Note on Mathematics of Imaging

A Preprint

Haocheng Dai

Contents

1 Calculus

Definition 1.1.

Definition 1.2. Let $f(x)$ be a function defined on an interval I, and let a be a point in I. Then, the $\boxed{\textbf{derivative}}$ of $f(x)$ at a is defined as:

$$f'(a) = \lim_{h \to 0} \frac{f(a+h) - f(a)}{h}$$

if this limit exists.

Remark 1.1. *The limit of a function is said to exist at a point a if the values of the function $f(x)$ get arbitrarily close to a specific value L as x approaches a from both the left and the right-hand sides, but not necessarily equal to L.*

Remark 1.2. *Below are four notations of derivative:*

- *Leibniz's notation: $\frac{df}{dx}, \frac{d^2 f}{dx^2}$;*
- *Lagrange's notation: $f'(x), f''(x)$;*
- *Newton's notation: $\dot{f}(x), \ddot{f}(x)$;*
- *Euler's notation: $Df(x), D^n f(x)$;*

Definition 1.3. The $\boxed{\textbf{fundamental theorem of calculus}}$ is a theorem that links the concept of differentiating a function with the concept of integrating a function.

- The first fundamental theorem of calculus: Let F be the function defined, for all $x \in [a, b]$, by

  $$F(x) = \int_a^x f(t)\, dt$$

  Then F is uniformly continuous on $[a, b]$ and differentiable on the open interval (a, b), and

  $$F'(x) = f(x)$$

  for all $x \in (a, b)$, so F is an antiderivative of f.

- The second fundamental theorem of calculus (Newton–Leibniz axiom): Let f be a real-valued function on a closed interval $[a, b]$ and F a continuous function on $[a, b]$ which is an antiderivative of f in (a, b):

  $$F'(x) = f(x).$$

  If f is Riemann integrable on $[a, b]$, then

  $$\int_a^b f(x)\, dx = F(b) - F(a).$$

Remark 1.3. *How to understand the association between the "area under the curve" and the "slope"? By looking at the Newton-Leibniz axiom, divide both side of the equation by $b - a$, we can have*

$$\begin{aligned} \int_a^b f(x)\, dx &= F(b) - F(a) \\ \underbrace{\frac{\int_a^b f(x)\, dx}{b-a}}_{\text{average height of the curve}} &= \underbrace{\frac{F(b) - F(a)}{b-a}}_{\text{average slope of the antiderivative}} . \end{aligned}$$

Namely the average height of the curve f is equivalent to the average slope of the antiderivative F in $[a, b]$.

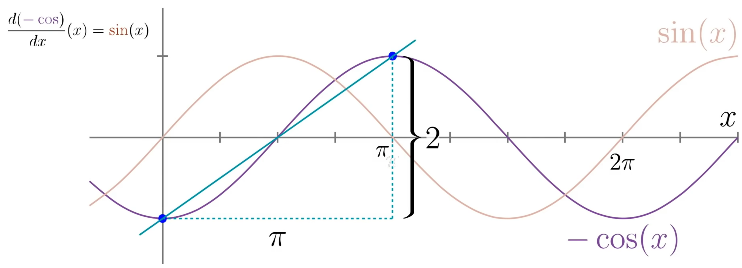


Figure 1: Association between the "area under the curve" and the "slope", where $\sin(x)$ is the curve and $-\cos(x)$ is the antiderivative of $\sin(x)$.

Definition 1.4. **Integration by parts (Change of variables)** formula states:

$$\int_a^b u\,dv = [uv]_a^b - \int_a^b v\,du$$
$$\int_a^b uv'\,dx = [uv]_a^b - \int_a^b u'v\,dx.$$

Example 1. Integration by parts twice:

$$\int_a^b f''(x)g(x)dx = \int_a^b g(x)\,df'(x) = [g(x)f'(x)]_a^b - \int_a^b f'(x)\,dg(x)$$

where

$$\begin{aligned}-\int_a^b f'(x)\,dg(x) &= -\int_a^b f'(x)g'(x)\,dx\\ &= -\int_a^b g'(x)\,df(x)\\ &= [-g'(x)f(x)]_a^b + \int_a^b f(x)\,dg'(x)\\ &= [-g'(x)f(x)]_a^b + \int_a^b f(x)g''(x)\,dx\end{aligned}$$

Hence

$$\int_a^b f''(x)g(x)dx = [g(x)f'(x)]_a^b - [g'(x)f(x)]_a^b + \int_a^b f(x)g''(x)\,dx$$

Definition 1.5. **Fourier series** is an expansion of a periodic function into a sum of trigonometric functions.

The Fourier series coefficients can be defined by the integrals in the sine-cosine form:

$$A_n = \frac{2}{P}\int_{-P/2}^{P/2} s(x)\cos\left(\frac{2\pi nx}{P}\right)dx \qquad \text{for } n \geq 1$$
$$B_n = \frac{2}{P}\int_{-P/2}^{P/2} s(x)\sin\left(\frac{2\pi nx}{P}\right)dx \qquad \text{for } n \geq 1$$

where P is the function's period. [1] With these coefficients defined the Fourier series is:

$$s(x) \sim A_0 + \sum_{n=1}^{\infty}\left(A_n\cos\left(\frac{2\pi nx}{P}\right) + B_n\sin\left(\frac{2\pi nx}{P}\right)\right)$$

[1] It is notable that, A_0 is the average value of the function $s(x)$. This is a property that extends to similar transforms such as the Fourier transform.

The Fourier series coefficients can also be defined by the integrals in the exponential form:

$$
\begin{aligned}
c_n &= \frac{1}{P}\int_{-P/2}^{P/2} s(x)e^{-\frac{2\pi i n x}{P}}\,dx && \text{for all integers } n\\
c_n &= \frac{1}{P}\int_{-P/2}^{P/2} s(x)\left(\cos\left(-\frac{2\pi n x}{P}\right) + i\sin\left(-\frac{2\pi n x}{P}\right)\right)\,dx && \text{for all integers } n\\
c_n &= \frac{1}{P}\int_{-P/2}^{P/2} s(x)\cos\left(-\frac{2\pi n x}{P}\right)\,dx + i\int_{-P/2}^{P/2} s(x)\sin\left(-\frac{2\pi n x}{P}\right)\,dx && \text{for all integers } n\\
c_n &= (A_n - iB_n)/2 && \text{for } n > 0\\
c_n &= (A_{-n} + iB_{-n})/2 && \text{for } n < 0\\
c_0 &= A_0 && \text{for } n = 0
\end{aligned}
$$

$$
\begin{aligned}
s(x) &= \sum_{n=-\infty}^{\infty} c_n \cdot e^{\frac{2\pi i n x}{P}} && \triangleright \text{ exponential form}\\
&= \sum_{n=1}^{\infty} \frac{A_n - iB_n}{2}\cdot\left(\cos\left(\frac{2\pi n x}{P}\right) + i\sin\left(\frac{2\pi n x}{P}\right)\right)\\
&+ \sum_{n=-\infty}^{-1} \frac{A_n + iB_n}{2}\cdot\left(\cos\left(\frac{2\pi n x}{P}\right) - i\sin\left(\frac{2\pi n x}{P}\right)\right)\\
&+ A_0\\
&= A_0 + \sum_{n=1}^{\infty} A_n\cos\left(\frac{2\pi n x}{P}\right) + B_n\sin\left(\frac{2\pi n x}{P}\right) && \triangleright \text{ sine-cosine form}
\end{aligned}
$$

Remark 1.4. *The Fourier series is an example of a trigonometric series, but not all trigonometric series are Fourier series.*

Remark 1.5. *When you express a function with a Fourier series you are actually performing the Gram-Schimdt process, by projecting a function onto a basis of Sine and Cosine functions*

$$
\begin{aligned}
A_n &= \left\langle s(x), \cos\left(\frac{2\pi n x}{P}\right)\right\rangle = \frac{2}{P}\int_{-P/2}^{P/2} s(x)\cos\left(\frac{2\pi n x}{P}\right)\,dx && \textit{for } n \geq 1\\
B_n &= \left\langle s(x), \sin\left(\frac{2\pi n x}{P}\right)\right\rangle = \frac{2}{P}\int_{-P/2}^{P/2} s(x)\sin\left(\frac{2\pi n x}{P}\right)\,dx && \textit{for } n \geq 1
\end{aligned}
$$

Example 2. Consider a sawtooth function:

$$
\begin{aligned}
s(x) &= \frac{x}{\pi}, \quad \text{for} - \pi < x < \pi,\\
s(x + 2\pi k) &= s(x), \quad \text{for} - \pi < x < \pi \text{ and } k \in \mathbb{Z}.
\end{aligned}
$$

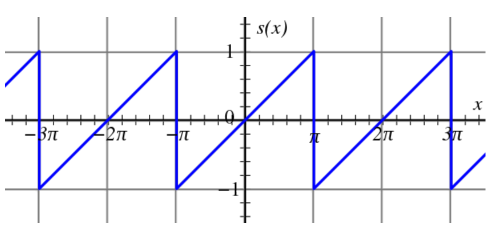


Figure 2: Plot of the sawtooth wave, a periodic continuation of the linear function $s(x) = x/\pi$ on the interval $(-\pi, \pi]$

In this case, the Fourier coefficients are given by

$$
\begin{aligned}
A_n &= \frac{2}{2\pi}\int_{-\pi}^{\pi} s(x)\cos(nx)\,dx = 0, \quad n \geq 0. && \triangleright \text{ Integration of odd function is 0.}\\
B_n &= \frac{2}{2\pi}\int_{-\pi}^{\pi} s(x)\sin(nx)\,dx \\
&= \frac{1}{\pi}\int_{-\pi}^{\pi} \frac{x}{\pi}\sin(nx)\,dx \\
&= \frac{1}{\pi^2}\int_{-\pi}^{\pi} x\sin(nx)\,dx \\
&= \frac{1}{n\pi^2}\int_{-\pi}^{\pi} x\,d(-\cos(nx)) && \triangleright \text{ Integration by parts.}\\
&= \frac{1}{n\pi^2}\left([-x\cos(nx)]_{-\pi}^{\pi} - \int_{-\pi}^{\pi} -\cos(nx)\,dx\right) \\
&= \frac{1}{n\pi^2}\left([-x\cos(nx)]_{-\pi}^{\pi} + \left[\frac{1}{n}\sin(nx)\right]_{-\pi}^{\pi}\right) \\
&= \frac{1}{n\pi^2}\left(-\pi\cos(n\pi) - (\pi\cos(-n\pi)) + \frac{1}{n}\sin(n\pi) - \frac{1}{n}\sin(-n\pi)\right) \\
&= \frac{1}{n\pi^2}\left(-\pi\cos(n\pi) - \pi\cos(n\pi) + \frac{1}{n}\sin(n\pi) + \frac{1}{n}\sin(n\pi)\right) \\
&= \frac{1}{n\pi^2}\left(-2\pi\cos(n\pi) + \frac{2}{n}\sin(n\pi)\right) \\
&= -\frac{2}{\pi n}\cos(n\pi) + \frac{2}{\pi^2 n^2}\sin(n\pi) \\
&= \frac{2\,(-1)^{n+1}}{\pi n}, \quad n \geq 1.
\end{aligned}
$$

It can be shown that the Fourier series converges to $s(x)$ at every point x where s is differentiable, and therefore:

$$
\begin{aligned}
s(x) &= A_0 + \sum_{n=1}^{\infty}\left(A_n\cos(nx) + B_n\sin(nx)\right) \\
&= \frac{2}{\pi}\sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n}\sin(nx), \quad \text{for} \quad x - \pi \notin 2\pi\mathbb{Z}.
\end{aligned}
$$

Below is the visualization of the evolution of Fourier approximation:

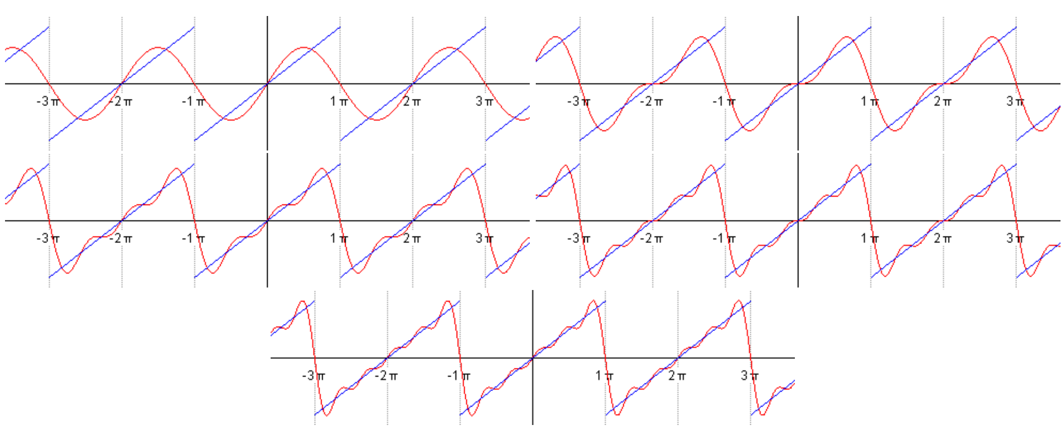


Figure 3: Plots of the first five successive partial Fourier series.

Definition 1.6. $\boxed{\textbf{Taylor series}}$ of a real or complex-valued function $f(x)$ that is infinitely differentiable at a real or complex number a is the power series

$$t(x)=\sum_{n=0}^{\infty}\frac{f^{(n)}(a)}{n!}(x-a)^n=f(a)+\frac{f'(a)}{1!}(x-a)+\frac{f''(a)}{2!}(x-a)^2+\frac{f'''(a)}{3!}(x-a)^3+\cdots,$$

where $f^{(n)}(a)$ denotes the nth derivative of f evaluated at the point a. Function $t(x)$ approximates the $f(x)$ around an arbitrarily small neighborhood of a. When $a=0$, the series is also called a Maclaurin series.

Proof. How are the coefficients ahead of the polynomial terms determined? Assuming the Taylor series of the function $f(x)$ is $t(x)=\sum_{n=0}^{\infty}c_n(x-a)^n$. In order to match the n-th derivative with the original function $f(x)$ at $x=a$, we have

$$t^{(n)}(x)=c_nn!+\sum_{m=n+1}^{\infty}\frac{m!}{(m-3)!}(x-a)^{m-3}=f^{(n)}(x)$$

When $x=a$, we have

$$c_nn!=f^{(n)}(a)$$
$$c_n=\frac{f^{(n)}(a)}{n!}$$

□

Remark 1.6. *When you look at the first-order approximation $t(x)=f(a)+f'(a)(x-a)$, it is very similar to what we did in Euler integration.*

Example 3.

- The Maclaurin series of the exponential function e^x is
$$\sum_{n=0}^{\infty}\frac{x^n}{n!}=\frac{x^0}{0!}+\frac{x^1}{1!}+\frac{x^2}{2!}+\frac{x^3}{3!}+\frac{x^4}{4!}+\frac{x^5}{5!}+\cdots.$$
It converges for all x, namely the radius of convergence is infinity.
- The Maclaurin series of the exponential function $\sin(x)$ is
$$\sin x=\sum_{n=0}^{\infty}\frac{(-1)^n}{(2n+1)!}x^{2n+1}=x-\frac{x^3}{3!}+\frac{x^5}{5!}-\frac{x^7}{7!}+\cdots.$$
It converges for all x, namely the radius of convergence is infinity.
- The Maclaurin series of the exponential function $\cos(x)$ is
$$\cos x=\sum_{n=0}^{\infty}\frac{(-1)^n}{(2n)!}x^{2n}=1-\frac{x^2}{2!}+\frac{x^4}{4!}-\frac{x^6}{6!}+\cdots.$$
It converges for all x, namely the radius of convergence is infinity.

Example 4. When we are propagating an integral curve, we are actually using the Taylor series:

$$t(x+a)=f(x)+\frac{f'(x)}{1!}(x+a-x)+\frac{f''(x)}{2!}(x+a-x)^2+\frac{f'''(x)}{3!}(x+a-x)^3+\cdots,$$

In practice, we use first-order expansion:

$$t(x+a)=f(x)+af'(x).$$

To be more precise, we can also use second-order expansion:

$$t(x+a)=f(x)+af'(x)+\frac{a^2}{2!}f''(x).$$

Definition 1.7. $\boxed{\textbf{Euler's formula}}$ states that for any real number x:

$$e^{ix} = \cos x + i \sin x.$$

Proof.

- Using Taylor series: According to the Maclaurin series of $e^x, \sin(x), \cos(x)$, we can have the following deviation

$$\begin{aligned}
e^{ix} &= \sum_{n=0}^{\infty} \frac{(ix)^n}{n!} \\
&= \frac{(ix)^0}{0!} + \frac{(ix)^1}{1!} + \frac{(ix)^2}{2!} + \frac{(ix)^3}{3!} + \frac{(ix)^4}{4!} + \frac{(ix)^5}{5!} + \frac{(ix)^6}{6!} + \frac{(ix)^7}{7!} + \cdots \\
&= \frac{x^0}{0!} + \frac{ix}{1!} - \frac{x^2}{2!} - \frac{ix^3}{3!} + \frac{x^4}{4!} + \frac{ix^5}{5!} - \frac{x^6}{6!} - \frac{ix^7}{7!} + \cdots \\
&= \cos(x) + i \sin(x)
\end{aligned}$$

- Using differentiation: Consider the function $f(\theta)$

$$f(\theta) = \frac{\cos\theta + i\sin\theta}{e^{i\theta}} = e^{-i\theta}\left(\cos\theta + i\sin\theta\right)$$

for real θ. Differentiating gives by the product rule

$$f'(\theta) = e^{-i\theta}\left(i\cos\theta - \sin\theta\right) - ie^{-i\theta}\left(\cos\theta + i\sin\theta\right) = 0$$

Thus, $f(\theta)$ is a constant. Since $f(0) = 1$, then $f(\theta) = 1$ for all real θ, and thus

$$e^{i\theta} = \cos\theta + i\sin\theta.$$

□

Definition 1.8. $\boxed{\textbf{Euler-Lagrange equation}}$ is defined as

$$\frac{\partial F}{\partial y} - \frac{\partial}{\partial x}\left(\frac{\partial F}{\partial y'}\right) = 0.$$

which is used to find a $y = f(x)$ making this integral

$$L(y) = \int_{x_1}^{x_2} F(x, y, y')dx$$

stationary.

Example 5. Suppose A and B are two points in an Euclidean space. We want to find the geodesic between A and B.

Solution. We would like to minimize

$$L = \int_A^B 1ds, \text{ where } ds = \sqrt{(dx)^2 + (dy)^2} = \sqrt{1 + (y')^2}dx$$

which can be written in another form

$$L = \int_A^B \sqrt{1 + (y')^2}dx$$

We need to find a $y(x)$ which minimize L, where

$$F = \sqrt{1 + (y')^2}$$

Substituting it into the Euler-Lagrange equation, we have

$$\begin{aligned}
\frac{\partial F}{\partial y} - \frac{d}{dx}\left(\frac{\partial F}{\partial y'}\right) &= 0 \\
-\frac{d}{dx}\left(\frac{y'}{\sqrt{1+(y')^2}}\right) &= 0 \\
\frac{y'}{\sqrt{1+(y')^2}} &= c \\
(y')^2 &= \frac{c^2}{1-c^2} \\
y' &= c_1 \\
y &= c_1 x + c_2
\end{aligned}$$

2 Vector Space

Definition 2.1. $\boxed{\textbf{Vector space}}$ over a field F is a non-empty set V together with two binary operations that satisfy the eight axioms listed below. In this context, the elements of V are commonly called vectors, and the elements of F are called scalars.

- Vector addition: for any $\mathbf{x}, \mathbf{y} \in V$, $\mathbf{x} + \mathbf{y} \in V$
- Scalar multiplication: for any $\mathbf{x} \in V, \alpha \in F$, $\alpha\mathbf{x} \in V$

To have a vector space, the following axioms must be satisfied:

- $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$
- $(\mathbf{x} + \mathbf{y}) + \mathbf{z} = \mathbf{x} + (\mathbf{y} + \mathbf{z})$
- There is a null vector $\theta \in X$ such that $\mathbf{x} + \theta = \mathbf{x}$ for every $\mathbf{x} \in X$
- $\alpha(\mathbf{x} + \mathbf{y}) = \alpha\mathbf{x} + \alpha\mathbf{y}$; $(\alpha + \beta)\mathbf{x} = \alpha\mathbf{x} + \beta\mathbf{x}$
- $(\alpha\beta)\mathbf{x} = \alpha(\beta\mathbf{x})$
- $0\mathbf{x} = \theta$; $1\mathbf{x} = \mathbf{x}$

Example 6. $\mathbb{R}^n, \mathbb{C}^n$, function spaces, and linear equations are all vector spaces.

Definition 2.2. $\boxed{\textbf{Subspace of } V}$ is a nonempty subset W of a vector space V that is closed under addition and scalar multiplication (and therefore contains the $\mathbf{0}$-vector of V)

Example 7. A hyperplane passing the origin point in the vector space V is a subspace of the V.

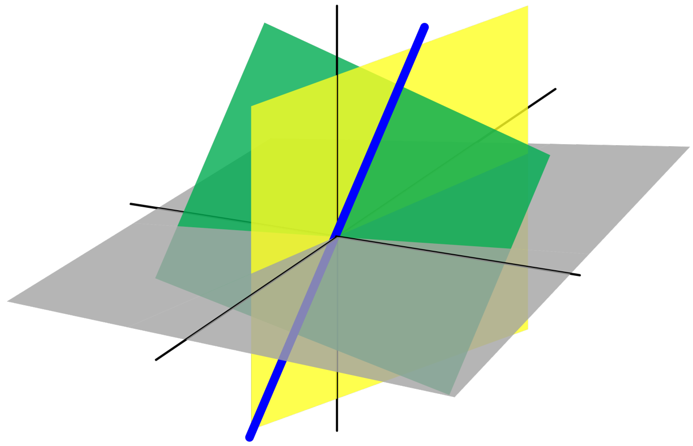


Figure 4: A line passing through the origin (blue, thick) in $\mathbb{R}^3$ is a linear subspace. It is the intersection of two planes (green and yellow).

Definition 2.3. $\boxed{\textbf{Norm}}$ $\|\cdot\| : X \times X \to \mathbb{R}$ is a mapping that satisfies the following axioms:

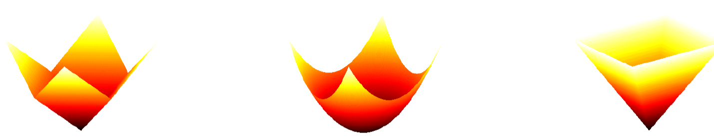
1. $\|\mathbf{x}\| \geq 0$ for all $\mathbf{x} \in X$, $\|\mathbf{x}\| = 0$ if and only in $\mathbf{x} = \theta$
2. $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ for each $\mathbf{x}, \mathbf{y} \in X$ ▷ Triangle inequality
3. $\|\alpha\mathbf{x}\| = |\alpha| \cdot \|\mathbf{x}\|$ for all scalars α and each $\mathbf{x} \in X$

where X is a vector space.

Remark 2.1. *Norm and inner product are two independent concepts. Norm is not necessarily defined by inner product. But when the Banach space's norm is defined by inner product (namely the norm satisfies parallelogram law), then it is called Hilbert space.*

Example 8. The norm is clearly an abstraction of our usual concept of length. For continuous situation, the **supremum norm** $\|f\|_\infty$ is the supremum (lowest upper bound) of all elements of its domain evaluated in f. For discrete situation, the sup norm equals to the maximum of absolute values of its components, namely $\|f\|_\infty = \max |f_i|$.

Remark 2.2. *If* $f : \mathbb{R}^n \to \mathbb{R}, f(\mathbf{x}) = \|\mathbf{x}\|_p, p \geq 1$, *then* f *is convex.*



(a) $p = 1$ (b) $p = 2$ (c) $p = \infty$

Figure 5: Image of 2D norm

Definition 2.4. $\boxed{\textbf{Normed linear vector space}}$ is a vector space X on which there is defined a real-valued function that maps each element $\mathbf{x}$ in X into a real number $\|\mathbf{x}\|$.

Definition 2.5. $\boxed{\textbf{Pre-Hilbert space}}$ is a linear vector space X together with an inner product defined on $X \times X$.

Definition 2.6. $\boxed{\textbf{Cauchy sequence}}$ is a sequence $\{x_n\}$ in a normed space such that $\|x_n - x_m\| \to 0$ as $n, m \to \infty$.

Remark 2.3. *In a normed space, every convergent sequence is a Cauchy sequence, however, a Cauchy sequence may not be convergent.*

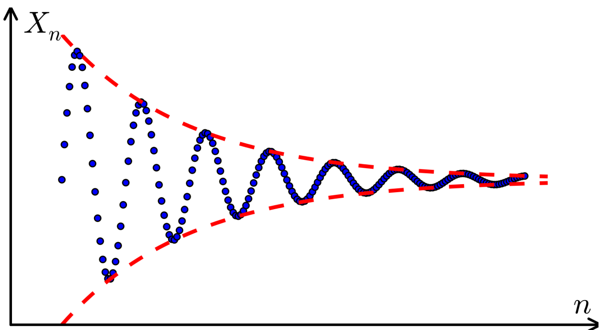


Figure 6: Example of Cauchy sequence

Definition 2.7. A normed linear vector space X is **complete** is every Cauchy sequence from X has a limit in X. The limit is also a vector.

Example 9.

- For a finite-dimensional vector space like $\mathbb{R}^n$, it is complete — every Cauchy sequence from $\mathbb{R}^n$ has a limit also lives in $\mathbb{R}^n$.
- For an infinite-dimensional vector space (a space of continuous functions) X, it is not complete as the limit of the below Cauchy sequence of x_n does not live in the space of continuous functions — the limit is a non-continuous function.

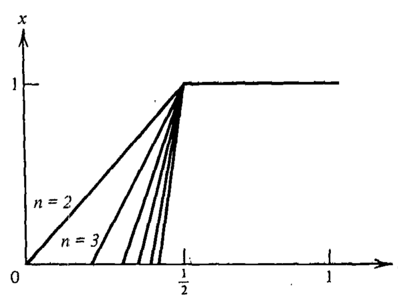


Figure 7: A Cauchy sequence makes continuous function space not complete.[Luenberger, 1997]

Definition 2.8. **Banach space** is a complete normed linear vector space.

Remark 2.4. *In a Banach space, norms are rigorously defined, which necessitates the validity of the triangle inequality. Therefore, it is a fundamental characteristic of Banach spaces that the triangle inequality holds.*

Definition 2.9. **Inner product** [Luenberger, 1997] $\langle \cdot, \cdot \rangle : X \times X \to \mathbb{R}$ is a mapping that satisfies the following axioms:

1. $\langle \mathbf{x}, y \rangle = \langle y, \mathbf{x} \rangle$
2. $\langle \mathbf{x} + y, z \rangle = \langle \mathbf{x}, z \rangle + \langle y, z \rangle$
3. $\langle \lambda \mathbf{x}, y \rangle = \lambda \langle \mathbf{x}, y \rangle$
4. $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$ and $\langle \mathbf{x}, \mathbf{x} \rangle = 0$ if and only if $\mathbf{x} = \theta$

where X is a vector space.

Definition 2.10. In a pre-Hilber space, two vector $\mathbf{v}_1, \mathbf{v}_2$ (function f, g) are said to be $\boxed{\textbf{orthogonal}}$ if

$$\langle \mathbf{v}_1, \mathbf{v}_2 \rangle = 0$$

or

$$\langle f, g \rangle = \frac{1}{L} \int_{-L}^{L} f(x) g(x) dx = 0$$

Example 10. $\sin(x)$ and $\cos(x)$ are two functions orthogonal to each other, as

$$\langle \sin(x), \cos(x) \rangle = \frac{1}{L} \int_{-L}^{L} \sin(x) \cos(x) \, dx = 0$$

Example 11. That is, $P_n(x)$ is a polynomial of degree n, such that

$$\int_{-1}^{1} P_m(x) P_n(x) \, dx = 0 \quad \text{if } n \neq m.$$

An especially compact expression for the Legendre polynomials is given by Rodrigues' formula:

n	$P_n(x)$
0	1
1	x
2	$\frac{1}{2}\left(3x^2 - 1\right)$
3	$\frac{1}{2}\left(5x^3 - 3x\right)$
n	$\frac{1}{2^n n!} \frac{d^n}{dx^n} (x^2 - 1)^n$

Example 12. For $A, B \in \mathrm{GL}(n)$, the inner product $\langle A, B \rangle = \mathrm{Tr}(A^\top B)$ and the associated norm $\|A\|_2 = \mathrm{Tr}(A^\top A)^{1/2}$.

Definition 2.11. $\boxed{\textbf{Hilbert space}}$ is a complete pre-Hilbert space or a Banach space whose norm is defined by the inner product.

Remark 2.5. *Any finite-dimensional inner product space (with inner-product-induced norm) is a Hilbert space. For example, every Cauchy sequence from $\mathbb{R}^n$ has a limit that also lives in $\mathbb{R}^n$.*

Theorem 2.1. ***Cauchy-Schwarz Inequality***. *If $p = 2$ and $q = 2$ and if $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \cdots]^\top \in l_2, \mathbf{y} = [\mathbf{y}_1, \mathbf{y}_2, \cdots]^\top \in l_2$, then*

$$\sum_{i=1}^{\infty} |\mathbf{x}_i \mathbf{y}_i| \leq \|\mathbf{x}\|_2 \cdot \|\mathbf{y}\|_2$$

$$\langle \mathbf{x}, \mathbf{y} \rangle^2 \leq \|\mathbf{x}\|_2^2 \cdot \|\mathbf{y}\|_2^2$$

Remark 2.6. *Cauchy-Schwarz inequality holds in Hilbert space.*

Theorem 2.2. ***Hölder Inequality***. *If p and q are positive numbers $1 \leq p \leq \infty, 1 \leq q \leq \infty$, such that $1/p + 1/q = 1$ and if $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \cdots]^\top \in l_p, \mathbf{y} = [\mathbf{y}_1, \mathbf{y}_2, \cdots]^\top \in l_q$, then*

$$\sum_{i=1}^{\infty} |\mathbf{x}_i \mathbf{y}_i| \leq \|\mathbf{x}\|_p \cdot \|\mathbf{y}\|_q$$

$$\|\mathbf{x} \odot \mathbf{y}\|_1 \leq \|\mathbf{x}\|_p \cdot \|\mathbf{y}\|_q$$

Equality holds if and only if $\left(\frac{|\mathbf{x}_i|}{\|\mathbf{x}\|_p}\right)^{1/q} = \left(\frac{|\mathbf{y}_i|}{\|\mathbf{y}\|_q}\right)^{1/p}$ for each i.

Remark 2.7. *Holder inequality is a generalization of Cauchy-Schwarz inequality.*

Remark 2.8. *Cauchy-Schwarz inequality is intuitive to understand as*

$$\begin{aligned} \cos(\theta)^2 &\leq 1 \\ |\mathbf{x}\|_2^2 \cdot \|\mathbf{y}\|_2^2 \cdot \cos(\theta)^2 &\leq |\mathbf{x}\|_2^2 \cdot \|\mathbf{y}\|_2^2 \\ \langle \mathbf{x}, \mathbf{y} \rangle^2 &\leq \|\mathbf{x}\|_2^2 \cdot \|\mathbf{y}\|_2^2 \end{aligned}$$

Theorem 2.3. ***Minkowski Inequality*** *(triangle inequality). If $\mathbf{x}$ and $\mathbf{y}$ are in l_p, $1 \leq p \leq \infty$, then*

$$\|\mathbf{x} + \mathbf{y}\|_p \leq \|\mathbf{x}\|_p + \|\mathbf{y}\|_p$$

Equality holds if and only if $k_1\mathbf{x} = k_2\mathbf{y}$ for some positive constants k_1 and k_2.

