1. Calculus and Analysis

Definition 1.1 Quadratic Formula: $ax^2+bx+c=0$ or in reduced form:
$x^2+px+q=0$ with $p=b/a$ and $q=c/a$

Definition 1.2 Discriminant: $\delta = b^2-4ac$

Definition 1.3 Solution to definition 1.1:
$$x_\pm = \frac{-b\pm\sqrt{\delta}}{2a} \quad \text{or} \quad x_\pm = \frac{1}{2}\left(-p\pm\sqrt{p^2-4q}\right)$$

Theorem 1.1
Fist Fundamental Theorem of Calculus: Let f be a continuous real-valued function defined on a closed interval $[a,b]$.
Let F be the function defined $\forall x\in[a,b]$ by:
$$F(X)=\int_a^x f(t)\,\mathrm{d}t \tag{1.1}$$
Then it follows:
$$F'(x)=f(x) \qquad \forall x\in(a,b) \tag{1.2}$$

Theorem 1.2
Second Fundamental Theorem of Calculus: Let f be a real-valued function on a closed interval $[a,b]$ and F an antiderivative of f in $[a,b]$: $F'(x)=f(x)$, then it follows if f is Riemann integrable on $[a,b]$:
$$\int_a^b f(t)\,\mathrm{d}t = F(b)-F(a) \iff \int_a^x \frac{\partial}{\partial x}F(t)\,\mathrm{d}t = F(x) \tag{1.3}$$

Definition 1.4 Domain of a function $\mathtt{dom}(\cdot)$:
Given a function $f:\mathcal{X}\to\mathcal{Y}$, the set of all possible input values $\mathcal{X}$ is called the domain of f $\mathtt{dom}(f)$.

Definition 1.5 Codomain/target set of a function $\mathtt{codom}(\cdot)$:
Given a function $f:\mathcal{X}\to\mathcal{Y}$, the codaomain of that function is the set $\mathcal{Y}$ into which all of the output of the function is **constrained** to fall.

Definition 1.6 Image of a function: Given a function $f:\mathcal{X}\to\mathcal{Y}$, the image of that function is the set to which the function can actually map:
$$\{y\in\mathcal{Y}|y=f(x),\quad \forall x\in\mathcal{X}\}:=f[\mathcal{X}] \tag{1.4}$$
Hence it is a subset of a function's codomain.

Example 1.1 :
Given $f:\mathbb{R}\to\mathbb{R}$
defined by $f:x\mapsto x^2 \iff f(x)=x^2$
$\mathtt{dom}(f)=\mathbb{R}$, $\mathtt{codom}(f)=\mathbb{R}$ **but** its image is $f[\mathbb{R}]=\mathbb{R}_+$.

Image (Range) of a subset

nospacing The image of a subset $A\subseteq\mathcal{X}$ under f is the subset $f[A]\subseteq\mathcal{Y}$ defined by:
$$f[A]=\{y\in\mathcal{Y}|y=f(x),\quad \forall x\in A\} \tag{1.5}$$

Note: Range

The term range is ambiguous as it may refer to the image or the codomain, depending on the definition.
However, modern usage almost always uses range to mean image.

Definition 1.7 (strictly) Increasing Functions:
A function f is called **monotonically increasing/increasing/non-decreasing** if:
$$x\leqslant y \iff f(x)\leqslant f(y) \qquad \forall x,y\in\mathtt{dom}(f) \tag{1.6}$$
And **strictly increasing** if:
$$x<y \iff f(x)<f(y) \qquad \forall x,y\in\mathtt{dom}(f) \tag{1.7}$$

Definition 1.8 (strictly) Decreasing Functions:
A function f is called **monotonically decreasing/decreasing/non-increasing** if:
$$x\geqslant y \iff f(x)\geqslant f(y) \qquad \forall x,y\in\mathtt{dom}(f) \tag{1.8}$$
And **strictly decreasing** if:
$$x>y \iff f(x)>f(y) \qquad \forall x,y\in\mathtt{dom}(f) \tag{1.9}$$

Definition 1.9 Monotonic Function: A function f is called monotonic iff either f is increasing or decreasing.

Definition 1.10 Linear Function: A function $L:\mathbb{R}^n\mapsto\mathbb{R}^m$ is linear if and only if:
$$L(\boldsymbol{x}+\boldsymbol{y})=L(\boldsymbol{x})+L(\boldsymbol{y})$$
$$L(\alpha\boldsymbol{x})=\alpha L(\boldsymbol{x}) \qquad \forall \boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^n,\quad \alpha\in\mathbb{R}$$

Corollary 1.1 Linearity of Differentiation: The derivative of **any** linear combination of functions equals the same linear combination of the derivatives of the functions:
$$\frac{\mathrm{d}}{\mathrm{d}x}\left(af(x)+bg(x)\right)=a\frac{\mathrm{d}}{\mathrm{d}x}f(x)+b\frac{\mathrm{d}}{\mathrm{d}x}g(x) \qquad a,b\in\mathbb{R} \tag{1.10}$$

Definition 1.11 Convex Function:
A function $f:\mathbb{R}^n\to\mathbb{R}$ is convex if $\mathtt{dom}(f)$ is convex, and
$$f(\lambda x+(1-\lambda)y)\leqslant\lambda f(x)+(1-\lambda)f(y)$$
with $\forall x,y\in\mathtt{dom}(f)$, $\forall\lambda\in[0,1]$.
(Replacing $\leqslant$ with $<$gives *strict* convexity)

Corollary 1.2 Title: A twice differentiable function of one variable $f:\mathbb{R}\to\mathbb{R}$ is convex on an interval $\mathcal{X}=[a,b]$ if and only if its second derivative is non-negative on that interval $\mathcal{X}$:
$$\in\mathcal{C}^1(\mathbb{R})\ \text{convex} \iff f''(x)\geqslant 0 \qquad \forall x\in\mathcal{X} \tag{1.11}$$

Definition 1.12 Smoothness/Continuity $\mathcal{C}^k$: Given a function $f:\mathcal{X}\to\mathcal{Y}$, the function is said to be of class k if it is differentiable up to order k **and** continuous, on its entire domain:
$$f\in\mathcal{C}^k(\mathcal{X}) \iff \exists f',f'',\ldots,f^{(k)}\ \text{continuous} \tag{1.12}$$

Note

- The class $\mathcal{C}^0$ consists of all continuous functions.
- P.w. continuous $\neq$ continuous.
- A function of that is k times differentiable must at least be of class $\mathcal{C}^{k-1}$.
- $\mathcal{C}^m(\mathcal{X})\subset\mathcal{C}^{m-1},\ldots\mathcal{C}^1\subset\mathcal{C}^0$
- Continuity is implied by the differentiability of all **derivatives** of up to order $k-1$.

Corollary 1.3 Smooth Function $\mathcal{C}^\infty$: Is a function $f:\mathcal{X}\to\mathcal{Y}$ that has derivatives infinitely many times differerntiable.
$$f\in\mathcal{C}^\infty(\mathcal{X}) \iff f',f'',\ldots,f^{(\infty)} \tag{1.13}$$

Corollary 1.4 Continuously Differentiable Function $\mathcal{C}^1$: Is the class of functions that consists of all differentiable functions whose derivative is continuous.
Hence a function $f:\mathcal{X}\to\mathcal{Y}$ of the class must satisfy:
$$f\in\mathcal{C}^1(\mathcal{X}) \iff f'\ \text{continuous} \tag{1.14}$$

Functions

Even Functions: have rotational symmetry with respect to the origin.
⇒**Geometrically:** its graph remains unchanged after reflection about the y-axis.
$$f(-x)=f(x) \tag{1.15}$$
Odd Functions: are symmetric w.r.t. to the y-axis.
⇒**Geometrically:** its graph remains unchanged after rotation of 180 degrees about the origin.
$$f(-x)=-f(x) \tag{1.16}$$

Theorem 1.3 Rules:
Let f be even and f odd respectively.
$g=:f\cdot f$ is even $\qquad g=:f\cdot f$ is even
$g=:f\cdot f$ is odd $\qquad$ the same holds for division

Examples

Even: $\cos x$, $|x|$, c, x^2, x^4,$\ldots\exp\left(-x^2/2\right)$.
Odd: $\sin x$, $\tan x$, x, x^3, x^5,$\ldots$.

x-**Shift**: $f(x-c)\Rightarrow$ shift to the right
$f(x+c)\Rightarrow$ shift to the left (1.17)
y-**Shift**: $f(x)\pm c\Rightarrow$ shift up/down (1.18)

Proof eq. (1.17) $f(x_n-c)$ we take the x-value at x_n but take the y-value at $x_o:=x_n-c$
⇒ we shift the function to x_n. □

Euler's formula
$$e^{\pm ix}=\cos x\pm i\sin x \tag{1.19}$$

Euler's Identity
$$e^{\pm i}=-1 \tag{1.20}$$

Note
$$e^n=1\Leftrightarrow n=\mathrm{i}\,2\pi k, \qquad k\in\mathbb{N} \tag{1.21}$$

Definition 1.13 Norm $\|\cdot\|_{\mathcal{Y}}$: A norm measures the **size** of its argument.
Formally let $\mathcal{Y}$ be a vector space over a field F, a norm on $\mathcal{Y}$ is a map:
$$\|\cdot\|_{\mathcal{Y}}:\mathcal{Y}\mapsto\mathbb{R}_+ \tag{1.22}$$
that satisfies: $\forall\boldsymbol{x},\boldsymbol{y}\in\mathcal{Y}$, $\alpha\in F\subseteq\mathbb{K}$ $\quad K=\mathbb{R}$ or $\mathbb{C}$
1. Definitness: $\|\boldsymbol{x}\|_{\mathcal{Y}}=0\iff\boldsymbol{x}=0.$
2. Homogenity: $\|\alpha\boldsymbol{x}\|_{\mathcal{Y}}=|\alpha|\|\boldsymbol{x}\|_{\mathcal{Y}}$
3. Triangular Inequality: $\|\boldsymbol{x}+\boldsymbol{y}\|_{\mathcal{Y}}\leqslant\|\boldsymbol{x}\|_{\mathcal{Y}}+\|\boldsymbol{y}\|_{\mathcal{Y}}$

Meaning: Triangular Inequality

States that for any triangle, the sum of the lengths of any two sides must be greater than or equal to the length of the remaining side.

