Note on Mathematics of Imaging

Haocheng Dai

1 Calculus

Definition 1.1. The 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 1.1. 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

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

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 1.2.** Integration by substitution (Change of variables) is a method for evaluating integrals, for one-dimensional case:

\[
\int_b^a f(\varphi(x))\varphi'(x)\,dx = \int_{\varphi(b)}^{\varphi(a)} f(\varphi(x))d\varphi(x) \rightarrow dy = y'(x)\,dx = \left| \frac{dy}{dx} \right| \,dx
\]

For \( n \)-dimensional case:

\[
dy_1 dy_2 \cdots dy_n = \det(D\varphi)(x_1, x_2, \cdots, x_n) \cdot dx_1 dx_2 \cdots dx_n
\]

**Definition 1.3.** 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 \( n \)th 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

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

When \( x = a \), we have

\[
c_n n! = f^{(n)}(a) \\
c_n = \frac{f^{(n)}(a)}{n!}
\]
Remark 1.2. 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 1.

- The Maclaurin series of the exponential function \( e^x \) is
  \[
  e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = \frac{x^0}{0!} + \frac{x^1}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \cdots.
  \]
  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 2. 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:
\[
 t(x + a) = f(x) + a f'(x).
\]
To be more precise, we can also use second-order expansion:
\[
 t(x + a) = f(x) + a f'(x) + \frac{a^2}{2!} f''(x).
\]

Definition 1.4. 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
\[
 e^{ix} = \sum_{n=0}^{\infty} \frac{(ix)^n}{n!} = \frac{(ix)^0}{0!} + \frac{(ix)^1}{1!} + \frac{(ix)^2}{2!} + \frac{(ix)^3}{3!} + \frac{(ix)^4}{4!} + \frac{(ix)^5}{5!} + \frac{(ix)^6}{6!} + \frac{(ix)^7}{7!} + \cdots
\]

\[
 = \frac{x^0}{0!} + ix \cdot \frac{x^1}{1!} - \frac{x^2}{2!} - ix^3 \cdot \frac{x^3}{3!} - \frac{x^4}{4!} + \frac{x^5}{5!} - \frac{x^6}{6!} - ix^7 \cdot \frac{x^7}{7!} + \cdots
\]

\[
 = \cos(x) + i \sin(x)
\]
• Using differentiation: Consider the function \( f(\theta) \)

\[
f(\theta) = \frac{\cos \theta + i \sin \theta}{e^{i\theta}} = e^{-i\theta} (\cos \theta + i \sin \theta)
\]

for real \( \theta \). Differentiating gives by the product rule

\[
f'(\theta) = e^{-i\theta} (i \cos \theta - \sin \theta) - ie^{-i\theta} (\cos \theta + i \sin \theta) = 0
\]

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

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

**Definition 1.5.** [Euler-Lagrange equation] is defined as

\[
\frac{\partial F}{\partial y} - \frac{d}{dx} \left( \frac{\partial F}{\partial y'} \right) = 0.
\]

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 3.** Suppose \( A \) and \( B \) are two points in a Euclidean space. We want to find the geodesic between \( A \) and \( B \).

**Solution.** We would like to minimize

\[
L = \int_{A}^{B} 1 ds, \quad \text{where} \quad ds = \sqrt{(dx)^2 + (dy)^2} = \sqrt{1 + (y')^2} dx
\]

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

\[
\frac{\partial F}{\partial y} - \frac{d}{dx} \left( \frac{\partial F}{\partial y'} \right) = 0
\]

\[
- \frac{d}{dx} \left( \frac{y'}{\sqrt{1 + (y')^2}} \right) = 0
\]

\[
\frac{y'}{\sqrt{1 + (y')^2}} = c
\]

\[
(y')^2 = \frac{c^2}{1 - c^2}
\]

\[
y' = c_1
\]

\[
y = c_1 x + c_2
\]
2 Vector Space

Definition 2.1. Vector space is a set of elements called vectors together with two operations: addition and scalar multiplication, which are assumed to satisfy the following axioms:

