Contents

1 Principles of Natural Sciences **2**
1.1 Mathematics . . . 2

2 Linear Algebra **4**
2.1 Overview of Linear Algebra . . . 4
2.2 Vectors . . . 8
2.3 System of Linear Equations . . . 10
2.3.1 Numeric Methods for SLE . . . 12
2.4 Matrices . . . 13
2.4.1 Special Matrices . . . 19
2.5 Vector Space . . . 19
2.6 Determinants and Trace . . . 22
2.7 Eigenvalues and Eigenvectors . . . 25
2.7.1 Numerical Methods for Computing Eigenvalues . . . 29
2.7.2 Finite State Markov Chains . . . 30
2.8 Orthogonality . . . 31
2.9 Symmetric Matrices and Quadratic Forms . . . 34
2.10 Distance and Approximation . . . 38
2.11 Linear Algebra Review . . . 42

3 Calculus & Analysis **45**
3.1 Calculus . . . 45
3.1.1 Calculus of Field . . . 48
3.1.2 Vector and Matrix Calculus . . . 49
3.2 Real and Functional Analysis . . . 50
3.3 Numerical Methods . . . 54
3.4 Optimization . . . 55
3.4.1 Convex Optimization . . . 59

4 Discrete Mathematics & Algorithms **69**

5 Differential Equations, Dynamics Systems & Stochastic Processes **70**
5.1 Dynamic Systems and Network Theory . . . 70
5.2 Stochastic Processes . . . 72

6 Thermodynamics & Statistical Mechanics **76**
6.1 Kinetic Theory of Ideal Gas . . . 76
6.2 Entropy and Free Energy . . . 77
6.3 Entropic Forces . . . 81
6.4 Chemical Systems . . . 82

Chapter 1

Principles of Natural Sciences

1.1 Mathematics

Development of a mathematical theory:

- Intuitive concepts/ideas. Ex. in Euclidean space, the minimal distance from a point to a plane is easy to obtain (the projection). Could be applicable to general situations if we properly generalize the concepts in Euclidean space.
- Formal constructs. Ex. vector space, inner product, projections.
- Propositions and theorems from the constructs. Ex. The same result holds: the projection of a point to a plane has the minimal distance. This can used to solve more problems, such as linear regression.

How to solve a mathematical problem?

- Scientific approach to mathematics (Investigative spirit): treat the mathematical problem as a scientific one, ask what would be implied for a mathematical object/proposition (hypothesis).
  - Ex. finding the maximum of an equation. Hypothesis: x maximizes function f. Finding evidence: if x maximizes f, its signature is that $f'(x)$ becomes 0.
  - Ex. suppose we want to find eigenvalues of a matrix, ask what are the signatures of eigenvalues (if eigenvalue is large, what would reflect that).
- Analytic/backward approach: to solve a problem, think of what is needed to solve it.
- Forward approach: given the conditions, think of what are their consequences, what we can learn/infer.

Representation:

- **Principle**: express an object in terms of basic objects (using basic operations defined)
- Most importantly: factorization; and polynomial (linear) expansion
- Note: the concept of space is closely related. The relations of objects are made clearer by working in a space of objects.

Approximation:

- **Principle**: an object is studied via other objects that are "close" to this object.
- Special case: limit or convergence of a sequence of objects

Generalization and Specialization:

- **Principle**: study first the basic/simple objects, then use them to build the general cases.

Connection/reformulation:

- **Principle**: recognize the connection with other mathematical branches, or view from the perspective of another branch.
- Most important example: geometric view.
- **Remark**: geometric view is good because it makes clear the relations of objects, or gives global view of objects. It is natural to define collections of objects as new objects and study them in geometry. For instance, a collection of points (basic objects) is a curve, and we can study the properties of curves.
- Ex. a function $y = f(x)$ is defined algebraiclly/analytically by a point-to-point mapping of x and y; the geometic view gives the overall shape.

Invariance:

- **Principle**: recognize the quantity/measure that does not change under transformations/changes. In general, if there exists such quantity, then explicitly define it, as it will be important to characterize a process.

Measure:

- **Principle**: characterize certain aspect of an object by a number.
- Natural questions to ask following definition of a new measure is: how to compute this measure; how the measures of related objects are related; how this measure is related to other properties of objects, etc.

Chapter 2

Linear Algebra

Reference: Leon [2006]; Poole [2006].

2.1 Overview of Linear Algebra

Motivation: why need linear algebra?

- Vectors: generalize the concept of scalar variables (numbers) - often we need to use multiple components (or a magnitude with direction) to represent an object of interest. Ex. we use vectors to represent the state of a system, the location in space, the multiple observations of a random variable, the different features of an object, and so on. We can then formulate many problems in terms of vectors. Some main classes of problems are listed below.
- Solving equations: we have multiple equations of multiple unknowns, and this is represented by a system of linear equations, $Ax = b$.
- Function optimiation: linear programming problem, e.g. max Ax s.t. $x_i \leq b_i, \forall i$. Optimization of quadratic forms, e.g. max $x^T Ax$ s.t. $\|x\| \leq 1$.
- Dynamic systems: let $x(t)$ be the state of the system at t, we have a linear system $dx/dt = Ax$, and we study the behavior of $x(t)$ as $t \to \infty$.
- Multivariate probability density function: the PDF may be centered around a point in high-D space, and the density spreads across an elliptic region (as in MVN), the density can thus be represented using ellipse or quadratic forms, $(x - \mu)^T \Theta (x - \mu)$.
- Formulating problems in matrix terms: the problems involving multiple unknowns in optimization, SLE, dynamic system, etc. can be stated in terms of operations between vectors and matrices. If we formulate the algebra of matrices, and the operations on matrices can be easiy solved, then the original problem is considerably simplified.
  - System of linear equations: suppose we need to solve $Ax = b$, where A is $n \times n$. The consistency of the SLE is equivalent to the invertibility of the matrix A, and the solution is $x = A^{-1}b$. Thus the problem is reduced to analyzing A^{-1}, which can solved by algebraic means. Ex. if we can write as BC, where the inverse of B and C are easy to obtain, then we have $x = C^{-1}B^{-1}b$.

Important problems that motivate linear algebra:

- Systems of linear equations: $Ax = b$. When does the system have a unique solution? Or generally, what is the dimension of the solution?

- Linear map: $f(x) = Ax$, what is the effect of this map on a vector x? Ex. does it make $\|x\|$ bigger?
- Optimization of functions of quadratic form: $\sum_{i,j} a_{ij}x_ix_j = x^TAx$, what is its maximum/minimum?

Key themes of linear algebra: linear map and linear/orthogonal representation

- **Finding a good representation/basis**: linear algebra concerns the effect of a linear map Ax, where $x \in \mathbf{R}^n$. If x has a linear representation $x = \sum_i c_iq_i$, where $\{q_i\}$ is the basis of $\mathbf{R}^n$, then $Ax = \sum_i c_i(Aq_i)$. So if we can find an appropriate basis, then Ax could have a simple representation. In particular, if q_i's are orthogonal and Aq_i are also orthogonal, the representation of Ax with Aq_i as basis would be simple.
- The importance of a good representation/basis is generally important. Ex 1. Fourier transform: use trigometric function as basis. Ex. 2. in statistics, use mixture of normal to represent more general shapes of distributions.
- Eigen-decomposition and Spectrum Decomposition: we choose the basis as the eigenvectors of A, denoted as q_i. Then $Aq_i = \lambda_iq_i$. So the effect of Ax is:

$$(c_i) \to (\lambda_ic_i) \tag{2.1}$$

  Using the basis of $\{q_i\}$, we simply scale the coordinates of x by λ_i.
- SVD: we choose the basis as the eigenvectors of A^TA, denoted as v_i. They form an orthogonal basis of $\mathbf{R}^n$. Then $Av_i = \sigma_iu_i$, where u_i's are eigenvectors of AA^T. Then u_i's form an orthogonal basis of $\mathbf{R}^m$. So the effect of Ax is:

$$(c_i) \to (\sigma_ic_i) \tag{2.2}$$

  So with basis of v_i in $\mathbf{R}^n$ for x, the coordinates of Ax is scale by σ_i using the basis of u_i in $\mathbf{R}^m$.
- Applications of matrix representations: we can study how norm of x changes by Ax, leading to quadratic forms. We can also study the effect of matrix power, which is determined by the largest eigenvalues.

General ideas of matrix and linear algebra:

- **Geometric perspective**: in linear algebra, we have correspondance among different mathematical objects, e.g. matrix A, linear map $f(x) = Ax$, SLE $Ax = 0$, and so on. In many situations, switch/refomulate the problem in a different perspective. The most widely used perspective is: view an algrberaic problem from geometric perspective. Examples:
  - Rank of a matrix and the number of independent variables in SLE: dimensions of the space of solutions. Ex. 2D SLE $Ax = 0$, normally (full ranked), it has a unique solution (dim. 0). If the row vectors of A are parallel, it has infinitely many solutions (dim. 1).
- **Equivalent Representations**: a fundamental idea of algebra. The same thing can be manifested as different combinations of the same underlying elements, using algebraic rules. For example, summation of many terms can often be written in different forms, e.g. summing of the same terms but in different orders. For instance, this result:

$$\operatorname{tr}(AB) = \operatorname{tr}(BA) \tag{2.3}$$

  is due to a simple switching of the order in summation:

$$\sum_i \sum_k a_{ik}b_{ki} = \sum_k \sum_i b_{ki}a_{ik}. \tag{2.4}$$

  Another example, $Ax = 0$ can be expressed in multiple ways.
- **Reduction**: to solve a general problem, first solve it for a special case, then reduce the general problem to the simpler forms. Two critical components of the strategy: special cases; and the rules of reduction.

  - System of linear equations: $Ax = b$, if A is upper triangular, then we can solve it. Rules: elmentary row operations.
  - Determinants: reduce the determinant of a general matrix to a triangular form, which is easy to solve. Rules: e.g. $\det(AB) = \det A \det B$.

  Ideas of factorization and expansion can be viewed as an application of the reduction principle.

- **Factorization**: writing matrix as a product of simpler matrices, this could help understand the original matrix. Ex. each simpler matrix may have a simple geometric interpretation (such as rotations), or each matrix is a simple row operation. In general, factorization is a fundamental algebraic idea: prime factorization, Fundamental Theorem of Algebra. Once a matrix is factorized, we can reduce the problem we are facing into simpler problems, e.g. solving SLE.

- **Expansion**: express a mathematical object as some function of more "basic" objects, often some kind of linear combinations. Examples: Fourier series, or expression of a vector as a linear combination of basic vectors. The idea can be expressed geometrically, where each dimension corresponds to a basic object.

- Mathematical **Transformations**: transformations or operations on objects: this is how "reduction" can be realized. Ex. LU factorization that expresses a matrix as product of elementary row operations. These transformations in linear algebra can often be represented by multiplication of matrices or vectors. This could include, for example, forming the covariance matrix (dot product of all columns, or all rows), extraction of specific rows or columns, rotation of vectors, orthogonalization of a basis (QR factorization), and so on.

  - Understanding the effect of mathematical transformations: an important idea is to see what is "conserved" from the transformations, or what is a simple way to characterize the effect of transformations. Ex. elementary row operations never change the linear dependency of row vectors; the effect of Ax can be understood using the eigenvalue of A.

- **Orthogonality**: a fundamental idea in linear algebra (and beyond). In general, find orthogonal representation of an object (i.e. in terms of orthogonal basis). If the vectors are orthogonal, often things are much easier, so a major approach is to cast the problem of a general matrix into that of orthogonal vectors.

  - Solving SLE: suppose we are solving $Ax = b$, if A is orthogonal, then we write this as: $\sum_i x_i A^i = b$, where A^i is column vector. So x_i can be easily determined by projection of b on A^i: $x_i \langle A_i, A_i \rangle = \langle b, A_i \rangle$. In general, we can use QR factorization to reduce A into orthogonal matrix.
  - Diagnolization of quadratic forms $x^T Ax$: if A is symmetric, we write $A = QDQ^T$, then $x^T Ax = (Q^T x)^T D(Q^T x)$, where D is diagonal.
  - PCA: we have data matrix X with p RVs. Some RVs are highly correlated. Intuitively, we want to find a set of independent RVs, i.e. orthogonal vectors, that "explain" the data. Ex. if X_1 is highly correlated with X_2, we find U_1 to be the "average of X_1 and X_2, and U_2 orthogonal to U_1 that explains the residuals.

- **Statistical perspective**: we can basic linear algebra objects in statistical terms. A vector can be viewed as a RV, norm can be viewed as variance of a RV, inner product of two vectors as covariance or correlation of two RVs. $X^T X$ can be viewed as covariance matrix. $x^T Ax$ where A is positive definite can be viewed as PDF of MVN distribution.

  - Example: rank preservation, $rank(A^T A) = rank(A)$. This can be understood as the covariance structure does not change the linear dependency of the variables.

Geometric perspectives of linear algebra:

- Concepts: the properties of and operations on vectors can be represented by geometric concepts including length, distance, angles, and so on. Most importantly, Ax where A is an matrix can be viewed as the application of linear map to x. Then an algebraic problem can be stated in geometric terms, and vice versa: a function can be represented by a geometric transformation (e.g. rotation) and an equation can be represented by a geometric shape/object(e.g. $\|x\| = 1$ is a circle).
- Properties of matrices/linear transformations: the most interesting problems include, the effect of A on the length of a vector, the direction of a vector, and the angle between two vectors.
- **System of linear equations**: suppose we need to solve $Ax = b$, we can view it in two ways: (1) Row vector view: for each m, $A_m \cdot x = b_m$, each equation corresponds to a hyperplane (line in 2D case), and the solution is the intersection of the hyperplanes. (2) Column vector view: we write this as: $x_1 A^1 + \cdots + x_n A^n = b$, where A_j is the column vector. The problem is thus to write b as a linear combination of vectors $A^1, \cdots, A^n$, and find the coordinates of b. In 2D, we can solve this geometrically (start at b, draw parallel lines of A^1 and A^2).
- Direction of vectors by transformation: directions that do not change by A: eigenvectors.
- Power of matrix: can be understood as successively application of the linear map corresponding to the matrix. Thus the power of the matrix can be understood using the geometry of linear map: e.g. in the direction of eigenvectors, the effect of multiple application of the map is simple (scaling).
- Determinants: the concept of volume. It quantifies the effect of a matrix on the volume of an object.
- Quadratic forms: can be understood as rotations of ellipse plus other transformations (translation).

Example: quadratic forms and ellipses

- Roation in 2D: suppose we have a point (x, y), after rotation counterclockwise by θ, the new point becomes (x', y') (using polar coordinate):

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \tag{2.5}$$

  Thus the rotation operation can be represented by a matrix shown above.
- Rotation matrix: in general, any rotation can be represented by a linear map (thus matrix). It is easy to show based on geometry that the rotation is a linear map: $f(u+v) = f(u)+f(v)$ and $f(\lambda v) = \lambda f(v)$. Next, rotation preserves the distance/norml of a vector, let Q be the matrix of the rotation, we have, for any vector u:

$$\|u\| = \|Qu\| \Rightarrow u^T u = (Qu)^T (Qu) = u^T Q^T Q u \tag{2.6}$$

  Since this is true for any u, we must have $Q^T Q = I$, i.e. Q is an orthogonal matrix. Furthermore, we can show that $\det Q = 1$ since rotations preserve handedness.
- Ellipse: it can be written in the form $\|D^{1/2} x\| = 1$, where D is a diagonal matrix $\text{Diag}(\sqrt{d_1}, \cdots, \sqrt{d_n})$.
- Rotation of ellipse: let x be a vector in the rotated ellipse, then to obtain the equation of x, we first rotate x back to the position perpidencular to the axix (clockwise by θ), then the new vector should satisfy the equation of ellipse. Let Q be the rotation matrix, we have:

$$\|D^{1/2} Q x\| = 1 \Rightarrow x^T Q^T D Q x = 1 \tag{2.7}$$

  Let $A = Q^T D Q$, we have $x^T A x = 1$. Thus any quadratic equation where A is a postive definive matrix can be viewed as an rotated ellipse.
- Lesson: when solving an algebraic problem, $x^T A x = 1$ in this case, represent it in terms of geometric objects, and use geometry to understand the transformations (norm does not change by rotations).

Questions of linear algebra and matrix [personal notes]:

- Low-rank matrices: eigenvalues and eigenvectors?

2.2 Vectors

Dot product and basic geometric concepts:

- Dot product: given two vectors $u, v \in \mathbb{R}^n$, the dot product is defined as:

$$u \cdot v = \sum_i u_i v_i \tag{2.8}$$

  Note that the dot product is a special case of inner product in Eucledian space.

- Length and distance: the length of a vector is $\|v\| = \sqrt{v \cdot v}$, and the distance between two vectors, $d(u, v) = \|u - v\|$.

- Angle: for non-zero vectors $u, v \in \mathbb{R}^n$, the angle is defined as:

$$\cos\theta = \frac{|u \cdot v|}{\|u\|\|v\|} \tag{2.9}$$

  Two vectors are orthogonal if $u \cdot v = 0$.

- Projection: given $u, v \in \mathbb{R}^n$, the projection of v onto u is defined as:

$$\text{proj}_u(v) = \frac{u \cdot v}{u \cdot u} u \tag{2.10}$$

  Thus it is the coordinate along the direction of u. An important special case is u is a unit vector, then the projection is simply $(u \cdot v)u$.

Fundamental Theorems:

- Pythagoras' Theorem: for all vectors $u, v \in \mathbb{R}^n$, $\|u + v\|^2 = \|u\|^2 + \|v\|^2$ if and only if u and v are orthogonal.

- Cauchy-Schwarz Inequality: for all vectors $u, v \in \mathbb{R}^n$, $|u \cdot v| \leq \|u\|\|v\|$.
  Proof 1: geometric proof, the LHS is the area of parallelogram bounded by u and v, and it is the product of $\|u\|$ and the height, which should be at most the norm of v. Consider the projection of v onto u, we have:

$$v = \frac{u \cdot v}{u \cdot u} u + \left(v - \frac{u \cdot v}{u \cdot u} u \right) \tag{2.11}$$

  By Pythagoras' Theorem, we have $\|\frac{u \cdot v}{u \cdot u} u\| \leq \|v\|$. Rewrite this inequality and we have the proof.

  Proof 2: algebraic proof. In the 2D case, we prove: $(u_1 v_1 + u_2 v_2)^2 \leq (u_1^2 + u_2^2)(v_1^2 + v_2^2)$. This can be extended to any dim. by induction.

- Triangle Inequality: for all vectors $u, v \in \mathbb{R}^n$, $\|u + v\| \leq \|u\| + \|v\|$.
  Proof: follows easily from the Cauchy-Schwarz Inequality.

Application of vector algebra to statistics:

- Sample variance and covariance: given a random variable X, with n iid. samples, $x_1, \cdots, x_n$, and $y_1, \cdots, y_n$. The sample variance of X is:

$$S_X = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2 = \frac{1}{n-1} \|x - \bar{x}\|^2 \tag{2.12}$$

where x is the n-dim. vector. If $\bar{x} \approx 0$, sample variance is simply the norm of x; otherwise, it is the norm of the vector $x - \mu\mathbf{1}$ (the identity vector of 1's). The sample covariance between X and Y:

$$S_{XY} = \frac{1}{n-1}\sum_i (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1}(x - \bar{x})^T(y - \bar{y}) \tag{2.13}$$

which is the dot product between the two vectors. The sample correlation coefficient is:

$$\rho_{XY} = \frac{S_{XY}}{\sqrt{S_X S_Y}} = \cos\theta \tag{2.14}$$

where θ is the angle between the two vectors, $x - \bar{x}$ and $y - \bar{y}$.

- Variance of sum of independent RVs: suppose $X = Y + Z$ is the sum of two independent random variables, then based on the Pythagoras' Theorem, we have:

$$S_X = S_Y + S_Z \tag{2.15}$$

  Or in terms of the random variables, $\operatorname{Var} X = \operatorname{Var} Y + \operatorname{Var} Z$.

- Covariance: we have

$$S_{XY} = \frac{1}{n-1}(x - \bar{x})^T(y - \bar{y}) \leq \frac{1}{n-1}\|x - \bar{x}\|\|y - \bar{y}\| = \sqrt{S_X S_Y} \tag{2.16}$$

  based on Cauchy-Schwarz Inequality. Or in terms of the random variables, $\operatorname{Cov}(X, Y) \leq \sqrt{\operatorname{Var} X} \cdot \sqrt{\operatorname{Var} Y}$.

- Remark: for any problem, where we can define a vector (a set of n components) and the corresponding operations make sense (addition, scalar multiplication, dot product), we could apply the results of vector algebra. The other applications may include, for example, the coefficients of functions (power series) as vectors.

Vector representation of lines and planes (hyperplanes): in general, geometric objects/shapes can be represented by equations defined on vectors. Generally, there are two ways: equation where RHS is 0 and equation using parameter(s).

- Normal equation representation: a line in $\mathbb{R}^2$, a plane in $\mathbb{R}^3$ or a hyperplane in $\mathbb{R}^n$ can be represented by a normal equation. Let $\vec{n}$ be the vector orthogonal to the line/plane/hyperplane (call it a plane), and let $\vec{p}$ be a point in the plane, then for any vector $\vec{x}$ in the plane, we have:

$$\vec{n} \cdot (\vec{x} - \vec{p}) = 0 \tag{2.17}$$

  This can be easily written in the form of the general equation of a hyperplane: $\sum_{i=1}^n a_i x_i = b$, or $a^T x = b$ where $a, x \in \mathbb{R}^n$. To write this in the normal equation form, we can write it as $a^T(x - p) = b$, where p is chosen s.t. $a^T p = b$ - we can choose a special value, e.g. all components are 0 except the last one, $p_n = a_n/b$.

- Vector equation representation: alternatively, one can also use the direction vector to represent a line. Let $\vec{d}$ be a vector along the direction of a line L, and $\vec{p}$ be a point in L, then any point in L can be represented by (called parametric equation):

$$\vec{x} = \vec{p} + t\vec{d} \tag{2.18}$$

  where $t \in \mathbb{R}$, and as t varies, we have the entire line. The vector representation of plane is similar, though $n - 1$ parameters are required for a hyperplane in $\mathbb{R}^n$, e.g. for $n = 3$:

$$\vec{x} = \vec{p} + s\vec{u} + t\vec{v} \tag{2.19}$$

  where $\vec{u}$ and $\vec{v}$ are two direction vectors of the plane.

- Distance of vector to lines and planes using normal equation: let $\vec{v}$ be a vector and we want to find the distance of $\vec{v}$ to the plane, P defined by $\vec{n} \cdot (\vec{x} - \vec{p}) = 0$. We form the projection of $\vec{v} - \vec{p}$ onto the vector $\vec{n}$, then the norm is the distance. So we have:

$$d(\vec{v}, P) = \|\text{proj}_{\vec{n}}(\vec{v} - \vec{p})\| = \frac{(\vec{v} - \vec{p}) \cdot \vec{n}}{\|\vec{n}\|} \tag{2.20}$$

- Distance of vector to lines using vector equation: let $\vec{v}$ be a vector and we want to find the distance of $\vec{v}$ to the line, L defined by $\vec{x} = \vec{p} + t\vec{d}$. Let O be the projection of $\vec{v} - \vec{p}$ onto d, then the distance is the norm of the vector $\vec{v} - O$. We have:

$$d(\vec{v}, L) = \|\vec{v} - \vec{p} - \frac{(\vec{v} - \vec{p}) \cdot \vec{d}}{\vec{d} \cdot \vec{d}} \vec{d}\| \tag{2.21}$$

- Remark: any linear equation or a linear function can be understood geometrically as a hyperplane. A system of linear equations is thus the intersection of multiple hyperplanes.

2.3 System of Linear Equations

Perspectives of system of linear equations (SLE):

- Intersection of hyperplanes (row vector view): we write the equations as:

$$A_i x = b_i, i = 1, \cdots, m \tag{2.22}$$

  where A_i is the i-th row vector. Each of the m equations represents a hyperplane in $\mathbb{R}^n$, and the solution is the intersection of m hyperplanes.

- Application: suppose we have three planes in $\mathbb{R}^3$, the intersection of the first two planes is a line L. The relationship between L and the last plane may fall into three cases: (1) single intersection: unique solution; (2) parallel: no solution; (3) within the plane: infinitely many solutions.

- Linear combination of column vectors (column vector view): we write the equations as:

$$\sum_{j=1}^{n} x_i A^j = b \tag{2.23}$$

  where A^j is the j-th column vector. Thus the problem is to write b as a linear combination of n column vectors. So the two fundamental problems are equivalent: SLE, and expressing a vector as a linear combination of a given set of vectors (of the same dim).

Interpretations of system of linear equations: the following problems (in terms of n-dimensational vector X) are all equivalent:

- System of linear equation:

$$\begin{cases} a_{11}x_1 + \cdots + a_{1n}x_n = 0 \\ \quad \cdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n = 0 \end{cases} \tag{2.24}$$

- Kernel of a linear map: suppose A is the matrix of some linear map L, then $X \in \text{Ker}L$:

$$AX = 0 \tag{2.25}$$

- Orthogonal space: X is a vector in the space orthogonal to the space of row vectors of A:

$$\begin{cases} A_1 X = 0 \\ \cdots \\ A_m X = 0 \end{cases} \tag{2.26}$$

- Linear dependence: X is not equal to 0 iff the column vectors of A are linearly dependent:

$$x_1 A^1 + \cdots x_n A^n = 0 \tag{2.27}$$

Solving systems of linear equations by reducing to row echelon form:

- Algebraic idea: variable elimination by algebraic manipulation of equations (adding, subtracting, etc. equations). Specifically, the goal is to create a system where one equation has one variable, another one has two equations, etc, and then back-substitution can be applied. And these operations can be understood in terms of operations on the coefficient or the augmented matrix $[A|b]$.
- Row echelon form: to create the system where back-substitution can be applied, we want: for each leading entry (the first non-zero entry) of each row, all entries below it (the left-bottom corner) must be 0. This can be expressed as two conditions:
  - Any rows consisting entirely of zeros are at the bottom.
  - In each non-zero row, the leading entry must be in a column to the left of any leading entries below it.
- Elementary row operations and row reduction:
  - Elementary row operations: interchange two rows, multiply a row by a non-zero constant, add a multiple of a row to another row.
  - Pivoting: the basic step of row reduction is pivoting. In this step, one entry is chosen to be a leading entry (pivot), and the elementary row operations are applied so that all entries below it are 0.
- Row equivalence: Definition: two matrices are row equivalent if there is a set of elementary row operations that convert one to the other. Two matrices are row equivalent if and only if they can be converted to the same row echelon form.

Row echelon form, matrix rank and solution set:

- Gaussian Elimination: the row reduction of the augmented matrix, and application of back-substitution. Analysis of complexity: for an $n \times n$ matrix,

$$T(n) = n^2 + (n-1)^2 + \cdots + 1^2 = O\left(\frac{1}{3}n^3\right) \tag{2.28}$$

  The row echelon form from Gaussian Elimination provides information of the solution set, the linear dependence of the row vectors, and so on.
- The Rank Theorem:
  - Definition: the rank of a matrix is the number of non-zero rows in its row echelon form.
  - Theorem: let A be the coefficient matrix of SLE with n variables, if the system is consistent, then:

$$\text{rank}(A) + \#\text{free variables} = n \tag{2.29}$$

    The intuition: in the row echelon form, any non-zero row allows one to solve one leading variable, and any extra variables cannot be solved, and are free variables.

  - Remark: the rank of A can be understood as the number of independent variables (the "true" number), and the Theorem says, this number plus the the number of free variables is n (the total number of variables).

- Rank of matrix transpose: the row rank and column rank of a matrix are equal:

$$\text{rank}(A) = \text{rank}(A^T) \tag{2.30}$$

  Proof: if x is a solution of $Ax = 0$, then $y = Ax$ is a solution of $A^T y = 0$; and vice versa.

- Gauss-Jordan Elimination: the idea is that in the Gaussian elimination, back substitution is still not easy (one step per variable). Ideally, we could reduce the matrix to such a form that any row contains only one variable. This is called the reduced row echelon form: row echelon form, all leading entries are 1 and all zero's elsewhere. The reduced row echelon form of a matrix is *unique*.

- Homogeneous system: it has either trivial or infinitely many solutions. Given a homogenous system $Ax = 0$ of m equations with n variables, if $m < n$, then it has infinitely many solutions.
  Proof: in the row echelon form, the number of non-zero row ($\text{rank}(A)$) must be less than or equal to m, thus less than n. Therefore, there must be free variables by the Rank Theorem.

Linear dependence of column or row vectors and the solution set of SLE:

- Linear dependence of column vectors: the column vectors of A, $A^1, \cdots, A^n$ are linearly dependent iff $Ax = 0$ has nontrivial solutions.
  Proof: by definition, linear dependence means there exists $x_1, \cdots, x_n$ s.t. $\sum_i x_i A^i = 0$.

- Linear dependence of row vectors: the row vectors of A, $A_1, \cdots, A_m$ are linearly dependent iff $\text{rank}(A) < m$.
  Proof: the linear dependence means that in the row echelon form, there exists a zero row, thus $\text{rank}(A) < m$. On the other hand, if $\text{rank}(A) < m$, then in the row echelon form, there is a zero row, and thus one can find a set of linear operations so that the linear combination of A_i's are 0.
  Intuition: if the number of independent variables (rank) is less than the number of equations, then there must be some dependence.