Corollary 1.5 Reverse Triangular Inequality:
$$-\|\boldsymbol{x}-\boldsymbol{y}\|_{\mathcal{Y}}\leqslant\|\boldsymbol{x}\|_{\mathcal{Y}}-\|\boldsymbol{y}\|_{\mathcal{Y}}\leqslant\|\boldsymbol{x}-\boldsymbol{y}\|_{\mathcal{Y}}$$
resp. $\quad \left|\|\boldsymbol{x}\|_{\mathcal{Y}}-\|\boldsymbol{y}\|_{\mathcal{Y}}\right|\leqslant\|\boldsymbol{x}-\boldsymbol{y}\|_{\mathcal{Y}}$

Corollary 1.6 Every norm is a convex function: By using definition definition 1.11 and the triangular inequality it follows (with the exception of the L0-norm):
$$\|\lambda x+(1-\lambda)y\|\leqslant\lambda\|x\|+(1-\lambda)\|y\|$$

1. Taylor Expansion

Definition 1.14 Taylor:
$$T_n(x)=\sum_{i=0}^n\frac{1}{n!}f^{(i)}(x_0)\cdot(x-x_0)^{(i)} \tag{1.23}$$
$$=f(x_0)+f'(x_0)(x-x_0)+\frac{1}{2}f''(x_0)(x-x_0)^2+\mathcal{O}(x^3) \tag{1.24}$$

Definition 1.15 Incremental Taylor: Goal: evaluate $T_n(x)$ (eq. (1.24)) at the point $x_0+\Delta x$ in order to propagate the function $f(x)$ by $h=\Delta x$:
$$T_n(x_0\pm h)=\sum_{i=0}^n\frac{h^i}{n!}f^{(i)}(x_0)i^{-1} \tag{1.25}$$
$$=f(x_0)\pm hf'(x_0)+\frac{h^2}{2}f''(x_0)\pm f'''(x_0)(h)^3+\mathcal{O}(h^4)$$

Note

If we chose Δx small enough it is sufficient to look only at the first two terms.

Definition 1.16 Multidimensional Taylor:
$$f(\boldsymbol{x})=f(\boldsymbol{x}_0)+Df(\boldsymbol{x}_0)(\boldsymbol{x}-\boldsymbol{x}_0) \tag{1.26}$$
$$+\frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_0)^\mathsf{T}H(\boldsymbol{x}-\boldsymbol{x}_0) \tag{1.27}$$
with $\quad H=\frac{\partial^2}{\partial\boldsymbol{x}\partial\boldsymbol{x}^\mathsf{T}}f(\boldsymbol{x}_0)$

Definition 1.17 Argmax: The argmax of a function defined on a set D is given by:
$$\arg\max_{x\in D}f(x)=\{x|f(x)\geqslant f(y),\forall y\in D\} \tag{1.28}$$

Definition 1.18 Argmin: The argmin of a function defined on a set D is given by:
$$\arg\min_{x\in D}f(x)=\{x|f(x)\leqslant f(y),\forall y\in D\} \tag{1.29}$$

Corollary 1.7 Relationship $\arg\min\leftrightarrow\arg\max$:
$$\arg\min_{x\in D}f(x)=\arg\max_{x\in D}-f(x) \tag{1.30}$$

Property 1.1 Argmax Identities:
1. **Shifting:**
$$\forall\lambda\ \text{const} \qquad \arg\max f(x)=\arg\max f(x)+\lambda \tag{1.31}$$
2. **Positive Scaling:**
$$\forall\lambda>0\ \text{const} \qquad \arg\max f(x)=\arg\max\lambda f(x) \tag{1.32}$$
3. **Negative Scaling:**
$$\forall\lambda<0\ \text{const} \qquad \arg\max f(x)=\arg\min\lambda f(x) \tag{1.33}$$
4. **Positive Functions:**
$$\forall\arg\max f(x)>0,\forall x\in\mathtt{dom}(f)$$
$$\arg\max f(x)=\arg\min\frac{1}{f(x)} \tag{1.34}$$
5. **Stricly Monotonic Functions:** for all strictly monotonic increasing functions (definition 1.7) g it holds that:
$$\arg\max g(f(x))=\arg\max f(x) \tag{1.35}$$

Definition 1.19 Max: The maximum of a function f defined on the set D is given by:
$$\max_{x\in D}f(x)=f(x^*) \quad \text{with} \quad \forall x^*\in\arg\max_{x\in D}f(x) \tag{1.36}$$

Definition 1.20 Min: The minimum of a function f defined on the set D is given by:
$$\min_{x\in D}f(x)=f(x^*) \quad \text{with} \quad \forall x^*\in\arg\min_{x\in D}f(x) \tag{1.37}$$

Corollary 1.8 Relationship $\min\leftrightarrow\max$:
$$\min_{x\in D}f(x)=-\max_{x\in D}-f(x) \tag{1.38}$$

Property 1.2 Max Identities:

1. **Shifting:**
$$\forall\lambda \text{ const} \qquad \max\{f(x)+\lambda\} = \lambda + \max f(x) \tag{1.39}$$
2. **Positive Scaling:**
$$\forall\lambda > 0 \text{ const} \qquad \max \lambda f(x) = \lambda \max f(x) \tag{1.40}$$
3. **Negative Scaling:**
$$\forall\lambda < 0 \text{ const} \qquad \max \lambda f(x) = \lambda \min f(x) \tag{1.41}$$
4. **Positive Functions:**
$$\forall \arg\max f(x) > 0, \forall x \in \mathtt{dom}(f) \qquad \max \frac{1}{f(x)} = \frac{1}{\min f(x)} \tag{1.42}$$
5. **Strictly Monotonic Functions:** for all strictly monotonic increasing functions (definition 1.7) g it holds that:
$$\max g(f(x)) = g(\max f(x)) \tag{1.43}$$

Definition 1.21 Supremum: The supremum of a function defined on a set D is given by:
$$\sup_{x\in D} f(x) = \{y | y \geqslant f(x), \forall x \in D\} = \min_{y|y\geqslant f(x),\forall x\in D} y \tag{1.44}$$
and is the smallest value y that is equal or greater $f(x)$ for any $x \iff$ smallest upper bound.

Definition 1.22 Infinmum: The infinmum of a function defined on a set D is given by:
$$\inf_{x\in D} f(x) = \{y | y \leqslant f(x), \forall x \in D\} = \max_{y|y\leqslant f(x),\forall x\in D} y \tag{1.45}$$
and is the biggest value y that is equal or smaller $f(x)$ for any $x \iff$ largest lower bound.

Corollary 1.9 Relationship sup ↔ inf:
$$\in_{x\in D} f(x) = -\sup_{x\in D} -f(x) \tag{1.46}$$

Note

The supremum/infinmum is necessary to handle unbound function that seem to converge and for which the max/min does not exist as the argmax/argmin may be empty.
E.g. consider $-e^x/e^x$ for which the max/min converges toward 0 but will never reached s.t. we can always choose a bigger $x \Rightarrow$ there exists no argmax/argmin $\Rightarrow$ need to bound the functions from above/below $\iff$ infinmum/supremum.

Definition 1.23 Time-invariant system (TIS): A function f is called time-invariant, if shifting the input in time leads to the same output shifted in time by the same amount.
$$y(t) = f(x(t), t) \xrightarrow[\forall\tau]{\text{time-invariance}} y(t-\tau) = f(x(t-\tau), t) \tag{1.47}$$

Definition 1.24 Inverse Function $g = f^{-1}$**:**
A function g is the inverse function of the function $f : A \subset \mathbb{R} \to B \subset \mathbb{R}$ if
$$f(g(x)) = x \qquad \forall x \in \mathtt{dom}(g) \tag{1.48}$$
and
$$g(f(u)) = u \qquad \forall u \in \mathtt{dom}(f) \tag{1.49}$$

Property 1.3
Reflective Property of Inverse Functions: f contains (a, b) if and only if f^{-1} contains (b, a).
The line $y = x$ is a symmetry line for f and f^{-1}.

Theorem 1.4 The Existence of an Inverse Function:
A function has an inverse function if and only if it is one-to-one.

Corollary 1.10 Inverse functions and strict monotonicity: If a function f is strictly monotonic definition 1.9 on its entire domain, then it is one-to-one and therefore has an inverse function.

1) Differential Calculus

Definition 1.25 Jacobi Matrix:
$$D\boldsymbol{u} = J_u(\boldsymbol{x}) = \frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}}(\boldsymbol{x}) = \frac{\partial(u_1,\dots,u_m)}{\partial(x_1,\dots,x_n)}\boldsymbol{x} = \tag{1.50}$$
$$= \begin{bmatrix} \frac{\partial u_1}{\partial x_1}(\boldsymbol{x}) & \frac{\partial u_1}{\partial x_2}(\boldsymbol{x}) & \cdots & \frac{\partial u_1}{\partial x_n}(\boldsymbol{x}) \\ \frac{\partial u_2}{\partial x_1}(\boldsymbol{x}) & & \ddots & \frac{\partial u_2}{\partial x_n}(\boldsymbol{x}) \\ \vdots & & & \vdots \\ \frac{\partial u_m}{\partial x_1}(\boldsymbol{x}) & \frac{\partial u_n}{\partial x_2}(\boldsymbol{x}) & \cdots & \frac{\partial u_n}{\partial x_n}(\boldsymbol{x}) \end{bmatrix}$$

2) Integral Calculus

Theorem 1.5 Important Integral Properties:

Addition
$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx \tag{1.51}$$
Reflection
$$\int_a^b f(x)\,dx = -\int_b^a f(x)\,dx \tag{1.52}$$
Translation
$$\int_a^b f(x)\,dx \overset{u:=x\pm c}{=} \int_{a\pm c}^{b\pm c} f(x \mp c)\,dx \tag{1.53}$$
f **Odd**
$$\int_{-a}^{a} f(x)\,dx = 0 \tag{1.54}$$
f **Even**
$$\int_{-a}^{a} f(x)\,dx = 2\int_0^a f(x)\,dx \tag{1.55}$$

Proof eqs. (1.55) and (1.56)
$$I := \int_{-a}^{a} f(x)\,dx = \int_{-a}^{0} f(x)\,dx + \int_0^a f(x)\,dx$$
$$\overset{\substack{t=-x\\dt=-dx}}{=} -\int_a^0 f(-x)\,dx + \int_0^a f(x)\,dx$$
$$= \int_0^a f(-x) + f(x)\,dx = \begin{cases} 0 & \text{if } f \text{ odd} \\ 2I & \text{if } f \text{ even}\end{cases}$$
□

2. Linear Algebra

Given a matrix $\boldsymbol{A} \in \mathbb{K}^{m,n}$

Rank: $\quad \mathfrak{rank}(\boldsymbol{A}) = \dim(\mathfrak{R}(\boldsymbol{A}))$
of a matrix is the dimension of the vector space generated (or spanned) by its columns/rows.
Span/Linear Hull: $\text{span}(\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_n) =$
$\{\lambda_1\boldsymbol{v}_1, \lambda_2\boldsymbol{v}_2, \dots, \lambda_n\boldsymbol{v}_n)\} = \{\boldsymbol{v} \mid \boldsymbol{v} = \sum_{i=1}^n \lambda_i\boldsymbol{v}_i), \lambda_i \in \mathbb{R}\}$
Is the set of vectors tha can be expressed as a linear combination of the vectors $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n$.
Note these vectors may be linearly independent.
Generating Set: Is the set of vectors which span the $\mathbb{R}^n$ that is: $\text{span}(\boldsymbol{v}_1, \dots, \boldsymbol{v}_m) = \mathbb{R}^n$.
e.g. $(4, 0)^\top, (0, 5)^\top$ span the $\mathbb{R}^n$.
Basis $\mathfrak{B}$: A lin. indep. generating set of the $\mathbb{R}^n$ is called basis of the $\mathbb{R}^n$.
The unit vectors $\boldsymbol{e}_1, \dots, \boldsymbol{e}_n$ build a standard basis of the $\mathbb{R}^n$
Vector Space
Image/Range: $\quad \mathfrak{R}(\boldsymbol{A}) := \{\boldsymbol{A}x \mid x \in \mathbb{K}^n\} \subset \mathbb{K}^n$
Null-Space/Kernel: $\quad \mathtt{N} := \{z \in \mathbb{K}^n \mid \boldsymbol{A}z = 0\}$
Dimension theorem:

Theorem 2.1 Rank-Nullity theorem: For any $\boldsymbol{A} \in \mathbb{Q}^{m\times n}$
$$n = \dim(\mathtt{N}[\boldsymbol{A}]) + \dim(\mathfrak{R}[\boldsymbol{A}])$$

From orthogonality it follows $x \in \mathfrak{R}(\boldsymbol{A})$, $y \in \mathtt{N}(\boldsymbol{A}) \Rightarrow x^\top y = 0$.