1. \( x + y = y + x \)
2. \( (x + y) + z = x + (y + z) \)
3. There is a null vector \( \theta \in X \) such that \( x + \theta = x \) for every \( x \in X \)
4. \( \alpha(x + y) = \alpha x + \alpha y \); \( (\alpha + \beta)x = \alpha x + \beta x \)
5. \( (\alpha \beta)x = \alpha (\beta x) \)
6. \( 0x = \theta; 1x = x \)

Definition 2.2. Inner product \( \langle \cdot , \cdot \rangle : X \times X \rightarrow \mathbb{R} \) is a mapping that satisfies the following axioms:

1. \( \langle x, y \rangle = \langle y, x \rangle \)
2. \( \langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle \)
3. \( \langle \lambda x, y \rangle = \lambda \langle x, y \rangle \)
4. \( \langle x, x \rangle \geq 0 \) and \( \langle x, x \rangle = 0 \) if and only if \( x = \theta \)

where \( X \) is a vector space.

Example 4. For \( A, B \in \text{GL}(n) \), the inner product \( \langle A, B \rangle = \text{Tr}(A^T B) \) and the associated norm \( \|A\|_2 = \text{Tr}(A^T A)^{1/2} \).

Definition 2.3. Positive definite matrix \( A \) is an \( n \times n \) symmetric matrix such that

1. \( \langle x^T Ax \rangle \geq 0 \) for all \( x \in \mathbb{R} \)
2. \( \langle x^T Ax \rangle = 0 \) holds if and only if \( x = 0 \)

Definition 2.4. 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 \( f \).

Definition 2.5. Norm \( \| \cdot \| : X \times X \rightarrow \mathbb{R} \) is a mapping that satisfies the following axioms:

1. \( \|x\| \geq 0 \) for all \( x \in X \), \( \|x\| = 0 \) if and only in \( x = \theta \)
2. \( \|x + y\| \leq \|x\| + \|y\| \) for each \( x, y \in X \) \( \triangleright \) Triangle inequality
3. \( \|\alpha x\| = |\alpha| \cdot \|x\| \) for all scalars \( \alpha \) and each \( x \in X \)

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, then it is called Hilbert space.

Example 5. 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 \( f : \mathbb{R}^n \to \mathbb{R}, f(x) = \|x\|_p, p \geq 1 \), then \( f \) is convex.

Figure 2: Image of 2D norm

Definition 2.6. \( l^p \) space consists of all sequences of scalars \( \{\xi_1, \xi_2, \ldots\} \) for which
\[
\sum_{i=1}^{\infty} |\xi_i|^p < \infty
\]
where \( 1 \leq p < \infty \).

The norm of an element \( x = \{\xi_i\} \) in \( l^p \) is defined as
\[
\|x\|_p = \left( \sum_{i=1}^{\infty} |\xi_i|^p \right)^{1/p}
\]

Definition 2.7. \( L^p[a, b] \) 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. For \( p \neq 2 \), the space of \( L^p \)-functions is a Banach space which is not a Hilbert space.

Remark 2.3. Always remember the absolute value sign in norm calculation.

Remark 2.4. \( L^p \) space is a space of measurable functions for which the \( p \)-th power of the absolute value is Lebesgue integrable.
Figure 3: Visualization of $\|x\|_p = 1$, namely the unit circle in different norms, which are the cross sections of Figure 2.

Definition 2.8. **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 $\mu$.

Remark 2.5. 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.

Definition 2.9. **Frobenius norm** is defined by

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

Definition 2.10. **Normed linear vector space** is a vector space $X$ on which there is defined a real-valued function that maps each element $x$ in $X$ into a real number $\|x\|$.

Definition 2.11. **Pre-Hilbert space** is a linear vector space $X$ together with an inner product defined on $X \times X$.

