COMP0199

1 Sequences of Functions

$(f_n)_{n=0}^{\infty}$ or (f_n) denotes the sequence of functions where $f_n : A \to \mathbb{R}$ for all $n \in \mathbb{N}$.

Sequence of Functions An assignment of a function $f_n : A \to \mathbb{R}$ to each $n \in \mathbb{N}$.

Hereafter, assume A as the domain / target interval.

1.1 Background

For numbers, the sequence u_n tends to l iff.

$$\forall \varepsilon > 0, \exists N \in \mathbb{N} : n > N \Rightarrow |u_n - l| < \varepsilon$$

A sequence of numbers u_n is Cauchy if

$$\forall \varepsilon > 0, \exists N \in \mathbb{N}, \forall m, n \geq N : |u_m - u_n| < \varepsilon$$

A sequence is Cauchy iff. it is convergent (in *complete* metric spaces). Note that the Cauchy Criterion does not refer to the limit itself.

1.2 Pointwise Convergence

If $\lim_{n\to\infty}(f_n(a))$ exists and is finite for all $a \in A$,
we can define the *limit function* $f : x \mapsto \lim_{n\to\infty}(f_n(x))$.
$(f_n(a))_{n\in\mathbb{N}}$ is said to **converge *pointwise*** towards f.

Pointwise Convergence The sequence $(f_n)_{n\in\mathbb{N}}$ converges pointwise to $f : A \to \mathbb{R}$ if

$$\forall x \in A, \forall \varepsilon > 0, \exists N \in \mathbb{N}, \forall n > N : |f(x) - f_n(x)| < \varepsilon$$

I.e., $\forall x \in A, f_n(x) \to f(x)$ as $n \to \infty$.

A sequence of functions (f_n) is pointwise Cauchy if $(f_n(x))$ is Cauchy for all x.
A sequence is pointwise Cauchy iff. it converges pointwise.

1.3 Uniform Convergence

Uniform Convergence The sequence $(f_n)_{n\in\mathbb{N}}$ **converges *uniformly*** to f on the interval A if

$$\forall \varepsilon > 0, \exists N \in \mathbb{N}, \forall x \in A, \forall n > N : |f(x) - f_n(x)| < \varepsilon$$

I.e., $\sup_{x\in A}|f_n(x) - f(x)| \to 0$ as $n \to \infty$.

Uniform convergence implies pointwise convergence, to the same limit. (It is *stronger*.)

Uniformity is preserved:
The limit f is continuous if all f_n are continuous and (f_n) converges uniformly.

Integrals are preserved:
If (f_n) converges uniformly to f on $[a, b]$, then

$$\lim_{n\to\infty} \int_a^b f_n(x)dx = \int_a^b f(x)dx$$

Derivatives are more complicated:
If $(f_n(x_0))$ converges for some $x_0 \in [a, b]$ and $(\boldsymbol{f_n'})$ converges uniformly on $[a, b]$, then

$$\lim_{n\to\infty} f_n'(x) = f'(x) \qquad \text{(for } x \in [a, b]\text{)}$$

...and the convergence of (f_n) is also uniform.

Note: the uniform convergence on the sequence of derivatives converges the "shape" of the function, while the single-point convergence on the sequence itself asserts that the solution is uniquely well-defined. (Consider $f_n(x) = n + x$.)

2 Series of Functions

$\sum f_n$ denotes the series of functions, i.e., the sum of the sequence (f_n).

2.1 easy stuff

$f = \sum f_n$ is more specifically $f : x \mapsto \sum_{n=0}^{\infty} f_n(x)$.
Clearly, this (infinite) sum may be undefined.

The partial sums of a series is a sequence itself: $\left(\sum_{k=0}^{n} f_k(x)\right)_n$

$\sum f_n$ **converges *pointwise*** on A if the sequence of *partial sums* converges pointwise on A.

Weierstrass M-Test A test for determining whether a series of functions converges *normally*:
$\sum f_n$ **converges *normally*** on A
if $\forall n \in \mathbb{N}, \forall x \in A : |f_n(x)| \leq M_n$
for $M_n \geq 0$, $\sum M_n$ converges.

Note that *normal convergence* implies ***uniform*** and ***absolute*** **convergence**.
The Weierstrass M-test may be seen as the Cauchy Criterion for uniform convergence.
Furthermore, the x-dependence is eliminated by effectively taking the suprema:

$$\sup_x \left| \sum_{n=N}^{\infty} f_n(x) \right| \leq \sum_{n=N}^{\infty} \sup_x |f_n(x)| \leq \sum_{n=N}^{\infty} M_n \to 0$$

- Does not help find the actual limit.
- Main tool used to test for uniform convergence.

2.2 Power Series

Power Series Series of functions of the form

$$\sum_{n \geq 0} a_n (x-b)^n$$

- “centered” around b
- effectively an infinite polynomial

A power series always converges for certain x:
- trivially true for $x = b$,
- then there exists a convergent neighborhood around b.

Radius of Convergence Some $r \geq 0$ where a power series converges uniformly on any closed interval $A \subset (b-r, b+r)$.
Divergence when $x \notin [b-r, b+r]$; unknown when $x = b \pm r$.
Can be infinite for series convergent everywhere.

D'Alembert's Ratio Test A test to find the radius of convergence r of a power series.
For some $\sum_{n \geq 0} a_n (x-b)^n$, consider the limit of the sequence of a-ratios $L = \lim_{n \to \infty} \left| \frac{a_{n+1}}{a_n} \right|$:

$$r = \begin{cases} \frac{1}{L} & \text{if } L \neq 0 \\ \infty & \text{if } L = 0 \\ 0 & \text{if } L = \infty \end{cases}$$

I.e., convergence when $|x - b| < \frac{1}{L}$.

2.3 Taylor Series

Taylor Series are a special case of power series.
They converge (within the radius) to arbitrary *analytic* functions.

For a function f infinitely differentiable at $b \in \mathbb{R}$, its Taylor Series centered at b is

$$\sum_{n \geq 0} \frac{f^{(n)}(b)}{n!} (x-b)^n$$

- converges to f in $(b-r, b+r)$ (if f is analytic at b)
- Maclaurin Series are a special case where $b = 0$

2.3.1 Well-known Functions

$$e^x = \sum_{n \geq 0} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots$$

$$\cos(x) = \sum_{n \geq 0} \frac{(-1)^n}{(2n)!} x^{2n} = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \ldots$$

$$\sin(x) = \sum_{n \geq 0} \frac{(-1)^n}{(2n+1)!} x^{2n+1} = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots$$

$$\ln(1+x) = \sum_{n \geq 1} \frac{(-1)^{n+1}}{n} x^n = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \ldots \quad \text{for } |x| < 1, x = 1$$

$$\frac{1}{1-x} = \sum_{n \geq 0} x^n = 1 + x + x^2 + x^3 + \ldots \quad \text{for } |x| < 1$$

$$\arctan(x) = \sum_{n \geq 0} \frac{(-1)^n}{2n+1} x^{2n+1} = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \ldots \quad \text{for } |x| < 1$$

2.3.2 Approximations

The degree-n Taylor polynomial of f centered at b is just the series truncated after $n+1$ terms:

$$T_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(b)}{k!} (x-b)^k$$

$f(x) \approx T_n(x)$ for $x \approx b$.

$R_n(x) = f(x) - T_n(x)$ is the tail of the terms in the series.
There are numerous ways to express this sub-series beyond a simple summation.

Lagrange Form of the Taylor Remainder

$$R_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} (x-b)^{n+1}$$

where ξ is a function of x and is strictly between b and x (open interval).

The derivation of such a term that “captures” an entire series comes from the Mean Value Theorem; the heavy lifting is done by ξ, which is a function itself and is *not* computable, but is bounded and thus provides an upper bound for $R_n(x)$. Let $M = \max_{\xi \in (b,x)} \left| f^{(n+1)}(\xi) \right|$:

$$|R_n(x)| \leq \frac{M}{(n+1)!} |x-b|^{n+1}$$

Since M is constant, we have, for $x \to b$:
- $R_n(x) \approx |x-b|^{n+1}$, or more specifically,
- $R_n(x) = O((x-b)^{n+1})$
- $R_n(x) = o((x-b)^n)$

3 Linear Algebra

Assume U, V, and W to be vector spaces.
Assume vector spaces to be over a field $\mathbb{F}$ (usually $\mathbb{R}$).
Assume $+$ over functions be *pointwise addition* – $(f + g)(x) = f(x) + g(x)$.

3.1 easy stuff

Linear Independence A subset $S = \{v_1, ..., v_k\} \subseteq V$ is linearly independent if
$\sum_{i=1}^{k} a_i v_i = \mathbf{0} \implies$ all $a_i = 0$

Alternatively, the map $(a_1, ..., a_k) \mapsto a_1 v_1 + ... + a_k v_k$ is injective.

Linear A map $f : V \to W$ is linear iff.
additivity $f(u + v) = f(u) + f(v)$ and
homogeneity $f(\lambda v) = \lambda f(v)$
are preserved (for arbitrary $u, v \in V$ and $\lambda \in \mathbb{F}$).

Additivity and homogeneity are commonly combined into $f(\lambda u + v) = \lambda f(u) + f(v)$.

Linear Map A *homomorphism* (see below) of a *vector space*.

Basis (…of a vector space V) is a subset $B \subseteq V$ that is linearly independent and *spans* V.

Direct Sum A structure-*propagating* Cartesian product[2] with operations defined element-wise. Notated with "oplus" $\oplus$.

The above is more specifically an **external** direct sum; an **internal** direct sum *conditionally* exists for *sub*structures (of a common abelian-group-like structure) iff. their intersection is *trivial*;
For a structure S with substructures A and B,
$A \oplus B$ is well-defined[1] ("A and B are in direct sum") iff. $A \cap B = \{\mathbf{0}\}$.

Decomposition Statements The expression $S = A \oplus B$ asserts that A and B are in direct sum and that their sum is exactly S.

$$S = A \oplus B \iff \text{"}A \oplus B\text{" } \textbf{and } S = A + B$$

Note: with internal direct sums, the $A \oplus B$ notation has overloaded meanings: it denotes the same sum set as $A + B$ while simultaneously *asserting* that $A \cap B = \{\mathbf{0}\}$.