1) Vector Algebra

1. **Planes**
https://math.stackexchange.com/questions/1485509/show-that-two-planes-are-parallel-and-find-the-distance-between-them

3. Geometry

Definition 3.1 Affine Transfromation/Map:

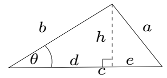
Corollary 3.1 Affine Transformation in 1D: Given: numbers $x \in \hat{\Omega}$ with $\hat{\Omega} = [a, b]$
The affine transformation of $\phi : \hat{\Omega} \to \Omega$ with $y \in \Omega = [c, d]$ is defined by:
$$y = \phi(x) = \frac{d-c}{b-a}(x-a) + c \tag{3.1}$$

Proof corollary 3.1 By definition 3.1 we want a function $f : [a, b] \to [c, d]$ that satisfies:
$$f(a) = c \qquad \textbf{and} \qquad f(b) = d$$
additionally $f(x)$ has to be a linear function (definition 1.10), that is the output scales the same way as the input scales.
Thus it follows:
$$\frac{d-c}{b-a} = \frac{f(x) - f(a)}{x-a} \iff f(x) = \frac{d-c}{b-a}(x-a) + c$$
□

Trigonometry

Law 3.1 Law of Cosine: relates the side of a triangle to the cosine of its angles.
$$a^2 = b^2 + c^2 - 2bc\cos\sphericalangle(b, c) \tag{3.2}$$

Proof We know: $\sin\theta = \frac{h}{b} \Rightarrow \underline{h}$ and $\cos\theta = \frac{d}{b} \Rightarrow d$
Thus $\underline{e} = c - d = c - b\cos\theta \Rightarrow a^2 = \underline{e}^2 + \underline{h}^2 \Rightarrow a$ □



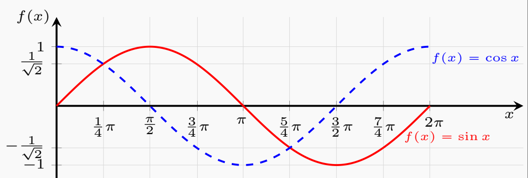
Law 3.2 Pythagorean theorem: special case of eq. (3.2) for right triangle:
$$a^2 = b^2 + c^2 \tag{3.3}$$

Euler's formula
$$e^{\pm ix} = \cos x \pm i\sin x \tag{3.4}$$

Euler's Identity
$$e^{\pm i} = -1 \tag{3.5}$$

Note
$$e^n = 1 \Leftrightarrow n = \mathrm{i}\,2\pi k, \qquad k \in \mathbb{N} \tag{3.6}$$

Sine and Cosine



$$\cos x \overset{(3.4)}{=} \frac{1}{2}\left[e^{ix} + e^{-ix}\right] \tag{3.7}$$
$$\sin x \overset{(3.4)}{=} \frac{1}{2i}\left[e^{ix} - e^{-ix}\right] = -\frac{i}{2}\left[e^{ix} - e^{-ix}\right] \tag{3.8}$$

Sinh and Cosh
$$\cosh x \overset{(3.4)}{=} \frac{1}{2}\left[e^x + e^{-x}\right] = \cos(\mathrm{i}\,x) \tag{3.9}$$
$$\sinh x \overset{(3.4)}{=} \frac{1}{2}\left[e^x - e^{-x}\right] = -\mathrm{i}\sin(\mathrm{i}\,x) \tag{3.10}$$

Note
$$e^x = \cosh x + \sinh x \qquad e^{-x} = \cosh x - \sinh x \tag{3.11}$$

Note
- $\cosh x$ is strictly positive.
- $\sinh x = 0$ has a unique root at $x = 0$.

Theorem 3.1 Addition Theorems:
$$\sin(\alpha \pm \beta) = \sin\alpha\cos\beta \pm \cos\alpha\sin\beta \tag{3.12}$$
$$\cos(\alpha \pm \beta) = \cos\alpha\cos\beta \mp \sin\alpha\sin\beta \tag{3.13}$$

Werner Formulas
$$\sin\alpha\cos\beta = \frac{1}{2}\left[\sin(\alpha+\beta) + \sin(\alpha-\beta)\right] \tag{3.14}$$
$$\sin\alpha\sin\beta = \frac{1}{2}\left[\cos(\alpha-\beta) - \cos(\alpha+\beta)\right] \tag{3.15}$$
$$\cos\alpha\cos\beta = \frac{1}{2}\left[\cos(\alpha+\beta) + \cos(\alpha-\beta)\right] \tag{3.16}$$

Note
Using theorem 3.1 if follows:
$$\cos(\alpha \pm \pi) = -\cos\alpha \qquad \text{and} \qquad \sin(\alpha \pm \pi) = -\sin\alpha \tag{3.17}$$

4. Topology

5. Numerics

1) Numerical Quadrature

Definition 5.1 Order of a Quadrature Rule: The order of a quadrature rule $\mathcal{Q}_n : \mathcal{C}^0([a, b]) \to \mathbb{R}$ is defined as:
$$\text{order}(\mathcal{Q}_n) := \max\left\{n \in \mathbb{N}_0 : \mathcal{Q}_n(p) = \in_a^b p(t)\,dt \quad \forall p \in \mathcal{P}_n\right\} + 1 \tag{5.1}$$
Thus it is the maximal degree+1 of polynomials (of degree maximal degree) $\mathcal{P}_{\text{maximal degree}}$ for which the quadrature rule yields exact results.

Note
Is a quality measure for quadrature rules.

1. Composite Quadrature

Definition 5.2 Composite Quadrature:
Given a mesh $\mathcal{M} = \{a = x_0 < x_1 < \ldots < x_m = b\}$ apply a Q.R. $\mathcal{Q}_n$ to each of the mesh cells $I_j := [x_{j-1}, x_j] \quad \forall j = 1, \dots, m \triangleq$ p.w. Quadrature:
$$\int_a^b f(t)\,dt = \sum_{j=1}^m \int_{x_{j-1}}^{x_j} f(t)\,dt = \sum_{j=1}^m \mathcal{Q}_n(f_{I_j}) \tag{5.2}$$

Lemma 5.1 Error of Composite quadrature Rules:
Given a function $f \in \mathcal{C}^k([a, b])$ with integration domain:
$$\sum_{i=1}^m h_i = |b - a| \qquad \text{for } \mathcal{M} = \{x_j\}_{j=1}^m$$
Let: $h_{\mathcal{M}} = \max_j |x_j, x_{j-1}|$ be the mesh-width
Assume an equal number of quadrature nodes for each interval $I_j = [x_{j-1}, x_j]$ of the mesh $\mathcal{M}$ i.e. $n_j = n$.
Then the error of a quadrature rule $\mathcal{Q}_n(f)$ of order q is given by:
$$\epsilon_n(f) = \mathcal{O}\left(n^{-\min\{k,q\}}\right) = \mathcal{O}\left(h_{\mathcal{M}}^{\min\{k,q\}}\right) \qquad \text{for } n \to \infty$$
$$\overset{\text{corollary 1.3}}{=} \mathcal{O}\left(n^{-q}\right) = \mathcal{O}\left(h_{\mathcal{M}}^q\right) \qquad \text{with } h_{\mathcal{M}} = \frac{1}{n} \tag{5.3}$$

Definition 5.3 Complexity W: Is the number of function evaluations $\triangleq$ number of quadrature points.
$$W(\mathcal{Q}(f)_n) = \#\text{f-eval} \triangleq n \tag{5.4}$$

Lemma 5.2 Error-Complexity $W(\epsilon_n(f))$: Relates the complexity to the quadrature error.
Assuming and quadrature error of the form :

$$\epsilon_n(f) = \mathcal{O}(n^{-q}) \qquad \iff \qquad \epsilon_n(f) = cn^{-q} \qquad c \in \mathbb{R}_+$$

the error complexity is algebraic (**??**) and is given by:

$$W(\epsilon_n(f)) = \mathcal{O}(\epsilon_n^{1/q}) = \mathcal{O}\left(\sqrt[q]{\epsilon_n}\right) \tag{5.5}$$

Proof lemma 5.2: **Assume**: we want to reduce the error by a factor of ϵ_n by increasing the number of quadrature points $n_{\text{new}} = a \cdot n_{\text{old}}$.
Question: what is the additional effort (#f-eval) needed in order to achieve this reduction in error?

$$\frac{c \cdot n_n^q}{c \cdot n_o^q} = \frac{1}{\epsilon_n} \qquad \Rightarrow \qquad n_n = n_o \cdot \sqrt[q]{\epsilon_n} = \mathcal{O}(\sqrt[q]{\epsilon_n}) \tag{5.6}$$

□

6. Stochastic

The probability that a discret random variable x is equal to some value $\bar{x} \in \mathcal{X}$ is:

$$p_x(\bar{x}) = \mathbb{P}(x = \bar{x})$$

addapet

Definition 6.1 Almost Surely (a.s.): Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. An event $\omega \in \mathcal{F}$ happens almost surely iff

$$\mathbb{P}(\omega) = 1 \quad \iff \quad \omega \text{ happens a.s.} \tag{6.1}$$

Definition 6.2 Probability Mass Function (PMF):

Definition 6.3 Discrete Random Variable (DVR): The set of possible values $\bar{x}$ of $\mathcal{X}$ is countable of finite.

$$\mathcal{X} = \{0, 1, 2, 3, 4, \dots, 8\} \qquad \mathcal{X} = \mathbb{N} \tag{6.2}$$

Definition 6.4 Probability Density Function (PDF): Is real function $f : \mathbb{R}^n \to [0, \infty)$ that satisfies:

Non-negativity:
$$f(x) \geqslant 0, \quad \forall x \in \mathbb{R}^n \tag{6.3}$$

Normalization:
$$\int_{-\infty}^{\infty} f(x)\,dx \overset{!}{=} 1 \tag{6.4}$$

Must be integrable (6.5)

Note: why do we need probability density functions

A continuous random variable X can realise an infinite count of real number values within its support B (as there are an infinitude of points in a line segment).
Thus we have an infinitude of values whose sum of probabilities must equal one.
Thus these probabilities must each be zero otherwise we would obtain a probability of ∞. As we can not work with zero probabilities we use the next best thing, infinitesimal probabilities (defined as a limit).
We say they are almost surely equal to zero:

$$\mathbb{P}(X = x) = 0 \qquad \text{a.s.}$$

To have a sensible measure of the magnitude of these infinitesimal quantities, we use the concept of probability density, which yields a probability mass when integrated over an interval.