Figure 4: Riemann-Darboux’s integration (in blue) and Lebesgue integration (in red).
Definition 2.12. **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 2.6.** In a normed space, every convergent sequence is a Cauchy sequence, however, a Cauchy sequence may not be convergent.

![Figure 5: Example of Cauchy sequence](image)

Definition 2.13. A normed linear vector space \( X \) is **complete** if every Cauchy sequence from \( X \) has a limit in \( X \). The limit is also a vector.

Definition 2.14. **Banach space** is a complete normed linear vector space.

Definition 2.15. **Hilbert space** is a complete pre-Hilbert space or a Banach space whose norm is defined by the inner product.

**Remark 2.7.** Actually, the hypothesis of completeness is weak, it is always possible to complete a space \( X \) endowed with an inner product. The “completed” norm is then associated with the inner product.

**Remark 2.8.** Hilbert spaces are complete infinite-dimensional spaces in which distances and angles can be measured. These spaces have a major impact on analysis and topology and will provide a convenient and proper setting for the functional analysis of partial differential equations.

**Theorem 2.1. Holder 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 \( x = [x_1, x_2, \cdots]^T \in l_p, y = [y_1, y_2, \cdots]^T \in l_q \), then

\[
\sum_{i=1}^{\infty} |x_i y_i| \leq \|x\|_p \cdot \|y\|_q
\]

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

**Theorem 2.2.** Cauchy-Schwarz Inequality. If \( p = 2 \) and \( q = 2 \) and if \( x = [x_1, x_2, \cdots]^T \in l_2, y = [y_1, y_2, \cdots]^T \in l_2 \), then

\[
\sum_{i=1}^{\infty} |x_i y_i| \leq \|x\|_2 \cdot \|y\|_2
\]

**Theorem 2.3.** Minkowski Inequality. If \( x \) and \( y \) are in \( l_p \), \( 1 \leq p \leq \infty \), then

\[
\|x + y\|_p \leq \|x\|_p + \|y\|_p
\]
Equality holds if and only if $k_1 x = k_2 y$ for some positive constants $k_1$ and $k_2$.

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

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

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

**Definition 2.16. 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 up to order $k$ have a finite $L^p$ norm.

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

**Remark 2.9.** 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 6.** Sobolev spaces with $p = 2$ are especially important because of their connection with 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 7.** The space $H^k$ can be defined naturally in terms of Fourier series whose coefficients decay sufficiently rapidly, namely,

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

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

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

**Remark 2.10. Overview of several spaces:**

---

1A 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.
<table>
<thead>
<tr>
<th>Spaces</th>
<th>Elements</th>
<th>Operations</th>
<th>Equivalents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vector</td>
<td>vectors</td>
<td>$\mathbf{x} + \mathbf{y}, \alpha \mathbf{x}$</td>
<td></td>
</tr>
<tr>
<td>Pre-Hilbert</td>
<td>vectors</td>
<td>$\mathbf{x} + \mathbf{y}, \alpha \mathbf{x}, \langle \cdot, \cdot \rangle$</td>
<td>vector space + $\langle \cdot, \cdot \rangle$</td>
</tr>
<tr>
<td>Banach</td>
<td>vectors</td>
<td>$\mathbf{x} + \mathbf{y}, \alpha \mathbf{x}, | \cdot |$</td>
<td>vector space(complete) + $| \cdot |$</td>
</tr>
<tr>
<td>Hilbert</td>
<td>vectors</td>
<td>$\mathbf{x} + \mathbf{y}, \alpha \mathbf{x}, \langle \cdot, \cdot \rangle$</td>
<td>vector space(complete) + $\langle \cdot, \cdot \rangle + | \cdot |$</td>
</tr>
<tr>
<td>Lebesgue</td>
<td>functions(vectors) s.t. $|f|_p &lt; \infty$, $p \in [1, \infty)$</td>
<td>$\mathbf{x} + \mathbf{y}, \alpha \mathbf{x}, | \cdot |$</td>
<td>Banach space</td>
</tr>
<tr>
<td>$H^1$ Sobolev</td>
<td>functions(vectors) s.t. $f'$ has $L^2$ norm</td>
<td>$\mathbf{x} + \mathbf{y}, \alpha \mathbf{x}, | \cdot |$</td>
<td>Banach space</td>
</tr>
</tbody>
</table>