- A general result of linear dependence of vectors: given any m vectors in $\mathbb{R}^n$, if $m > n$, then the vectors are linearly dependent.
  Intuition: any three vectors in $\mathbb{R}^2$ must be linearly dependent. The reason is that to solve the linear system, we have three unknowns but only two equations (one per coordinate), thus the system must have infinitely many solutions.
  Proof: consider the matrix $A = [v_1, \cdots, v_m]$, where v_i is one of the m column vectors. Since $m > n$, there are more unknowns (m) than equations (n), thus the system has non-trivial solution, thus the vectors are linearly dependent.

2.3.1 Numeric Methods for SLE

Iterative methods for solving SLE: Section 2.5 of Poole [2006]

- General idea: if we write the equations to be solved as $x = f(x)$ (where x could be a vector), then we could form the recurrence: $x_{n+1} = f(x_n)$, and solve it iteratively. The convergence can often be shown by Fixed Point theorem. To apply this idea to SLE $Ax = b$, we first solve x_1 in the first equation assuming other variables are known, $x_1 = f_1(x_2, \cdots, x_n)$, then do this for $x_2 = f_2(x_1, x_3, \cdots, x_n)$, and so on.

- Jacobi's method: suppose we have an initial solution $x^{(0)}$. At the k-step, we have $x^{(k)}$ we first use the first equation to solve x_1 (using the current values of all other unknowns), and use the second equation to solve x_2, and so on, until we solve all x_i's. Then we update $x^{(k+1)}$, and repeat.

  - Gauss-Seidel method: as soon as we obtain a new value of x_i at each step, we use it in the iteration to solve other unknowns.

- Geometric interpretation: in a 2D case, the goal is to find the intersection of two lines. We first intersect the first line with $y = y^{(k)}$, find the value of x, say $x^{(k)}$, then we intersect the second line with $x = x^{(k)}$, and find the value of $y = y^{(k+1)}$. Repeat this process, and if it converges, it converges to the solution.

- Convergence: we first define that, for a $n \times n$ matrix A, it is strictly diagonally dominant if: for each i,

$$|a_{ii}| > |a_{i1}| + \cdots + |a_{in}| \tag{2.31}$$

  Theorem: if the coefficient matrix A of a SLE is strictly diagonally dominant, then it has a unique solution and both Jacobi's and Gauss-Seidel method converge to the solution.

  - One can prove that: if the Jacobi or Gauss-Seidel method converge for a system of n equations with n variables, then it must converge to the solution.
  - However, there are cases where the method does not converge (diverge instead).

- Relation to conditional maximization (CM): suppose we are optimizing a function $f(x)$, and the derivative of f is: $f'(x) = Ax - b$, then our goal is to solve $f'(x) = 0$. The CM procedure for maximization is: maximize x_1 assuming x_2 is known, and vice versa. This is equivalent to solving the two equations iteratively:

$$\begin{cases} \partial f/\partial x_1 &= A_1 x - b_1 = 0 \\ \partial f/\partial x_2 &= A_2 x - b_2 = 0 \end{cases} \tag{2.32}$$

  which is exactly the iterative method here.

- **Lesson**: iterative methods for solving any kind of equations.

- Remark:

  - Advantages of iterative method over Gaussian elimination: (1) Efficiency: for sparse matrices. (2) Robustness to numerical errors: since the results are reached gradually.
  - One may also design the iterative methods from a different perspective, e.g. the solution of $Ax = b$ is the stationary point of a linear dynamic system, and then we can solve it by simulating the time evolution of the system.

2.4 Matrices

Problems of matrix algebra:

- Algebraic properties of matrices: similar to those of real numbers. such as: associativity, commutativity. More complex problems: polynomials, fractional.

- Properties of special matrices: diagnonal, upper triangular, etc.

- What does matrix algebra say about SLE: what matrix has a unique solution?

Product of matrix and a vector: let $x = (x_1 \cdots x_n)^T$ be a n-dimensional column vector.

- Multiplication of row vector and a matrix is a row vector: $x^T A$ is a row vector of the same dimension of A.

- Multiplication of matrix and column vector is a column vector: Ax is a column vector of the same dimension of A.

- Row-vector representation of Ax (SLE representation):

$$Ax = \begin{pmatrix} A_1 \\ A_2 \\ \cdots \\ A_m \end{pmatrix} x = \begin{pmatrix} A_1 x \\ A_2 x \\ \cdots \\ A_m x \end{pmatrix} \tag{2.33}$$

  Or the i-th component $(Ax)_i = A_i x$. This is the representation used in SLE: $Ax = b$.

- Column-row representation of Ax (inner product representation): linear combination of column vector of A.

$$Ax = (A^1 \cdots A^n) \begin{pmatrix} x_1 \\ \cdots \\ x_n \end{pmatrix} = \sum_{i=1}^{n} x_i A^i \tag{2.34}$$

  where A^i is the i-th column vector of A, i.e. Ax is a linear combination of the column vectors of A.

Matrix multiplication: assume A is a $m \times n$ matrix and B is a $n \times r$ matrix. Let A_i be the i-th row matrix and a_j be the j-th column matrix (similar for B). The product AB can be written in four forms:

- Row-column representation of matrix product: this is the definition of matrix product:

$$(AB)_{ij} = A_i b_j \tag{2.35}$$

- Column-row representation of matrix product: the product of $a_i B_i$ is a $m \times r$ matrix, and this is called the outer product. The product is also called the outer product expansion.

$$AB = (a_1 \cdots a_n) \begin{pmatrix} B_1 \\ \cdots \\ B_n \end{pmatrix} = \sum_{i=1}^{n} a_i B_i \tag{2.36}$$

  This is similar to dot product between vectors: suppose x and y are n-dim column vectors, then

$$x^T y = (x_1 \cdots x_n) \begin{pmatrix} y_1 \\ \cdots \\ y_n \end{pmatrix} = \sum_{i} x_i y_i \tag{2.37}$$

- Row-matrix representation of matrix product:

$$AB = \begin{pmatrix} A_1 \\ A_2 \\ \cdots \\ A_m \end{pmatrix} B = \begin{pmatrix} A_1 B \\ A_2 B \\ \cdots \\ A_m B \end{pmatrix} \tag{2.38}$$

- Matrix-column representation of matrix product:

$$AB = A(b_1 \cdots b_r) = (Ab_1 \cdots Ab_r) \tag{2.39}$$

- Block multiplication: an even more general form of matrix multiplication is the block multiplication. Essentially, if dimensions of blocks of A and B match, then the blocks can be multiplied. Suppose:

$$A = \begin{pmatrix} A_{11} \cdots A_{1t} \\ \cdots \\ A_{s1} \cdots A_{st} \end{pmatrix} \tag{2.40}$$

$$B = \begin{pmatrix} B_{11} \cdots A_{1r} \\ \cdots \\ B_{t1} \cdots A_{tr} \end{pmatrix} \tag{2.41}$$

Then if we define $C_{ij} = \sum_{k=1}^{t} A_{ik} B_{kj}$, we have:

$$AB = \begin{pmatrix} C_{11} \cdots C_{1r} \\ \cdots \\ C_{s1} \cdots C_{sr} \end{pmatrix} \tag{2.42}$$

- Remark: in specific problems, choose the best form of matrix product could simplify the analysis.

Properties of matrix transpose:

- Transpose of matrix product: For matrices A and B,

$$(AB)^T = B^T A^T \tag{2.43}$$

  Proof: $(AB)_{ij} = A_i b_j = b_j A_i$, and b_j is the j-th row of B^T, and A_i is the i-th column of A^T. Thus the RHS is the ji-th term of $B^T A^T$.

- If A is a square matrix, then $A + A^T$ is symmetric.

- Product of matrix and its transpose: suppose A is a $m \times n$ matrix, let A_j be the j-th column of A, then

$$A^T A = \begin{pmatrix} A_1^T \\ \cdots \\ A_n^T \end{pmatrix} (A_1 \cdots A_n) \tag{2.44}$$

  And $(A^T A)_{jk} = A_j^T A_k$ is the dot product of the two vectors A_j and A_k. The same can be done for AA^T: we simply use row vectors of A above. AA^T **and** $A^T A$ **are symmetric matrices**. Applications:

  - If the columns of A are orthogonal to each other, then $A^T A$ is a diagonal matrix.
  - Statistical interpretation: suppose X is the $n \times D$ data matrix (mean 0), then $X^T X$ is the sample covariance matrix of D variables.
  - **Geometric interpretation**: suppose we are interested in the effect of A on the norm of a vector, this is given by $A^T A$: $\|Ax\| = \langle Ax, Ax \rangle = x^T A^T A x$.
  - If we need to deal with matrix $A^T U A$, where U is a real symmetric matrix, then we can use eigen-decomposition of U to write it in the form of $B^T B$, where B is some matrix.

Properties of matrix product:

- Extraction of any row or column of a matrix: Let A be $m \times n$ matrix, e_i be $1 \times m$ unit vector, and e_j be $n \times 1$ unit vector, then:

$$e_i A = A_i \tag{2.45}$$

  i.e. the product is the i-th row of A, and:

$$A e_j = a_j \tag{2.46}$$

  i.e. the product is the j-th column of A.
  Proof: only consider the second part. We use the outer-product expansion:

$$A e_j = (a_1 \cdots a_n)(0 \cdots 1 \cdots 0)^T = a_j \cdot 1 = a_j \tag{2.47}$$

  Remark: how do we generalize these results? Ex. how do we extract the diagnoal vector of a matrix? Or choose one element per row/column (permutation of indices)?

- Product of a matrix and a diagonal matrix: each column (or row) of the matrix is multiplied by the diagonal element. If $D = \text{Diag}(d_1, \cdots, d_n)$, then $AD = [d_1 a_1 \qquad d_n a_n]$, where a_j is the j-th column of A. Similarly, the i-th row of DA would be $d_i A_i$, where A_i is the i-th row of A.

- Upper triangular matrix: if both A and B are upper triangular matrices, then AB is also an upper triangular matrix.
  Proof 1: we consider the ij entry of (AB), where $i > j$:

$$(AB)_{ij} = A_i b_j = \sum_k a_{ik} b_{kj} \tag{2.48}$$

  When $k < i$, $a_{ik} = 0$, when $k \geq i$, we have $k > j$, thus $b_{kj} = 0$.
  Proof 2: consider linear map $f(x) = ABx$. Let $u = Bx$, then the effect of u is that: u_1 depends on x_1 to x_n, u_2 depends only on x_2 to x_n, and so on. Next, the effect of $y = Au$ is s.t. y_1 depends on u_1 to u_n (thus x_1 to x_n), y_2 depends on u_2 to u_n (thus x_2 to x_n), and so on.

- Trace: if A and B are square matrices, then

$$\text{tr}\,(AB) = \text{tr}\,(BA) \tag{2.49}$$

  Proof: switching of the order of indices in summation:

$$\sum_i \sum_k a_{ik} b_{ki} = \sum_k \sum_i b_{ki} a_{ik}. \tag{2.50}$$

- Rank: the rank of the product is less than or equal to the rank of each matrix. Suppose A is $n \times p$ matrix and B an $p \times r$ matrix, we have

$$\text{rank}(AB) \leq \min\{\text{rank}A, \text{rank}B\} \tag{2.51}$$

  Proof: first, if $Bx = 0$, then $ABx = 0$, thus $\text{Ker}B \subseteq \text{Ker}(AB)$. Because rank and kernel of B sum to r, and the same for the matrix AB, we have $\text{rank}B \geq \text{rank}(AB)$.
  Next, $\text{Im}(AB) = \{A(Bx) : x \in R^r\} \subseteq \{Au, u \in R^p\} = \text{Im}(A)$, so we have $\text{rank}(AB) \leq \text{rank}(A)$.
  Remark: the geometric intuition is simple: AB is the product (composition) of two linear maps A and B. Both A and B may reduce the dimension in the image, and total reduction of dimension from AB must be bigger than reduction of A or B alone.

Motivation: why inverse is important?

- The basic idea is the relationship between the SLE $Ax = b$, and the matrix inverse A^{-1}. Thus if we can derive the properties of A^{-1} from an algebraic perspective, then we immediately know the properties of the SLE.

- Example: the consistency of the SLE is equivalent to the invertibility of A. The dimension of the solution set of the SLE is related to the rank of A. The solution is given by $x = A^{-1}b$.

- Geometric intuition: inverse function (map). If $y = Ax$ is the linear map corresponding to A, then $x = A^{-1}y$ is the inverse map.

- Ideas for finding inverse of a matrix: if A can be expressed in terms of simpler matrices (factorization, expansions, etc.), can we express A^{-1} in terms of inverses of these simpler matrices.

Basic properties of matrix inverse:

- 2×2 matrix: let A be a square matrix:

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \tag{2.52}$$

Its inverse is:

$$A^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \tag{2.53}$$

- Some useful properties:

$$(AB)^{-1} = B^{-1}A^{-1} \tag{2.54}$$

$$(A^T)^{-1} = (A^{-1})^T \tag{2.55}$$

- Inverse of block matrix: the inverse of an upper triangular matrix is (Exercise 64 of 3.3. in Poole [2006])

$$\begin{bmatrix} A & B \\ O & D \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} & -A^{-1}BD^{-1} \\ O & D^{-1} \end{bmatrix} \tag{2.56}$$

Matrix inverse and SLE:

- Elementary matrix: elementary row operation applied to the identity matrix. The elementary row operation can be equivalently represented by an elementary matrix:
  Theorem: let E be an elementary matrix from a certain elementary row operation. The result of the same operation applied on an $n \times r$ matrix A is EA.

- Product and inverse of elementary matrices: the successive applications of elementary row operations (as in Gaussian elimination) can be represented by product of elementary matrices. Any elementary matrix is invertible and its inverse is an elementary matrix of the same type.

- **The Fundamental Theorem of Invertible Matrices** (Version 1): let A be an $n \times n$ matrix, the following statements are equivalent:

  1. A is invertible.
  2. $Ax = b$ has a unique solution for any $b \in \mathbb{R}^n$.
  3. $Ax = 0$ has only the trivial solution.
  4. The reduced row echelon form of A is I_n.
  5. A is a product of elementary matrices.

  Proof: (a) $\Rightarrow$ (b): we multiply A^{-1} to $Ax = b$, and obtain a unique solution $x = A^{-1}b$.
  (b) $\Rightarrow$ (c), (c) $\Rightarrow$ (d): from Gassuian-Jordan elimination (application of elementary row operations).
  (d) $\Rightarrow$ (e): write the process of GJ-elimination as $(E_k \cdots E_1)A = I_n$, thus $A = (E_1)^{-1} \cdots (E_k)^{-1}$.
  (e) $\Rightarrow$ (a): obvious.

- Remark: this theorem reveals the relationship between an inverible matrix and SLE, thus the solution of SLE can be reduced to matrix analysis; and similarly, the matrix inverse can be analyzed from the SLE perspective (imagine an unknown vector x to solve). Furthermore, the GJ-elimination provides a representation of A: a product of elementary matrices. Representation like this could be a powerful tool: e.g. representation of a polynomial as a product of first-order terms $(x - a)$.

- Theorem: let A be a square matrix, if B is a square matrix s.t. $AB = I$ or $BA = I$, then A is invertible and $B = A^{-1}$.
  Proof: suppose we have $BA = I$. Consider the equation $Ax = 0$, we multiply B, and have $BAx = 0$, or $Ix = 0$, thus $x = 0$. So $Ax = 0$ has only trivial solution, thus A is invertible.

- Solving matrix inverse by GJ Elimination: if a set of elementary row operations convert A to I_n (from GJ Elimination), the same operations will convert I to A^{-1}. Intuition: we are trying to solve the inverse map: given $y = Ax$, we want to obtain x in terms of y. But to solve this, we use GJE, and we have $x = \text{GJ}(A) \cdot y$, where $\text{GJ}(A)$ is GJE applied to A. This is exactly A^{-1}.

LU factorization:

- Gaussian elimination in the form of matrix factorization: suppose we do Gaussian elimination to obtain an upper triangular matrix U, this is equivalent to multiplying elementary row matrices on A, so we may have:
$$E_3 E_2 E_1 A = U \tag{2.57}$$
where each E_i involves no row substitutions. Then we can see that:
$$A = E_1^{-1} E_2^{-1} E_3^{-1} U = LU \tag{2.58}$$
where L is a unit lower triangular matrix
$$L = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ * & 1 & \cdots & 0 \\ \cdot & \cdot & \cdots & \\ * & * & \cdots & 1 \end{pmatrix} \tag{2.59}$$
Not all matrices have LU factorization. Ex.
$$A = \begin{bmatrix} 0 & 1 \\ 2 & 3 \end{bmatrix} \tag{2.60}$$
There is no way to make the matrix upper triangular (the term 2 cannot become 0), unless we allow row interchange.

- Theorem (LU factorization): if A is a square matrix that can be reduced to row echelon form without using any row interchanges, then A has an LU factorization where L is unit lower triangular and U upper triangular. If A has an LU factorization, then it is unique.
Proof: the uniqueness part follows from both L and U are invertible. Two ways of proving this: (1) use the fact that L can be written as a product of elementry row matrices; (2) consider the SLE: $Lx = 0$, clearly, it has a unique solution.

- Solving SLE using LU factorization: write $Ax = b$ as $LUx = b$, define $y = Ux$, then we solve the SLE in two steps: $Ly = b$ by forward substitution and $Ux = y$ by backward substitution.
  - Remark: this is an example illustrating the power of "factorization". We write a matrix as a product of two simpler matrices, and the corresponding SLE can be solved by solving two simpler SLE.

- Finding LU factorization: to obtain LU facotorization, we apply the row operations of the form $R_i - kR_j$. It can be shown that $L_{ij} = k$.

- $P^T LU$ factorization: sometimes a matrix A cannot be reduced to row echelon without using row interchanges, e.g. multiple leading entries of a top row are all 0. In this case, we first apply row interchange (through permutation matrix), then do LU: $PA = LU \Rightarrow A = P^T LU$, where P^T is the inverse of P (a property of permutation matrix).

2.4.1 Special Matrices

Diagonal matrix:

- Linear map: Suppose D is a n-by-n diagnoal matrix, with i-th diagonal element d_i. Then D is a linear map that scales each component by d_i. In other words, given $x \in \mathbb{R}^n$, we have:

$$Dx = \text{diag}(d_1, \cdots, d_n) \cdot \begin{pmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{pmatrix} = \begin{pmatrix} d_1 x_1 \\ d_2 x_2 \\ \cdots \\ d_n x_n \end{pmatrix} \tag{2.61}$$

- Pre- and post-multiplication of diagonal matrix: let $D = \text{diag}(d_i)$ be $n \times n$ diagonal matrix, and A be $n \times m$ matrix, then DA is the matrix, where each row of A is scaled by d_i. If A is $m \times n$ matrix, then AD is the matrix, where each column of A is scaled by d_i.
  Ref: `https://solitaryroad.com/c108.html`

- Matrix product ADB: where D is diagonal with elements d_i, and A is $m \times n$ matrix and B is $n \times p$ matrix. Let a_i be the column vectors of A and b_i^T row vector of B, we have:

$$ADB = [d_1 a_1 \cdots d_n a_n] \begin{bmatrix} b_1^T \\ \cdots \\ b_n^T \end{bmatrix} = \sum_i d_i a_i b_i^T \tag{2.62}$$

  where $a_i b_i^T$ is a $m \times p$ matrix. An important special case is Spectrum Decomposition: for symmetric matrix, we have $A = QDQ^T = \sum_i \lambda_i q_i q_i^T$.

- Matrix scaling: suppose we have $n \times n$ symmetric matrix R. We want to scale each element r_{ij} be s_i and s_j. This can be accomplished by: let $S = \text{diag}(s_i)$

$$SRS = (s_i s_j r_{ij})_{ij} \tag{2.63}$$

  To see this: SR would scale each row of R by s_i. Then multiply S will scale each column of SR by s_j.

Elementary matrices [Wiki]: any row or column switch operation (interchange two rows or columns) corresponds to an elementary matrix, which is formed by applying the same operation on the identity matrix. Given a matrix A, the matrix $\tilde{A}$ after applying a switch operation (between rows i and j) is:

$$\tilde{A} = AE_{ij} \tag{2.64}$$

where $E_{i,j}$ is the elementary matrix corresponding to the operation. The properties of the matrix:

$$\det(E_{ij}) = -1 \tag{2.65}$$

2.5 Vector Space

Geometric intuitions of matrix, linear map, dimension and rank:

- A matrix can be viewed as a linear map, and the rank of a matrix represents the effect of the linear map on dimensions. A full-ranked matrix A, preserves the dimension of vector x.

- Why a matrix/linear map may reduce the dimension? Ex. projection of an plane usually leads to a plane, but may lead to a line. Algebraically, suppose x is 2D, the equation $Ax = 0$, where elements of A are variables, have many solutions (projection of line into a dot).

Subspace and dimension

- Definition of row and column space of a matrix: consider an $m \times n$ matrix A, define row space of A as $\text{Span}(A_1, \cdots, A_m)$ and column space as $R(A) = \text{Span}(A^1, \cdots, A^n)$ (also called the range of A).

- Row space: elementary row operations do not change the row space (after each operation, an row is a linear combination of other rows). So to determine the base of the row space of A, we obtain the reduced echelon form of A.

- Column space: the equation $Ax = 0$ describes the linear dependency of column vectors: $x_1 a_1 + \cdots x_n a_n = 0$ where a_i is the i-th column vector. So we obtain the reduced echelon form of A, it preserves the linear dependency.

- The Basis Theorem: let S be a subspace, then any two bases of S have the same number of vectors (called dimension).
  Proof idea: suppose we have two bases $\{u_1, \cdots, u_r\}$ and $\{v_1, \cdots, v_s\}$, we show that $r = s$. Proof by contradiction, if $r > s$, then each of u_i can be expressed as a linear combination of v_j's, but the number of independent v_j's is smaller, so u_i's must be linearly dependent.

Rank of matrices:

- Theorem: dimension of the row and column space of A are equal. This leads to the defintion of rank of A.
  Proof: use the reduced row echelon form of A.
  Intuition: consider a special case of "proportional" matrix. If rows are proportional, then columns should also be proportional. Ex. we have a matrix:

$$\begin{bmatrix} a & b & c \\ 2a & 2b & 2c \end{bmatrix} \tag{2.66}$$

  The row vectors are proportional with basis $[a, b, c]$, and the column vectors are also proportional with the basis $[1, 2]^T$.

- Rank of matrix transpose: $\text{rank}(A) = \text{rank}(A^T)$.
  Proof: follow from the equal dimension of row and column space.

- Null space and nullity: define null space as $\text{null}(A) = \{X \in \mathbf{R}^n : AX = 0\}$, and its dimension $\text{nullity}(A)$.

- **The Rank Theorem**: if A is an $m \times n$ matrix, then

$$\text{rank}(A) + \text{nullity}(A) = n \tag{2.67}$$

  Proof: consider the reduce row echelon form of A, the number of independent variables is rank of A, and the remaining number of free parameters is nullity of A.
  Remark: the Rank Theorem establishes the basic one-to-one correspondable between matrix rank and the solution of SLE (null space). This is one of the most important theorems in linear algebra.

- **The Fundamental Theorem of Invertible Matrices** (Version 2): let A be an $n \times n$ matrix, the following statements are equivalent:

  1. A is invertible.
  2. $Ax = b$ has a unique solution for any $b \in \mathbb{R}^n$.
  3. $Ax = 0$ has only the trivial solution.
  4. The reduced row echelon form of A is I_n.
  5. A is a product of elementary matrices.
  6. $\text{rank}(A) = n$.

7. nullity$(A) = 0$.
8. The column vectors of A are linearly independent.
9. The column vectors of A span $\mathbf{R}^n$.
10. The column vectors of A form a basis of $\mathbf{R}^n$.
11. The row vectors of A are linearly independent.
12. The row vectors of A span $\mathbf{R}^n$.
13. The row vectors of A form a basis of $\mathbf{R}^n$.

- Rank of A^TA: let A be an $m \times n$ matrix, then $\text{rank}(A^TA) = \text{rank}(A)$, and the matrix A^TA is invertible if and only if $\text{rank}(A) = n$.
  Proof: by the Rank Theorem, we will only need to show that $\text{nullity}(A^TA) = \text{nullity}(A)$. To see this, let $x \in \text{null}(A)$, i.e. $Ax = 0$, then $A^T(Ax) = 0$, so $x \in \text{null}(A^TA)$. Conversely, if $A^TAx = 0$, then $x^TA^TAx = 0$, thus $(Ax)^T(Ax) = 0$, so $Ax = 0$.
  Remark:
  - Covariance matrix perspective: A^TA can be viewed as covariance matrix, and we can say that the covariance matrix preserves the linear dependency of the original matrix. Ex. if $A_1 = A_2$, then $\text{Cov}(A_1, A_j) = \text{Cov}(A_2, A_j), \forall j$.
  - Inner product: $\langle Ax, Bx\rangle = x^TA^TBx$, thus A^TB **can be viewed in terms of inner product and quadratic form**.

- Rank of matrix product:
  - Theorem: $\text{rank}(AB) \leq \text{rank}(B)$ and $\text{rank}(AB) \leq \text{rank}(A)$.
    Proof: use the fact that if $Bx = 0$, then $ABx = 0$. And take the tranpose.
  - Theorem: if U is invertible then $\text{rank}(UA) = \text{rank}(A)$.
    Proof: on the one hand, $\text{rank}(UA) \leq \text{rank}(A)$, on the other hand, $\text{rank}(A) = \text{rank}(U^{-1}UA) \leq \text{rank}(UA)$.

  Remark: the general intuition is that when you apply successive linear transformations, you tend to "shrink" the solution space. It preserves the solutions iff the transformation is invertible.

Linear transformations:

- **Theorem** (matrix representation of linear transformations): let $T : \mathbf{R}^n \to \mathbf{R}^m$ be a linear map, then T can be represented by $m \times n$ matrix $A = [T(e_1), \cdots, T(e_n)]$, where $e_1, \cdots, e_n$ are the standard basis of $\mathbf{R}^n$.
  Proof: consider a vector x, $T(x) = x_1T(e_1) + \cdots x_nT(e_n)$, and this can be written as Ax, where A is given above.

- The Rank Theorem for linear transformations: let $L : V \to W$ be a linear map, let $\text{Ker}L = \{v \in V : L(v) = 0\}$ and $\text{Im}L = \{L(v) : v \in V\}$, then:

$$\dim V = \dim \text{Ker}L + \dim \text{Im}L \tag{2.68}$$

  Remark: the intuition is the total dimension of V should be preserved. If $\dim \text{Im}L < \dim V$, then the lost dimension must be due to: multiple points of V map to a single point in W, and this is simply the dimension of $\text{Ker}L$.

Linear map: preserve operations of vector addition and scale multiplication. Some examples of linear map:

- Integral: a linear map from the space of all real-valued integrable functions on some interval to $\mathbf{R}$. And similar expectation of random variables.

- Geometric operations in $\mathbf{R}^2$: reflection, ratotation, projection, stretching, shear, etc.

Representation of linear maps:

- Change of basis: suppose we know the coordinates of a vector v with one basis B, we want to determine its coordinates under another basis B'. Let $M_B(v)$ and $M_{B'}(v)$ be the coordinates of v in B and B' respectively. Let $M_{B'}^B$ be the matrix of B under B' (i.e. each column of $M_{B'}^B$ corresponds to the coordinates of one basis vector of B under B'). Then we have:

$$M_{B'}(v) = M_{B'}^B M_B(v) \tag{2.69}$$

  Proof: write $v = \sum_i x_i v_i = (v_1 \cdots v_n)(x_1 \cdots x_n)^T$ where v_i are basis and x_i are coordinates, use B' as the basis, then the first term is the matrix $M_{B'}^B$ and the second $M_B(v)$.

- Linear map and matrix: given vector spaces V and W where $\dim V = n$ and $\dim W = m$, and the basis of V and W given, then there is a one-to-one mapping between a linear map from V to W, and a $m \times n$ matrix. Consider a linear map $F : V \to W$, let B and B' be the basis of V and W respectively, and $M_{B'}^B(F)$ be the matrix associated with F (i.e. if v_i is a basis of V, then a column of this matrix is the coordinates of $F(v_i)$ under B'), we have:

$$M_{B'}(F(v)) = M_{B'}^B(F) M_B(v) \tag{2.70}$$

  Proof: suppose $B' = (w_1 \cdots w_m)$, then we have $F(v) = F(\sum_i x_i v_i) = \sum_i x_i F(v_i)$, thus $F(v)$ has coordinates $(x_1 \cdots x_n)$ under $F(v_i)$. The problem is to convert these coordinates into the basis B'.

- Change of basis for linear map: consider a linear map $F : V \to V$, and let B and B' be two different basis of V. Suppose $N = M_B^{B'}(\mathrm{id})$, then the two matrices of F under two basis are similar:

$$M_{B'}^{B'}(F) = N^{-1} M_B^B(F) N \tag{2.71}$$

  Remark: conversely, whenever we have similar matrices, e.g. $B = U^{-1}AU$, we know that A and B represent the same underlying linear map, except the basis are different.