Definition 6.5 Continuous Random Variable (CRV): A real random variable (rrv) X is said to be (absolutely) continuous if there exists a pdf (definition 6.4) f_X s.t. for any subset $B \subset \mathbb{R}$ it holds:

$$\mathbb{P}(X \in B) = \int_B f_X(x)\,dx \tag{6.6}$$

Property 0.1 Zero Probability: If X is a continuous rrv (definition 6.5), then:

$$\mathbb{P}(X = a) = 0 \qquad \forall a \in \mathbb{R} \tag{6.7}$$

Open vs. Closed Interval

Property 0.2 : For any real numbers a and b, with $a < b$ it holds:

$$\begin{aligned}\mathbb{P}(a \leqslant X \leqslant b) &= \mathbb{P}(a \leqslant X < b) = \mathbb{P}(a < X \leqslant b)\\ &= \mathbb{P}(a < X < b)\end{aligned} \tag{6.8}$$

$\iff$ including or not the bounds of an interval does not modify the probability of a continuous rrv.

Note

Changing the value of a function at finitely many points has no effect on the value of a definite integral.

Corollary 6.1 : In particular for any real numbers a and b with $a < b$, letting $B = [a, b]$ we obtain:

$$\mathbb{P}(a \leqslant X \leqslant b) = \int_a^b f_x(x)\,dx$$

Proof Property 0.1:

$$\begin{aligned}\mathbb{P}(X = a) &= \lim_{\Delta x \to 0} \mathbb{P}(X \in [a, a + \Delta x])\\ &= \lim_{\Delta x \to 0} \int_a^{a+\Delta x} f_X(x)\,dx = 0\end{aligned}$$

□

Proof Property 0.2:

$$\begin{aligned}\mathbb{P}(a \leqslant X \leqslant b) &= \mathbb{P}(a \leqslant X < b) = \mathbb{P}(a < X \leqslant b)\\ &= \mathbb{P}(a < X < b) = \int_a^b f_X(x)\,dx\end{aligned}$$

□

Definition 6.6 Support of a probability density function: The support of the density of a pdf $f_X(.)$ is the set of values of the random variable X s.t. its pdf is non-zero:

$$\operatorname{supp}(()f_X) := \{x \in \mathcal{X} | f(x) > 0\} \tag{6.9}$$

Note: this is not a rigorous definition.

Theorem 6.1 RVs are defined by a PDFs: A probability density function f_X completely determines the distribution of a continuous real-valued random variable X.

Corollary 6.2 Identically Distributed: From theorem 6.1 it follows that to RV X and Y that have exactly the same pdf follow the same distribution.
We say X and Y are identically distributed.

1. Cumulative Distribution Fucntion

Definition 6.7 Cumulative distribution function (CDF): Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space.
The (cumulative) distribution function of a real-valued random variable X is the function given by:

$$F_X(x) = \mathbb{P}(X \leqslant x) \qquad \forall x \in \mathbb{R}$$

Property 1.1 :

Monotonically Increasing
$$x \leqslant y \iff F_X(x) \leqslant F_X(y) \quad \forall x, y \in \mathbb{R} \tag{6.10}$$

Upper Limit
$$\lim_{x \to \infty} F_X(x) = 1 \tag{6.11}$$

Lower Limit
$$\lim_{x \to -\infty} F_X(x) = 0 \tag{6.12}$$

Definition 6.8 CDF of a discret rv X: Let X be discret rv with pdf p_X, then the CDF of X is given by:

$$F_X(x) = \mathbb{P}(X \leqslant x) = \sum_{t=-\infty}^{x} p_X(t)$$

Definition 6.9 CDF of a continuous rv X: Let X be continuous rv with pdf f_X, then the CDF of X is given by:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt \quad \iff \quad \frac{\partial F_X(x)}{\partial x} = f_X(x)$$

Lemma 6.1 Probability Interval: Let X be a continuous rrv with pdf f_X and cumulative distribution function F_X, then it holds that:

$$\mathbb{P}(a \leqslant X \leqslant b) = F_X(b) - F_X(a) \tag{6.13}$$

Proof definition 6.9:

$$F_X(x) = \mathbb{P}(X \leqslant x) = \mathbb{P}(X \in (-\infty, x)) = \int_{-\infty}^{x} f_X(t)\,dt$$

□

Proof lemma 6.1:

$$\mathbb{P}(a \leqslant X \leqslant b) = \mathbb{P}(X \leqslant b) - \mathbb{P}(X \leqslant a)$$

or by the fundamental theorem of calculus (theorem 1.2):

$$\mathbb{P}(a \leqslant X \leqslant b) = \int_a^b f_X(t)\,dt = \int_a^b \frac{\partial F_X(t)}{\partial t}\,dt = [F_X(t)]|_a^b$$

□

Theorem 6.2 A continuous rv is fully characterized by its CDF: A cumulative distribution function completely determines the distribution of a continuous real-valued random variable.

Theorem 6.3 (Scalar Discret) Change of Variables: Let X be a discret rv $X \in \mathcal{X}$ with pmf p_X and define $Y \in \mathcal{Y}$ as $Y = g(x)$ s.t. $\mathcal{Y} = \{y | y = g(x), \forall x \in \mathcal{X}\}$. **Where** g is an arbitrary strictly monotonic (definition 1.9) function.
Let: $\mathcal{X}_y = x_i$ be the set of all $x_i \in \mathcal{X}$ s.t. $y = g(x_i)$.
Then the pmf of Y is given by:

$$p_Y(y) = \sum_{x_i \in \mathcal{X}_y} p_X(x_i) = \sum_{x \in \mathcal{Y} : g(x) = y} p_X(x) \tag{6.14}$$

Proof theorem 6.3:

$$Y = g(X) \iff \mathbb{P}(Y = y) = \mathbb{P}(x \in \mathcal{X}_y) = p_Y(y)$$

□

Theorem 6.4 (Scalar Continuous) Change of Variables: Let X be a continuous rv $X \in \mathcal{X}$ with pdf f_X and define $Y \in \mathcal{Y}$ as $Y = g(x)$ s.t. $\mathcal{Y} = \{y | y = g(x), \forall x \in \mathcal{X}\}$. **Where** g is an arbitrary strictly monotonic (definition 1.9) function.
Then the pdf of Y is given by:

$$f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right| = f_X(x)\left|\frac{d}{dy}\left(g^{-1}(y)\right)\right| \tag{6.15}$$

$$= f_X(x)\frac{1}{\left|\frac{dy}{dx}\right|} = \frac{f_X(g^{-1}(y))}{\left|\frac{dg}{dx}(g^{-1}(y))\right|} \tag{6.16}$$

Theorem 6.5 (Continuous) Change of Variables: Let X be a continuous rv $X \in \mathcal{X}$ with pdf f_X and define $Y \in \mathcal{Y}$ as $Y = g(x)$ s.t. $\mathcal{Y} = \{y | y = g(x), \forall x \in \mathcal{X}\}$. **Where** g is an arbitrary strictly monotonic (definition 1.9) function.
Then the pdf of Y is given by:

$$f_Y(\mathbf{y}) = f_X(\mathbf{x})\left|\det\left(\frac{\partial g}{\partial \mathbf{x}}\right)\right|^{-1} \tag{6.17}$$

$$= f_X(g^{-1}(\mathbf{y}))\left|\det\left(\frac{\partial g}{\partial \mathbf{x}}\right)\right|^{-1}$$

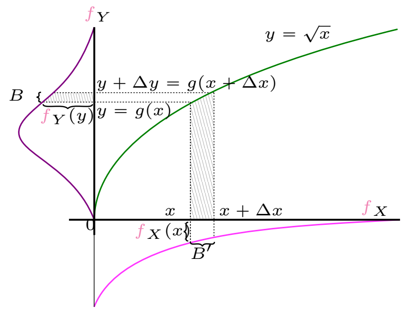
Where $\frac{\partial g}{\partial \mathbf{x}}$ is the Jaccobian (definition 1.25).

Note

A monotonic function is required in order to satisfy inevitability.

Proof theorem 6.4 (non-formal): The probability contained in a differential area must be invariant under a change of variables that is:

$$|f_Y(y)\,dy| = |f_x(x)\,dx|$$



□

Proof theorem 6.4 from CDF:

$$\mathbb{P}(Y \leqslant y) = \mathbb{P}(g(X) \leqslant y) = \begin{cases}\mathbb{P}(X \leqslant g^{-1}(y)) & \text{if } g \text{ is increas.}\\ \mathbb{P}(X \geqslant g^{-1}(y)) & \text{if } g \text{ is decreas.}\end{cases}$$

If g is monotonically increasing:

$$F_Y(y) = F_X(g^{-1}(y))$$

$$f_Y(y) = \frac{d}{dy}F_X(g^{-1}(y)) = f_X(x) \cdot \frac{d}{dy}g^{-1}(y)$$

If g is monotonically decreasing:

$$F_Y(y) = 1 - F_X(g^{-1}(y))$$

$$f_Y(y) = \frac{d}{dy}F_X(g^{-1}(y)) = -f_X(x) \cdot \frac{d}{dy}g^{-1}(y)$$

□

Proof theorem 6.4: **Let** $B = [x, x + \Delta x]$ and $B' = [y, y + \Delta y] = [g(x), g(x + \Delta x)]$ we know that the probability of equal events is equal:

$$y = g(x) \quad \Rightarrow \quad \mathbb{P}(y) = \mathbb{P}(g(x)) \text{ (for disc. rv.)}$$

Now lets consider the probability for the continuous r.v.s:

$$\mathbb{P}(X \in B) = \int_x^{x+\Delta x} f_X(t)\,dt \xrightarrow{\Delta x \to 0} |\Delta x \cdot f_x(x)|$$

For y we use Taylor (definition 1.14)

$$\begin{aligned}g(x + \Delta x) &\overset{\text{eq. (1.24)}}{=} g(x) + \frac{dg}{dx}\Delta y \quad \text{for } \Delta x \to 0\\ &= y + \Delta y \quad \text{with } \Delta y := \frac{dg}{dx} \cdot \Delta x\end{aligned} \tag{6.18}$$

Thus for $\mathbb{P}(Y \in B')$ it follows:

$$\begin{aligned}\mathbb{P}(X \in B') &= \int_y^{y+\Delta y} f_Y(t)\,dt \xrightarrow{\Delta y \to 0} |\Delta y \cdot f_Y(y)|\\ &= \left|\frac{dg}{dx}(x)\Delta x \cdot f_Y(y)\right|\end{aligned}$$

Now we simply need to related the surface of the two pdfs:

$$B = [x, x + \Delta x] \overset{\text{same surfaces}}{\propto} [y, y + \Delta y] = B'$$

$$\mathbb{P}(Y \in B) = \mathbb{P}(X \in B')$$

$$\overset{\Delta y \to 0}{\iff} |f_Y(y) \cdot \Delta y| = \left|f_Y(y) \cdot \frac{dg}{dx}(x)\Delta x\right| = |f_X(x) \cdot \Delta x|$$

$$f_Y(y)\left|\frac{dg}{dx}(x)\right||\Delta x| = f_X(x) \cdot |\Delta x|$$

$$\Rightarrow f_Y(y) = \frac{f_X(x)}{\left|\frac{dg}{dx}(x)\right|} = \frac{f_X(g^{-1}(y))}{\left|\frac{dg}{dx}g^{-1}(y)\right|}$$