Alternative definitions of internal direct sum:
Substructures A and B of a common S are in direct sum iff. every element of $T = A + B$ can be written in exactly one way as the sum of some $a \in A$ and $b \in B$. I.e.,

$$A \oplus B \iff \forall t \in T\ \exists!(a, b) \in A \times B : t = a + b$$

Again, $S = A \oplus B$ if also $T = S$. A more general version, assuming $V = \sum_i V_i$:

$$V = \bigoplus_i V_i \iff \forall v \in V, \exists!(v_i)_i \in \bigoplus_i V_i : v = \sum_i v_i$$

Note that the second $\oplus$ denotes the external direct sum.
The generalized intersection criteria: $\bigoplus_i V_i \iff \forall i : V_i \cap \sum_{j \neq i} V_j = \{\mathbf{0}\}$.

Note: the *direct product* is a structure-propagating Cartesian product;
the external direct sum is a direct product with the restriction that the resulting tuples can only have finitely many non-zero elements (which matters if there are infinite operands).

If $(u_1, ..., u_n)$ is a basis of U and $(v_1, ..., v_m)$ is a basis of V, then $U \oplus V$ iff. $(u_1, ..., u_n, v_1, ..., v_m)$ are linearly independent. (U and V are in the same ambient space.)
The internal direct sum is associative; the external direct sum is associative up to isomorphism.

Composition For functions $f : B \to C$ and $g : A \to B$,
composition is a binary operator where $f \circ g : A \to C$, $x \mapsto f(g(x))$.

Composition verifies (for arbitrary functions f, g, h):
- associativity $f \circ (g \circ h) = (f \circ g) \circ h$,
- identity $f \circ \mathrm{id} = \mathrm{id} \circ f = f$, and
- distributivity (over $+$)
  right $(f + g) \circ h = (f \circ h) + (g \circ h)$ and
  left $f \circ (g + h) = (f \circ g) + (f \circ h)$ **only if** f is *linear*.

$(\mathrm{End}(V), +)$ is an abelian group;
$(\mathrm{End}(V), +, \circ)$ is a ring.
$\mathrm{End}(V) := \{f : V \to V \mid f \text{ is linear}\}$; see below section on Endomorphisms.

Note that, for linear f (so symmetric distributivity), we obtain the properties as matrix multiplication.

Dimension The dimension of a vector space V, notated $\dim V$, is the cardinality of (any) basis of V.
Note that the Steinitz exchange lemma guarantees any two bases of V have the same cardinality.
$\dim \{0\} = 0$.

For a function $f : A \to B$:
Injectivity $\forall x, y \in A : f(x) = f(y) \Rightarrow x = y$ i.e., $\forall y \in \mathrm{im}\, f, \exists! x \in A : f(x) = y$
Surjectivity $\forall y \in B, \exists x \in A : f(x) = y$ i.e., $\mathrm{im}\, f = B$
Bijectivity Injectivity *and* Surjectivity.

3.2 Linear Maps

Assume f, g, h to be linear maps, and by default $f : V \to W$.

Image $\mathrm{im}\, f := \{f(v) \mid v \in V\} \quad = f(V) \subseteq W$
Kernel $\ker f := \{v \in V \mid f(v) = \mathbf{0}\}$
Rank $\mathrm{rank}\, f = \dim \mathrm{im}\, f$
Rank-Nullity Theorem $\mathrm{rank}\, f + \dim \ker f = \dim V$

Note that Rank is only defined for images of finite dimension, and
Rank-Nullity Theorem assumes V to have finite dimension too.

Kernel criterion for Injectivity f is injective iff. $\ker f = \{\mathbf{0}\}$.

Clearly, injectivity implies $\dim W \geq \dim V$. (Can be proved using the Rank-Nullity Theorem.)

Matrix Column Rank criterion for Injectivity f is injective iff. for a matrix M representing f, its columns are linearly independent.

Clearly, surjectivity implies $\dim V \geq \dim W$. (Can be proved using the Rank-Nullity Theorem.)

Note that any linear map can be made surjective trivially by setting its codomain (W) to its image.

Isomorphism Both-way-*structure-preserving* bijections.

Examples of structure properties to preserve: linearity, group operation.
The *structure* of vector spaces is linearity, and linear maps are, well, linear bidirectionally, so all **bijective linear maps are isomorphisms**.

Note[2]: The concept of *structure-preserving maps* come from category theory, where categories explicitly define its such maps, called *morphisms*; vector spaces' morphisms are linear maps by definition, so bijective linear maps are isomorphisms :shrug:.

Clearly, bijectivity implies $\dim V = \dim W$.
Intuitive way to express bijectivity: $\forall w \in W, \exists! v \in V : f(v) = w$.

Two structures are *isomorphic* iff. there exists an isomorphism between them.
For V and W over the same field, $\dim V = \dim W$ iff. they are isomorphic.

Matrix representations of a bijection must be square.
A matrix is invertible iff. it represents a bijection.

3.3 Endomorphisms

Endomorphism A map $f : V \to V$; same domain and codomain.
Automorphism An endomorphism that is also an isomorphism.

For endomorphisms, **injectivity, surjectivity, and bijectivity are equivalent** (in finite dimension).

f is bijective iff. it maps a basis to a basis.

3.3.1 Change of Basis

Consider an endomorphism f of V represented by matrix M in the basis B_1.

$N = P^{-1} M P$ is the matrix representing f in B_2 where P is a matrix whose columns are the coordinate vectors of B_2 in B_1.

Transition Matrix (aka. change-of-basis matrix) what P is above.

Observe that if $P^{-1} M P = N$, then $P N P^{-1} = M$.

4 Matrix Reduction

Assume V denotes some finite-dimensional vector space over $\mathbb{F}$ ($= \mathbb{R}$ or $\mathbb{C}$, as specified).
Assume $T : V \to V$ denotes some endomorphism of V.
Assume $M \in \mathcal{M}_{n,n}$ is the matrix representation of any such T.

Observe that as T is an endomorphism and $n = \dim V$, we can always represent it as some square matrix.

4.1 Abstract ahh Stuff

Invariant Subspace (aka. stable subspace) $U \subseteq V$ is an invariant subspace under T if $T(U) \subseteq U$.
I.e., $\forall u \in U : T(u) \in U$.

Observe that V and $\{\mathbf{0}\}$ are always invariant subspaces under *any* T;
$\ker T$ and $\operatorname{im} T$ are always invariant subspaces under a given T.
Note: it is also written "U is T-invariant" for some subspace U and transformation T.

Highly relevant to the concepts below: the COMP0147 notes on *cosets* and *quotient groups*.

Assume $U \subseteq V$ a subspace.

Coset For any $v \in V$, the coset $v + U$ (or $[v]$ in *equivalence class* notation) is the set $\{v + u \mid u \in U\}$.

Quotient Space A quotient (vector) space V/U is a new vector space over the set of all cosets of U.
I.e., V/U's carrier set is $\{\{v + u \mid u \in U\} \mid v \in V\}$.
Vector operations are defined as $[x] + [y] = [x + y]$ and $c[x] = [cx]$.
Observe that $\dim(V/U) = \dim V - \dim U$.

Note: there are no conditions on U; one can form a quotient vector space from *any* subspace. (Due to addition being abelian and multiplication being external.) This is contrary to general groups or rings, which require specific properties (normal subgroups or ideals) on the substructure.

Complementary Subspace A subspace $W \subseteq V$ is complementary to U iff. $V = U \oplus W$.
Construction from a Quotient Space:
Pick a basis $\{[v_1], ..., [v_k]\}$ for V/U.
Pick a $\tilde{v}_i \in [v_i]$ as a *representative* for every coset in the basis.
$W = \operatorname{span}\{\tilde{v}_1, ..., \tilde{v}_k\}$ is a subspace complementary to U.

Complements to a subspace are generally **non-unique**, and a complementary subspace can be viewed geometrically as a *section* of the corresponding quotient space. However, the addition of an inner product (to define orthogonality) allows the construction of a unique, **canonical** complement $U^\perp$.
(See *Orthogonality* subsection below.)

Observe that $W \cong (V/U)$ for any complement W. Specifically, the W-restriction of the projection map $\pi|_W : W \to V/U$ is an isomorphism.

The construction above produces a linear map $s : V/U \to V$ with $\pi \circ s = \mathrm{id}_{V/U}$, called a *section* (or *splitting*) of the projection $\pi : V \to V/U$. Its image is the complement W. Given a choice of s, the map $(u, [v]) \mapsto u + s([v])$ provides an *explicit* isomorphism $U \oplus (V/U) \cong V$. Because this isomorphism relies on the arbitrary choice of representatives to build s, there is no *canonical* isomorphism between V and $U \oplus (V/U)$.

4.2 Block Matrices

For a T-invariant U, one can construct, in a deliberate basis, a *block triangular* matrix.
Let W be a subspace complementary to U.

Pick a basis $B_U = (u_1, ..., u_k)$ for U and a basis $B_W = (w_1, ..., w_{n-k})$ for W.
$B = (u_1, ..., u_k, w_1, ..., w_{n-k})$ is then a basis for V. (Since $V = U \oplus W$.)

Now, we construct a matrix M for T in B.
Recall that a matrix can be constructed column-wise for $b \in B$ as $T(b)$'s coordinates (in B):

$$M = \begin{bmatrix} \uparrow & & \uparrow & \uparrow & & \uparrow \\ T(u_1) & \cdots & T(u_k) & T(w_1) & \cdots & T(w_{n-k}) \\ \downarrow & & \downarrow & \downarrow & & \downarrow \end{bmatrix} = \left[\begin{array}{c|c} A & B \\ \hline 0 & C \end{array}\right]$$

...where A is a $k \times k$ block; 0 is a $(n - k) \times k$ block of zeroes.

Observe:
- A is the B_U-coordinates of $T(U)$
- 0 is the B_W-coordinates of $T(U)$, which are zero since U is invariant
- B is the B_U-coordinates of $T(W)$
- C is the B_W-coordinates of $T(W)$

Observe that B will be zeroes if W is *also* invariant, yielding a *block diagonal* matrix.

Usefully, $\det(M) = \det(A) \times \det(C)$ and $\chi_M = \chi_A \times \chi_C$.

