overview

Here is a quick run-down on these notes, with various terms to be learned in **boldface**.

Much of scientific work involves relationships called **map**s:

$$f : X \to Y : x \mapsto f(x)$$

For example,

- ∘ time $\mapsto$ the population of the US;
- ∘ temperature $\mapsto$ pressure in a bottle;
- ∘ location (longitude, latitude, altitude) $\mapsto$ (barometric pressure, humidity, temperature);
- ∘ mother's age $\mapsto$ frequency of newborn with Down syndrom
- ∘ available resources (capital, raw materials, labor pool, etc) $\mapsto$ output of the US economy
- ∘ etc.

All this is part of our hope to understand effects in terms of causes.

Once we feel we understand such a relationship, we are eager to put it to use in order to find out how to cause certain effects. Mathematically, we are trying to solve the equation:

$$f(?) = y$$

for given $f : X \to Y$ and given $y \in Y$.

In this generality and vagueness, nothing much can be said other than to urge familiarity with basic map terms, such as, **domain**, **target** and **range** of a map, the map properties **1-1** (equivalent to **uniqueness** of solutions), **onto** (equivalent to **existence** of a solution for any y), **invertible** (equivalent to having exactly one solution for any $y \in Y$, the best-possible situation), and the notions of **left inverse**, **right inverse** and **inverse** related to the earlier notions by the concept of **map composition**.

Often, though, the map f is a **smooth** map, from some subset X of **real** n**-dimensional coordinate space** $\mathbb{R}^n$ to $\mathbb{R}^m$, say. With the list $x = (x_1, \ldots, x_n)$ our notation for $x \in \mathbb{R}^n$, this means that, first of all,

$$f(x) = (f_1(x), f_2(x), \ldots, f_m(x)) \ \in \mathbb{R}^m$$

with each f_j a **scalar**-valued function, and, secondly, at any point $p \in X$, we can expand each f_j into a Taylor series:

$$f_j(p + h) = f_j(p) + Df_j(p)^{\mathrm{t}} h + o(h), \quad j = 1, \ldots, m,$$

with

$$Df_j(p) = (D_1 f_j(p), \ldots, D_n f_j(p)) \ \in \mathbb{R}^n$$

the **gradient** of f_j at p, and $x^{\mathrm{t}} y$ the **scalar product** of the n**-vector**s x and y, and the $o(h)$ denoting 'higher-order' terms that we eventually are going to ignore in best scientific fashion.

This implies that

$$f(p + h) = f(p) + Df(p)h + o(h),$$

with

$$Df(p) = \begin{bmatrix} D_1 f_1(p) & \cdots & D_n f_1(p) \\ \vdots & \cdots & \vdots \\ D_1 f_m(p) & \cdots & D_n f_m(p) \end{bmatrix}$$

the **Jacobian** matrix of f at p.

With this, a standard approach to finding a solution to the equation

$$f(?) = y$$

is **Newton's method**: If x is our **current guess** at the solution, we are looking for a **correction** h so that

$$y = f(x+h) = f(x) + Df(x)h + o(h);$$

we ignore the 'higher-order' terms that hide behind the expression $o(h)$, and so get a *linear equation* for h:

$$y - f(x) = Df(x)?,$$

which we solve for h, add this correction to our current x to get a new guess

$$x \leftarrow x + h = x + Df(x)^{-1}(y - f(x))$$

and repeat. Under suitable circumstances, the process converges, to a solution.

The *key idea* here is the reduction, from solving a general equation $f(?) = y$ to solving a sequence of **linear** equations, $Df(x)? = z$. This works since, in principle, we can always solve a linear system.

Most equations $f(?) = y$ that can be solved are actually solved by this process or a variant thereof, hence the importance of knowing how to solve *linear* equations.

For this reason, our first task will be to introduce **linear map**s and **linear space**s, especially **linear spaces of functions**., i.e., vector spaces in which the basic **vector operations**, namely **vector addition** and **multiplication by a scalar**, are defined **pointwise**. These provide the proper means for expressing the concept of **linearity**. Then we recall **elimination** as the method for solving a **homogeneous** linear system

$$A? = 0$$

with $A \in \mathbb{R}^{m\times n}$. Specifically, we recall that elimination classifies the unknowns as **bound** and **free**, and this leads to **row echelon forms**, in particular the **rrref** or **really reduced row echelon form**, from which we can obtain a complete description of the solution set of $A? = 0$, i.e., for $\operatorname{null} A$, the **nullspace** of A, as well as an efficient description of $\operatorname{ran} A$, the **range** of A. Thus equipped, we deal with the general linear system $A? = b$ via the homogeneous linear system $[A, b]? = 0$.