□

Rules of Probability

Definition 6.10 Marginalization/Sum Rule:

Given: $p_{x,y}(\bar{x}, \bar{y})$
$$p_x(\bar{x}) := \sum_{\bar{y} \in \mathcal{Y}} p_{x,y}(\bar{x}, \bar{y}) \tag{6.19}$$

Definition 6.11 Conditioning:

Given: p_{xy}
$$p_{xy}(x | y = \bar{y}) := \frac{p_{xy}(x, y = \bar{y})}{p_y(y = \bar{y})}$$

if
$$p_y(\bar{y}) \neq 0 \tag{6.20}$$

Definition 6.12 Product Rule: follows directly from eq. (6.20)

$$p(x, y) = p(y|x)p_x(x) = p(x|y)p(y) \tag{6.21}$$

Theorem 6.6 Total Probability Theorem: Given: $p_{x,y}(\bar{x}, \bar{y})$ with eq. (6.19) and eq. (6.21) it follows:

$$\begin{aligned}p_x(\bar{x}) &\overset{\text{eq. (6.19)}}{=} \sum_{\bar{y} \in \mathcal{Y}} p_{x,y}(\bar{x}, \bar{y})\\ &\overset{\text{eq. (6.21)}}{=} \sum_{y \in \mathcal{Y}} p_{x|y}(\bar{x}|\bar{y})\,p_y(\bar{y})\end{aligned} \tag{6.22}$$

Definition 6.13 Independence: Two random variables x and y are said to be **independent** if:
$$p(x|y) = p(x) \overset{eq.\ (6.20)}{\iff} p(x,y) = p(x)\,p(y) \qquad (6.23)$$

Corollary 6.3 eq. (6.23):
$$p(x|y) = p(x) \overset{\text{implies}}{\iff} p(y|x) = p(y) \qquad (6.24)$$

add mutual independence

2. Conditional PDF

Let x, y, z be R.V. (which themselves may be collections of random variables)

Definition 6.14 Marginalization:
$$p_{x|z}(\bar{x}|\bar{z}) = \sum_{\bar{y}\in\mathcal{Y}} p_{xy|z}(\bar{x},\bar{y}|\bar{z}) \qquad (6.25)$$

Definition 6.15 Conditioning:
$$p_{x|yz}(\bar{x}|\bar{y},\bar{z}) = \frac{p_{xy|z}(\bar{x},\bar{y}|\bar{z})}{p_{y|z}(\bar{y}|\bar{z})} \qquad (6.26)$$

Definition 6.16 Product Rule: follows directly from eq. (6.26)
$$p_{xyz}(\bar{x},\bar{y}|\bar{z}) = p_{x|yz}(\bar{x}|\bar{y},\bar{z})\,p_{y|z}(\bar{y}|\bar{z}) \qquad (6.27)$$

Note

z basically parameterizes the pdf.

Definition 6.17 Conditional Independence: Two random variables x and y are said to be conditionally independent on z if
$$p(x|y,z) = p(x|z) \overset{eq.\ (6.26)}{=} p(x,y|z) = p(x|z)p(y|z) \qquad (6.28)$$
Hence, knowledge of z makes x and y independent.

Note

Conditional independence does not imply $p(x,y) = p(x)p(y)$

Rule 6.1 Bayes' Rule: Given: the prior $p(X)$ and the liklihood $p(Y|X)$, we can find the posterior by:
$$\mathbb{P}(X|Y) = \frac{\mathbb{P}(Y,X)}{\mathbb{P}(Y)} = \frac{\mathbb{P}(X)\mathbb{P}(Y|X)}{\mathbb{P}(Y)} \overset{\text{normalization}}{=} \frac{\mathbb{P}(X)\mathbb{P}(Y|X)}{\sum_{X=x}\mathbb{P}(X=x)\mathbb{P}(Y|X=x)}$$
$$\textbf{Posterior} = \frac{\text{Prior}\cdot\text{Liklihood}}{\text{Normalization}}$$

Proof Equation (6.25)
$$p_{x|z}(\bar{x}|\bar{z}) \overset{eq.\ (6.20)}{=} \frac{p_{xz}(\bar{x},\bar{z})}{p_z(\bar{z})} \overset{eq.\ (6.19)}{=} \frac{\sum_{y\in\mathcal{Y}} p_{xyz}(\bar{x},\bar{y},\bar{z})}{p_z(\bar{z})} \overset{eq.\ (6.21)}{=} \frac{\sum_{y\in\mathcal{Y}} p_{xy|z}(\bar{x},\bar{y}|\bar{z})\,\cancel{p_z(\bar{z})}}{\cancel{p_z(\bar{z})}}$$
□

Proof Equation (6.26)
$$p_{x|yz}(\bar{x}|\bar{y},\bar{z}) \overset{eq.\ (6.20)}{=} \frac{p_{xyz}(\bar{x},\bar{y},\bar{z})}{p_{yz}(\bar{y},\bar{z})} \overset{eq.\ (6.21)}{=} \frac{p_{xy|z}(\bar{x},\bar{y}|\bar{z})\,\cancel{p_z(\bar{z})}}{p_{y|z}(\bar{y}|\bar{z})\,\cancel{p_z(\bar{z})}}$$
□

Proof Equation (6.24)
$$p(y|x) \overset{eq.\ (6.20)}{=} \frac{p(x,y)}{p(x)} \overset{p(x,y)=p(x)p(y)}{=} p(y)$$
□

Proof Equation (6.28)
$$p(x|y,z) = p(x|z)$$
$$p(x|y,z) = \frac{p(x,y,z)}{p(y,z)} = \frac{p(x,y|z)\,\cancel{p(z)}}{p(y|z)\,\cancel{p(z)}} = \frac{p(x,y|z)}{p(y|z)}$$
$$\Rightarrow p(x,y|z) = p(x|z)p(y|z)$$
□

Key figures

Expectation

Definition 6.18 Expectation (disc. case):
$$\mu_X := \mathbb{E}_x[x] := \sum_{\bar{x}\in\mathcal{X}} \bar{x}\,p_x(\bar{x}) \qquad (6.29)$$

Definition 6.19 Expectation (cont. case):
$$\mathbb{E}_x[x] := \int_{\bar{x}\in\mathcal{X}} \bar{x}\,f_x(\bar{x})\,\mathrm{d}\bar{x} \qquad (6.30)$$

Law 6.1 Expectation of independent variables:
$$\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y] \qquad (6.31)$$

Property 2.1 Translation and scaling: If $\boldsymbol{X} \in \mathbb{R}^n$ and $\boldsymbol{Y} \in \mathbb{R}^n$ are random vectors, and $a, b, a \in \mathbb{R}^n$ are constants then it holds:
$$\mathbb{E}[a + b\boldsymbol{X} + c\boldsymbol{Y}] = a + b\mathbb{E}[\boldsymbol{X}] + c\mathbb{E}[\boldsymbol{Y}] \qquad (6.32)$$
Thus $\mathbb{E}$ is a linear operator (definition 1.10).

Note: Expectation of the expectation

The expectation of a r.v. X is a constant hence with Property 2.1 it follows:
$$\mathbb{E}[\mathbb{E}[X]] = \mathbb{E}[X] \qquad (6.33)$$

Property 2.2 Matrix×Expectation: If $\boldsymbol{X} \in \mathbb{R}^n$ is a random vector and $\boldsymbol{A} \in \mathbb{R}^{m\times n}, \boldsymbol{B} \in \mathbb{R}^{n\times m}$ are constant matrices then it holds:
$$\mathbb{E}[\boldsymbol{AXB}] = \boldsymbol{A}\mathbb{E}[(\boldsymbol{XB})] = \boldsymbol{A}\mathbb{E}[\boldsymbol{X}]\,\boldsymbol{B} \qquad (6.34)$$

Proof eq. (6.31):
$$\mathbb{E}[XY] = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{X,Y}(x,y)xy \overset{\text{definition 6.13}}{=} \sum_{x\in\mathcal{X}} p_X(x)x \sum_{y\in\mathcal{Y}} p_Y(y)y = \mathbb{E}[X]\,\mathbb{E}[Y]$$
□

Law 6.2 of the Unconscious Statistician: Let X be a random variable $X \in \mathcal{X}$ and define $Y \in \mathcal{Y}$ as $Y = g(x)$ s.t. $\mathcal{Y} = \{y|y = g(x), \forall x \in \mathcal{X}\}$, then Y is a random variable with expectation:
$$\mathbb{E}_Y[y] = \sum_{y\in\mathcal{Y}} y\,p_Y(y) = \sum_{x\in\mathcal{X}} g(x)p_x(x) \quad \text{or integral for CRV} \qquad (6.35)$$

Consequence

Hence if we p_X we do not have to first calculate p_Y in order to calculate $\mathbb{E}_Y[y]$.

Variance

Definition 6.20 Variance $\mathbb{V}(X)$: The variance of a random variable X is the expected value of the squared deviation from the expectation of X ($\mu = \mathbb{E}[X]$).
It is a measure of how much the actual values of a random variable X fluctuate around its executed value $\mathbb{E}[X]$ and is defined by:
$$\mathbb{V}(X) := \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 \qquad (6.36)$$

Proof eq. (6.36)
$$\mathbb{V}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}\left[X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2\right]$$
$$\overset{\text{Property 2.1}}{=} \mathbb{E}\left[X^2\right] - 2\mathbb{E}[X]\,\mathbb{E}[X] + \mathbb{E}[X]^2 = \mathbb{E}\left[X^2\right] - \mu^2$$
□

Property 2.3 Variance of a Constant: If $a \in \mathbb{R}$ is a constant then it follows that its expected value is deterministic ⇒ we have no uncertainty ⇒ no variance:
$$\mathbb{V}(a) = 0 \quad \textbf{with} \quad a \in \mathbb{R} \qquad (6.37)$$

Property 2.4 Affine Transformation: If $\boldsymbol{X} \in \mathbb{R}^n$ is a random vector, $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ a constant matrix and $b \in \mathbb{R}^n$ then it holds:
$$\mathbb{V}(\boldsymbol{AX} + b) = \boldsymbol{A}\mathbb{V}(\boldsymbol{X})\,\boldsymbol{A}^\intercal \qquad (6.38)$$

Proof Property 2.4
$$\mathbb{V}(\boldsymbol{AX} + b) = \mathbb{E}\left[(\boldsymbol{AX} - \mathbb{E}[\boldsymbol{XA}])^2\right] + 0 =$$
$$= \mathbb{E}\left[(\boldsymbol{AX} - \mathbb{E}[\boldsymbol{AX}])(\boldsymbol{AX} - \mathbb{E}[\boldsymbol{AX}])^\intercal\right]$$
$$= \mathbb{E}\left[\boldsymbol{A}(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{A}(\boldsymbol{X} - (\mathbb{E}[\boldsymbol{X}]))^\intercal\right]$$
$$= \mathbb{E}\left[\boldsymbol{A}(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{X} - (\mathbb{E}[\boldsymbol{X}])^\intercal \boldsymbol{A}^\intercal\right]$$
$$= \boldsymbol{A}\mathbb{E}\left[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{X} - (\mathbb{E}[\boldsymbol{X}])^\intercal\right]\boldsymbol{A}^\intercal = \boldsymbol{A}\mathbb{V}(\boldsymbol{X})\,\boldsymbol{A}^\intercal$$
□