2.6 Determinants and Trace

Determinant:

- Motivation: geometrically we have the concept of area and volume. The correspoding concept in algebra is determinant. Another way: need a way to measure the size of a linear map (how the map changes the magnitude of the vectors).

- Definition/Leibniz formula: the determinant of a $n \times n$ matrix A is:

$$\det A = \sum_{\sigma \in S_n} \mathrm{sgn}(\sigma) \prod_{i=1}^{n} A_{i,\sigma_i} \tag{2.72}$$

  where σ is a permutation of the set $\{1, 2, \cdots, n\}$.

- Laplace expansion: given a $n \times n$ matrix A, denote M_{ij} the (i, j) minor (the determinant of the remaining square matrix after removing the i-th row and the j-th column). The (i, j) cofactor is: $C_{ij} = (-1)^{i+j} M_{ij}$. We have the following expansion on the i-th row or the j-th column:

$$\det A = a_{i1}C_{i1} + \cdots a_{i2}C_{i2} + \cdots a_{in}C_{in} = a_{1j}C_{1j} + \cdots a_{2j}C_{2j} + \cdots a_{nj}C_{nj} \tag{2.73}$$

  Proof idea: plug-in the determinant definition of the co-factors in the above equation, it is easy to see that the RHS is a sum of product terms (over permutations). Then one only need to verify the signs are correct.

- Co-factor and adjugate matrix: the co-factor matrix of A is (C_{ij}), where C_{ij} is the (i,j) co-factor of A. The adjugate matrix is the transpose of the co-factor matrix: $\text{adj}(A) = C^T$. From the Laplace expansion, we have:

$$A\text{adj}(A) = \text{adj}(A)A = \det(A)I_n \tag{2.74}$$

  This implies that the inverse of A (if it exists) can be expressed in terms of the adjugate matrix:

$$A^{-1} = \frac{1}{\det(A)}\text{adj}(A) \tag{2.75}$$

Determinant and volume: [The Determinant: a Means to Calculate Volume]

- Volume of parallelepiped: the volume is defined recursively, as the product of the volume of the "base" and the "altitude". Given a parallelepiped defined by k vectors $\alpha_1, \cdots, \alpha_k$ in $\mathbb{R}^n$, suppose we decompose α_1 by: $\alpha_1 = B + C$, where B is perpendicular to the subspace spanned by $\alpha_2, \cdots, \alpha_k$ and C is a vector in the subspace, then the volume of the parallelepiped is the product of the volume of the parallelepiped defined by $\alpha_2, \cdots, \alpha_k$ and the length of B.

- Theorem: given an m-dimensional parallelepiped P in n-dimensional space, let $V(P)$ be the volume of P, and A be the matrix formed by the m vectors (row vector), then:

$$V(P)^2 = \det(AA^T) \tag{2.76}$$

  Proof idea: if α_1 is perpendicular to the subspace, then the result is easy to show; for the general case, we write $\alpha_1 = B + C$, and show that this would not change the volume or determinant.

- Proof by induction: let $\tilde{A}$ be the matrix defined by $B, \alpha_2, \cdots, \alpha_m$, it is easy to see that A is related to $\tilde{A}$ by a series of row operations (since C is a linear combination of $\alpha_2, \cdots, \alpha_m$), so we have $\tilde{A}$ is a product of A and elementary matrices, and:

$$\det(AA^T) = \det(E_1 \cdots E_{m-1}\tilde{A}\tilde{A}^T E_{m-1}^T \cdots E_1^T) = \det(\tilde{A}\tilde{A}^T) \tag{2.77}$$

  Now we show that for the matrix $\tilde{A}$ (the special case), the relation of determinant and volume holds. Let D be matrix defined by $\alpha_2, \cdots, \alpha_m$, we have:

$$\tilde{A}\tilde{A}^T = \begin{pmatrix} B \\ D \end{pmatrix}(B^T D^T) = \begin{pmatrix} BB^T & BD^T \\ DB^T & DD^T \end{pmatrix} \tag{2.78}$$

  The nondiagonal terms are 0 as B is diagonal to D. So the determinant is $BB^T \det(DD^T)$, and we apply the induction hypothesis.

- Square matrix: for n-dimensional square matrix, A, and the corresponding parallelepiped P, we have $V(P) = \det(A)$.

- A special case: suppose we have vectors $u, v, w \in \mathbf{R}^3$, the volumne of the parallelepiped formed by the vectors is:

$$V = u \cdot (v \times w) = \det(u, v, w) \tag{2.79}$$

  where (u, v, w) is the matrix formed by the column vectors.
  Proof: the first part follows from the definition of cross product. The second part follows from the coordinate representation of cross product.

- Geometirc intepretation: scale factor for measure when the matrix is regarded as a linear transformation. Thus a matrix with determinant 3 when applied to a set of points with finite area will transform those points into a set with three times the volume.

Properties and computation of determinant:

- Theorem: Let A be a square matrix:
  - Adding a scalar multiple of one column (or row) to another column (or row) does not change the value of the determinant.
  - Interchanging two columns of a matrix multiplies its determinant by -1.
  - Transpose: $\det A^T = \det A$. Intuition: definition of determinant is symmetric wrt. rows or columns.
- Triangular matrix: if A is a triangular matrix, then $\det A$ is equal to the product of diagonal elements of A.
  Proof: follows from repeated application of Laplace expansion.
  Remark: this leads to a main strategy of computing determinant: reduce the matrix to row-echelon form, then obtain the determinant from the resulting triangular matrix. More generally, we use elementary matrix factorization of a matrix to study its determinant.
- Elementary matrices: let E be an $n \times n$ elementary matrix,
  - If E results from interchanging two rows of I_n, then $\det E = -1$.
  - If E results from multiplying one row of I_n by k, then $\det E = k$.
  - If E results from adding a multiple of one row of I_n to another row, then $\det E = 1$.

  We also have this: if E is an elementary matrix, then $\det(EB) = (\det E)(\det B)$.
- **Theorem**: A is an invertible matrix iff $\det A \neq 0$, and $\det A^{-1} = 1/\det A$.
  Proof: consider the elementary matrix factorization of A, and use the properties of determinants of elementary matrices.
- Matrix product: $\det AB = \det A \det B$.
  Proof: if A or B is not invertible, the proof is rather trivial. If they are invertible, then they can be expressed as products of elementary matrices.
  Intuition: the effect on the volume of an area by AB is the product of the effect by A and by B.
- Similar matrices: if A and B are similar, then: $\det A = \det B$.

Cramer's rule:

- Consider a linear equation $AX = b$, where A is $n \times n$ square matrix, then it has a unique solution iff $\det A \neq 0$. Furthermore, when $\det A \neq 0$:

$$x_j = \frac{\det(A^1, A^2, \cdots, A^{j-1}, b, A^{j+1}, \cdots, A^n)}{\det A} \tag{2.80}$$

- Proof: the idea is that the determiants of A and of A where column j is replaced are related using the properites of determinant. First note that $\sum_i x_i A^i = b$. We multiply the i-th column of A by x_i (except the j-th column), and add them to the j-th column:

$$x_i \det A = \det(A^1, \cdots, A^{j-1}, \sum_i x_i A^i, A^{j+1}, \cdots, A^n) = \det(A^1, \cdots A^{j-1}, b, A^{j+1}, \cdots, A^n) \tag{2.81}$$

Properties of matrix trace: trace is the sum of main diagonal of a square matrix.

- Trace is a linear map:

$$\operatorname{tr}(A + B) = \operatorname{tr}(A) + \operatorname{tr}(B) \tag{2.82}$$

$$\operatorname{tr}(cA) = c \cdot \operatorname{tr}(A) \tag{2.83}$$

- Transpose: $\mathrm{tr}A^T = \mathrm{tr}A$.
  Proof: follow the definition.

- Product: $\mathrm{tr}(AB) = \mathrm{tr}(BA)$. The proof simply follows the definition:

$$\mathrm{tr}\,(AB) = \sum_{i,j} A_{ij}B_{ji} = \mathrm{tr}\,(BA) \tag{2.84}$$

- Invariance under cyclic permutations: e.g. for matrices A, B, C:

$$\mathrm{tr}\,(ABC) = \mathrm{tr}\,(BCA) = \mathrm{tr}\,(CAB) \tag{2.85}$$

  Proof: $\mathrm{tr}\,(A(BC)) = \mathrm{tr}\,((BC)A) = \mathrm{tr}\,(B(CA)) = \mathrm{tr}\,(CA)B)$.
  Note: this theorem can be very useful to study quadratic forms:

$$x^TAx = \mathrm{tr}\,(x^TAx) = \mathrm{tr}\,(x^TAx) = \mathrm{tr}\,(Axx^T) \tag{2.86}$$

  When x is a random vector, then xx^T is the covariance of x.

- Similar matrices: if $B = P^{-1}AP$, then $\mathrm{tr}\,(A) = \mathrm{tr}\,(B)$.
  Proof: $\mathrm{tr}\,(B) = \mathrm{tr}\,(P^{-1}AP) = \mathrm{tr}\,(APP^{-1}) = \mathrm{tr}\,(A)$.

- Trace commutes with expectation and derivative: because trace is a linear operator. So suppose X a matrix of random variables, we have:

$$\mathrm{tr}\,(\mathrm{E}\,(X)) = \mathrm{E}\,(\mathrm{tr}\,(X)) \tag{2.87}$$

  The proof just follows the definition: intuitively, both LHS and RHS are sum of expectation of diagonal elements of X. Also for derivatives of a random vector x:

$$d\mathrm{tr}\,(x) = \mathrm{tr}\,(dx) \tag{2.88}$$

2.7 Eigenvalues and Eigenvectors

Motivations of eigenvalues and eigenvectors:

- Geometric motivation: for a given linear map, it may be possible to find certain directions, along which the effect of the linear map is simply strecthing/shrinking. Ex. (1) Horizontal shear: the direction is $u = (1, 0)$. (2) Scaling: the directions are simply the axis. (3) the example in Wiki, "Eigenvalues and eigenvectors".

- Idea: if we use these directions (eigenvectors) as the basis, then the linear map would have a simple representation. In particular, if the directions from an orthogonal basis, then the map of any basis vector is another basis vector (of the same direction), thus the linear map is represented by a diagonal matrix.

- Motivation from dynamic systems: the steady state of a system is often represented by equations such as $Ax = \lambda x$ ($\lambda = 1$ is a special case). Thus the eigenvectors are interpreted as steady states of a system specified by A.

- Understanding the effect on the norm of vector by linear map: suppose we want to know in general, how the vector norms are changed by a linear map. We can use eigenvectors. Ex. suppose $x = c_1u_1 + c_2u_2$ where u_1, u_2 are orthonormal eigenvectors, then $Ax = \lambda_1c_1u_1 + \lambda_2c_2u_2$. The norm of this vector is $\lambda_1^2c_1^2 + \lambda_2^2c_2^2$, which is largely determined by the largest eigenvalues.

Reserach program of eigenvalues and eigenvectors:

- Problem: suppose we are solving a system $A^n x = x$ e.g. the steady state of a dynamic system. How do we solve it without numerically calculating A^n? The general problem is to understand the effect of repeated application of a matrix/linear map.
- Geometric perspective: find special vectors/directions where Ax has a simple scaling effect.
- Reduction: suppose we find such eigenvectors, to understand the effect on any vector x, we will expand x as a linear combination of the eigenvectors, and show that the effect on x is scaling each coordinate of x by the eigenvvalues.
- Theoretical guarantees: to apply this reduction program, we will need to some guarantee, about the eigenspace of A. This leads to the study of its properties: whether eigenvectors of distinct eigenvalues are linearly independent, the dimension of eigenspace for each eigenvalue, and so on.
- Determining eigenvalues: once we establish this program of finding matrix power and determining the effect of linear map, we will need to solve eigenvalues. For this, we study special matrices where eigenvalues are easy to determine, and the rules of reducing to special matrices (without changing eigenvalues). And the algorithms of numerically determining eigenvalues.

Questions of eigenvalues and eigenvectors:

- What matrices are diagonalizable?
- If a matrix A has fewer than n eigenvectors, what can we say about Ax? Perhaps for some x, the effect of Ax can be viewed using eigenvectors?

Matrix rank and eigenvectors [personal throughts]

- If a $n \times n$ matrix has eigendecomposition, then it must be full ranked (similar to a diagonal matrix), but the inverse is not true.
- Question: what are eigenvalues and eigenvectors of low-rank matrices? The general idea is that a low-rank matrix projects vectors into a low-dim. space, which is defined by the eigenvectors. In other words, applying low-rank matrix on a vector is equivalent to a vector in a lower-dim. space defined by eigenvectors.
- Ex. suppose we have
$$A = \begin{bmatrix} a & b \\ a & b \end{bmatrix} \tag{2.89}$$
Then given (x_1, x_2), we have $y_1 = ax_1 + bx_2 = y_2$. So y falls in the diagonal lines (eigenvector).

Characteristic polynomial:

- Eigenvalues and characteristic polynomial: for a given square matrix A, its characteristic polynormial is defined as a polynomial of λ: $\det(A - \lambda I)$. An eigenvalue of A must satisfy the characteristic equation (from the definition of eigenvalues):
$$\det(A - \lambda I) = 0 \tag{2.90}$$
- Basic relationship of eigenvalues: if $\lambda_1, \cdots, \lambda_n$ are eigenvalues of A, then:
$$\prod_{i=1}^{n} \lambda_i = \det A \qquad \sum_{i=1}^{n} \lambda_i = \mathrm{tr} A \tag{2.91}$$
Proof: follow from the characteristic equation.

Properties of eigenvalues and eigenvectors:

- **Eigenvalues of triangular matrix**: if A is triangular, then eigenvalues of A are diagonal elements of A.
  Proof: using the characteristic polynomial and the property that determinant of upper triangular matrix is the product of diagonal elements.
  Remark: while triangular matrices are easy to solve, the elementary row operations do not preserve eigenvalues, so we cannot easily reduce a matrix to upper triangular to obtain eigenvalues.

- Power of matrix: if λ is an eigenvalue of A with corresponding eigenvector x, then λ^n is an eigenvalue of A^n with corresponding eigenvector x ($n > 0$).
  Proof: $A^n x = A^{n-1}(Ax) = \lambda A^{n-1}x$, and do this repeatedly.

- Matrix inverse: suppose λ is an eigenvalue of an invertible matrix A, and x is one corresponding eigenvector, then λ^{-1} is the eigenvalue of the matrix A^{-1}:

$$A^{-1}x = A^{-1}(\lambda^{-1}Ax) = \lambda^{-1}x \tag{2.92}$$

- Matrix transpose: A and A^T have the same eigenvalues, and if x is an eigenvector of A belonging to λ_1 and y is an eigenvector of A^T belonging to λ_2, $\lambda_1 \neq \lambda_2$, then:

$$\langle x, y \rangle = 0 \tag{2.93}$$

  Proof: for any eigenvalue of A: $\det(A^T - \lambda I) = \det(A - \lambda I)^T = 0$. For the second part:

$$\lambda_1 x^T y = x^T A^T y = x^T \lambda_2 y = \lambda_2 x^T y \tag{2.94}$$

  Since $\lambda_1 \neq \lambda_2$, x and y must satisfy: $x^T y = 0$.
  Remark: the proof exploits fundamental symmetry of A and A^T. A simple corollary is: if A is symmetric (e.g. $A^T A$), then any two eigenvectors belonging to two eigenvalues are orthogonal.

- Matrix product: AB and BA have the same eigenvalues.
  Proof: suppose $ABv = \lambda v$, we have: $BA(Bv) = B(\lambda v) = \lambda(Bv)$, thus λ is an eigenvalue of BA with eigenvector Bv.

Similar matrices:

- Motivation: we need some reduction of matrices that preserve eigenvalues (elementary row operations not), and this leads to the concept of similarity.

- Definition: two $n \times n$ matrices A and B are similar if there is an invertible matrix P s.t. $P^{-1}AP = B$, written as $A \sim B$. It is easy to show that similarity satisfies equivalence relation.

- Properties of similar matrices: if A and B are similar, then:
  - $\det A = \det B$.
  - A and B have the same rank, and A is invertible iff B is invertible.
  - A and B have the same characteristic polynomial and thus the same eigenvalues:

$$\det(A - \lambda I) = \det[S^{-1}(A - \lambda I)S] = \det(S^{-1}AS - \lambda I) \tag{2.95}$$

- Remark: geometrically, similar matrices differ only in the basis, so their eigenvalues should be the same.

Eigen Decomposition and its conditions:

- Motivation: the dimensions of eigenspaces, and the condition when A has n linearly independent eigenvectors (which serves as a basis of $\mathbf{R}^n$).

- Theorem: let A be an $n \times n$ matrix, and $\lambda_1, \cdots, \lambda_m$ be distinct eigenvalues of A with corresponding eigenvectors $v_1, \cdots, v_m$, then $v_1, \cdots, v_m$ are linearly independent.
  Intuition: each v_j represents one distinct effect of A, thus should be linearly independent. Consider a simple case of $n = 2$, then if the condition does not hold, $v_2 = kv_1$. Clearly,

$$\lambda_2 v_2 = Av_2 = A(kv_1) = k(Av_1) = k\lambda_1 v_1 = \lambda_1 v_2 \tag{2.96}$$

  So this leads to $\lambda_1 = \lambda_2$.
  Proof idea: by contradiction. If it does not hold, then one eigenvector can be written as a linear combination of other eigenvectors, use similar idea to demonstrate that this must lead to the equality of some λ_i's.

- Diagnoalizable: defintion, A is diagonalizable if there is a diagnoal matrix D s.t. A is similar to D, $P^{-1}AP = D$.

- **Theorem**: A is $n \times n$ matrix, then A is diagonalizable iff A has n linearly-independent eigenvectors. Let $x_1, \cdots, x_n$ be the n eigenvectors, $D = \text{diag}(\lambda_1, \cdots, \lambda_n)$ and $P = (x_1 \cdots x_n)$ (x_i as column vector), we can write A as:

$$A = PDP^{-1} \tag{2.97}$$

  Proof: use $Ax_i = \lambda_i x_i$, we have: $AP = (\lambda_1 x_1 \cdots \lambda_n x_n) = PD$.
  Remark: the problem is then when does A has n independent eigenvectors. One scenario is that A has n distinct eigenvalues - this is trivial. In general, A may have fewer eigenvalues, for example, it may happen that the stretching factors are the same in two directions. The questions are: whether one eigenvalue can only generate one eigenvector, and the relationship between the *algebraic multiplicity* and *geometric multiplicity* (dimension of eigenspace).

- Lemma: if A is an $n \times n$ matrix, then the geometric multiplicity of each eigenvalue is less than or equal to its algebraic multiplicity.
  Intuition: clearly, the sum of algebraic multiplicity of all eigenvalues is n, so the sum of geometric multiplicity must be equal or smaller - otherwise, the eigenvectors are not independent. So in general, geometric multiplicity is less than or equal to algebraic multiplicity.

- **The Diagonalization Theorem**: let A be an $n \times n$ matrix with distinct eigenvalues $\lambda_1, \cdots, \lambda_k$. These statements are equivalent:
  - A is diagonalizable.
  - The union of the bases of the eigenspaces of A contains n vectors.
  - The algebraic multiplicity of each eigenvalue is equal to its geometric multiplicity.

- Application of Eigen-Decomposition Theorem: suppose A is diagnolizable, $A = PDP^{-1}$. Consider any x, we write $x = \sum_i c_i u_i$ where u_i is the i-th eigenvector. We have then:

$$Ax = \sum_i \lambda_i c_i u_i \tag{2.98}$$

  Or in matrix form: $x = Pc$, where $c = [c_1, \cdots, c_n]^T$, then $Ax = APc = PDc$. In plane language: suppose x has coordinates c in the eigenspace, then the effect of A is a new vector with coordinates $\lambda_i c_i$ in the eigen-space.

- Example of non-diagonalizable matrices: rotation matrix, under no direction the effect of the matrix is stretching.

Power and exponential of a matrix: simple form for a diagonalizable matrix:

- Suppose $A = SDS^{-1}$, then for an integer k:

$$A^k = SD^kS^{-1} = S \cdot \text{diag}(\lambda_1^k, \cdots, \lambda_n^k) \cdot S^{-1} \tag{2.99}$$

Similarly, for the exponential:

$$e^A = Se^DS^{-1} = S \cdot \text{diag}(e^{\lambda_1}, \cdots, e^{\lambda_n}) \cdot S^{-1} \tag{2.100}$$

- **Theorem** (effect of power of matrix): suppose A is diagonalizable $n \times n$ matrix, with eigenvalues, $\lambda_1, \cdots, \lambda_n$, and let $x_1, \cdots, x_n$ be the corresonding eigenvectors. Consider a vector x expressed as: $x = c_1x_1 + \cdots c_nx_n$, then effect of A^k on x is:

$$A^kx = c_1\lambda_1^kx_1 + \cdots + c_n\lambda_n^kx_n \tag{2.101}$$

Proof: first, $Ax = c_1Ax_1 + \cdots c_nAx_n = c_1\lambda_1x_1 + \cdots c_n\lambda_nx_n$. Extend this step, it is easy to have the results of A^kx.

Schur Decomposition:

- Schur Decomposition Theorem: for an $n \times n$ matrix A with complex entries, there exists a unitary matrix Q and an upper triangular matrix U (diagonal elements are eigenvalues of A) s.t. A can be expressed as:

$$A = QUQ^{-1} \tag{2.102}$$

Proof: since A is complex matrix, it must have eigenvalues. Let x_1 be an eigenvector belonging to λ_1, expand x_1 into orthonomal basis $x_1, w_2, \cdots, w_n$, then:

$$A(x_1w_2 \cdots w_n) = (\lambda_1x_1Aw_2 \cdots Aw_n) = \begin{pmatrix} \lambda_1 & A_{11} \\ 0 & A_{22} \end{pmatrix} (x_1 \cdots w_n) \tag{2.103}$$

Apply the same procedure to A_{22} recursively.

- Remark: the theorem states that for any complex matrix A, there exists an orthonomal basis s.t. the linear map of A can be represented by an upper triangular matrix. Comparing with Eigen Decomposition: the eigenvectors in general are not orthogonal to each other.

2.7.1 Numerical Methods for Computing Eigenvalues

Finding eigenvalues and eigenvectors [personal thoughts]: it is not easy to directly solve eigenvalues, so we find "signatures" of eigenvalues.

- Use the definition: eigenvector does not change direction. So we can start with a random vector x_0, compare x_0 and Ax_0, and try to close the direction.
- Use the fact that eigenvalues is the scaling factor of the effect of A in the direction of eigenvectors: so if we apply A repeatedly, eventually, the direction of the dominant eigenvector will be revealed.
- Use MVN: eigenvector is related to the direction of maximum PDF, so we can explore the eigenvectors by exploring the density plot. Remark: this only applies to symmetric matrices.

The Power Method:

- Intuition: from the power of matrix, Equation 2.101, we know that A^kx, as k approaches infinity, will be dominated by the largest eigenvalue, so we can simply obtain A^kx at large k, and the results will tell the dominant eigenvalue (largest in absolute value).

- Theorem: let A be $n \times n$ diagonalizable matrix with dominant eigenvalue λ_1, then there exsists vector x_0 s.t. the sequence $\{x_k = A^k x_0\}$ approaches a dominant eigenvector of A.
  Proof: from Equation 2.101, we factorize λ_1^k, clearly, the remaining terms all approach 0, so we have:

$$x_k = A^k x_0 \to c_1 \lambda_1^k v_1, \text{ as } k \to \infty \tag{2.104}$$

  where v_1 is the dominant eigenvector.

- The power method: while we can simply apply this Theorem to any x_0, the problem is that the power may grow too fast, causing numerical problem. So we use scaling at each step, i.e. consider the sequence $\{y_k\}$, $y_k = x_k/m_k$, where m_k is the scaling factor. We choose m_k be the largest component of x_k (absolute value).

The Shifted Power method: the power method only solves the dominant eigenvalue. To solve the other eigenvalues, we reduce the problem so that the second largest eigenvalue is the dominant eigenvalue of a new matrix.

- Theorem: if λ is an eigenvalue of A, then $\lambda - \alpha$ is an eigenvalue of $A - \alpha I$ for any scalar α.
  Proof: trivial.

- The shifted power method: suppose λ_1 is the dominant eigenvalue, we consider the matrix $A - \lambda_1 I$, then it has eigenvalues $0, \lambda_2 - \lambda_1, \lambda_3 - \lambda_1, \cdots$. We can then apply the power method to obtain $\lambda_2 - \lambda_1$, and so on.

2.7.2 Finite State Markov Chains

Markov chain:

- Probability vector: x is an n-dim. vector, with the properties that $0 \le x_i \le 1$ and $\sum_i x_i = 1$.

- Stochastic matrix: this is the transition matrix of a Markov chain. P is stochastic if any column of P is a probability vector. For Markov chain, $P_{ij} = P(j \to i)$, the transition prob. from state j to i.

- Markov chain dynamics: let x_k be the probability vector at the k-th iteration, then the probability of being in state i at time $k+1$ is: $x_{k+1,i} = \sum_j p_{ij} x_{k,j} = p_i x_k$, where p_i is the i-th row of the transition matrix P. Write in matrix form:

$$x_{k+1} = P x_k \tag{2.105}$$

  In particular, we have: $x_k = P^k x_0$ where x_0 is the starting state, and P_{ij}^k is the probability of moving from state j to i after k iterations.

Convergence of Markov chain:

- Problem: from the power of matrix, we know that $P^k x_0$ will be dominated by the largest eigenvalues, so the problem is to determine if P is diagonalizable and its dominant eigenvalues.

- **Theorem (largest eigenvalue)**: if P is the $n \times n$ transition matrix of a MC, then 1 is an eigenvalue of P. Furthmore, any other eigenvalues $|\lambda| \le 1$. If P is regular, and $\lambda \ne 1$, then $|\lambda| < 1$.
  Proof: first to show 1 is an eigenvalue of P, we consider its transpose P^T. Using the fact that row sum of $P^T = 1$, it is easy to see that 1 is an eigenvalue with eigenvector $\mathbf{1}$ (all 1's).
  Second part: show that 1 is the largest eigenvalue. Intuitively, all the terms of P are probabilities, so λ cannot be very large in order to satisfy $Px = \lambda x$. Let k be the largest component of x, the eigenvector of λ, and $|x_k| = m$. The k-th component of Px:

$$|(Px)_k| = |p_{1k}x_1 + \cdots p_{nk}x_k| \le m(p_{1k} + \cdots p_{nk}) = m \tag{2.106}$$

  So the eigenvalue cannot be greater than 1.

- **Theorem (convergence of MC)**: let P be a regular $n \times n$ transition matrix, then as $k \to \infty$, P^k approaches an $n \times n$ matrix L whose columns are identical, each corresponding to a vector x, which is the steady state probability vector for P.
  Proof: follow from Equation 2.101, we consider $P^k e_i$ where e_i is the i-th standard basis vector (this is the i-th column of P^k).
  Remark: the rate of convergence is determined by the second largest eigenvalue (< 1). If it is small, the contribution of this eigenvalue (and the even smaller ones) would converge quickly to 0.

2.8 Orthogonality

Motivation:

- Projection provides a **simple representation** of vectors: coordinates, which allow simple algebra of vectors. Ex. if a problem is expressed in the form of non-orthogonal basis, then by first recasting the problem in orthogonal basis, we may simplify the problem.
- Motivating problems: (1) given a vector, find its representation using an orthogonal basis. (2) Orthogonality has many applications: e.g. minimizing distance, volume.
- Remark: finding simple representations is a fundamental problem of mathematics. Ex. primer factorization of numbers, Fourier series of functions.

Orthogonality and projection:

- Inner product: suppose x and u are two vectors. The inner product of x on u can be written as:

$$\langle x, u \rangle = u^T x = x^T u \tag{2.107}$$

  It is the area defined by the parallelogram of x and u.

- Projection and angle: the angel between x and u is given by:

$$\cos(\theta) = \frac{\langle x, u \rangle}{\|x\|\|u\|} \tag{2.108}$$

  Suppose the projection of x on u is given by βu, where β is a scalar, then the vector $x - \beta u$ is perpendicular to u:

$$\langle x - \beta u, u \rangle = 0 \tag{2.109}$$

  Thus:

$$\langle x, u \rangle - \beta \langle u, u \rangle = 0 \Rightarrow \beta = \frac{\langle x, u \rangle}{\|u\|^2} \tag{2.110}$$

  When u is a unit vector, we have the coordinate of x on u is simply $\langle x, u \rangle$.

- **Projection in matrix form**: suppose U is a set of n basis vectors of $\mathbb{R}^m$ (m and n can be different) where $u_i, 1 \leq i \leq n$ (column vector) is the i-th basis vector. The projection of v on the basis U can be written in the matrix form:

$$v = \sum_i x_i u_i = (u_1 \cdots u_n) \begin{pmatrix} x_1 \\ \cdots \\ x_n \end{pmatrix} = Ux \tag{2.111}$$

  where x_i is a scalar. The interpretation is: suppose we have a vector v, its coordinates in basis U is given by $v = Ux$. Note that U does not have to be orthogonal (should be invertible).