Both $\operatorname{null} A$ and $\operatorname{ran} A$ are typical examples of **linear subspace**s, and these efficient descriptions for them are in terms of a **basis**, i.e., in terms of an invertible linear map V from some **coordinate space** $\mathbb{F}^n$ to the linear subspace in question. This identifies bases as particular **column map**s, i.e., linear maps from some coordinate space, i.e., maps of the form

$$\mathbb{F}^n \to X : a \mapsto a_1v_1 + \cdots a_nv_n =: [v_1, \ldots, v_n]a$$

for some sequence $v_1, \ldots, v_n$ in the linear space X in question.

We'll spend some time recalling various details about bases, how to construct them, how to use them, and will also mention their generalization, **direct sum**s and their associated **linear projector**s or **idempotent**s. We stress the notion of **dimension** (= number of columns or elements in a basis), in particular the **Dimension Formula**

$$\dim \operatorname{dom} A = \dim \operatorname{ran} A + \dim \operatorname{null} A,$$

valid for any linear map A, which summarizes much of what is important about dimension.

We'll also worry about how to determine the **coordinates** of a given $x \in X$ with respect to a given basis V for X, i.e., how to solve the equation

$$V? = x.$$

This will lead us to **row maps**, i.e., linear maps from some linear space to coordinate space, i.e., maps of the form

$$X \to \mathbb{F}^n : x \mapsto (\lambda_1 x, \cdots, \lambda_n x) =: [\lambda_1, \ldots, \lambda_n]^{\mathrm{t}} x$$

for some sequence $\lambda_1, \ldots, \lambda_n$ of **linear functional**s on the linear space X in question. It will also lead us to **interpolation** aka **change of basis**, and will make us single out **inner product space**s as spaces with a ready supply of suitable row maps, and thence to **least-squares**, to particularly good bases, namely **o.n.** (:= **orthonormal**) bases (which are the **isometries** for the standard **norm**, the **Euclidean norm** $\|x\|_2 = \sqrt{x^{\mathrm{t}}x}$ associated with the standard **inner product** and which can be constructed from an arbitrary basis by **Gram-Schmidt**).

We'll find that bases also show up naturally when we try to **factor** a given linear map $A \in L(X,Y)$ in the most efficient way, as a product

$$A = V\Lambda^{\mathrm{t}}$$

with $\Lambda^{\mathrm{t}} \in L(X, \mathbb{F}^r)$ and $V \in L(\mathbb{F}^r, Y)$ and r as small as possible. It will be one of my tasks to convince you that you have actually carried out such factorizations, in fact had to do this in order to do certain standard operations, like differentiating or integrating polynomials and other functions. Such factorizations are intimately connected with the **rank** of A (since the smallest possible r is the rank of A) and lead, for a matrix A, to the **SVD**, or **Singular Value Decomposition**,

$$A = V\Sigma W^{\mathrm{c}}$$

with V, W o.n. and Σ diagonal, a factorization that is, in a certain sense, a best way of describing the action of the linear map A. Other common factorizations for matrices are the **PLU factorization** with P a **permutation matrix**, L **unit lower triangular**, and U **upper triangular** (generated during elimination); and the (more stable) **QR factorization**, with Q **unitary** (i.e., an o.n. basis) and R **upper**, or, **right triangular**, obtained by elimination with the aid of specific **elementary matrices** called **Householder reflection**s.

For *square* matrices, one hopes to (but does not always) get factorizations of the form $A = V\Sigma V^{-1}$ with Σ diagonal (the simplest example of a matrix without such a factorization is the **nilpotent** matrix $\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$), but often must be (and is) content to get the **Schur form**, which is available for any square matrix and is of the form $A = VUV^{\mathrm{c}}$ with V an o.n. basis and U upper triangular. In either case, A is then said to be **similar** to Σ and U, respectively. These latter factorizations, or **similarities**, are essential for an understanding of the **power sequence**

$$A^0 = \mathrm{id}, A^1 = A, A^2 = AA, A^3 = AAA, ...$$

of the square matrix A and, more generally, for an understanding of the **matrix polynomial** $p(A)$, since, e.g.,

$$A = V\,\mathrm{diag}(\mu_1, \ldots, \mu_n)V^{-1} \quad \Longrightarrow \quad p(A) = V\,\mathrm{diag}(p(\mu_1), \ldots, p(\mu_n))V^{-1},$$

for any polynomial p and even for some well-behaved functions p like the **exponential** $p : t \mapsto \exp(t)$. In particular, then

$$A^k = V\,\mathrm{diag}(\mu_1^k, \ldots, \mu_n^k)V^{-1}, \quad k = 0, 1, 2, \ldots,$$

therefore we can describe the behavior of the *matrix* sequence $(A^k : k = 0, 1, 2, \ldots)$ entirely in terms of the *scalar* sequences $(\mu_j^k : k = 0, 1, 2, \ldots)$. Specifically, we can characterize **power-boundedness**, **convergence**, and **convergence to** 0.