Definition 6.21 Covariance: The Covariance is a measure of how much two or more random variables vary linearly with each other.
$$\mathrm{Cov}[X,Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] \qquad (6.39)$$

Proof eq. (6.39)
$$\mathrm{Cov}[X,Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$
$$= \mathbb{E}[XY - X\mathbb{E}[Y] - \mathbb{E}[X]\,Y + \mathbb{E}[X]\,\mathbb{E}[Y]]$$
$$= \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] - \mathbb{E}[X]\,\mathbb{E}[Y] + \mathbb{E}[X]\,\mathbb{E}[Y]$$
$$= \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$$
□

Definition 6.22 Covariance Matrix: The variance of a k-dimensional random vector $\boldsymbol{X} = (X_1 \ \ldots \ X_k)$ is given by the Covariance Matrix.
The Covariance is a measure of how much two or more random variables vary linearly with each other and the Variance on the diagonal is again a measure of how much a variable varies:
$$\mathbb{V}(\boldsymbol{X}) := \Sigma(\boldsymbol{X}) := \mathrm{Cov}[\boldsymbol{X},\boldsymbol{X}] := \qquad (6.40)$$
$$= \mathbb{E}\left[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])^\intercal\right]$$
$$= \mathbb{E}\left[\boldsymbol{XX}^\intercal\right] - \mathbb{E}[\boldsymbol{X}]\,\mathbb{E}[\boldsymbol{X}]^\intercal \in [-\infty,\infty]$$
$$= \begin{bmatrix} \mathbb{V}(X_1) & \cdots & \mathrm{Cov}[X_1,X_k] \\ \vdots & \ddots & \\ \mathrm{Cov}[X_k,X_1] & \cdots & \mathbb{V}(X_k) \end{bmatrix}$$
$$= \begin{bmatrix} \mathbb{E}[(X_1-\mu_1)(X_1-\mu_1)] & \cdots & \mathbb{E}[(X_1-\mu_1)(X_k-\mu_k)] \\ \vdots & \ddots & \\ \mathbb{E}[(X_k-\mu_k)(X_1-\mu_1)] & \cdots & \mathbb{E}[(X_k-\mu_k)(X_k-\mu_k)] \end{bmatrix}$$

Note: Covariance and Variance

The variance is a special case of the covariance in which two variables are identical:
$$\mathrm{Cov}[X,X] = \mathbb{V}(X) \equiv \sigma^2(X) \equiv \sigma_X^2 \qquad (6.41)$$

add http://www.visiondummy.com/2014/04/geometric-interpretation-covariance-matrix/

Property 2.5 Translation and Scaling:
$$\mathrm{Cov}(a + bX, c + dY) = bd\,\mathrm{Cov}(X,Y) \qquad (6.42)$$

Definition 6.23 Correlation Coefficient: Is the standardized version of the covariance:
$$\mathrm{Corr}[\boldsymbol{X}] := \frac{\mathrm{Cov}[\boldsymbol{X}]}{\sigma_{X_1}\cdots\sigma_{X_k}} \in [-1,1] \qquad (6.43)$$
$$= \begin{cases} +1 & \text{if } Y = aX + b \text{ with } a > 0, b \in \mathbb{R} \\ -1 & \text{if } Y = aX + b \text{ with } a < 0, b \in \mathbb{R} \end{cases}$$

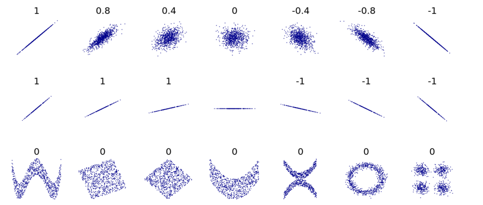


Figure 1: Several sets of (x,y) points, with their correlation coefficient

Law 6.3 Translation and Scaling:
$$\mathrm{Corr}(a + bX, c + dY) = \mathrm{sign}(b)\mathrm{sign}(d)\mathrm{Cov}(X,Y) \qquad (6.44)$$

Note

- The correlation/covariance reflects the noisiness and direction of a linear relationship (top row fig. 1), **but** not the slope of that relationship (middle row fig. 1) nor many aspects of nonlinear relationships (bottom row)
- The set in the center of fig. 1 has a slope of 0 but in that case the correlation coefficient is undefined because the variance of Y is zero.
- Zero covariance/correlation $\mathrm{Cov}(X,Y) = \mathrm{Corr}(X,Y) = 0$ implies that there does not exist a **linear** relationship between the random variables X and Y.

Difference Covariance&Correlation

1. Variance is affected by scaling and covariance not **??** and law 6.3.
2. Correlation is dimensionless, whereas the unit of the covariance is obtained by the product of the units of the two RV variables.

Law 6.4 Covariance of independent RVs: The covariance/correlation of two independent variable's (definition 6.13) is zero:
$$\mathrm{Cov}[X,Y] = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] \overset{eq.\ (6.31)}{=} \mathbb{E}[X]\,\mathbb{E}[Y] - \mathbb{E}[X]\,\mathbb{E}[Y] = 0$$

Zero covariance/correlation⇏ independence

$$\mathrm{Cov}(X,Y) = \mathrm{Corr}(X,Y) = 0 \not\Rightarrow p_{X,Y}(x,y) = p_X(x)p_Y(y)$$

For example: let $X \sim \mathcal{U}([-1,1])$ and let $Y = X^2$.

1. Clearly X and Y are dependent
2. **But** the covariance/correlation between X and Y is non-zero:
$$\mathrm{Cov}(X,Y) = \mathrm{Cov}(X,X^2) = \mathbb{E}\left[X\cdot X^2\right] - \mathbb{E}[X]\,\mathbb{E}\left[X^2\right]$$
$$= \mathbb{E}\left[X^3\right] - \mathbb{E}[X]\,\mathbb{E}\left[X^2\right] \overset{eq.\ (6.49)}{\underset{eq.\ (6.47)}{=}} 0 - 0\cdot\mathbb{E}\left[X^2\right]$$

⇒ the relationship between Y and X must be non-linear.

Definition 6.24 Autocorrelation/Crosscorrelation $\gamma(t_1,t_2)$: Describes the covariance (definition 6.21) between the two values of a stochastic process $(\boldsymbol{X}_t)_{t\in T}$ at different time points t_1 and t_2.
$$\gamma(t_1,t_2) = \mathrm{Cov}\left[\boldsymbol{X}_{t_1}, \boldsymbol{X}_{t_2}\right] = \mathbb{E}\left[\left(\boldsymbol{X}_{t_1} - \mu_{t_1}\right)\left(\boldsymbol{X}_{t_2} - \mu_{t_2}\right)\right] \qquad (6.45)$$

For zero time differences $t_1 = t_2$ the autocorrelation functions equals the variance:
$$\gamma(t,t) = \mathrm{Cov}[\boldsymbol{X}_t, \boldsymbol{X}_t] \overset{eq.\ (6.41)}{=} \mathbb{V}(\boldsymbol{X}_t) \qquad (6.46)$$

Notes

- **Hence** the autocorrelation describes the correlation of a function or signal with itself at a previous time point.
- **Given** a random time dependent variable $\boldsymbol{x}(t)$ the autocorrelation function $\gamma(t, t-\tau)$ describes how *similar* the time translated function $\boldsymbol{x}(t-\tau)$ and the original function $\boldsymbol{x}(t)$ are.
- If there exists some relation between the values of the time series that is non-random then the autocorrelation is non-zero.
- The autocorrelation is maximized/most similar for no translation $\tau = 0$ at all.

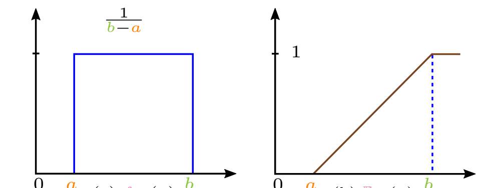
1) Distributions

Definition 6.25 Uniform Distribution $\mathcal{U}(a, b)$:
Is probability distribution, where all intervals of the **same** length on the distribution's support (definition 6.6) $\operatorname{supp}(\mathcal{U}[a, b]) = [a, b]$ are equally probable/likely.

$$f(x) = \frac{1}{b-a}\mathbb{1}_{x\in[a;b)} = \begin{cases} \frac{1}{b-a} = \text{const} & \text{if} \quad a \leqslant x \leqslant b \\ 0 & \text{else} \end{cases} \tag{6.47}$$

$$F(x) = \begin{cases} 0 & & x < a \\ \frac{x-a}{b-a} & \text{if} & a \leqslant x \leqslant b \\ 1 & & x > b \end{cases} \tag{6.48}$$

$$\mathbb{E}[X] = \frac{a+b}{2} \qquad \mathbb{V}(X) = \frac{(b-a)^2}{12} \tag{6.49}$$



(a) $f_X(x)$ (b) $F_X(x)$

Definition 6.26 Exponential Distribution $\exp(\lambda)$:

Definition 6.27 Bernoullidistribution $X \sim \operatorname{Bern}(p)$: X is a binary variable i.e. can only attain the values 0 (failure) or 1 (success) with a parameter p that signifies the success probability:

$$\operatorname{Bern}(x; p) = \begin{cases} p & \text{for } x = 1 \\ 1-p & \text{for } x = 0 \end{cases} \iff \begin{cases} P(X=1) = p \\ P(X=0) = 1-p \end{cases}$$
$$= p^x \cdot (1-p)^{1-x} \quad \text{for } x \in \{0, 1\}$$

Definition 6.28 Laplace Distribution:

Laplace Distibution $\qquad f(\boldsymbol{x}; \mu, \sigma) = \dfrac{1}{2\sigma}\exp\left(-\dfrac{|\boldsymbol{x}-\mu|}{\sigma}\right)$ (6.50)

Definition 6.29
Multivariate Normal distribution $\boldsymbol{X} \sim \mathcal{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$:
The k-multivariate Normal distribution of:
$\boldsymbol{X} = (x_1 \ \dots \ x_k)^\intercal$ a k-dimensional random vector with:
$\boldsymbol{\mu} = (\mathbb{E}[\boldsymbol{x}_1] \ \dots \ \mathbb{E}[\boldsymbol{x}_k])^\intercal$ a k-dim mean vector
and $k \times k$ **p.s.d.** covariance matrix:
$\boldsymbol{\Sigma} := \mathbb{E}[(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^\intercal] = [\operatorname{Cov}[\boldsymbol{x}_i, \boldsymbol{x}_j], 1 \leqslant i, j \leqslant k]$
is given by:

$$f_{\boldsymbol{X}}(\boldsymbol{x}_1, \dots, \boldsymbol{x}_k) = \underbrace{\frac{1}{\sqrt{(2\pi)^k \det(\boldsymbol{\Sigma})}}}_{\text{Normalisation}} \exp\left(-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})\right) \tag{6.51}$$

Definition 6.30 Jointly Gaussian Random Variables:
Two random variables $\boldsymbol{x}, \boldsymbol{y}$ both scalars or vectors, are said to be jointly Gaussian if the joint vector random variable $[\boldsymbol{x} \quad \boldsymbol{y}]^\intercal$ is again a GRV.