4.3 Eigenvalues and Eigenvectors

Eigenvector A nonzero vector $v \in V$ is an eigenvector of some T iff. there exists a scalar $\lambda \in \mathbb{F}$ where $T(v) = \lambda v$.

Eigenvalue The scalar λ above is the eigenvalue of T *corresponding* to the eigenvector v.

Other equations characterizing eigenvalues and eigenvectors wrt. matrix representations include

$$(\lambda I - M)v = \mathbf{0} \quad \text{or} \quad v \in \ker(\lambda I - M)$$

(See section below on the computational utility of these.)

An eigenvector's span is an invariant subspace where the associated endomorphism acts like a scaling: An eigenvector spans a 1-dimensional invariant subspace, and a linear transformation on 1-dimensional invariant subspaces can *only* be a scaling.

Eigenspace For a given eigenvalue λ, we have its eigenspace $E_\lambda = \ker(\lambda I - M) \subseteq V$.
Equivalently, $E_\lambda = \operatorname{span}\{v_1, ..., v_k\}$ where $\{v_1, ..., v_k\}$ is a basis of eigenvectors for λ (and k is the geometric multiplicity).

Observe that any $v \in E_\lambda \setminus \{\mathbf{0}\}$ is an eigenvector corresponding to λ.
Observe that $\{\text{eigenspaces of } T\} \subsetneq \{\text{invariant subspaces of } T\}$. (Unequal in nontrivial cases.)

Geometric Multiplicity The geometric multiplicity of an eigenvalue λ of M, $\gamma_M(\lambda)$, is $\dim E_\lambda$.
This is the "physical" dimension of the associated eigenspace.

Algebraic Multiplicity The algebraic multiplicity of an eigenvalue λ of M, $\alpha_M(\lambda)$, is its multiplicity as a root of χ_M (M's characteristic polynomial; see below).

Spectrum The spectrum of some T is the multiset of all its eigenvalues (with algebraic multiplicity).

Linear Independence of Eigenvectors For *distinct* eigenvalues $\lambda_1, ..., \lambda_k$, their corresponding eigenvectors $v_1, ..., v_k$ are linearly independent.
Alternatively: Take a basis for each eigenspace; the union of all these bases, across distinct eigenvalues, is linearly independent.
This allows the construction of bases out of eigenvectors if we have enough eigenvalues.

4.3.1 Computing Eigenvalues

Assume $v \in V \neq \mathbf{0}$ and $\lambda \in \mathbb{F}$.

Derivation from $Mv = \lambda v$ to the *characteristic polynomial*:

$$\begin{aligned} Mv &= \lambda v \\ \lambda v - Mv &= \mathbf{0} \\ \lambda(Iv) - Mv &= \mathbf{0} \\ (\lambda I - M)v &= \mathbf{0} \end{aligned}$$

Observe that a given λ is a solution
$\Longleftrightarrow \ker(\lambda I - M)$ is non-trivial
$\Longleftrightarrow \lambda I - M$ is not injective ($\Leftrightarrow$ not surjective $\Leftrightarrow$ not bijective)
$\Longleftrightarrow \lambda I - M$ is not *invertible*
$\Longleftrightarrow \det(\lambda I - M) = 0$.
I.e., any λ that satisfies any statement above is an eigenvalue, and the set of all such solutions are the eigenvalues of the transformation represented by M.

Characteristic Polynomial The characteristic polynomial χ_M for matrix M is

$$\chi_M(\lambda) = \det(\lambda I - M)$$

The roots of χ_M (expanding det with the Leibniz formula) are exactly the eigenvalues of M.

$$\begin{aligned} \chi_M(\lambda) = \det(\lambda I - M) &= \det\left(\begin{bmatrix} \lambda & 0 & \dots & 0 \\ 0 & \lambda & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \lambda \end{bmatrix} - M\right) \\ &= \lambda^n + a_{n-1}\lambda^{n-1} + \dots + a_2\lambda^2 + a_1\lambda + a_0 \end{aligned}$$

Aside[2]: there's a very subtle (and mostly insignificant) difference between $\det(M - \lambda I)$ and $\det(\lambda I - M)$. I refuse to elaborate further.

Observe that $T : V \to V$ cannot have more than $n = \dim V$ eigenvalues (or eigenspaces).
Observe that χ_M has degree n.
0 is a root of χ_M iff. $\det(M) = 0$

4.3.2 Computing Eigenvectors

Obviously, we just compute $\ker(\lambda I - M)$ for a given λ to obtain its eigenspace.

4.4 Polynomials and Linear Maps

The characteristic polynomial computes eigenvalues.
The roots and their multiplicity provide information on possible reductions (see below).

Complex vs Real fields
Per the fundamental theorem of algebra, not all roots of the characteristic polynomial may be real.
For $\mathbb{F} = \mathbb{R}$, some $n \times n$ matrix M has $\leq n$ eigenvalues.
For $\mathbb{F} = \mathbb{C}$, M has exactly n eigenvalues (counted with algebraic multiplicity).

Recall that polynomials in $\mathbb{R}$ factor into linear *and* irreducible quadratic factors, while polynomials in $\mathbb{C}$ factor fully into linear factors.

If $\mathbb{F} = \mathbb{R}$, χ_M factors into distinct irreducible quadratic terms P_i and linear terms.

$$\chi_M(x) = P_1^{r_1}(x) \times \dots \times P_k^{r_k}(x) \times (x - \lambda_1)^{s_1} \times \dots \times \left(x - \lambda_j\right)^{s_j}$$

If $\mathbb{F} = \mathbb{C}$:

$$\chi_M(x) = (x - \lambda_1)^{r_1} \times \dots \times (x - \lambda_m)^{r_m}$$

Annihilating Polynomial A polynomial P over $\mathbb{F}$ is an annihilating polynomial for M if $P(M) = 0$.

Minimal Polynomial The *monic* polynomial μ_M over $\mathbb{F}$ of least degree that annihilates M; $\mu_M(M) = 0$

Cayley-Hamilton Theorem Square matrices are annihilated by their own characteristic polynomials; $\forall M \in \mathcal{M}_{i,i} : \chi_M(M) = 0$
Effectively, this means $\{I, M, M^2, ..., M^n\}$ is linearly dependent, and χ_M itself demonstrates the exact dependence (via the coefficients). E.g., one can:
- compute matrix inverses cheaply (by rearranging $\chi_M(M)$)
- represent high powers M^k where $k \geq n$ as some linear combination of the lower powers

4.5 Matrix Reduction

Recall that a given matrix is just *one* representation of a linear transformation and is implicitly associated with a chosen basis.
Different bases thus correspond to different matrices that still represent the same transformation.

- We want diagonal matrices as they're computationally advantageous.
- While some T's matrix M in the standard basis may not be diagonal, it may have a diagonal representation under a different basis.
- Such a basis is precisely an *eigenbasis* – a basis consisting only of eigenvectors.

Eigenbasis For a transformation T, an *eigenbasis* is a basis of V consisting only of eigenvectors of T.

Diagonalization For a matrix M and its transformation T, M is *diagonalizable* iff. an eigenbasis exists for T.
Constructing P with the eigenbasis as columns, we obtain a change-of-basis matrix s.t. $D = P^{-1}MP$ is diagonal.

Further properties:
- the entries of D are the eigenvalues of M (with multiplicity)
- M is diagonalizable iff. μ_M splits into *distinct* linear factors.
- M is diagonalizable iff. χ_M splits over $\mathbb{F}$ *and* $\alpha_M(\lambda) = \gamma_M(\lambda)$ for all eigenvalues λ.
- M is diagonalizable iff. $\sum_i \gamma_M(\lambda_i) = n$.
- Powers of M are easy to compute: $M^k = PD^kP^{-1}$ (where D^k is just diagonal entries raised to k).
- Similarly, for any analytic function f (e.g., e^M): $f(M) = Pf(D)P^{-1}$.

Kernel Decomposition Theorem If the minimal polynomial of T is split into pairwise coprime factors $p(x) = p_1(x)^{r_1}...p_k(x)^{r_k}$, then V decomposes into a direct sum of the factors' kernels:

$$V = \ker(p_1(T)^{r_1}) \oplus \dots \oplus \ker(p_k(T)^{r_k})$$

Following the complete factorization of the characteristic polynomial, we have...
For $\mathbb{F} = \mathbb{R}$:

$$V = \ker\left(P_1^{r_1}(M)\right) \oplus \dots \oplus \ker\left(P_k^{r_k}(M)\right) \oplus \ker\left((\lambda_1 I - M)^{s_1}\right) \oplus \dots \oplus \ker\left(\left(\lambda_j I - M\right)^{s_j}\right)$$

For $\mathbb{F} = \mathbb{C}$:

$$V = \ker\left((\lambda_1 I - M)^{r_1}\right) \oplus \dots \oplus \ker\left((\lambda_m I - M)^{r_m}\right)$$

Note that every term and its power r or s directly correspond to the factorization of χ_M.

Generalized Eigenspace For an eigenvalue λ with algebraic multiplicity $m = \alpha_M(\lambda)$, the generalized eigenspace is

$$K_\lambda = \ker((\lambda I - M)^m)$$

- Its elements are *generalized eigenvectors.*
- $\dim K_\lambda = \alpha_M(\lambda)$ (always, unlike eigenspaces where $\dim E_\lambda \leq \alpha_M(\lambda)$).
- While the direct sum of eigenspaces $\bigoplus E_\lambda$ may not span V, the direct sum of generalized eigenspaces always does (if χ_M splits, e.g., over $\mathbb{C}$):

$$V = \bigoplus_i K_{\lambda_i}$$

Triangularizable A matrix M is similar to an upper-triangular matrix iff. its characteristic polynomial χ_M splits fully into linear factors over $\mathbb{F}$.
- Since $\mathbb{C}$ is algebraically closed, all matrices over $\mathbb{F} = \mathbb{C}$ are upper-triangularizable.
- A real matrix is triangularizable over $\mathbb{R}$ iff. all its eigenvalues are real. If it has irreducible quadratic factors, it cannot be triangularized (but can be block-triangularized with 2×2 blocks).