**Definition 2.17.** [Fourier transform](#) of function $f(x)$ can be expressed as

$$F(\omega) = \int_{-\infty}^{\infty} f(x) e^{-2\pi i \omega x} \, dx.$$  

And inverse of Fourier transform reads as

$$f(x) = \int_{-\infty}^{\infty} F(\omega) e^{2\pi i \omega x} \, d\omega.$$  

**Definition 2.18.** [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

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

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 2.11.** Fourier transform of Dirac delta function is

$$\Delta(\omega) = \mathcal{F}(\delta(x - x_0)) = \int_{-\infty}^{\infty} \delta(x - x_0) e^{-2\pi i \omega x} \, dx = e^{-2\pi i \omega x_0}.$$  

When $x_0 = 0$,

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

**Theorem 2.5.** 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

\[
L G(x,s) = \delta(x-s),
\]

where \( \delta \) is the Dirac delta function.

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

\[
\int L G(x,s)f(s)ds = \int \delta(x-s)f(s)ds = f(x)
\]

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 2.19. **Linear operators** \( L \) are the operators such that for every pair of functions \( f \) and \( g \) and scalar \( t \), it has

\[
L(f + g) = L(f) + L(g),
\]

\[
L(tf) = tL(f).
\]

Definition 2.20. **Eigenfunctions** \( u \) of linear operators \( D \) are the functions such that

\[
Du = \lambda u,
\]

where \( \lambda \) is the eigenvalue and \( u \) is the corresponding eigenfunction.

**Example 8.** Below are a few examples of eigenfunctions of linear operator:

- **Differentiation:** \( \frac{d}{dx} e^{\lambda x} = \lambda e^{\lambda x} \)
- **Gradient:** \( \nabla e^{\lambda x} = \lambda e^{\lambda x} \)
- **Laplacian:**
  \[
  - \nabla^2 \sin(ax + b) = \lambda \sin(x^2)
  \]
  \[
  - \nabla^2 j_l(r)Y^m_l(\theta, \phi) = 0, \text{ where } Y^m_l(\theta, \varphi) = \sqrt{\frac{(2\ell+1)(\ell-m)!}{(\ell+m)!}} P^m_\ell(\cos \theta) e^{im\varphi}, j_l(x) = \sqrt{\frac{x}{2\pi}} J_{\ell+1/2}(x).
  \]

**Example 9.** The following operators are all linear:

\[https://www.math.mcgill.ca/jakobson/papers/soup.pdf\]
<table>
<thead>
<tr>
<th>Operator</th>
<th>$L(f + g) = L(f) + L(g)$</th>
<th>$L(tf) = tL(f)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Differential</td>
<td>$\frac{df}{dx} + \frac{dg}{dx}$</td>
<td>$\frac{df}{dx} = t\frac{df}{dx}$</td>
</tr>
<tr>
<td>Integral</td>
<td>$\int (f + g)dx = \int f dx + \int g dx$</td>
<td>$\int (tf)dx = t \int f dx$</td>
</tr>
<tr>
<td>Gradient</td>
<td>$\nabla(f + g) = \nabla f + \nabla g$</td>
<td>$\nabla(tf) = t \nabla f$</td>
</tr>
<tr>
<td>Fourier</td>
<td>$F(f + g) = Ff + Fg$</td>
<td>$F(tf) = tFf$</td>
</tr>
<tr>
<td>Laplacian</td>
<td>$\Delta(f + g) = \Delta f + \Delta g$</td>
<td>$\Delta(tf) = t \Delta f$</td>
</tr>
<tr>
<td>Expectation</td>
<td>$E(f + g) = E(f) + E(g)$</td>
<td>$E(tf) = tE(f)$</td>
</tr>
</tbody>
</table>

**Definition 2.21.** 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 $x, x' \in \mathcal{X}$, we have

$$k(x, x') : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$$

$$k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$$

**Remark 2.12.** 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 2.13.** 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 $x, x'$ represent two different books, we can’t take an inner product between books, but we can take an inner product between the feature maps $\phi(x), \phi(x')$ corresponding to $x, x'$.