Corollary 6.4 Jointly GRV of GRVs: If $\boldsymbol{x}$ and $\boldsymbol{y}$ are both independent GRVs $\boldsymbol{x} \sim \mathcal{N}(\mu_x, \Sigma_x)$, $\boldsymbol{y} \sim \mathcal{N}(\mu_y, \Sigma_y)$, **then** they are jointly Gaussian (definition 6.30).

$$p(\boldsymbol{x}, \boldsymbol{y}) = p(x)p(y) \tag{6.52}$$
$$\propto \exp\left(-\frac{1}{2}\left\{(\boldsymbol{x}-\mu_x)^\intercal \Sigma_x^{-1}(\boldsymbol{x}-\mu_x) + (\boldsymbol{y}-\mu_y)^\intercal \Sigma_y^{-1}(\boldsymbol{y}-\mu_y)\right\}\right)$$
$$= \exp\left(-\frac{1}{2}\left[(\boldsymbol{x}-\mu_x)^\intercal \quad (\boldsymbol{y}-\mu_y)^\intercal\right]\begin{bmatrix} 0 & \Sigma_x^{-1} \\ \Sigma_y^{-1} & 0 \end{bmatrix}\begin{bmatrix} \boldsymbol{x}-\mu_x \\ \boldsymbol{y}-\mu_y \end{bmatrix}\right)$$

Property 0.1 Scalar Affine Transformation of GRVs:
Let $\boldsymbol{y} \in \mathbb{R}^n$ be GRV, $a \in \mathbb{R}_+, b \in \mathbb{R}$ and let $\boldsymbol{x}$ be defined by the affine transformation (definition 3.1):

$$\boldsymbol{x} = a\boldsymbol{y} + b \qquad a \in \mathbb{R}_+, b \in \mathbb{R}^d$$

Then $\boldsymbol{x}$ is a GRV with:

$$\boxed{x \sim \mathcal{N}(\mu_x, \sigma_x^2) = \mathcal{N}(a\mu + b, a^2\sigma^2)} \tag{6.53}$$

Property 0.2 Affine Transformation of GRVs: Let $\boldsymbol{y} \in \mathbb{R}^n$ be GRV, $\boldsymbol{A} \in \mathbb{R}^{d\times n}, b \in \mathbb{R}^d$ and let $\boldsymbol{x}$ be defined by the affine transformation (definition 3.1):

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{y} + b \qquad \boldsymbol{A} \in \mathbb{R}^{d\times n}, b \in \mathbb{R}^d$$

Then $\boldsymbol{x}$ is a GRV:

Property 0.3 Linear Combination of jointly GRVs:
Let $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{y} \in \mathbb{R}^m$ two jointly GRVs, and let $\boldsymbol{z}$ be defined as:

$$\boldsymbol{z} = \boldsymbol{A}_x\boldsymbol{x} + \boldsymbol{A}_y\boldsymbol{y} \qquad \boldsymbol{A}_x \in \mathbb{R}^{d\times n}, \boldsymbol{A}_x \in \mathbb{R}^{d\times m}$$

Then $\boldsymbol{z}$ is GRV.

Note

- **Joint vs. multivariate**: a joint normal distribution can be a multivariate normal distribution or a product of univariate normal distributions **but**
- Multivariate refers to the number of variables that are placed as inputs to a function.

Diagonal Covariance Matrix

For i.i.d. data the covariance matrix becomes diagonal:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_k^2 \end{bmatrix} \quad \text{and} \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{bmatrix} \tag{6.54}$$

eq. (6.51) decomposed s.t. $x_1, \dots, x_k$ become mutal independent (??):

$$p(\boldsymbol{X}) = \prod_{i=1}^{k} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(x_i - \mu_i)^2}{2\sigma_i^2}\right) \tag{6.55}$$

Proof Property 0.2 scalar case
Let $y \sim p(y) = \mathcal{N}(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$ and define $\boldsymbol{x} = a\boldsymbol{y} + b \qquad a \in \mathbb{R}_+, b \in \mathbb{R}$
Using the Change of variables formula it follows:

$$p_x(\bar{x}) \overset{??}{=} \frac{p_y(\bar{y})}{\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right|} \overset{\bar{y} = \frac{\bar{x}-b}{a}}{=} \frac{1}{a}\frac{1}{\sqrt{2\pi\mu^2}}\exp\left(-\frac{1}{2\sigma^2}\Big(\overbrace{\frac{\bar{x}-b}{a}}^{\bar{y}(\bar{x})} - \mu\Big)^2\right)$$
$$= \frac{1}{\sqrt{2\pi a^2\mu^2}}\exp\left(-\frac{1}{2\sigma^2 a^2}\Big(\bar{x} - \underbrace{b - a\mu}_{\mu_x}\Big)^2\right)$$

Hence $\qquad x \sim \mathcal{N}(\mu_x, \sigma_x^2) = \mathcal{N}(a\mu + b, a^2\sigma^2)$

□

Note

We can also verify that we have calculated the right mean and variance by:

$$\mathbb{E}[x] = \mathbb{E}[ay + b] = a\mathbb{E}[y] + b = a\mu + b$$
$$\mathbb{V}(x) = \mathbb{V}(ay + b) = a^2\mathbb{V}(y) = a^2\sigma^2$$

Proof Property 0.3
From Property 0.2 it follows immediately that $\boldsymbol{z}$ is GRV $\boldsymbol{z} \sim \mathcal{N}(\mu_z, \Sigma_z)$ with:

$$\boldsymbol{z} = \boldsymbol{A}\xi \quad \textbf{with} \quad \boldsymbol{A} = [\boldsymbol{A}_x \quad \boldsymbol{A}_y] \text{ and } \xi = (\boldsymbol{x} \quad \boldsymbol{y})$$

Knowing that $\boldsymbol{z}$ is a GRV it is sufficient to calculate μ_z and Σ_z in order to characterize its distribution:

$$\mathbb{E}[\boldsymbol{z}] = \mathbb{E}[\boldsymbol{A}_x x + \boldsymbol{A}_y y] = \boldsymbol{A}_x\mu_x + \boldsymbol{A}_y\mu_y$$
$$\mathbb{V}(\boldsymbol{z}) = \mathbb{V}(\boldsymbol{A}\xi) \overset{\text{Property 2.4}}{=} \boldsymbol{A}\mathbb{V}(\xi)\boldsymbol{A}^\intercal$$
$$= [\boldsymbol{A}_x \quad \boldsymbol{A}_y]\begin{bmatrix} \mathbb{V}(x) & \operatorname{Cov}[x, y] \\ \operatorname{Cov}[y, x] & \mathbb{V}(y) \end{bmatrix}[\boldsymbol{A}_x \quad \boldsymbol{A}_y]^\intercal$$
$$= [\boldsymbol{A}_x \quad \boldsymbol{A}_y]\begin{bmatrix} \mathbb{V}(x) & \operatorname{Cov}[x, y] \\ \operatorname{Cov}[y, x] & \mathbb{V}(y) \end{bmatrix}\begin{bmatrix} \boldsymbol{A}_x^\intercal \\ \boldsymbol{A}_y^\intercal \end{bmatrix}$$
$$= \boldsymbol{A}_x\mathbb{V}(x)\boldsymbol{A}_x^\intercal + \boldsymbol{A}_y\mathbb{V}(y)\boldsymbol{A}_y^\intercal$$
$$+ \underbrace{\boldsymbol{A}_y\operatorname{Cov}[y, x]\boldsymbol{A}_x^\intercal}_{=0\text{by independence}} + \underbrace{\boldsymbol{A}_x\operatorname{Cov}[x, y]\boldsymbol{A}_y^\intercal}_{=0\text{by independence}}$$
$$= \boldsymbol{A}_x\Sigma_x\boldsymbol{A}_x^\intercal + \boldsymbol{A}_y\Sigma_y\boldsymbol{A}_y^\intercal$$

□

Note

Can also be proofed by using the normal definition of definition 6.20 and tedious computations.

Definition 6.31 The delta function $\delta(\boldsymbol{x})$:
The delta/dirac function $\delta(\boldsymbol{x})$ is defined by:

$$\int_{\mathbb{R}} \delta(\boldsymbol{x})f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = f(0)$$

for any integrable function f on $\mathbb{R}$.
Or alternativly by:

$$\delta(x - x_0) = \lim_{\sigma\to 0}\mathcal{N}(x|x_0, \sigma) \tag{6.56}$$
$$\approx \infty\mathbb{1}_{\{x=x_0\}} \tag{6.57}$$

Property 0.4 Properties of δ:
- **Normalization**: The delta function integrates to 1:
$$\int_{\mathbb{R}}\delta(x)\,\mathrm{d}x = \int_{\mathbb{R}}\delta(x)\cdot c_1(x)\,\mathrm{d}x = c_1(0) = 1$$
where $c_1(x) = 1$ is the constant function of value 1.
- **Shifting**:
$$\int_{\mathbb{R}}\delta(x - x_0)f(x)\,\mathrm{d}x = f(x_0) \tag{6.58}$$
- **Symmetry**: $\qquad \int_{\mathbb{R}}\delta(-x)f(x)\,\mathrm{d}x = f(0)$
- **Scaling**: $\qquad \int_{\mathbb{R}}\delta(\alpha x)f(x)\,\mathrm{d}x = \frac{1}{|\alpha|}f(0)$

Note

- In mathematical terms δ is not a function but a **gernalized function**.
- We may regard $\delta(x - x_0)$ as a density with all its probability mass centered at the signle point x_0.
- Using a box/indicator function s.t. its surface is one and its width goes to zero, instead of a normaldistribution eq. (6.56) would be a non-differentiable/discret form of the dirac measure.