- **Coordinates under an orthogonal basis**: let $\{u_1, \cdots, u_n\}$ be an orthogonal basis for a subspace W of $\mathbb{R}^m$, and let v be a vector in W, then there are unique scalars $x_1, \cdots, x_n$ s.t.

$$v = x_1u_1 + \cdots + x_nu_n, \quad x_i = \frac{v \cdot u_i}{u_i \cdot u_i}, \text{where } i = 1, \cdots, n \tag{2.112}$$

  When the basis is orthonomal, i.e. $\|u_i\| =$, we have: $x_i = v \cdot u_i = u_i^T v$. We can write this in the matrix form:

$$x = \begin{pmatrix} u_1^T v \\ \cdots \\ u_n^T v \end{pmatrix} = \begin{pmatrix} u_1^T \\ \cdots \\ u_n^T \end{pmatrix} v = U^T v \tag{2.113}$$

Orthogonal matrices:

- An orthogonal matrix: is a square matrix with real entries whose columns are orthogonal unit vectors (i.e., orthonormal). May also be called orthonormal matrix. An orthogonal matrix Q satisfies:

$$QQ^T = Q^TQ = I \tag{2.114}$$

  i.e. the inverse of an orthogonomal matrix is its transpose.

- Preservation of dot product of vectors: Q is an orthogonal matrix if any only if for any two vectors x and y:

$$\langle Qx, Qy\rangle = \langle x, y\rangle \tag{2.115}$$

  And a special case is: for any $x \in \mathbb{R}^n, \|Qx\| = \|x\|$.
  Proof: First, it is easy to check that if Q is orthogonal, then it will have the stated properties. The converse can be proven by choosing special vectors e_i and e_j, and use $q_i \cdot q_j = Qe_i \cdot Qe_j = e_i \cdot e_j$.
  Thus **the geometric effect of orthogonal matrix** is: shape-preserving map as it preserves both distance and angle. Examples: reflection and rotation.

- If Q is orthogonal, then $Q^{-1} = Q^T$ is orthogonal, $\det Q = 1$ or -1. If λ is an eigenvalue of Q, then $|\lambda| = 1$. If Q_1 and Q_2 are orthogonal, then so is Q_1Q_2.
  Proof: suppose $Qv = \lambda v$, then we have: $\|v\| = \|Qv\| = \|\lambda v\| = \lambda\|v\|$. The intuition should be clear, if Q has an eigenvector, then the effect along the direction must preserve the distance.
  Remark: the geometric interpretations are clear: orthogonal matrix has volume 1 (preserving distance for identity matrix), eigenvalue 1 or -1 (no scaling) and the product of two orthogonal matrices still preserve distance.

- Unitary matrix: the complex analogue of orthogonal matrix. An unitary matrix U satisfies: $UU^H = U^HU = I$.

- Remark: intuitively, an orthonormal matrix is like identity matrix: orthogonal with unit length. Conjecture: any orthonormal matrix can be transformed to identity matrix by certain operations: translation, rotation.

Orthogonal complements and orthogonal projections:

- Definition: let W be a subspace of $\mathbb{R}^n$, the set of vectors orthogonal to W is called the orthogonal complement:

$$W^\perp = \{v \in \mathbb{R}^n : v \cdot w = 0, \forall w \in W\} \tag{2.116}$$

- Orthogonal complements for row and column space of matrix: let A be $m \times n$ matrix, then

$$\text{row}(A)^\perp = \text{null}(A) \quad \text{col}(A)^\perp = \text{null}(A^T) \tag{2.117}$$

  Proof: follow from the defintion, if x is orthogonal to the row space of A, then we have $A_ix = 0, \forall i$ where A_i is a row vector. So x is in the null space of A.

- Orthogonal projections: let W be a subspace of $\mathbb{R}^n$, and $\{u_1, \cdots, u_k\}$ be an orthogonal basis of W. Then for any vector v, its orthogonal projection into W is defined as:

$$\text{proj}_W(v) = \left(\frac{u_1 \cdot v}{u_1 \cdot u_1}\right) u_1 + \cdots + \left(\frac{u_k \cdot v}{u_k \cdot u_k}\right) u_k \tag{2.118}$$

  The component of v orthogonal to W is the vector:

$$\text{perp}_W(v) = v - \text{proj}_W(v) \tag{2.119}$$

- **The Orthogonal Decomposition Theorem**: let W be a subspace of $\mathbb{R}^n$ and v be a vector, then there are unique vectors w in W and $w^\perp$ in $W^\perp$ s.t. $v = w + w^\perp$.
  Remark: the vector w and $w^\perp$ does not depend on the basis of W. The corollary is: in n-dim. space, $\dim(W) + \dim(W^\perp) = n$.
  Remark: the theorem provides the foundation for simple representation of vectors: any vector can be decomposed into the sum of vectors from orthogonal complements.

- **The Rank Theorem from Geometric perspective**: if A is an $m \times n$ matrix, then:

$$\text{rank}(A) + \text{nullity}(A) = n \tag{2.120}$$

  Proof: rank A is the dim. of the row space, and its orthogonal space is null(A). The sum of the dim. of the two space should be n.

Gram-Schmidt Proess and QR factorization:

- Gram-Schmidt process: convert any basis into an orthogonal basis. Let $\{x_1, \cdots, x_k\}$ be the basis of a subspace W of $\mathbb{R}^n$. Then we can define the following vectors: starting with $v_1 = x_1$,

$$v_2 = x_2 - \frac{x_2 \cdot v_1}{v_1 \cdot v_1} v_1 \tag{2.121}$$

$$v_3 = x_3 - \frac{x_3 \cdot v_1}{v_1 \cdot v_1} v_1 - \frac{x_3 \cdot v_2}{v_2 \cdot v_2} v_2 \tag{2.122}$$

  and so on. In other words, for each x_k, we find v_k that is perpendicular to the subspace defined by v_1 to v_{k-1}.

- **Theorem (QR factorization)**: express the Gram-Schmidt process in matrix terms. A $m \times n$ matrix A of rank n (linearly independent column vectors), it can be written as: $A = QR$, where Q is a $m \times n$ orthonomal matrix (the column vectors are orthogonal to each other, and are unit vectors), and R is $n \times n$ upper triangular, invertible matrix.
  Proof: we express a_j's (column vector) in terms of q_j's - the orthonomal basis coming from Gram-Schmidt Process. So $a_1 = r_{11}q_1$, $a_2 = r_{12}q_1 + r_{22}q_2$, and so on. Write this in matrix form.

- Remark: (1) the significance is that it provides *another (possibly simpler) representation* of full-ranked matrices (or linearly-independent vectors), using the orthogonal basis. (2) The general idea of matrix factorization: transformation of vectors (row or columns). Ex. LU factorization.

Applications of QR factorization:

- Intuition: QR factorization reduces a matrix into the product of two much simpler matrices, so we can use it to address some problems of the original matrix.

- Finding the determinant: if $A = QR$, then $\det A = \det Q \det R$, where $\det Q$ is 1 or -1, and $\det R$ is the product of the diagonal elements. Remark: the geometric intuition is that, to find the volume of a parallelepiped, we obtain the heights, and the volume is easy to determine.

- Finding the inverse: suppose A is invertible, then $A^{-1} = R^{-1}Q^{-1} = R^{-1}Q^T$. The inverse of R is simply determined by back-substitution.

Least squares problem: let A be a $m \times n$ matrix of rank n, and $\hat{x}$ be the solution to the linear squares problem $Ax = b$, i.e. $\hat{x}$ minimizes $\|Ax - b\|$. Then $b - A\hat{x}$ should be orthogonal to the column space of A (projectio of b to the space). Note that: $N(A^T) = R(A)^{\perp}$. Thus $b - A\hat{x}$ should be in $N(A^T)$ and satisfy:

$$A^T(b - A\hat{x}) = 0 \tag{2.123}$$

Or $\hat{x} = (A^TA)^{-1}A^Tb$.

2.9 Symmetric Matrices and Quadratic Forms

Connection between orthogonal eigenvectors and symmetric matrices [personal thoughts]:

- Motivation: suppose the eigenvectors of a matrix are orthogonal (desired), what does it say about the matrix?
- If A has orthogonal eigenvectors, then A must be symmetric. Wlog, suppose all eigenvectors v_i have norm 1, then the matrix $P = (v_1, \cdots, v_n)$ is orthogonal matrix. So we have $A = PDP^{-1} = PDP^T$, where D is the diagonal matrix of eigenvalues. It is easy to prove that $A^T = A$.

Symmetric and Hermitian matrices:

- Definition: A is a Hermitian matrix if $A^H = A$, where H stands for complex conjugate transpose. This is a general case of real symmetric matrix.
- **Eigenvectors of Hermitian matrices**: the eigenvalues of an Hermitian matrix are all real, and eigenvectors belonging to distinct eigenvalues are orthogonal.
  Proof: follows the results of eigenvectors of matrix transpose. In particular, suppose λ_1 and λ_2 are distinct eigenvalues with eigenvectors v_1 and v_2, then

  $$\lambda_1 v_1^T v_2 = (Av_1)^T v_2 = v_1^T A^T v_2 = v_1^T \lambda_2 v_2 = \lambda_2 v_1^T v_2 \tag{2.124}$$

  Since $\lambda_1 \neq \lambda_2$, it must have $v_1^T v_2 = 0$.
- **Spectral Theorem for Hermitian Matrices**: let A be a $n \times n$ matrix, then A is symmetric if and only if it is orthogonally diagnolizable, i.e. there exists an orthonomal basis, Q, consisting of eigenvectors of A, and A has Eigen Decomposition ($\lambda_i, 1 \leq i \leq n$ is real):

  $$A = QDQ^H = Q \cdot \text{diag}(\lambda_1, \cdots, \lambda_n) \cdot Q^H \tag{2.125}$$

  Proof: (1) follow from Eigen Decomposition and from the orthogonality of eigenvectors. (2) Or, use Schur Decomposition, and show that the upper triangular matrix is diagonal.
- **Spectral Theorem for Real Symmetric Matrices**: apply the above Theorem to a real symmetric matrix, A is symmetric if and only if A has a simple Eigen Decomposition, where Q is orthonogal matrix:

  $$A = QDQ^T \tag{2.126}$$

  Proof: it is known D is real. If λ is real, then we can show that its corresponding eigenvectors are real (from solving the linear equation $(A - \lambda I)x = 0$). Suppose n eigenvectors (with norm 1) are $q_1, \cdots, q_n$, then they form an orthonomal basis. We have:

  $$A[q_1 \cdots q_n] = [Aq_1 \cdots Aq_n] = [\lambda_1 q_1 \cdots \lambda_n q_n] = [q_1 \cdots q_n]D \tag{2.127}$$

  From $AQ = QD$, we have $A = QDQ^{-1} = QDQ^T$.

- **Spectral Decomposition**: we can write the spectrum theorem in a different form, if A is orthonogally diagonizable, $A = QDQ^T$, then we can write it as:

$$A = \lambda_1 q_1 q_1^T + \cdots + \lambda_n q_n q_n^T \tag{2.128}$$

  Each of the terms $\lambda_i q_i q_i^T$ is a symmetric, rank 1 matrix (easy to prove: each row is a multiple of q_i). This is similar to Fourier series expansion of some function, and we can see that A would be dominated by terms from large eigenvalues (since q_i are orthonomal basis, their norms are all similar).

- Inverse of symmetric matrices: if A is symmetric, then A^{-1} is also symmetric. In fact, $A = QDQ^T$, and $A^{-1} = QD^{-1}Q^T$.

Quadratic forms:

- Definition: a quadratic form is a homogeneous polynomial of degree 2 (inhomogeneous polynomials can be mapped to homogeneous ones by translation):

$$q(x_1, \cdots, x_n) = \sum_{i,j} a_{ij} x_i x_j \tag{2.129}$$

- Quadratic forms in matrix notation: let $A = (a_{ij}$ be an $n \times n$ sqaure matrix, we have:

$$x^T Ax = \sum_i a_{ii} x_i^2 + \sum_{i<j} (a_{ij} + a_{ji}) x_i x_j \tag{2.130}$$

  If A is not symmetric, we could define a symmetric matrix A' as: $a'_{ij} = a'_{ji} = \frac{1}{2}(a_{ij} + a_{ji})$. With A being symmetric, we could write quadratic form as:

$$x^T Ax = \langle Ax, x \rangle \tag{2.131}$$

- A special quadratic form: when there is no $x_i x_j$ term ($i \neq j$), the quadratic from can be written as:

$$\sum_i \lambda_i x_i^2 = \sum_i (\lambda_i x_i) x_i = \langle Dx, x \rangle = x^T Dx \tag{2.132}$$

  where D is a diagonal matrix: $D = \text{diag}(\lambda_1, \cdots, \lambda_n)$.

- **Standardization of quadratic forms in a diagonal form**: Since A is symmetric, we could write A as: $A = UDU^T$, plug in this:

$$x^T Ax = x^T UDU^T x = (U^T x)^T D (U^T x) \tag{2.133}$$

  Thus let $x' = U^T x$, or $x = Ux' = \sum_i x'_i u_i$, in other words, x' is the coordinates of x on U. We have:

$$x^T Ax = \sum_{i=1}^n \lambda_i x_i'^2 = x'^T Dx' \tag{2.134}$$

- Geometric view of diagonalization: when A is pd, these two views are equivalent: the contour $x^T Ax = c$ is an ellipse; the curve $x^T Ax$ formed by $\|x\| = 1$ is an ellipse with the axis defined by eigenvector of A. To prove the latter, we consider the shape of the curve in terms of $x' = Q^T x$.

- Complete the square: suppose A is a symmetric matrix, b is a vector, and c a scalar, we have:

$$x^T Ax + b^T x + c = \left(x + \frac{1}{2} A^{-1} b\right)^T A \left(x + \frac{1}{2} A^{-1} b\right) + (c - \frac{1}{4} b^T A^{-1} b) \tag{2.135}$$

- Cholesky decomposition: suppose A is a symmetric, positive definite matrix, and C is the Cholesky decomposition of A, i.e. $CC^T = A$, then we could write quadratic form as:

$$x^T Ax = x^T CC^T x = (C^T x)^T (C^T x) \tag{2.136}$$

  Thus the quadratic form is simply the norm of the vector $C^T x$. The advantage of this representation is that eigenvalues are not explicitly used.

- Rank of quadratic form: we have this theorem:

$$\text{rank}(AA^T) = \text{rank}(A^T A) = \text{rank}(A) \tag{2.137}$$

  Proof: we show that $Ax = 0$ iff $A^T Ax = 0$. First, it is obvious that when $Ax = 0$, we have $A^T(Ax) = 0$. Next, if $A^T Ax = 0$, we have $x^T(A^T A)x = 0$, or $\|Ax\| = 0$, so $Ax = 0$.

Maximization/minimization of quadratic forms [Wiki]:

- Rayleigh quotient: let A be an $n \times n$ Hermitian matrix, then Rayleigh quotient for a column-vector x, is defined as:

$$R(A, x) = \frac{x^T Ax}{x^T x} \tag{2.138}$$

  The normalization by $x^T x$ (the norm of x) is necessary as otherwise, $x^T Ax$ can be arbitrarily large when $\|x\|$ is large. Suppose $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ be the eigenvalues of A, then:

$$\max_{x \in \mathbb{R}^n} R(A, x) = \lambda_1 \qquad \min_{x \in \mathbb{R}^n} R(A, x) = \lambda_n \tag{2.139}$$

  And the solutions are obtained at the first and the last eigenvectors of A, respectively.
  Proof: use the standarization of the quadratic form above:

$$x^T Ax = \sum_i \lambda_i x_i'^2 \qquad x^T x = x'^T x' = \sum_i x_i'^2 \tag{2.140}$$

  Thus we have:

$$\lambda_1 \geq \sum_i \lambda_i x_i'^2 / \sum_i x_i'^2 \geq \lambda_n \tag{2.141}$$

- **Geometric view of Rayleign quotient**: we want to max. or min. $x^T Ax$ subject to $\|x\| = 1$. We consider contours of ellipse $x^T Ax$, and intersections with the circle $\|x\| = 1$. In 2D case, it is clear that the maximum and minimum occur where the ellipse and the circle are tangent. At max, x is in the direction of the longer axis of the ellipse (i.e. larger eigenvalue), and min. is achieved at the direction of the smaller eigenvalue.

- Rayleigh quotient can also be proven using Lagrange's multiplier method: we have the optimization problem:

$$\max_x x^T Ax \text{ subject to } \|x\| = 1 \tag{2.142}$$

  The Lagrange's multiplier method would optimize the function $f(x, \lambda) = x^T Ax - \lambda(x^T x - 1)$. Solving the derivative at 0:

$$\frac{\partial f(x, \lambda)}{\partial x} = 2x^T A - 2\lambda x^T = 0 \Rightarrow Ax = \lambda x \tag{2.143}$$

  where we use the result that $\partial(x^T Ax)/\partial x = x^T(A + A^T)$. Thus the solution x is an eigenvalue of A.

- Generalized Rayleigh quotient: let A and B be real symmetric positive-definite matrices, the generalized Rayleigh quotient is defined as:
$$R(A, B, x) = \frac{x^T Ax}{x^T Bx} \tag{2.144}$$
It can be reduced to Rayleigh quotient by Cholseky decomposition of B: $B = D^T D$. Let $C = D^T AD^{-1}$ and $y = Dx$:
$$\frac{x^T Ax}{x^T Bx} = \frac{x^T D^T D^{-T} AD^{-1} Dx}{x^T D^T Dx} = \frac{y^T Cy}{y^T y} \tag{2.145}$$
Thus the solution is the first eigenvector of the real symmetric positive definite matrix C, and the maximum is the first eigenvalue of C. Note that the eigenvalues of C and $B^{-1}A$ are equal: if x is an eigenvector of $B^{-1}A$ belonging to λ, then Dx is an eigenvector of C belonging to λ. More generally, we have (see Theorem 2.5. of Applied Multivariate Statistical Analysis by [Hardle et al, 4ed, 2015]):
Theorem: let $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_p$ be the eigenvalues of $B^{-1}A$, we then have:
$$\max_x \frac{x^T Ax}{x^T Bx} = \lambda_1 \geq \lambda_2 \geq \cdots \lambda_p = \min_x \frac{x^T Ax}{x^T Bx} \tag{2.146}$$

- Application of Rayleigh quotient in graph: suppose we have a graph with the adjacency matrix A. We want to put (continuous) labels on the graph x, with fixed norm, s.t. the labels of adjacent nodes are similar. This is equivalent to maximizing $\sum_{i,j} A_{ij} x_i x_j = x^T Ax$. The results are achieved at x parallel to the largest eigenvector of A.

Positive definite matrices:

- Definition: a real symmetric matrix A is positive definite if $\forall x \in \mathbb{R}^n$, $x^T Ax \geq 0$, with equality iff $x = 0$.
- Conditions of positive definite: given an $n \times n$ symmetric matrix A, then the following conditions are equivalent to A being positive definite:
  - All eigenvalues of A are positive. This follows from: $x^T Ax = \sum_i \lambda_i x_i'^2$, thus positive eigenvalues imply the positive definiteness, and vice visa.
  - The leading principal submatrices (the upper left 1-by-1, 2-by-2, etc. corners) all have positive determinants.
    Proof: each of principal submatrix is positive definite, and thus has positive eigenvalues. The determinant must thus be positive (similarity of submatrix and diagonal matrix).
  - Cholesky Decomposition: there exists a lower triangular matrix (unique) with positive diagonal elements s.t.
    $$A = LL^T \tag{2.147}$$
    Proof: if A is P.D., then $A = UDU^T = U\sqrt{D}\sqrt{D}^T U^T$, where $\sqrt{D} = \text{diag}(\sqrt{\lambda_1}, \cdots, \sqrt{\lambda_n})$. Let $L = U\sqrt{D}$, then $A = LL^T$. If Cholesky decomposition exists for A, then for any x, we have:
    $$x^T Ax = x^T LL^T x = \|L^T x\| \tag{2.148}$$
    Thus it is nonnegative for all x.
- Application of positive definite matrices: e.g. function optimization. Let $F(x)$ be a function in $\mathbb{R}^n$, let x_0 be a stationary point, i.e. $F'(x_0) = 0$, and $H(x_0)$ be the Hessian matrix of F at x_0 (i.e. second partial derivative), we have:
  - If $H(x_0)$ is positive definite, then x_0 is a local minimum.
  - If $H(x_0)$ is negative definite, then x_0 is a local maximum.

  Proof: Taylor expansion of F at x_0: $F(x) \approx F(x_0) + (x - x_0)^T H(x_0)(x - x_0)$.

- Remark:
  - The equivalent conditions of positive definite matrices can be applied to complex matrices, where transpose needs to be replaced with complex conjugate tranpose.
  - A positive-definite matrix is in many ways analogous to a positive real number: all eigenvalues are positive, Cholesky decomposition, etc.

2.10 Distance and Approximation

Reference: [Poole, 2ed, Chapter 7]

Properties of A^TA and AA^T: both are symmetric matrices and have special properties.

- Interpretation of A^TA: the length of Ax can be written as:

$$\|Ax\|^2 = \langle Ax, Ax\rangle = x^TA^TAx \tag{2.149}$$

  which is a quadratic form defined by the matrix A^TA.

- Rank of A^TA is equal to rank of A and A^T.

- If v is an eigenvector of A^TA belonging to λ, then Av is an eigenvector of AA^T belonging to λ.
  Proof: $A^TAv = \lambda v \Rightarrow A(A^TAv) = A(\lambda v) \Rightarrow AA^T(Av) = \lambda(Av)$.

- **Orthogonality**: let v_1 and v_2 be eigenvectors belonging to two distinct eigenvalues of A^TA, then v_1, v_2 are orthogonal and furthermore, Av_1 and Av_2 are also orthogonal.
  Proof: since A^TA is symmetric, we know that v_1, v_2 are orthogonal. For the second part,

$$\langle Av_1, Av_2\rangle = v_1A^TAv_2 = \lambda_2(v_1v_2) = 0 \tag{2.150}$$

- A^TA has orthogonal eigenvectors, and also all the eigenvalues are nonnegative.
  Proof: let λ_i be the eigenvalue of A^TA, corresponding to eigenvector v_i, then

$$\|Av_i\|^2 = v_i^TA^TAv_i = \lambda_i\|v_i\|^2 \geq 0 \tag{2.151}$$

  Thus $\lambda_i \geq 0$.

Singular Value Decomposition (SVD):

- Motivation: any square matrix has eigendecomposition, so Ax has simple linear or orthogonal (if A is symmetric) representation. Using eigenvectors of A as basis, Ax has coordiates λ_ic_i, where c_i is the coordinate of x along the i-th eigenvectors. How can we generalize to arbitrary matrix so that we have simple representation of Ax?

- Motivation: What is the effect of $f(x) = Ax$ in terms of the length of Ax? Suppose $\|x\| = 1$, what would be the norm of Ax?

- **Singular values**: suppose A is $m \times n$ matrix, and we have $\|x\| = 1$ in R^n. The vectors Ax form ellipse in R^m, with axis defined by the eigenvectors of A^TA and length of the half axis given by $\sigma_i = \sqrt{\lambda_i}$, where λ_i is the i-th eigenvalue of A^TA. We call σ_i singular values of A.
  Proof: we have $\|Ax\|^2 = \langle Ax, Ax\rangle = x^TA^TAx$, and this is an ellipise defined by A^TA. Let v_i be the i-th eigevector of A^TA, then $\|Av_i\|^2 = \lambda_i\|v_i\|^2 = \lambda_i$, or $\|Av_i\| = \sigma_i$.

- **Theorem: Singular Value Decomposition**: if A is an $m \times n$ matrix, $m > n$, then there exist $m \times m$ orthogonal matrix U, $n \times n$ orthogonal matrix V, and $m \times n$ diagonal matrix Σ s.t. A can be expressed as:

$$A = U\Sigma V^T \tag{2.152}$$

  Let $\sigma_1, \cdots, \sigma_r$ be the r non-zero singular values of A. The form of Σ is: a diagonal matrix with elements σ_1 to σ_r, and the rest 0. Note: the rank of U cannot be greater than n, so U cannot have more than n basis vectors, effectively it will have r basis and the rest 0.
  Proof (by construction): let $v_i \in \mathbf{R}^n, 1 \leq i \leq n$ be unit eigenvectors of A^TA, and $V = (v_1 \cdots v_n)$, then V is an orthogonal matrix. We show that $Av_i \in \mathbf{R}^m$ are orthogonal: $(Av_i)^T(Av_j) = v_i^T A^T A v_j = v_i^T \lambda_j v_j = \lambda_j v_i^T v_j = 0$, where λ_j is the eigenvalue corresponding to v_j. From the definition of singular values, we know $\|Av_i\| = \sigma_i$. Now let u_i be unit vector along the direction of $Av_i, 1 \leq i \leq r$: $Av_i = \sigma_i u_i$, then u_i are orthogonal. Rewrite $Av_i = \sigma_i u_i$ in matrix form leads to SVD.
  When $r < m$, we only have r vectors for U. To make the difference, we can use Gram-Schimdt process to expand U to orthogonal matrix.

- Interpretations of the matrices: V and U represent the eigenvectors of A^TA and AA^T respectively.
  Proof: to see the later, we plug in the definition of u_i:

$$AA^T u_i = AA^T Av_i/\sigma_i = A(\lambda_i v_i)/\sigma_i = \lambda_i u_i \tag{2.153}$$

- Algorithm of SVD:

  1. Find the n eigenvalues of eigenvectors of A^TA. We have $V = [v_1, \cdots, v_n]$.
  2. Find signular values $\sigma_, \cdots, \sigma_n$, and we have Σ.
  3. Use $Av_i = \sigma_i u_i, 1 \leq i \leq r$, we find u_i. Then expand to orthonomal basis to get U.

Geometric view of SVD:

- The effect of Ax: see Figure 7.18 and 7.19 [Poole]. Suppose $v_i \in \mathbf{R}^n$ is the i-th eigenvectors of A^TA, then v_i's form orthonormal basis of $\mathbf{R}^n$. The effect of A on v_i is stretching by a factor of $\sigma_i = \sqrt{\lambda_i}$, where λ_i is the i-th eigenvector of A^TA. The direction Av_i generally changes, but Av_i's are orthogonal in $\mathbf{R}^m$. We can then normalize Av_i by dividing σ_i, or:

$$u_i = \frac{Av_i}{\sigma_i} \tag{2.154}$$

  With these, we are ready to consider Ax for general x. Our idea is to express x in terms of v_i's: $x = \sum_{i=1}^n c_i v_i$, then:

$$Ax = A\sum_{i=1}^n c_i v_i = \sum_{i=1}^n c_i Av_i = \sum_{i=1}^n c_i \sigma_i u_i \tag{2.155}$$

  So if we use u_i's as orthonormal basis of $\mathbf{R}^m$, the coordinates of Ax are $c_i\sigma_i$ (scaling of original coordinates by σ_i).

- Geometric view of matrix decomposition: `https://blogs.sas.com/content/iml/2017/08/28/singular-value-deco` `html`. See [Gilbert Strang, 1993]. The effect of any matrix A can be viewed as:

  1. Rotation through V^T: V^Tx transforms x to the coordinates of x with v_i's as the basis: $(c_1, \cdots, c_n)^T$
  2. Scaling: convert the vector c into $(c_1\sigma_1, \cdots, c_n\sigma_n)^T$.
  3. Rotation through U: convert this vector into new representation with u_i's as basis.

SVD in outproduct expansion/low rank approximation:

- Approximation: We expand the SVD as

$$A = \sum_{i=1}^{n} \sigma_i u_i v_i^T \tag{2.156}$$

  Note that $u_i v_i^T$ is $m \times n$ matrix with rank 1. If we order $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$, and most σ_i's are small (rank of A is small), then choose the top k ones, and we have the approximation: $A \approx \sum_{i=1}^{k} \sigma_i u_i v_i^T$.

- Geometric view: suppose σ_1 is much larger than all other singular values, then the effect of Ax is mostly on the direction of the first SV, so A can be approximated by considering only its first singular value.

Applications of SVD:

- Rank of matrix: let A be an $m \times n$ matrix, and $A = U\Sigma V^T$ be the SVD of A, then the rank of A is the number of non-zero signular values of A, i.e.:

$$\text{rank}A = \text{rank}\Sigma \tag{2.157}$$

  Proof 1: multiply a matrix by a nonsingular matrix does not change its rank (both U and V are orthognoal, thus nonsingular).
  Proof 2: we can also do this by proof-by-construction. Let r be the number of non-zero signular values of A. Consider the row matrix of A. Its i-th row:

$$A_i = (\sigma_1 u_1 v_1^T + \cdots \sigma_r u_r v_r^T)_i = \sigma_1 u_{1i} v_1^T + \cdots + \sigma_r u_{mi} v_i^T \tag{2.158}$$

  Thus any row of A is linear combination of $v_i^T, 1 \leq i \leq r$.

- Solving linear systems $Ax = 0$ and $A^T x = 0$: let $\sigma_1, \cdots, \sigma_r$ be all the nonzero singular values of A, and $A = U\Sigma V$, then: (a) $\{u_{r+1}, \cdots, u_m\}$ is an orthonormal basis for null(A^T) and (b) $\{v_{r+1}, \cdots, v_n\}$ is an orthonormal basis for null(A).
  Proof: (a) similar to Proof 2 above, we consider the column matrix of A, and we can show that $\{u_1, \cdots, u_r\}$ an orthonormal basis for column space of A, then u_{r+1} to u_m an orthonormal basis of its orthogonal complement. (b) By definition, we know that $Av_{r+1} = \cdots = Av_n = 0$.