There are many reasons for wanting to understand the power sequence of a matrix; here is one. Often, elimination is not the most efficient way to solve a linear system. Rather, the linear system

$$A? = b$$

itself is solved by iteration, by splitting $A =: M - N$ with M 'easily' invertible, and looking at the **equivalent equation**

$$M? = N? + b$$

which leads to the **iteration**

$$x \leftarrow M^{-1}(Nx + b) =: Bx + c.$$

Convergence of this process depends crucially on the behavior of the power sequence for B (and does not at all depend on the particular **vector norm** or **map norm** used).

The factorization

$$A = V \operatorname{diag}(\mu_1, \ldots, \mu_n) V^{-1}$$

is equivalent to having $AV = V \operatorname{diag}(\mu_1, \ldots, \mu_n)$, i.e.,

$$[Av_1, \ldots, Av_n] = [\mu_1 v_1, \ldots, \mu_n v_n]$$

for some invertible $V = [v_1, \ldots, v_n] : \mathbb{F}^n \to \operatorname{dom} A$, i.e., to having a basis V consisting of **eigenvector**s for A, with the μ_j the corresponding **eigenvalue**s. For this reason, we'll study the **eigenstructure** of A and the **spectrum** of A, as well as **similarity**, i.e., the **equivalence relation**

$$A \sim C \;:=\; \exists V \;\; A = VCV^{-1}.$$

In this study, we make use of polynomials, particular the **annihilating polynomial**s (which are the nontrivial polynomials p for which $p(A) = 0$) and their cousins, the nontrivial polynomials p for which $p(A)x = 0$ for some $x \neq 0$, and the unique **monic** annihilating polynomial of minimal degree, called the **minimal polynomial** for A, as well as the **Krylov sequence** $x, Ax, A^2x, \ldots$.

We'll discuss the most important classification of eigenvalues, into **defective** and **non-defective** eigenvalues, and give a complete description of the asymptotic behavior of the power sequence $A^0, A^1, A^2, \ldots$ in terms of the eigenstructure of A, even when A is not **diagonalizable**, i.e., is not similar to a diagonal matrix (which is equivalent to some eigenvalue of A being defective).

We'll also discuss standard means for locating the spectrum of a matrix, such as **Gershgorin's circles** and the **characteristic polynomial** of a matrix, and give the Perron-Frobenius theory concerning the dominant eigenvalue of a positive matrix.

From the Schur form (*vide supra*), we derive the basic facts about the eigenstructure of **hermitian** and of **normal** matrices. We give the **Jordan form** only because of its mathematical elegance since, in contrast to the Schur form, it cannot be constructed reliably numerically.

As a taste of the many different applications of Linear Algebra, we discuss briefly: the solution of a system of constant-coefficient ODEs, Markov processes, subdivision in CAGD, Linear Programming, the Discrete Fourier Transform, approximation by broken lines, and the use of flats in analysis and CAGD.

Further, we also consider briefly **minimization** of a real-valued map

$$f : K \to \mathbb{R}$$

with $K \subset \mathbb{R}^n$. Returning to our Taylor expansion

$$f(p+h) = f(p) + Df(p)^{\mathrm{t}} h + o(h),$$

we notice that, usually, p cannot be a minimum point for f unless it is a **critical point**, i.e., unless the gradient, $Df(p)$, is the zero vector. However, even with $Df(p) = 0$, we only know that f is 'flat' at p. In particular, a critical point could also be a (local) maximum point, or a saddle point, etc. To distinguish between the various possibilities, we must look at the **second-order** terms, i.e., we must write and know, more explicitly, that

$$f(p+h) = f(p) + Df(p)^{\mathrm{t}} h + h^{\mathrm{t}} D^2 f(p) h/2 + o(h^{\mathrm{t}} h),$$

with

$$H := D^2 f := \begin{bmatrix} D_1 D_1 f & \cdots & D_1 D_n f \\ \vdots & \cdots & \vdots \\ D_n D_1 f & \cdots & D_n D_n f \end{bmatrix}$$

the **Hessian** for f, hence

$$h \mapsto h^{\mathrm{t}} D^2 f(p) h = \sum_{i,j} D_i D_j f(p) h_i h_j$$

the associated **quadratic form**.

We will learn to distinguish between maxima, minima, and saddle points by the signs of the eigenvalues of the Hessian, mention **Sylvester's Law of Inertia**, and show how to estimate the effect of **perturbations** on H on the spectrum of H, using ideas connected with the **Rayleigh quotient**.

At this point, you will realize that these notes are strongly influenced by the use of Linear Algebra in Analysis, with important applications, e.g., in Graph Theory, ???, or ???, being ignored (partly through ignorance).

Finally, although **determinant**s have little to contribute to Linear Algebra at the level of this book, we'll give a complete introduction to this very important Linear Algebra tool, and then discuss the **Schur complement**, **Sylvester's determinant identity**, and the **Cauchy-Binet formula**.

Throughout, we'll rely on needed material from prerequisite courses as collected in an appendix called **Background**.