Theorem 2.4. ***Divergence Theorem.*** *Letting φ be a C^1 vector field, defined on Ω, which is a region in the plane with boundary $\partial\Omega$, then*

$$\int_\Omega \mathrm{div}\varphi dx = \int_{\partial\Omega} \langle \varphi, N \rangle dl$$

where N is the outward normal to Ω and $\mathrm{div}(\varphi) = \mathrm{trace}(D\varphi)$.

Definition 2.12. $\boxed{l^p \textbf{ space}}$ consists of all sequences of scalars $\{\xi_1, \xi_2, \cdots\}$ for which

$$\sum_{i=1}^{\infty} |\xi_i|^p < \infty$$

where $1 \leq p < \infty$.

The norm of an element $\mathbf{x} = \{\xi_i\}$ in l^p is defined as

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{\infty} |\xi_i|^p \right)^{1/p}$$

Definition 2.13. $\boxed{L^p[a,b] \textbf{ space (Lebesgue space)}}$ consists of all functions $f(u)$ for which

$$\int_a^b |f(u)|^p du < \infty$$

where $1 \leq p < \infty$.

The norm of an element $f(u)$ in L^p is defined as

$$\|f\|_p = \left(\int_a^b |f(u)|^p du \right)^{1/p}$$

The L^p-functions are the functions for which this integral converges.

Remark 2.9. *Always remember the* ***absolute value sign*** *in norm calculation.*

Remark 2.10. *According to Minkoski inequality, the triangle inequality, which is essential for inner product-induced norms, holds only when $p \geq 1$ (convex). That is why L^p and l^p spaces requires $1 \leq p < \infty$.*

Proof. Let $\mathbf{a} = [1, 0]^\top, \mathbf{b} = [0, 1]^\top$, we have $\mathbf{a} + \mathbf{b} = [1, 1]^\top$

$$\begin{aligned} \|\mathbf{a}\|_{\frac{1}{2}} &= 1 \\ \|\mathbf{b}\|_{\frac{1}{2}} &= 1 \\ \|\mathbf{a}\|_{\frac{1}{2}} + \|\mathbf{b}\|_{\frac{1}{2}} &= 2 \\ \|\mathbf{a} + \mathbf{b}\|_{\frac{1}{2}} &= (1^{\frac{1}{2}} + 1^{\frac{1}{2}})^2 = 4 \\ \|\mathbf{a}\|_{\frac{1}{2}} + \|\mathbf{b}\|_{\frac{1}{2}} &< \|\mathbf{a} + \mathbf{b}\|_{\frac{1}{2}} \end{aligned}$$

The triangle inequality does not hold. □

Remark 2.11. *l^p, L^p is Banach space (complete and norm defined). l^p, L^p space is Hilbert space (Banach space with inner-product-induced norm) if and only if $p = 2$.*

Proof. Consider two simple functions:

$$\begin{aligned}
f(x) &= 1[0, 1/2] \\
g(x) &= 1[1/2, 1] \\
\|f\|_p^2 &= \left(\left(\int |f(x)|^p \, dx\right)^{\frac{1}{p}}\right)^2 \\
&= \left(\left(\frac{1}{2}\right)^{\frac{1}{p}}\right)^2 = \left(\frac{1}{2}\right)^{\frac{2}{p}} = \|g\|_p^2 \\
2\|f\|_p^2 + 2\|g\|_p^2 &= 4 \times \left(\frac{1}{2}\right)^{\frac{2}{p}} \\
\|f+g\|_p^2 &= \left(\left(\int |f(x)+g(x)|^p \, dx\right)^{\frac{1}{p}}\right)^2 = 1 \\
\|f-g\|_p^2 &= \left(\left(\int |f(x)-g(x)|^p \, dx\right)^{\frac{1}{p}}\right)^2 = 1 \\
\|f+g\|_p^2 + \|f-g\|_p^2 &= 2
\end{aligned}$$

Hence, parallelogram law $\|f+g\|_p^2 + \|f-g\|_p^2 = 2\|f\|_p^2 + 2\|g\|_p^2$ holds if and only if $p = 2$. □

Remark 2.12.

l^p/Lebesgue space L^p with $p \geq 1$ →Banach space (norm defined, hence triangle inequality holds)

l^p/Lebesgue space L^p with $p = 2$ →Banach space with parallelogram law holds

⇔Banach space with inner product induced norm

⇔Hilbert space

Remark 2.13. *The parallelogram law is a necessary and sufficient condition for a norm to be defined by an inner product. See proof below.*

Theorem 2.5. *Norm satisfying parallelogram law → norm is induced by inner product:*

Let V be a vector space over $\mathbb{R}$ and $\|\cdot\| : V \to \mathbb{R}$ be a norm on V such that:

$$\|\mathbf{x}+\mathbf{y}\|^2 + \|\mathbf{x}-\mathbf{y}\|^2 = 2\|\mathbf{x}\|^2 + 2\|\mathbf{y}\|^2$$

for each $\mathbf{x}, \mathbf{y} \in V$.

Then the function $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ defined by:

$$\langle \mathbf{x}, \mathbf{y} \rangle = \frac{\|\mathbf{x}+\mathbf{y}\|^2 - \|\mathbf{x}-\mathbf{y}\|^2}{4}$$

for each $\mathbf{x}, \mathbf{y} \in V$, is an inner product on V.

Proof. 1. $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$:

$$\begin{aligned}
\langle \mathbf{y}, \mathbf{x} \rangle &= \frac{\|\mathbf{y}+\mathbf{x}\|^2 - \|\mathbf{y}-\mathbf{x}\|^2}{4} \\
&= \frac{\|\mathbf{x}+\mathbf{y}\|^2 - \|-(\mathbf{x}-\mathbf{y})\|^2}{4} \\
&= \frac{\|\mathbf{x}+\mathbf{y}\|^2 - \|\mathbf{x}-\mathbf{y}\|^2}{4} && \triangleright \|\alpha\mathbf{x}\| = |\alpha| \cdot \|\mathbf{x}\| \\
&= \langle \mathbf{x}, \mathbf{y} \rangle
\end{aligned}$$

2. $\langle \mathbf{x}+\mathbf{y}, \mathbf{z}\rangle = \langle \mathbf{x}, \mathbf{z}\rangle + \langle \mathbf{y}, \mathbf{z}\rangle$:

$$\begin{aligned}\langle \mathbf{x}, \mathbf{z}\rangle + \langle \mathbf{y}, \mathbf{z}\rangle &= \frac{1}{4}\left(\|\mathbf{x}+\mathbf{z}\|^2 - \|\mathbf{x}-\mathbf{z}\|^2\right) + \frac{1}{4}\left(\|\mathbf{y}+\mathbf{z}\|^2 - \|\mathbf{y}-\mathbf{z}\|^2\right)\\ &= \frac{1}{4}\left(\|\mathbf{x}+\mathbf{z}\|^2 - \|\mathbf{x}-\mathbf{z}\|^2 + \|\mathbf{y}+\mathbf{z}\|^2 - \|\mathbf{y}-\mathbf{z}\|^2\right)\\ &= \frac{1}{4}\left(\|\mathbf{x}+\mathbf{z}\|^2 + \|\mathbf{y}\|^2 - \|\mathbf{x}-\mathbf{z}\|^2 - \|\mathbf{y}\|^2 + \|\mathbf{x}\|^2 + \|\mathbf{y}+\mathbf{z}\|^2 - \|\mathbf{x}\|^2 - \|\mathbf{y}-\mathbf{z}\|^2\right)\end{aligned}$$

Note that by hypothesis of inner product, we have $\|f+g\|^2 + \|f-g\|^2 = 2\left(\|f\|^2 + \|g\|^2\right)$ for each $f, g \in V$.

Setting $f = \mathbf{x}+\mathbf{z}$ and $g = \mathbf{y}$, we have:

$$\|\mathbf{x}+\mathbf{y}+\mathbf{z}\|^2 + \|\mathbf{x}-\mathbf{y}+\mathbf{z}\|^2 = 2\left(\|\mathbf{x}+\mathbf{z}\|^2 + \|\mathbf{y}\|^2\right)$$

Setting $f = \mathbf{x}-\mathbf{z}$ and $g = \mathbf{y}$, we have:

$$\|\mathbf{x}+\mathbf{y}-\mathbf{z}\|^2 + \|\mathbf{x}-\mathbf{y}-\mathbf{z}\|^2 = 2\left(\|\mathbf{x}-\mathbf{z}\|^2 + \|\mathbf{y}\|^2\right)$$

Setting $f = \mathbf{x}$ and $g = \mathbf{y}+\mathbf{z}$, we have:

$$\|\mathbf{x}+\mathbf{y}+\mathbf{z}\|^2 + \|\mathbf{x}-\mathbf{y}-\mathbf{z}\|^2 = 2\left(\|\mathbf{x}\|^2 + \|\mathbf{y}+\mathbf{z}\|^2\right)$$

Setting $f = \mathbf{x}$ and $g = \mathbf{y}-\mathbf{z}$, we have:

$$\|\mathbf{x}+\mathbf{y}-\mathbf{z}\|^2 + \|\mathbf{x}-\mathbf{y}+\mathbf{z}\|^2 = 2\left(\|\mathbf{x}\|^2 + \|\mathbf{y}-\mathbf{z}\|^2\right)$$

By putting this together, we have:

$$\begin{aligned}\langle \mathbf{x}, \mathbf{z}\rangle + \langle \mathbf{y}, \mathbf{z}\rangle =&\frac{1}{4}\left(\|\mathbf{x}+\mathbf{z}\|^2 + \|\mathbf{y}\|^2 - \|\mathbf{x}-\mathbf{z}\|^2 - \|\mathbf{y}\|^2 + \|\mathbf{x}\|^2 + \|\mathbf{y}+\mathbf{z}\|^2 - \|\mathbf{x}\|^2 - \|\mathbf{y}-\mathbf{z}\|^2\right)\\ =&\frac{1}{8}\left(\|\mathbf{x}+\mathbf{y}+\mathbf{z}\|^2 + \|\mathbf{x}-\mathbf{y}+\mathbf{z}\|^2 - \|\mathbf{x}+\mathbf{y}-\mathbf{z}\|^2 - \|\mathbf{x}-\mathbf{y}-\mathbf{z}\|^2\right)\\ +&\frac{1}{8}\left(\|\mathbf{x}+\mathbf{y}+\mathbf{z}\|^2 + \|\mathbf{x}-\mathbf{y}-\mathbf{z}\|^2 - \|\mathbf{x}+\mathbf{y}-\mathbf{z}\|^2 - \|\mathbf{x}-\mathbf{y}+\mathbf{z}\|^2\right)\\ =&\frac{2}{8}\left(\|\mathbf{x}+\mathbf{y}+\mathbf{z}\|^2 - \|\mathbf{x}+\mathbf{y}-\mathbf{z}\|^2\right)\\ =&\frac{1}{4}\left(\|\mathbf{x}+\mathbf{y}+\mathbf{z}\|^2 - \|\mathbf{x}+\mathbf{y}-\mathbf{z}\|^2\right)\\ =&\langle \mathbf{x}+\mathbf{y}, \mathbf{z}\rangle\end{aligned}$$

3. $\langle \lambda\mathbf{x}, \mathbf{y}\rangle = \lambda\langle \mathbf{x}, \mathbf{y}\rangle$:

- When $\lambda = 0$, $\langle 0\cdot\mathbf{x}, \mathbf{y}\rangle = \langle \mathbf{0}, \mathbf{y}\rangle = \frac{\|\mathbf{0}+\mathbf{y}\|^2 - \|\mathbf{0}-\mathbf{y}\|^2}{4} = \frac{\|\mathbf{y}\|^2 - \|\mathbf{y}\|^2}{4} = 0 = 0\cdot\langle \mathbf{x}, \mathbf{y}\rangle$;
- When $\lambda = 1$, $\langle 1\cdot\mathbf{x}, \mathbf{y}\rangle = \langle \mathbf{x}, \mathbf{y}\rangle = 1\cdot\langle \mathbf{x}, \mathbf{y}\rangle$;
- When $\lambda > 1, \lambda \in \mathbb{N}$, if $\langle n\mathbf{x}, \mathbf{y}\rangle = n\langle \mathbf{x}, \mathbf{y}\rangle$, then

$$\begin{aligned}\langle (n+1)\mathbf{x}, \mathbf{y}\rangle &= \langle n\mathbf{x}+\mathbf{x}, \mathbf{y}\rangle\\ &= \langle n\mathbf{x}, \mathbf{y}\rangle + \langle \mathbf{x}, \mathbf{y}\rangle\\ &= n\langle \mathbf{x}, \mathbf{y}\rangle + \langle \mathbf{x}, \mathbf{y}\rangle && \triangleright \text{ Previous property}\\ &= (n+1)\langle \mathbf{x}, \mathbf{y}\rangle\end{aligned}$$

- For $\lambda \in \mathbb{Q}, \mathbb{R}$, refer to here.

4. $\langle \mathbf{x}, \mathbf{x}\rangle = 0$ if and only if $\mathbf{x} = \mathbf{0}$:

$$\begin{aligned}\langle \mathbf{x}, \mathbf{x}\rangle &= \frac{\|\mathbf{x}+\mathbf{x}\|^2 - \|\mathbf{x}-\mathbf{x}\|^2}{4}\\ &= \frac{\|2\mathbf{x}\|^2 - \|\mathbf{0}\|^2}{4} && \triangleright \text{ Norm's axiom 1}\\ &= \frac{4\|\mathbf{x}\|^2}{4} && \triangleright \text{ Norm's axiom 2}\\ &= \|\mathbf{x}\|^2\end{aligned}$$

Hence, $\langle \mathbf{x}, \mathbf{x} \rangle = \|\mathbf{x}\|^2 = 0$ if and only if $\mathbf{x} = \mathbf{0}$.

□

Theorem 2.6. *Inner-product-induced norm → parallelogram law:*

Let $\|\cdot\|$ be the inner product norm of inner product space $(V, \langle \cdot, \cdot \rangle), \mathbf{x}, \mathbf{y} \in V$.

Then

$$\|\mathbf{x}+\mathbf{y}\|^2 + \|\mathbf{x}-\mathbf{y}\|^2 = 2\left(\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2\right).$$

Proof.

$$\begin{aligned}
\langle \mathbf{x}+\mathbf{y}, \mathbf{x}+\mathbf{y} \rangle + \langle \mathbf{x}-\mathbf{y}, \mathbf{x}-\mathbf{y} \rangle &= \|\mathbf{x}+\mathbf{y}\|^2 + \|\mathbf{x}-\mathbf{y}\|^2 && \triangleright \text{ Definition of inner product induced norm} \\
\langle \mathbf{x}+\mathbf{y}, \mathbf{x}+\mathbf{y} \rangle + \langle \mathbf{x}-\mathbf{y}, \mathbf{x}-\mathbf{y} \rangle &= \langle \mathbf{x}, \mathbf{x} \rangle + \langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{y}, \mathbf{x} \rangle + \langle \mathbf{y}, \mathbf{y} \rangle \\
&\quad + \langle \mathbf{x}, \mathbf{x} \rangle - \langle \mathbf{x}, \mathbf{y} \rangle - \langle \mathbf{y}, \mathbf{x} \rangle + \langle \mathbf{y}, \mathbf{y} \rangle && \triangleright \text{ Linearity of inner product} \\
&= 2\langle \mathbf{x}, \mathbf{x} \rangle + 2\langle \mathbf{y}, \mathbf{y} \rangle \\
&= 2\left(\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2\right) && \triangleright \text{ Definition of inner product induced norm}
\end{aligned}$$

Hence, $\|\mathbf{x}+\mathbf{y}\|^2 + \|\mathbf{x}-\mathbf{y}\|^2 = 2\left(\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2\right)$. □

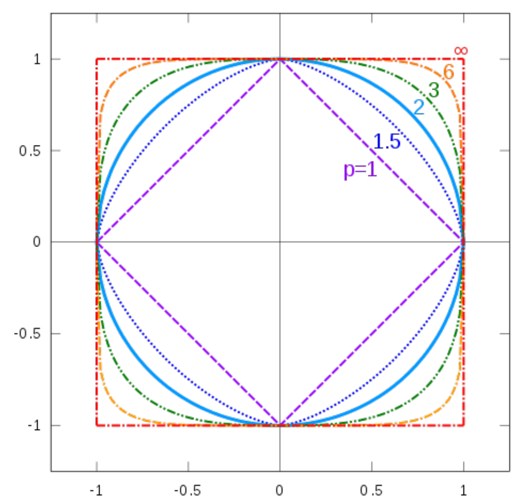


Figure 8: Visualization of $\|\mathbf{x}\|_p = 1$, namely the unit circle in different norms, which are the cross sections of Figure 2.

Definition 2.14. **Lebesgue integral** of a function f over a measure space X is written

$$\int_X f d\mu$$

to emphasize that the integral is taken with respect to the measure μ.

Remark 2.14. *In mathematics, the integral of a non-negative function of a single variable can be regarded, in the simplest case, as the area between the graph of that function and the x-axis. The Lebesgue integral extends the integral to a larger class of functions.*

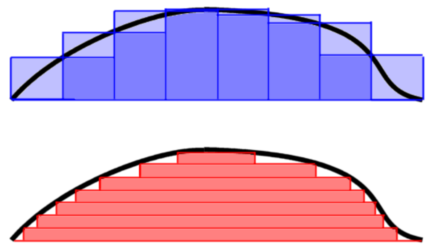


Figure 9: Riemann-Darboux's integration (in blue) and Lebesgue integration (in red).

Definition 2.15. $\boxed{\textbf{Frobenius norm}}$ is defined by

$$\|A\|_{\mathrm{F}} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2} = \sqrt{\operatorname{trace}(A^*A)} = \sqrt{\sum_{i=1}^{\min\{m,n\}}\sigma_i^2(A)}$$

Definition 2.16. $\boxed{\textbf{Gram–Schmidt process}}$ is a method for orthonormalizing a set of vectors $\{\mathbf{v}_1,\ldots,\mathbf{v}_n\}$ (vector space) in an inner product space. We define the projection[2] operator by

$$\operatorname{proj}_{\mathbf{u}}(\mathbf{v}) = \langle \mathbf{v},\mathbf{u}\rangle \cdot \frac{\mathbf{u}}{\langle \mathbf{u},\mathbf{u}\rangle},$$

The Gram–Schmidt process then works as follows:

$$\begin{aligned}
\mathbf{u}_1 &= \mathbf{v}_1, & \mathbf{e}_1 &= \frac{\mathbf{u}_1}{\|\mathbf{u}_1\|}\\
\mathbf{u}_2 &= \mathbf{v}_2 - \operatorname{proj}_{\mathbf{e}_1}(\mathbf{v}_2), & \mathbf{e}_2 &= \frac{\mathbf{u}_2}{\|\mathbf{u}_2\|}\\
\mathbf{u}_3 &= \mathbf{v}_3 - \operatorname{proj}_{\mathbf{e}_1}(\mathbf{v}_3) - \operatorname{proj}_{\mathbf{e}_2}(\mathbf{v}_3), & \mathbf{e}_3 &= \frac{\mathbf{u}_3}{\|\mathbf{u}_3\|}\\
\mathbf{u}_4 &= \mathbf{v}_4 - \operatorname{proj}_{\mathbf{e}_1}(\mathbf{v}_4) - \operatorname{proj}_{\mathbf{e}_2}(\mathbf{v}_4) - \operatorname{proj}_{\mathbf{e}_3}(\mathbf{v}_4), & \mathbf{e}_4 &= \frac{\mathbf{u}_4}{\|\mathbf{u}_4\|}\\
&\ \vdots & &\ \vdots\\
\mathbf{u}_k &= \mathbf{v}_k - \sum_{j=1}^{k-1}\operatorname{proj}_{\mathbf{e}_j}(\mathbf{v}_k), & \mathbf{e}_k &= \frac{\mathbf{u}_k}{\|\mathbf{u}_k\|}.
\end{aligned}$$

where $\{\mathbf{u}_1,\ldots,\mathbf{u}_n\}$ is the orthogonal basis of the vector space and $\{\mathbf{e}_1,\ldots,\mathbf{e}_n\}$ is the orthonormal basis of the vector space.

Definition 2.17. $\boxed{\textbf{Positive definite matrix}}$ A is an $n\times n$ symmetric matrix such that

1. $\langle \mathbf{x}^\top A\mathbf{x}\rangle \geq 0$ for all $\mathbf{x}\in\mathbb{R}$
2. $\langle \mathbf{x}^\top A\mathbf{x}\rangle = 0$ holds if and only if $\mathbf{x}=0$

Definition 2.18. $\boxed{\textbf{Positive definite function}}$ k is an $n\times n$ symmetric function such that

$$\int\int f(x)k(x,y)f(y)\,dxdy > 0$$

for all L^2 function f.

Remark 2.15. *The Gram-Schmidt matrix* $\mathbf{G}$ *of a set of vectors* $\{\mathbf{v}_1,\ldots,\mathbf{v}_n\}$ *is defined as below*

$$\begin{bmatrix}
\langle \mathbf{v}_1,\mathbf{v}_1\rangle & \langle \mathbf{v}_2,\mathbf{v}_1\rangle & \cdots & \langle \mathbf{v}_n,\mathbf{v}_1\rangle\\
\langle \mathbf{v}_1,\mathbf{v}_2\rangle & \langle \mathbf{v}_2,\mathbf{v}_2\rangle & \cdots & \langle \mathbf{v}_n,\mathbf{v}_2\rangle\\
\vdots & \vdots & \ddots & \vdots\\
\langle \mathbf{v}_1,\mathbf{v}_n\rangle & \langle \mathbf{v}_2,\mathbf{v}_n\rangle & \cdots & \langle \mathbf{v}_n,\mathbf{v}_n\rangle
\end{bmatrix}$$

[2]The definition of projection guarantees the $\mathbf{u}$ is orthogonal to previous $\mathbf{e}$.

It is a symmetric semi-positive-definite matrix the determinant of which indicates whether the set of vectors are linearly independent.

Proof. The proof of Gram-Schmidt matrix $\mathbf{G}$'s semi-positive-definitness reads as

$$\begin{aligned}\mathbf{G} &= \mathbf{V}^\top\mathbf{V}\\ \mathbf{u}^\top\mathbf{G}\mathbf{u} &= \mathbf{u}^\top\mathbf{V}^\top\mathbf{V}\mathbf{u}\\ \mathbf{u}^\top\mathbf{G}\mathbf{u} &= (\mathbf{V}\mathbf{u})^\top\mathbf{V}\mathbf{u} \geq 0\end{aligned}$$

□

Definition 2.19. **Sobolev space** $W^{k,p}(\mathbb{R})$ for $1 \leq p \leq \infty$ in one-dimensional case is defined as the subset of functions f in $L^p(\mathbb{R})$ such that f and its weak derivatives[3] up to order k have a finite L^p norm.

$$\|f\|_{k,p} = \left(\sum_{i=0}^{k}\int \left|f^{(i)}(t)\right|^p \, dt\right)^{\frac{1}{p}}.$$

Remark 2.16. *In the one-dimensional problem, it is enough to assume that the $(k-1)$-th derivative $f^{(k-1)}$ is differentiable almost everywhere and is equal almost everywhere to the Lebesgue integral of its derivative (this excludes irrelevant examples such as Cantor's function).*

Example 13. Sobolev spaces with $p = 2$ are especially important because of their connection with the Fourier series and because they form a Hilbert space. A special notation has arisen to cover this case since the space is a Hilbert space:

$$H^k = W^{k,2}.$$

Thereby, the frequently occurring H^1 denotes the Sobolev space is constituted by the functions f such that its first derivative have a finite L^2 norm.

Example 14. The space H^k can be defined naturally in terms of Fourier series whose coefficients decay sufficiently rapidly, namely,

$$H^k(\mathbb{T}) = \left\{ f \in L^2(\mathbb{T}) : \sum_{n=-\infty}^{\infty} \left(1 + n^2 + n^4 + \cdots + n^{2k}\right) \left|\widehat{f}(n)\right|^2 < \infty \right\}$$

where $\widehat{f}$ is the Fourier series of f, and $\mathbb{T}$ denotes the 1-torus. As above, one can use the equivalent norm

$$\|f\|_{k,2}^2 = \sum_{n=-\infty}^{\infty} \left(1 + |n|^2\right)^k \left|\widehat{f}(n)\right|^2.$$

Remark 2.17. *Overview of several spaces:*

Spaces	Elements	Operations	Equivalents
Vector	vectors	$\mathbf{x}+\mathbf{y}, \alpha\mathbf{x}$	
Pre-Hilbert	vectors	$\mathbf{x}+\mathbf{y}, \alpha\mathbf{x}, \langle\cdot,\cdot\rangle$	vector space + $\langle\cdot,\cdot\rangle$
Banach	vectors	$\mathbf{x}+\mathbf{y}, \alpha\mathbf{x}, \|\cdot\|$	vector space (complete) +$\|\cdot\|$
Lebesgue	functions (vectors) s.t. $\|f\|_p < \infty, p \in [1, \infty)$	$\mathbf{x}+\mathbf{y}, \alpha\mathbf{x}, \|\cdot\|$	Banach space
H^1 Sobolev	functions (vectors) s.t. f' has L^2 norm	$\mathbf{x}+\mathbf{y}, \alpha\mathbf{x}, \|\cdot\|$	Banach space
Hilbert	vectors	$\mathbf{x}+\mathbf{y}, \alpha\mathbf{x}$, $\|\cdot\|$ defined by $\langle\cdot,\cdot\rangle$	vector space (complete)+ $\langle\cdot,\cdot\rangle + \|\cdot\|$

[3]A weak derivative is a generalization of the concept of the derivative of a function (strong derivative) for functions not assumed differentiable, but only integrable to lie in the L^p space.

Definition 2.20. $\boxed{\textbf{Fourier transform}}$ of function $f(x)$ can be expressed as

$$F(\omega) = \int_{-\infty}^{\infty} f(x)\, e^{-i2\pi\omega x}\, dx.$$

And inverse of Fourier transform reads as

$$f(x) = \int_{-\infty}^{\infty} F(\omega)\, e^{i2\pi x\omega}\, d\omega.$$

Remark 2.18. *Let's consider a 2D spatial domain signal, denoted by the function $f(x, y)$, where x and y represent the spatial coordinates.*

The Fourier transform of the signal $f(x, y)$ is given by:

$$F(u, v) = \iint_{-\infty}^{\infty} f(x, y) \cdot \exp(-i2\pi(ux + vy))\, dx\, dy$$

where $F(u, v)$ is the complex-valued function in the Fourier domain and u and v represent the frequency coordinates.

Now, let's consider a rotation of the spatial domain signal $f(x, y)$ by an angle θ in the counterclockwise direction. The rotated signal, denoted by $g(x, y)$, is given by:

$$g(x, y) = f(x\cos\theta - y\sin\theta, x\sin\theta + y\cos\theta)$$

To understand the effect of this rotation in the Fourier domain, we need to compute the Fourier transform of the rotated signal $g(x, y)$.

Substituting the expression for $g(x, y)$ in the Fourier transform formula, we have:

$$G(u, v) = \iint_{-\infty}^{\infty} g(x, y) \cdot \exp(-i2\pi(ux + vy))\, dx\, dy$$

$$= \iint_{-\infty}^{\infty} f(x\cos\theta - y\sin\theta, x\sin\theta + y\cos\theta) \cdot \exp(-i2\pi(ux + vy))\, dx\, dy$$

Now, we can introduce a change of variables by substituting $x' = x\cos\theta - y\sin\theta$ and $y' = x\sin\theta + y\cos\theta$. This allows us to express the integral in terms of the new variables x' and y':

$$G(u, v) = \iint_{-\infty}^{\infty} f(x', y') \cdot \exp(-i2\pi(ux + vy))\, dx'\, dy'$$

$$= \iint_{-\infty}^{\infty} f(x', y') \cdot \exp(-i2\pi((u\cos\theta - v\sin\theta)x' + (u\sin\theta + v\cos\theta)y'))\, dx'\, dy'$$

Comparing the above equation with the Fourier transform formula, we can see that the expression inside the exponential function is of the form $-i2\pi(ux' + vy')$, which corresponds to the Fourier transform of the original signal $f(x', y')$.

Therefore, we can rewrite the equation as:

$$G(u, v) = F(u\cos\theta - v\sin\theta, u\sin\theta + v\cos\theta)$$

This result shows that rotating the signal $f(x, y)$ in the spatial domain corresponds to a phase shift in the Fourier domain, where the phase shift is determined by the rotation angle θ.

Hence, rotation in the spatial domain leads to the same rotation in the Fourier domain, preserving the rotational symmetry property of the Fourier transform.

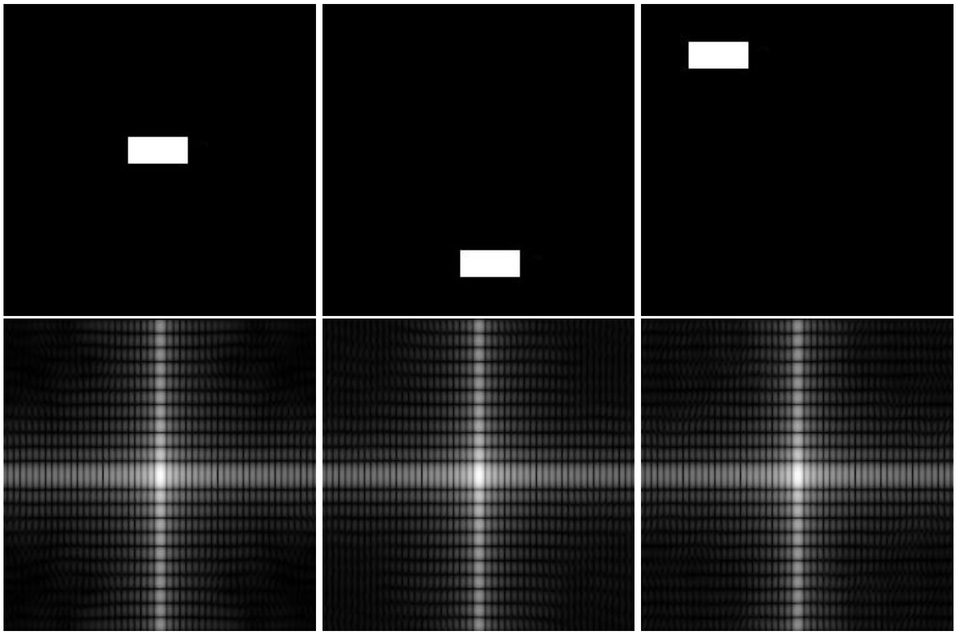


Figure 10: Top row from left to right: bar image with no translation, up translation, down translation rotation in spatial domain; bottom row from left to right: corresponding Fourier spectrum of the bar image in spatial domain.[Chaudhuri, 2004]

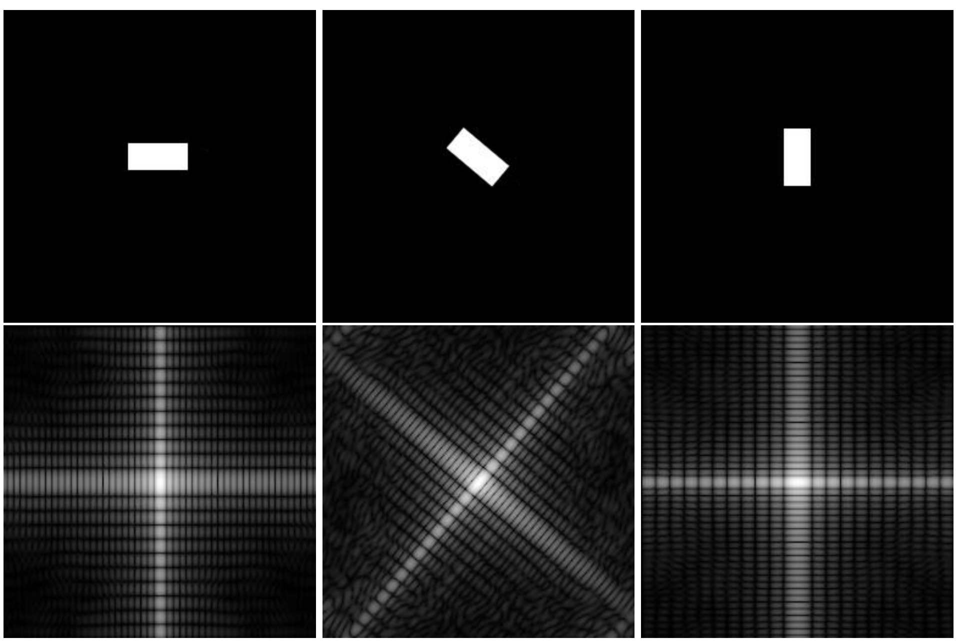


Figure 11: Top row from left to right: bar image with 0°, 40°, 90° rotation in spatial domain; bottom row from left to right: corresponding Fourier spectrum of the bar image with 0°, 40°, 90° rotation in spatial domain.