- Find pseudoinverse: suppose A has SVD: $A = U\Sigma V^T$, and Σ^+ is the pseudoinverse of Σ, then the pseudoinverse of A is $V\Sigma^+ U^T$. When A is invertible, then we have:

$$A^{-1} = V\Sigma^{-1} U^T \tag{2.159}$$

Vector norm: [Wiki]

- Definition: given a vector space V over a field F, norm is a function $p : V \to F$ that satisfies positive scalability, triangle inequality and the condition: if $p(v) = 0$ then v is a zero-vector.
- Eucledian (L_2) norm: $\|x\| := \sqrt{\sum_i x_i^2}$.
- Manhattan (L_1) norm: $\|x\|_1 := \sum_i |x_i|$.
- L_p norm: $\|x\|_p := (\sum_i x_i^p)^{1/p}$.
- Maximum (L_∞) norm: $\|x\|_\infty := \max(|x_1|, \cdots, |x_n|)$.

Matrix norms [Wiki]

- Sub-multiplicative matrix norm: a matrix norm (for square matrix) is sub-multiplicative if for all matrices A and B in $K^{n \times n}$:

$$\|AB\| \leq \|A\|\|B\| \tag{2.160}$$

- Induced norm: suppose a vector norm is given, we could then define the induced norm for matrices:

$$\|A\| = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|} \tag{2.161}$$

  Thus the norm of A defines the maximum effect (in terms of the size/norm of the vectors) of the linear map A on any vector.

- Property of induced norm:
  - When $m = n$ and one uses the same norm on the domain and the range, the induced operator norm is a sub-multiplicative matrix norm.
  - Induced norm and eigenvalue: for any eigenvalue λ, $|\lambda| \leq \|A\|$. The proof follows from the definition of matrix norm.

- Special induced norms: suppose we have p-norm for vectors. In the case of $p = 1$ and $p = \infty$:

$$\|A\|_1 = \max_j \sum_{i=1}^{m} |a_{ij}| \tag{2.162}$$

$$\|A\|_\infty = \max_i \sum_{j=1}^{n} |a_{ij}| \tag{2.163}$$

  Thus the two correspond to the maximum of the absolute column and row sum, respectively.

Perron-Frobenius Theorem of positive matrices [Wiki]:

- Theorem: A is a $n \times n$ positive matrix, i.e. $a_{ij} > 0$ for all i, j. Then the following statements hold:
  - Perron root or the Perron-Frobenius eigenvalue: there exists a positive eigenvalue, $r > 0$, of A and any other eigenvalue λ (possibly, complex) is strictly smaller than r in absolute value, $|\lambda| < r$.
  - The Perron-Frobenius eigenvalue r is simple: r is a simple root of the characteristic polynomial of A.
  - The eigenvector (both left and right) of A with eigenvalue r is positive, i.e. there exists v s.t. $Av = rv$, $v_i > 0$ for all i; and there exists w s.t. $wA = rw$, $w_i > 0$ for all i.
  - The Perron-Frobenius eigenvalue r satisfies:

$$\min_i \sum_j a_{ij} \leq r \leq \max_i \sum_j a_{ij} \tag{2.164}$$

- Proof of the eigenvalue inequalities: the general idea is that r describes the effect of the linear map A, thus it is related to the matrix elements, especially the matrix norm. Intuitively, when the elements are all small, it's unlikely that the linear map increases the norm by a large extent, i.e. it has a large eigenvalue. First, we use the result of matrix induced norm (choose the infinity norm):

$$|r| \leq \|A\|_\infty = \max_i \sum_{j=1}^{n} |a_{ij}| \tag{2.165}$$

  Next, we need a lower bound of r. Suppose $Aw = rw$ where w is the positive eigenvector, the idea is that if r is too small, the equality cannot hold. Let w_i be the smallest element of w, let it be 1, then $r = (Aw)_i = \sum_j a_{ij} w_j \geq \sum_j a_{ij}$.

- Remark: a simple application to the stochastic matrix, the row sums are all equal to 1, thus $r = 1$.

Perron-Frobenius Theorem of non-negative matrices [Wiki]: with some additional conditions, the Perron-Frobenius results can be applied to non-negative matrices as well:

- Irreducible matrices: a matrix A is irreducible if and only if its associated (directed) graph G is strongly connected.
- Period of irreducible matrices: for any node i, we could define its period d_i as the greatest common divisor of m, where $A_{ii}^m > 0$. One can show that all nodes of A have the same period, called the period of A.
- Theorem: let A be an irreducible non-negative $n \times n$ matrix with period h, then the following statements hold:
  - Perron root or the Perron-Frobenius eigenvalue: there exists a positive eigenvalue, $r > 0$, of A and any other eigenvalue λ (possibly, complex) is strictly smaller than r in absolute value, $|\lambda| < r$.
  - The Perron-Frobenius eigenvalue r is simple: r is a simple root of the characteristic polynomial of A.
  - The eigenvector (both left and right) of A with eigenvalue r is positive, i.e. there exists v s.t. $Av = rv$, $v_i > 0$ for all i; and there exists w s.t. $wA = rw$, $w_i > 0$ for all i.
  - The Perron-Frobenius eigenvalue r satisfies:

$$\min_i \sum_j a_{ij} \leq r \leq \max_i \sum_j a_{ij} \tag{2.166}$$

2.11 Linear Algebra Review

General lessons:

- Functional perspective: the goal of linear algebra is to understand the function $y = Ax$. The key questions are:
  - How does this function affect the dimension? Ex. if x is in 2D, what is the space of Ax? This leads to the study of matrix rank, linear dependency of vectors, etc.
  - What is the scale of Ax? If we know $\|x\|$, what can we say about $\|Ax\|$? If x is in a rectangle, what is the volume of Ax? This leads to the study of eigenvalues, determinant, etc.
  - What is the direction of Ax? To characterize the effect of Ax, we need to understand both the length and direction. The study of eigenvalues/vectors addresses this question.

  The general approach we take is: if A can be viewed as transformation from other matrix (or matrices) B, and we know the property of B, what can we say about the property of A? For instance, suppose we are interested in the dimension (rank):
  - Multiplying a row/column of A by a constant does not affect its rank.
  - $\text{rank}(A^T) = \text{rank}(A)$.
  - $\text{rank}(AB) \leq \min \text{rank}(A), \text{rank}(B)$.

  The important transformations are: product (factorization), transpose, inverse, change of basis (similarity).
- Central role of **representations**: to understand a matrix A, we can find better representations of A. Important representations in linear algebra include: LU factorization, eigendecomposition, QR factorization.

- Geometric perspective: often the best way to understand representations and algebraic transformations is through geometric view. Matrix product: successive transformation; matrix inverse: inverse map; similar matrix: change of basis; quadratic form: ellipse; linear relationship between vectors: dimensions of hyperplanes. More generally:
  - Relationship between vectors: geometric relationship, e.g. parallel, orthogonal.
  - Algebraic transformations on vectors: rotation/translation, etc. on vectors.
  - Functions may be viewed/interpreted as geometric objects: e.g. least square can be viewed as distance. Quadratic form as ellipse.
- Iterative methods for numerical algorithms. In general, the idea is to improve the current solution and the key is to come up with directions for doing this. Examples:
  - Gradient descent: we use gradient to move towards the direction that minimizes the function.
  - Numerical methods for SLE: the intuition is: suppose we have x^{t-1}, then we update it with x^t so that "most" equations are satisfied (greedy solution given values of x^{t-1}), then we will have $\|x^t - x^*\| < \|x^{t-1} - x^*\|$, where x^* is the true solution.

Statistical interpretations:

- A vector can be viewed as a sample from a random variable and vice versa. Thus: variance of a random variable (sample variance) can be viewed as the norm of a vector, and covariance inner product (mean 0), correlation $\cos\theta$. Ex. if $X = Y + Z$, then the variance $S_X = S_Y + S_Z$, this is basically Pythagoras Theorem.
- Linear dependence of in random variables: we can similarly defined linear dependence of RVs, but we cannot easily study the sample vectors of these RVs, as they are not entirely linear-dependent in the same way because of sampling variations. Consider a $m \times n$ matrix A, where row is a RV. When the RVs are linearly dependent (low rank), we expect the rows of A are "approximately" linearly dependent. This is the idea behind low-rank approximation, where we remove the noises.
- If we have multiple RVs, then we can represent the samples as a matrix. The relationship of RVs, e.g. independence, correlation/linear dependency, can be interpreted in terms of the property of matrix, e.g. if all RVs are independent, then the matrix is diagonal.

Linear systems and vector space:

- The effect of a linear map on dimensions: preserve the dimension or reduce the dimension, but cannot increase it. An example that a linear map reduces dimension: projection onto lower dimension, $(x_1, x_2) \to (x_1)$.
- Behavior of linear system: we have $Ax = b$, where A is $m \times n$ matrix:
  - If $m < n$: the system is always undertermined, it always has infinitely many solution.
  - If $m = n$: (1) When A is full ranked, single solution; (2) When $\text{rank}(A) < n$, infinitely many solutions.
  - If $m > n$: overdetermined. Whether it has solutions depend on if rank of A and rank of augemented matrix $A|b$ are the same.
- Linear dependence of vectors, SLE and matrix inverse: a linear map (or matrix) A is invertible iff it does not reduce dimension, i.e. its vectors are linearly independent. To test linear dependency of $A_1, \cdots, A_n$, we solve if there exist $x_1, \cdots, x_n$ s.t. $\sum_i x_i A_i = 0$, and this is just SLE.

Eigenvalues and eigenvectors:

- The central goal is to understand the effect of $y = Ax$. General ideas that are very powerful for understanding the effect of a function:
  - Finding a **representation** of function s.t. we can understand what are the main "components" of a function. Ex. Fourier transform, we write $f(x) = \sum_i c_i f_i(x)$, where each $f_i(x)$ is a periodic function. The coefficients would tell what f_i contributes most to $f(x)$.
  - Consider the effect of f on some special values of x: and try to link $f(x)$ on general x to these special values. Ex. integral of a function (a map from function to real value): if $f = \sum_i c_i x^i$, then the integral of f can be easily obtain from integral of x^i.
- Eigenvalues/eigenvectors provide new representation of Ax: find special u_i's s.t. Au_i is a simple scaling. This leads to a new representation of Ax: we write $x = \sum_i c_i u_i$, then $Ax = \sum_i \lambda_i c_i u_i$. This allows to analyze the main effects: the directions with the largest λ_i's.
- Geometric view of Ax according to eigendecomposition: suppose we have $v = x_1 v_1 + x_2 v_2$, then the effect of Av is: it scales coordinates x_1, x_2 to $\lambda_1 x_1$ and $\lambda_2 x_2$ in the direction of v_1, v_2.
- Important property is that generally, eigenvectors of A and A^T are orthogonal. Intuitively, transpose and orthogonality are related: e.g. $(a, b) \cdot (-b, a) = 0$. More specifically, we can show in 2D, if $x = (x_1, x_2)$ is an eigenvector of A, then x' orthogonal to x, $(-x_2, x_1)$, is an eigenvector of A^T.
- Question: is having eigendecomposition "typical" (or common) for $n \times n$ matrices? Ex. a matrix with random numbers in entries.

Orthogonality:

- Orthogonality describes some kind of independence relationship between objects. Often, we desire some independence, e.g. when creating a basis to express other objects (vectors, functions). Orthogonality provides a powerful means of interpreting (geometrically) equations in the form of $x \cdot y = 0$.
- Functional perspective: we can ask whether the function $f(x) = Ax$ changes the relationship of two vectors x and y. Ex. if x and y are orthogonal, will Ax and Ay be too? This leads to the idea of orthogonal matrix.
- QR factorization: geometrically, this is the transformation of parallelogram to rectangle.
- Application of orthogonalization in statistics: suppose we want to study the relationship between y and independent variables x_j's, we may want to create orthogonal variables from x_j's, and study how y is related to these variables (if they are not independent, we cannot learn their true effects on y).

Quadratic forms:

- Geometric interpretation of quadratic forms: in general, to study the property of a function $f(x)$, we can ask what is the solution (geometrically) of $f(x) = b$. Under this interpretation, the solutions of $x^T Ax = b$ form ellipse: whose axis is defined by the eigenvectors of A, and length defined by $\sqrt{\lambda_i}$.
- Standardization of quadratic forms and its geometric interpretation: Let $A = UDU^T$ be the eigen-decomposition of A, then the contour $x^T Ax = b$ can be viewed as $\sum_i \lambda_i x_i'^2 = b$, where λ_i is the eigenvalue and $x = Ux' = \sum_i x_i' u_i$. So we project x on u_i's, and the coordinates x_i''s form an ellipse with dimensions defined by λ_i'.
- Rayleigh quotient: it is defined as $\langle Ax, x\rangle / \langle x, x\rangle$, so it is the projection of Ax on x. Intuitively, this is the scaling effect of Ax, and we expect it should lie in between the smallest and large eigenvalues of A.

Chapter 3

Calculus & Analysis

3.1 Calculus

Derivative:

- Gradient: for a function $y = f(x)$, where $f : \mathbb{R}^n \to \mathbb{R}$, the derivative is called gradient ($\nabla f(x)$). It is a n-dim. row vector:
$$\frac{\partial y}{\partial \mathbf{x}} = \left(\frac{\partial y}{\partial x_1} \cdots \frac{\partial y}{\partial x_n} \right) \tag{3.1}$$
Note: a different convention is to have the gradient as a column vector.

- Derivative/Jacobian: a function $\mathbf{y} = f(\mathbf{x})$, where $f : \mathbb{R}^n \to \mathbb{R}^m$. The derivative of $\mathbf{y}$ wrt. $\mathbf{x}$ is an $m \times n$ matrix:
$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ & \cdots & \\ \frac{\partial y_m}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_n} \end{pmatrix} \tag{3.2}$$
The Jacobian determinant (often simply called the Jacobian) is the determinant of the Jacobian matrix (if $m = n$).

- Geometric interpretation of Jacobian: we consider a rectangular parallelepiped in the neighborhood of $x_0 \in \mathbb{R}^n$, defined by the length of edges: $\Delta x_1, \cdots, \Delta x_n$. Its image in $\mathbb{R}$ under f is a parallelepiped, where the j-th edge is defined by the vector:
$$\left(\frac{\partial f_1}{\partial x_j} \Delta x_j, \cdots, \frac{\partial f_m}{\partial x_j} \Delta x_j \right) = \left(\frac{\partial f_1}{\partial x_j}, \cdots, \frac{\partial f_m}{\partial x_j} \right) \Delta x_j \tag{3.3}$$
Thus the image is a parallelepiped in $\mathbb{R}^n$ with the directions of edge defined by the Jacobian matrix, and the volume of the original rectangular parallelepiped is scaled by $\det(J)$, where J is the Jacobian matrix of f at x_0.

Techniques for integrations:

- Auxiliary variable: suppose we want to calculate integral of $f(x, t)$ where t is an auxiliary variable. If we can write
$$f(x, t) = \frac{\partial}{\partial t} g(x, t) \tag{3.4}$$
and the integral of $g(x, t)$ is easy to find, then we have:
$$\frac{d}{dt} \int g(x, t) dx = \int \frac{\partial}{\partial t} g(x, t) = \int f(x, t) dx \tag{3.5}$$

Variable substitution for integrals:

- Theorem: let f be a continuous function in $\mathbb{R}^n$ with compact support, and $y = \phi(x)$ defined on a set S, where $x \in \mathbb{R}^n$. And let $T = \phi(S)$ be the image of S, then:

$$\int_T f(y)dy = \int_S f(\phi(x))J(x)dx \tag{3.6}$$

  where J is the Jacobian determinant of ϕ:

$$J(x) = \det\left(\frac{\partial \phi_j}{\partial x_i}\right) \tag{3.7}$$

- Proof by the finite sum approximation of integral: the integral in the LHS can be approximated by a sum of $f(y_i)\Delta y_i$, where Δy_i is the volume of the parallelepiped near y_i. Suppose y_i is the image of x_i, then by the geometrical interpretation of the Jacobian matrix, we have $\Delta y_i = J(x_i)\Delta x_i$, plug-in this to the integral and we have the equality.

- Proof by the Fundamental Theorem of Calculus: see [Change of variables in Multiple Integrals, AMM, 1999].

Representations of a hyperplane: [Bishop, Section 4.1] in general, two ways of representing a geometric object (set) are: (1) the equation that the points in the object must satisfy; (2) parameteric form: as the values of the parameters vary, all points in the object are covered. For a hyperplane in $\mathbb{R}^n$:

- Orthogonal vector: any point x in the hyperplane, $x \in \mathbb{R}^n$ satisfies the equation:

$$w^T x + b = 0 \tag{3.8}$$

  We show that w is orthogonal to the plane. In fact, if x_1, x_2 are in the hyperplane, then:

$$w^T x_1 + b = 0, w^T x_2 + b = 0 \Rightarrow w^T(x_2 - x_1) = 0 \tag{3.9}$$

  To find the distance (signed) of any point $x \in \mathbb{R}^n$ to the hyperplane: we define the projection of x on the plane $x_\perp$, and let r be the distance, then we have:

$$x = x_\perp + r\frac{w}{\|w\|} \tag{3.10}$$

  Multiply w^T and plus b in both sides, sovling r:

$$r = \frac{w^T x + b}{\|w\|} \tag{3.11}$$

- Basis vectors: let $v_1, v_2, \cdots, v_q$ be the basis vectors of a hyperplane, then the linear combination of these vectors, plus the location parameter describes all points in the plane:

$$f(\lambda) = \mu + \sum_{j=1}^{q} \lambda_j v_j \tag{3.12}$$

  where λ is the coordinates. In particular, a line can be represented as a function of $t \in \mathbb{R}$:

$$f(t) = x_0 + tv \tag{3.13}$$

  where x_0 is a point in the line and v is a vector parallel to the line.

Curves:

- Tangent of a curve: suppose we have a curve (a contour line of some function) in $\mathbb{R}^n$, $f(x) = c$, where c is a constant. Near the point x_0 at the curve, the function at x can be approximated by:

$$f(x) = f(x_0) + \nabla f(x_0)(x - x_0) \tag{3.14}$$

  But since both x and x_0 are in the curve, i.e. $f(x) = f(x_0) = c$, the tangent can be represented by the line (hyperplane):

$$\nabla f(x_0)(x - x_0) = 0 \tag{3.15}$$

  In other words, the gradient is perpendicular to the tangent of the curve (contour line).

Gaussian integrals: [A Brief Look at Gaussian Integrals, Straub]

- Simple Gaussian integral:

$$I = \int_{-\infty}^{\infty} \exp(-\frac{1}{2}ax^2)dx = \sqrt{\frac{2\pi}{a}} \tag{3.16}$$

  Proof: first we solve the case for $a = 1$ by polar coordinate transform; and variable subsitution for ax^2.

- Gaussian integral with power term:

$$\int_{-\infty}^{\infty} x^{2n} \exp(-\frac{1}{2}ax^2)dx = \frac{(2n)!}{a^n 2^n n!}\sqrt{\frac{2\pi}{a}} \tag{3.17}$$

  Note that when the power term is odd, the integral is 0 because of symmetry.
  Proof: take derivative of I wrt. a on both sides of the Equation 3.16, we have:

$$\frac{dI}{da} = -\frac{1}{2}\int_{-\infty}^{\infty} x^2 \exp(-\frac{1}{2}ax^2)dx = -\frac{1}{2}\sqrt{2\pi}a^{-3/2} \tag{3.18}$$

  We could do repeated differentiation to get the Equation above.

- Gaussian integral with linear term:

$$\int_{-\infty}^{\infty} \exp(-\frac{1}{2}ax^2 + Jx)dx = \exp\left(\frac{J^2}{2a}\right)\sqrt{\frac{2\pi}{a}} \tag{3.19}$$

  Proof: complete-the-square.

- Multivariate Gaussian integral: let A be a symmetric $n \times n$ matrix, and x is n-dim. column vector. Denote $|A|$ as the determinant of A, then

$$\int_{-\infty}^{\infty} \exp(-\frac{1}{2}x^T Ax)d^n x = \frac{(2\pi)^{n/2}}{|A|^{1/2}} \tag{3.20}$$

  Proof: diagonalization of A:

$$A = SDS^T \tag{3.21}$$

  where D is the diagnoal matrix with eigenvalues, λ_i, as diagonal terms, and S be orthonomal matrix ($\det S = 1$). Let the ingeral be I, and we do variable subsitution $x = Sy$. Note that $dx = dy$ since $\det S = 1$.

$$I = \int_{-\infty}^{\infty} \exp(-\frac{1}{2}y^T Dy)d^n y = \prod_{i=1}^{n}\int_{-\infty}^{\infty} \exp(-\frac{1}{2}\lambda_i y_i^2)dy_i = \prod_{i=1}^{n}\sqrt{\frac{2\pi}{\lambda_i}} = \frac{(2\pi)^{n/2}}{|A|^{1/2}} \tag{3.22}$$

- Multivariate Gaussian integral with linear term:

$$\int_{-\infty}^{\infty} \exp(-\frac{1}{2}x^T Ax + J^T x)d^n x = \frac{(2\pi)^{n/2}}{|A|^{1/2}} \exp\left(\frac{1}{2}J^T A^{-1} J\right) \tag{3.23}$$

  Proof: plug in $x = Sy$, and use the complete-the-square trick for quadratic forms:

$$-\frac{1}{2}x^T Ax + J^T x = -\frac{1}{2}(y - S^T A^{-1})^T D(y - S^T A^{-1}) + \frac{1}{2}J^T A^{-1} J \tag{3.24}$$

3.1.1 Calculus of Field

Vectors:

- Cross product: given a vector $\vec{u}$ and $\vec{v}$, the cross product:

$$\vec{u} \times \vec{u} = \vec{n}\|u\|\|v\| \sin\theta \tag{3.25}$$

  where θ is the angle between u and v, and $vecn$ the direction orthogonal to the plane of $\vec{u}$ and $\vec{v}$. Thus the cross-product represent the area of the paralleloid formed by $\vec{u}$ and $\vec{v}$.

Divergence Theorem:

- Divergence: suppose $\mathbf{F}$ is a vector field, the divergence of $\mathbf{F}$ at a point p is defined as: the limit of the net flow across the smooth boundary of a three dimensional region V divided by the volume of V as V shrinks to p:

$$\nabla \cdot \mathbf{F}(p) = \lim_{V \to p} \int \int_{S(V)} \frac{\mathbf{F} \cdot \mathbf{n}}{|V|} dS \tag{3.26}$$

  where $S(V)$ is the surface of V, and the integral is the surface integral over $S(V)$. In the Cartesian coordinate, suppose $\mathbf{F} = U\mathbf{i} + V\mathbf{j} + W\mathbf{k}$, then:

$$\nabla \cdot \mathbf{F} = \frac{\partial U}{\partial x} + \frac{\partial V}{\partial y} + \frac{\partial W}{\partial k} \tag{3.27}$$

- Divergence Theorem: suppose $\mathbf{F}$ is a continuous-differentiable vector field, we have:

$$\int_V \nabla \cdot \mathbf{F} dV = \int_S \mathbf{F} \cdot \mathbf{n} dS \tag{3.28}$$

  The intuition is that: to calculate fluid flows out of a certain region, we need to add up the sources inside the region and subtract the sinks. The divergence at a given point describes the strength of the source or sink there (net flow at the point). Thus Divergence Theorem is essentially a conservation law.

- Gauss's law in electrostatics: let $\mathbf{F}$ be the electric field, then the electric flux across a surface is equal to (up to a constant) the electric charge within the area bounded by the surface. Written in the differential form:

$$\nabla \cdot \mathbf{F} = \frac{\rho}{\epsilon_0} \tag{3.29}$$

  where ρ is the total electric charge density.

Laplace operator:

- Laplace operator: suppose f is a twice-differentiable function in $\mathbb{R}^n$, the Laplacian operator of f is defined as the divergence of the gradient of f:

$$\Delta f = \nabla \cdot \nabla f = \nabla^2 f \tag{3.30}$$

  In the Cartesian coordinates, we could write it as:

$$\Delta f = \sum_i \frac{\partial^2 f}{\partial x_i^2} \tag{3.31}$$

- Application in Electrostatics: suppose $\mathbf{E}$ is the electric field associated with a charge density q, and ϕ is the electrostatic potential. By the Divergence Theorem:

$$\int_S (V)\mathbf{E}\cdot\mathbf{n} = \int_S (V)\mathbf{E}\cdot\mathbf{n} = \int_V \Delta\phi dV = \int_V q dV \tag{3.32}$$

  We could write this as:

$$\Delta\phi = q \tag{3.33}$$

  Thus the potential function is given by the Poisson's Equation.

- Application to diffusion at equilibrium: suppose u is some density at equilibrium, then the net flux of u through a boundary of a small region is 0 at equilibrium, thus:

$$\Delta u = 0 \tag{3.34}$$

  This is the Laplace's Equation.

- Energy minimization [Wiki]: the Dirichlet energy of a function f on a region U is defined as:

$$E(f) = \frac{1}{2}\int_U \|\nabla f\|^2 dV \tag{3.35}$$

  The function f that minimizes $E(f)$ is given by $\Delta f = 0$.

3.1.2 Vector and Matrix Calculus

Reference: [Matrix-calculs-R.pdf, from Google], [Matrix calculus and MLE for the multivariate Normal, Berkeley CS 281A notes]

Definition of derivatives and chain rule:

- Definition: we follow the convention of Jacobian matrix. Let $y = f(x)$ be m-dimensional function of the n-dimensional vector x. Both x and y are represented as column vectors. The derivative of y over x is $m \times n$ matrix:

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ & \cdots & \\ \frac{\partial y_m}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_n} \end{pmatrix} \tag{3.36}$$

  In particular, if y is scalar, then the derivative is the gradient (n-dim. row vector): $\nabla f = (\frac{\partial y}{\partial x_1}, \cdots, \frac{\partial y}{\partial x_n})$.

- Chain rule: let $\mathbf{x}$ be a vector of $\mathbb{R}^n$, $\mathbf{y}$ be a r-dim function of $\mathbf{x}$, and $\mathbf{z}$ be an m-dim function of $\mathbf{y}$, then the derivative of $\mathbf{z}$ wrt $\mathbf{x}$ is an $m \times n$ matrix given by:

$$\frac{\partial \mathbf{z}}{\partial \mathbf{x}} = \frac{\partial \mathbf{z}}{\partial \mathbf{y}}\frac{\partial \mathbf{y}}{\partial \mathbf{x}} \tag{3.37}$$

- Remark: some books define the derives as the transpose of Jacobian defined above (e.g. "Matrix-calculs-R.pdf").

Derivatives for some important functions:

- $y = \mathbf{a}^T\mathbf{x}$: $\mathbf{a}$ is an n-dim column vector, and y is a scalar, we have:

$$\frac{\partial(\mathbf{a}^T\mathbf{x})}{\partial \mathbf{x}} = \mathbf{a}^T \tag{3.38}$$

- $y = A\mathbf{x}$: A is an $m \times n$ matrix, we have:
$$\frac{\partial(A\mathbf{x})}{\partial \mathbf{x}} = A \tag{3.39}$$
Proof: expand A as row vectors, and apply the previous result.

- $y = \mathbf{x}^T A\mathbf{x}$: A is $n \times n$ square matrix, y is a scalar function (quadratic form):
$$\frac{\partial(\mathbf{x}^T A\mathbf{x})}{\partial \mathbf{x}} = \mathbf{x}^T(A + A^T) \tag{3.40}$$
Proof: follow the expansion of the quadratic form, or use the product rule.

Matrix derivatives:

- Theorem: for two matrices A and B,
$$\frac{\partial \mathrm{tr}\,(AB)}{\partial A} = B^T \qquad \frac{\partial \mathrm{tr}\,(AB)}{\partial B} = A^T \tag{3.41}$$
Proof: we have $\mathrm{tr}\,(AB) = \sum_{i,j} A_{ij}B_{ji}$, thus,
$$\frac{\partial \mathrm{tr}\,(AB)}{\partial A_{ij}} = B_{ji} \tag{3.42}$$
The latter follows from the fact that $\mathrm{tr}\,(AB) = \mathrm{tr}\,(BA)$.

- Theorem: A is a square matrix with positive determinant, then:
$$\frac{\partial \log\det(A)}{\partial A} = (A^{-1})^T \tag{3.43}$$
Proof: use the Laplace expansion of the determinant, we have:
$$\frac{\partial \det(A)}{\partial A} = \mathrm{adj}(A)^T \tag{3.44}$$
where $\mathrm{adj}(A)$ is the adjugate matrix. Plugin the relation: $A^{-1} = 1/\det(A) \cdot \mathrm{adj}(A)$, and apply the chain rule:
$$\frac{\partial \log\det(A)}{\partial A} = \frac{1}{\det(A)} \cdot \frac{\partial \det(A)}{\partial A} = \frac{1}{\det(A)}\mathrm{adj}(A)^T = (A^{-1})^T \tag{3.45}$$

3.2 Real and Functional Analysis

Generating functions [Marsden, Elementary Classical Analysis]:

- Idea: the information contained in a sequence can be represented as a function, e.g. series, and then the properties of the original sequence can be studied via the new function. Most commonly, the sequence is expressed in a formal power series.