Jordan Canonical Form Every matrix whose characteristic polynomial splits is similar to a block-diagonal matrix of *Jordan blocks*:

$$J = \begin{bmatrix} J_{d_1}(\lambda_1) & 0 & \cdots & 0 \\ 0 & J_{d_2}(\lambda_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & J_{d_k}(\lambda_k) \end{bmatrix} \quad \text{where} \quad J_d(\lambda) = \begin{bmatrix} \lambda & 1 & 0 & \cdots & 0 \\ 0 & \lambda & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda & 1 \\ 0 & 0 & \cdots & 0 & \lambda \end{bmatrix}$$

5 Differential Calculus

5.1 Homogeneous Linear DEs

Homogeneous Linear Differential Equation (HLDE) A linear differential equation of the form

$$a_n y^{(n)} + a_{n-1} y^{(n-1)} + \dots + a_1 y' + a_0 y = 0$$

Principle of superposition The linear combination of any two solutions to a HLDE is also a solution. This is due to the differential operator being linear.

The general solution to an n-th order HLDE is a linear combination of n linearly independent solutions.

Let $D = \frac{\mathrm{d}}{\mathrm{d}x}$ be the linear differential operator. The HLDE can be written as:

$$P(D)y = 0$$

where $P(z) = a_n z^n + \dots + a_1 z + a_0$ is the characteristic polynomial.
Clearly, solving the HLDE is equivalent to solving for $\ker(P(D))$.

Spectral Mapping Theorem For any T and polynomial P, if v is an eigenvector of T with eigenvalue λ, then v is also an eigenvector of $P(T)$ with eigenvalue $P(\lambda)$:

$$Tv = \lambda v \implies P(T)v = P(\lambda)v$$

Applying this to $T = D$, whose eigenfunctions are $y = e^{\lambda x}$ with eigenvalue λ (since $D(e^{\lambda x}) = \lambda e^{\lambda x}$), we get:

$$P(D)e^{\lambda x} = P(\lambda)e^{\lambda x}$$

Thus, if some λ satisfies $P(\lambda) = 0$, then $P(D)e^{\lambda x} = 0$, meaning the eigenfunction $e^{\lambda x} \in \ker(P(D))$ is a solution.

The general solution is built from the roots of the characteristic polynomial $\chi(\lambda) = P(\lambda) = 0$:

1. **Distinct Real Roots**: For each real root λ_i, there is a solution $e^{\lambda_i x}$.
2. **Repeated Real Roots**: For a real root λ with multiplicity m, the kernel contains the m linearly independent solutions:
$$e^{\lambda x}, x e^{\lambda x}, \dots, x^{m-1} e^{\lambda x}$$
3. **Complex Conjugate Roots**: Since the coefficients a_i are real, complex roots appear in conjugate pairs $\lambda = a \pm ib$.
   - These yield complex solutions $e^{(a \pm ib)x} = e^{ax} e^{\pm ibx} = e^{ax}(\cos(bx) \pm i \sin(bx))$.
   - Taking the real and imaginary parts (which are linear combinations) yields two real, linearly independent solutions:
   $$e^{ax}\cos(bx) \quad \text{and} \quad e^{ax}\sin(bx)$$
   - If repeated with multiplicity m, multiply these by powers x^k for $k < m$.

For complex conjugate solutions $y_1 = u + iv$ and $y_2 = u - iv$, their real and imaginary parts are linear combinations of the conjugate pair themselves:

$$u = \frac{y_1 + y_2}{2} \quad \text{and} \quad v = \frac{y_1 - y_2}{2i}$$

By superposition, these linear combinations are also solutions, and they are guaranteed to be real-valued and linearly independent.

5.2 Multivariate Functions

Fubini's Theorem For a function $f : \mathbb{R}^2 \to \mathbb{R}$ that is continuous over a rectangle $X \times Y$ (X and Y are intervals), the double integral can be computed as an iterated integral:

$$\iint_{X \times Y} f(x,y)\,\mathrm{d}(x,y) = \int_X \left(\int_Y f(x,y)\,\mathrm{d}y \right) \mathrm{d}x = \int_Y \left(\int_X f(x,y)\,\mathrm{d}x \right) \mathrm{d}y$$

This extends to non-rectangular regions with appropriate limits of integration.

Gradient A representation of the first derivative of a multivariate function.
For a function $f : \mathbb{R}^n \to \mathbb{R}$, the gradient of f at some point $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ is the vector of partial derivatives:

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \dots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}$$

The gradient points in the direction of steepest ascent, and its magnitude is the rate of increase in that direction.

Hessian A representation of the second derivative of a multivariate function.
For a function $f : \mathbb{R}^n \to \mathbb{R}$, the Hessian of f at some point $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ is the matrix of second partial derivatives:

$$(\mathcal{H}_f)_{i,j} = \frac{\partial^2 f}{\partial x_i \partial x_j}$$

$$\mathcal{H}_f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix} = \mathbf{J}(\nabla f)^\top$$

The Hessian is used to analyze the local curvature of f and classify stationary points.

5.3 Stationary Points

If the second partial derivatives of f are continuous near a point x, then $\frac{\partial^2 f}{\partial x \partial y}(x) = \frac{\partial^2 f}{\partial y \partial x}(x)$.
Consequently, the Hessian is symmetric for any f with continuous second partial derivatives.

Classification of Stationary point x:

- $\det(\mathcal{H}_f(x)) > 0$
  - If $\frac{\partial^2 f}{\partial x_1^2}(x) > 0$, then x is a local minimum.
  - If $\frac{\partial^2 f}{\partial x_1^2}(x) < 0$, then x is a local maximum.
- $\det(\mathcal{H}_f(x)) < 0$
  x is a saddle point.
- $\det(\mathcal{H}_f(x)) = 0$
  inconclusive

More generally (beyond 2 dimensions), by the *Real Spectral Theorem*, the classification is performed from the signs of the eigenvalues (aka, *definiteness*) of the Hessian:

- positive definite (all eigenvalues > 0): local minimum
- negative definite (all eigenvalues < 0): local maximum
- indefinite (eigenvalues of mixed sign): saddle point
- singular (some eigenvalue $= 0$): inconclusive

6 Euclidean Spaces

Continue assuming V is a finite-dimensional vector space over $\mathbb{F} = \mathbb{R}$ *or* $\mathbb{F} = \mathbb{C}$;
$u, v, w \in V$ and $a, b \in \mathbb{F}$.

6.1 Inner Products

The *canonical* inner product is the dot product.

Inner Product Additional structure on a V as a binary operation $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{F}$ satisfying:

1. Conjugate symmetry: $\langle v, w \rangle = \overline{\langle w, v \rangle}$
2. Linearity in the first argument: $\langle av + bw, u \rangle = a\langle v, u \rangle + b\langle w, u \rangle$
3. Positive-definiteness: $u \neq 0 \Longrightarrow \langle u, u \rangle > 0$

Observe that for $\mathbb{F} = \mathbb{R}$, $x = \overline{x}$, so 1. is reduced to simple symmetry, and 2. extends to bilinearity.
Also, for $\mathbb{F} = \mathbb{C}$, 1. implies that $\langle u, u \rangle \in \mathbb{R}$.

Implied properties:
- $\langle 0, u \rangle = \langle u, 0 \rangle = 0$
- $\langle u, u \rangle = 0 \Longleftrightarrow u = 0$ (A way to prove zero-ness of a vector)
- $\langle u, av + bw \rangle = \overline{a}\langle u, v \rangle + \overline{b}\langle u, w \rangle$; i.e., conjugate-linearity in the second argument (*sesquilinear*).

The inner product is a *symmetric positive-definite bilinear form* for $\mathbb{F} = \mathbb{R}$, and a *positive-definite sesquilinear form* for $\mathbb{F} = \mathbb{C}$.

Inner Product Space A vector space equipped with an inner product.

Euclidean Space An inner product space over $\mathbb{R}$.

Matrix Representation of Inner Product A matrix $M \in \mathcal{M}_{n,n}$ represents the inner product

$$\langle u, v \rangle = u^\top M v$$

for $u, v \in \mathbb{R}^n$ iff. M is symmetric and positive-definite; i.e.,
- $M^\top = M$
- $v \neq 0 \Longrightarrow v^\top M v > 0$

For $\mathbb{F} = \mathbb{C}$, replace "transposition" $x^\top$ with "conjugate transposition" $x^\dagger = \overline{x}^\top$ and these conditions remain true.

I.e., (in finite dimension), every inner product can be represented as a matrix, and every matrix satisfying said conditions represents an inner product.

$\langle \cdot, \cdot \rangle_M$ denotes the inner product represented by M.

6.1.1 Infinite Dimension

Let's for example consider functions $f, g : \mathbb{R} \to \mathbb{R}$ on the interval $[a, b]$.

The *canonical* inner product is

$$\langle f, g \rangle = \int_a^b f(x) g(x) \, \mathrm{d}x$$

Symmetry comes from the commutativity of multiplication;
linearity is inherited from that of integration;
positive-definiteness is inherited from $\mathbb{R}$ as a group.

6.2 Norms

Norm A function $\|\cdot\| : V \to \mathbb{R}$ satisfying:

1. Subadditivity (triangle inequality): $\|v + w\| \leq \|v\| + \|w\|$
2. Absolute homogeneity: $\|av\| = |a| \, \|v\|$
3. Positive-definiteness: $u \neq 0 \Longrightarrow \|u\| > 0$

Induced Norm Every inner product naturally *induces* a norm via

$$\|u\| = \sqrt{\langle u, u \rangle}$$

Note that the proof of the triangle inequality relies on the Cauchy-Schwarz inequality (below).

The other direction, i.e., whether a norm induces an inner product, requires the *parallelogram law* $\|u + v\|^2 + \|u - v\|^2 = 2(\|u\|^2 + \|v\|^2)$.

Euclidean Norm The norm induced by the canonical inner product in Euclidean spaces. E.g.,

$$\left\| \begin{bmatrix} x \\ y \end{bmatrix} \right\| = \sqrt{\langle \begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} x \\ y \end{bmatrix} \rangle} = \sqrt{\begin{bmatrix} x \\ y \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix}} = \sqrt{x^2 + y^2}$$

Cauchy-Schwarz Inequality In any inner product space V,
$|\langle v, w \rangle|^2 \leq \langle v, v \rangle \langle w, w \rangle$
or, assuming the induced norm,
$|\langle v, w \rangle| \leq \|v\| \|w\|$.