Definition 2.21. $\boxed{\textbf{Dirac delta function}}$ can be loosely thought of as a function on the real line which is zero everywhere except at the origin, where it is infinite,

$$\delta(x) \simeq \begin{cases} +\infty, & x = 0 \\ 0, & x \neq 0 \end{cases}$$

and which is also constrained to satisfy the identity

$$\int_{-\infty}^{\infty} \delta(x)\,\mathrm{d}x = 1.$$

For any function $f(x)$ that is continuous at $x = x_0$, the delta distribution is defined as

$$\int_{-\infty}^{\infty} f(x)\delta(x - x_0)\,dx = f(x_0).$$

Remark 2.19. *Fourier transform of Dirac delta function is*

$$\begin{aligned}\Delta(\omega) = \mathcal{F}(\delta(x - x_0)) &= \int_{-\infty}^{\infty} \delta(x - x_0)e^{-i2\pi\omega x}\,dx \\ &= e^{-i2\pi\omega x_0}\end{aligned}$$

When $x_0 = 0$,

$$\Delta(\omega) = \mathcal{F}(\delta(x - x_0)) = 1$$

Theorem 2.7. *The property of a **Green's function** can be exploited to solve differential equations of the form*

$$Lu(x) = f(x),$$

where L and $f(x)$ are given. If the kernel of L is non-trivial, then the Green's function is not unique.

A Green's function, $G(x, s)$ of a linear differential operator $L = L(x)$ at point s, is any solution of

$$LG(x, s) = \delta(x - s),$$

where δ is the Dirac delta function.

If such function G can be found for the operator L, then we obtain

$$\int LG(x, s)f(s)ds = \int \delta(x - s)f(s)ds = f(x)$$

Because the operator $L = L(x)$ is linear and acts only on the variable x, one may take the operator L outside of integration, yielding

$$L\left(\int G(x, s)f(s)ds\right) = f(x),$$

which means that

$$u(x) = \int G(x, s)f(s)ds$$

is the solution to $Lu(x) = f(x)$.

Definition 2.22. $\boxed{\textbf{Linear operators}}$ L are the operators such that for every pair of functions f and g and scalar t, it has

$$\begin{aligned}L(f + g) &= L(f) + L(g), \\ L(tf) &= tL(f).\end{aligned}$$

Definition 2.23. $\boxed{\textbf{Eigenfunctions } u \textbf{ of linear operators } D}$ are the functions such that

$$Du = \lambda u,$$

where λ is the eigenvalue and u is the corresponding eigenfunction.

Example 15. Below are a few examples of eigenfunctions of linear operator:

- Differentiation: $\frac{d}{dx}e^{\lambda x} = \lambda e^{\lambda x}$
- Gradient: $\nabla e^{\lambda \mathbf{x}} = \lambda e^{\lambda \mathbf{x}}$

- Laplacian:
  - $\nabla^2 \sin(ax+b) = \lambda \sin(x)$[4]
  - $\nabla^2 \sin(ax)\sin(by) = -(a^2+b^2)\sin(ax)\sin(by)$

Example 16. The following operators are all linear:

Operator	$L(f+g) = L(f) + L(g)$	$L(tf) = tL(f)$
Differential	$\frac{d(f+g)}{dx} = \frac{df}{dx} + \frac{dg}{dx}$	$\frac{d(tf)}{dx} = t\frac{df}{dx}$
Integral	$\int(f+g)\,dx = \int f\,dx + \int g\,dx$	$\int(tf)\,dx = t\int f\,dx$
Gradient	$\nabla(f+g) = \nabla f + \nabla g$	$\nabla(tf) = t\nabla f$
Fourier	$\mathcal{F}(f+g) = \mathcal{F}f + \mathcal{F}g$	$\mathcal{F}(tf) = t\mathcal{F}f$
Laplacian	$\Delta(f+g) = \Delta f + \Delta g$	$\Delta(tf) = t\Delta f$
Expectation	$E(f+g) = E(f) + E(g)$	$E(tf) = tE(f)$

Definition 2.24. **Boundary condition** is a mathematical condition that is imposed on the solution of the PDE at the boundary of the domain in which the PDE is defined. The boundary condition specifies how the solution behaves at the boundary of the domain and is necessary to obtain a unique solution of the PDE.

Example 17. Below are the most commonly used boundary conditions:

- Dirichlet boundary conditions on a domain $\Omega \subset \mathbb{R}^n$ take the form
$$y(x) = f(x) \quad \forall x \in \partial\Omega,$$
where f is a known function defined on the boundary $\partial\Omega$.
- Neumann boundary conditions on a domain $\Omega \subset \mathbb{R}^n$ take the form
$$\frac{\partial y}{\partial \mathbf{n}}(\mathbf{x}) = f(\mathbf{x}) \quad \forall \mathbf{x} \in \partial\Omega,$$
where $\mathbf{n}$ denotes the (typically exterior) normal to the boundary $\partial\Omega$, and f is a given scalar function.

Definition 2.25. **Least-norm problem** can be formulated as

$$\begin{aligned} &\arg\min \|\mathbf{x}\| \\ &\quad \text{s.t. } \langle \mathbf{x}, \mathbf{e}^{(i)} \rangle = c_i \end{aligned}$$

Example 18. Discrete least-norm problem:

$$\begin{aligned} &\arg\min \|\mathbf{x}\| \\ &\quad \text{s.t. } x_1 + x_2 = 1 \\ &\quad\quad x_3 = 1 \end{aligned}$$

1. Find the plane represented by each $\langle \mathbf{x}, \mathbf{e}^{(i)} \rangle = c_i$: When the dot product between a fixed vector $\mathbf{e}^{(i)}$ and a vector variable $\mathbf{x}$ is a constant c_i, it entails that the projection of $\mathbf{x}$ on $\mathbf{e}^{(i)}$ is a fixed value — namely $\mathbf{e}^{(i)}$ is perpendicular to the plane represented by $\mathbf{x}$, see Figure 12. Hence, the plane induced by $\langle \mathbf{x}, \mathbf{e}^{(i)} \rangle = c_i$ is perpendicular to $\mathbf{e}^{(i)}$, with the intercept indicated by c_i.
2. Find the subspace satisfies all $\langle \mathbf{x}, \mathbf{e}^{(i)} \rangle = c_i$: Once we have found out all the plane that is perpendicular to $\mathbf{e}^{(i)}$ with intercept c_i, we can have the subspace that satisfies all $\langle \mathbf{x}, \mathbf{e}^{(i)} \rangle = c_i$ by intersecting the planes.

[4] https://www.math.mcgill.ca/jakobson/papers/soup.pdf

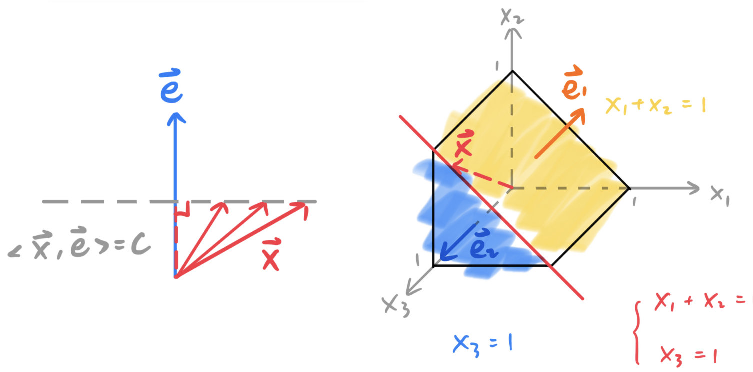


Figure 12: Left: plane represented by $\langle \mathbf{x}, \mathbf{e} \rangle = c$. Right: 3D case with multiple constraints.

3. Shortest $\|\mathbf{x}\|$ lives in the span of $\{\mathbf{e}^{(i)}\}$: As the span of $\mathbf{e}^{(\mathbf{i})}$ is orthogonal to the subspace, to determine the $\mathbf{x}$ that minimizes $\|\mathbf{x}\|$, we can express $\hat{\mathbf{x}}$ as a decomposition into $\hat{\mathbf{x}} = \sum \beta_i \mathbf{e}^{(i)}$. Consequently, $\hat{\mathbf{x}}$ resides within the span of $\mathbf{e}^{(\mathbf{i})}$.

4. Calculate the β coefficient by incorporating the constraints and $\hat{\mathbf{x}} = \sum \beta_i \mathbf{e}^{(i)}$:

$$\begin{aligned}
\begin{bmatrix} \mathbf{e}^{(1)} \\ \vdots \\ \mathbf{e}^{(n)} \end{bmatrix} \times \hat{\mathbf{x}} &= \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} \\
\begin{bmatrix} \mathbf{e}^{(1)} \\ \vdots \\ \mathbf{e}^{(n)} \end{bmatrix} \times \begin{bmatrix} \mathbf{e}^{(1)} & \cdots & \mathbf{e}^{(n)} \end{bmatrix} \times \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} &= \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} \\
\begin{bmatrix} \langle \mathbf{e}_1, \mathbf{e}_1 \rangle & \cdots & \langle \mathbf{e}_n, \mathbf{e}_1 \rangle \\ \vdots & \ddots & \vdots \\ \langle \mathbf{e}_1, \mathbf{e}_n \rangle & \cdots & \langle \mathbf{e}_n, \mathbf{e}_n \rangle \end{bmatrix} \times \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} &= \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} \\
\mathbf{G}\beta &= \mathbf{c} \\
\beta &= \mathbf{G}^{-1}\mathbf{c}
\end{aligned}$$

Example 19. Continuous least-norm problem:

$$\begin{aligned}
&\arg\min \|f\| \\
&\quad \text{s.t. } f(t_i) = c_i
\end{aligned}$$

If f lives in a reproducing kernel Hilbert space with kernel $k(\cdot,\cdot)$, then we have

$$\begin{aligned}
&\arg\min \|f\| \\
&\quad \text{s.t. } \langle k(\cdot,t_i), f(\cdot) \rangle_{\mathcal{H}} = f(t_i) = c_i
\end{aligned}$$

1. Find the plane represented by each $\langle k(\cdot,t_i), f(\cdot) \rangle_{\mathcal{H}} = c_i$

2. Find the subspace satisfies all $\langle k(\cdot,t_i), f(\cdot) \rangle_{\mathcal{H}} = c_i$

3. Shortest $\|f\|_{\mathcal{H}}$ lives in the span of $\{k(\cdot,t_i)\}$

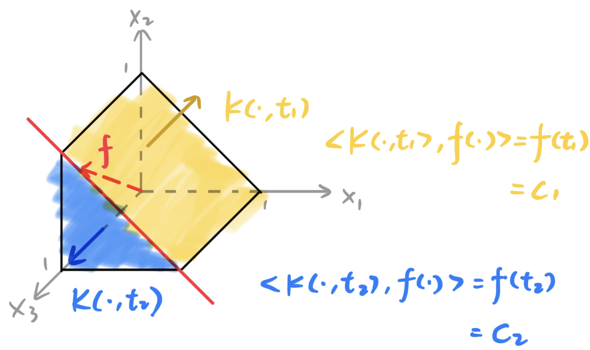


Figure 13: Right: Continuous case with multiple constraints.

4. Calculate the β coefficient by incorporating the constraints and $\hat{f} = \sum \beta_i k(\cdot, t_i)$:

$$
\begin{aligned}
\begin{bmatrix} k(\cdot,t_1) \\ \vdots \\ k(\cdot,t_n) \end{bmatrix} \times \hat{f} &= \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} \\
\begin{bmatrix} k(\cdot,t_1) \\ \vdots \\ k(\cdot,t_n) \end{bmatrix} \times \begin{bmatrix} k(\cdot,t_1) & \cdots & k(\cdot,t_n) \end{bmatrix} \times \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} &= \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} \\
\begin{bmatrix} \langle k(\cdot,t_1), k(\cdot,t_1)\rangle_{\mathcal{H}} & \cdots & \langle k(\cdot,t_1), k(\cdot,t_n)\rangle_{\mathcal{H}} \\ \vdots & \ddots & \vdots \\ \langle k(\cdot,t_n), k(\cdot,t_1)\rangle_{\mathcal{H}} & \cdots & \langle k(\cdot,t_n), k(\cdot,t_n)\rangle_{\mathcal{H}} \end{bmatrix} \times \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} &= \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} \\
\begin{bmatrix} k(t_1,t_1) & \cdots & k(t_1,t_n) \\ \vdots & \ddots & \vdots \\ k(t_n,t_1) & \cdots & k(t_n,t_n) \end{bmatrix} \times \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} &= \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} \\
\mathbf{K}\beta &= \mathbf{c} \\
\beta &= \mathbf{K}^{-1}\mathbf{c}
\end{aligned}
$$

Example 20. For landmark registration, we formulate the problem:

$$
\begin{aligned}
&\arg\min \|u_i\|_{\mathcal{H}} \\
&\text{s.t. } u_i\left(\mathbf{x}^{(j)}\right) = \mathbf{c}_i^{(j)}
\end{aligned}
$$

where the deformation field $\phi(\cdot): \mathbb{R}^n \to \mathbb{R}^n$ read as

$$
\phi\left(\mathbf{x}^{(j)}\right) = \mathbf{x}^{(j)} + u\left(\mathbf{x}^{(j)}\right) = \mathbf{x}^{(j)} + \mathbf{c}^{(j)}
$$

and u_i is the i-th component of the output of the displacement function. Here, we solve the deformation field in different components individually. For 3D registration, we are going to perform the process below three times.

We assume u_i lives in a reproducing kernel Hilbert space with kernel $k(\cdot,\cdot)$, then we have

$$
\begin{aligned}
&\arg\min \|u_i\| \\
&\text{s.t. } \langle k(\cdot, \mathbf{x}^{(j)}), u_i(\cdot)\rangle_{\mathcal{H}} = u_i(\mathbf{x}^{(j)}) = \mathbf{c}_i^{(j)}
\end{aligned}
$$

1. Find the plane represented by each $\langle k(\cdot, \mathbf{x}^{(j)}), u_i(\cdot)\rangle_{\mathcal{H}} = \mathbf{c}_i^{(j)}$

2. Find the subspace satisfies all $\langle k(\cdot,\mathbf{x}^{(j)}), u_i(\cdot)\rangle_{\mathcal{H}} = \mathbf{c}_i^{(j)}$

3. Shortest $\|u_i\|_{\mathcal{H}}$ lives in the span of $\{k(\cdot,\mathbf{x}^{(j)})\}$

4. Calculate the β coefficient by incorporating the constraints and $\hat{u}_i = \sum \beta_i k(\cdot,\mathbf{x}^{(j)})$:

$$
\begin{aligned}
\begin{bmatrix} k(\cdot,\mathbf{x}^{(1)}) \\ \vdots \\ k(\cdot,\mathbf{x}^{(n)}) \end{bmatrix} \times \hat{u}_i &= \begin{bmatrix} \mathbf{c}^{(1)} \\ \vdots \\ \mathbf{c}^{(n)} \end{bmatrix} \\
\begin{bmatrix} k(\cdot,\mathbf{x}^{(1)}) \\ \vdots \\ k(\cdot,\mathbf{x}^{(n)}) \end{bmatrix} \times \begin{bmatrix} k(\cdot,\mathbf{x}^{(1)}) & \cdots & k(\cdot,\mathbf{x}^{(n)}) \end{bmatrix} \times \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} &= \begin{bmatrix} \mathbf{c}^{(1)} \\ \vdots \\ \mathbf{c}^{(n)} \end{bmatrix} \\
\begin{bmatrix} \langle k(\cdot,\mathbf{x}^{(1)}), k(\cdot,\mathbf{x}^{(1)})\rangle_{\mathcal{H}} & \cdots & \langle k(\cdot,\mathbf{x}^{(1)}), k(\cdot,\mathbf{x}^{(n)})\rangle_{\mathcal{H}} \\ \vdots & \ddots & \vdots \\ \langle k(\cdot,\mathbf{x}^{(n)}), k(\cdot,\mathbf{x}^{(1)})\rangle_{\mathcal{H}} & \cdots & \langle k(\cdot,\mathbf{x}^{(n)}), k(\cdot,\mathbf{x}^{(n)})\rangle_{\mathcal{H}} \end{bmatrix} \times \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} &= \begin{bmatrix} \mathbf{c}^{(1)} \\ \vdots \\ \mathbf{c}^{(n)} \end{bmatrix} \\
\begin{bmatrix} k(\mathbf{x}^{(1)},\mathbf{x}^{(1)}) & \cdots & k(\mathbf{x}^{(1)},\mathbf{x}^{(n)}) \\ \vdots & \ddots & \vdots \\ k(\mathbf{x}^{(n)},\mathbf{x}^{(1)}) & \cdots & k(\mathbf{x}^{(n)},\mathbf{x}^{(n)}) \end{bmatrix} \times \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} &= \begin{bmatrix} \mathbf{c}^{(1)} \\ \vdots \\ \mathbf{c}^{(n)} \end{bmatrix} \\
\mathbf{K}\beta &= \mathbf{c}_i \\
\beta &= \mathbf{K}^{-1}\mathbf{c}_i
\end{aligned}
$$

Definition 2.26. [ucl, 2019] $\boxed{\textbf{Kernel}}$ $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a function, where $\mathcal{H}$ is a Hilbert space and $\mathcal{X}$ is a non-empty set, if there exists a function $\phi : \mathcal{X} \to \mathcal{H}$ such that for any $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, we have

$$
\begin{aligned}
&k(\mathbf{x},\mathbf{x}') : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \\
&k(\mathbf{x},\mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}')\rangle_{\mathcal{H}}
\end{aligned}
$$

Remark 2.20. *The kernel can be regarded as a distance function, which tells you the similarity of the two samples. In the Gaussian process, the covariance function is exactly a kernel.*

Remark 2.21. *We imposed almost no conditions on $\mathcal{X}$: we don't even require there to be an inner product defined on the elements of $\mathcal{X}$. Defining the inner product on $\mathcal{H}$ is enough. For example, let $\mathbf{x}, \mathbf{x}'$ represent two different books, we can't take an inner product between books, but we can take an inner product between the feature maps $\phi(\mathbf{x}), \phi(\mathbf{x}')$ corresponding to $\mathbf{x}, \mathbf{x}'$.*

Remark 2.22. *The kernel gives a way to compute inner products in some feature space without even knowing what this space is and what is ϕ. In most cases, we care more about the inner product than the feature mapping itself. We never need the coordinates of the data in the feature space. One example is the Gaussian kernel $k(\mathbf{x},\mathbf{y}) = \exp(-\gamma\|\mathbf{x}-\mathbf{y}\|^2)$. If we Taylor-expand this function, we'll see that it corresponds to an infinite-dimensional codomain of ϕ.*

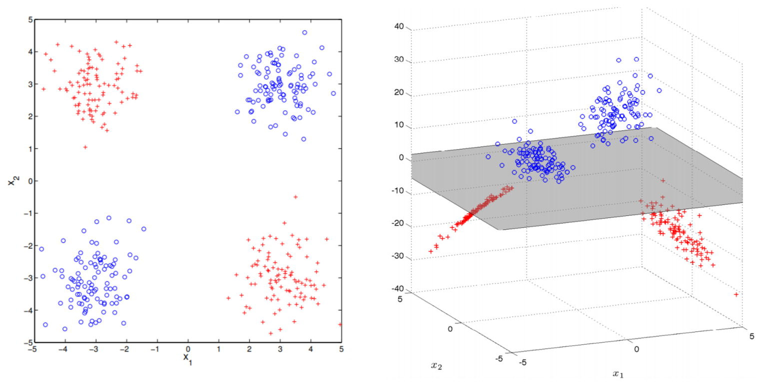


Figure 14: $\phi(\mathbf{x}) = [\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_1\mathbf{x}_2]^\top$ example of the kernel: on the left, the points are plotted in the original space; on the right, the points are plotted into a higher dimensional feature space by ϕ.

Definition 2.27. $\boxed{\textbf{Reproducing kernel Hilbert Space } \mathcal{H}}$

- is a Hilbert space, i.e., a vector space equipped with inner product $\langle \cdot, \cdot \rangle$ and norm $\| \cdot \|$;
- There exists an operator $\delta_{\mathbf{x}} : f \to f(\mathbf{x})$, for any $\mathbf{x} \in X$ (typically X will be $\mathbb{R}^n$), $f \in \mathcal{H}$, $\delta_{\mathbf{x}}$ is bounded, i.e., there exists δ_x, M such that $\|\delta_{\mathbf{x}} f\| \le M \|f\|_{\mathcal{H}}$;
- For any $\mathbf{x} \in X, f \in \mathcal{H}$, there exists a unique function (vector) $k_{\mathbf{x}} = k(\cdot, \mathbf{x}) \in \mathcal{H}$, s.t. $f(\mathbf{x}) = \delta_{\mathbf{x}}(f) = \langle f, K_{\mathbf{x}} \rangle_{\mathcal{H}}$, namely the reproducing ability, which is guaranteed by Riesz representation theorem.

Definition 2.28. $\boxed{\textbf{Reproducing kernel}}$ $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a function, where $\mathcal{H}$ is a Hilbert space and $\mathcal{X}$ is a non-empty set, if k satisfies

1. $\forall \mathbf{x} \in \mathcal{X}, k(\cdot, \mathbf{x}) \in \mathcal{H}$ ▷ Feature map of every point is in the feature space
2. $\forall \mathbf{x} \in \mathcal{X}, \forall f \in \mathcal{H}, f(\mathbf{x}) = \langle f, k(\cdot, \mathbf{x}) \rangle_{\mathcal{H}}$ ▷ Reproducing property
3. $k(\mathbf{x}, \mathbf{y}) = \langle k(\cdot, \mathbf{x}), k(\cdot, \mathbf{y}) \rangle_{\mathcal{H}} = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle_{\mathcal{H}}$

Remark 2.23. *From a discrete perspective, $k(\cdot, \cdot)$ can be regarded as a "matrix"; $k(\cdot, \mathbf{x}^{(i)})$ can be viewed as a "vector" designated at $\mathbf{x}^{(i)}$ column; and $k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$ is a scalar designated at $\mathbf{x}^{(i)}$ "row" and $\mathbf{x}^{(j)}$ "column".*

Remark 2.24. *The feature map is not unique, only the kernel is.* ***RKHS functions*** *can be written as linear combination of feature maps $k(\cdot, \mathbf{x})$, which we can regard as "basis function":*

$$
\begin{aligned}
f(\cdot) &= \sum_{i=1}^{m} \alpha_i \phi_i(\cdot) = \sum_{i=1}^{m} \alpha_i k(\cdot, \mathbf{x}^{(i)}) \\
f(\mathbf{x}) &= \langle f(\cdot), k(\cdot, \mathbf{x}) \rangle_{\mathcal{H}} \\
&= \left\langle \sum_{i=1}^{m} \alpha_i k(\cdot, \mathbf{x}^{(i)}), k(\cdot, \mathbf{x}) \right\rangle_{\mathcal{H}} \\
&= \sum_{i=1}^{m} \alpha_i \left\langle k(\cdot, \mathbf{x}^{(i)}), k(\cdot, \mathbf{x}) \right\rangle_{\mathcal{H}} \\
&= \sum_{i=1}^{m} \alpha_i k(\mathbf{x}, \mathbf{x}^{(i)})
\end{aligned}
$$

For shorter notation:

$$
\begin{aligned}
f &= \sum_{i=1}^{m} \alpha_i \phi_i = \sum_{i=1}^{m} \alpha_i k(\cdot, \mathbf{x}^{(i)}) \\
f(\mathbf{x}) &= \langle f, k(\cdot, \mathbf{x}) \rangle_{\mathcal{H}} \\
&= \left\langle \sum_{i=1}^{m} \alpha_i k(\cdot, \mathbf{x}^{(i)}), k(\cdot, \mathbf{x}) \right\rangle_{\mathcal{H}} \\
&= \sum_{i=1}^{m} \alpha_i \left\langle k(\cdot, \mathbf{x}^{(i)}), k(\cdot, \mathbf{x}) \right\rangle_{\mathcal{H}} \\
&= \sum_{i=1}^{m} \alpha_i k(\mathbf{x}, \mathbf{x}^{(i)})
\end{aligned}
$$

Remark 2.25. *When comparing the expression of Green's function and RKHS, we found that Green's function is the kernel of the inverse of the operator, where $g(s)$ corresponds to α_i and $G(\mathbf{x}, s)$ corresponds to RKHS $k(\mathbf{x}, \mathbf{x}_i)$.*

$$
\begin{aligned}
Lu(x) &= g(x) \\
u(x) &= \int g(s) G(x, s) ds
\end{aligned}
$$

$$
g(x) = \sum_{i=1}^{m} \alpha_i k(x, \mathbf{x}_i)
$$

Example 21. We define a feature map $\phi : \mathbb{R}^2 \to \mathbb{R}^3$

$$
\phi(\mathbf{x}) = [\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_1 \mathbf{x}_2]^\top
$$

For the reproducing property, we define an RKHS function $f : \mathbb{R}^2 \to \mathbb{R}$

$$
\begin{aligned}
f(\mathbf{x}) &= \sum_{l=1}^{\infty} f_l \phi_l(\mathbf{x}) && \triangleright \text{ Remark 3} \\
&= a\mathbf{x}_1 + b\mathbf{x}_2 + c\mathbf{x}_1\mathbf{x}_2 \\
f(\cdot) &= [a, b, c]^\top
\end{aligned}
$$

where $f(\cdot)$ or f stands for a function while $f(\mathbf{x})$ means the value of function f at $\mathbf{x}$. With this, we can write

$$
\begin{aligned}
f(\mathbf{x}) &= f(\cdot)^\top \phi(\mathbf{x}) \\
&= \langle f(\cdot), \phi(\mathbf{x}) \rangle_{\mathcal{H}}
\end{aligned}
$$

The reproducing property tells us that the evaluation of f at $\mathbf{x}$ can be written as an inner product in feature space.

Example 22. The kernel $k(\mathbf{x}, \mathbf{y}) = \frac{1}{2\sigma} e^{-\sigma|\mathbf{x}-\mathbf{y}|}$ associates with the inner product $\langle Lk(\mathbf{x}, \mathbf{y}), g \rangle$, where $L = \left(-\frac{\partial^2}{\partial x^2} + \alpha\right)$?

Proof.

$$
\begin{aligned}
Lk(\mathbf{x},0) &= \delta(\mathbf{x}) \\
\mathcal{F}\left(Lk(\mathbf{x})\right) &= \mathcal{F}\left(\delta(\mathbf{x})\right) = 1 \\
\mathcal{F}\left(\left(-\frac{\partial^2}{\partial x^2}+\alpha\right)k(\mathbf{x})\right) &= \int_{-\infty}^{+\infty}\left(-\frac{\partial^2}{\partial x^2}+\alpha\right)k(\mathbf{x})e^{-j\omega x}dx \\
&= -\int_{-\infty}^{+\infty}\frac{\partial^2}{\partial x^2}k(\mathbf{x})e^{-j\omega x}\,dx+\alpha\int_{-\infty}^{+\infty}k(\mathbf{x})e^{-j\omega x}\,dx \\
&= -\int_{-\infty}^{+\infty}k(\mathbf{x})\frac{\partial^2}{\partial x^2}e^{-j\omega x}\,dx+\alpha K(\omega) \\
&= -\int_{-\infty}^{+\infty}k(\mathbf{x})(-j\omega)^2e^{-j\omega x}\,dx+\alpha K(\omega) && \triangleright\ \text{Integration by part twice} \\
&= \omega^2\int_{-\infty}^{+\infty}k(\mathbf{x})e^{-j\omega x}\,dx+\alpha K(\omega) \\
&= \omega^2K(\omega)+\alpha K(\omega) \\
&= (\omega^2+\alpha)K(\omega) = 1
\end{aligned}
$$

$$
\begin{aligned}
K(\omega) &= \frac{1}{\omega^2+\alpha} \\
\mathcal{F}^{-1}(K(\omega)) = k(\mathbf{x}) &= \frac{1}{2\sqrt{\alpha}}e^{-\sqrt{\alpha}|\mathbf{x}|} \\
k(\mathbf{x},\mathbf{y}) &= \frac{1}{2\sigma}e^{-\sigma|\mathbf{x}-\mathbf{y}|}
\end{aligned}
$$

□

Definition 2.29. $\boxed{\textbf{Primal-Dual method}}$

Assume the primal problem as below:

$$
\begin{aligned}
&\text{maximize } z(x) \\
&\text{subject to } G(x)\leq\theta && x\in\Omega
\end{aligned}
$$

which is equivalent to

$$
\begin{aligned}
&\text{minimize } w(y)=\sup_{x\in\Omega}\{z(x)+\langle G(x),y\rangle\} \\
&\text{subject to } y\geq\theta && x\in\Omega
\end{aligned}
$$

More specifically,

Primal

$$
\begin{aligned}
\text{maximize } z &= \sum_{j=1}^{n}c_jx_j \\
\text{subject to } \sum_{j=1}^{n}a_{ij}x_j &\leq b_i && (i=1,2,\cdots,m) \\
x_j &\geq 0 && (j=1,2,\cdots,n)
\end{aligned}
$$

Dual

$$\begin{aligned}
\text{minimize } w &= \sum_{i=1}^{m} b_i y_i \\
\text{subject to } \sum_{i=1}^{m} a_{ij} y_i &\geq c_j \qquad (j = 1, 2, \cdots, n) \\
y_i &\geq 0 \qquad (i = 1, 2, \cdots, m)
\end{aligned}$$

Remark 2.26. *The difference between supremum (resp. infimum) and maximum (resp. minimum) is that for bounded, infinite sets, the maximum (resp. minimum) may not exist, but the supremum (resp. infimum) always does.*

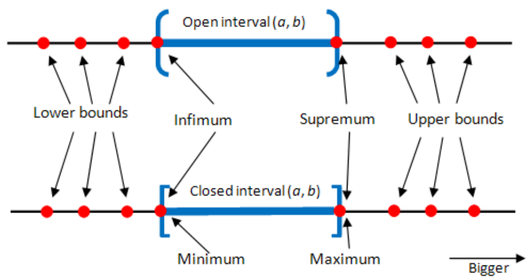


Figure 15: Supremum and Infimum

Example 23.

