 12 *Suppose $A \in \mathbb{R}^{n\times n}$.*
If a nonzero vector v satisfies

$$Av = \lambda v \text{ for some number } \lambda$$

then, v is called eigenvector and λ is called (corresponding) eigenvalue.

Note that even if $A \in \mathbb{R}^{n\times n}$, the eigenvalue λ might have to be complex,
then: look for v too in $\mathbb{C}^n$.
Most vectors change when A acts on the vector: in both length and direction. View eigenvector v as a direction that A does not 'change': but just scales v.

Eigenvector: direction that A does not change, just scales: scaling = eigenvalue

Fact 13 *Eigenvalues: roots of polynomial* $\det(sI - A)$*, and then for each root* λ*, look for nonzero vectors v in kernel of* $\lambda I - A$*.*

This fact is used as a computational procedure for finding eigenvalues and eigenvectors.

Fact 14 *Suppose* $A \in \mathbb{R}^{n \times n}$

- *Do eigenvalues always exist? Yes, if complex eigenvalues are allowed too.*
- *Do real eigenvalues also exist? May be or may be not. Like roots of a polynomial.*
- *Are eigenvectors unique? Just direction: so scaled vector also is eigenvector. (Scaling should be nonzero scaling.)*
- *Are eigenvectors unique 'up to' direction? If eigenvalues are distinct, then yes.*
- *Can eigenvalue be repeated? Repetition is possible, like roots of a polynomial.*
- *Do eigenvectors form a 'basis'? (≡ diagonalizable)*
  Yes, if distinct eigenvalues. If not, still possible (under certain conditions).
- *Repeated eigenvalue case ≡ Jordan canonical form*
  Matrices with repeated eigenvalues could be 'diagonalizable'
- *Simple eigenvalue: not-repeated*
- *Semi-simple eigenvalue: repeated, but yield that many eigenvectors*
  (not a hurdle to diagonalizability).
- *'Defective' eigenvalue: repeated eigenvalues and insufficient eigenvectors. Then, need to go for 'generalized eigenvalues'.*
- *Eigenvector can be normalized to length 1. (With respect to a pre-defined notion of length.)*
- *v corresponding to distinct eigenvalue* λ *is then (almost) unique.*
- *In order to achieve uniqueness, after length is made 1, we can still choose v and* $-v$ *(or* αv *with* $\alpha \in \mathbb{C}$ *and* $|\alpha| = 1$*).*
- *If* $\lambda \in \mathbb{R}$*, then* $v \in \mathbb{R}^n$ *(and* $v \neq 0$*). One can choose sign of one (nonzero) component.*

Projection matrices
A square matrix $P \in \mathbb{R}^{n \times n}$ is called a projection matrix if $P^2 = P$. Eigenvalues of P are 0 and 1 only: counted with multiplicity, we will have n of them. Image of P is the subspace to which projection happens. Reflectors can be viewed as $(I - 2P)$, i.e. along kernel of P, no change happens, and along image of P, we have reflection.

Invariant subspaces
Suppose A is a map from $V \to V$. A subspace $W \subseteq V$ is called A-invariant if $AW \subseteq W$. Of course, the set containing only 0 vector is A-invariant and the entire space V is also A-invariant: these are trivial. Other non-trivial examples of A-invariant subspaces are spans of one or more eigenvectors.

Algebraic and geometric multiplicity
The characteristic polynomial χ_A of a matrix $A \in \mathbb{R}^{n \times n}$ is defined as determinant $\det(sI - A) =: \chi_A(s)$.

The number of times a root λ is repeated as a root of the polynomial χ_A is called the eigenvalue's algebraic multiplicity ($n_a(\lambda)$). The dimension of the kernel of $\lambda I - A$ equals $n - \text{rank}\ (\lambda I - A)$, and this number is called the geometric multiplicity of λ ($n_g(\lambda)$). In general, for any eigenvalue λ of A,

$$\text{geometric multiplicity of } \lambda \leqslant \text{ algebraic multiplicity of } \lambda.$$

Algebraic multiplicities of all the eigenvalues add up to n (the size of A, same as degree of characteristic polynomial.)