Note that the Cauchy-Schwarz inequality is a consequence of the axioms of inner products; it is not an axiom itself, but it is a non-trivial result that helps prove, e.g., $\sqrt{\langle u, u \rangle}$ is indeed a norm.

6.3 Metrics

Metric (function) (Aka. distance) A function $d : A \times A \to \mathbb{R}$ satisfying:

1. Symmetry: $d(x, y) = d(y, x)$
2. Positive-definiteness: $d(x, y) \geq 0$ and $d(x, y) = 0 \Longleftrightarrow x = y$
3. Triangle inequality: $d(x, z) \leq d(x, y) + d(y, z)$

for all $x, y, z \in A$.

Metric Space A structure (A, d) where $d : A \times A \to \mathbb{R}$ is a metric.

Induced Metric Inner product spaces induce the metric $d(u, v) = \|u - v\|$.

Every inner product space is canonically a metric space using its induced norm and thus induced metric.

6.4 Orthogonality

Orthogonality For an inner product space, two vectors are orthogonal $v \perp w$ iff. $\langle v, w \rangle = 0$.

Generalized Pythagorean Theorem For an inner product space with the induced norm $\|\cdot\|$,

$$v \perp w \quad \Longrightarrow \quad \|v + w\|^2 = \|v\|^2 + \|w\|^2$$

Orthonormal Basis A basis where vectors are pairwise orthogonal and of unit norm.

The coordinates of a vector in an orthonormal basis are easily computed via the inner product; for a basis $(b_1, ..., b_n)$ and a vector v, $v = \langle v, b_1 \rangle b_1 + ... + \langle v, b_n \rangle b_n$.

Every orthonormal family of vectors is linearly independent, and every finite-dimensional inner product space has an orthonormal basis.

Orthogonal Complementary Subspace A given inner product subspace $W \subseteq V$ has a unique orthogonal complement (aka., the canonical complement)

$$W^\perp = \{ v \in V \mid \langle v, w \rangle = 0 \; \forall w \in W \}$$

Clearly,
$V^\perp = \{0\}$ and $\{0\}^\perp = V$
$V = W \oplus W^\perp$

For an orthonormal basis $(w_1, ..., w_k)$ of $W \subseteq V$, we can define the *orthogonal projection* s.t.

$$P_W(v) = \langle v, w_1 \rangle w_1 + ... + \langle v, w_k \rangle w_k$$

Clearly, $P_W(v) \in W$ and $v - P_W(v) \in W^\perp$.

Orthogonal Matrix A matrix $M \in \mathcal{M}_{n,n}$ is orthogonal iff. its columns form an orthonormal basis. Alternatively, $M^\top M = I$ or $M^\top = M^{-1}$.

Real Spectral Theorem For a real symmetric matrix M,
- M has a full eigenbasis
- The eigenbasis can be chosen to be orthonormal

Consequently, M is orthogonally diagonalizable: there exists an orthogonal matrix P s.t. $D = P^\top M P$ is diagonal.

For **real symmetric matrices**,
- eigenvectors are orthogonal (and eigenvalues are real)
- positive-definiteness $\Longleftrightarrow$ (strictly) positive eigenvalues (**useful property**)

This eigenvalue condition for positive-definiteness allows the *easier* classification of matrices as inner products.

Recall that matrix representations of inner products are symmetric;
they can thus always be diagonalized.
Furthermore, positive eigenvalues allows normalization of an diagonal form into the identity matrix, so **every inner product can be represented as the canonical inner product in some basis**.

Advantages of an orthonormal basis:
1. Coordinates are easily computed via the inner product.
2. Inner products become the canonical dot product.
3. Orthogonal projections have a closed form (inner product).
4. Change of basis matrices between orthonormal bases are orthogonal.
5. A bunch more…

7 Numerical Methods

Numerical methods approximate continuous mathematical problems (such as finding roots, solving differential equations, optimizing functions, or integrating) using discrete algorithms.

7.1 Preliminaries and Error Analysis

Absolute Error Let x be the exact mathematical quantity and $\hat{x}$ be its numerical approximation. The absolute error is defined as $e = |x - \hat{x}|$.

Order of Convergence *q* Let a sequence (x_n) converge to a limit l. The sequence converges with **order of convergence** $q \geq 1$ and **asymptotic error constant** $M > 0$ if:

$$\lim_{n \to \infty} \frac{|x_{n+1} - l|}{|x_n - l|^q} = M$$

For $q = 1$, we require $M < 1$ for convergence (linear convergence).

The order q governs the asymptotic speed of convergence:
- Linear ($q = 1$): The error is reduced by a constant factor M at each step.
- Quadratic ($q = 2$): The distance to the limit is squared at each step, which roughly doubles the number of correct decimal digits per iteration.

The constant M acts as a scaling factor on the error propagation.

Asymptotic Error Bounding (Big-O Notation) In step-based or partition-based methods with step/subinterval size $h > 0$, we say the error $e(h)$ is $O(h^p)$ as $h \to 0$ if there exist constants $C > 0$ and $h_0 > 0$ such that:

$$|e(h)| \leq Ch^p \quad \forall h \in (0, h_0]$$

The integer $p \geq 1$ is the **order of accuracy** of the method.

An error bound of $O(h^p)$ implies that halving the step size h reduces the asymptotic error bound by a factor of 2^p. Equivalently, for a partition of $[a, b]$ into n subintervals where h is proportional to $\frac{1}{n}$, the error is $O\left(\frac{1}{n^p}\right)$.

7.2 Newton's Method

Newton's Method is an iterative numerical technique to find roots/zeroes of a differentiable function $f : \mathbb{R} \to \mathbb{R}$. The goal is to find values a such that $f(a) = 0$.

Geometrically, each iteration approximates the curve by its tangent line at the current point, and finds the intersection of this tangent with the x-axis.

For an initial guess $x_0 \in \mathbb{R}$ and a desired tolerance $\delta > 0$, while $|f(x_n)| > \delta$, compute:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

Fixed Point Characterization Let f be differentiable with $f'(x) \neq 0$ for all x. A point l satisfies $f(l) = 0$ iff. l is a fixed point of the sequence-defining function $g : x \mapsto x - \frac{f(x)}{f'(x)}$.

Convergence to Zero If the sequence (x_n) defined by Newton's method converges to a limit l, f' is continuous on an interval containing l, and $f'(l) \neq 0$, then $f(l) = 0$.

- **Sufficient Conditions for Local Convergence**: If $f(a) = 0$, f' is continuous on an open interval I containing a, and $f'(a) \neq 0$, then there exists an interval $J \subset I$ containing a such that for any starting point $x_0 \in J$, the sequence (x_n) converges to a.
- **Quadratic Convergence**: If $f(a) = 0$, $f'(a) \neq 0$, and f is twice continuously differentiable ($f \in \mathcal{C}^2$) in a neighborhood of a, then Newton's method converges quadratically (with order $q = 2$). As established in the preliminaries, this means the distance to the root is squared at each step, doubling the number of correct digits.
- **Limitations**:
  - Newton's method fails to converge if $f'(x_n) = 0$ for some iteration.
  - Convergence is highly sensitive to the initial guess x_0. Poor choices can result in oscillation or divergence (e.g., $f(x) = |x|^{0.25}$).

7.3 Euler's Method

Euler's Method is a first-order numerical procedure for solving ordinary differential equations (ODEs). Consider an ODE of the form:

$$y' = f'(x) = G(x, f(x)) \quad \text{with initial condition} \quad f(x_0) = y_0$$

To approximate the exact solution f, we partition the domain with step size $h > 0$ and grid points $x_k = x_0 + kh$. We approximate the values $y_k \approx f(x_k)$ recursively via:

$$y_{k+1} = y_k + hG(x_k, y_k)$$

- **Error Analysis via Taylor Expansion**: Let f be the exact solution to the ODE. Assuming $f \in \mathcal{C}^2$, its Taylor expansion is:

  $$f(x_0 + h) = f(x_0) + hf'(x_0) + \frac{1}{2}h^2 f''(x_0) + O(h^3)$$

  Using the ODE relation $f'(x_0) = G(x_0, f(x_0))$ and substituting $y_0 = f(x_0)$:

  $$f(x_0 + h) = y_0 + hG(x_0, y_0) + \frac{1}{2}h^2 f''(x_0) + O(h^3)$$

  Comparing this to a single Euler step $y_1 = y_0 + hG(x_0, y_0)$, we obtain the **local truncation error**:

  $$e_1 = f(x_0 + h) - y_1 = \frac{1}{2}h^2 f''(x_0) + O(h^3)$$

Local Truncation Error (LTE) The error introduced by a single step assuming the previous step was exact. For Euler's method, LTE is $O(h^2)$ asymptotically.

Global Truncation Error (GTE) The cumulative error over the entire integration interval $[x_0, x_N]$. Since the number of steps $N = \frac{x_N - x_0}{h}$ is proportional to $\frac{1}{h}$, the errors propagate and accumulate. Consequently, Euler's method is a first-order method with a GTE of $O(h)$.

Thus, Euler's method is a first-order method: halving the step size h roughly halves the global approximation error.

7.4 Gradient Descent

Gradient Descent is a first-order iterative optimization algorithm for finding local minima of a differentiable multivariate function $f : \mathbb{R}^d \to \mathbb{R}$.

Because the gradient $\nabla f(x)$ points in the direction of steepest **ascent**, moving in the opposite direction $-\nabla f(x)$ slides down the function's surface.

Choose a starting point $x_0 \in \mathbb{R}^d$, learning rate (step size) $r > 0$, and convergence tolerance $\delta > 0$. Iterate:

$$x_{n+1} = x_n - r\nabla f(x_n)$$

Exit the loop when $\|x_{n+1} - x_n\| < \delta$.

- **Convergence Analysis**: Let the gradient of f be L-Lipschitz continuous; i.e., there exists $L > 0$ such that:

  $$\|\nabla f(y) - \nabla f(x)\| \leq L\|y - x\| \quad \forall x, y \in \mathbb{R}^d$$

  Convergence Rate Under the Lipschitz condition, if the learning rate satisfies $r \leq \frac{1}{L}$, then after k iterations:

  $$f(x_k) - f(x^*) = O\left(\frac{1}{k}\right)$$

  where x^* is a local minimum. This represents a sublinear rate of convergence, which is slower than linear convergence.