Primal

$$\begin{array}{lrcrl}
\max & z = 30x_1 & + & 100x_2 & \\
\text{s.t.} & x_1 & + & x_2 & \leq 7 \\
 & 4x_1 & + & 10x_2 & \leq 40 \\
 & x_1 & & & \geq 3 \\
 & x_1 & & & \geq 0 \\
 & & & x_2 & \geq 0
\end{array}$$

Multiply constraints i by a factor y_i. Choose the sign of y_i such that all inequalities are $\leq$ after multiplication:

$$\begin{array}{lrcrll}
\max & z = 30x_1 & + & 100x_2 & & \\
\text{s.t.} & x_1 & + & x_2 & \leq 7 & \times y_1 \\
 & 4x_1 & + & 10x_2 & \leq 40 & \times y_2 \\
 & x_1 & & & \geq 3 & \times(-y_3) \\
 & x_1 & & & \geq 0 & \times(-y_4) \\
 & & & x_2 & \geq 0 & \times(-y_5)
\end{array}$$

Add up all the obtained inequalities into a resultant inequality:

$$(y_1 + 4y_2 - y_3 - y_4)x_1 + (y_1 + 10y_2 - y_5)x_2 \leq 7y_1 + 40y_2 - 3y_3$$

Make the coefficients of the resultant constraint match the objective function. Then, the right hand side of the resultant constraints is an upper bound of z^*:

Dual

$$\begin{array}{lrcrcrcrcrcr}
\min & w = 7y_1 & + & 40y_2 & - & 3y_3 & & & & & & \\
\text{s.t.} & y_1 & + & 4y_2 & - & y_3 & - & y_4 & & & = & 30 \\
 & y_1 & + & 10y_2 & & & & & - & y_5 & = & 100
\end{array}$$

Remark 2.27. *Finding the max problem is equivalent to finding the min of the upper bound, that's why in the above example we should have the right sign of factor to make all inequalities are $\leq$ after multiplication. Likewise, finding the min problem is equivalent to find the max of the lower bound.*

Definition 2.30. $\boxed{\textbf{Thin plate splines}}$ are a spline-based technique for data interpolation and smoothing, which has the natural representation in terms of radial basis functions

$$f(x) = \sum_{i=1}^{K} w_i \varphi(\|x - c_i\|),$$

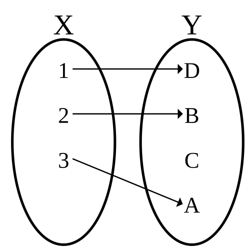
where w_i is a set of mapping coefficient, c_i is a set of control points and corresponding φ for TPS is $\varphi(r) = r^2 \log(r)$.

For 2D case, the energy function is defined as below

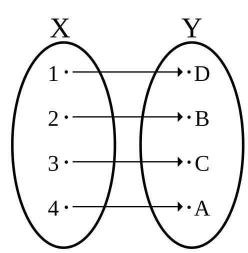
$$E_{\text{TPS,smooth}}(f) = \sum_{i=1}^{K} \|y_i - f(x_i)\|^2 + \lambda \int\int \left[\left(\frac{\partial^2 f}{\partial x_1^2}\right)^2 + 2\left(\frac{\partial^2 f}{\partial x_1 \partial x_2}\right)^2 + \left(\frac{\partial^2 f}{\partial x_2^2}\right) \right] dx_1 dx_2$$

where the tuning parameter λ is to control the rigidity of the deformation, balancing the aforementioned criterion with the measure of goodness of fit. If the interpolant pass through the data points exactly, then the first term of the energy function below should be zero. For this variational problem, it can be shown that there exists a unique minimizer f.

Definition 2.31. $\boxed{\textbf{Injection}}$ (injective function) is a function f that maps distinct elements of its domain to distinct elements, i.e. $f(x_1) = f(x_2)$ implies $x_1 = x_2$.



Definition 2.32. $\boxed{\textbf{Bijection}}$ (bijective/invertible function) is a function between the elements of two sets, where each element of one set is paired with exactly one element of the other set, and each element of the other set is paired with exactly one element of the first set. There are no unpaired elements.



Remark 2.28. *Bijective functions are essential to many areas of mathematics including the definitions of isomorphism, homeomorphism, diffeomorphism.*

Definition 2.33. If $x \in X$, then the $\boxed{\textbf{image}}$ of x under f, is denoted as $f(x)$.

3 Differential Geometry

Overview. Riemannian manifolds is a space that locally resembles[5] Euclidean space and is equipped with a Riemannian metric, which defines the inner product at each point on the tangent space and is in the form of a metric tensor. Tangent space is a vector(Euclidean) space that associates with each point on the manifold. The distance between two points on a Riemannian manifold is called geodesic, which is also called the shortest path. This distance makes a manifold a metric space.

Definition 3.1. $\boxed{\textbf{Euclidean space}}$ $\mathbb{R}^n$ is a space with metric tensor $\mathbf{I}$ everywhere.

Definition 3.2. $\boxed{\textbf{Isomorphism}}$ [Fletcher et al., 2004a, Hao, 2014, Zhang, 2016] is a function between two structures of the same type that can be reversed by an inverse function, i.e. bijective.

Example 24. In various areas of mathematics, isomorphisms have received specialized names, depending on the type of structure under consideration.

- An **isometry** (Def. 3.40) is an isomorphism of metric spaces, also a metric-preserving diffeomorphism.
- A **homeomorphism** (Def. 3.3) is an isomorphism of topological spaces.
- A **diffeomorphism** (Def. 3.9) is an isomorphism of spaces equipped with a differential structure, typically differentiable manifolds.

Remark 3.1. *Isometry, homeomorphism and diffeomorphism are all bijective, i.e. one-to-one.*

Definition 3.3. $\boxed{\textbf{Homeomorphism}}$ $f : M \to N$ is a bijective function between two topological spaces M, N, such that f and f^{-1} are both continuous function.

Definition 3.4. $\boxed{\textbf{Manifold}}$ is a Hausdorff space M with a countable basis such that for each point $p \in M$ there is a neighborhood U of p that is homeomorphic to $\mathbb{R}^n$ for some integer n. In other words, locally, a manifold is like a Euclidean space.

Definition 3.5. $\boxed{\textbf{Immersion}}$ is a smooth mapping $f : M \to N$ if for all $p \in M$, the differential $f_{*,p} : T_pM \to T_{f(p)}N$ is injective.

Definition 3.6. $\boxed{\textbf{Embedding}}$ is a smooth mapping $f : M \to N$ if

1. it is a one-to-one (bijective) immersion;
2. the image $f(M)$ with the subspace topology is homeomorphic (bijective) to M under f.

Definition 3.7. $\boxed{\textbf{Submanifold}}$ (or immersed submanifold) N of smooth manifold M together with an injective immersion $\iota : N \to M$. Identifying N with its image $\iota(N) \subset M$, we can consider N as a subset of M.

Definition 3.8. $\boxed{\textbf{Rank}}$ of a smooth map $f : M \to N$ at a point $p \in M$ is the rank of its **differential** (Jacobian) at p.

Remark 3.2. *Let m be the dimension of M and n be the dimension of N, in case $f : M \to N$ has maximal rank at p, there are three not mutually exclusive possibilities:*

1. *If $m = n$, then by the inverse function theorem, f is a local diffeomorphism at p;*
2. *If $m \leq n$, then the maximal rank is m and f is an immersion at p;*
3. *If $m \geq n$, then the maximal rank is n and f is a submersion at p.*

Definition 3.9. $\boxed{\textbf{Diffeomorphism}}$ $f : M \to N$ is a bijective function between two smooth manifolds M, N, such that f and f^{-1} (f is full ranked hence invertable) are both smooth functions.

[5]When we say that a Riemannian manifold "locally resembles Euclidean space," we mean that if you zoom in closely enough to a small region on the manifold, that region will look like a small piece of Euclidean space in terms of distances, angles, and curvature — the local region can be approximated by the Euclidean space, but not exactly the smae.

Remark 3.3. *Composing function with a diffeomorphism is a linear operator:*

$$(I_1 + I_2) \circ \phi = I_1 \circ \phi + I_2 \circ \phi (tI) \circ \phi = t(I \circ \phi)$$

Definition 3.10. For two manifolds M and N, a smooth mapping $f : M \to N$ induces a linear mapping of the tangent spaces $f_* : T_pM \to T_{f(p)}N$ called the $\boxed{\textbf{differential}}$ (Jacobian) of f.

Definition 3.11. $\boxed{\textbf{Metric-preserving mapping}}$ $f : M \to N$ is a smooth mapping for all $p \in M$ and tangent vectors $\mathbf{u}, \mathbf{v} \in T_pM$, we have

$$\langle \mathbf{u}, \mathbf{v} \rangle_p^M = \langle f_*\mathbf{u}, f_*\mathbf{v} \rangle_{f(p)}^N.$$

Definition 3.12. $\boxed{\textbf{Tangent space}}$ T_pM is a vector space attached to each point on a manifold M, which is isomorphic (not equivalent) to the Euclidean space. Intuitively, it is thought of as the linear space that best approximates M in a neighborhood of point p. Vectors in this space are called tangent vectors.

Remark 3.4. *Tangent space means for each and every point* p *in* $\mathbb{R}^n$*, we introduce a new coordinate system where all the vectors originated at* p *will reside.*

Example 25. The rotation group is presented as

$$\mathrm{SO}(3) = \{\mathbf{R} \in \mathbb{R}^{3\times 3} | \mathbf{R}^\top \mathbf{R} = I, |\mathbf{R}| = 1\}$$

In order to derive the form of elements in its Lie Algebra, $\mathfrak{so}(3)$, take a generic curve $\mathbf{R}(t)$ through the identity in $\mathrm{SO}(3)$ with derivative $X \in \mathfrak{so}(3)$ at $t = 0$ and consider the derivative of the constraint at $t = 0$. The product rule yields

$$\left.\frac{d}{dt}\right|_{t=0} \mathbf{R}(t)^\top \mathbf{R}(t) = \mathbf{X}^\top + \mathbf{X} = 0$$

This implies that any element of $\mathfrak{so}(3)$ is a **skew-symmetric matrix**.

Remark 3.5. *The* ***cross product*** *of vector* $\mathbf{a}$ *and* $\mathbf{b}$ *can be written as below*

$$\mathbf{a} \times \mathbf{b} = [\mathbf{a}]_\times \mathbf{b} = \begin{pmatrix} 0 & -\mathbf{a}_3 & \mathbf{a}_2 \\ \mathbf{a}_3 & 0 & -\mathbf{a}_1 \\ -\mathbf{a}_2 & \mathbf{a}_1 & 0 \end{pmatrix} \cdot \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \mathbf{b}_3 \end{pmatrix},$$

namely a skew-symmetric matrix times the vector $\mathbf{b}$*.*

Example 26. The group of symmetric positive definite matrices is presented as

$$\mathrm{SPD}(n) = \{\mathbf{X} \in \mathbb{R}^{n\times n} | \mathbf{X} = \mathbf{X}^\top, \mathbf{X} > 0\}.$$

The tangent space $T_A\mathrm{SPD}(n)$ at A, is the space of symmetric matrices [Yger et al., 2016].

Definition 3.13. $\boxed{\textbf{Tangent bundle}}$ TM consists of the tangent space T_pM at all points p in M.

$$TM = \{(p, \mathbf{v}) | p \in M, \mathbf{v} \in T_pM\}$$

Since a tangent space T_pM is the set of all tangent vectors to M at p, the tangent bundle is the collection of all tangent vectors, along with the information of the point to which they are tangent.

Definition 3.14. $\boxed{\textbf{Hessian matrix}}$ of a differentiable, multivariable function $f : \mathbb{R}^n \to \mathbb{R}$ at p is defined by

$$Hf_p = \begin{pmatrix} \partial_1^2 f_p & \partial_1\partial_2 f_p & \cdots & \partial_1\partial_n f_p \\ \partial_2\partial_1 f_p & \partial_2^2 f_p & \cdots & \partial_2\partial_n f_p \\ \vdots & \vdots & \ddots & \vdots \\ \partial_n\partial_1 f_p & \partial_n\partial_2 f_p & \cdots & \partial_n^2 f_p \end{pmatrix}.$$

Remark 3.6. *The Hessian matrix of a function* f *is the Jacobian matrix of the gradient of the function* f *; that is:* $H(f) = D(\nabla f)$*.*

Proof.

$$\nabla f(x^1, \cdots, x^n) = \begin{pmatrix} \partial_1 f(x^1, \cdots, x^n) \\ \vdots \\ \partial_m f(x^1, \cdots, x^n) \end{pmatrix}$$

$$D(\nabla f) = \begin{pmatrix} \partial_1\partial_1 f & \partial_2\partial_1 f & \cdots & \partial_n\partial_1 f \\ \partial_1\partial_2 f & \partial_2\partial_2 f & \cdots & \partial_n\partial_2 f \\ \vdots & \vdots & \ddots & \vdots \\ \partial_1\partial_n f & \partial_2\partial_n f & \cdots & \partial_n\partial_n f \end{pmatrix} = Hf$$

□

Remark 3.7. *If multivariable function $f : \mathbb{R}^n \to \mathbb{R}, \nabla f(x) = 0, Hf(x)$ is positive (resp. negative) definite, then x is the isolated local minimum (resp. maximum).*

Remark 3.8. *Relationship between convexity and positive-definiteness:*

$$f \text{ is convex} \Leftrightarrow \forall p, Hf_p \text{ is positive semi–definite}$$
$$f \text{ is strictly convex} \Leftrightarrow \forall p, Hf_p \text{ is positive definite}$$

Definition 3.15. **Laplacian** of a differentiable, multivariable function $f : \mathbb{R}^n \to \mathbb{R}$ at p is defined by

$$\Delta f_p = \mathrm{tr}(Hf_p) = \sum_{i=1}^{n} \partial_i^2 f_p.$$

Remark 3.9. *The eigenvector of a 2D Laplacian operator is an image s.t. the effect of applying the Laplacian operator to the image is equivalent to scaling the image by a scalar.*

Definition 3.16. **Jacobian matrix** (differential) of a differentiable, vector-valued function $f : \mathbb{R}^n \to \mathbb{R}^m$ at p is defined by [sta, 2013d]

$$Df_p = \begin{pmatrix} \partial_1 f_p^1 & \partial_2 f_p^1 & \cdots & \partial_n f_p^1 \\ \partial_1 f_p^2 & \partial_2 f_p^2 & \cdots & \partial_n f_p^2 \\ \vdots & \vdots & \ddots & \vdots \\ \partial_1 f_p^m & \partial_2 f_p^m & \cdots & \partial_n f_p^m \end{pmatrix} = \begin{pmatrix} (\nabla f_p^1)^\top \\ (\nabla f_p^2)^\top \\ \vdots \\ (\nabla f_p^m)^\top \end{pmatrix},$$

where

$$f(x^1, \cdots, x^n) = \begin{pmatrix} f^1(x^1, \cdots, x^n) \\ \vdots \\ f^m(x^1, \cdots, x^n) \end{pmatrix}.$$

Remark 3.10. *The Jacobian of ∇f, where $f : \mathbb{R}^n \to \mathbb{R}$, is the Laplacian of f.*

Remark 3.11. *A matrix can be thought of as a linear transformation. Hence, we can think of Df_p as a linear function $Df_p : T_p\mathbb{R}^n \to T_p\mathbb{R}^m$, which maps a vector in the tangent space at the source point p to a vector in the tangent space at the target point $f(p)$.*

In other words, the Jacobian matrix Df_p tells you how the change (v) in the domain p will be reflected in the range $f(p)$. More formally, we have

$$\frac{d}{dt} f(\gamma(t))|_{t=0} = Df_p \cdot v_p \text{ ,where } v \in T_p\mathbb{R}^n$$

Example 27. When $f : \mathbb{R} \to \mathbb{R}$, we have $\frac{d}{dp} f(p) = Df_p \cdot v_p = f'(p) \cdot v_p$, namely the first order Taylor expansions.

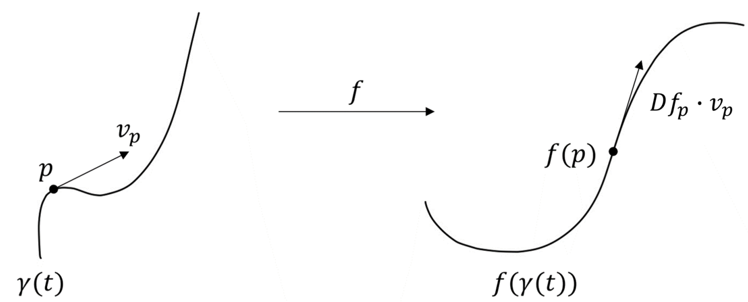


Figure 16: Jacobian

Definition 3.17. $\boxed{\textbf{Divergence}}$ of a differentiable, vector-valued function $f : \mathbb{R}^n \to \mathbb{R}^m$ at p is defined by

$$\nabla \cdot f_p = \operatorname{tr}(Df_p) = \sum_{i=1}^{n} \partial_i f_p^i.$$

Remark 3.12. *The* ***determinant*** *of Jacobian at a given point gives important information about the behavior of f near that point.*

- *If the Jacobian determinant at p is non-zero, then f is invertible near a point $p \in \mathbb{R}^n$.*
- *If the Jacobian determinant at p is positive (resp. negative), then f preserves (resp. reverses) orientation near p.*
- *The absolute value of the Jacobian determinant at p gives us the factor by which the function f expands or shrinks volumes near p.*

Example 28. For vector field $f : \mathbb{R}^2 \to \mathbb{R}^2$:

$$Df = \begin{pmatrix} \frac{\partial f^1}{\partial x^1} & \frac{\partial f^1}{\partial x^2} \\ \frac{\partial f^2}{\partial x^1} & \frac{\partial f^2}{\partial x^2} \end{pmatrix}$$

Example 29. Integrating $\int_b^a (2x^3+1)^7(x^2)\,dx$.

Solution. Making

$$y = \varphi(x) = 2x^3 + 1$$
$$\frac{dy}{dx} = \varphi'(x) = 6x^2$$

Therefore we have

$$\begin{aligned} \int_b^a (2x^3+1)^7(x^2)\,dx &= \frac{1}{6}\int_b^a f(\varphi(x))\varphi'(x)\,dx \\ &= \frac{1}{6}\int_{\varphi(b)}^{\varphi(a)} f(\varphi(x))d\varphi(x) \\ &= \frac{1}{6}\int_{\varphi(b)}^{\varphi(a)} f(y)dy \\ &= \frac{1}{6}\int_{\varphi(b)}^{\varphi(a)} y^7 dy \\ &= \frac{1}{48}[y^8]_{\varphi(b)}^{\varphi(a)} \end{aligned}$$

Definition 3.18. $\boxed{\textbf{Vector field}}$ $\mathfrak{X}(M)$ is a function on a manifold M that smoothly assigns to each point $p \in M$ a tangent vector $v \in T_pM$.

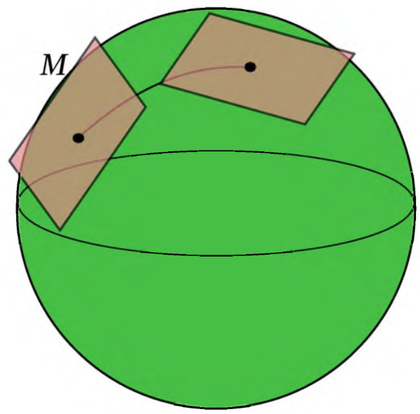


Figure 17: Tangent Space

Definition 3.19. $\boxed{\textbf{Mahalanobis distance}}$ is a measure of the distance between a point $\mathbf{s}^{(i)}$ and a distribution, can be written as

$$d_M(\mathbf{s}^i) = \sqrt{(\mathbf{s}^{(i)} - \mu)^\top \mathbf{\Sigma}^{-1}(\mathbf{s}^{(i)} - \mu)},$$

where μ is the mean vector and $\mathbf{\Sigma}$ is the covariance matrix of the distribution.

It can also be defined as a dissimilarity measure between two random vectors S^i and S^j of the same distribution with the covariance matrix $\mathbf{\Sigma}$:

$$d_M(\mathbf{s}^{(i)}, \mathbf{s}^{(j)}) = \sqrt{(\mathbf{s}^{(i)} - \mu)^\top \mathbf{\Sigma}^{-1}(\mathbf{s}^{(j)} - \mu)}.$$

Definition 3.20. $\boxed{\textbf{Metric}}$ is a mapping $d : X \times X \to \mathbb{R}$ over a vector space X if for all $\mathbf{x}^{(i)}, \mathbf{x}^{(j)}, \mathbf{x}^{(k)} \in X$, it satisfies the properties:

- Triangular inequality: $d(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) + d(\mathbf{x}^{(j)}, \mathbf{x}^{(k)}) \geq d(\mathbf{x}^{(i)}, \mathbf{x}^{(k)})$
- Non-negativity: $d(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) \geq 0$
- Symmetry: $d(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = d(\mathbf{x}^{(j)}, \mathbf{x}^{(i)})$
- Distinguishability: $d(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = 0 \Leftrightarrow \mathbf{x}^{(i)} = \mathbf{x}^{(j)}$

The ordered pair (X, d) is called a **metric space**. Strictly speaking, if a mapping satisfies the first three properties but not the fourth, it is called **pseudometric**.

Remark 3.13. *From a Mahalanobis distance perspective $d_M(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \sqrt{(\mathbf{x}^{(i)} - \mathbf{x}^{(j)})^\top M(\mathbf{x}^{(i)} - \mathbf{x}^{(j)})}$, if d_M is a pseudometric (namely $d(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = 0 \not\Leftrightarrow \mathbf{x}^{(i)} = \mathbf{x}^{(j)}$), $\mathbf{M}$ is not full-rank[Suárez et al., 2021], which arises to the problem of dimensionality reduction.*

Remark 3.14. ***Another interpretation of Mahalanobis distance** d_M. Due to the positive definiteness of covariance matrix $\mathbf{M} = \Sigma^{-1} = \mathbf{J}^\top\mathbf{J}$, we can have*

$$\begin{aligned}
d_M^2(\mathbf{s}^{(i)}, \mathbf{s}^{(j)}) &= (\mathbf{s}^{(i)} - \mathbf{s}^{(j)})^\top \mathbf{M}(\mathbf{s}^{(i)} - \mathbf{s}^{(j)}) \\
&= (\mathbf{s}^{(i)} - \mathbf{s}^{(j)})^\top \mathbf{J}^\top\mathbf{J}(\mathbf{s}^{(i)} - \mathbf{s}^{(j)}) \\
&= (\mathbf{J}(\mathbf{s}^{(i)} - \mathbf{s}^{(j)}))^\top \mathbf{J}(\mathbf{s}^{(i)} - \mathbf{s}^{(j)}) \\
&= \|\mathbf{J}(\mathbf{s}^{(i)} - \mathbf{s}^{(j)})\|_2^2
\end{aligned}$$

Learning a Mahalanobis distance is equivalent to learning a linear mapping $\mathbf{J}$ that transforms the data into a new space. And the corresponding distance is basically the Euclidean distance.

Remark 3.15. *Metric tensor $\mathbf{M}$ can be decomposed to $\mathbf{J}^\top\mathbf{J}$ due to its positive definiteness, where $\mathbf{J}$ is namely the Jacobian matrix representing the diffeomorphism maps the tangent space to the Euclidean space $\mathbb{R}^n$. $\mathbf{J}v$ is the image of v in the Euclidean space, hence the norm is calculated via Euclidean metric $\|v\|_g = \|\mathbf{J}v\| = (\mathbf{J}v)^\top\mathbf{I}(\mathbf{J}v)$.*

Definition 3.21. $\boxed{\textbf{Riemannian metric}}$ on a differential manifold M is a smooth function that assigns to each point p of M an inner product $\langle\cdot,\cdot\rangle$ on the tangent space T_pM.

Remark 3.16. *Assuming A is the metric at point $p \in M$, and v, λ are the unit eigenvector and its square root corresponding eigenvalue of A.*

$$
\begin{gathered}
Av = \lambda v \\
\|v\|^2 = v^\top Av = v^\top\lambda v = \lambda v^\top v = \lambda
\end{gathered}
$$

The deviation above tells you that the length of unit vector in the direction of eigenvector is scored as its corresponding eigenvalue. The unit vector points to other direction may be scored at different length.

Remark 3.17. ***Covariance matrix and Riemannian metric.[Arvanitidis et al., 2016]*** *A local covariance matrix can be used to represent the local structure of the data: the inverse of a local diagonal covariance matrix Σ^{-1} can be treated as the metric tensor, as the eigenvector corresponds to the smallest eigenvalue of the metric tensor is the direction of the geodesic. The geodesic is "pulled" towards the data where the metric is small.*

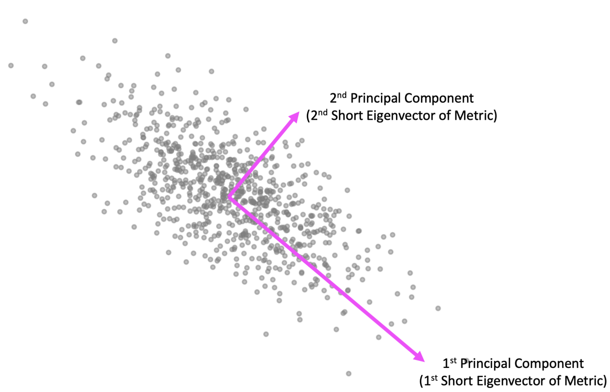


Figure 18: Correspondance between eigenvectors of coveriancce matrix and the Riemannian metric.

Definition 3.22. $\boxed{\textbf{Metric tensor}}$ [YouTube, 2018d] is a function that tells how to compute the distance between any two points in a given space.

Example 30. Typically, we calculate the arc length as below

$$
\text{arc length} = \int \|\dot{\gamma}(t)\| dt
$$

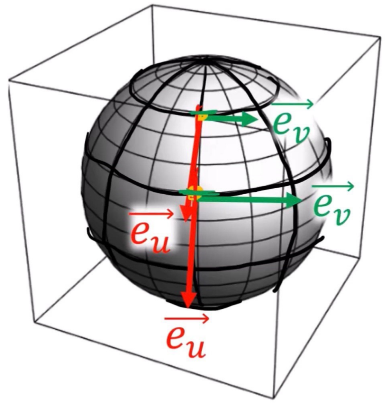


Figure 19: Base vectors on tangent space

By introducing the position vector, and expand it intrinsically, we can have

$$\begin{aligned}
\left\|\frac{d\vec{R}}{d\lambda}\right\|^2 &= \frac{d\vec{R}}{d\lambda}\cdot\frac{d\vec{R}}{d\lambda} \\
&= \left(\frac{d\vec{R}}{du}\cdot\frac{du}{d\lambda}+\frac{d\vec{R}}{dv}\cdot\frac{dv}{d\lambda}\right)\cdot\left(\frac{d\vec{R}}{du}\cdot\frac{du}{d\lambda}+\frac{d\vec{R}}{dv}\cdot\frac{dv}{d\lambda}\right) \\
&= \left(\frac{du}{d\lambda}\right)^2\left(\frac{d\vec{R}}{du}\cdot\frac{d\vec{R}}{du}\right)+\frac{du}{d\lambda}\cdot\frac{dv}{d\lambda}\left(\frac{d\vec{R}}{du}\cdot\frac{d\vec{R}}{dv}\right) \\
&+ \frac{dv}{d\lambda}\cdot\frac{du}{d\lambda}\left(\frac{d\vec{R}}{dv}\cdot\frac{d\vec{R}}{du}\right)+\left(\frac{dv}{d\lambda}\right)^2\left(\frac{d\vec{R}}{dv}\cdot\frac{d\vec{R}}{dv}\right) \\
&= \begin{pmatrix}\frac{du}{d\lambda} & \frac{dv}{d\lambda}\end{pmatrix}\underbrace{\begin{pmatrix}\frac{d\vec{R}}{du}\cdot\frac{d\vec{R}}{du} & \frac{d\vec{R}}{du}\cdot\frac{d\vec{R}}{dv} \\ \frac{d\vec{R}}{dv}\cdot\frac{d\vec{R}}{du} & \frac{d\vec{R}}{dv}\cdot\frac{d\vec{R}}{dv}\end{pmatrix}}_{\text{metric tensor}}\begin{pmatrix}\frac{du}{d\lambda} \\ \frac{dv}{d\lambda}\end{pmatrix} && \triangleright \vec{e}_u=\frac{d\vec{R}}{du}, \vec{e}_v=\frac{d\vec{R}}{dv} \\
&= \begin{pmatrix}\frac{du}{d\lambda} & \frac{dv}{d\lambda}\end{pmatrix}\underbrace{\begin{pmatrix}\vec{e}_u\cdot\vec{e}_u & \vec{e}_u\cdot\vec{e}_v \\ \vec{e}_v\cdot\vec{e}_u & \vec{e}_v\cdot\vec{e}_v\end{pmatrix}}_{\text{metric tensor}}\begin{pmatrix}\frac{du}{d\lambda} \\ \frac{dv}{d\lambda}\end{pmatrix}
\end{aligned}$$

Remark 3.18. *Actually, without seeing the manifold extrinsically, it's hard to derive the metric, as we don't know the position vector. Provided that the metric is already given, so can we calculate what we want intrinsically.*

The parametric equation of a sphere is shown as below:

$$\begin{aligned}
\vec{R} &= [X,Y,Z]^\top \\
\text{where } X &= \cos(v)\sin(u) = \cos(\lambda)\sin(\lambda) \\
Y &= \sin(v)\sin(u) = \sin(\lambda)\sin(\lambda) \\
Z &= \cos(u) = \cos(\lambda) \\
\text{when } u &= \lambda, v = \lambda.
\end{aligned}$$

After expanding the base vectors extrinsically, we can have the base vectors expressed as below:

$$\begin{aligned}
\vec{e}_u = \frac{d\vec{R}}{du} &= \frac{\partial \vec{R}}{\partial X}\frac{\partial X}{\partial u} + \frac{\partial \vec{R}}{\partial Y}\frac{\partial Y}{\partial u} + \frac{\partial \vec{R}}{\partial Z}\frac{\partial Z}{\partial u} \\
&= \cos(v)\cos(u)\frac{\partial \vec{R}}{\partial X} + \sin(v)\cos(u)\frac{\partial \vec{R}}{\partial Y} - \sin(u)\frac{\partial \vec{R}}{\partial Z} \\
&= \cos(v)\cos(u)\vec{e}_X + \sin(v)\cos(u)\vec{e}_Y - \sin(u)\vec{e}_Z \\
\vec{e}_v = \frac{d\vec{R}}{dv} &= \frac{\partial \vec{R}}{\partial X}\frac{\partial X}{\partial u} + \frac{\partial \vec{R}}{\partial Y}\frac{\partial Y}{\partial u} + \frac{\partial \vec{R}}{\partial Z}\frac{\partial Z}{\partial u} \\
&= -\sin(v)\sin(u)\frac{\partial \vec{R}}{\partial X} + \cos(v)\sin(u)\frac{\partial \vec{R}}{\partial Y} \\
&= -\sin(v)\sin(u)\vec{e}_X + \cos(v)\sin(u)\vec{e}_Y
\end{aligned}$$