Jordan canonical form
When for some eigenvalue λ of A, $n_g(\lambda) < n_a(\lambda)$, then that matrix A cannot be diagonalized by a similarity transformation, i.e. $(A \to S^{-1}AS)$, but we can obtain a canonical form that has all entries zero except the diagonal (which containing eigenvalues) and first super-diagonal (containing only zeros and ones). This nice form is called the Jordan canonical form. The presence of ones along the first super-diagonal is linked to the need to have generalized eigenvectors instead of just eigenvectors, since there are an insufficient number of independent eigenvectors for that repeated eigenvalue.

Finding solution x to $Ax = b$ when b is not in the image of A
Suppose A is full column rank and b is not in the image of A. Then, there does not exist an x such that $Ax = b$, but we can try to find x such that $\|Ax - b\|_2$ is minimized: this is least in the total least square sense.
When A is full column rank, then such an x is unique and can be found by solving:

$$A^T(Ax - b) = 0.$$

This above equation should be viewed as formulating the expectation that when x minimizes the error (i.e. $Ax - b$), then the error vector is perpendicular to all the columns of A, and hence $(Ax - b) \perp$ each column of A. In other words, each row of A^T is perpendicular to the error $Ax - b$, solving for such an x gives us the best x.
Thus, $x = (A^TA)^{-1}A^Tb$ minimizes the error in total-square-error sense.

Linear fit (least squares fit)
Given (x_j, y_j) for $j = 1, \ldots, N$, find w (weight) and b (bias) such that

$$y_j = wx_j + b \qquad \text{Linear fit with a bias: linear regression}$$

Higher order polynomial? Fix degree: say p

$$y_j = w_p x_j^p + w_{p-1}x_j^{p-1} + \cdots + w_1 x_j^1 + w_0$$

Coefficients w_i (parameters/unknowns enter linearly)
Easily solved using SVD: helps in minimizing the total square error.
When we are looking for degree $p = N - 1$, i.e. exact fit, we get the van-der-Monde matrix V

$$V := \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^p \\ 1 & x_2 & x_2^2 & \cdots & x_2^p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N-1} & x_{N-1}^2 & \cdots & x_{N-1}^p \end{bmatrix}, \text{ define } y := \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{N-1} \end{bmatrix}, \text{ and define } w := \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_p \end{bmatrix}.$$

Then, $y = Vw$, and matrix V is invertible (for $p = N-1$) if the x_i are all distinct. Thus, exact fit of a polynomial of sufficiently high order is always possible.
We are often more interested in a lower-order fit, but best in a total least square sense.

Recursive least squares
We will consider the case of a lower-order fit. Suppose $p \ll N$, i.e. we have many more measurements than unknowns. N is the number of measurements, and number of 'components per sample' is m (like $p+1$). In this case, $V \in \mathbb{R}^{N \times m}$ is very tall. SVD of $V^T V$ have some large singular values and some small (negligible) singular eigenvalues: look for 'knee'.
Need a way to update V, $V^T V$ and the current estimate $(V^T V)^{-1} V^T y$ as more measurements come.

- New data comes: measurement by measurement: one by one
- Need a 'fast' update: not a complete recalculation
- Rank one update $P_{N+1} = P_N + vv^T$ (v: new column vector)
- Can introduce slow 'forgetting' by giving more weightage to new data
- $P_{N+1} = (1-\alpha)P_N + \alpha vv^T$: with $\alpha > \frac{1}{N+1}$
- $\alpha = \frac{1}{N+1}$ means equal weightage to all measurements so far
- Also need to update the inverse of a matrix with new measurements
- Sherman/Morrison/Woodbury (or just matrix inverse lemma):

$$(A+UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1}$$

- Use matrix inverse lemma (special case):

$$(A+vv^T)^{-1} = A^{-1} - \frac{A^{-1}vv^T A^{-1}}{1+v^T A^{-1} v}$$

  (Verify this formula by evaluating: $(A+vv^T)\times$ RHS of above and use the fact that $v^T A^{-1} v$ is a scalar and this scalar can be pulled to the left from within long matrix/vector products.)