- **Remarks & Limitations**:
  - The behavior depends heavily on the learning rate r. If r is too large (e.g., $r > \frac{2}{L}$), the steps overshoot and can cause divergence. If r is too small, convergence is excessively slow.
  - In non-convex landscapes, gradient descent can easily get trapped in sub-optimal local minima or stall at saddle points.
- **Common Extensions**:

  Stochastic Gradient Descent (SGD) Approximates the true gradient using a subset of data or adds controlled noise to the gradient updates. This enables the algorithm to escape local minima and saddle points.

  Momentum (Heavy Ball Method) Accelerates convergence by incorporating a fraction of the previous step's update, mimicking physical momentum:

  $$v_{n+1} = \beta v_n - r\nabla f(x_n) \quad \text{where } \beta \in [0, 1)$$
  $$x_{n+1} = x_n + v_{n+1}$$

7.5 Numerical Integration

Numerical Integration methods approximate definite integrals when the antiderivative of the integrand is analytically intractable. The objective is to approximate the definite integral:

$$I = \int_a^b f(x)\,\mathrm{d}x$$

Let the interval $[a, b]$ be partitioned into n equal subintervals of width $h = \frac{b-a}{n}$. Let $x_k = a + kh$ for $k = 0, 1, \ldots, n$.

7.5.1 Riemann Sums & Trapeziums

Midpoint Rule Approximates f on each subinterval $[x_k, x_{k+1}]$ by a constant value evaluated at the midpoint $m_k = x_k + \frac{h}{2}$:

$$I_n^M = h\sum_{k=0}^{n-1} f(m_k)$$

Midpoint Error Bound If f is twice continuously differentiable ($f \in \mathcal{C}^2([a, b])$) and $|f''(x)| \leq M$ for all $x \in [a, b]$, then the approximation error satisfies:

$$|I - I_n^M| \leq M\frac{(b-a)^3}{24n^2}$$

This indicates quadratic convergence ($O\left(\frac{1}{n^2}\right)$ or $O(h^2)$) as defined in the preliminaries.

Trapezium Rule Approximates f on each subinterval by a linear interpolant (piecewise linear approximation):

$$I_n^T = h\sum_{k=0}^{n-1} \frac{f(x_k) + f(x_{k+1})}{2} = h\left(\frac{f(a) + f(b)}{2} + \sum_{k=1}^{n-1} f(x_k)\right)$$

Trapezium Error Bound If $f \in \mathcal{C}^2([a, b])$ and $|f''(x)| \leq M$ for all $x \in [a, b]$, then the approximation error satisfies:

$$|I - I_n^T| \leq M\frac{(b-a)^3}{12n^2}$$

The trapezium rule also achieves quadratic convergence ($O\left(\frac{1}{n^2}\right)$), but carries a larger error bound constant than the midpoint rule.

Simpson's Rule Approximates f on pairs of subintervals using quadratic interpolants (piecewise quadratic approximation). For an even number of subintervals n:

$$I_n^S = \frac{h}{3}\left[f(a) + f(b) + 4\sum_{j=1}^{\frac{n}{2}} f(x_{2j-1}) + 2\sum_{j=1}^{\frac{n}{2}-1} f(x_{2j})\right]$$

Simpson Error Bound If f is four-times continuously differentiable ($f \in \mathcal{C}^4([a, b])$) and $|f^{(4)}(x)| \leq M_4$ for all $x \in [a, b]$, then the approximation error satisfies:

$$|I - I_n^S| \leq M_4\frac{(b-a)^5}{180n^4}$$

This yields a higher order of accuracy of $O\left(\frac{1}{n^4}\right)$ or $O(h^4)$.

This $O(h^4)$ convergence rate makes Simpson's rule highly efficient: doubling the number of intervals reduces the error bound by a factor of 16.

7.5.2 Monte Carlo

Monte Carlo Integration A randomized algorithm for estimating integrals, particularly valuable in multi-dimensional domains where standard grid-based methods suffer from the **curse of dimensionality**.

- **Algorithm (Geometric Interpretation)**:
  1. Bound the integration region within a rectangle $D = [a, b] \times [y_{\min}, y_{\max}]$ such that $y_{\min} \leq f(x) \leq y_{\max}$ for all $x \in [a, b]$.
  2. Generate N independent and identically distributed (i.e. i.i.d.) random points (X_i, Y_i) uniformly on D.
  3. Count the number of points C that fall under the curve (i.e., $Y_i \leq f(X_i)$, assuming $f(x) \geq 0$).
  4. Estimate the integral by scaling the bounding area:

  $$I \approx \text{Area}(D) \times \frac{C}{N}$$

- **Probabilistic Derivation via Expectation**: By the Law of Large Numbers, the integral can be estimated as:

  $$I = (b-a)E[f(X)] \approx \frac{b-a}{N}\sum_{i=1}^{N} f(X_i)$$

  where $X_i \sim U(a, b)$ are i.i.d. random variables.
- **Convergence Rate**: By the Central Limit Theorem, the error converges at a rate of $O\left(\frac{1}{\sqrt{N}}\right)$ regardless of the dimensionality d of the integration domain.

  While $O\left(\frac{1}{\sqrt{N}}\right)$ is slow compared to $O\left(\frac{1}{n^2}\right)$ or $O\left(\frac{1}{n^4}\right)$ in 1D, Monte Carlo remains highly scalable for high-dimensional integration where standard grid methods require $O(n^d)$ points.

8 Probabilities

8.1 Basics

Random Variable A *measurable* function $X : \Omega \to T$.

Ω belongs to a *probability triple* $(\Omega, \mathcal{F}, P)$ where the *probability measure* $P : \mathcal{F} \to [0, 1]$ assigns probabilities to *events* $a \in \mathcal{F} \subseteq \mathcal{P}(\Omega)$. Naturally, an event, being a subset of outcomes, is considered to have occurred iff. the actual outcome is an element of it.

Conceptually, a random variable maps abstract outcomes to concrete "numeric" values. Measurability ensures this mapping is well-behaved wrt. the structure of $\mathcal{F}$ and P, and is analogous to continuity in standard analysis.

Study *measure theory* for further rigor.

Sample Space Ω, the set of possible outcomes (symbolic).
Target Space T, a measurable space (numeric repr. of outcomes). Often, $T = \mathbb{R}^n$.

X may be discrete or continuous.

It is important to realize that random variables can be freely *transformed*:
$Y = f(X)$ is a new random variable for any measurable function $f : T \to T'$;
and *combined*:
$Z = g(X, Y)$ is a new random variable for any measurable function $g : T \times T \to T'$.

Continuous functions on $\mathbb{R}^n$ are measurable. In fact, you'd rarely encounter a non-measurable function unless you're looking for one.

Probability Density Function The PDF of a random variable is a function $f : T \to \mathbb{R}$ mapping a given target space outcome to its *relative* probability.
A PDF $f : T \to \mathbb{R}$ must satisfy
1. $f(x) \geq 0$
2. $\int_T f(x)\,\mathrm{d}x = 1$

PDFs are non-unique wrt. a given random variable.
A random variable can be "reconstructed" from a PDF, also non-uniquely.

Concretely, a PDF specifies, via its integral, the probability of the random variable taking values within a given interval; if a random variable X has PDF f, then for any $a, b \in T$ with $a \leq b$,

$$P(a \leq X \leq b) = \int_a^b f(x)\,\mathrm{d}x$$

Clearly, $P(X = a) = 0$ for any a, and the strictness of the inequalities does not matter.

Cumulative Distribution Function The CDF of a random variable X is the function $F : T \to \mathbb{R}$ s.t.

$$F(x) = P(X \leq x)$$

If X has a PDF f, then also

$$F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t$$

The association between a random variable and its CDF is unique.

Inverse CDF For a "distribution" $\mathcal{X}$, we notate its inverse CDF as $\mathcal{X}[c]$. I.e.,
$\mathcal{X}[c] = x \iff P(X \leq x) = c$.

Expected Value (Mean) For a random variable X with PDF f, the expected value is

$$\mu = E[X] = \int_T x f(x)\,\mathrm{d}x$$

If X is discrete, then $E[X] = \sum_T x P(X = x)$.

Note that the expected value (and thus variance) can be defined for any random variable, even without a PDF, but this requires measure theory; for any random variable X on $(\Omega, \mathcal{F}, P)$,

$$E[X] = \int_\Omega X(\omega)\,\mathrm{d}P(\omega)$$

In the following section on statistics, this justifies the computation of means and variances on arbitrary statistical samples (which are assumed to be of some unspecified random variable).

Covariance For random variables X and Y with means μ_X and μ_Y and a *joint* PDF $f_{X,Y}$,

$$\begin{aligned}\mathrm{Cov}(X, Y) &= E[(X - E[X])(Y - E[Y])] \\ &= E[XY] - E[X]E[Y] \\ &= \int_T \int_T (x - \mu_X)(y - \mu_Y) f_{X,Y}(x, y)\,\mathrm{d}x\,\mathrm{d}y\end{aligned}$$

The "product" XY is a new random variable of the pointwise product of X and Y, i.e., $(XY)(\omega) = X(\omega)Y(\omega)$ for all $\omega \in \Omega$.

But of course this doesn't directly yield useful computational formulas... so you should learn measure theory.

Variance For a random variable X with PDF f and mean μ,

$$\begin{aligned}\mathrm{Var}(X) &= \mathrm{Cov}(X, X) \\ &= E[(X - E[X])^2] \\ &= E[X^2] - E[X]^2 \\ &= \int_T x^2 f(x)\,\mathrm{d}x - \mu^2 = \int_T (x - \mu)^2 f(x)\,\mathrm{d}x\end{aligned}$$

Covariance measures the linear dependence between X and Y;
variance measures the spread of variable around its mean.

Law of the Unconscious Statistician For a random variable X with PDF f and any function $g : T \to U$ (U may be different target space),

$$E[g(X)] = \int_T g(x) f(x)\,\mathrm{d}x$$

This follows from the definition of the expected value but is not immediately obvious.
This explains why $E[X^2] = \int_T x^2 f(x)\,\mathrm{d}x$ above.