**Remark 2.14.** The kernel gives a way to compute inner products in some feature space without even knowing what this space is and what is $\phi$. 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 $k(x, y) = \exp(-\gamma \|x - y\|^2)$. If we Taylor-expand this function, we’ll see that it corresponds to an infinite-dimensional codomain of $\phi$.

**Figure 6:** $\phi(x) = [x_1, x_2, x_1x_2]^T$ 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 $\phi$.

**Definition 2.22.** 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 \( \delta_x : f \to f(x) \), for any \( x \in X \) (typically \( X \) will be \( \mathbb{R}^n \)), \( f \in \mathcal{H} \), \( \delta_x \) is bounded, i.e., there exists \( \delta_x, M \) such that \( \| \delta_x f \| \leq M \| f \|_\mathcal{H} \);

• For any \( x \in X, f \in \mathcal{H} \), there exists a unique function (vector) \( k_x = k(\cdot, x) \in \mathcal{H} \), s.t. \( f(x) = \delta_x(f) = \langle f, k_x \rangle_\mathcal{H} \), namely the reproducing ability, which is guaranteed by Riesz representation theorem.

**Definition 2.23.** Reproducing kernel \( k : X \times X \to \mathbb{R} \) is a function, where \( \mathcal{H} \) is a Hilbert space and \( X \) is a non-empty set, if \( k \) satisfies

1. \( \forall x \in X, k(\cdot, x) \in \mathcal{H} \) \( \triangleright \) Feature map of every point is in the feature space

2. \( \forall x \in X, \forall f \in \mathcal{H}, f(x) = \langle f, k(\cdot, x) \rangle_\mathcal{H} \) \( \triangleright \) Reproducing property

3. \( k(x, y) = \langle k(\cdot, x), k(\cdot, y) \rangle_\mathcal{H} = \langle \phi(x), \phi(y) \rangle_\mathcal{H} \)

**Remark 2.15.** From a discrete perspective, \( k(\cdot, \cdot) \) can be regarded as a “matrix”; \( k(\cdot, x^{(i)}) \) can be viewed as a “vector” designated at \( x^{(i)} \) column; and \( k(x^{(i)}, x^{(j)}) \) is a scalar designated at \( x^{(i)} \) “row” and \( x^{(j)} \) “column”.

**Remark 2.16.** The feature map is not unique, only the kernel is. RKHS functions can be written as linear combination of feature maps \( k(\cdot, x) \), which we can regard as “basis function”:

\[
\begin{align*}
f(\cdot) &= \sum_{i=1}^{m} \alpha_i \phi_i(\cdot) = \sum_{i=1}^{m} \alpha_i k(\cdot, x^{(i)}) \\
f(x) &= \langle f(\cdot), k(\cdot, x) \rangle_\mathcal{H} \\
&= \left\langle \sum_{i=1}^{m} \alpha_i k(\cdot, x^{(i)}), k(\cdot, x) \right\rangle_\mathcal{H} \\
&= \sum_{i=1}^{m} \alpha_i \left\langle k(\cdot, x^{(i)}), k(\cdot, x) \right\rangle_\mathcal{H} \\
&= \sum_{i=1}^{m} \alpha_i k(x, x^{(i)})
\end{align*}
\]