Since $\vec{e}_X, \vec{e}_Y, \vec{e}_Z$ are perpendicular to each other, so the metric tensor is yielded as below:

$$\begin{pmatrix} \vec{e}_u \cdot \vec{e}_u & \vec{e}_u \cdot \vec{e}_v \\ \vec{e}_v \cdot \vec{e}_u & \vec{e}_v \cdot \vec{e}_v \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & \sin^2(u) \end{pmatrix}$$

Substituting the metric tensor back into the expression of norm of velocity, we get

$$\begin{aligned}
\left\| \frac{d\vec{R}}{d\lambda} \right\|^2 &= \begin{pmatrix} \frac{du}{d\lambda} & \frac{dv}{d\lambda} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \sin^2(u) \end{pmatrix} \begin{pmatrix} \frac{du}{d\lambda} \\ \frac{dv}{d\lambda} \end{pmatrix} \\
&= \left(\frac{du}{d\lambda}\right)^2 + \sin^2(u)\left(\frac{dv}{d\lambda}\right)^2
\end{aligned}$$

- For $u = \frac{\pi}{4}, v = \lambda$

$$\begin{aligned}
\left\| \frac{d\vec{R}}{d\lambda} \right\|^2 &= \left(\frac{du}{d\lambda}\right)^2 + \sin^2(u)\left(\frac{dv}{d\lambda}\right)^2 \\
&= 0^2 + \sin^2\left(\frac{\pi}{4}\right) \cdot 1^2 = \frac{1}{2}
\end{aligned}$$

The functional of arc length is

$$\text{arc length} = \int \left\| \frac{d\vec{R}}{d\lambda} \right\| dt = \int \frac{\sqrt{2}}{2} dt = \frac{\sqrt{2}}{2} t$$

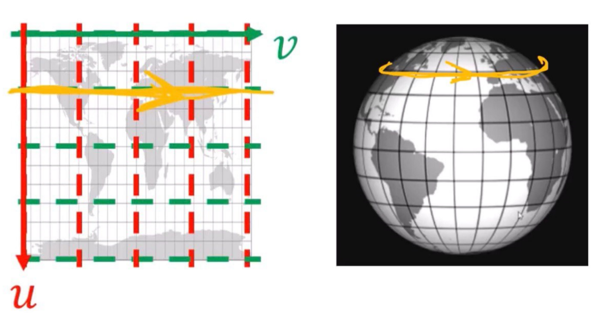


Figure 20: $u = \frac{\pi}{4}, v = \lambda$

- For $u = \frac{\pi}{2}, v = \lambda$

$$\left\|\frac{d\vec{R}}{d\lambda}\right\|^2 = \left(\frac{du}{d\lambda}\right)^2 + \sin^2(u)\left(\frac{dv}{d\lambda}\right)^2$$
$$= 0^2 + \sin^2\left(\frac{\pi}{2}\right) \cdot 1^2 = 1$$

The functional of arc length is

$$\text{arc length} = \int \left\|\frac{d\vec{R}}{d\lambda}\right\| dt = \int 1dt = t$$

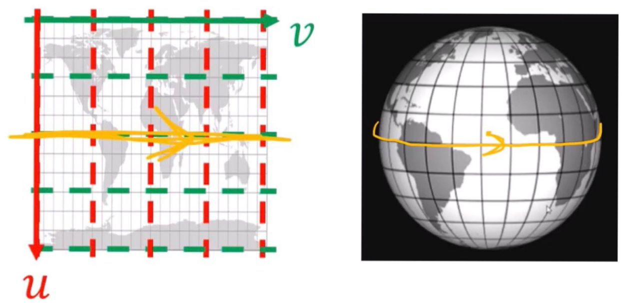


Figure 21: $u = \frac{\pi}{2}, v = \lambda$

Remark 3.19. *So the figure below is a good illustration of the role metric tensor plays:*

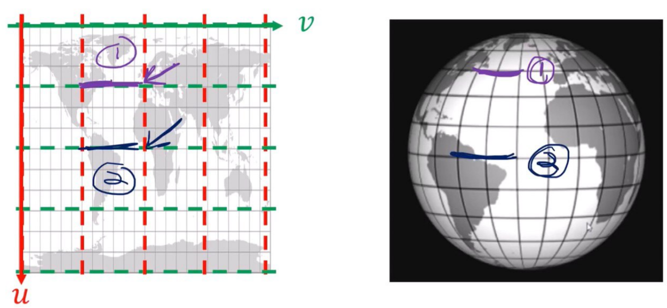


Figure 22: What we usually see in expression convenience vs. What actually it is

The sub-figure left is what we usually see in practice, which gives us the illusion that this is an Euclidean space, but a distorted one. However, the actual shape of the manifold is the sub-figure right, which can be more arbitrary than this sphere in most cases. So, if we want to measure the distance between two points, what we need is simply the metric tensor on each points. With the metric tensor, we can derive the inner product of the velocity vector, then integrate the norm of velocity by t*, we can have the distance we want.*

In other words, the metric tensor is the tool to describe the shape of a manifold.

Remark 3.20. *As figure 11 shows, longer axis stands for higher time cost, while shorter axis represents lower time cost, namely a shorter distance. And figure 12 illustrates the previous property well - the closer to the polars, the lower time cost would be.*

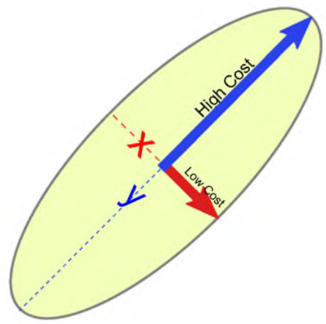


Figure 23: Visualization of metric tensor

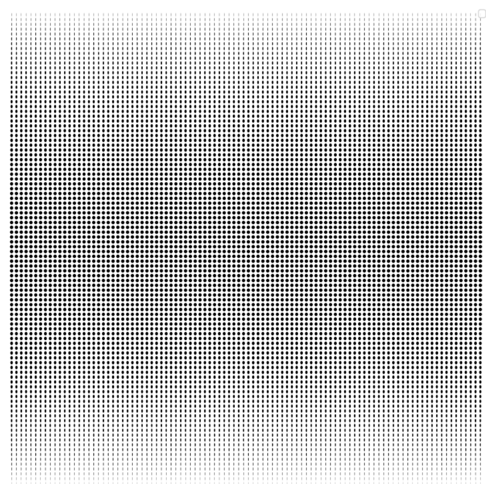


Figure 24: Visualization of sphere metric field

Definition 3.23. **Riemannian manifold** (M, g) is a differentiable (smooth) manifold M equipped with a Riemannian metric g.

Example 31. The Riemannian metric can be equated with a smoothly varying positive-definite symmetric matrix g, called the metric tensor, defined at each point. For two vectors $v, w \in T_pM$, given local coordinates $(x^1, x^2, \cdots, x^n)$ in a neighborhood of p, the entry in g ($n \times n$ matrix) can be expressed like below

$$g_{ij} = \langle E_i, E_j \rangle,$$

where $E_i = \frac{\partial}{\partial x^i}$ are the coordinate basis vectors at p. With this definition, we can compute the inner product $\langle v, w \rangle$ as $v^\top g w$. Also, for a vector v, we can compute the length of the vector as $\langle v, v \rangle^{\frac{1}{2}}$, which is the L^2 norm. Sometimes, people utilize the inverse of the diffusion tensor, D^{-1}, to define a local cost function as

$$\langle v, w \rangle = v^\top D^{-1} w,$$

where $v, w \in T_pM$. In this case, since the inverse of the diffusion tensor are positive-definite symmetric and they are also Riemannian metric, a DTI is actually wrapped into a Riemannian manifold.

Definition 3.24. **Geodesic** between two points $p, q \in M$ can be defined by the minimization of the energy functional

$$E(\gamma) = \int_0^1 \|\dot{\gamma}(t)\|^2 dt$$

where $\gamma : [0, 1] \to M$ is a curve with fixed endpoints $\gamma(0) = p, \gamma(1) = q$. The inner product between two tangent vectors $v, w \in T_x M$ is given by $\langle v, w\rangle = v^\top g(x) w$, where $g(x)$ is the Riemannian metric at point x.

Remark 3.21. *[sta, 2013a]* ***Intrinsic*** *distance is measured as 'an ant would walk along the surface'.* ***Extrinsic*** *distance is defined as the L^2 norm between two points, basically 'how a surface looks from the outside'. From the beginning and through the middle of the 18th century, differential geometry was studied from the extrinsic point of view: curves and surfaces were considered as lying in a Euclidean space of higher dimension. Starting with the work of Riemann, the intrinsic point of view was developed, in which one cannot speak of moving "outside" the geometric object because it is considered to be given in a free-standing way.*

Example 32. [Bhatia, 2009] Let A and B be any two elements of $\mathbb{P}^n$. Then there exists a unique geodesic $[A, B]$ joining A and B. This geodesic has a parametrization

$$\gamma(t) = A^{\frac{1}{2}}(A^{-\frac{1}{2}}BA^{-\frac{1}{2}})^\top A^{\frac{1}{2}}, t \in [0, 1].$$

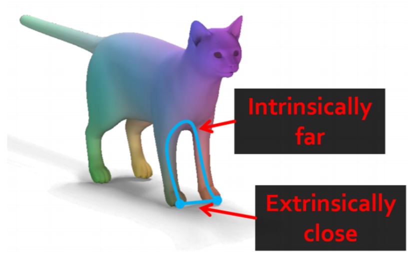


Figure 25: Intrinsic vs. Extrinsic Distance

Definition 3.25. $\boxed{\textbf{Geodesic Equation}}$ guarantees the acceleration vector normal to the surface

$$\frac{d^2u^k}{dt^2} + \Gamma^k_{ij}\frac{du^i}{dt}\cdot\frac{du^j}{dt} = 0$$

Proof. In this section, all the computations are conducted in 2D situation.

$$\text{Velocity Vector: } \frac{d\vec{R}}{dt} = \frac{\partial\vec{R}}{\partial u}\cdot\frac{du}{dt} + \frac{\partial\vec{R}}{\partial v}\cdot\frac{dv}{dt}$$
$$\text{Acceleration Vector: } \frac{d^2\vec{R}}{dt^2} = \frac{d}{dt}\left(\frac{\partial\vec{R}}{\partial u}\cdot\frac{du}{dt} + \frac{\partial\vec{R}}{\partial v}\cdot\frac{dv}{dt}\right)$$

Expend the expression of acceleration vector, we have

$$\begin{aligned}
\frac{d^2\vec{R}}{dt^2} &= \frac{d}{dt}\left(\frac{\partial\vec{R}}{\partial u}\cdot\frac{du}{dt}+\frac{\partial\vec{R}}{\partial v}\cdot\frac{dv}{dt}\right)\\
&= \frac{\partial\vec{R}}{\partial u}\cdot\frac{d^2u}{dt^2}+\frac{du}{dt}\left(\frac{d}{dt}\cdot\frac{\partial\vec{R}}{\partial u}\right)+\frac{\partial\vec{R}}{\partial v}\cdot\frac{d^2v}{dt^2}+\frac{dv}{dt}\left(\frac{d}{dt}\cdot\frac{\partial\vec{R}}{\partial v}\right)\\
&= \frac{\partial\vec{R}}{\partial u}\cdot\frac{d^2u}{dt^2}+\frac{du}{dt}\left(\frac{\partial}{\partial u}\cdot\frac{d\vec{R}}{dt}\right)+\frac{\partial\vec{R}}{\partial v}\cdot\frac{d^2v}{dt^2}+\frac{dv}{dt}\left(\frac{\partial}{\partial v}\cdot\frac{d\vec{R}}{dt}\right)\\
&= \frac{\partial\vec{R}}{\partial u}\cdot\frac{d^2u}{dt^2}+\frac{du}{dt}\left[\frac{\partial}{\partial u}\cdot\left(\frac{\partial\vec{R}}{\partial u}\cdot\frac{du}{dt}+\frac{\partial\vec{R}}{\partial v}\cdot\frac{dv}{dt}\right)\right]\\
&\quad+\frac{\partial\vec{R}}{\partial v}\cdot\frac{d^2v}{dt^2}+\frac{dv}{dt}\left[\frac{\partial}{\partial v}\cdot\left(\frac{\partial\vec{R}}{\partial u}\cdot\frac{du}{dt}+\frac{\partial\vec{R}}{\partial v}\cdot\frac{dv}{dt}\right)\right]\\
&= \frac{\partial\vec{R}}{\partial u}\cdot\frac{d^2u}{dt^2}+\frac{\partial^2\vec{R}}{\partial u^2}\cdot\left(\frac{du}{dt}\right)^2+\frac{\partial^2\vec{R}^2}{\partial u\partial v}\cdot\frac{du}{dt}\cdot\frac{dv}{dt}\\
&\quad+\frac{\partial\vec{R}}{\partial v}\cdot\frac{d^2v}{dt^2}+\frac{\partial^2\vec{R}^2}{\partial u\partial v}\cdot\frac{du}{dt}\cdot\frac{dv}{dt}+\frac{\partial^2\vec{R}}{\partial v^2}\cdot\left(\frac{dv}{dt}\right)^2
\end{aligned}$$

By using Einstein Notation, and making $u^1 = u, u^2 = v$, we can denote the **acceleration vector** as

$$\boxed{\frac{d^2\vec{R}}{dt^2} = \frac{d^2u^i}{dt^2}\cdot\frac{\partial\vec{R}}{\partial u^i}+\frac{du^i}{dt}\cdot\frac{du^j}{dt}\cdot\frac{\partial^2\vec{R}}{\partial u^i\partial u^j}} \tag{1}$$

Assuming that $\frac{\partial^2\vec{R}}{\partial u^i\partial u^j}$ is consist of three components, so we can express it like

$$\frac{\partial^2\vec{R}}{\partial u^i\partial u^j} = \Gamma^1_{ij}\frac{\partial\vec{R}}{\partial u^1}+\Gamma^2_{ij}\frac{\partial\vec{R}}{\partial u^2}+L_{ij}\vec{n}$$

where the Christoffel symbol Γ^k_{ij}, gives us the tangential component of $\frac{\partial^2\vec{R}}{\partial u^i\partial u^j}$ and the second fundamental form L_{ij}, gives us the normal component of $\frac{\partial^2\vec{R}}{\partial u^i\partial u^j}$. By using the Einstein Notation, we can have a more concise form as below:

$$\boxed{\frac{\partial^2\vec{R}}{\partial u^i\partial u^j} = \Gamma^k_{ij}\frac{\partial\vec{R}}{\partial u^k}+L_{ij}\vec{n}} \tag{2}$$

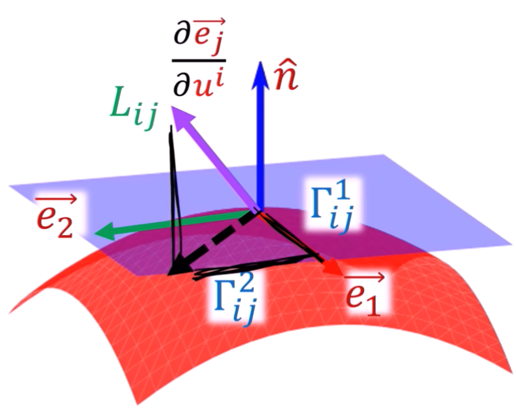


Figure 26: meaning of L_{ij}, Γ^k_{ij}

Finally, by substituting Eq.(2) into Eq.(1), we can have acceleration vector as below

$$\begin{aligned}\frac{d^2\vec{R}}{dt^2} &= \frac{d^2u^i}{dt^2}\cdot\frac{\partial\vec{R}}{\partial u^i} + \frac{du^i}{dt}\cdot\frac{du^j}{dt}\cdot\frac{\partial^2\vec{R}}{\partial u^i\partial u^j} \\ &= \frac{d^2u^i}{dt^2}\cdot\frac{\partial\vec{R}}{\partial u^i} + \frac{du^i}{dt}\cdot\frac{du^j}{dt}\cdot\left(\Gamma^k_{ij}\frac{\partial\vec{R}}{\partial u^k} + L_{ij}\vec{n}\right) \\ &= \underbrace{\left(\frac{d^2u^k}{dt^2} + \Gamma^k_{ij}\frac{du^i}{dt}\cdot\frac{du^j}{dt}\right)\frac{\partial\vec{R}}{\partial u^k}}_{\text{tangential part}} + \underbrace{L_{ij}\cdot\frac{du^i}{dt}\cdot\frac{du^j}{dt}\cdot\vec{n}}_{\text{normal part}}\end{aligned}$$

That acceleration vector normal to the surface requires

$$\frac{d^2u^k}{dt^2} + \Gamma^k_{ij}\frac{du^i}{dt}\cdot\frac{du^j}{dt} = 0,$$

which is the Geodesic Equation.

Derivation of Γ^k_{ij} and L_{ij} As $\vec{n}$ is perpendicular to the tangent vectors, therefore, by multiplying $\frac{\partial\vec{R}}{\partial u^l}$ on both sides of equation above, we can yield that

$$\frac{\partial^2\vec{R}}{\partial u^i\partial u^j}\cdot\frac{\partial\vec{R}}{\partial u^l} = \left(\Gamma^k_{ij}\frac{\partial\vec{R}}{\partial u^k} + L_{ij}\vec{n}\right)\cdot\frac{\partial\vec{R}}{\partial u^l} = \Gamma^k_{ij}\frac{\partial\vec{R}}{\partial u^k}\cdot\frac{\partial\vec{R}}{\partial u^l} \tag{3}$$

Since $\frac{\partial\vec{R}}{\partial u^k}\cdot\frac{\partial\vec{R}}{\partial u^l} = \vec{e}_k\cdot\vec{e}_l = g_{kl}$, then we get

$$\frac{\partial^2\vec{R}}{\partial u^i\partial u^j}\cdot\frac{\partial\vec{R}}{\partial u^l} = \Gamma^k_{ij}g_{kl},$$

By substituting the metric form below into Eq.(3)

$$\begin{pmatrix}\frac{\partial\vec{R}}{\partial u^1}\cdot\frac{\partial\vec{R}}{\partial u^1} & \frac{\partial\vec{R}}{\partial u^1}\cdot\frac{\partial\vec{R}}{\partial u^2} \\ \frac{\partial\vec{R}}{\partial u^2}\cdot\frac{\partial\vec{R}}{\partial u^1} & \frac{\partial\vec{R}}{\partial u^2}\cdot\frac{\partial\vec{R}}{\partial u^2}\end{pmatrix} = \begin{pmatrix}g_{11} & g_{12} \\ g_{21} & g_{22}\end{pmatrix}$$

and with Kronecker delta cancellation rule $g_{kl}\cdot g^{lm} = \delta^m_k$, we can have

$$\begin{aligned}\Gamma^k_{ij}g_{kl}g^{lm} &= \frac{\partial^2\vec{R}}{\partial u^i\partial u^j}\cdot\frac{\partial\vec{R}}{\partial u^l}\cdot g^{lm} \\ \Gamma^k_{ij}\delta^m_k &= \frac{\partial^2\vec{R}}{\partial u^i\partial u^j}\cdot\frac{\partial\vec{R}}{\partial u^l}\cdot g^{lm} \\ \Gamma^m_{ij} &= \left(\frac{\partial\vec{e_j}}{\partial u^i}\cdot\vec{e_l}\right)g^{lm}\end{aligned} \tag{4}$$

Likewise, by multiplying $\vec{n}$ at both side of Eq.(2), we can yield the extrinsic expression of second fundamental form

$$L_{ij} = \frac{\partial^2\vec{R}}{\partial u^i\partial u^j}\cdot\vec{n}$$

□

Definition 3.26. Given a vector space V and a functional $f : V \to \mathbb{R}$, $x, h \in V, \alpha \in \mathbb{R}$, if the limit

$$\delta f(x) = \lim_{\alpha\to 0}\frac{1}{\alpha}[f(x+\alpha h) - f(x)]$$

exists, it's called the **Gâteaux derivative** of f at x with increment h. If the limit exists for $\forall h \in V$, the functional f is said to be Gâteaux differentiable at x.

Example 33. Given $x \in \mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}$, which has continuous partial derivatives with respect to each components of x. Then, the Gateaux derivative of f is

$$\delta f(x) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} h_i = \langle \nabla f, h \rangle$$

Definition 3.27. $\boxed{\textbf{Directional derivative}}$ of a multivariate differentiable function along a given vector v at a given point x intuitively represents the instantaneous rate of change of the function, moving through x with a velocity specified by h.

$$\nabla_h f(x) = Df(x)(h) = \langle \nabla f, h \rangle$$

Remark 3.22. *Relationship between* ***partial derivative(scalar), directional derivative(scalar) and gradient(vector)****:*

- *The vector consists of partial derivatives is the gradient.*
- *The linear combination of partial derivatives is directional derivative.*
- *Partial derivative is a special directional derivative, which is along the axis.*

Remark 3.23. *Relationship between* ***Gâteaux derivative(scalar), directional derivative(scalar) and covariant derivative(vector)****:*

- *Gâteaux derivative(differential) is a generalization of the concept of directional derivative in differential calculus.*
- *Covariant derivative is a generalization of the directional derivative from vector calculus. The covairant derivative of a function is directional*
- *Gâteaux and directional derivative are applicable to functional $f : \mathbb{R}^n \to \mathbb{R}$, so their output are both scalars. While the covariant derivative is for vector field $v : \mathbb{R}^n \to \mathbb{R}^n$, so its output is still a vector.*

$$\begin{aligned} \nabla_h f(x) &= h^i \nabla_{\frac{\partial}{\partial x^i}} f \\ &= h^i \frac{\partial f}{\partial x^i} \end{aligned}$$

$$\begin{aligned} \nabla_h v &= h^i \nabla_{\frac{\partial}{\partial u^i}} (v^j \vec{e_j}) \\ &= h^i \left(\frac{\partial v^j}{\partial u^i} \vec{e_j} + v^j \nabla_{\frac{\partial}{\partial u^i}} \vec{e_j} \right) \\ &= h^i \left(\frac{\partial v^j}{\partial u^i} \vec{e_j} + v^j \frac{\partial \vec{e_j}}{\partial u^i} \right) \\ &= h^i \left(\frac{\partial v^j}{\partial u^i} \vec{e_j} + v^j \Gamma^k_{ij} \vec{e_k} \right) && \triangleright \frac{\partial \vec{e_j}}{\partial u^i} = \Gamma^k_{ij} \vec{e_k} \\ &= h^i \left(\frac{\partial v^k}{\partial u^i} \vec{e_k} + v^j \Gamma^k_{ij} \vec{e_k} \right) \\ &= h^i \left(\frac{\partial v^k}{\partial u^i} + v^j \Gamma^k_{ij} \right) \vec{e_k} \end{aligned}$$

Definition 3.28. $\boxed{\textbf{Covariant derivative}}$ $\nabla_{\vec{w}} \vec{v}$, refers to Levi-Civita connection generally,

- is the ordinary derivative for Euclidean space.

- is the rate of change vector at $\vec{v}$ of a vector field in a direction $\vec{w}$ with the normal component subtracted, extrinsically.

Levi-Civita connection has following properties:

1. $\nabla_{a\vec{w}+b\vec{t}}\vec{v} = a\nabla_{\vec{w}}\vec{v} + b\nabla_{\vec{t}}\vec{v}$
2. $\nabla_{\vec{w}}(\vec{v}+\vec{u}) = \nabla_{\vec{w}}\vec{v} + \nabla_{\vec{w}}\vec{u}$ ▷ Distributive Property
3. $\nabla_{\vec{w}}(\vec{v}\cdot\vec{u}) = (\nabla_{\vec{w}}\vec{v})\cdot\vec{u} + \vec{v}\cdot(\nabla_{\vec{w}}\vec{u})$ ▷ Product Rule
4. $\nabla_{\vec{w}}(a\vec{v}) = (\nabla_{\vec{w}}a)\vec{v} + a(\nabla_{\vec{w}}\vec{v})$
5. $\nabla_{\partial_i}(a) = \frac{\partial a}{\partial u^i}$
6. $\nabla_{\vec{w}}\vec{v} = \nabla_{\vec{v}}\vec{w}$ ▷ Commutative Property

Remark 3.24.
- *In Euclidean space, the covariant derivative is simply the change of the vector fields that take changing basis vectors into account.*
- ***Parallel transport*** *provides a way to compare a vector in one tangent plane to a vector in another, by moving the vector along a curve without changing it.*
- *Different expressions of Γ^m_{kj} give us different ways of "parallel transport". If $\Gamma^m_{kj} = \frac{1}{2}g^{im}\left(\frac{\partial g_{ij}}{\partial u^k} + \frac{\partial g_{ki}}{\partial u^j} - \frac{\partial g_{jk}}{\partial u^i}\right)$, then it's Levi-Civita connection. If $\Gamma^m_{kj} = 0$, it's another connection.*

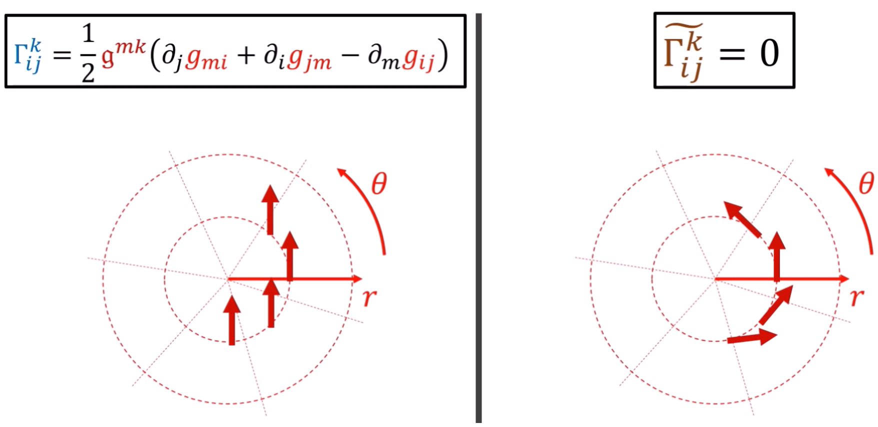


Figure 27: Different definitions of Γ^m_{kj} give us different kinds of "parallel transport".

- *Covariant derivative helps us find parallel transported vector fields. $\nabla_{\vec{w}}\vec{v} = \vec{0}$ means the vector $\vec{v}$ is parallel transported in the direction $\vec{w}$ at $\vec{v}$'s position.*
- *[rg,] Covariant derivative $\nabla_{\vec{w}}\vec{v}$ is the difference between a vector field v and its parallel transport in the direction w.*

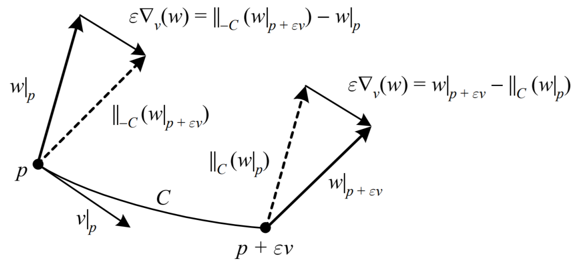


Figure 28: Difference between a vector and the parallel transported one

- *Covariant derivative provides a connection between tangent spaces in a curved space.*
- *In curved space, a geodesic has zero tangential acceleration when we travel along it at constant speed. To compute geodesic curves, we need to find curves where the acceleration vector is normal to the space, namely* $\nabla_{\dot{\gamma}(t)}\dot{\gamma}(t) = \vec{0}$ *holds along the curve* γ.
- *In other words, geodesic is a curve resulting from parallel transporting a vector along itself.*
- *In the special case of a manifold isometrically embedded into a higher-dimensional Euclidean space, the covariant derivative can be viewed as the orthogonal projection of the Euclidean directional derivative onto the manifold's tangent space. In this case the Euclidean derivative is broken into two parts, the extrinsic normal component (dependent on the embedding) and the intrinsic covariant derivative component.*

Example 34. [YouTube, 2018a] In a 2D Euclidean space, we represent a vector as below:

$$\vec{v} = v^1\vec{e_1} + v^2\vec{e_2} = \sum_i v^i\vec{e_i} = v^i\vec{e_i},$$

where v^1, v^2 are constant.

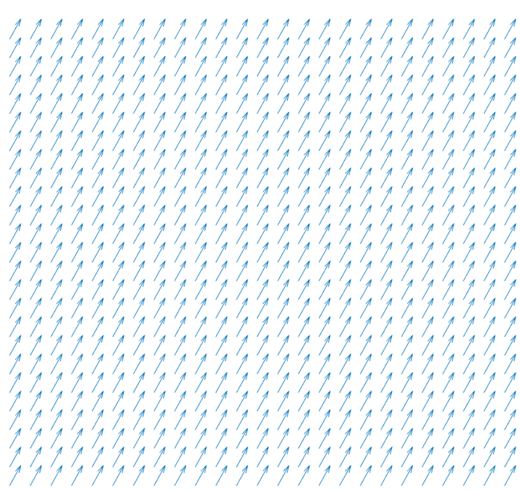


Figure 29: Constant vector space

The covariant derivative defined in Euclidean space is just intuitive:

$$
\begin{aligned}
\nabla_{\frac{\partial}{\partial u^1}} \vec{v} &= \frac{\partial}{\partial u^1}\vec{v} \\
&= \frac{\partial}{\partial u^1}(v^1\vec{e_1} + v^2\vec{e_2}) \\
&= \frac{\partial}{\partial u^1}(v^1\vec{e_1}) + \frac{\partial}{\partial u^1}(v^2\vec{e_2}) \\
&= \frac{\partial v^1}{\partial u^1}\vec{e_1} + v^1\frac{\partial \vec{e_1}}{\partial u^1} + \frac{\partial v^2}{\partial u^1}\vec{e_2} + v^2\frac{\partial \vec{e_2}}{\partial u^1} \\
&= \frac{\partial v^1}{\partial u^1}\vec{e_1} + \frac{\partial v^2}{\partial u^1}\vec{e_2} \\
&= \vec{0} \\
\nabla_{\frac{\partial}{\partial u^2}} \vec{v} &= \frac{\partial}{\partial u^2}\vec{v} \\
&= \frac{\partial v^1}{\partial u^2}\vec{e_1} + \frac{\partial v^2}{\partial u^2}\vec{e_2} \\
&= \vec{0}
\end{aligned}
$$

In conclusion, in Euclidean space, the covariant derivative of a vector field is just the ordinary derivative. We need to make sure to differentiate both the vector components and the basis vectors.

$$
\begin{aligned}
\frac{\partial}{\partial u^i}\vec{v} &= \frac{\partial}{\partial u^i}v^j\vec{e_j} \\
&= \underbrace{\frac{\partial v^j}{\partial u^i}\vec{e_j}}_{\text{components}} + \underbrace{\frac{\partial \vec{e_j}}{\partial u^i}v^j}_{\text{basis vectors}}
\end{aligned}
$$

Example 35. [YouTube, 2018b] In extrinsic case, the covariant derivative is defined as below

$$
\begin{aligned}
\nabla_{\frac{\partial}{\partial u^i}} \vec{v} &= \frac{\partial \vec{v}}{\partial u^i} - \vec{n} \\
&= \frac{\partial}{\partial u^i}(v^1\vec{e_1} + v^2\vec{e_2}) - \vec{n} \\
&= \frac{\partial}{\partial u^i}v^j\vec{e_j} - \vec{n} \\
&= \frac{\partial v^j}{\partial u^i}\vec{e_j} + \frac{\partial \vec{e_j}}{\partial u^i}v^j - \vec{n} \\
&= \frac{\partial v^j}{\partial u^i}\vec{e_j} + (\Gamma^k_{ij}\vec{e_k} + L_{ij}\hat{n})v^j - \vec{n} && \triangleright\ \frac{\partial \vec{e_j}}{\partial u^i} = \Gamma^1_{ij}\vec{e_1} + \Gamma^2_{ij}\vec{e_2} + L_{ij}\hat{n} \\
&= \frac{\partial v^j}{\partial u^i}\vec{e_j} + \Gamma^k_{ij}\vec{e_k}v^j \\
&= \frac{\partial v^k}{\partial u^i}\vec{e_k} + \Gamma^k_{ij}\vec{e_k}v^j \\
&= \left(\frac{\partial v^k}{\partial u^i} + \Gamma^k_{ij}v^j\right)\vec{e_k}
\end{aligned} \tag{5}
$$

where L_{ij} is the second fundamental form and Γ^k_{ij} is in form of Eq.(4).