Standard Deviation $\sigma(X) = \sqrt{\mathrm{Var}(X)}$

If X and Y are independent, then $\mathrm{Cov}(X, Y) = 0$.

Independence X and Y are independent iff. their joint distribution factorizes; $f_{X,Y}(x, y) = f_X(x) f_Y(y)$. I.e.,

$$P(X \leq x \cap Y \leq y) = P(X \leq x) P(Y \leq y).$$

Equivalently for discrete variables, $P(X = x \cap Y = y) = P(X = x)P(Y = y)$.

Also, $E[XY] = E[X]E[Y]$ and thus $\mathrm{Cov}(X, Y) = 0$.

8.2 Classic Distributions

8.2.1 Exponential

$$\begin{aligned}X &\sim \mathrm{Exp}(\lambda) && \{\lambda \in \mathbb{R}^+\} \\ f(x) &= \lambda e^{-\lambda x} && \{x \geq 0\} \\ T &= [0, \infty) \\ E[X] &= \frac{1}{\lambda} \\ \mathrm{Var}(X) &= \frac{1}{\lambda^2}\end{aligned}$$

8.2.2 Uniform

$$\begin{aligned}X &\sim U(a, b) && \{a < b \in \mathbb{R}\} \\ f(x) &= \frac{1}{b - a} && \{a \leq x \leq b\} \\ T &= [a, b] \\ E[X] &= \frac{a + b}{2} \\ \mathrm{Var}(X) &= \frac{(b - a)^2}{12}\end{aligned}$$

8.2.3 Normal

$$\begin{aligned}X &\sim N(\mu, \sigma^2) && \{\mu \in \mathbb{R}, \sigma \in \mathbb{R}^+\} \\ f(x) &= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \\ T &= \mathbb{R} \\ E[X] &= \mu \\ \mathrm{Var}(X) &= \sigma^2\end{aligned}$$

Aka. Gaussian distribution.
No closed-form CDF.

8.2.4 Student's t-Distribution

$$\begin{aligned}X &\sim t(\nu) && \{\nu \in \mathbb{R}^+\} \\ f(x) &= \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \\ T &= \mathbb{R} \\ E[X] &= \begin{cases} 0 & \text{for } \nu > 1 \\ \text{undefined otherwise} \end{cases} \\ \mathrm{Var}(X) &= \begin{cases} \frac{\nu}{\nu-2} & \text{for } \nu > 2 \\ \infty & \text{for } 1 < \nu \leq 2 \\ \text{undefined otherwise} \end{cases}\end{aligned}$$

Its sole parameter ν is called the *degrees of freedom* and is usually an integer.

Similar to the normal distribution but with heavier tails.
As $\nu \to \infty$, $t(\nu) \to N(0, 1)$.

Definition via χ^2 and normal distributions:

$$\frac{Z}{\sqrt{V/\nu}} \sim t(\nu)$$

where $Z \sim N(0, 1)$ and $V \sim \chi^2(\nu)$ (and are independent).

8.2.5 Chi-Squared

$$\begin{aligned}X &\sim \chi^2(k) && \{k \in \mathbb{Z}^+\} \\ f(x) &= \frac{x^{\frac{k}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{k}{2}}\,\Gamma(k/2)} && \{x > 0\} \\ T &= [0, \infty) \\ E[X] &= k \\ \mathrm{Var}(X) &= 2k\end{aligned}$$

Its sole parameter k is called the *degrees of freedom.*

$$\sum_{i=1}^{k} Z_i^2 \sim \chi^2(k)$$

where $Z_i \overset{\text{iid}}{\sim} N(0, 1)$.

$\chi^2(k) = \mathrm{Gamma}\left(\alpha = \frac{k}{2}, \theta = 2\right)$

8.3 Multivariate

...joint probability distributions...

$$P(X = a, Y = b) = P(X = a \cap Y = b)$$

Joint distro of two continuous variables:

$$P(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \int_c^d f_{X,Y}(x, y)\,\mathrm{d}y\,\mathrm{d}x$$

for the *joint* PDF $f_{X,Y}$ of X and Y.

Marginal Distributions Given a multivariate distribution, the single-variate distribution of one variable is obtained by "integrating out" (summing over) all other variables. E.g. for the joint PDF $f_{X,Y}$,

$$\begin{aligned}f_X(x) &= \int_T f_{X,Y}(x, y)\,\mathrm{d}y \\ f_Y(y) &= \int_T f_{X,Y}(x, y)\,\mathrm{d}x\end{aligned}$$

This "recovers" the individual distributions.

Conditional Probabilities

$$\begin{aligned}P(X \leq x \mid Y = y) &= \frac{P(X \leq x \cap Y = y)}{P(Y = y)} \\ &= \frac{\int_{-\infty}^{x} f_{X,Y}(t, y)\,\mathrm{d}t}{f_Y(y)}\end{aligned}$$

8.4 Useful Stuff

Law of Large Numbers For a sequence of *independent and identically distributed* (i.i.d.) random variables, the average converges to the individual expected value as the number of variables goes to infinity.
Formally, if $X_1, X_2, \ldots, X_n$ are i.i.d. with expected value μ, then

$$P\left(\lim_{n \to \infty} \overline{X} = \mu\right) = 1$$

where $\overline{X} = \frac{1}{n}(X_1 + \ldots + X_n)$.

Central Limit Theorem (CLT) For i.i.d.s $X_1, X_2, \ldots, X_n$ with mean μ and *finite* variance σ^2, the standardized sample mean converges in distribution to a standard normal:

$$\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \overset{\sim}{\to} N(0, 1) \quad \text{as } n \to \infty$$

where $\overline{X}_n = \frac{1}{n}(X_1 + \ldots + X_n)$.

Equivalently, for large (finite) n, $\overline{X}_n$ is approximately $N(\mu, \sigma^2/n)$-distributed.
We consider the approximation valid for $n \geq 30$.

9 Statistics

Statistics can be seen as the physical manifestation of associated theoretical concepts in probability.

> Interesting: see *frequentist* vs *Bayesian* interpretations of probability.

Model (Statistical Model) The probability distribution (or family of distributions) from which a sample is drawn (aka. the *underlying parent distribution*).

When describing a physical process, the model is almost always an assumption.

Sample (Random Sample) A sample $\boldsymbol{X}$ of size n is a sequence of random variables $\boldsymbol{X} = (X_1, ..., X_n)$ drawn from the model.
A *realized* sample is an observation of $\boldsymbol{X}$, i.e. a sequence of real values $\boldsymbol{x} = (x_1, ..., x_n)$ where x_i is the observed realization of X_i.

Simple Random Sample (SRS) A sample where X_i are i.i.d.
E.g., a sample drawn uniformly at random from the model, with replacement.

Clearly, the CLT can be used to approximate the mean of SRSs.

9.1 The Normal Distribution

The normal distribution frequently appears in various statistical processes.
Different processes may produce different normal distributions, but they can be normalized to the *standard* normal distribution $N(\mu = 0, \sigma^2 = 1)$.

Normalization For some $X \sim N(\mu, \sigma^2)$, the normalized variable $Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$.

z	$P(\lvert Z\rvert \leq z)$
1	68.27%
1.96	95.00%
2	95.45%
3	99.73%

9.2 Parameter Estimation

To use a statistical model to represent a physical random process, we must estimate its defining parameters θ from sample data. E.g., under a normal model, we estimate the mean μ and variance σ^2; under an exponential model, we estimate the rate λ.

> If a realized sample of n observations is drawn from the distribution of one random variable X, we can model the sample prior to realization as a random sample: a sequence of i.i.d.s $X_1, ..., X_n$ sharing the same distribution as X.
> Per the CLT, the distribution of the sample mean $\overline{X}$ is approximated by a normal distribution (for sufficient n).
> It is therefore useful to estimate the μ and σ^2 of an arbitrarily-distributed variable X, as these two parameters directly dictate the behavior of the sample mean even when X itself is not normal.

Statistic Any function T that maps a random sample $\boldsymbol{X}$ to a *statistic* (a new random variable).
When evaluated on a realized sample, $T(\boldsymbol{x})$ yields a real value.

E.g., the sample mean estimator $\overline{X} = \frac{1}{n}\sum_i X_i$ is a random variable, while the realized sample mean $\overline{x} = \frac{1}{n}\sum_i x_i$ is a fixed number.

Estimator An estimator $T(\boldsymbol{X})$ is a statistic used to estimate a parameter θ of the underlying distribution. The realized value $\hat{\theta} = T(\boldsymbol{x})$ is an **estimate**.

Bias An estimator $T(\boldsymbol{X})$ for θ is unbiased iff. $E[T(\boldsymbol{X})] = \theta$.

E.g. the sample mean estimator $\overline{X}$ is an unbiased estimator of the true mean μ since $E[\overline{X}] = \mu$.

Sampling Distribution The sampling distribution of an estimator $T(\boldsymbol{X})$ is its probability distribution as a random variable.
Equivalently, it is the distribution of realized estimates $T(\boldsymbol{x})$ across all possible realizations of the random sample $\boldsymbol{X}$.

Knowledge of the sampling distribution of an estimator allows us to e.g. compute *confidence intervals* for the estimate, and thus is a crucial part of statistics.

Note that statistics and their sampling distributions are usually tied to the sample size n.

> **Normality Assumption**: The exact sampling distributions of the estimated mean (Student's t) and variance (χ^2) below assume a normal model, i.e., $X_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$. For non-normal models, these exact finite-sample results do not hold, although the CLT does permit the normal approximation of the sample mean for large sample sizes.

9.2.1 Mean

Mean Estimator The canonical sample mean estimator is the arithmetic mean:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

It is **unbiased**:

$$E[\overline{X}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}(n\,E[X]) = \frac{1}{n}(n\mu) = \mu$$

Also, $\mathrm{Var}(\overline{X}) = E\left[\left(\overline{X} - \mu\right)^2\right] = \frac{\sigma^2}{n}$.