- Such rank-one updates need just $O(N^2)$ operations instead of $O(N^3)$ flops to recompute

Principal Component Analysis (PCA)
Unsupervised learning method:

- PCA: dimensionality reduction and for finding underlying laws
- It is used for spreading data along high-variation directions (to find groups)
- Suppose features/variables/measurements x_i, for $i = 1, \ldots, m$ are collected N times/samples.
- Number of samples = N (large)
- Number of features = m: each sample contains m components
- We are looking for (linear) relations between the m variables
- Along directions $\perp$ to linear relations: maximum variation: good for classification (when multiple classes)
- Thus, dimensionality reduction: unsupervised
- Given $x_j \in \mathbb{R}^m$ for $j = 1, \ldots, N$, find matrix $A \in \mathbb{R}^{q \times m}$ such that

$$Ax_j = 0 \qquad \text{Linear fit with zero bias: subtract mean}$$

- (Higher order polynomial? Include powers of x_j components
and 'lengthen' x_j not relevant/pursued further)
- Define $X := [x_1\ x_2\ \cdots\ x_j\ \cdots x_N] \in \mathbb{R}^{m \times N}$ and let $U\Sigma V^T = X$
- Σ: (largest) $\sigma_1, \ldots, \sigma_{m-q}$ (non-negligible), and (negligible) $\sigma_{m-q+1}, \ldots, \sigma_m$.(Choice of q: 'knee')
- Last q rows of U^T: valuable: the equations that each x_j satisfy
- Use U as new coordinates:
- first $m - q$ coordinates: principal 'components':
- directions with largest 'variation': degree of freedom is within these directions
- Unsupervised: after reducing the dimension: see classes getting separated automatically (if the data contains such separation).

Perron Frobenius theorem

- Many applications involve square matrices $A \in \mathbb{R}^{n \times n}$ with every entry positive
- Eigenvalue farthest (from origin) is positive & that left/right eigenvectors are significant
- All entries of A are not positive, but non-negative $\implies$ similar conclusions
(under graph-theoretic assumptions)

Largest magnitude $\equiv$ 'farthest' from the origin.

Fact 15 *(Perron Frobenius theorem) Consider $A > 0$ (entry-wise). Then,*

- *A has at least one real eigenvalue*
- *A has at least one real and positive eigenvalue*
- *Farthest λ_P is real and positive (farthest $\equiv$ farthest from origin)*
- *λ_P is not repeated. (Call λ_P: Perron root)*
- *No other eigenvalue λ of A is on the circle $|z| = \lambda_P$*
  (unique farthest)
- *All components in eigenvector v_P corresponding to λ_P has same sign: +ve*
- *Every other eigenvector (if real) has mixed signs.*
  (Unique eigenvector in positive 'orthant'.)

Many many applications of this theorem: Claim in Perron Frobenius theorem:
(Under some conditions): eigenvalue farthest (from 0) and corresponding eigenvector: both real and positive (components).
What quantities have to be real, and positive?

- Fraction of population in different age-groups: for one speci
- Population of various species: interacting with each other
- Population growth rates (for discrete time systems)
- Probabilities: between 0 and 1: for example, Markov chain, and probability of being in a state
- Rankings: relative values: but need positive
  Webpage rankings (for search engines) and ranking after a 'pairwise comparison'
- Prices of commodities: say a barter economy: 'pure' exchange
- Employee 'effort values' and corresponding distribution of net-profit

Google's Page Rank Algorithm: how to rank various webpages containing sought information?

- Assumption: more important webpages have more links coming into that page
- Construct graph with info about which webpage links to which, and construct a row-stochastic matrix A containing 'probabilities' of jumping to another page.
- A has all positive (in fact, non-negative) entries with $\lambda_P = 1$ and v_P has webpage's ranks: highest value: most important webpage
- Need 'fast convergence' in the power method: trade-off between high spectral gap and relevance of the computed rankings