- An important property of the generating function: if the sequences $\{a_n\}$ and $\{b_n\}$ are represented by the function $f(x)$ and $g(x)$ respectively, then:
$$f(x)g(x) = \left(\sum_i a_i x^i\right)\left(\sum_j b_j x^j\right) = \sum_k \left(\sum_{i+j=k} a_i b_j\right) x^k \tag{3.46}$$

- Probability generating function (PGF): for discrete probability distributions, the information may be represented as a sequence, i.e. $p_k = P(X = k)$, thus the RV can be represented as a power series:

$$G(z) = \mathrm{E}\,(z^X) = \sum_k p_k z^k \tag{3.47}$$

Basis function and representation in a different domain [Marsden, Elementary Classical Analysis]:

- Idea: a function may be represented in terms of a set of basis functions, then the function can be written in a different form using the basis function. If the basis functions are not continuous, the representation is a vector (finite dimension); if the basis functions are continusous, the representations form a new function (transform).
- Fourier transform: a function (periodic) may be written as a sum of multiple (basis) periodic function, say, $e^{ik\omega x}$, this forms the discrete Fourier series. When ω is continuous, we have the Fourier transform:

$$\hat{f}(\omega) = \int_{-\infty}^{+\infty} f(x) e^{i\omega x} dx \tag{3.48}$$

  The physical meaning of this function at ω is the coefficient of f at ω, if writting f as a sum of basis functions.

- Other examples:
  - A multi-modal p.d.f. may be represented as a sum of normal distributions with different location parameters.
  - A electrical field (3D function) may be represented as a sum of fields with source charge placed at different spatial points.
- Remark: a function is represented in different domains. Ex. for a function in the time domain, its Fourier transform is a new function in the frequency domain.
  - This is different from the idea of generating functions where a new function is formed that represent different things, while the idea of transform represents the same function in different forms.

Convergence of sequence of functions [Marsden, Elementary Classical Analysis]:

- Motivation: we may want to define or approximate a function via an infinite sequence or series, thus need a definition of convergence for functions.
- Pointwise convergence: a sequence of functions $\{f_n\}$ converges to a function f pointwise if and only if:

$$\lim_{n\to\infty} f_n(x) = f(x) \tag{3.49}$$

  for every x in the domain.

- Uniform convergence: a sequence of functions $\{f_n\}$ converges uniformly to a function f if:

$$\lim_{n\to\infty} \sup\{|f_n(x) - f(x)|\} = 0 \tag{3.50}$$

  An alternatie defintion is $f_n(x)$ and $f(x)$ can be arbitrarily close, using the ϵ-N language: for every $\epsilon > 0$, there exists N, s.t. for any $n > N$ and $x \in S$, we have $|f_n(x) - f(x)| < \epsilon$. Geometrically, uniform convergence means f_n all lies in the ribbon of size ϵ around the function f.

- Why we need two definitions? Uniform convergence (stronger) implies pointwise convergence, but not the other way. Ex. the sequence of function x^n defined in $[0, 1)$ converges pointwise to 0, but not uniformly.

Fourier analysis [An introduction to wavelets, Amara Graps; personal notes]

- Motivation: electromagnetic signals consisting of waves with different frequencies. Understand the contribution of each frequency.
- Idea of **Analyzing function**: to investigate a function, we probe the function by using another function, in this case, the similarity of the function with the analyzing function. The similarity here is the inner product of functions, which can be interpreted as correlation if we have evenly-spaced sampled sequences from the two functions.
- Intuition of approximating a periodic function with sin and cos: suppose we consider x near 0, $\sin(x)$ approximates an increasing function, while $\cos(x)$ decreasing. Suppose f is increasing, we use $\sin(kx)$ to approximate, where k matches the slope of f. We can then consider the residual, and repeat this process.
- Relation to regression: (1) similar to iterative regression in statistics. Here, instead of regression over covariates, we regress over basis functions. We can write down this process in an explicit form: $f(x)$ as the sum of $sin(kx)$ and $cos(kx)$. (2) Generalized Additive Model in regression analysis. (3) A special case where covariates are independent: orthogonal basis functions.
- **Theorem: Fourier Series**. Any 2π-periodic function $f(x)$ is the sum of:

$$f(x) = a_0 + \sum_{k=1}^{\infty}(a_k \cos kx + b_k \sin kx) \tag{3.51}$$

  where the coefficients are given by:

$$a_0 = \frac{1}{2\pi}\int_0^{2\pi} f(x)dx \qquad a_k = \frac{1}{\pi}\int_0^{2\pi} f(x)\cos(kx)dx \qquad b_k = \frac{1}{\pi}\int_0^{2\pi} f(x)\sin(kx)dx \tag{3.52}$$

  The interpretation is that: the time domain signal $f(x)$ can be considered in the frequency domain, defined by the Fourier coefficients a_k and b_k (which represent the contribution of each frequency).
- Orthogonal basis function and proof of Fourier series: if we define the inner product of two functions as:

$$\langle f, g\rangle = \int f(x)g(x)dx, \tag{3.53}$$

  we can show that $\cos kx$ and $\sin kx$ are orthogonal to each other. Then the coefficients a_k and b_k are coordinates of f with the orthogonal basis.
- Extension to the continuous case: Fourier transform of function f is given by:

$$\hat{f}(\omega) = \int_{-\infty}^{+\infty} f(x)e^{i\omega x}dx \tag{3.54}$$

  here ω is similar to k, and represents "frequency", and $\hat{f}(\omega)$ is similar to a_k and b_k. The inverse Fourier transform is just $f(x)$.
- Discrete Fourier Transform (DFT) [Discrete Fourier Transform - Simple Step by Step, Youtube]: suppose we do not have actual f, but rather data points from function f, let them be $x_1, \cdots, x_N$. We will approximate the integral in Fourier coefficients using summation:

$$X_k = \sum_n x_n \cdot e^{-2\pi ikn/N} \tag{3.55}$$

  The DFT can be written as to is the multiplication of $\{x_n\}$ by a matrix to obtain $\{X_k\}$. This can be done efficiently via the FFT algorithm.

- Windowed Fourier Transform (WFT): when the function is not periodic, we can chop function into windows, and apply FT in each window.
- Applications of Fourier analysis: sound editing, removing high frequency pitch.

But what is the Fourier Transform? A visual introduction [YouTube]

- Motivation: demixing sound. Consider pure frequency sound waves: mixing of a few different frequencies can lead to complex signal.
- Physical intuitions: **Winding graph**. Suppose we have a wave with a given frequency, consider the vector of its amplitude, it oscillates over time. Imagine we rotate the vector in 2D with a certain period, and we consider the "center of mass" of the vector. The location of the center thus becomes a function of the rotating frequency. When the frequency is different from the signal frequency, the center is close to 0; and it is large only when the two frequencies match. This simple transformation allows us to consider the signal in the frequency domain: at this domain, mixing signals becomes simple.
- Translating the intuition: the rotating vector can be expressed as $e^{-2\pi ift}$ (where f is the rotating frqeuency), and the center of mass as $\int g(t)e^{-2\pi it}dt$, where $g(t)$ is the signal. This gives the Fourier coefficients. Intuitively, we try all possible frequencies in the winding graph, and see which frequencies show up.

Introduction to wavelet analysis [Understanding Wavelets, Youtube; Easy Introduction to Wavelets, Youtube; An introduction to wavelets, Amara Graps]

- Motivation: signals at different scales. Ex. a smoothing varying (low frequency) background, with sharp (short) signal, and noises. Fourier (sin and cos) are not good at approximating localized signals.
- Idea of wavelet analysis: use more general basis function s.t. we better capture signals at multiple scales, e.g. short spikes. The main difference with Fourier analysis is: the wavelet functions (basis) are localized in space.
- Wavelet basis functions: requirements are compact support (localized), integral is 0 and orthogonal. Many wavelet families are distinguished by: vanishing moments, where moment of k is defined as $\int f(x)x^k dx$. High vanishing moments: cannot approximate simple functions (e.g. constant). A related concept is: regularity, which captures smoothness of function.
- How wavelet analysis works? [video: sliding wavelets matching the original signal] A family of wavelet function is defined by **scale** and **shift** (translation): at a particular scale, a wavelet captures a local signal at a shift (time); and the fact that we have many scales allow us to capture signals at different levels: e.g. a low frequency varying background as well as a high frequency (short scale) signal at a particular time. The wavelet coefficient of scale s and location l:

$$X_{s,l} = \int_{-\infty}^{+\infty} x(t)\psi_{s,l}(t)dt \tag{3.56}$$

  where $x(t)$ is the original signal. The coefficients tell the amplitude at each scale and shift (time).
- Continuous wavelet transform (CWT) and discrete wavelet transform (DWT).
- Denoising data by wavelet: DWT, thresholding (removing/shrinking small coefficient), and inverse DWT. Example: earthquake signal analysis. High frequency (very short scale), captures "white noise". Intermediate frequency: true earthquake signal.

- Harr wavelet [Lorenzo Sadun, Youtube]: Level 0: $h_0(t) = 1$ is a constant. Level 1: $h_1(t)$ is step wise function +1 and -1 in [0,1]. Level 2: $h_2(t)$ and $h_3(t)$ are the same function (+1 or -1) with interval size 1/2 (the functions have non-zero values only in [0,1] or [-1,0]). Level 3: $h_4(t)$ to $h_7(t)$, the same function with interval size 1/4 (functions have non-zero values only in [-1, -1/2], [-1/2,0], [0, 1/2] or [1/2,1]). It's easy to check the functions are orthogonal as: at the same scale, $h_i(t)$ and $h_j(t)$ are defined on different intervales; at different scales, the integral is 0. The wavelet coefficients are:

$$C_n = \int_0^1 f(t)h_n(t)dt \tag{3.57}$$

  C_0: mean of the signal; C_1: the difference of the first and second halves; and so on.

3.3 Numerical Methods

Reference: [Heath, Scientific Computing: an Introductory Survey]
Background:

- Contraction Mapping Theorem (Banach Fixed-point Theorem): let f be a function defined on the metric space of M to itself. f is a contraction if there exists some $k < 1$ s.t. for all $x, y \in M$:

$$d(f(x), f(y)) \leq kd(x, y) \tag{3.58}$$

  Then the mapping f has one and only one fixed point, x^*, i.e. $f(x^*) = x^*$. The proof follows from the fact that the distance between any two consective x_n and x_{n+1} is an exponentially decreasing sequence:

$$d(x_n, x_{n+1}) \leq k^n d(x_0, x_1) \tag{3.59}$$

- Fixed-point iteration for solving nonlinear equations [Heath, Scientific Computing: an Introductory Survey, Chapter 5]: a general algorithm is to write the equation $f(x) = 0$ (for optimization problem, it is derivative) as $x = g(x)$, and then define the iteration:

$$x_{n+1} = g(x_n) \tag{3.60}$$

  If g is a contraction, then the algorithm will converge to a unique x^*. However, note that not all such functions are contractions.

System of nonlinear equations [Heath, Chapter 5]:

- Fixed point iteration: consider $g : \mathbb{R}^n \to \mathbb{R}^n$, let $G(x)$ be the Jacobian of g. If the spectral radius (largest eigenvalue) of $G(x) < 1$ near x^*, i.e. $\rho(G) < 1$, then the fixed point iteration $x_{k+1} = g(x_k)$ converges to the fixed point $x^* = g(x^*)$.
  Proof: near x^*, we have

$$g(x_k) \approx g(x^*) + G(x^*)(x_k - x^*) \tag{3.61}$$

  Let $e_k = x_k - x^*$, and plug in $x_{k+1} = g(x_k)$, we have this relation for the error:

$$e_{k+1} \approx G(x^*)e_k \tag{3.62}$$

  To show that it converges, we consider the norm of error:

$$\|e_{k+1}\|^2 \approx \|G(x^*)e_k\|^2 = e_k^T G(x^*)^T G(x^*)e_k \leq \rho^2 \|e_k\|^2 \tag{3.63}$$

  where ρ is the spectral radius of $G(x^*)$. The last step is based on: (1) Rayleign quotient, and (2) the square of the largest eigenvalue of G is also the largest eigenvalue of G^2 (G is symmetric). When $\rho < 1$, we have $\|e_k\|$ decreases geometrically.

- Newton's method: we approximate the function $f(x)$ near x as a linear function and solve it:

$$f(x+s) \approx f(x) + J_f(x)s \tag{3.64}$$

where $J_f(x)$ is the Jacobian of f. We view this as a linear function of s, and the solution is given by the linear system $J_f(x)s = -f(x)$. The solution in terms of x is thus given by $x + s$. Our iteration rule is:

$$x_{k+1} = x_k + s_k, \text{ where } J_f(x_k)s_k = -f(x_k) \tag{3.65}$$

We can show that the Jacobian of the fixed point form is equal to 0. The fixed point equation $g(x) = x - J_f(x)^{-1}f(x)$, and its Jacobian at x^*:

$$G(x^*) = I - J_f(x^*)^{-1}J_f(x^*) + \sum_{i=1}^{n} f_i(x^*)H_i(x^*) = 0 \tag{3.66}$$

So Newton's method has quadratic convergence. The algorithm requires $O(n^2)$ computation for Jacobian, and $O(n^3)$ for solving the linear system at each step.

- Quasi-Newton methods: approximate Jacobian with finite difference like method. Secant Updating Methods.

- Robust Newton-like methods: the challenge of Newton's method is that it is guaranteed to converge, so we can use a *damped Newton method*:

$$x_{k+1} = x_k + \alpha_k s_k \tag{3.67}$$

where $0 < \alpha_k \leq 1$. At the beginning $\alpha < 1$, so we make smaller step, and when it is close to the optimum, we make $\alpha = 1$. We can monitor the change of $\|f(x_k)\|$.

3.4 Optimization

Questions of optimization:

- The proof of strong duality from Slater's condition.
- The proof that KTT is necessary and sufficient condition for primal and dual optimality if the Slater's condition is satisified.

Theory of unconstrained optimization [Heath, Chapter 6]

- Gradient: consider $f : \mathbb{R}^n \to \mathbb{R}$, near any point x, we use Taylor expansion:

$$f(x+s) \approx f(x) + \nabla f(x) \cdot s \tag{3.68}$$

Suppose the norm of s is fixed, then using Cauchy-Schwarz Inequality, we know that $\nabla f(x) \cdot s$ is maximized at $s \propto \nabla f(x)$ and minimized at $s \propto -\nabla f(x)$. So **the negative gradient point to the direction of steepest descent**. In particular, in this direction, $f(x+s) \leq f(x)$, so the function f decreases.

- Necessary condition of optimum: when x^* is a minimum, the function value cannot decrease, so we have:

$$\nabla f(x^*) = 0 \tag{3.69}$$

Such x^* is called a *critical point* of f (also called stationary point). This condition is called **first order necessary condition** for optimality.

- Sufficient condition of optimum: let x^* be a critical point of f, for $s \in \mathbb{R}^n$, we have:

$$f(x^* + s) - f(x^*) \approx \nabla f(x^*)^T s + \frac{1}{2} s^T H_f(x^*) s = \frac{1}{2} s^T H_f(x^*) s \tag{3.70}$$

  Thus when $H_f(x^*)$ is positive definite, this term is positive when $s \neq 0$, so x^* is a local minimum of f. This leads to **second order sufficient condition** for optimality:

  - If $H_f(x^*)$ is positive definite, then x^* is a local minimum.
  - If $H_f(x^*)$ is negative definite, then x^* is a local maximum.
  - If $H_f(x^*)$ is indefinite, then x^* is a saddle point.

  To check if a symmatric matrix is positive definite, we can use Cholesky factorization.

- Sensitivity: we only analyze the case of 1D. Suppose x^* is the true optimum, and our solution is $\hat{x}$. The condition number of the optimization problem is given by how much x changes in response to change of function values. Let $h = \hat{x} - x^*$, we use Taylor expansion:

$$f(\hat{x}) \approx f(x^*) + f'(x^*)h + \frac{1}{2} f''(x^*)h^2 \tag{3.71}$$

  So if our error of y is $|f(\hat{x})| - f(x^*) \leq \epsilon$, our error of x is $h \approx \sqrt{2\epsilon / |f''(x^*)|}$. This means generally our error is much larger than our error of function values. This is not surprising, as $f(x^*) = 0$, so the function value is not sensitive to change of x. If we can directly solve $f'(x) = 0$, then we have better accuracy as the derivative of $f'(x)$ is generally not 0.

Methods for 1D optimization:

- Golden section method: we assume that f is unimodal. Suppose we want to find minimum in an interval $[a, b]$, we note first that we cannot use bisection search method. Instead, suppose $x_1, x_2 \in [a, b]$ with $x_1 < x_2$, if $f(x_1) < f(x_2)$, then the minimum cannot lie in $(x_2, b]$ and if $f(x_1) > f(x_2)$, the minimum cannot lie in $[a, x_1)$. This would allow us to have a triplet that must contain the minimum: either (x_1, x_2, b) or (a, x_1, x_2). We will need to then choose an interior point in each case to shrink the triplet. Based on Algorithm 6.1 in [Heath, 6.4] (the top case), we choose the location (parameterized by τ) as:

$$\frac{b - x_2}{b - x_1} = \frac{1 - \tau}{\tau} = \tau \tag{3.72}$$

  This leads to $\tau^2 = 1 - \tau$.

- Newton's method for 1D optimization [Heath, Chapter 6]: we minimize function near x^* using quadratic approximation of f:

$$f(x + h) \approx f(x) + f'(x)h + \frac{1}{2} f''(x)h^2 \tag{3.73}$$

  The minimum of this function is given by $h = -f'(x)/f''(x)$. This leads to the recurrence:

$$x_{k+1} = x_k - f'(x_k)/f''(x_k) \tag{3.74}$$

  Note that this is equivalent to solving the equation $f'(x) = 0$ using Newton's method. The convergence rate is again quadratic.

Methods for unconstrained optimization [Heath, Scientific computing: an introductory Survery, 2ed, Chapter 6]

- Univariate search [Chapra & Canale, Chapter 14]: to minimize $f(x, y)$, where x and y are two sets of parameters. Iteratively minimize f when x is fixed; and then minimize f when y is fixed. Also called conditional minimization. It is particulary useful in statistics where conditional distributions are often easier to obtain.

- Nelder-Mead simplex method: the function needs to be unimodal. The idea of the method: choose $n+1$ point Simplex. At each step, shrink the simplex towards optimum. Intuitively, at 1D, if $f(x_1) > f(x_2)$, we should shrink x_1. At 2D, we move from the worst point, in the line towards the centroid. The method works in low dim ($n \leq 3$).

- Steepest descent: at any step where the value of x is x_k, we choose the direction: $s_k = -\nabla f(x_k)$, and define the 1D function:

$$\phi(\alpha) = f(x_k + \alpha_k s_k) \tag{3.75}$$

  Find α_k that minimizes the above function (line search) and set:

$$x_{k+1} = x_k + \alpha_k s_k \tag{3.76}$$

  At the new point, the gradient is orthogonal to s (i.e. the line is tangent with the contour of f). Proof: $\phi'(\alpha) = 0$ implies that $\nabla f(x + \alpha s) \cdot s = 0$. The behavior of Steepest descent: zigzag towards the minimum, however, it does not have any global view and converge rather slowly.

- Newton's method: near x^*, the function can be approximated by the quadratic function:

$$f(x + s) \approx f(x) + \nabla f(x)^T s + \frac{1}{2} s^T H_f(x) s \tag{3.77}$$

  where $H_f(x)$ is the Hessian matrix (the second partial derivative) of f at x. Minimizing the quadratic function of s (using complete the square or vector calculus):

$$H_f(x)s + \nabla f(x) = 0 \tag{3.78}$$

  This gives the recurrence:

$$x_{k+1} = x_k + s_k, \text{ where } H_f(x_k)s_k = -\nabla f(x_k) \tag{3.79}$$

  It is easy to see that this is equivalent to solving the nonlinear system $\nabla f(x) = 0$. The convergence of Newton's method follows from Contraction Mapping Theorem. For 1D case, Newton's method is a fixed point iteration with the function g defined as:

$$g(x) = x - f(x)/f'(x) \tag{3.80}$$

  Near the point $f'(x^*) = 0$, it is easy to show that $g'(x^*) = 0$, so g is a contraction near x^*.

  - Direction of curvature: if the Hessian matrix is not positive definite, the direction may not be the direction of descent.
  - Solving the linear equation: directly solving the matrix inverse is expensive. In practice, this is done via solving a linear equation; furthermore, if H is positive definite, could use Cholesky decomposition to solve the linear equation.

- How to address the convergence problem? (1) Damped Newton $x_{k+1} = x_k + \alpha_k s_k$ with $0 < \alpha_k \leq 1$. (2) Trust region. (3) test if $H_f(x)$ is positive definite.

- Quasi-Newton method: the motivation is that in Newtons' method, we need $O(n^2)$ computation for evaluating Hessian and $O(n^3)$ for solving the linear system. The general idea to save the computation cost is to avoid computing Hessian, and the update rule requires $O(n^2)$ to solve. One of the most popular one is Secant updating method (BFGS): approximate Hessian at each iteration.

Theory for constrained optimization [Heath, Chapter 6]

- Background: tangent of a contour/curve. Suppose we have a curve or contour represented as $f(x) = c$. Let s be a direction along the tangent line, then $f(x+s) = f(x)$ by the definition of tangent (function value does not change). But $f(x+s) = f(x) + \nabla f(x) \cdot s$, so we have:

$$\nabla f(x) \cdot s = 0 \tag{3.81}$$

This means that the gradient is orthogonal to the tangent line.

- Lagrange multipliers method for one equality constraint: suppose we have the optimization problem for $x \in \mathbb{R}^n$:

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & g(x) = 0 \end{array} \tag{3.82}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}$ (i.e. we have only one constraint). Thus we are minimizing f over the curve $g(x) = 0$. Suppose s is a direction along the tangent line of $g(x) = 0$, and we consider the point x^*, then

$$\nabla g(x^*)s = 0 \tag{3.83}$$

On the other hand, since x^* is a local minimum, its function value should not change, by the Taylor expansion, we have:

$$\nabla f(x^*)s = 0 \tag{3.84}$$

This implies that: $\nabla f \parallel \nabla g$, or

$$\nabla f(x^*) + \lambda \nabla g(x^*) = 0 \tag{3.85}$$

for some λ. We define the Lagrangian function $L : \mathbb{R}^{n+1} \to \mathbb{R}$:

$$L(x, \lambda) = f(x) + \lambda g(x) \tag{3.86}$$

The condition that x^* satisfies can be written as:

$$\frac{\partial}{\partial x} L(x, \lambda) = \nabla f(x) + \nabla g(x) = 0 \tag{3.87}$$

$$\frac{\partial}{\partial \lambda} L(x, \lambda) = g(x) = 0 \tag{3.88}$$

which is exactly what we have shown above: parallel gradient and constraint. This is the **first order necessary condition for constrained optimum**.

- Lagrange multiplier method for multiple equality constraints: our problem is to:

$$\min_x f(x) \qquad \text{subject to } g(x) = 0 \tag{3.89}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}^m$ (i.e. we have m constraints). The Lagrangian function is defined as: $L : \mathbb{R}^{n+m} \to \mathbb{R}$:

$$L(x, \lambda) = f(x) + \lambda^T g(x) \tag{3.90}$$

where $\lambda \in \mathbb{R}^m$. The first order necessary condition for optimality is:

$$\nabla f(x) + J_g^T(x)\lambda = 0 \qquad g(x) = 0 \tag{3.91}$$

To obtain the second order sufficient condition, we note that:

$$H_L(x, \lambda) = \begin{bmatrix} B(x, \lambda) & J_g^T(x) \\ J_g(x) & O \end{bmatrix} \tag{3.92}$$

where

$$B(x, \lambda) = \nabla_{xx} L(x, \lambda) = H_f(x) + \sum_{i=1}^{m} \lambda_i H_{g_i}(x) \tag{3.93}$$

In general, the matrix $H_L(x, \lambda)$ is symmetric but not positive definite. A sufficient condition for x^* to be the minimum is that $B(x^*, \lambda^*)$ is positive definite on the tangent space to the constraint surface. Specifically, let Z be a matrix whose columns form a basis for the tangent subspace, if $Z^T BZ$ is positive definite, then x^* is a local minimum.

  - Remark: consider the case of one constraint, we have $B(x, \lambda) = H_f(x) + \lambda H_g(x)$. We consider $f(x + s)$ along the direction of tangent line to $g(x) = 0$, denoted as s. We have $\nabla g(x) \cdot s = 0$. We only need to show that $s^T H_f(x) s \geq 0$, for s along the tangent direction (instead of considering all x's).

- Remark: a geometric way of constrained optimization, draw feasibility set, and see how the contour line of the (unconstrained) objective function intersects with the feasiblity set.

Methods for constrained optimization [Heath, Chapter 6]

- Sequential quadratic programming: we consider only equality constraint, and we solve x^* at critical point using Equation 3.91. This is done via Newton's method. This is equivalent to minimizing a quadratic function (in terms of s) satisfying the constraint that $J_g(x)s + g(x) = 0$.

Techniques for searching in constrained space [Bussemaker & Siggia, ISBM, 2000; Bauer & Baily, Bioinformatics, 2009]:

- Problem: the variables of the objective function may be constrained (within an interval), however, many standard procedures (e.g. Gradient Descent) only search in the unconstrained fashion.

- Variable transformation: suppose we want to minimize $f(x)$, where $x \in [a, b]$, we could transform x to z s.t. $z \in (-\infty, +\infty)$. First map $[a, b]$ to $[0, 1]$:

$$y = \frac{x - a}{b - a} \tag{3.94}$$

  Then map $[0, 1]$ to $(-\infty, +\infty)$:

$$z = \ln \frac{y}{1 - y} \tag{3.95}$$

  To obtain the values of the variables in the original space, do inverse transformation:

$$y = \frac{e^z}{1 + e^z} \tag{3.96}$$

$$x = a + y(b - a) \tag{3.97}$$

3.4.1 Convex Optimization

Reference: [CMU 10-725, Optimization], [Boyd & Vandenberghe, Convex Optimization].

Application of convex optimization in machine learning:

- SVM: maximum margin, can be solved by quadratic programming.

- Probabilistic graphical model: variational approximation, solved by convex optimization.

- Regression with regularization: least square

- Maximum Entropy: suppose we want to estimate the unknown distribution (p) of a RV X, we know the expectation of some functions defined on X, f_i. Then we are solving the optimization problem:

$$\max_p H(p) \qquad \text{s.t. } \mathrm{E}[f_i] = \theta_i \tag{3.98}$$

- Clustering and other unsupervised learning problems.

Classes of optimization problems:

- Least square problem: find $x \in \mathbf{R}^n$ that minimize:

$$f(x) = \|Ax - b\|^2 = \sum_{i=1}^{k} (a_i^T x - b_i)^2 \tag{3.99}$$

  where $A \in \mathbf{R}^{k \times n}$. The solution is given by the normal equation:

$$A^T A x = A^T b \tag{3.100}$$

  Several extensions of least square problem:

  - Weighted least square: when the errors have different variances for different observations/samples, we need to minimize:

$$f(x) = \sum_{i=1}^{k} w_i (a_i^T x - b_i)^2 \tag{3.101}$$

    This can be easily transformed to a least square problem.

  - Regularization: we need to minimize, e.g. with $\rho > 0$:

$$f(x) = \sum_{i=1}^{k} (a_i^T x - b_i)^2 + \rho \sum_{i=1}^{n} x_i^2 \tag{3.102}$$

- Linear programming (LP): many problems can be tranformed to linear program. For example, the Chebyshev approximation problem, minimize:

$$f(x) = \max_{i=1}^{k} |a_i^T x - b_i| \tag{3.103}$$

  This is simlar to least square problem. This can be converted to LP by defining the size (maximum absolute error) as t:

$$\text{Minimize } t \qquad \text{s.t. } a_i^T x - t \leq b_i, -a_i^T x - t \leq -b_i, i = 1, 2, \cdots, k \tag{3.104}$$

- Convex optimization: the goal is:

$$\text{Minimize } f_0(x) \qquad \text{s.t. } f_i(x) \leq -b_i, i = 1, 2, \cdots, m \tag{3.105}$$

  where the functions $f_0, f_1, \cdots, f_m$ are all convex.

Concerns/Strategies of optimization problems and the benefit of convex optimization:

- Feasibility of solutions: a solution that satisfies all constraints may not exist, and checking feasibility is generally difficult. For convex optimization, feasibility is easy to find.
- Local vs. global optimum: a function may have many local optima. For convex optimization problem, local optimum is always the global optimum.
- Convergence of the algorithm and the stopping criteria: could be a major problem. For convex optimization problem, the stopping criteria is easy to define.
- Numerical stability: a concern for all optimization algorithms.

- General strategy: a key is to convert an optimization problem into a convex optimization problem, which then can be solved.