Parameterize the space with tangent space basis

$$
\begin{aligned}
\vec{R} &= [X, Y, Z]^{\top} \\
\text{where } X &= \cos(u^2)\sin(u^1) \\
Y &= \sin(u^2)\sin(u^1) \\
Z &= \cos(u^1),
\end{aligned}
$$

where $\vec{R}$ is the position vector and u^1, u^2 represents the latitude and longitude, respectively.

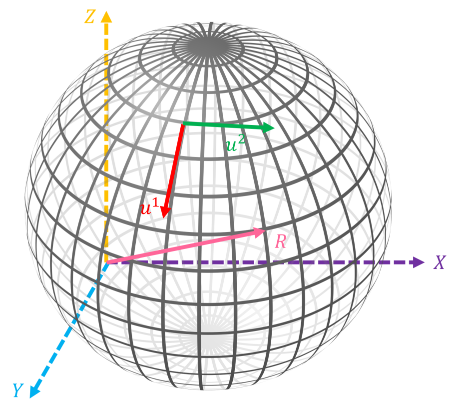


Figure 30: Parametric equations for sphere

By using the chain rule, we have

$$
\begin{aligned}
\vec{e}_1 = \frac{\partial \vec{R}}{\partial u^1} &= +\cos(u^2)\cos(u^1)\frac{\partial \vec{R}}{\partial X} + \sin(u^2)\cos(u^1)\frac{\partial \vec{R}}{\partial Y} - \sin(u^1)\frac{\partial \vec{R}}{\partial Z} \\
&= +\cos(u^2)\cos(u^1)\vec{e}_X + \sin(u^2)\cos(u^1)\vec{e}_Y - \sin(u^1)\vec{e}_Z
\end{aligned} \tag{6}
$$

$$
\begin{aligned}
\vec{e}_2 = \frac{\partial \vec{R}}{\partial u^2} &= -\sin(u^2)\sin(u^1)\frac{\partial \vec{R}}{\partial X} + \cos(u^2)\sin(u^1)\frac{\partial \vec{R}}{\partial Y} \\
&= -\sin(u^2)\sin(u^1)\vec{e}_X + \cos(u^2)\sin(u^1)\vec{e}_Y
\end{aligned} \tag{7}
$$

With Eq.(6,7), we can yield the metric as below

$$
g_{ij} = \begin{pmatrix} \vec{e}_1 \cdot \vec{e}_1 & \vec{e}_1 \cdot \vec{e}_2 \\ \vec{e}_2 \cdot \vec{e}_1 & \vec{e}_2 \cdot \vec{e}_2 \end{pmatrix} = \begin{pmatrix} \frac{\partial \vec{R}}{\partial u^1} \cdot \frac{\partial \vec{R}}{\partial u^1} & \frac{\partial \vec{R}}{\partial u^1} \cdot \frac{\partial \vec{R}}{\partial u^2} \\ \frac{\partial \vec{R}}{\partial u^2} \cdot \frac{\partial \vec{R}}{\partial u^1} & \frac{\partial \vec{R}}{\partial u^2} \cdot \frac{\partial \vec{R}}{\partial u^2} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & \sin^2(u^1) \end{pmatrix}
$$

$$
g^{ij} = \begin{pmatrix} \vec{e}_1 \cdot \vec{e}_1 & \vec{e}_1 \cdot \vec{e}_2 \\ \vec{e}_2 \cdot \vec{e}_1 & \vec{e}_2 \cdot \vec{e}_2 \end{pmatrix}^{-1} = \begin{pmatrix} \frac{\partial \vec{R}}{\partial u^1} \cdot \frac{\partial \vec{R}}{\partial u^1} & \frac{\partial \vec{R}}{\partial u^1} \cdot \frac{\partial \vec{R}}{\partial u^2} \\ \frac{\partial \vec{R}}{\partial u^2} \cdot \frac{\partial \vec{R}}{\partial u^1} & \frac{\partial \vec{R}}{\partial u^2} \cdot \frac{\partial \vec{R}}{\partial u^2} \end{pmatrix}^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{\sin^2(u^1)} \end{pmatrix}
$$

Substituting Eq.(6,7) into the second derivative of position vector, we get

$$
\begin{aligned}
\frac{\partial \vec{e_1}}{\partial u^1} &= \frac{\partial}{\partial u^1}\left(\frac{\partial \vec{R}}{\partial u^1}\right) \\
&= -\cos(u^2)\cos(u^1)\frac{\partial \vec{R}}{\partial X} - \sin(u^2)\sin(u^1)\frac{\partial \vec{R}}{\partial Y} - \cos(u^1)\frac{\partial \vec{R}}{\partial Z} \\
&= -\cos(u^2)\cos(u^1)\vec{e}_X - \sin(u^2)\sin(u^1)\vec{e}_Y - \cos(u^1)\vec{e}_Z \qquad (8) \\
\frac{\partial \vec{e_2}}{\partial u^2} &= \frac{\partial}{\partial u^2}\left(\frac{\partial \vec{R}}{\partial u^2}\right) \\
&= -\cos(u^2)\sin(u^1)\frac{\partial \vec{R}}{\partial X} - \sin(u^2)\sin(u^1)\frac{\partial \vec{R}}{\partial Y} \\
&= -\cos(u^2)\sin(u^1)\vec{e}_X - \sin(u^2)\sin(u^1)\vec{e}_Y \qquad (9) \\
\frac{\partial \vec{e_2}}{\partial u^1} &= \frac{\partial}{\partial u^1}\left(\frac{\partial \vec{R}}{\partial u^2}\right) \\
&= -\sin(u^2)\cos(u^1)\frac{\partial \vec{R}}{\partial X} + \cos(u^2)\cos(u^1)\frac{\partial \vec{R}}{\partial Y} \\
&= -\sin(u^2)\cos(u^1)\vec{e}_X + \cos(u^2)\cos(u^1)\vec{e}_Y \qquad (10)
\end{aligned}
$$

Substituting g^{ij} and Eq.(6,7,8,9,10) into Eq.(4), we can yield the Christoffel symbols as below

$$
\begin{array}{llll}
\Gamma^1_{11} = 0 & \Gamma^1_{12} = 0 & \Gamma^1_{21} = 0 & \Gamma^1_{22} = -\frac{1}{2}\sin(2u^1) \\
\Gamma^2_{11} = 0 & \Gamma^2_{12} = \cot(u^1) & \Gamma^2_{21} = \cot(u^1) & \Gamma^2_{22} = 0
\end{array}
$$

Substituting the Christoffel symbols into Eq.(5), we finally get the extrinsic expression of covariant derivative on sphere as below

$$
\begin{aligned}
\nabla_{\vec{e_1}}\vec{v} &= \left(\frac{\partial v^2}{\partial u^1} + v^2\cot(u^1)\right)\vec{e_2} \\
\nabla_{\vec{e_2}}\vec{v} &= \left(\frac{\partial v^1}{\partial u^2} - \frac{1}{2}\sin(2u^1)v^2\right)\vec{e_1} + \left(\frac{\partial v^2}{\partial u^2} + v^1\cot(u^1)\right)\vec{e_2} \qquad (11)
\end{aligned}
$$

We initialize two different vector field along the equator to see what does covariant derivative exactly mean:

- The first vector field along the equator is

$$\vec{v} = \cos(u^2)\vec{e_1} + \sin(u^2)\vec{e_2} \qquad \text{where } u^1 = \frac{\pi}{2}, u^2 = \lambda \in [0, \frac{\pi}{2}]$$

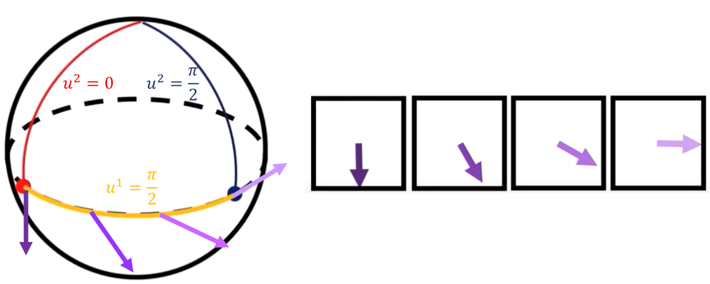


Figure 31: Exponential and log map

Substitute the u^1, u^2 into Eq.(11), we have

$$\begin{aligned}\nabla_{\vec{e_2}}\vec{v} &= \left(\frac{\partial v^1}{\partial u^2} - \frac{1}{2}\sin(2u^1)v^2\right)\vec{e_1} + \left(\frac{\partial v^2}{\partial u^2} + v^1\cot(u^1)\right)\vec{e_2} \\ &= -\sin(u^2)\vec{e_1} + \cos(u^2)\vec{e_2},\end{aligned}$$

since $\nabla_{\vec{e_2}}\vec{v} \neq \vec{0}$, which means the rate of change is not completely normal to the tangent space.

- The second vector field along the equator is

$$\vec{v} = 0\vec{e_1} + 1\vec{e_2} \qquad\qquad \text{where } u^1 = \frac{\pi}{2}, u^2 = \lambda \in [0, \frac{\pi}{2}]$$

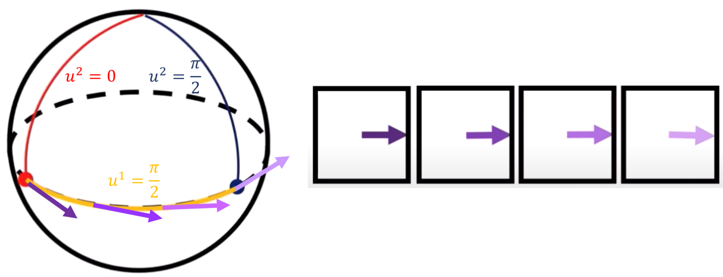


Figure 32: Exponential and log map

Since the vector field has nothing to do with u^1, u^2, we have

$$\begin{aligned}\nabla_{\vec{e_2}}\vec{v} &= \left(\frac{\partial v^1}{\partial u^2} - \frac{1}{2}\sin(2u^1)v^2\right)\vec{e_1} + \left(\frac{\partial v^2}{\partial u^2} + v^1\cot(u^1)\right)\vec{e_2} \\ &= (0-0)\vec{e_1} + (0+0)\vec{e_2} = \vec{0},\end{aligned}$$

which means the rate of change doesn't exist in the tangent space. And this is exactly the geodesic which is resulted from parallel transporting a vector along itself.

Example 36. [YouTube, 2018c] In extrinsic case, we have to subtract the normal component, however, in intrinsic case, such normal component doesn't exist, so we have

$$\begin{aligned}\nabla_{\frac{\partial}{\partial u^i}}\vec{v} &= \frac{\partial \vec{v}}{\partial u^i} \\ &= \frac{\partial}{\partial u^i}(v^j\vec{e_j}) \\ &= \frac{\partial v^j}{\partial u^i}\vec{e_j} + v^j\frac{\partial \vec{e_j}}{\partial u^i} \\ &= \frac{\partial v^k}{\partial u^i}\vec{e_k} + v^j\Gamma^k_{ij}\vec{e_k} && \triangleright\ \frac{\partial \vec{e_j}}{\partial u^i} = \Gamma^k_{ij}\vec{e_k} \\ &= \left(\frac{\partial v^k}{\partial u^i} + v^j\Gamma^k_{ij}\right)\vec{e_k}\end{aligned}$$

The only difference between the extrinsic and intrinsic cases lies in the calculation of Christoffel symbol. Previously, we derived the christoffel symbol in Eq.(4), by inner product between Eq.(6,7,8,9,10), given the position vector. However, in intrinsic case, there's no longer a position vector, so we have to find another way to derive the Christoffel symbol. And it turns out using the metric:

$$
\begin{aligned}
\frac{\partial}{\partial u^k} g_{ij} &= \frac{\partial}{\partial u^k}(\vec{e_i} \cdot \vec{e_j}) && \triangleright\, g_{ij} = \frac{\partial}{\partial u^i} \cdot \frac{\partial}{\partial u^j} = \vec{e_i} \cdot \vec{e_j} \\
&= \frac{\partial \vec{e_i}}{\partial u^k} \cdot \vec{e_j} + \vec{e_i} \cdot \frac{\partial \vec{e_j}}{\partial u^k} \\
&= (\Gamma^l_{ik}\vec{e_l}) \cdot \vec{e_j} + \vec{e_i} \cdot (\Gamma^l_{jk}\vec{e_l}) \\
&= \Gamma^l_{ik}(\vec{e_l} \cdot \vec{e_j}) + \Gamma^l_{jk}(\vec{e_i} \cdot \vec{e_l}) \\
&= \Gamma^l_{ik} g_{lj} + \Gamma^l_{jk} g_{il}
\end{aligned}
$$

Similarly, we can yield other two expressions:

$$
\begin{aligned}
\frac{\partial g_{ij}}{\partial u^k} &= \Gamma^l_{ik} g_{jl} + \Gamma^l_{jk} g_{il} \\
\frac{\partial g_{ki}}{\partial u^j} &= \Gamma^l_{kj} g_{il} + \Gamma^l_{ij} g_{kl} \\
\frac{\partial g_{jk}}{\partial u^i} &= \Gamma^l_{ji} g_{kl} + \Gamma^l_{ki} g_{jl}
\end{aligned}
$$

Add the two of them up and subtract the left one:

$$
\begin{aligned}
\frac{\partial g_{ij}}{\partial u^k} + \frac{\partial g_{ki}}{\partial u^j} - \frac{\partial g_{jk}}{\partial u^i} &= \Gamma^l_{ik} g_j l + \Gamma^l_{jk} g_{il} + \Gamma^l_{kj} g_{il} + \Gamma^l_{ij} g_{kl} - \Gamma^l_{ji} g_{kl} - \Gamma^l_{ki} g_{jl} \\
&= \Gamma^l_{jk} g_{il} + \Gamma^l_{kj} g_{il} \\
&= 2\Gamma^l_{kj} g_{il}
\end{aligned}
$$

Times g^{im} on both sides, we can finally get the intrinsic expression of the Christoffel symbol:

$$
\begin{aligned}
2\Gamma^l_{kj} g_{il} g^{im} &= g^{im}\left(\frac{\partial g_{ij}}{\partial u^k} + \frac{\partial g_{ki}}{\partial u^j} - \frac{\partial g_{jk}}{\partial u^i}\right) \\
\Gamma^l_{kj}\delta^m_l &= \frac{1}{2} g^{im}\left(\frac{\partial g_{ij}}{\partial u^k} + \frac{\partial g_{ki}}{\partial u^j} - \frac{\partial g_{jk}}{\partial u^i}\right) \\
\Gamma^m_{kj} &= \frac{1}{2} g^{im}\left(\frac{\partial g_{ij}}{\partial u^k} + \frac{\partial g_{ki}}{\partial u^j} - \frac{\partial g_{jk}}{\partial u^i}\right)
\end{aligned}
$$

The derivation below illustrates the extrinsic and intrinsic expressions of the Christoffel symbols are actually the same:

$$
\begin{aligned}
\frac{\partial g_{ij}}{\partial u^k} + \frac{\partial g_{ki}}{\partial u^j} - \frac{\partial g_{jk}}{\partial u^i} &= \frac{\partial \vec{e_i}}{\partial u^k} \cdot \vec{e_j} + \vec{e_i} \cdot \frac{\partial \vec{e_j}}{\partial u^k} && \triangleright\, \frac{\partial g_{ij}}{\partial u^k} = \frac{\partial \vec{e_i}}{\partial u^k} \cdot \vec{e_j} + \vec{e_i} \cdot \frac{\partial \vec{e_j}}{\partial u^k} \\
&+ \frac{\partial \vec{e_k}}{\partial u^j} \cdot \vec{e_i} + \vec{e_k} \cdot \frac{\partial \vec{e_j}}{\partial u^i} \\
&- \frac{\partial \vec{e_j}}{\partial u^i} \cdot \vec{e_k} - \vec{e_j} \cdot \frac{\partial \vec{e_i}}{\partial u^k} \\
&= 2\vec{e_i} \cdot \frac{\partial \vec{e_k}}{\partial u^j} && \triangleright\, \frac{\vec{e_j}}{u^i} = \frac{\vec{e_i}}{u^j}
\end{aligned}
$$

$$
\begin{aligned}
\Gamma^m_{kj} &= \frac{1}{2} g^{im}\left(\frac{\partial g_{ij}}{\partial u^k} + \frac{\partial g_{ki}}{\partial u^j} - \frac{\partial g_{jk}}{\partial u^i}\right) \\
&= \frac{1}{2} g^{im} \cdot 2\vec{e_i} \cdot \frac{\partial \vec{e_j}}{\partial u^j} \\
&= \left(\vec{e_i} \cdot \frac{\partial \vec{e_k}}{\partial u^j}\right) g^{mi}
\end{aligned}
$$

Definition 3.29. $\boxed{\textbf{First fundamental form}}$ on manifold M is the field which assigns to each $p \in M$ the bilinear map

$$\begin{aligned} g_p(v,w) &: T_pM \times T_pM \to \mathbb{R} \\ g_p(v,w) &= \langle v, w \rangle \end{aligned}$$

where $v, w \in T_pM$.

Definition 3.30. $\boxed{\textbf{Second fundamental form}}$ on manifold M is defined by

$$\begin{aligned} h_p(v,w) &: T_pM \times T_pM \to T_pM^{\perp} \\ h_p(v,w) &= (d\Pi(p)v)w = (d\Pi(p)w)v \end{aligned}$$

where $p \in M$ and $v, w \in T_pM$.

Definition 3.31. $\boxed{\textbf{Riemannian exponential map}}$ takes the position $p = \gamma(0) \in M$ and velocity $v = \dot{\gamma}(0) \in T_pM$ as input and returns the point at time 1 along the geodesic with these initial conditions. When γ is defined over the interval $[0, 1]$, the Riemannian exponential map at p is defined as

$$\begin{aligned} \mathrm{Exp}_p(v) &: T_pM \to M \\ \mathrm{Exp}_p(v) &= \mathrm{Exp}(p, v) = \gamma(1) \end{aligned}$$

Remark 3.25. *The exponential map is a diffeomorphism in a neighborhood of zero. The inverse of it in this neighborhood is the Riemannian log map.*

Example 37. For a Lie group with bi-invariant metric, the Lie group exponential map is the same with the Riemannian exponential map at the identity, that is, for any tangent vector $X \in \mathfrak{g}$, we have

$$\exp(X) = \mathrm{Exp}_e(X).$$

For matrix groups, the Lie group exponential map of a matrix $X \in \mathfrak{gl}(n)$ is computed by the formular

$$\exp(X) = \sum_{k=0}^{\infty} \frac{1}{k!} X^k.$$

This series converges absolutely for all $X \in \mathfrak{gl}(n)$

Definition 3.32. $\boxed{\textbf{Riemannian log map}}$ is the inverse of Riemannian exponential map, defined in the neighborhood $\mathrm{Exp}_p(v)$

$$\begin{aligned} \mathrm{Log}_p &: M \to T_pM \\ \mathrm{Log}_p(\gamma(1)) &= v \end{aligned}$$

Remark 3.26. *The matrix logarithm of M is defined as*

$$\log(M) = X^{-1} \log(D) X,$$

where $M \in \mathcal{R}^{n \times n}$ is a diagonalizable matrix, $X \in \mathcal{R}^{n \times n}$ and $D \in \mathcal{R}^{n \times n}$ is a diagonal matrix. $\log(D) \in \mathcal{R}^{n \times n}$ is also a diagonal matrix with diagonal elements equals to the logarithm of the corresponding diagonal elements of D.

Remark 3.27. *According to the property above, we can also have*

$$\begin{aligned} \mathrm{tr}(\log(M)) &= \mathrm{tr}(X^{-1} \log(D) X) \\ &= \mathrm{tr}(X X^{-1} \log(D)) \\ &= \mathrm{tr}(\log(D)) \end{aligned}$$

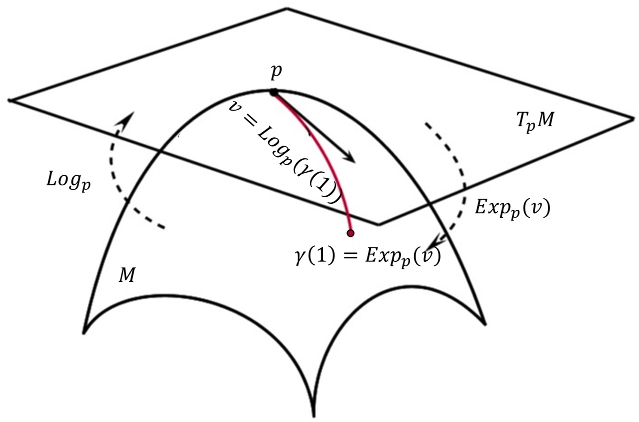


Figure 33: Exponential and log map

Definition 3.33. $\boxed{\textbf{Group}}$ G is a set of elements with a binary operation $\cdot$, such that

1. $\forall x, y \in G, x \cdot y \in G$
2. $\forall x, y \in G, (x \cdot y) \cdot z = x \cdot (y \cdot z)$
3. $\exists e \in G, \forall x \in G$ satisfy $e \cdot x = x \cdot e = x$, where e is unique
4. $\forall x \in G, \exists y \in G$ such that $y \cdot x = x \cdot y = e$

Definition 3.34. $\boxed{\textbf{Lie group}}$ is a smooth manifold equipped with group structures, where the two group operations

$$\begin{aligned}
&(x, y) \to x \cdot y : G \times G \to G && \text{Multiplication} \\
&x \to x^{-1} : G \to G && \text{Inverse}
\end{aligned}$$

are both smooth mappings. In other words, a Lie group adds a smooth manifold structure to a group.

Remark 3.28. *The group action should be thought of as a transformation of the manifold M, just as matrices are transformations of Euclidean space.[Fletcher et al., 2004b]*

Example 38. Many common geometric transformations in Euclidean space form Lie groups. For example, rotations, translations, magnifications, and affine transformations of $\mathbb{R}^n$ all form Lie groups. More generally, Lie groups can be used to describe transformations of smooth manifolds.[Fletcher et al., 2004b]

Definition 3.35. $\boxed{\textbf{Orbit}}$ of a point $p \in M$ is defined as

$$G(p) = \{g \cdot p : g \in G\}$$

Example 39. If $G = SO(2)$, the orbit of point p is a circle.

Definition 3.36. $\boxed{\textbf{Lie algebra}}$ is a **vector space** $\mathfrak{g}$ together with an operation called Lie bracket $[\cdot, \cdot]$, a alternating bilinear map $\mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}$. $\forall x, y, z \in \mathfrak{g}$ and $a, b \in \mathbb{R}$, the following axioms are satisfied:

1. Linearity: $[ax + by, z] = a[x, z] + b[y, z]$
2. Anticommutativity: $[x, y] = -[y, x] = x \cdot y - y \cdot x$
3. Jacobi identity: $[x, [y, z]] + [z, [x, y]] + [y, [z, x]] = 0$

Definition 3.37. $\boxed{\textbf{Lie bracket (Lie derivative) of vector fields}}$ $[\cdot,\cdot]$ is an operator that assigns to any two vector fields X and Y on a smooth manifold M a third vector field denoted $[X,Y]$, and is sometimes denoted $\mathcal{L}_X Y$ ("Lie derivative of Y along X"):

$$[X,Y] = \sum_{i=1}^{n}\sum_{j=1}^{n}\left(X^j\partial_j Y^i - Y^j\partial_j X^i\right)\partial_i.$$

If M is (an open subset of) $\mathbb{R}^n$, then the vector fields X and Y can be written as smooth maps of the form $X : M \to \mathbb{R}^n$ and $Y : M \to \mathbb{R}^n$, and the Lie bracket $[X,Y] : M \to \mathbb{R}^n$ is given by:

$$[X,Y] = \mathbf{J}_Y X - \mathbf{J}_X Y$$

where $\mathbf{J}_Y$ and $\mathbf{J}_X$ are $n \times n$ Jacobian matrices ($\partial_j Y^i$ and $\partial_j X^i$ respectively using index notation) multiplying the $n \times 1$ column vectors X and Y.

Remark 3.29. *Geometrically, the Lie bracket of vector fields gives information about how they "fail to commute" when applied to different points on the manifold.*

Remark 3.30. *Coordinate lines are just flow curves along the basis vector. Coordinate flow curves always close, which means Lie bracket of basis vectors always has to be zero vector.*

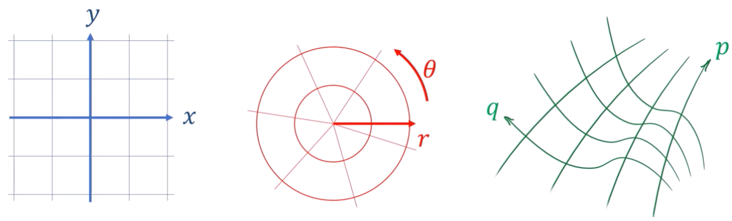


Figure 34: Coordinate lines

Lie bracket(commutator) measures how much vector field flow curves fail to close.

Example 40. The example below shows how to calculate the flow curve, given the flow field.

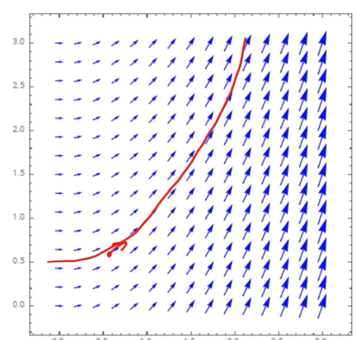


Figure 35: Flow field

Vector arrows tell you velocity at each point:

$$\vec{w} = 1\vec{e_x} + x\vec{e_y}.$$

To express the vector in the way of position vector $\vec{R}$, we have

$$\begin{aligned}\vec{w} = \frac{d\vec{R}}{d\lambda} &= \frac{dx}{d\lambda}\frac{\partial\vec{R}}{\partial x} + \frac{dy}{d\lambda}\frac{\partial\vec{R}}{\partial y}\\ &= \frac{dx}{d\lambda}\vec{e_x} + \frac{dy}{d\lambda}\vec{e_y}\end{aligned}$$

Associating the two expressions above, we can yield the respective expressions of x, y.

$$\frac{dx}{d\lambda} = 1 \rightarrow x = \lambda + c_1$$
$$\frac{dy}{d\lambda} = x = \lambda + c_1 \rightarrow y = \frac{1}{2}\lambda^2 + c_1\lambda + c_2$$

Assuming $c_1 = c_2 = 0$, this can be a possible flow curve

$$x(\lambda) = \lambda, y(\lambda) = \frac{1}{2}\lambda^2,$$

which is tangent to all vectors in a vector field.

Example 41. The Lie bracket of the vector field is defined below:

$$[\vec{u}, \vec{v}] = \underbrace{\vec{u}(\vec{v})}_{\text{derivative of } \vec{v} \text{ in the direction of } \vec{u}} - \vec{v}(\vec{u})$$

We separate the flow field in the example above into two fields

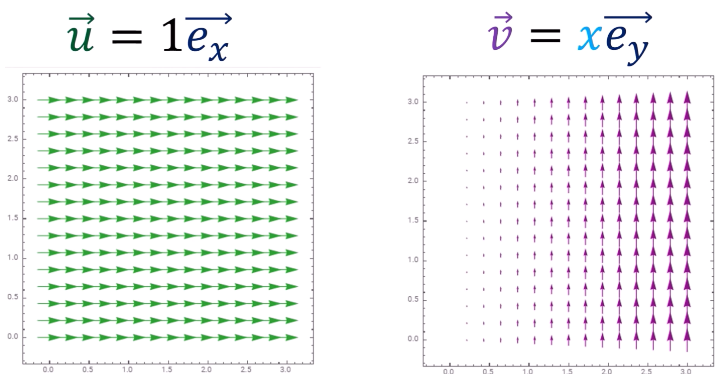


Figure 36: Two vector fields

Derivative of $\vec{u}$ in the direction of $\vec{v}$ is shown below

$$\begin{aligned}
\vec{v}(\vec{u}) &= v^i \vec{e_i}(u^j \vec{e_j}) \\
&= v^i \partial_i (u^j \partial_j) \\
&= v^i [(\partial_i u^j)\partial_j + u^j (\partial_i \partial_j)] && \triangleright \text{ product rule} \\
&= v^i (\partial_i u^j)\partial_j + v^i u^j (\partial_i \partial_j) \\
&= v^y (\partial_y u^x)\partial_x + v^y u^x (\partial_y \partial_x) \\
&= x(\partial_y 1)\partial_x + x \cdot 1(\partial_y \partial_x) \\
&= x(\partial_y \partial_x) \\
&= x(\partial_y \vec{e_x}) \\
&= x\frac{\vec{e_x}}{\partial_y} \\
&= \vec{0}
\end{aligned}$$

Derivative of $\vec{v}$ in the direction of $\vec{u}$ is shown below

$$\begin{aligned}
\vec{u}(\vec{v}) &= u^i \vec{e_i}(v^j \vec{e_j}) \\
&= u^i \partial_i (v^j \partial_j) \\
&= u^i [(\partial_i v^j)\partial_j + v^j(\partial_i \partial_j)] && \triangleright \text{ product rule} \\
&= u^i (\partial_i v^j)\partial_j + u^i v^j (\partial_i \partial_j) \\
&= u^x (\partial_y v^y)\partial_y + u^x v^y (\partial_x \partial_y) \\
&= 1(\partial_x x)\partial_y + 1 \cdot x(\partial_x \partial_y) \\
&= \partial_y + x(\partial_y \partial_x) \\
&= \partial_y + x(\partial_y \vec{e_x}) \\
&= \partial_y + x\frac{\vec{e_x}}{\partial_y} \\
&= \partial_y \\
&= \vec{e_y}
\end{aligned}$$

$$[\vec{u}, \vec{v}] = \vec{u}(\vec{v}) - \vec{v}(\vec{u}) = \vec{e_y} - \vec{0} = \vec{e_y}$$

These four lines don't give us a closed rectangle. Lie bracket is defined like this to computes the difference between these two derivatives. $\vec{e_y}$ is the separation vector.