Recall that in general, $E[\overline{X}] = \mu$ and $\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}$.
Under the parent normality assumption $X_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, the exact sampling distribution of the sample mean is

$$\overline{X} \sim N(\mu, \sigma^2/n)$$

which can be standardized:

$$Y = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \sqrt{n}\frac{\overline{X} - \mu}{\sigma} \sim N(0, 1)$$

...and rearranged to yield the realized $(1 - \alpha)$ confidence interval for μ:

$$\left[\overline{x} + z\left[\frac{\alpha}{2}\right]\frac{\sigma}{\sqrt{n}}, \overline{x} + z\left[1 - \frac{\alpha}{2}\right]\frac{\sigma}{\sqrt{n}}\right] \quad \text{or} \quad \left[\overline{x} - \left|z\left[\frac{\alpha}{2}\right]\right|\frac{\sigma}{\sqrt{n}}, \overline{x} + \left|z\left[\frac{\alpha}{2}\right]\right|\frac{\sigma}{\sqrt{n}}\right]$$

($z[c]$ is the inverse CDF of the standard normal distribution.)

Studentized Mean Often, the true variance σ^2 required in the above interval is unknown. If σ is substituted with the unbiased sample standard deviation S, the resulting statistic instead follows Student's t-distribution with $n - 1$ degrees of freedom:

$$T = \sqrt{n}\frac{\overline{X} - \mu}{S} \sim t(n - 1)$$

> Because we are using an estimator S instead of a constant σ, T is the quotient of a normal distribution (the mean estimator) and the square root of a chi-squared distribution (the variance estimator, see below), which results in a t-distribution. The t-distribution exhibits heavier tails than the normal distribution to account for this uncertainty in the variance.

Using this exact sampling distribution, the $(1 - \alpha)$ realized confidence interval for μ is now

$$\left[\overline{x} - \left|t_{n-1}\left[\frac{\alpha}{2}\right]\right|\frac{s}{\sqrt{n}}, \overline{x} + \left|t_{n-1}\left[\frac{\alpha}{2}\right]\right|\frac{s}{\sqrt{n}}\right]$$

where $\overline{x}$ and s are the mean and standard deviation of the realized sample.

Clearly, $\left|t_{n-1}\left[\frac{\alpha}{2}\right]\right| \geq \left|z\left[\frac{\alpha}{2}\right]\right|$, so the realized confidence intervals are wider when the variance is also estimated.

9.2.2 Variance

Variance Estimator The estimator for sample variance following the discrete formula is

$$\tilde{S}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

However, $\tilde{S}^2$ is a **biased** estimator of the population variance σ^2 (and thus called *uncorrected*):

$$E\left[\tilde{S}^2\right] = \text{...complex derivation...} = \frac{n-1}{n}\sigma^2$$

> The bias arises because the sample mean $\overline{X}$ is used in place of the unknown true population mean μ. The sample observations X_i are closer to their own sample mean $\overline{X}$ than to the true mean μ, causing $\tilde{S}^2$ to systematically underestimate the true variance.

Bessel's Correction $\tilde{S}^2$'s bias is corrected by scaling by $\frac{n}{n-1}$, yielding the **unbiased sample variance estimator**:

$$S^2 = \frac{n}{n-1}\tilde{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

Now,

$$E[S^2] = \frac{n}{n-1}E\left[\tilde{S}^2\right] = \frac{n}{n-1}\left(\frac{n-1}{n}\sigma^2\right) = \sigma^2$$

We denote the corresponding realized estimates (calculated from a realized sample $\boldsymbol{x}$) as $\tilde{s}^2$ and s^2.

Under the parent normality assumption ($X_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$), the sampling distribution of S^2 is related to the chi-squared χ^2 distribution.
Specifically, the standardized statistic K follows a χ^2 distribution with $n - 1$ degrees of freedom:

$$K = \frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \sim \chi^2(n-1)$$

(Note that χ^2 is asymmetric.)
For a confidence level of $1 - \alpha$, we have:

$$P\left(\chi^2_{n-1}\left[\frac{\alpha}{2}\right] \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{n-1}\left[1 - \frac{\alpha}{2}\right]\right) = 1 - \alpha$$

$$P\left(\frac{(n-1)S^2}{\chi^2_{n-1}\left[1 - \frac{\alpha}{2}\right]} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{n-1}\left[\frac{\alpha}{2}\right]}\right) = 1 - \alpha$$

For a realized sample with unbiased variance estimate s^2, the confidence interval for σ^2 is

$$\left[\frac{(n-1)s^2}{\chi^2_{n-1}\left[1 - \frac{\alpha}{2}\right]}, \frac{(n-1)s^2}{\chi^2_{n-1}\left[\frac{\alpha}{2}\right]}\right]$$

To obtain the realized confidence interval for the standard deviation σ, we simply take the square root of the bounds.

9.3 Hypothesis Testing

> Hypothesis testing provides a structured statistical framework to make decisions about a model parameter using sample data.
>
> A *null hypothesis* making a claim about a parameter is stated, then a test statistic is computed from a realized sample, and finally a p-value is calculated to quantify the strength of evidence against the null hypothesis under the model's assumed sampling distribution.

9.3.1 Procedure

> The hypothesis testing framework operates on top of the model's assumptions. The hypotheses make assertions about the parameter θ itself, while the model provides the reference sampling distribution of the test statistic under H_0 (which is core to calculating p-values used to draw conclusions).
> While a rejection of H_0 technically rejects the *combination* of the hypothesis and the model, the framework assumes the model correct and attributes the rejection to H_0 alone. Similarly, failing to reject does *not* validate the model.

Let θ be the parameter of interest. Note that we do not know the true value of θ.

Null Hypothesis (H_0) The baseline or default assumption that $\theta = \theta_0$, an assumed value.

Alternative Hypothesis (H_1) A claim contradicting H_0 that we wish to seek statistical evidence for.
There are three common forms of H_1:
- **Two-tailed**: $\theta \neq \theta_0$ (true θ is different from θ_0)
- **Left-tailed**: $\theta < \theta_0$ (true θ is smaller than θ_0)
- **Right-tailed**: $\theta > \theta_0$ (true θ is larger than θ_0)

Test Statistic A statistic T whose sampling distribution under H_0 is (completely) specified by the model. Its evaluation on a realized sample yields an observed value $t_0 = T(\boldsymbol{x})$.

The test statistic is typically built from an estimator of θ, re-centered and rescaled under H_0.

> E.g., for a normal model ($X_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$) with unknown variance, the appropriate test statistic for the mean $H_0 : \mu = \mu_0$ is
>
> $$T(\boldsymbol{X}) = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} = \sqrt{n}\frac{\overline{X} - \mu_0}{S} \sim t(n-1) \quad \text{under } H_0$$
>
> Evaluating this on our realized sample yields the observed value $t_0 = \sqrt{n}\frac{\overline{x} - \mu_0}{s}$.

***p*-value** The probability, under H_0, of obtaining a test statistic T at least as extreme as the observed value t_0. Clearly, this depends on the type of H_1:
- For **two-tailed**: $p = P(|T| \geq |t_0|)$
- For **left-tailed**: $p = P(T \leq t_0)$
- For **right-tailed**: $p = P(T \geq t_0)$

> E.g., (continuing with the above example) Under the Student's $t(n-1)$ distribution of T under H_0:
> - Two-tailed: $p = 2P(T \geq |t_0|)$ (by symmetry of the t-distribution)
> - Left-tailed: $p = P(T \leq t_0)$
> - Right-tailed: $p = P(T \geq t_0)$

Significance Level (α) The maximum tolerated probability of a false positive (i.e., type I error); used to determine the threshold for rejecting H_0.
- $p < \alpha$: we **reject $\boldsymbol{H_0}$** in favor of H_1 (the observed effect is statistically significant).
- $p \geq \alpha$: we **fail to reject $\boldsymbol{H_0}$** (we cannot conclude anything significant; we do not "accept" H_0, but rather find insufficient evidence to discard it).

α is commonly 5%.

9.4 Bootstrapping

> Traditional statistical inference relies on parametric assumptions (e.g., population normality) or large-sample asymptotics (the CLT) to analytically derive the sampling distribution of an estimator. However, in many practical scenarios, these models fail:
> - The underlying distribution may be highly skewed, multi-modal, or simply non-normal (e.g., alcohol consumption or wealth distributions).
> - The sample size n may be too small for asymptotic theorems (like the CLT) to provide a reliable approximation.
> - Sourcing additional physical samples may be impossible or prohibitively expensive (e.g., a one-off historical survey).

Bootstrapping provides a computational method to estimate the sampling distribution of any statistic directly from a single observed sample.
The only assumption is that the observed sample is *representative* of the underlying distribution.

9.4.1 Bootstrap Algorithm

Given:
- An unknown population distribution with a true parameter of interest θ.
- An observed sample $\boldsymbol{x} = (x_1, ..., x_n)$ of size n.
- An statistic $T(\boldsymbol{X})$ estimating θ, yielding an initial estimate $\hat{\theta} = T(\boldsymbol{x})$.

Repeat B times:
1. Draw a uniform random sample $\boldsymbol{x}^* = (x_1^*, ..., x_n^*)$ of size n from $\boldsymbol{x}$ uniformly **with replacement**.
2. Compute the statistic $\hat{\theta}^* = T(\boldsymbol{x}^*)$.

This yields a sequence of B bootstrap estimates: $\hat{\theta}_1^*, \hat{\theta}_2^*, ..., \hat{\theta}_B^*$.
As $B \to \infty$, the empirical distribution of the bootstrap estimates converges to the true sampling distribution of $\hat{\theta}$, allowing statistical inference (e.g. confidence intervals) without parametric assumptions.

9.4.2 Constructing Bootstrap Confidence Intervals

Under the **Percentile Bootstrap Method**, we can construct a $(1 - \alpha)$ confidence interval for θ directly from the empirical distribution of our bootstrap estimates:
1. Sort the B bootstrap estimates $\hat{\theta}_i^*$ in increasing order.
2. Find the lower percentile index at $\frac{\alpha}{2}$ and the upper percentile index at $1 - \frac{\alpha}{2}$.
3. The resulting confidence interval is

$$\left[\hat{\theta}^*_{[B\,\frac{\alpha}{2}]}, \hat{\theta}^*_{[B\,(1-\frac{\alpha}{2})]}\right]$$