Convex sets: definitions and important examples

- Definition: a set C is convex, if for any $x_1 \in C$ and $x_1 \in C$, the line segment between x_1 and x_2 also falls in C, i.e.
$$\lambda x_1 + (1-\lambda)x_2 \in C \tag{3.106}$$
- Convex hull: given n points $x_1, \cdots, x_n$, the convex hull is the set of points that are convex combinations of the x_i's: $\lambda_1 x_1 + \cdots + \lambda_n x_n$, where $\lambda_1 + \cdots \lambda_n = 1$. More generally, for any set C, one can define the convex hull of C:
$$\{\lambda_1 x_1 + \cdots + \lambda_k x_k | x_i \in C, \lambda_i \geq 0, \lambda_1 + \cdots \lambda_n = 1\} \tag{3.107}$$
- Convex cone: a cone is a set of rays passing through the origin. S is a convex cone if:
$$\forall x, y \in S, \lambda, \mu > 0 \Rightarrow \lambda x + \mu y \in S \tag{3.108}$$
- Hyperplanes and half spaces: hyperplanes are defined by a linear equation (below).
- Eucledian balls and ellipsoids: a ball centered on x_c is defined as:
$$B(x_c, r) = \{x | \|x - x_c\|_2 \leq r\} \tag{3.109}$$
An Eucledian ball is a convex set (triangle inequality). An ellipsoid is defined as:
$$C = \left\{x | (x - x_c)^T P^{-1} (x - x_c) \leq 1\right\} \tag{3.110}$$
where P is symmetric and positive definite. The proof of convex set: linear transformation of convex set (ellipsoid is a linear transformation of ball, defined by the matrix P) is still a convex set.
- Norm balls: defined by $\{x | \|x - x_c\| \leq r\}$, where we could use L_1 norm, infinity norm, etc. That norm balls are convex sets can be proven by the triangle inequality: given $x_1, x_2 \in C$, let $x = \lambda x_1 + (1-\lambda)x_2$, we have:
$$\|x - x_c\| = \|\lambda(x_1 - x_c) + (1-\lambda)(x_2 - x_C)\| \leq \lambda \|x_1 - x_C\| + (1-\lambda)\|x_2 - x_C\| \leq \lambda r + (1-\lambda)r = r \tag{3.111}$$
Norm balls are important in machine learning for regularization.
- Norm cones: defined by:
$$C = \{(x, t) | \|x\| \leq t\} \subseteq \mathbb{R}^{n+1} \tag{3.112}$$
- Polyhedra: defined as the solution set of a finite number of linear equalities and inequalities. A polyhedron is thus the intersection of a finite number of halfspaces and hyperplanes.
- Positive semi-definite cones: defined as:
$$S_+^n = \left\{\Sigma \in S^n | \forall x \in \mathbb{R}^n : x^T \Sigma x \geq 0\right\} \tag{3.113}$$
where S^n is the set of symmetric $n \times n$ matrices. The set S_+^n is a convex cone. This can be proven by:
$$x^T(\lambda_1 \Sigma_1 + \lambda_2 \Sigma_2)x = \lambda_1 x^T \Sigma_1 x + \lambda_2 x^T \Sigma_2 x \geq 0 \tag{3.114}$$
This set is important in machine learning: e.g. the set of covariance matrices.

Operations that preserve convexity:

- Intersection: if S_1 and S_2 are convex sets, then $S_1 \bigcap S_2$ is also a convex set. Examples: polydedron, positive semidefinite cone (intersection of infinitely many linear constraints).
- Affine function: given a vector x, $f(x) = Ax + b$ is an affine function, where $A \in \mathbb{R}^{n \times n}$. If S is convex, then the image of S, $f(S)$ is also convex. Examples: translation, scaling, ellipsoids.
- Linear fractional functions:
$$f(x) = \frac{Ax + b}{c^T x + d} \tag{3.115}$$
where $c^T x + d \geq 0$.
- Theorem: any closed convex set can be represented as the intersection of halfspaces (possibly uncountably infinite) which contain it.
Remark: a generalization of the fact that polyhedron is an intersection of a finite number of halfplanes.

Two theorems about convex sets:

- Separating Hyperplane Theorem: Every two non-intersecting convex sets C and D have a separating hyperplane. Proof by construction: find points $c \in C$ and $d \in D$ that minimizes the Eucledian distance between C and D, then the hyperplane that is orthogonal to $d - c$, and passes the midpoint, separates C and D.
- Supporting Hyperplane Theorem: for any non-empty convex set C, and any point x_0 in the boundary of C, there exists (at least one) supporting hyperplane at x_0.
  - Intuition: tangent of C at x_0.
  - Another intuition: a convex set is an intersection of halfspaces, so we only need to choose one of these hyperplanes passing x_0.
  - Relation to Separating Hyperplane Theorem: we could construct one separating hyperplane between two convex sets that is a supporting hyperplane for one set.

Convex functions:

- Motivation: find an example (or a class of functions) where the global minimum of f exists, and there is an efficient algorithm to find it. Then for any new problem, try to map it to this class of functions.
- Definition: a function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if for any $x, y \in \text{dom} f$ and $\alpha \in [0, 1]$:
$$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha) f(y) \tag{3.116}$$
- Restriction to lines: $f : \mathbb{R}^n \to \mathbb{R}$ is convex iff the restriction of f on any line is convex, i.e. $g : \mathbb{R} \to \mathbb{R}$, $g(t) = f(x + tv)$ is convex in t.
  - Remark: a very useful property. Ex. used for the proof that $f(X) = \log \det X$, where X is a positiive definite matrix, is concave (page 74 of the book)
- Epigraph: for a convex function f is defined as the region bounded ("enclosed") by the function:
$$\text{epi} f = \{(x, t) | f(x) \leq t\} \tag{3.117}$$
By definition, it is easy to show that f is a convex function iff epi$f$ is a convex set.
- Sublevel set: the α-sublevel set of $f : \mathbb{R}^n \to \mathbb{R}$ is the set:
$$C_\alpha = \{x \in \text{dom} f | f(x) \leq \alpha\} \tag{3.118}$$
By definition, it is easy to show that C_α is convex for any α (however, the converse is not true).

Conditions of convexity in terms of derivative:

- First-order condition of convexity: suppose $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable, then f is convex iff for all $x, y \in \text{dom} f$:
$$f(y) \geq f(x) + \nabla f(x)^T (y - x) \tag{3.119}$$
That is, $f(y)$ is above $L(y)$ where $L(\cdot)$ is the hyperplane tangent on f at x. Proof ideas:
  - If: Let $x^* = \alpha x + (1 - \alpha) y$, we consider the hyperplane $L(\cdot)$ passing x^* that is tangent at f. By the condition, $f(x) \geq L(x)$ and $f(y) \geq L(y)$. , we have:
$$\alpha f(x) + (1 - \alpha) f(y) \geq \alpha L(x) + (1 - \alpha) L(y) = L(x^*) = f(x^*) \tag{3.120}$$
  - Only if: if f is convex, then epi$f$ is a convex set. By the Supporting Hyperplane Theorem, for any $x \in \text{dom} f$, there exists a hyperplane passing x s.t. epi$f$ is in the one side of the hyperplane, and it is easy to show that this translates to the inequality.
- Second-order condition of convexity: suppose $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable, then f is convex iff for all $x, y \in \text{dom} f$: $\nabla^2 f(x)$ is positive semidefinite (PSD).
Proof: consider the neighborhood of x, by the first-order condition, for any point y, we have:
$$f(y) \geq f(x) + \nabla f(x)^T (y - x) \tag{3.121}$$
However, $f(y)$ can also be expressed in terms of $f(x)$, $\nabla f(x)$ and $\nabla^2 f(x)$ using Taylor series, so for any y, we must have $(y - x)^T \nabla^2 f(x)(y - x) \geq 0$.

Examples of convex functions:

- Linear (affine) function: $f(x) = b^T x + c$.
- Quadratic function: for a PSD matrix A,
$$f(x) = x^T A x + b^T x + c \tag{3.122}$$
- Norms: $f(x) = \|x\|$, where the norm could be L_1, L_2, etc. The proof follows the triangle inequality.
- Log-sum-exp: the function f is convex:
$$f(x) = \log \left(\sum_{i=1}^{n} \exp(x_i) \right) \tag{3.123}$$

Properties of convex functions:

- Nonnegative weighted sum: If f_1 and f_2 are convex, then $f_1 + f_2$ is also convex. If f is convex, and $c > 0$, then cf is also convex.
- Composition with an affine function: if f is convex, then $f(Ax + b)$ is also convex. Example:
  - Norm: $f(x) = \|Ax + b\|$
- Pointwise maximum: if $f_1, \cdots, f_m$ are convex, then $f(x) = \max\{f_1(x), \cdots, f_m(x)\}$ is convex. Example:
  - Piecewise linear function: $f(x) = \max_{i=1,2,\cdots,m}(a_i^T x + b_i)$.
- Pointwise supremum: if $f(x, y)$ is convex in x for each $y \in A$, then, $g(x) = \sup_{y \in A} f(x, y)$ is convex.
  - A generalization of the preivous results: if A is finite, then $g(x) = \max\{f(x, y_1), \cdots, f(x, y_m)\}$, and g is convex by the pointwise maximum.

- Proof: the intersection of the epigraphs is also convex.
- Furthest distance to a point in a set C: $f(x) = \sup_{y \in C} \|x - y\|$.
- Maximum eigenvalues of real symmetric matrix: for $X \in S^n$, the function is convex:

$$\lambda_{\max} = \sup_{\|y\|=1} y^T X y \tag{3.124}$$

- Composition with scalar function: $g : \mathbb{R}^n \to \mathbb{R}$, and $h : \mathbb{R} \to \mathbb{R}$, then

$$f(x) = h(g(x)) \tag{3.125}$$

  is convex if g is convex, h is convex and nondecreasing. The proof follows from the second derivative of f and the chain rule. Examples:

  - If g is convex, then $\exp g(x)$ is convex.
  - If g is concave and positive, then $1/g(x)$ is convex.

- Vector composition: similar to before, $g : \mathbb{R}^n \to \mathbb{R}^k$, and $h : \mathbb{R}^k \to \mathbb{R}$, then

$$f(x) = h(g_1(x), \cdots, g_k(x)) \tag{3.126}$$

  is convex if g_i is convex, h is convex and h is nondecreasing in each argument. Examples:

  - $\sum_i \log g_i(x)$ is concave if g_i are concave and positive.
  - $f(x) = \log\left(\sum_{i=1}^n \exp(g_i(x))\right)$ is convex if g_i are convex.

- Minimization over some variable: $f(x, y)$ is convex on (x, y), and C is a convex set, then

$$g(x) = \inf_{y \in C} f(x, y) \tag{3.127}$$

  is convex. Proof: projection of $f(x, y)$ on x forms a convex set, and g is the boundary of this set. Examples:

  - Distance to a set: $\text{dist}(x, S) = \inf_{y \in S} \|x - y\|$ is convex if S is convex.

- Jensen's inequality: suppose f is convex, and $\sum_i \theta_i = 1, \theta_i \geq 0$, then:

$$f\left(\sum_i \theta_i x_i\right) \leq \sum_i \theta_i f(x_i) \tag{3.128}$$

  The continuous version of the inequality: suppose X is a random variable, then:

$$f(E(X)) \leq E(f(X)) \tag{3.129}$$

Definitions of convex optimization problem:

- Convex optimization problem: the standard form:

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, i = 1, \cdots, m \\ & h_i(x) = 0, i = 1, \cdots, p \end{array} \tag{3.130}$$

  where f_0, f_i and h_i are all convex functions.

- Feasibility: the set of x that satisfies all the constraints is called the feasibility set. And x is feasible if x satisfies the constraints.

- Theorem: if $\hat{x}$ is a local minimizer of a convex optimization problem, then $\hat{x}$ is a global minimizer. Proof: if $\nabla f(\hat{x}) = 0$, then by the first-order condition of convexity, we have, for any x,

$$f(x) \geq f(\hat{x}) + \nabla f(\hat{x})(x - \hat{x}) = f(\hat{x}) \tag{3.131}$$

  If $f(\hat{x}) \neq 0$, there is a direction of descent and we can find x' s.t. $f(x') \neq f(\hat{x})$, this is contradictory to local minimality unless $\hat{x}$ is in the boundary.

Conversion of forms of convex optimization problems:

- Eliminating equality constraints:

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, i = 1, \cdots, m \\ & Ax = b \end{array} \tag{3.132}$$

  Note that from $Ax = b$, we could have $x = Fz + x_0$ for some z, thus we could restate the problem in terms of z: minimize $f_0(Fz + x_0)$ over z, subject to $f_i(Fz + x_0) \leq 0$.

- Introducing slack variables for inequality constraints:

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & a_i^T x \leq b_i, i = 1, \cdots, m \end{array} \tag{3.133}$$

  is equivalent to:

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & a_i^T x + s_i = b_i, i = 1, \cdots, m \\ & s_i \geq 0, i = 1, \cdots, m \end{array} \tag{3.134}$$

- Epigraph form: the standard form is equivalent to:

$$\begin{array}{ll} \text{minimize over}(x, t) & t \\ \text{subject to} & f_0(x) - t \leq 0 \\ & f_i(x) \leq 0, i = 1, \cdots, m \\ & h_i(x) = 0, i = 1, \cdots, p \end{array} \tag{3.135}$$

- Mixing constrained and unconstrained optimization:

$$\begin{array}{ll} \text{minimize} & f_0(x_1, x_2) \\ \text{subject to} & f_i(x_1) \leq 0, i = 1, \cdots, m \end{array} \tag{3.136}$$

  is equivalent to:

$$\begin{array}{ll} \text{minimize} & \tilde{f}_0(x_1) \\ \text{subject to} & f_i(x_1) \leq 0, i = 1, \cdots, m \end{array} \tag{3.137}$$

  where $\tilde{f}_0(x_1) = \inf_{x_2} f_0(x_1, x_2)$.

- Remark: a number of ways to converting the optimization problems, e.g. representing the (equality constraint) in parameteric form and state the problem in terms of new parameters, slack variable for inequality constraints, introducing variables for the objective function and constraints, etc.

Classes/examples of convex optimization problems:

- Linear programming (LP): affinie objective and constraint functions:

$$\begin{array}{ll} \text{minimize} & c^T x + d \\ \text{subject to} & Gx \preceq h \\ & Ax = b \end{array} \tag{3.138}$$

  The feasibility set is a polyhedron.

- LP example: piecewise-linear minimization, this is equivalent to the LP problem:

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & a_i^T x_i + b_i \leq t, i = 1, \cdots, m \end{array} \tag{3.139}$$

- LP example: Chebyshev center of a polyhedron, defined as the center of the largest inscribed ball. Suppose the polyhedron is defined as:

$$P = \{x | a_i^T x \leq b_i, i = 1, \cdots, m\} \tag{3.140}$$

  Then the center (x_C) and r can be solved by the LP:

$$\begin{array}{ll} \text{minimize} & r \\ \text{subject to} & a_i^T x_C + r\|a_i\|_2 \leq b_i, i = 1, \cdots, m \end{array} \tag{3.141}$$

  where the constraint specifies that the ball must be inside the polyhedron (i.e. any point in the ball must satisfy the inequalities of the polyhedron).

- Quadratic programming (QP), Geometric programming (GP): see the slides/book.

Lagrange dual function: [Hindi, A Tutorial on Convex Optimization II: Duality and Interior Point Methods]

- Optimization problem: we have a problem of the form:

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, i = 1, \cdots, m \\ & h_i(x) = 0, i = 1, \cdots, p \end{array} \tag{3.142}$$

  Note that the functions are not necessarily convex.

- Motivation: our basic idea is to design an (unconstrained) optimization problem s.t. the solution of the new problem would give rise to the solution to the original problem. For simplicity, we consider only one inequality constraint $f_1(x) \leq 0$. We consider the function of this form:

$$L(x, \lambda) = f_0(x) + \lambda f_1(x) \tag{3.143}$$

  Then minimizing $L(x, \lambda)$ tends to lead to small $f_0(x)$ as desired; meanwhile, if $\lambda \geq 0$, then minimizing L would also require a small $f_1(x)$, possibly satisfying the constraint.

- Lagrangian: we define the Lagrangian as:

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \tag{3.144}$$

  We refer to λ_i and ν_i's as Lagrange mulipliers associated with inequality and equality constraints. Note that λ_i should be nonnegative according to our intuition.

- Lagrange dual function: we hypothesize that minimizing Lagrangian above w.r.t. x is related to the original problem. We define the dual function of the original (primal) problem as:

$$g(\lambda, \nu) = \inf_{x \in D} \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \right) \tag{3.145}$$

  where D is the domain of x (common domain of all f_i and h_i's). We next study how $g(\lambda, \nu)$ is related to the solution of the primal problem.

Lagrange dual problem:

- Lower bounds on optimal value: suppose p^* is the optimum of the primal problem. For any $\lambda \succeq 0$, we have:
$$g(\lambda, \nu) \leq p^* \tag{3.146}$$
To show this, let $\tilde{x}$ be a feasible point for the original problem. We have:
$$L(\tilde{x}, \lambda, \nu) = f_0(\tilde{x}) + \sum_{i=1}^{m} \lambda_i f_i(\tilde{x}) + \sum_{i=1}^{p} \nu_i h_i(\tilde{x}) \leq f_0(\tilde{x}) \tag{3.147}$$
Thus the minimum of the LHS, $g(\lambda, \nu) \leq f_0(\tilde{x})$ for any feasible point $\tilde{x}$, hence it is a lower bound of p^*.

- Dual problem: since the lower bound holds for any λ, ν, the maximum of $g(\lambda, \nu)$ must also be a (tighter) lower bound of p^*. We define the dual problem as:
$$\begin{array}{ll} \text{maximize} & g(\lambda, \nu) \\ \text{subject to} & \lambda \succeq 0 \end{array} \tag{3.148}$$
Note that the dual problem is a convex optimization problem. In fact, $g(\lambda, \nu)$ is a concave function. It is pointwise minimum of the function $L(x, \lambda, \nu)$: for any specific value of x, L is convex of λ and ν.

- Weak duality: Let d^* be the solution of the dual problem, clearly,
$$d^* \leq p^* \tag{3.149}$$
This is true for all problems. We are interested in when the equality holds (thus we can solve the dual problem instead).

- Strong duality and Slater's constraint qualification: if the primal problem (assuming the equality constraint is linear $Ax = b$) is convex, i.e. f_0 and f_i are convex, then usually strong duality holds. The Slater's condition states that: if the problem is strictly feasible, i.e. there exists x s.t.
$$f_i(x) < 0, i = 1, \cdots, m \qquad Ax = b \tag{3.150}$$
Then the strong duality holds.

- Example: Lagrange dual of LP, see the example in [Hindi]. The dual problem is another LP in inequality form.

Optimality conditions: let x^* be the primal optimal, and λ^*, ν^* be the dual optimal, we want to find out the conditions that they must satisfy.

- Complementary slackness: from the definition of $g(\lambda, \nu)$, we have:
$$g(\lambda^*, \nu^*) \leq f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* f_i(x^*) + \sum_{i=1}^{p} \nu_i^* h_i(x^*) \tag{3.151}$$
Since x^* satisfies the constraints, and $\lambda_i \geq 0$, we also have:
$$f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* f_i(x^*) + \sum_{i=1}^{p} \nu_i^* h_i(x^*) \leq f_0(x^*) \tag{3.152}$$
If the strong duality holds, the two equality signs above must be true as $g(\lambda^*, \nu^*) = f_0(x^*)$. This implies that: (1) x^* minimizes $L(x, \lambda^*, \nu^*)$; and (2) complementary slackness:
$$\lambda_i^* f_i(x^*) = 0 \text{ for } i = 1, \cdots, m \tag{3.153}$$

- Interpretation of complementary slackness: another way to see this is λ^*, ν^* minimizes $g(\lambda, \mu)$, thus either λ_i^* is in the boundary, i.e. $\lambda_i^* = 0$ or $\partial g / \partial \lambda = 0$. The latter implies that: $f_i(\tilde{x}) = 0$, where $\tilde{x}$ minizes $L(x, \lambda^*, \nu^*)$. With strong duality, $\tilde{x} = x^*$.

- KTT condition: suppose f_i and h_i are differentiable, since x^* minimizes $L(x, \lambda^*, \nu^*)$, the gradient vanishes at x^*:

$$\nabla f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*) + \sum_{i=1}^{p} \nu_i^* \nabla h_i(x^*) = 0 \tag{3.154}$$

  This condition and the other constaints constitute the KKT condition that the primal and dual optimal must satisfy (necessary):

$$\begin{array}{lcll} f_i(x^*) & \leq & 0 & i = 1, \cdots, m \\ h_i(x^*) & = & 0 & i = 1, \cdots, p \\ \lambda_i^* & \geq & 0 & i = 1, \cdots, m \\ \lambda_i^* f_i(x^*) & = & 0 & i = 1, \cdots, m \end{array} \tag{3.155}$$

  For convex problems, if the Slater's condition holds, then KTT is a necessary and condition for primal and dual optimality. Thus the original problem can be reduced to the problem of solving KTT condition (no optimization involved).

- Solving the primal via the dual: sometimes it is easier to solve the dual problem (which is always convex), than the primal problem. So we first solve the dual, and use the fact that x^* minimizes $L(x, \lambda^*, \nu^*)$ to solve x^*. See the example of "Minimizing a separable function subject to an equality constraint" in [Hindi].

Chapter 4

Discrete Mathematics & Algorithms

1. Graphs

   Eigenvalues of graphs:

   - Matrices associated with graphs: the weight (or adjacency matrix) W, the graph Laplacian $L = D - W$, and the transition matrix, $P = D^{-1}W$. Note that P is not symmetric, but P is similar to the matrix $S = D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$. First, we note that S is the normalized weight matrix:

   $$s_{ij} = \frac{w_{ij}}{\sqrt{d_i}\sqrt{d_j}} \tag{4.1}$$

   Then we have:

   $$P = D^{-1}W = D^{-\frac{1}{2}}(D^{-\frac{1}{2}}WD^{-\frac{1}{2}})D^{\frac{1}{2}} = D^{-\frac{1}{2}}SD^{\frac{1}{2}} \tag{4.2}$$

   - Eigenvalue of adjacency (weight) matrix: the Perron-Frobenius Theorem implies that if G is connected, then the largest eigenvalue has multiplicity 1. This eigenvalue, $\lambda_{\max}$ is a kind of "average degree":

   $$\max\{\bar{d}, d_{\max}\} \leq \lambda_{\max} \leq d_{\max} \tag{4.3}$$

   - Eigenvalues and eigenvectors of normalized graph Laplacian and transition matrix: let $L_{\text{rw}} = D^{-1}L = I - P$ be the graph Laplacian (random walk version), then λ is an eigenvalue of L_{rw} with eigenvector u if and only if $1 - \lambda$ is an eigenvalue of P with eigenvector u.
   - Reference: [Lovasz, Eigenvalues of graphs]

Chapter 5

Differential Equations, Dynamics Systems & Stochastic Processes

5.1 Dynamic Systems and Network Theory

1. Ecological networks

   Structural features of ecological networks:

   - Reference: Ecological networks and their fragility, [Montoya & Sole, Nature, 2006]
   - Connectivity: species in the networks are closely connected: in 7 food webs, the average degree of the majority of species is two to three.
   - Compartments exist in ecol. networks, corresponding to habitat boundaries. However, it's difficult to detect them and there is debate on their existence (especially within a habitat).
   - Clusters are more common: two examples, acaquatic system (a fish that feed from several trophic levels at various life stages), host-parasitoid systems, where a parasitoid may feed on a host, but also hyper-parasitize other parasitoids.
   - Link distribution: do not match other networks quantitatively. Moreover, the 'rich get richer' mechanism is at odds with ecological principles: e.g. as more species feed upon a fruit species, then competition for that fruit will increase. Species with many links to other species in the web may get that way simply by being the most abundant.
   - The nested diet structure: predicted by the cascase model. The top predator potentially exploiting all the other species, the next predator exploiting all but the top predator, and so on.
   - Trophic level, body size, species abundance and links: the larger a species' body size, the more species on which it can feed, and thus the higher its trophic level. Large body size and high trophic level mean lower abundance.

   Connectivity and stability of ecological networks:

   - Reference: Ecological networks and their fragility, [Montoya & Sole, Nature, 2006]. Anticipating Critical Transitions [Scheffer & Vandermeer, Science, 2012]. Systemic risk in banking ecosystems [Haldane & May, Nature, 2011].
   - Robert May's model of complexity(connectivity) vs. stability: complexity generally reduces stability of the network. A random assembly of N species, each of which had feedback mechanisms that would ensure the population stability were it alone, showed a sharp transition from overall

stability to instability as the number and strength of interactions among species increased. This transition occurs once:

$$m\alpha^2 > 1 \tag{5.1}$$

where m is the average number of links per species, and α their average strength.

  - This is an example of one broad class of networks: the units can flip between alternative stable states and where the probability of being in one state is promoted by having neighbors in that state. Other examples: banks (solvent or not).
  - The commonness of short paths in food webs suggests that disturbances spread rapidly throughout the food web.

- Implication: a trade-off between local and systemic resilience. Strong connectivity promotes local resilience, because effects of local perturbations are eliminated quickly through subsidiary inputs from the broader system. The same prerequisites that allow recovery from local damage may set a system up for large-scale collapse (Figure 1 of Scheffer12).
- The effect of removing one species: (1) most-connected species: many species lose their only prey source, and the web quickly breaks into many disconnected sub-webs. (2) Random species: these webs are robust, showing both little fragmentation and few secondary extinctions. Perhaps well-connected species are those that are relatively abundant at their particular trophic level and so are unlikely to be lost.

Detecting the early warning signs of a critical tipping point:

- Reference: Anticipating Critical Transitions [Scheffer & Vandermeer, Science, 2012].
- In the vicinity of many kinds of tipping points, the rate at which a system recovers from small perturbations becomes very slow, a phenomenon known as 'Critical slowing down'
- Combining this with network structure: e.g. what nodes are more likely to have an early warning sign?

Explaining stability: many weak-few strong interactions, modularity, hierarchical model.

- Reference: Systemic risk in banking ecosystems [Haldane & May, Nature, 2011].
- Challenge: a search for special food-web structures that may help reconcile complexity with persistence or stability.
- Importance of link strength: for highly connected food webs, there is a "many weak and few strong" pattern of interactions (as a species feed on more species, the strength of intersctions decreases). Thus the impact of disturbance will be attenuated as it passes through the network.
- Disassortative networks or modularity promotes robustness: cnce the system is appropriately compartmentalized, by firebreaks, or vaccination of "superspreaders", disturbance or risk is more easily countered.
- Nested hierarchy: networks with strong mutualistic interactions (e.g., pollination) are more robust if they have nested structures where specialists are preferentially linked in their mutualism to generalists that act as hubs of connectivity.

Application of ecological network analysis on financial system stability:

- Model of banking system: nodes are banks, each bank has several activities/states: deposit, interbank borrowing, interbank loans, external assets, and net worth. Several key parameters: the number of creditors, the capital reserver ratio.
- Mechanism of shock propagation in financial system:
  - Creditor banks will lose part or all of their loans. For this mechanism, each subsequent phase of loan-driven shocks is attenuated(many creditors).

  - Losses in the value of a bank's external assets
  - The diminished availability of interbank loans (perhaps most important)
- Insights from the model: e.g. excessive homogeneity within a financial system - the banks doing the same thing - can minimize risk for each individual bank, but maximize the probability of the entire system collapsing. Thus diversity and modularity could promote the stability of the banking system.

5.2 Stochastic Processes

1. Finite Markov chains

   Reference: [Lowler, Introduction to Stochastic Processes, Chapter 1-2]

   Markov chain dynamics:

   - Dynamics: let P be the transition matrix of the chain (stochastic matrix), $P = (p_{ij})$, where p_{ij} is the transition probability, $P(X_{t+1} = j | X_t = i)$. The dynamics of the chain can be represented by the probability, $\phi_i(t) = P(X_t = i)$. We have the dynamics:

     $$\phi_j(t+1) = \sum_i \phi_i(t) p_{ij} \tag{5.2}$$

     We could write this in a matrix form. Let $\phi(t)$ be the row vector at time t, we have:

     $$\phi(t+1) = \phi(t)P \tag{5.3}$$

     Thus at time t, we have:

     $$\phi(t) = \phi(0)P^t \tag{5.4}$$

     The power of P^t has the interpretation: the (i, j) entry of P^t is $P(X_t = j | X_0 = i)$.
   - Long-range behavior: as $n \to \infty$, what does P^n converge to? Intuitively: it could converge to some distribution over all states, or alternate between distributions, or stay forever in some states.

   Catalog of Markov chains:

   - Communication class: two states i and j communicate with each other, written as $i \leftrightarrow j$, if $\exists m, n \geq 0$ s.t. $p_m(i, j) > 0$ and $p_n(i, j) > 0$. Intuitively, the state j can be reached from i and vice versa. It is easy to prove that the communication relation is an equivalence class.
   - Irreducible MC: if a MC has only one communication class, it is irreducible. Intuitively, the chain will stay at any state, and must converge to some distribution π over all states, $\pi_i > 0, \forall i$. Define the period of a state i as:

     $$d_i := \text{GCD}\{n : p_i(i, i) > 0\} \tag{5.5}$$

     where GCD stands for the greatest common divisor. To see why a period is needed, consider a two-state MC with no self-loop, then clearly, it always alternate between the two states, and one can say that the period is equal to 2. We can prove that: all states of an irreducible MC have the same period d, defined as the period of the MC.
   - Reducible MC: multiple communication class. For a reducible MC, it is possible that some are states are sinks, and other states are "transient" (eventually, the chain will only stay in the sinks). In general, we define: transient classes (with probability 1, the chain will leave the class, and never get back) and recurrent classes. Intuitively, a reducible MC will setlle in one of the recurrent class, and within each recurrent class, follow the stationary distribution.
   - Random walk on a graph: the period is either 1 or 2 (bipartite graph).