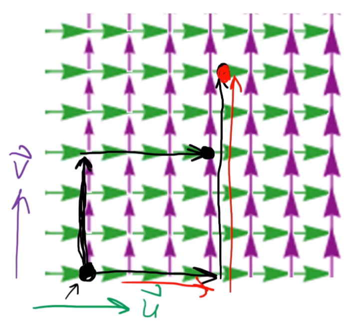


Figure 37: Separation vector

Definition 3.38. $\boxed{\textbf{Lie derivative of tensor fields } T}$ w.r.t. smooth vector field X is defined by

$$(\mathcal{L}_X T)_p = \frac{d}{dt}\bigg|_{t=0} (\phi_t^* T)_p = \lim_{t \to 0} \frac{\phi_t^*(T_{\phi_t(p)}) - T_p}{t},$$

where X is on smooth manifold M, T is the covariant tensor field on M and let $\phi_t(p) = \phi(t,p)$ be a diffeomorphism parameterized by point p and "time" t. X is induced by ϕ

Remark 3.31. *Intuitively, if you have a tensor field T and a vector field X, then $\mathcal{L}_X T$ is the infinitesimal change you would see when you flow T using the vector field $-X$, which is the same thing as the infinitesimal change you would see in T if you flowed along the vector field X.*

Definition 3.39. $\boxed{\textbf{The curvature of a Riemannian manifold}}$ can be described in various ways; the most standard one is the curvature tensor, given in terms of a Levi-Civita connection (or covariant differentiation) ∇ and Lie bracket $[\cdot,\cdot]$ by the following formula:

$$R(\mathbf{u},\mathbf{v})\mathbf{w} = \nabla_\mathbf{u}\nabla_\mathbf{v}\mathbf{w} - \nabla_\mathbf{v}\nabla_\mathbf{u}\mathbf{w} - \nabla_{[\mathbf{u},\mathbf{v}]}\mathbf{w}.$$

Here $R(\mathbf{u},\mathbf{v})$ is a linear transformation of the tangent space of the manifold; it is linear in each argument. If $u=\partial/\partial x^i$ and $v=\partial/\partial x^j$ are coordinate vector fields then $[\mathbf{u},\mathbf{v}]=0$ and therefore the formula simplifies to

$$R(\mathbf{u},\mathbf{v})\mathbf{w}=\nabla_{\mathbf{u}}\nabla_{\mathbf{v}}\mathbf{w}-\nabla_{\mathbf{v}}\nabla_{\mathbf{u}}\mathbf{w}$$

i.e. the curvature tensor measures the non-commutativity of the covariant derivative.

Remark 3.32. *Curvature in Riemannian geometry measures how much the manifold deviates from being flat (like Euclidean space). If the sectional curvature is zero everywhere on the manifold, it implies that the manifold is locally flat in every direction. In other words, if you zoom in closely enough at any point on the manifold, it looks like a piece of Euclidean space. It's important to note that a manifold with zero curvature is not necessarily globally flat.*

Definition 3.40. $\boxed{\textbf{Isometry}}$ $\phi : M \to N$ is a function which preserves distance between manifold M and N. $\forall x\in M$, it has

1. The derivative of ϕ at x is an isomorphism of tangent space $D\phi_x : T_xM \to T_{\phi_x}N$
2. $\forall v,w\in T_xM$, the Riemannian metric preserves as $\langle v,w\rangle=\langle D\phi_x\cdot v, D\phi_x\cdot w\rangle$

Example 42. [sta, 2013b] If S and S' are surfaces with metric g and g', then the surfaces are isometric if there exists $\phi : S\to S'$ such that for all tangent vector $X_p, Y_p\in T_pS$ and all $p\in S$, we have

$$\langle X_p, Y_p\rangle_g=\langle D\phi\cdot X_p, D\phi\cdot Y_p\rangle_{g'}$$

Notice that $D\phi$ is the Jacobian matrix pushes forward tangent vectors from T_pS to $T_{\phi(p)}S'$. We can understand an isometry as preserving the intrinsic geometry at corresponding points.

Definition 3.41. $\boxed{\textbf{Isometry group}}$ G is group of M such that $\forall p,q\in M, g\in G, d(p,q)=d(g\cdot p, g\cdot q)$ holds.

Definition 3.42. $\boxed{\textbf{Conformality}}$ $\phi : S_1\to S_2$ is a function that for all $X,Y\in T_pM$, there exists a function $u : M\to\mathbb{R}$ such that

$$e^{2u(p)}\langle X,Y\rangle_{g_1}=\langle D\phi_p\cdot X, D\phi_p\cdot Y\rangle_{g_2},$$

where g_1, g_2 are the metrics of S_1, S_2 at points $p, \phi(p)$.

Remark 3.33. *[sta, 2013c] Compared to isometries that preserve both lengths and angles, conformality is a weaker condition that preserves only angles. Conformality is very flexible, in fact, all surfaces are locally conformal to the Euclidean metric.*

Definition 3.43. $\boxed{\textbf{Isotropy subgroup}}$ of p is defined as $G_p=\{g\in G|g\cdot p=p\}$. In other words, G_p is the subgroup of G which leaves p fixed.

Definition 3.44. $\boxed{\textbf{Symmetric space}}$ is a connected Riemannian manifold M such that $\forall p\in M$, there is an involutive isometry $\phi_p : M\to M$ that has p as an isolated fixed point. A point $x\in X$ is called a fixed point of ϕ if $\phi(x)=x$.

Definition 3.45. $\boxed{\textbf{Automorphism group}}$ [Singh, 2013] $\Psi : G\to G$ is defined as

$$\Psi_g(h)=ghg^{-1}$$

given $\forall g,h\in G$.

Definition 3.46. $\boxed{\textbf{Inner automorphism group}}$ $\mathrm{Inn}(G)$ is the collection of all inner automorphisms of the form $\Psi_g, \forall g\in G$. $\mathrm{Inn}(G)$ is a Lie group and commutative.

Definition 3.47. $\boxed{\textbf{Dual pairing}}$ (m,v), where $m\in V^*$, the dual space to V, and $v\in V$

Definition 3.48. $\boxed{\textbf{Adjoint action}\ \mathrm{Ad}_g}$ is the derivative of $\Psi_g(h)$ with respect to h at the identity, which is

$$\mathrm{Ad}_g : G\times\mathfrak{g}\to\mathfrak{g}$$
$$\mathrm{Ad}_g=d(\Psi_g)_e$$

- For matrix group, Ad_g is derived as

$$\begin{aligned}
\mathrm{Ad}_g w &= \frac{\partial}{\partial \xi}\Psi_g(w) \\
&= \frac{\partial}{\partial \xi}\Psi_g(h_\xi|_{\xi=0}) \\
&= \frac{\partial}{\partial \xi}\left(g(h_\xi|_{\xi=0})g^{-1}\right) \\
&= g\left(\frac{\partial}{\partial \xi}h_\xi|_{\xi=0}\right)g^{-1} \\
&= gwg^{-1}
\end{aligned}$$

where h_ξ denotes the variation of h by ξ such that $h_0 = e$ and $\frac{\partial}{\partial \xi}h_\xi|_{\xi=0} = w$ for $\Psi_g(h_\xi) = gh_\xi g^{-1}$.

- For conjugation of operators under dual pairing, using $\mathrm{Ad}_g w = gwg^{-1}$, we have

$$\begin{aligned}
(m, \mathrm{Ad}_g w) &= (m, gwg^{-1}) \\
&= (g^\top m, wg^{-1}) \\
&= (g^\top m g^{-\top}, w) \\
\mathrm{Ad}^*_g m &= g^\top m g^{-\top},
\end{aligned}$$

as if A is a linear operator from V to V, its conjugate $A^* : V^* \to V^*$ is defined by $(A^* m, v) = (m, Av)$. For detail, see [Singh, 2013].

- For $\mathrm{Diff}(\Omega)$, Ad_ϕ is derived as

$$\begin{aligned}
\mathrm{Ad}_\phi w &= \frac{\partial}{\partial \xi}(\Psi_\phi h_\xi)|_{\xi=0} \\
&= \frac{\partial}{\partial \xi}(\phi \circ h_\xi \circ \phi^{-1})|_{\xi=0} \\
&= D\phi|_{h_\xi \circ \phi^{-1}} w|_{\phi^{-1}} \\
&= (D\phi \circ \phi^{-1}) w \circ \phi^{-1}
\end{aligned}$$

where h_ξ denotes the variation of h by ξ such that $h_0 = \mathrm{Id}$ and $\frac{\partial}{\partial \xi}h_\xi|_{\xi=0} = w$ for $\Psi_\phi(h_\xi) = \phi \circ h_\xi \circ \phi^{-1}$.

Definition 3.49. $\boxed{\textbf{Infinitesimal adjoint action}\ \mathrm{ad}}$ is the derivative of the adjoint map Ad with respect to g at identity, which is

$$\begin{aligned}
&\mathrm{ad} : \mathfrak{g} \times \mathfrak{g} \to \mathfrak{g} \\
&\mathrm{ad} = d(\mathrm{Ad}_g)_e
\end{aligned}$$

- For matrix group, ad_g is derived as

$$\begin{aligned}
\mathrm{ad}_v w &= \frac{\partial}{\partial \xi}\mathrm{Ad}_{g_\xi} w|_{\xi=0} \\
&= \frac{\partial}{\partial \xi}(g_\xi w g_\xi^{-1})|_{\xi=0} \\
&= \left(\frac{\partial}{\partial \xi}g_\xi w g_\xi^{-1}\right)\Big|_{\xi=0} + \left(g_\xi w \frac{\partial}{\partial \xi}g_\xi^{-1}\right)\Big|_{\xi=0} \\
&= vw - \left(g_\xi w g_\xi^{-1}\frac{\partial}{\partial \xi}g_\xi g_\xi^{-1}\right)\Big|_{\xi=0} \\
&= vw - wv
\end{aligned}$$

where g_ξ is the variation of g by ξ with $g_0 = e$ and $\frac{\partial}{\partial \xi}g_\xi|_{\xi=0} = v$ for $\mathrm{Ad}_{g_\xi} w = g_\xi w g_\xi^{-1}$.

- For conjugation of operators under dual pairing, using $\mathrm{ad}_v w = vw - wv$, we have

$$\begin{aligned}(m, \mathrm{ad}_v w) &= (m, vw - wv) \\ &= (m, vw) - (m, wv) \\ &= (v^\top m, w) - (mv^\top, w) \\ &= (v^\top m - mv^\top, w) \\ \mathrm{ad}_v^* m &= v^\top m - mv^\top,\end{aligned}$$

as if A is a linear operator from V to V, its conjugate $A^* : V^* \to V^*$ is defined by $(A^*m, v) = (m, Av)$. For detail, see [Singh, 2013].

- For $\mathrm{Diff}(\Omega)$, ad_v is derived as

$$\begin{aligned}\mathrm{ad}_v w &= \frac{\partial}{\partial \xi}((D\phi_\xi \circ \phi_\xi^{-1})w \circ \phi_\xi^{-1})|_{\xi=0} \\ &= \frac{\partial}{\partial \xi}((D\phi_\xi \circ \phi_\xi^{-1}))|_{\xi=0} w + D\mathrm{Id}\frac{\partial}{\partial \xi}(w \circ \phi_\xi^{-1})|_{\xi=0} \\ &= \left((\frac{\partial}{\partial \xi} D\phi_\xi | \phi_\xi^{-1} + DD\phi_\xi|_{\xi=0} D\phi_\xi^{-1}) w \circ \phi_\xi^{-1} + Dv|_{\phi_\xi^{-1}} \frac{\partial \phi_\xi^{-1}}{\partial \xi} \right) \Bigg|_{\xi=0} \\ &= (Dv + 0)w - Dwv \\ &= Dvw - Dwv\end{aligned}$$

where ϕ_ξ is the variation of ϕ by ξ with $\phi_0 = e$ and $\frac{\partial}{\partial \xi}\phi_\xi|_{\xi=0} = v$ for $\mathrm{Ad}_\phi w = (D\phi \circ \phi^{-1})w \circ \phi^{-1}$.

Definition 3.50. **Left/Right multiplication** is a diffeomorphism such that

$$\begin{aligned}L_y &: x \to y \cdot x && \text{Left Multiplication} \\ R_y &: x \to x \cdot y && \text{Right Multiplication}\end{aligned}$$

where $y \in G$, G is a Lie group.

Definition 3.51. **Left/Right-invariant** means $\forall y \in G$, we have $L_{y*}X = X$ or $R_{y*}X = X$.

Example 43. The metric G^I has the property of right-invariance: if $U, V \in T_\phi \mathrm{Diff}(M)$ then

$$G^I_\phi(U, V) = G^I_{\phi\circ\psi}(U \circ \psi, V \circ \psi) \qquad \forall \psi \in \mathrm{Diff}(M)$$

Definition 3.52. **Inertia operator** $L : \mathfrak{g} \to \mathfrak{g}^*$ is defined by

$$\langle v, w\rangle = (Lv, w), \forall v, w \in \mathfrak{g}$$

L must be invertible and

$$(Lv, w) = (Lw, v), \forall v, w \in \mathfrak{g}$$

in order to satisfy the properties of a well-formed Riemannian metric.

Definition 3.53. A linear operator $f : \mathfrak{g} \to \mathfrak{g}$ is **transposed** with respect to the inner product defined by L, using the formula

$$\langle f^\dagger v, w\rangle = \langle v, fw\rangle, \forall v, w \in \mathfrak{g}$$

We use this to define the adjoint-transpose action $\mathrm{Ad}^\dagger : G \times \mathfrak{g} \to \mathfrak{g}$ via the transpose of Ad_g and the infinitesimal adjoint-transpose $\mathrm{ad}^\dagger : \mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}$ via the transpose of ad_v

Remark 3.34. *For an operator like a matrix:*

adjoint = (conjugate) transpose

classic adjoint = adjugate

In most linear algebra discussions, "adjoint" refers to the transpose of a matrix.

Proof. To prove the adjugate of a real matrix $\mathbf{A}$ is transpose, we have

$$\begin{aligned}\langle \mathbf{A}f, g\rangle &= (\mathbf{A}f)^\top g = f^\top \mathbf{A}^\top g \\ \langle f, \mathbf{A}^* g\rangle &= \langle f, \mathbf{A}^\top g\rangle = f^\top \mathbf{A}^\top g \\ \Rightarrow \langle \mathbf{A}f, g\rangle &= \langle f, \mathbf{A}^* g\rangle\end{aligned}$$

□

Example 44.

- Adjoint of the gradient is the negative divergence: $\langle \nabla f, g\rangle = \langle f, -\nabla \cdot g\rangle$
- Adjoint of the Fourier transform is its inverse: $\langle \mathcal{F}(f), g\rangle = \langle f, \mathcal{F}^{-1}(g)\rangle$
- Adjoint of the Laplacian is itself: $\langle \Delta f, g\rangle = \langle f, \Delta g\rangle$
- Adjoint of the linear interpolation is the splatting: $\langle f \circ \phi, g\rangle = \langle f, \phi^* g\rangle$

Definition 3.54. $\boxed{\textbf{Information metric}}$ G^I is defined by

$$G^I_\phi(U,V) = -\int_M \langle \Delta u, v\rangle_g \mathrm{vol} + \lambda \sum_{i=1}^k \left(\int_M \langle u, \xi_i\rangle_g \mathrm{vol} \cdot \int_M \langle v, \xi_i\rangle_g \mathrm{vol}\right),$$

where $U = u \circ \phi, V = v \circ \phi, \lambda > 0, \Delta$ is the Laplace-de Rham operator lifted to vector fields, and $\xi_1, \cdots, \xi_k$ is an orthonormal basis of the harmonic 1-form on M.

Definition 3.55. $\boxed{\textbf{Fisher-Rao metric}}$ is the Riemannian metric on $\mathrm{Dens}(M)$ given by

$$G^F_\mu(\alpha, \beta) = \frac{1}{4}\int_M \frac{\alpha}{\mu} \cdot \frac{\beta}{\mu}\mu,$$

for tangent vectors $\alpha, \beta \in T_\mu \mathrm{Dens}(M)$. It can be interpreted as the Hessian of relative entropy, or information divergence.

Definition 3.56. The background metric g on manifold M is called $\boxed{\textbf{compatible with}}$ μ if $\mathrm{vol}_g = \mu$, for $\mu \in \mathrm{Dens}(M)$.

Definition 3.57. $\boxed{\textbf{Set of vector space isomorphisms}}$ $\mathcal{L}_{iso}$ is defined by

$$\mathcal{L}_{iso}(\mathbb{R}^m, V) = \{e : \mathbb{R}^m \to V | e \text{ is a vector space isomorphism}\}.$$

Definition 3.58. $\boxed{\textbf{Frame}}$ of m-dimension real vector space V is the basis $e_1, \cdots, e_m$ of V.

Definition 3.59. $\boxed{\textbf{Frame bundle}}$ $\mathcal{F}$ of a smooth m-dimensional submanifold M is defined by

$$\mathcal{F}(M) = \{(p, e) | p \in M, e \in \mathcal{F}(M)_p\},$$

where $\mathcal{F}(M)_p = \mathcal{L}_{iso}(\mathbb{R}^m, T_p M)$ is the space of frames of tangent space at p.

Definition 3.60. A smooth curve $\beta : \mathbb{R} \to \mathcal{F}(M)$ is called a $\boxed{\textbf{lift}}$ of smooth curve $\gamma : \mathbb{R} \to M$ if [Robbin and Salamon, 2011]

$$\pi \circ \beta = \gamma.$$

Example 45. The general linear group $GL(m, \mathbb{R})$ acts on this space by composition on the right via

$$GL(m) \times \mathcal{L}_{iso}(\mathbb{R}^m, V) \to \mathcal{L}_{iso}(\mathbb{R}^m, V) : (a, e) \to a^* e = e \circ a$$

Definition 3.61. A curve $\beta(t) = (\gamma(t), e(t)) \in \mathcal{F}(M)$ is called $\boxed{\textbf{horizontal lift}}$ of γ if the vector field $X(t) = e(t)\xi$ along γ is parallel for every $\xi \in \mathbb{R}^m$. Thus a horizontal lift of γ has the form

$$\beta(t) = (\gamma(t), \Phi_\gamma(t, 0)e)$$

for some $e \in \mathcal{L}_{iso}(\mathbb{R}^m, T_{\gamma(0)}M)$.

Definition 3.62. Suppose $\phi : M \to N$ is a differential map from M to N, $\gamma : (-\epsilon, \epsilon) \to M$ is a curve on M, $\gamma(0) = p, \gamma'(0) = v \in T_pM$, then $\phi \circ \gamma$ is a curve on N, $\phi \circ \gamma(0) = \phi(p)$, we define the tangent vector

$$\phi_*(v) = (\phi \circ \gamma)'(0) \in T_{\phi(p)}N,$$

as the $\boxed{\textbf{pushforward}}$ tangent vector of v induced by ϕ.

Definition 3.63. $\boxed{\textbf{Pullback and pushforward}}$ are defined as below

$$\begin{aligned}
\text{Pullback(right action): } \varphi^*\rho &= |D\varphi|\rho(\varphi(\cdot)) \\
&= |D\varphi|\rho \circ \varphi \\
\text{Pushforward(left action): } \varphi_*\rho &= (\varphi^{-1})^*\rho \\
&= |D\varphi^{-1}|\rho(\varphi^{-1}(\cdot)) \\
&= |D\varphi^{-1}|\rho \circ \varphi^{-1}
\end{aligned}$$

where $(\varphi, \rho) \in \text{Diff}(M) \times \text{Dens}(M)$.

Remark 3.35. *Given a ϕ, there can be two effects or two directions of the ϕ. The pullback is the one from distorted to checkered, while the pushforward is the one from checkered to distorted, which is more intuitive.*

Remark 3.36. *The scaling factor of $|D\phi|$ in pullback or pushforward actually tells us the intensity change after the diffeomorphism action. For density matching, if the volume is squeezed, the intensity would increase, which is reflected in the value of Jacobian determinant.*

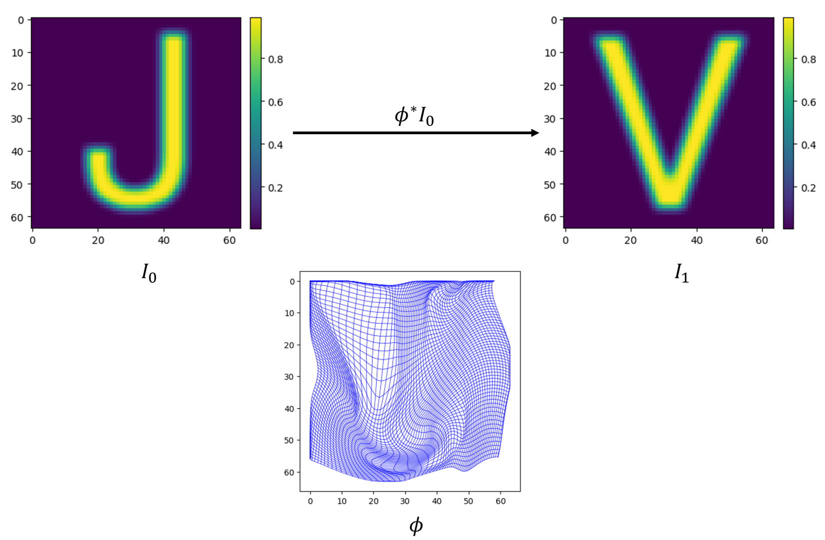


Figure 38: Pullback

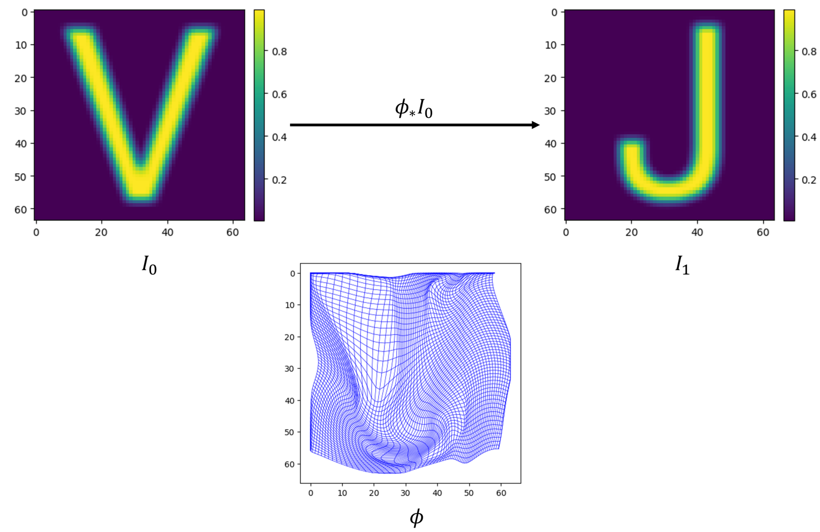


Figure 39: Pushforward

Remark 3.37. ***Relationship between*** $I_0, I_1, \phi, \phi^{-1}$***:*** *Intuitively, we may think that ϕ demonstrate the right way to distort $I_0 : V$ to $I_1 : J$, like distorting the "painted flat tablecloth". In a sense, that's right, if we write it as $I_1 = \phi_* I_0$. However, we always use compose operation to distort the density map, which is $|D\phi^{-1}| I_0 \circ \phi^{-1}$, so when we are using the composing, always remember that we are composing ϕ^{-1}, instead of the more intuitively-correct ϕ.*

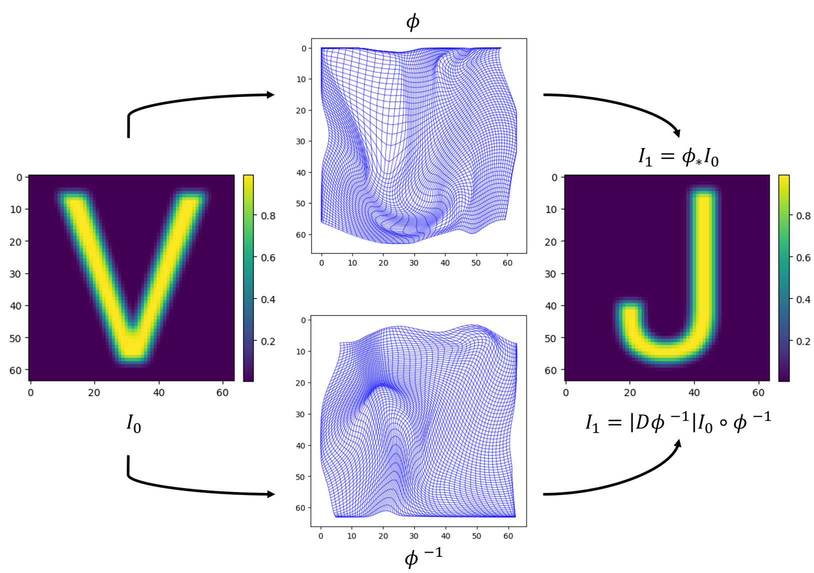


Figure 40: Pushforward

Remark 3.38. ***"Blank Space"*** *You may have the confusion that if we deform the image using the diffeomorphism φ in Fig.(31), namely $\phi_* I$, what to fill in the blank triangle area at the bottom, since I analogize the diffeomorphism to distorting the "painted flat tablecloth".*

How about we think this way, by implementing the algorithm through Python discretely, we can have $I \circ \varphi^{-1} : \mathbb{Z}^2 \to \mathbb{R}^2 \to \mathbb{R}$. The space of $\mathbb{Z}^2$ consists of all the integer tuple among $(0 : M, 0 : N)$, where M, N are the height and *width of I. The space of $\mathbb{R}^2$ can be arbitrary real number tuple, but still in the range of $(0:M, 0:N)$, as I is not defined beyond this range. For simplicity, $I \circ \varphi^{-1}: \mathbb{Z}^2 \to \mathbb{R}$ is still a function defined all over the $(0:M, 0:N)$, which won't cause the blank space.*

The essential cause of the above phenomena is that, due to the limitation in computer, we typically store the diffeomorphism in the data structure of array, which means $\phi: \mathbb{Z}^2 \to \mathbb{R}^2$, where $\mathbb{Z}^2$ is consists of all the possible tuple among $(0:M, 0:N)$, that's how we index the entries in the array. So we should never worry about the blank space, as how we plot the diffeomorphism is different from how to express it in algorithm.

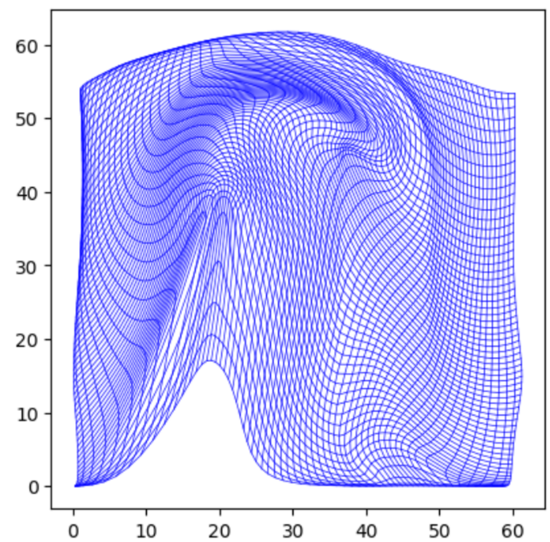


Figure 41: φ

4 Statistics for Images

Definition 4.1. $\boxed{\textbf{Principal geodesic analysis(PGA)}}$ seeks a sequence of geodesic submanifolds that maximize the variance of the data. These submanifolds are called the principal geodesic submanifolds.

- **Definition 1.** The principal geodesic submanifolds are defined by first constructing an orthonormal basis $e_1, \cdots, e_d$ of $T_\mu M$. Then, these vectors are used to form a sequence of nested subspaces $V_k = \mathrm{span}(\{e_1, \cdots, e_k\}) \cap U$, where $U \subset T_\mu M$ is a neighbourhood of 0, such that projection is well-defined for all geodesic submanifolds of $\mathrm{Exp}_\mu(U)$.

  The principal geodesic submanifolds are given by
  $$H_k = \mathrm{Exp}_\mu(V_k)$$
  The first principal direction is now chosen to maximize the projected variance along the corresponding geodesic:
  $$\begin{aligned} e_1 &= \arg\max_{\|e\|=1} \sum_{i=1}^n \|\mathrm{Log}_\mu(\pi_H(x_i))\|^2, \qquad \text{where } H = \mathrm{Exp}_\mu(\mathrm{span}(\{e\}) \cap U) \\ e_k &= \arg\max_{\|e\|=1} \sum_{i=1}^n \|\mathrm{Log}_\mu(\pi_H(x_i))\|^2, \qquad \text{where } H = \mathrm{Exp}_\mu(\mathrm{span}(\{e_1, \cdots, e_{k-1}, e\}) \cap U) \end{aligned}$$

  where we define the **projection operator** $\pi_H : M \to H$ as
  $$\begin{aligned} \pi_H(x) &= \arg\min_{y \in H} d(x,y)^2 \\ &= \arg\min_{y \in H} \|\mathrm{Log}_x(y)\|^2 \\ &\approx \arg\min_{y \in H} \|\mathrm{Log}_p(x) - \mathrm{Log}_p(y)\|^2 \end{aligned}$$

- **Definition 2.** The intent of principal geodesic analysis is to find an orthonormal basis $\{e_1, \cdots, e_k\}$ of a set of points $\{x_1, \cdots, x_n\} \in \mathbb{R}^d$, which satisfies the recursive relationship
  $$\begin{aligned} e_1 &= \arg\min_{\|e\|=1} \sum_{i=1}^n \|x_i - \langle e, x_i \rangle e\|^2 \\ e_2 &= \arg\min_{\|e\|=1} \sum_{i=1}^n \|x_i - \langle e_1, x_i \rangle e_1 - \langle e, x_i \rangle e\|^2 \\ &\vdots \\ e_k &= \arg\min_{\|e\|=1} \sum_{i=1}^n \|x_i - \langle e_1, x_i \rangle e_1 - \cdots - \langle e_{k-1}, x_i \rangle e_{k-1} - \langle e, x_i \rangle e\|^2 \end{aligned}$$
  In other words, the subspace $V_k = \mathrm{span}(\{e_1, \cdots, e_k\})$ is the k-dimensional subspace that minimizes the sum-of-squared distances to the data. The principal geodesic submanifolds are given by
  $$H_k = \mathrm{Exp}_\mu(V_k)$$
  The first principal direction is now chosen to minimize the sum-of-squared distance of the data to the corresponding geodesic:
  $$\begin{aligned} e_1 &= \arg\max_{\|e\|=1} \sum_{i=1}^n \|\mathrm{Log}_{x_i}(\pi_H(x_i))\|^2, \qquad \text{where } H = \mathrm{Exp}_\mu(\mathrm{span}(\{e\}) \cap U) \\ e_k &= \arg\max_{\|e\|=1} \sum_{i=1}^n \|\mathrm{Log}_{x_i}(\pi_H(x_i))\|^2, \qquad \text{where } H = \mathrm{Exp}_\mu(\mathrm{span}(\{e_1, \cdots, e_{k-1}, e\}) \cap U) \end{aligned}$$

Example 46.