For shorter notation:

\[
\begin{align*}
f &= \sum_{i=1}^{m} \alpha_i \phi_i = \sum_{i=1}^{m} \alpha_i k(\cdot, x^{(i)}) \\
f(x) &= \langle f, k(\cdot, x) \rangle_\mathcal{H} \\
&= \left\langle \sum_{i=1}^{m} \alpha_i k(\cdot, x^{(i)}), k(\cdot, x) \right\rangle_\mathcal{H} \\
&= \sum_{i=1}^{m} \alpha_i \left\langle k(\cdot, x^{(i)}), k(\cdot, x) \right\rangle_\mathcal{H} \\
&= \sum_{i=1}^{m} \alpha_i k(x, x^{(i)})
\end{align*}
\]
Remark 2.17. 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 \( \alpha_i \) and \( G(x, s) \) corresponds to RKHS \( k(x, x_i) \).

\[
Lu(x) = g(x) \\
u(x) = \int g(s)G(x, s)ds \\
g(x) = \sum_{i=1}^{m} \alpha_i k(x, x_i)
\]

Example 10. We define a feature map \( \phi : \mathbb{R}^2 \rightarrow \mathbb{R}^3 \)

\[
\phi(x) = [x_1, x_2, x_1x_2]^T
\]

For the reproducing property, we define an RKHS function \( f : \mathbb{R}^2 \rightarrow \mathbb{R} \)

\[
f(x) = \sum_{l=1}^{\infty} f_l \phi_l(x) \\
= ax_1 + bx_2 + cx_1x_2 \\
f(\cdot) = [a, b, c]^T
\]

Remark 3

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

\[
f(x) = f(\cdot)^T \phi(x) \\
= \langle f(\cdot), \phi(x) \rangle_{\mathcal{H}}
\]

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

Definition 2.24. [Primal-Dual method]

Assume the primal problem as below:

\[
\begin{align*}
\text{maximize} & \quad z(x) \\
\text{subject to} & \quad G(x) \leq \theta & x \in \Omega
\end{align*}
\]

which is equivalent to

\[
\begin{align*}
\text{minimize} & \quad w(y) = \sup_{x \in \Omega} \{ z(x) + \langle G(x), y \rangle \} \\
\text{subject to} & \quad y \geq \theta & x \in \Omega
\end{align*}
\]

More specifically,
Primal

maximize \( z = \sum_{j=1}^{n} c_j x_j \)

subject to \( \sum_{j=1}^{n} a_{ij} x_j \leq b_i \) \((i = 1, 2, \cdots, m)\)

\( x_j \geq 0 \) \((j = 1, 2, \cdots, n)\)

Dual

minimize \( w = \sum_{i=1}^{m} b_i y_i \)

subject to \( \sum_{i=1}^{m} a_{ij} y_i \geq c_j \) \((j = 1, 2, \cdots, n)\)

\( y_i \geq 0 \) \((i = 1, 2, \cdots, m)\)

**Remark 2.18.** 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 7: Supremum and Infimum](image)

Example 11.

Primal

\[ \begin{align*}
\text{max} \quad & z = 30x_1 + 100x_2 \\
\text{s.t.} \quad & x_1 + x_2 \leq 7 \\
& 4x_1 + 10x_2 \leq 40 \\
& x_1 \geq 3 \\
& x_1 \geq 0 \\
& x_2 \geq 0
\end{align*} \]

Multiply constraints \( i \) by a factor \( y_i \). Choose the sign of \( y_i \) such that all inequalities are \( \leq \) after multi-
plication:

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

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**

\[
\begin{align*}
\text{min} \quad w = 7y_1 + 40y_2 - 3y_3 \\
\text{s.t.} \quad & y_1 + 4y_2 - y_3 - y_4 = 30 \\
& y_1 + 10y_2 - y_5 = 100
\end{align*}
\]

**Remark 2.19.** 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 2.25.** 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 \(\varphi\) for TPS is \(\varphi(r) = r^2 \log(r)\).

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

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

where the tuning parameter \(\lambda\) 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 2.26.** 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 2.27. **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 2.20. Bijective functions are essential to many areas of mathematics including the definitions of isomorphism, homeomorphism, diffeomorphism.