Definition 6.32 Discrete-time white noise: Is a random signal $\{\epsilon_t\}_{t\in T_{\text{discret}}}$ having equal intensity at different frequencies and is defined by:
- Having zero tendencies/expectation (otherwise the signal would not be random):
$$\mathbb{E}[\boldsymbol{\epsilon}[k]] = 0 \qquad \forall k \in T_{\text{discret}} \tag{6.59}$$
- Zero autocorrelation γ (definition 6.24) i.e. the signals of different times are in no-way correlated:
$$\gamma(\boldsymbol{\epsilon}[k], \boldsymbol{\epsilon}[k+n]) = \mathbb{E}[\boldsymbol{\epsilon}[k]\boldsymbol{\epsilon}[k+n]^\intercal] = \mathbb{V}(\boldsymbol{\epsilon}[k])\,\delta_{\text{discret}}[n] \qquad \forall k, n \in T_{\text{discret}} \tag{6.60}$$

With $\qquad \delta_{\text{discret}}[n] := \begin{cases} 1 & \text{if } n = 0 \\ 0 & \text{else} \end{cases}$

Proof eq. (6.62):

$$\gamma(\boldsymbol{\epsilon}[k], \boldsymbol{\epsilon}[k+n]) = \operatorname{Cov}[\boldsymbol{\epsilon}[k], \boldsymbol{\epsilon}[k+1]]$$
$$= \mathbb{E}\left[(\boldsymbol{\epsilon}[k] - \mathbb{E}[\boldsymbol{\epsilon}[k]])(\boldsymbol{\epsilon}[k+n] - \mathbb{E}[\boldsymbol{\epsilon}[k+n]])^\intercal\right]$$
$$\overset{\text{eq. (6.59)}}{=} \mathbb{E}[(\boldsymbol{\epsilon}[k])(\boldsymbol{\epsilon}[k+n])]$$

□

Definition 6.33 Continuous-time white noise: Is a random signal $(\epsilon_t)_{t\in T_{\text{continuous}}}$ having equal intensity at different frequencies and is defined by:
- Having zero tendencies/expectation (otherwise the signal would not be random):
$$\mathbb{E}[\boldsymbol{\epsilon}(t)] = 0 \qquad \forall t \in T_{\text{continuous}} \tag{6.61}$$
- Zero autocorrelation γ (definition 6.24) i.e. the signals of different times are in no-way correlated:
$$\gamma(\boldsymbol{\epsilon}(t), \boldsymbol{\epsilon}(t+\tau)) = \mathbb{E}[\boldsymbol{\epsilon}(t)\boldsymbol{\epsilon}(t+\tau)^\intercal] \tag{6.62}$$
$$\overset{\text{eq. (6.57)}}{=} \mathbb{V}(\boldsymbol{\epsilon}(t))\,\delta(t-\tau) = \begin{cases} \mathbb{V}(\boldsymbol{\epsilon}(t)) & \text{if } \tau = 0 \\ 0 & \text{else} \end{cases}$$
$$\forall t, \tau \in T_{\text{continuous}} \tag{6.63}$$

2) Sampling Random Numbers

Most math libraries have uniform **random number generator** (**RNG**) i.e. functions to generate uniformly distributed random numbers $U \sim \mathcal{U}[a, b]$ (eq. (6.47)).
Furthermore repeated calls to these RNG are independent, that is:

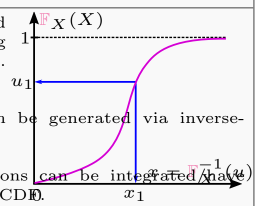
$$p_{U_1,U_2}(u_1, u_2) \overset{\text{eq. (6.23)}}{=} p_{U_1}(u_1)\cdot p_{U_2}(u_2)$$
$$= \begin{cases} 1 & \text{if } u_1, u_2 \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$

Question: using samples $\{u_1, \dots, u_n\}$ of these CRVs with uniform distribution, how can we create random numbers with arbitrary discreet or continuous PDFs?

3) Inverse-transform Technique

Idea

Can make use of section 1 and the fact that CDF are increasing functions (definition 1.7).



Advantage:
- Simple to implement
- All discrete distributions can be generated via inverse-transform technique

Drawback:
- Not all continuous distributions can be integrated/have closed form solution for their CDF.
E.g. Normal-,Gamma-,Beta-distribution.

1. Continuous Case

Definition 6.34 One Continuous Variable: Given: a desired continuous pdf f_X and uniformly distributed rn $\{u_1, u_2, \dots\}$:
1. Integrate the desired pdf f_X in order to obtain the desired cdf F_X:
$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,\mathrm{d}t \tag{6.64}$$
2. Set $F_X(X) \overset{!}{=} U$ on the range of X with $U \sim \mathcal{U}[0, 1]$.
3. Invert this equation/find the inverse $F_X^{-1}(U)$ i.e. solve:
$$U = F_X(X) = F_X\big(\underbrace{F_X^{-1}(U)}_{X}\big) \tag{6.65}$$
4. Plug in the uniformly distributed rn:
$$x_i = F_X^{-1}(u_i) \qquad \text{s.t.} \qquad x_i \sim f_X \tag{6.66}$$

Definition 6.35 Multiple Continuous Variable:
Given: a pdf of multiple rvs $f_{X,Y}$:
1. Use the product rule (eq. (6.21)) in order to decompose $f_{X,Y}$:
$$f_{X,Y} = f_{X,Y}(x, y) = f_{X|Y}(x|y)f_Y(y) \tag{6.67}$$
2. Use definition 6.36 to first get a rv for y of $Y \sim f_Y(y)$.
3. Then with this fixed y use definition 6.36 again to get a value for x of $X \sim f_{X|Y}(x|y)$.

Proof definition 6.36:
Claim: if U is a uniform rv on $[0,1]$ then $\mathbb{F}_X^{-1}(U)$ has $\mathbb{F}_X$ as its CDF.
Assume that $\mathbb{F}_X$ is strictly increasing (definition 1.7).
Then for any $u \in [0,1]$ there must exist a **unique** x s.t. $\mathbb{F}_X(x) = u$.
Thus $\mathbb{F}_X$ must be invertible and we may write $x = \underline{\mathbb{F}_X^{-1}(u)}$.
Now let a arbitrary:

$$\mathbb{F}_X(a) = \mathbb{P}(\underline{x} \leqslant a) = \mathbb{P}(\mathbb{F}_X^{-1}(U) \leqslant a)$$

Since $\mathbb{F}_X$ is strictly increasing:

$$\mathbb{P}\left(\mathbb{F}_X^{-1}(U) \leqslant a\right) = \mathbb{P}(U \leqslant \mathbb{F}_X(a)) \overset{\text{eq. (6.47)}}{=} \int_0^{\mathbb{F}_X(a)} 1 \, \mathrm{d}t = \mathbb{F}_X(a)$$

□

Note

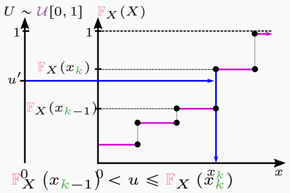
Strictly speaking we may not assume that a CDF is strictly increasing but we as all CDFs are weakly increasing (definition 1.7) we may always define an auxiliary function by its infimum:

$$\hat{\mathbb{F}}_X^{-1} := \inf\{x | \mathbb{F}_X(X) \geqslant 0\} \qquad u \in [0,1] \tag{6.68}$$

2. Discret Case

Idea

Given: a desired discret pmf p_X s.t. $\mathbb{P}(X = x_i) = p_X(x_i)$ and uniformly distributed rn $\{u_1, u_2, \ldots\}$.
Goal: given a uniformly distributed rn u determine k s.t.:



$$\sum_{i=1}^{k-1} < U \leqslant \sum_{i=1}^{k} \iff \mathbb{F}_X(x_{k-1}) < u \leqslant \mathbb{F}_X(x_k) \tag{6.69}$$

and return x_k.

Definition 6.36 One Discret Variable:

1. Compute the CDF of p_X (definition 6.8)

$$\mathbb{F}_X(x) = \sum_{t=-\infty}^{x} \mathrm{p}_X(t) \tag{6.70}$$

2. Given the uniformly distributed rn $\{u_i\}_{i=1}^n$ find k^i (≙ inversion) s.t.:

$$\mathbb{F}_X\left(x_{k(i)-1}\right) < u_i \leqslant \mathbb{F}_X\left(x_{k(i)}\right) \qquad \forall u_i \tag{6.71}$$

Proof ??: First of all notice that we can always solve for an unique x_k.

Ask: why, are Discret CRV always strictly increasing/unique?

Given a fixed x_k determine the values of u for which:

$$\mathbb{F}_X(x_{k-1}) < u \leqslant \mathbb{F}_X(x_k) \tag{6.72}$$

Now observe that:

$$u \leqslant \mathbb{F}_X(x_k) = \mathbb{F}_X(x_{k-1}) + \mathrm{p}_X(x_k)$$
$$\Rightarrow \mathbb{F}_X(x_{k-1}) < u \leqslant \mathbb{F}_X(x_{k-1}) + \mathrm{p}_X(x_k)$$

The probability of U being in $(\mathbb{F}_X(x_{k-1}), \mathbb{F}_X(x_k)]$ is:

$$\mathbb{P}\left(U \in [\mathbb{F}_X(x_{k-1}), \mathbb{F}_X(x_k)]\right) = \int_{\mathbb{F}_X(x_{k-1})}^{\mathbb{F}_X(x_k)} \mathrm{p}_U(t) \, \mathrm{d}t$$
$$= \int_{\mathbb{F}_X(x_{k-1})}^{\mathbb{F}_X(x_k)} 1 \, \mathrm{d}t = \int_{\mathbb{F}_X(x_{k-1})}^{\mathbb{F}_X(x_{k-1}) + \mathrm{p}_X(x_k)} 1 \, \mathrm{d}t = \mathrm{p}_X(x_k)$$

Hence the random variable $x_k \in \mathcal{X}$ has the pdf p_X. □

Definition 6.37
Multiple Continuous Variables (Option 1):
Given: a pdf of multiple rvs $\mathrm{p}_{X,Y}$:

1. Use the product rule (eq. (6.21)) in order to decompose $\mathrm{p}_{X,Y}$:

$$\mathrm{p}_{X,Y} = \mathrm{p}_{X,Y}(x,y) = \mathrm{p}_{X|Y}(x|y)\mathrm{p}_Y(y) \tag{6.73}$$

2. Use **??** to first get a rv for y of $Y \sim \mathrm{p}_Y(y)$.
3. Then with this fixed y use **??** again to get a value for x of $X \sim \mathrm{p}_{X|Y}(x|y)$.

Definition 6.38
Multiple Continuous Variables (Option 2):
Note: this only works if $\mathcal{X}$ and $\mathcal{Y}$ are finite.
Given: a pdf of multiple rvs $\mathrm{p}_{X,Y}$ **let** $N_x = |\mathcal{X}|$ and $N_y = |\mathcal{Y}|$ the number of elements in $\mathcal{X}$ and $\mathcal{Y}$.

Define $\mathrm{p}_Z(1) = \mathrm{p}_{X,Y}(1,1), \mathrm{p}_Z(2) = \mathrm{p}_{X,Y}(1,2), \ldots$
$\ldots, \mathrm{p}_Z(N_x \cdot N_y) = \mathrm{p}_{X,Y}(N_x, N_y)$

Then simply apply **??** to the auxillary pdf p_Z

1. Use the product rule (eq. (6.21)) in order to decompose $f_{X,Y}$:

$$f_{X,Y} = f_{X,Y}(x,y) = f_{X|Y}(x|y)f_Y(y) \tag{6.74}$$

2. Use definition 6.36 to first get a rv for y of $Y \sim f_Y(y)$.
3. Then with this fixed y use definition 6.36 again to get a value for x of $X \sim f_{X|Y}(x|y)$.

4) Examples

Example 6.1 Theorem 6.4: Let x be uniformly distributed on $[0,1]$ (definition 6.25) with pmf $\mathrm{p}_X(x)$ then it follows:

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{1}{\mathrm{p}_Y(y)} \Rightarrow \mathrm{d}x = \mathrm{d}y\,\mathrm{p}_y(y) \Rightarrow x = \int_{-\infty}^{y} \mathrm{p}_y(t)\,\mathrm{d}t = \mathbb{F}_Y(x)$$

Example 6.2 Theorem 6.4: Let

add https://www.youtube.com/watch?v=WUUb7VIRzgg

5) Proofs