Algorithm 1 Principal Geodesic Analysis

Inputs: $x_1, \cdots, x_n \in M$
$\mu \leftarrow$ intrinsic mean of $\{x_i\}$ ▷ Remark 4.6
$u_i \leftarrow \mathrm{Log}_\mu(x_i)$ ▷ Map $x_i \in M$ to $T_\mu M$ as u_i
$\Sigma \leftarrow \frac{1}{n}\sum_{i=1}^n u_i u_i^\top$
$\{e_k, \lambda_k\} \leftarrow$ eigenvectors and eigenvalues of Σ
return $\{e_k, \lambda_k\}$

Remark 4.1.
- *The sample variance of the data is the expected value of the squared Riemannian distance from the mean.*
- *For data in $\mathbb{R}^n$ the two definitions are equivalent since PGA reduces to PCA in the linear case.*

Definition 4.2. $\boxed{\textbf{Atlas}}$ is the image of the average anatomy of a collection of anatomical images.

Remark 4.2. *Motivation of atlas building:*

- *Map population into common coordinate space;*
- *Learn about variability of brain anatomy;*
- *Describe difference from normal;*
- *Use as normative atlas for segmentation.*

Example 47. Atlas-based segmentation:

1. Build atlas with segmented structures;
2. Transform the atlas to case;
3. Make decision voxel-wise according to atlas.

Remark 4.3. *The space of diffeomorphisms is not a vector space, despite the linear average $\mu = \frac{1}{N}\sum_{i=1}^N x_i$, we need a more general notion of average, and here comes the Fréchet mean.*

Definition 4.3. $\boxed{\textbf{Fréchet mean and variance}}$ are defined as

$$p_* = \arg\min_p \phi(p), \qquad \triangleright \text{ Fréchet mean}$$

$$\phi(p) = \sum_{i=1}^N d^2(p, x_i) w_i, \qquad \triangleright \text{ Fréchet variance}$$

where (M, d) is a complete metric space, $x_1, x_2, \cdots, x_n$ are the random points in M and $p \in M$.

Remark 4.4. *The* ***Karcher*** *means are then those points, p of M, which locally minimize ϕ, while the* ***Fréchet*** *means are then those points, which globally minimize ϕ.*

Remark 4.5. *Actually, the Fréchet mean is the generalization of the arithmetic mean, median, geometric mean, and harmonic mean, by using different distance functions.*

- *For real numbers, the* ***arithmetic*** *mean is a Fréchet mean, by using the usual Euclidean distance as the distance function.*
- *For positive real numbers, the* ***geometric*** *mean is a Fréchet mean, by using the (hyperbolic) distance function $d(x, y) = |\log(x) - \log(y)|$*

- *For positive real numbers, the* ***harmonic*** *mean is a Fréchet mean, by using the distance function* $d(x,y) = |\frac{1}{x} - \frac{1}{y}|$

Remark 4.6. ***How to compute the intrinsic mean of manifold data:*** *According to 4.1.2 in [Fletcher, 2004], Karcher[Karcher, 1977] shows that the gradient of* ϕ *above is*

$$\nabla\phi(p) = -2\sum_{i=1}^{N} \mathrm{Log}_p(x_i) w_i.$$

If $w_i = \frac{1}{2N}$*, then we have*

$$\phi(p) = \frac{1}{2N}\sum_{i=1}^{N} d^2(p, x_i),$$
$$\nabla\phi(p) = -\frac{1}{N}\sum_{i=1}^{N} \mathrm{Log}_p(x_i).$$

The gradient descent algorithm takes successive steps in the negative gradient direction. Given a current estimate μ_j*, say* $\mu_1 = x_1$*, as the intrinsic mean, the equation for updating the mean by taking a step in the negative gradient direction is*

$$\mu_{j+1} = \mathrm{Exp}_{\mu_j}\left(\frac{\tau}{N}\sum_{i=1}^{N} \mathrm{Log}_{\mu_j}(x_i)\right),$$

where τ *is the step size. And this updating equation is easy to understand: Log map all the* $x_i \in M$ *to the tangent space of* μ_j*, after derived the mean in the tangent space, exp map this mean back to the* M*, namely the final intrinsic mean we want.*

Definition 4.4. $\boxed{\textbf{Fréchet median}}$ is defined as

$$p_* = \arg\min_p \psi(p), \qquad \triangleright \text{ Fréchet median}$$
$$\psi(p) = \sum_{i=1}^{N} |d^2(p, x_i)|,$$

where $x_1, x_2, \cdots, x_n$ are the random points in M and $p \in M$.

Definition 4.5. $\boxed{\textbf{Weighted geometric median}}$ is defined as

$$p_* = \arg\min_p \psi(p),$$
$$\psi(p) = \sum_{i=1}^{N} d^2(p, x_i) w_i,$$

where $x_1, x_2, \cdots, x_n$ are the random points in M and $p \in M$.

Definition 4.6. $\boxed{\textbf{Fréchet distance}}$ between A and B is defined as infimum over all reparameterizations α and β of the maximum over all $t \in [0,1]$ of the distance between $A(\alpha(t))$ and $B(\beta(t))$.

$$F(A,B) = \inf_{\alpha,\beta} \max_{t\in[0,1]} \left\{ d\left(A(\alpha(t)), B(\beta(t))\right)\right\}$$

where d is the distance function, e.g. Euclidean distance.

Example 48. Informally, we can think of the t in parameterization as "time". In the discrete situation, we can think of $\alpha(t), \beta(t)$ as two sequences of same size. Say integer index $t \in [0, 100)$, so the sequences $\alpha(\cdot) : \mathbb{Z}^1 \to \mathbb{Z}^1, \beta(\cdot) : \mathbb{Z}^1 \to \mathbb{Z}^1$ can be presented as follows:

$$\alpha(\cdot) = [\alpha(0), \alpha(1), \alpha(2), \cdots, \alpha(99)],$$
$$\beta(\cdot) = [\beta(0), \beta(1), \beta(2), \cdots, \beta(99)],$$

where $\alpha(0) \leq \alpha(1) \leq \alpha(2) \leq \cdots \leq \alpha(99)$ and $\beta(0) \leq \beta(1) \leq \beta(2) \leq \cdots \leq \beta(99)$. We can view the $\alpha(\cdot), \beta(\cdot)$ as two displacement-time graphs in the way of sequence, which tells you the moving points move forward in what manner.

$A(\cdot) : \mathbb{Z}^1 \to \mathbb{R}^n, B(\cdot) : \mathbb{Z}^1 \to \mathbb{R}^n$ are also two sequences of same size(not necessary the same size as α, β), which can be exhibited as follows:

$$
\begin{aligned}
A(\cdot) &= [A(0), A(1), A(2), \cdots], \\
B(\cdot) &= [B(0), B(1), B(2), \cdots],
\end{aligned}
$$

and $A(i)$ and $B(i)$ are the coordinates in the metric space.

So for calculating the Fréchet distance between the curves, the key is to find all the possible reparameterizations α, β. And for each combination of α, β, we can then find the maximum distance across the whole "time period", followed by finding the infimum among all the maximum distance under each combination.

Remark 4.7. *In a nutshell, the Fréchet distance between two given fixed paths can be found in this way: we want to find moving patterns of the two points on two paths that the maximun distance during this travel is minimized. The satisfied moving patterns are often making the two points moving forward "simultenously".*

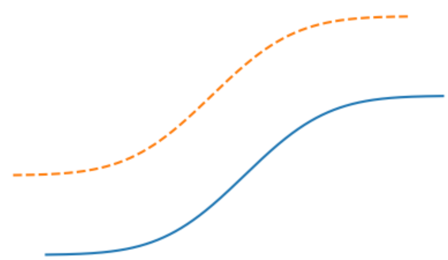


Figure 42: The Fréchet distance between two same shape curves is the norm of the translation vector.

Definition 4.7. $\boxed{\textbf{Wasserstein distance}}$ is defined as

$$
\begin{aligned}
W_p(\mu, \nu) &= \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{M \times M} \operatorname{dist}(x, y)^p \, \mathrm{d}\gamma(x, y) \right)^{1/p}, \\
&= \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{M \times M} \operatorname{dist}(x, y)^p \gamma(x, y) \mathrm{d}x \mathrm{d}y \right)^{1/p}, \\
W_2(\mu, \nu) &= \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{M \times M} \operatorname{dist}(x, y)^2 \gamma(x, y) \mathrm{d}x \mathrm{d}y \right)^{1/2},
\end{aligned}
$$

where $\Gamma(\mu, \nu)$ denotes the set of all coupling of marginal distribution μ and ν, and x, y are actually indicating the position in the respective distribution.

In discrete case, the distance reads as

$$
W_p(\mu, \nu) = \min_{T \in \Gamma(\mu,\nu)} \langle T, M_{\mu\nu} \rangle = \min_{T \in \Gamma(\mu,\nu)} \operatorname{tr}(T^\top M_{\mu\nu}),
$$

where $M_{\mu\nu} = [\operatorname{dist}(x_i, y_j)^p]_{ij} \in \mathbb{R}^{m \times n}$ and $\Gamma(\mu, \nu) = \{T \in \mathbb{R}_+^{m \times n} | T\mathbb{1}_m = a, T^\top \mathbb{1}_n = b\}$, namely the transport map matrix T is under the constraint that the sum of each row and column equals to a and b, respectively. The two matrices correspond to $\operatorname{dist}(x, y)$ and $\gamma(x, y)$ in continuous form.

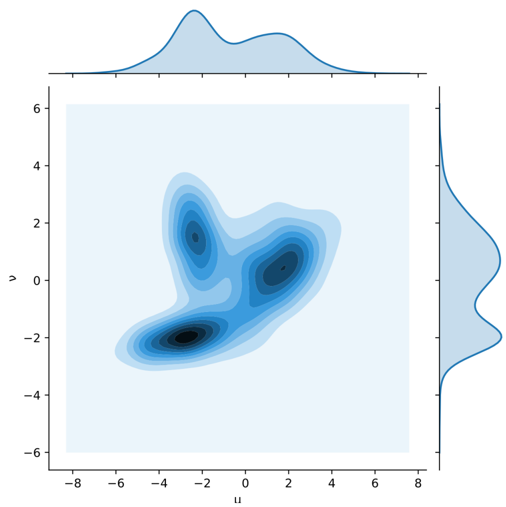


Figure 43: Wasserstein distance, credit to Wikipedia

Remark 4.8. *Wasserstein distance is closely related to optimal transport problem. That is, for two distributions of mass $\mu(x), \nu(y)$ in the space S, where $x, y \in S$, we wish to transport the mass at the lowest cost. The problem only makes sense when the sums of two distributions are identical, fortunately $\mu(x), \nu(y)$ are two probability distributions, namely the sum of mass equals to 1, this premise will be satisfied.*

Assuming there is also a cost function $c(x, y) \to [0, +\infty)$ which indicates the cost of transporting ***unit mass*** *from point x to point y. Function $\gamma(x, y)$ depicts a transport plan which gives the amount of mass moved from point x to point y. Therefore, the cost of the whole transport plan equals to*

$$\int\int c(x, y)\gamma(x, y)\, dxdy,$$

and the Wasserstein distance is exactly the cost of optimal transport plan.

Lemma 4.1. *The p-Wasserstein distance between the two probability measures μ and ν on $\mathbb{R}^1$ has the following closed-form expression:*

$$\begin{aligned} W_p(\mu, \nu) &= \left(\int_{-\infty}^{+\infty} |U(s) - V(s)|^p ds\right)^{1/p} && \triangleright \int \textit{quantity} \times \textit{unit distance} \\ &= \left(\int_0^1 |U^{-1}(t) - V^{-1}(t)|^p dt\right)^{1/p}, && \triangleright \int \textit{distance} \times \textit{unit quantity} \end{aligned}$$

where U and V are the CDFs of μ, ν respectively.

Proof. Assuming $\{(x_1, y_1), (x_2, y_2)\} \subset (\gamma^*)^6, x_1 < x_2$, where γ^* denotes the optimal transport plan. Given the previous assumption, we claim $y_1 \leq y_2$.

Supposing that $y_1 \leq y_2$ is not the case, namely $y_1 > y_2$, which yields[7]

$$|x_1 - y_2|^p + |x_2 - y_1|^p < |x_1 - y_1|^p + |x_2 - y_2|^p$$

However this inequality suggests that $\{(x_1, y_2), (x_2, y_1)\} \subset (\gamma^*)$, rather than $\{(x_1, y_1), (x_2, y_2)\} \subset (\gamma^*)$, which contradicts the initial assumption, namely the optimality of γ^*, as it indicates that γ^* is no cyclically monotone.

[6]The support of a probability distribution can be loosely thought of as the closure of the set of possible values of a random variable having that distribution.

[7]For a more detailed deviation of this inequality in the case of $p > 1$, refers to Appendix A in `https://arxiv.org/pdf/1509.02237.pdf`

Now, for $x \in (\mu), y \in (\nu)$, we claim that $(x, y) \in (\gamma^*)$ if and only if $U(x) = V(y)$. To see this, note that form the monotonicity property we just built, we deduce that $(x, y) \in (\gamma^*)$ if and only if

$$\gamma^*(\mathbb{R}, (-\infty, y]) = \gamma^*((-\infty, x], (-\infty, y]) = \gamma^*((-\infty, x], \mathbb{R})$$

In turn, the fact that $\gamma^* \in \Gamma(\mu, \nu)$ implies that $\gamma^*((-\infty, x], \mathbb{R}) = F(x)$ and $\gamma^*(\mathbb{R}, (-\infty, y]) = G(y)$. From previous relation, we conclude that

$$W_p(\mu, \nu) = \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{M \times M} \mathrm{dist}(x, y)^p \gamma(x, y) \mathrm{d}x\mathrm{d}y \right)^{1/p} = \left(\int_0^1 |F^{-1}(t) - G^{-1}(t)|^p dt \right)^{1/p}$$

$\square$

Example 49. For one dimensional discrete case, to transport ν to μ,

1. 4 extra squares would be moved from 0 to 1;
2. 3 extra squares would be moved from 1 to 2;
3. 2 extra squares would be moved from 2 to 3;
4. 1 extra squares would be moved from 3 to 4.

The "earth" need to be moved is exactly the difference between the two CDFs at each location. Therefore the p-Wasserstein distance equals to $(\sum |U(s) - V(s)|^p ds)^{1/p} = (4^p \times 1 + 3^p \times 1 + 2^p \times 1 + 1^p \times 1)^{1/p}$

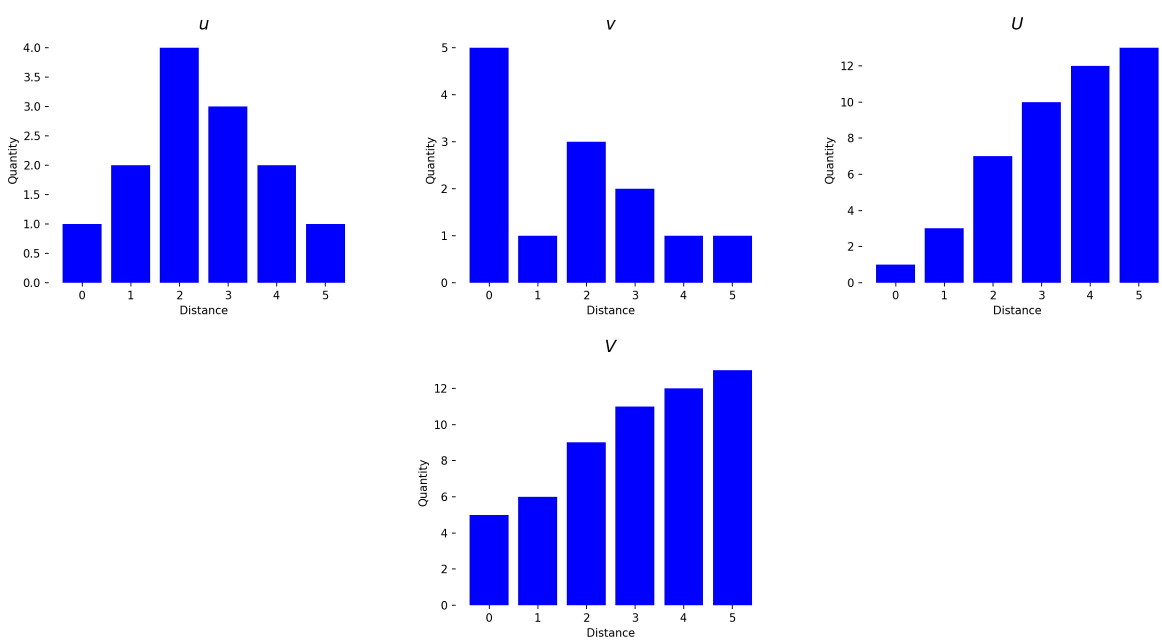


Figure 44: Two distribution μ and ν.

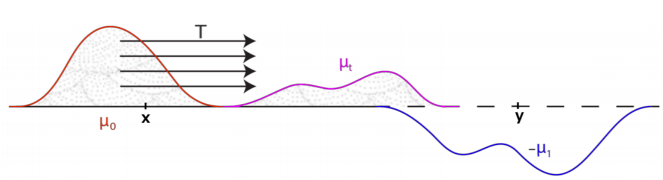


Figure 45: Optimal transport measures the minimal effort required for filling $-\mu_1$ with μ_0, i.e. transporting one distribution to another.

Remark 4.9. *Relationship between **KL divergence**, **JS divergence** and **Wasserstein distance**:*

- *Intuitively, KL divergence looks like a distance between two distributions, however $D_{KL}(p,q) \neq D_{KL}(q,p)$, namely it is asymmetric. So comes the JS divergence.*
- *When the two distributions are far apart, the KL divergence cannot reflect the distance between the distributions while JS divergence is constant, which is deadly for backpropagation in neural network. Nevertheless, the Wasserstein distance can tackle this drawback of KL/JS divergence, as the optimal transport plan of two distant distributions would always make sense and variable.*

Remark 4.10. *Advantages of Wasserstein distance:*[8]

- *By leveraging Wasserstein distance, we can get a better average/summary image of two distribution.*

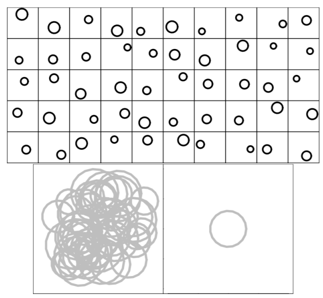


Figure 46: Top: Some random circles. Bottom left: Euclidean average of the circles. Bottom right: Wasserstein barycenter.

- *When we are creating a geodesic between two distributions P_0, P_1, and P_t interpolates between them, Wasserstein distance can preserve the basic structure of the distribution.*

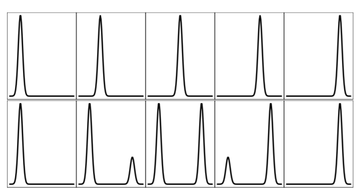


Figure 47: Top row: Geodesic path from P_0 to P_1. Bottom row: Euclidean path($P_t = tP_0 + (1-t)P_1$) from P_0 to P_1.

- *Wasserstein distance is insensitive to small wiggles.*

Definition 4.8. $\boxed{\textbf{Wasserstein barycenter}}$ is defined as

$$\mu^* = \arg\min_{\mu} \sum_i W_2^2(\mu, \mu_i)$$

where W_2 is the L^2 Wasserstein distance and μ_i is the sample distribution. For computation in discrete situation, please refer to `https://arxiv.org/pdf/1310.4375.pdf`.

Remark 4.11. *Euclidean averaging does not contain geometric information. Barycenters under the Wasserstein distance are more intuitive.*[9]

[8]Figures in this remark credit to `https://www.stat.cmu.edu/~larry/=sml/Opt.pdf`

[9]`https://people.csail.mit.edu/eddchien/presentations/Stochastic_Wasserstein_Barycenters.pdf`

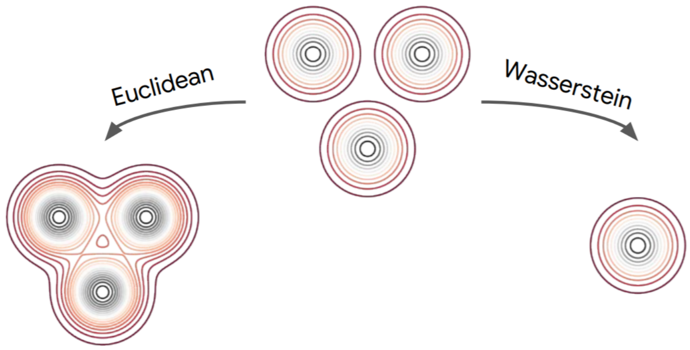


Figure 48: Euclidean mean vs. Wasserstein barycenter of distributions.

Corollary 4.1. *Given* $\Delta f = g$, $f(x,y) = \mathcal{F}^{-1}\left[\frac{1}{2\cos(2\pi u/M)+2\cos(2\pi v/N)-4}G(u,v)\right]$.

Proof.

$$\begin{aligned}
f(x+1,y)+f(x-1,y)+f(x,y+1)+f(x,y-1)-4f(x,y) &= g(x,y)\\
\mathcal{F}[f(x+1,y)+f(x-1,y)+f(x,y+1)+f(x,y-1)-4f(x,y)] &= \mathcal{F}[g(x,y)]\\
e^{-2\pi ju/M}F(u,v)+e^{2\pi ju/M}F(u,v)+e^{-2\pi jv/N}F(u,v)+e^{2\pi jv/N}F(u,v)-4F(u,v) &= G(u,v)\\
\triangleright\text{time shifting[shi,]: } \mathcal{F}(f(x+a,y+b)) &= e^{-2\pi jau/M-2\pi jbv/N}F(u,v)
\end{aligned}$$

After withdrawing the common factor, we have

$$\begin{aligned}
F(u,v) &= \frac{1}{e^{-2\pi ju/M}+e^{2\pi ju/M}+e^{-2\pi jv/N}+e^{2\pi jv/N}-4}G(u,v)\\
f(x,y) &= \mathcal{F}^{-1}\left[\frac{1}{e^{-2\pi ju/M}+e^{2\pi ju/M}+e^{-2\pi jv/N}+e^{2\pi jv/N}-4}G(u,v)\right]
\end{aligned}$$

As $e^{ju} = \cos(u) + j\sin(u)$, we can write $f(x,y)$ in this way:

$$f(x,y) = \mathcal{F}^{-1}\left[\frac{1}{2\cos(2\pi u/M)+2\cos(2\pi v/N)-4}G(u,v)\right]$$

Likewise, you can find the 3D one like below:

$$f(x,y,z) = \mathcal{F}^{-1}\left[\frac{1}{2\cos(2\pi u/M)+2\cos(2\pi v/N)+2\cos(2\pi w/O)-6}G(u,v,w)\right]$$

□

Remark 4.12. *The formula below defines MATLAB's discrete Fourier transform G of an* $m\times n$ *matrix g:*

$$G(u,v) = \sum_{x=1}^{m}\sum_{y=1}^{n} w_m^{(x-1)(u-1)} w_n^{(y-1)(v-1)} g(x,y),$$

where $w_m = e^{-2\pi i/m}, w_n = e^{-2\pi i/n}$ *and* $x,u\in[1,m], y,v\in[1,n]$.

In a more specific way, we have

$$G(u,v) = \sum_{x=1}^{m}\sum_{y=1}^{n} e^{-2\pi i\left(\frac{(x-1)(u-1)}{m}+\frac{(y-1)(v-1)}{n}\right)} g(x,y)$$

5 Linear Algebra for Images

5.1 Geometric Transformations

Transformation	Matrix	#DoF	Preserves
Translation	$\begin{pmatrix} 1 & 0 & t_1 \\ 0 & 1 & t_2 \\ 0 & 0 & 1 \end{pmatrix}$	2	orientation
Scaling	$\begin{pmatrix} s_1 & 0 & 0 \\ 0 & s_2 & 0 \\ 0 & 0 & 1 \end{pmatrix}$	2	orientation
Shearing	$\begin{pmatrix} 1 & s_1 & 0 \\ s_2 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$	2	orientation
Rotation	$\begin{pmatrix} \cos(\theta) & -\sin(\theta) & 0 \\ \sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{pmatrix}$	1	lengths
Affine	$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ 0 & 0 & 1 \end{pmatrix}$	6	parallelism

Remark 5.1. *The transformations can be combined by matrix multiplication, of which the order matters,*

$$\begin{aligned} &\begin{pmatrix} 1 & 0 & t_1 \\ 0 & 1 & t_2 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} \cos(\theta) & -\sin(\theta) & 0 \\ \sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} s_1 & 0 & 0 \\ 0 & s_2 & 0 \\ 0 & 0 & 1 \end{pmatrix} \\ =& \begin{pmatrix} \cos(\theta) & -\sin(\theta) & t_1 \\ \sin(\theta) & \cos(\theta) & t_2 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} s_1 & 0 & 0 \\ 0 & s_2 & 0 \\ 0 & 0 & 1 \end{pmatrix} \\ =& \begin{pmatrix} s_1\cos(\theta) & -\sin(\theta) & t_1 \\ \sin(\theta) & s_2\cos(\theta) & t_2 \\ 0 & 0 & 1 \end{pmatrix} \end{aligned}$$

especially when there is a translation operation.

$$\begin{aligned} &\begin{pmatrix} \cos(\theta) & -\sin(\theta) & 0 \\ \sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} s_1 & 0 & 0 \\ 0 & s_2 & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & t_1 \\ 0 & 1 & t_2 \\ 0 & 0 & 1 \end{pmatrix} \\ =& \begin{pmatrix} s_1\cos(\theta) & -\sin(\theta) & 0 \\ \sin(\theta) & s_2\cos(\theta) & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & t_1 \\ 0 & 1 & t_2 \\ 0 & 0 & 1 \end{pmatrix} \\ =& \begin{pmatrix} s_1\cos(\theta) & -\sin(\theta) & s_1t_1\cos(\theta) - t_2\sin(\theta) \\ \sin(\theta) & s_2\cos(\theta) & t_1\sin(\theta) + s_2t_2\cos(\theta) \\ 0 & 0 & 1 \end{pmatrix} \end{aligned}$$

Definition 5.1. **Affine transformation** is a geometric transformation of a Euclidean space that preserves lines and parallelism (but not necessarily distances and angles).

Remark 5.2. *Affine transformations include scaling, rotation, translation, shear mapping, reflection, or compositions of them in any combination and sequence.*

Definition 5.2. $\boxed{\textbf{Rigid transformation}}$ is a geometric transformation of a Euclidean space that preserves the Euclidean distance between every pair of points.

Remark 5.3. *Rigid transformations include rotation, translation, reflection, or compositions of them in any combination and sequence.*

Remark 5.4. *Any rigid transformation T can be decomposed into translation and rotation.*

5.2 Matrix Derivative

- For $f(x) : \mathbb{R}^n \to \mathbb{R}, f(x)' \in \mathbb{R}^n$, we can have:

$$\begin{aligned}
\frac{\partial}{\partial x}(b^\top x) = \frac{\partial}{\partial x}(x^\top b) = b \\
\frac{\partial}{\partial x}(x^\top x) = 2x \\
\frac{\partial}{\partial x}(x^\top Ax) = 2A \\
\frac{\partial}{\partial x}\|Ax - b\|_2^2 = 2A^\top(Ax - b) \\
\frac{\partial}{\partial x}\|Ax - b\|_2 = \frac{A^\top(Ax - b)}{\|Ax - b\|_2}
\end{aligned}$$

- For $f(X) : \mathbb{R}^{n\times m} \to \mathbb{R}, f(X)' \in \mathbb{R}^{n\times m}$, we can have:

$$\begin{aligned}
\frac{\partial}{\partial X}(a^\top Xb) = ab^\top \\
\frac{\partial}{\partial X}(a^\top X^\top b) = ba^\top \\
\frac{\partial}{\partial X}\operatorname{tr}(A^\top XB) = AB^\top \\
\frac{\partial}{\partial X}\operatorname{tr}(A^\top X^\top B) = BA^\top \\
\frac{\partial}{\partial X}|X| = |X|(X^{-1})^\top
\end{aligned}$$

References

[shi,] Properties of the dft. PDF file.

[rg,] Riemannian geometry: Definitions, pictures, and results. PDF file.

[sta, 2013a] (2013a). Lecture 10: Computing geodesics. PDF file.

[sta, 2013b] (2013b). Lecture 15: Isometries, rigidity and curvature. PDF file.

[sta, 2013c] (2013c). Lecture 19: Conformal geometry. PDF file.

[sta, 2013d] (2013d). Lecture 4 note: Surface theory i. PDF file.

[ucl, 2019] (2019). Introduction to rkhs, and some simple kernel algorithms. PDF file.

[Arvanitidis et al., 2016] Arvanitidis, G., Hansen, L. K., and Hauberg, S. r. (2016). A locally adaptive normal distribution. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

[Bhatia, 2009] Bhatia, R. (2009). *Positive definite matrices*, volume 24. Princeton University Press.

[Chaudhuri, 2004] Chaudhuri, P. K. (2004). Effect of translation & rotation on the fourier transform.

[Fletcher, 2004] Fletcher, P. (2004). *Statistical Variability in Nonlinear Spaces: Application to Shape Analysis and DT-MRI*. PhD thesis, Department of Computer Science, University of North Carolina.

[Fletcher et al., 2004a] Fletcher, P., Pizer, S., and Joshi, S. (2004a). Statistical variability in nonlinear spaces: Application to shape analysis and dt-mri. In *University of North Carolina at Chapel Hill*, pages 6469–6469.

[Fletcher et al., 2004b] Fletcher, P. T., Lu, C., Pizer, S. M., and Joshi, S. (2004b). Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE transactions on medical imaging*, 23(8):995–1005.

[Hao, 2014] Hao, X. (2014). *Improved segmentation and analysis of white matter tracts based on adaptive geodesic tracking*. PhD thesis, The University of Utah.

[Karcher, 1977] Karcher, H. (1977). Riemannian center of mass and mollifier smoothing. *Communications on pure and applied mathematics*, 30(5):509–541.

[Luenberger, 1997] Luenberger, D. (1997). *Optimization by vector space methods*. John Wiley & Sons.

[Robbin and Salamon, 2011] Robbin, J. and Salamon, D. (2011). Introduction to differential geometry. Lecture notes, ETH. Preliminary version.

[Singh, 2013] Singh, N. (2013). *Multivariate regression of shapes via deformation momenta: Application to quantifying brain atrophy in aging and dementia*. PhD thesis, The University of Utah.

[Suárez et al., 2021] Suárez, J. L., García, S., and Herrera, F. (2021). A tutorial on distance metric learning: Mathematical foundations, algorithms, experimental analysis, prospects and challenges. *Neurocomputing*, 425:300–322.

[Yger et al., 2016] Yger, F., Berar, M., and Lotte, F. (2016). Riemannian approaches in brain-computer interfaces: a review. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(10):1753–1762.

[YouTube, 2018a] YouTube (2018a). Tensor calculus 17: The covariant derivative (flat space).

[YouTube, 2018b] YouTube (2018b). Tensor calculus 18: Covariant derivative (extrinsic) and parallel transport.

[YouTube, 2018c] YouTube (2018c). Tensor calculus 19: Covariant derivative (intrinsic) and geodesics.

[YouTube, 2018d] YouTube (2018d). Tensors for beginners 9: The metric tensor.

[Zhang, 2016] Zhang, M. (2016). *Bayesian models on manifolds for image registration and statistical shape analysis*. PhD thesis, The University of Utah.