Irreducible and aperiodic MCs:

- Theorem: if a MC is irreducible and aperiodic, i.e. $d = 1$, then there exists an invariant distribution, π, where $\forall i : \pi_i > 0$:

$$\pi P = \pi \tag{5.6}$$

  If ϕ is any initial distribution, then

$$\lim_{n\to\infty} \phi P^n = \pi \tag{5.7}$$

  We prove the theorem in three steps.

- Lemma: there exists $M > 0$ s.t. for all $n > M$, P^n is positive (i.e. all entries are positive).
  Proof: we first consider pairs. For any pair i, j, there exists $M(i, j)$ s.t. for all $n > M(i, j)$, $p_n(i, j) > 0$. Next we consider indiividual nodes, for each i, there exists $M(i)$ s.t. for all $n > M(i)$, $p_n(i, i) > 0$. We then choose M as the maximum of all $M(i, j)$ and $M(i)$.

- Application of Perron-Frobenius Theorem: we apply the theorem to P^n, the positive matrix. Note that it is still a stochastic matrix (rows summed to 1). The Perron root or the spectral radius of P^n is $r = 1$. Specifically, we have: (1) 1 is a simple (multiplicity one) eigenvalue of P^n, and all other eigenvalues have absolute value less than 1; (2) the left eigenvector of 1 is positive (all entries are positive).

- Convergence of P^n: show that P^n converges to the left eigenvector π. The idea is: we write the matrix in Jordon normal form, and take its power. The contributions of all other eigevalues except 1 disappear as $n \to \infty$.

Other MCs:

- Irreducible and periodic ($d \geq 2$) MC: the state space will split into d subsets $A_1, \cdots, A_d$ s.t. the chain cycle through the d subsets. In other words, in one step, it stays with certain distribution in the set A_1, and in next step, it moves to the set A_2 with another distribution, and so on. The average distribution is still π, the unique left eigenvector of 1. In general, P has d eigenvalues with absolute value 1, $z^d = 1$, and for any initial distribution ϕ:

$$\lim_{n\to\infty} \frac{1}{d} \left(\phi P^{n+1} + \cdots + \phi P^{n+d}\right) = \pi \tag{5.8}$$

  – Example: a bipartite graph.

- Reducible MC: suppose the chain has r recurrent classes, $R_1, \cdots, R_r$, and s transient classes $T_1, \cdots, T_s$. As $n \to \infty$, the chain will fall in one of the recurrent classes, and let π^k be the stationary distribution of the class R_k. We are interested in $p_n(i, j)$ for any i, j. Clearly, it is 0 when j is in a transient class. Otherwise, suppose $j \in R_k$, let $\alpha_k(i)$ be the probability that, starting at i, the chain ends up at the class R_k, we have:

$$\lim_{n\to\infty} p_n(i, j) = \alpha_k(i)\pi^k(j) \tag{5.9}$$

Detailed balance and time reversibility:

- Detailed balance: given a finite state Markov chain with rate matrix Q, if there exists a distribution π on the states s.t.:

$$\pi_i q_{ij} = \pi_j q_{ji} \qquad \forall i \neq j \tag{5.10}$$

  Then π is the equlibrium distribution of the chain (verifying the equation $\pi Q = \pi$). And the Markov chain is time reversible: at equilibrium, we have:

$$P(X_t = i, X_{t+1} = j) = \pi_i P(i \to j) = \pi_j P(j \to i) = P(X_t = j, X_{t+1} = i) \tag{5.11}$$

  Note that Q is reversible does not mean Q is symmetric; in fact a symmetric matrix implies uniform π according to detailed balance.

- Rate matrix parameterization: if Q is reversible, we could parameterize $q_{ij} = \pi_i s_{ij}$, where π is the equilibrium distribution and $s_{ij} = s_{ji}$ symmetric (verify the detailed balance equation). This is equivalent to say: $Q = S\Pi$, where S is a symmetric matrix and $\Pi = \text{diag}(\pi_1, \cdots, \pi_n)$.

2. Stochastic processes on graphs

   Stationary distribution of random walk on graphs:

   - Transition matrix: given an undirected graph with weights $W = (w_{ij})$. Let D be the degree matrix, $d_i = \sum_j w_{ij}$. The transition probability from i to j is proportional to w_{ij}, with the normalization constant d_i. Thus the transition probability matrix (stochastic matrix) is given by:

   $$p_{ij} = \frac{w_{ij}}{d_i} \tag{5.12}$$

   In a matrix from:

   $$P = D^{-1}W \tag{5.13}$$

   - Stationary distribution: if G is connected and not bi-partite (thus irreducible and aperiodic), then it has a unique stationary distribution π, with $\pi_i = d_i / \sum_i d_i$.

   PageRank:

   - Model: random walk on a (directed) graph G. Let P be the transition matrix of G, $p_{ij} = w_{ij}/d_i$. At each step, a probability $1 - d$ of random start, i.e. returning to the initial distribution. For the simple case, the initial distribution is uniform over all n nodes. Let $\phi_i(t)$ be the probability of being in the node i at the i-th step, then:

   $$\phi(t+1) = d \cdot \phi(t)P + (1-d)\frac{\mathbf{1}}{N} \tag{5.14}$$

   where $\mathbf{1}$ is the unit row vector. The PageRank socre of a node i is the probability of being at the i-th node in the stationary distribution.

   Epidemic Thresholds in Real Networks [CHAKRABARTI & FALOUTSOS, ACMT, 2008]:

   - Movitation: suppose a virus is propagated in a network, how fast the virus will be spread? What is the condition of an epidemic?
   - Basic model of virus propagation (SIS model): given a network G, a rate of infection, called the birth rate, β is associated with each edge, and a death rate δ is associated with each infected node. Let η_t be the size of the infected population at t, then the system can be described as:

   $$\frac{d\eta_t}{dt} = \beta\langle k\rangle \eta_t \left(1 - \frac{\eta_t}{N}\right) - \delta\eta_t \tag{5.15}$$

   The steady state solution is $\eta = N\left(1 - \frac{\delta}{\beta\langle k\rangle}\right)$. Importantly, there exists an epidemic threshold (condition): if $\beta\langle k\rangle < \delta$, then any viral outbreak will die out quickly.
   - New propagation model: for any node i, its state at time t, $X_{i,t}$ is binary (1 if infected, 0 if suspectible). The system can be described as a Markov chain of 2^N states (each state is a configuration of the graph). Let $p_{i,t} = P(X_{i,t} = 1)$, we derive the approximation (upper bound) of the recurrence equation of $p_{i,t}$. To be infected at t, there are two cases: (1) infected at $t-1$, and not cured, this probability is $p_1 = p_{i,t-1}(1-\delta)$; (2) not infected at $t-1$, but receive new infection at t, this probability satisfies:

   $$p_2 \le (1 - p_{i,t-1}) \cdot \beta \sum_{j \in \text{Nei}(i)} p_{j,t-1} \le \beta \sum_j a_{ij} p_{j,t-1} \tag{5.16}$$

Write in the matrix form, we have:

$$p(t) \leq Sp(t-1) \tag{5.17}$$

where $S = \beta A + (1-\delta)I$. Note: the new infection probability is approximated, and the paper has the exact form.

- Approximation and sufficient condition of epidemic die out: clearly, if the largest eigenvalue of S, $\lambda_{1,S} < 1$, the probability vector $p(t)$ will converge. It can be shown that the eigenvalues/eigenvectors of S and A are closely related: the i-th eigenvalue of S is related to the i-th eigenvalue of A:

$$\lambda_{i,S} = 1 - \delta + \beta\lambda_{i,A} \tag{5.18}$$

  and the two have the same eigenvectors. The condition $\lambda_{1,S} < 1$ is equivalent to $\beta\lambda_{1,A} < \delta$.

- Necessary condition of epidemic die out: to show the condition is also necessary, use the stability of the steady state (0). The epidemic die out would imply that the steady state is stable, and use the derivative, we can show the stability implies that $\lambda_{1,S} < 1$, or $\beta\lambda_{1,A} < \delta$.

- Remark:
  - Intuition of the epidemic threshold: the epidemic will die out if the birth rate is less than the death rate. The birth rate is roughly the number of neighbors, given by $\lambda_{1,A}$, times β, and the death rate is δ.
  - Why the result does not depend on initial condition: the only absorbing state of the Markov chain is the state where everyone is uninfected. Suppose we have a large number of initial infected nodes, then the number of deaths is high and outcompete the new births, so η_t will reduce and converge to the steady state; and similarly, if the initial η_0 is low, it will increase and converge to the steady state.

Chapter 6

Thermodynamics & Statistical Mechanics

6.1 Kinetic Theory of Ideal Gas

This section is based on Chapter 3 of Nelson [2004]

1. Overview

   Physical idea:

   - Ideal gas is a system of particles in random motion, thus everything about the gas is ultimately determined by the mechanics of particles.
   - Energy distribution of the particles determine the macroscopic properties of the system.

   Intuition: why energy distribution is important? Macroscopic properties depend on how a system interact with its environment, where interaction essentially means the exchange of energy (not consider particle exchange for now).

   - If system A has higher average energy than B, then the energy of particles in A will be transmitted to particles in B
   - If system A has many particles whose energy is greater than some threshold (needed for something to occur, e.g. a chemical reaction), then A will react faster

   Case studies:

   - Water evaporation rate: how it depends on temperature? Given two bottle of hot and cold waters, does putting them together increase the total evaporation rate?

2. Temperature and ideal gas law

   Definition: temperature of ideal gas is defined as:

   $$\left\langle \frac{1}{2}mv^2 \right\rangle = \frac{3}{2}k_B T \tag{6.1}$$

   Ideal gas law: pressure of the gas is created from the change of momentum of particles hitting the wall of the container. Apply Newton's law to the particles hitting the wall in Δt:

   $$PV = \frac{3}{2}N\left\langle \frac{1}{2}mv^2 \right\rangle = Nk_B T \tag{6.2}$$

3. Boltzman distribution

   Distribution of potential energy: consider gas under a potential field, then number of particles would not be uniform: need more particles in one side s.t. the pressure from this side exactly balances the effect of potential on the particle flow Feynman [1963]. Replace particle density with probability of finding a particle at a certain position x:

$$P(x) \propto e^{-\Phi(x)/k_B T} \tag{6.3}$$

   Distribution of kinetic energy: similarly consider gas under a potential field, the flow of particles is conserved (i.e. same at different positions). The flow rate depends on the velocity distribution, and use the distribution of potential energy to derive the following Feynman [1963]:

$$P(u) \propto e^{-mu^2/2k_B T} \tag{6.4}$$

   where u is the velocity (only one-dimensional case).

   Boltzman distribution: more generally, for any kind of energy, we will have the same kind of distribution:

$$P(s) \propto e^{-E(s)/k_B T} \tag{6.5}$$

   where $E(s)$ is the total energy of a particle at state s. For ideal gas, s refers to x and u. The distribution also applies not just to a single particle, but any larger systems or subsystems. For example, for a system with N particles whose energy only depends on the positions and velocities, then:

$$P(\mathbf{x_1}, \cdots, \mathbf{x_N}, \mathbf{v_1}, \cdots, \mathbf{v_N}) \propto e^{-E(\mathbf{x_1}, \cdots, \mathbf{x_N}, \mathbf{v_1}, \cdots, \mathbf{v_N})/k_B T} \tag{6.6}$$

   Application to reactions: (Arrhenius Law) suppose E^* is the energy barrier of a reaction, then the rate of reaction is proportional to the ratio of particles whose energy is higher than E^*. From Boltzman distribution, we have rate $\propto \exp(-E^*/k_B T)$

6.2 Entropy and Free Energy

This section is based on Chapter 6 of Nelson [2004]

1. Statistical postulate and entropy

   Problem: how to explain the macroscopic properties/behavior in terms of microscopic states/changes?

   Physical idea:

   - There are a very large number of possible micro-states of a given system, and any micro-state is equally probable. We could imagine that the system is constantly switching among these states, because of internal or external interactions.
   - The number and distribution of micro-states with certain properties determine the macroscopic properties of the system. A general way to analyze the system is: identify all micro-states and group micro-states by their properties, then analyze the distribution of different groups.
   - Remark: given a system, the collective properties of the underlying microscopic states lead to the (observed) macroscopic proerties. We could define the state of system using state variables, say x, Examples:
     - Protein-DNA system: many possible configurations, in some of these: the two are bound; and in the other states: not bound. Then x is a binary variable.
     - Solution: the density of the solution characterizes the state, it is a function over space.

To understand the behavior of the system, what matters is the number of micro-states subject to the constraint x, denoted as $\Omega(x)$, and according to the statistical postulate, the fraction of micro-states in x is $p(x) \propto \Omega(x)$. For micro. system, this $p(x)$ distribution is important (when E is the state variable, this is the Boltzman distribution); for macro. system, the number of states differ vastly across different x, thus only need to consider maximum $\Omega(x)$ (Second Law).

Entropy:

- Definition: suppose Ω is the number of micro-states of a system (as a function of some macroscopic states), then the disorder of the system is measured by its entropy, defined as:

$$S = k_B \ln \Omega \tag{6.7}$$

- The function ln is motivated by additivity requirement: for 2 independent systems, one has $\Omega = \Omega_1 \Omega_2$, thus $S = S_1 + S_2$.

Entropy of ideal gas: the entropy depends on the total kinetic energy (ignore potential energy), number of particles and volume. To count the number of micro-states (defined by velocities of all particles), note that a micro-state must be subject to the energy constraint:

$$E = \frac{1}{2m} \sum_{i=1}^{N} \sum_{J=1}^{3} (p_{i,J})^2 \tag{6.8}$$

where $p_{i,J}$ is the momentum of the i-th particle along the J axis. Since all micro-states lie in a high-dimensional space of all possible values of $p_{i,J}$, the problem of couting the actual micro-states is equivalent to computing the probability of the surface subject to the above constraint in this space. The result is Sakur-Tetrode equation.

Second Law of Thermodynamics:

- Law: for an isolated system (without exchange of particles and energy), any spontaneous process will change the entropy of the system by $\Delta S \geq 0$. At equilibrium, then $\Delta S = 0$ for any spontaneous process.
- Interpretation: for a given physical system, the macro-state of a system evolves s.t. the entropy of the macrostate is maximum. Suppose the macro-state of a system can be defined by a variable V (e.g. pressure of a system of gas), then among all possible states of the system, the state with V that maximizes $S(V)$ (the entropy of the system at the value V) is the most probable.
- Application of the law: if we define some variable that measures the macrostate of the system, then at equilibrium, this variables reaches a certain value that maximizes the total entropy.
- Examples:
  - Ideal gas: its free expansion will increase S because there will be more uncertainty/disorder (moleculars could occupy more spaces, thus more micro-states). We could define p (pressure) as the measure of macrostate, and at equilibrium $S(p)$ is maximized.
  - Mixing water and solutes: define a variable μ that measures the extent of mixture, then at equilibrium $S(\mu)$ is maximized.
- Remark: the state variable itself could be a function, e.g. $\rho(\vec{r})$ describes the density distribution of a system, then one need to maximize $S(\rho)$, a problem of optimization with functions as variables (calculus of variation).
  - Example: given a solution, at equilibrium, the concentration of solute should be spatially uniform, this can be proven by the Second Law. In general, if the molecules are also subject to the external field, the Second Law can be similarly applied, even though the distribution is not uniform.

Remark:

- Under some simple cases, one may be able to compute entropy through counting number of micro-states under certain constraint (macroscopic variables such as energy). Similar counting problems/arguments will be found when the micro-state analysis is needed (e.g. partition function).
- The behavior of a system must be constrained by the Second Law, to apply it, analyze the change of ΔS_{tot} for the system and its environment (if it is not isolated).

Questions:

- Exactly how entropy is defined? Ex 1. a system of N particles, where each particles has two possible states. Clearly the system where the two states are equally likely, and the system where one state is predominantly preferred, should have different entropies. Ex 2. in general, for a system with contiuous variables, how the number of microstates is defined?

2. Temperature

   Problem: put two systems into thermal contact, then they will excange energy through interaction of particles, and eventually reach equlibrium. What will be the state of thermal equlibrium (how energy is distributed)?

   Model: 2 systems A and B in thermal contact (together isolated), at equilibrium, its total entropy must be maximized from the Second Law. Let E_A and E_B be the energy of A and B respectively, and S_A, S_B be the entropy of A and B respectively, then:

   $$S_{tot}(E_A) = S_A(E_A) + S_B(E_B) = S_A(E_A) + S_B(E - E_A) \tag{6.9}$$

   where E is the total energy (conserved). Take derivative with respect to E_A and LHS should be equal to 0, we have:

   $$\frac{dS_A}{dE_A} = \frac{dS_B}{dE_B} \tag{6.10}$$

   Defintion: the above equation motivates the definition of temperature to characterize thermal equilibrium:

   $$T = (\frac{dS}{dE})^{-1} \tag{6.11}$$

   Temperature could be interpreted as the "availability of energy".

3. Open systems

   Motivation: given an open system that exchanges energy with its environment, what will be the analog of the Second Law that prescribes the change of the system?

   Physical idea: treat the system and its environment (heat bath) together, and analyze its change of entropy.

   Fixed volume system: an open system (a) with a large heat bath (B) at constant temperature T, and the volume of the system is constant. Suppose E_a changes by ΔE, then the entropy change of B is $-\Delta E/T$ by the definition of T. Apply the Second Law, we have:

   $$\Delta S_{tot} = \Delta S - \frac{\Delta E}{T} \geq 0 \tag{6.12}$$

   If we define $F = E - TS$ as the Helmholtz free energy of a system, then we have $\Delta F \leq 0$. Or at equilibrium, the free energy of the system is minimized.

   Fixed pressure system: similar to the above analysis, but need to consider the work done to the environment when computing its energy change. If we define $G = E + PV - TS$ as the Gibbs free energy, then we have $\Delta G \leq 0$. Or the Gibbs free energy of the system is minimized.

Maximum work: if a subsystem is in a state of greater than minimum free energy, it can do external work. The maximum possible work is $F - F_{min}$ or $G - G_{min}$.

Remark: the distinction between fixed volume and fixed pressure system does not matter for biological systems which do not involve gas phase.

4. Microscopic systems

Problem: if it is a small system that is inside a heat bath, then the fluctuation of the system will be important, will need to talk about the probability distribution of the states of the system.

Physical idea: consider the micro-states of the entire system: subsystem a and the heat bath (B). The energy difference in the subsystem will be transmitted to B, which will increase or decrease the entropy of B. Thus, the subsystem with different energy will have different probabilities. Similar to open macroscopic system, instead of maximizing number of micro-states, now it is needed to talk about probability distribution of micro-states.

Equilibrium distribution of a microscopic system:

- Boltzmann distribution: apply the above idea, the probability of a being in a state with energy E_a is proportional to the number of micro-states of B with energy $E - E_a$, i.e. $\Omega_B(E - E_a)$. This number could be be computed by (E_a is small):

$$S_B(E - E_a) = S_B(E) - E_a \frac{dS_B}{dE_B} = S_B(E) - \frac{E_a}{T} \tag{6.13}$$

  Speaking in terms of probability of the system a, we have: the probability of the system being in a state s is:

$$P(s) = \frac{1}{Z} e^{-E_s/k_B T} \tag{6.14}$$

  where $Z = \sum_s e^{-E(s)/k_B T}$ is the partition function.
- Kinetic interpretation of Boltzman distribution: consider a simple system with 2 types of chemical species (each corresponding to state of some particle). Let ΔE be the energy difference between the two states, and let E^* be the energy barrier of the conversion, then by the Arrhenius Law, the rates of reactions are $e^{-E^*/k_B T}$ and $e^{-E^* + \Delta E/k_B T}$ respectively. At equilibrium, the number of particles must satisfy:

$$N_2/N_1 = e^{-\Delta E/k_B T} \tag{6.15}$$

  Or P_2/P_1 is equal to the same ratio, where P_2, P_1 are interpreted as the probability of a randomly chosen particle being state 2 and 1 respectively.

Minimum free energy principle:

- Evolution of a microscopic system: the system has probability P_j of being in the state j, thus the evolution of the system is defined by $P = (P_1, P_2, \cdots, P_J)$, which changes over time. We know the equilibrium distribution is specified by the Boltzmann distribution, can we define a quantity as a function of P changes monotonically over time (thus Boltzmann distribution would optimize this quantity)?
- Free energy: the free energy of a subsystem a can be defined as:

$$F_a = \langle E_a \rangle - T S_a \tag{6.16}$$

  where $\langle E_a \rangle = \sum_j P_j E_j$ is the average energy of the subsystem, and $S_a = -k_B \sum_j P_j \ln P_j$ is the entropy of the subsystem. We can prove that: F_a decreases in a spontaneous process, and at equilibrium, F_a is minimized. In fact, we can solve this problem:

$$\min_P \sum_j P_j E_j + k_B T \sum_j P_j \ln P_j \qquad \text{subject to: } \sum_j P_j = 1 \tag{6.17}$$

  This can be solved by Lagrange's multiplier method, and the solution is Boltzmann distribution.

- Free energy at equilibrium: by plugging in the Boltzmann distribution into the equation of F_a, we can easily show that $F_a = -k_B T \ln Z$, where Z is the partition function of the subsystem a:

$$Z = \sum_j e^{-E_j/k_B T} \tag{6.18}$$

  The equlibrium free energy reflects the competition of energy and the entropy: a system will try to lower its energy (which favors the lowest energy state), and at the same time increases its entropy (which favors an equal partition of all states).

- Two-state systems: for a microscopic system, suppose the micro-states can be grouped into two classes I and II, then the probability of being in one class vs the other is the ratio of paritition functions in class I and II. In terms of the free energy, we have:

$$\frac{P_I}{P_{II}} = e^{-\Delta F/k_B T} \tag{6.19}$$

  where ΔF is the free energy difference between the two classes. It could be similarly interpreted in kinetic terms:

$$\frac{k_{I\to II}}{k_{II\to I}} = e^{\Delta F/k_B T} \tag{6.20}$$

  Example: a protein with many possible configurations, some of them represen the OPEN state (e.g. of a channel), and the others CLOSE state.

6.3 Entropic Forces

This section is based on Chapter 7 of Nelson [2004]

1. Overview: analysis of ΔG

   Problem: the principles of statistical mechanics specify how a system behaves according to its free energy (minimization of G for macroscopic systems, and Boltzmann distribution for microscopic systems). But given a system, how to analyze its ΔG in terms of its structural properties?

   Principle: for a system under constant temperature, $\Delta G = \Delta H - T\Delta S$, thus the two components can drive a spontanesou change:

   - ΔH: the loss of energy. Note that $\Delta H = \Delta E$, the change of internal energy for solutions where no volumn change (and external work) is involved.
   - ΔS: the increase of entropy.

   Example: hydrophobic interaction is driven by entropy. Suppose we have nonpolar molecules in water, they will tend to merge together because:

   - The nonpolar molecules disrupt the H-bond network of water molecules. For water molecules to still form H-bond in the surface of contact with nonpolar molecules, the orientations of water molecules are constrained.
   - Becuase the larger contact surface means more constrained micro-states, the tendency will be to minimize the contact surface.
   - In the end, once we compare the states where nonpolar molecules are isolated vs clustered: ΔH will be neglectable because the same number of H-bonds are formed among water molecules, but the entropy will be different.

6.4 Chemical Systems

This section is based on Chapter 8 of Nelson [2004]

1. Chemical potential

   Physical idea: put 2 systems A and B together that could exchange energy and particles, then N_A and N_B will be "balanced" at equilibrium (e.g. that all particles will be in one system will be unlikely).

   Chemical potential: two systems A and B could exchange energy and particles, the entropies of A and B are $S_A(N_A, E_A)$ and $S_B(N_B, E_B)$ respectively. Similar to the analysis of thermal equilibrium, the total entropy is maximized at equilibrium, therefore:

   $$\left.\frac{\partial S_A}{\partial E_A}\right|_{N_A} = \left.\frac{\partial S_B}{\partial E_B}\right|_{N_B} \tag{6.21}$$

   In general, when there are multiple chemcial species, then the partial derivative should be conditioned on fixed number of particles of all other species and fixed total energy. We have the definition of the chemical potential for a chemical species α:

   $$\mu_\alpha = -T\left.\frac{\partial S}{\partial N_\alpha}\right|_{E, N_\beta, \beta\neq\alpha} \tag{6.22}$$

   Remark: to understand the partial derivative wrt N_A while fixing the energy E_A, imagine that only "static" particles are added into A (no kinetic energy); or if the potential energy needs to be considered, extract the same amount of energy of the added particles

   Chemcial potential of gas and dilute solutions: let c be the concentration or density of the particles N/V, then the chemical potential is a function of c and T:

   $$\mu = k_B T \ln\frac{c}{c_0} + \mu^0(T) \tag{6.23}$$

   where c_0 is the reference concentration and $\mu^0(T)$ is the chemical potential of the reference concentration at T. The idea of the proof: for ideal gas, we know $S(N, E_{\text{kin}}, V)$, take derivative wrt to E_{kin}, to fix the total energy, extract the potential energy introduced. Suppose ϵ is the potential energy of one particle, then:

   - add ΔN particles holding E_{kin}: $\Delta S_1 = \Delta N \frac{dS}{dN}|_{E_{\text{kin}}}$
   - extract $\epsilon\Delta N$ kinetic energy holding N: $\Delta S_2 = -\epsilon\Delta N \frac{dS}{dE_{\text{kin}}}|_N$

   So we have:

   $$\left.\frac{\partial S}{\partial N}\right|_E = \left.\frac{\partial S}{\partial N}\right|_{E_{\text{kin}}} - \epsilon\left.\frac{\partial S}{\partial E_{\text{kin}}}\right|_N \tag{6.24}$$

   Remark: for some complex process of interest, design an equivalent process/device etc, to break it down into multiple small steps, components.

   Interpretation of chemical potential: the availability of "particles". Also it will be greater for molecules with large internal energy (more likely to dump the energy into the world as heat). So: a molecular species will be highly available for chemical reactions if its concentration c is big or its internal energy ϵ is big.

2. Gibbs distribution

   Problem: similar to the consideration of Boltzman distribution, the fluctation of a microscopic system inside a big heat bath with exchange of both energy and particles.

Gibbs distribution: similar to the analysis of Boltzman distribution, analyze the entropy of the heat bath (B), which now depends on both energy and the number of particles of the system: any energy or particle flow from the system into B will increase its entropy, whose value can be found via the defintion of T and μ. We have: the probability of being in state s with energy E_s and the number of particles N_s is:

$$P(s) = \frac{1}{Z} e^{(-E_s + \mu N_s)/k_B T} \tag{6.25}$$

Gibbs distribution for microscopic subsystems: when the microscopic system is complex, i.e. a single "state" of interest actually corresponds to many microscopic states, we will need to consider the free energy of the subsystem, as in 4. So we have:

$$P(s) = \frac{1}{Z} e^{(-G_s + \mu N_s)/k_B T} \tag{6.26}$$

where G_s is the free energy of the state s.

Remark: intuitively, the probability of a state depends on:

- Energy of the state (E_s): lower energy means higher energy of the heat bath, thus more likely;
- Particles of the state (N_s): smaller number of particles means more particles in the heat bath, thus more likely;
- Entropy of the state (S_s): if the entropy of the system itself at that state is larger, then the total entropy of system and heat bath is larger, thus the state is more likely.

3. Chemical reactions

   Problem: for a chemical reaction, the distribution of molecular species at equilibrium?

   A simple two-state system: suppose the two states of the same molecule could be converted, let μ_1 and μ_2 be the chemical potentials of the two states, then one molecule is converted from state 1 to 2, the world entropy change is $(-\mu_2 + \mu_1)/T$. By the Second Law, we must have $\mu_1 > \mu_2$. At equilibrium, we have $\mu_1 = \mu_2$.

   Equilibrium of chemical reactions: for a chemical reaction, define

$$\Delta G = \sum_j \nu_j \mu_j \tag{6.27}$$

   where ν_j is the stoichiometric coefficient of the species j (signed), then by our preceding analysis, we know that the reaction will run forward if $\Delta G < 0$ and backward if $\Delta G < 0$; at equilibrium $\Delta G = 0$. For gas or dilute solutions, plug in μ_j, we will have the concentrations of all species will be determined by the constant K_{eq}, which is:

$$K_{eq} = e^{-\Delta G^0/k_B T} \tag{6.28}$$

   where ΔG^0 is the ΔG of the reaction when all species are in reference condition.

   Remark: chemical potential can be equivalently defined as $\frac{\partial G}{\partial N}|_{T,p}$. Could show that for a chemical system, its change of free energy is exactly the quantity ΔG defined above: compute the total energy change in two ways (i) by two molecular species; (ii) by the subsystem and its environment.

Bibliography

Richard Feynman. *Feynman Lectures in Physics.* 1963.

Steven Leon. *Linear Algebra with Applications.* 2006.

Philip Nelson. *Biological Physics: Energy, Information, Life.* 2004.

David Poole. *Linear Algebra: a Modern Introduction, 2ed.* 2006.