Definition 2.28. If $x \in X$, then the **image** of $x$ under $f$, is denoted as $f(x)$. 
3 Differential Geometry for Images

Overview. Riemannian manifolds is a space that locally resemble Euclidean space and equipped with a Riemannian metric, which defines the inner product at each point on the tangent space and is in form of a metric tensor. Tangent space is a vector(Euclidean) space that associate with each point on the manifold. 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. [Isomorphism\textsuperscript{2,3,4}] is a function between two structures of the same type that can be reversed by an inverse function, i.e. bijective.

Example 12. In various areas of mathematics, isomorphisms have received specialized names, depending on the type of structure under consideration.

- An isometry (Def. \textsuperscript{3.38}) is an isomorphism of metric spaces, also a metric-preserving diffeomorphism.
- A homeomorphism (Def. \textsuperscript{3.2}) is an isomorphism of topological spaces.
- A diffeomorphism (Def. \textsuperscript{3.8}) 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 3.2. [Homeomorphism\textsuperscript{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 3.3. [Manifold\textsuperscript{M}] 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 3.4. [Immersion\textsuperscript{f:}\mathit{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 3.5. [Embedding\textsuperscript{f:}\mathit{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 3.6. [Submanifold\textsuperscript{N}] (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 3.7. [Rank\textsuperscript{f:}] 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 3.8. **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.

Definition 3.9. For two manifolds \( M \) and \( N \), a smooth mapping \( f : M \to N \) induces a linear mapping of the tangent spaces \( f_* : T_p M \to T_{f(p)} N \) called the **Jacobian** of \( f \).

Definition 3.10. **Metric-preserving mapping** \( f : M \to N \) is a smooth mapping for all \( p \in M \) and tangent vectors \( u, v \in T_p M \), we have

\[
\langle u, v \rangle^M_p = \langle f_*u, f_*v \rangle^{N}_{f(p)}.
\]

Definition 3.11. **Tangent space** \( T_p M \) is a vector space attached to each point on a manifold \( M \), which is equivalent to the Euclidean space. Intuitively, it’s 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 3.3. **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 in.

Example 13. The rotation group is presented as

\[
\text{SO}(3) = \{ R \in \mathbb{R}^{3\times 3} | R^T R = I, |R| = 1 \}
\]

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

\[
\frac{d}{dt} \bigg|_{t=0} R(t)^T R(t) = X^T + X = 0
\]

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

Remark 3.4. The **cross product** of vector \( \mathbf{a} \) and \( \mathbf{b} \) can be written as below

\[
\mathbf{a} \times \mathbf{b} = [\mathbf{a}] \times \mathbf{b} = \begin{pmatrix} 0 & -a_3 & a_2 \\ a_3 & 0 & -a_1 \\ -a_2 & a_1 & 0 \end{pmatrix} \cdot \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix},
\]

namely a skew-symmetric matrix times the vector \( \mathbf{b} \).

Example 14. The group of symmetric positive definite matrices is presented as

\[
\text{SPD}(n) = \{ X \in \mathbb{R}^{n \times n} | X = X^T, X > 0 \}.
\]

The tangent space \( T_A \text{SPD}(n) \) at \( A \), is the space of symmetric matrices \( \mathbb{R} \).
Definition 3.12. **Tangent bundle** \( TM \) consists of the tangent space \( T_pM \) at all points \( p \) in \( M \).

\[
TM = \{(p,v) | p \in M, v \in T_pM\}
\]

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 3.13. **Hessian matrix** of a differentiable, multivariable function \( f : \mathbb{R}^n \to \mathbb{R} \) at \( p \) is defined by

\[
H_f(p) = \begin{pmatrix}
\frac{\partial^2 f}{\partial x_1 \partial x_1}(p) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(p) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(p) \\
\frac{\partial^2 f}{\partial x_2 \partial x_1}(p) & \frac{\partial^2 f}{\partial x_2 \partial x_2}(p) & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n}(p) \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial^2 f}{\partial x_n \partial x_1}(p) & \frac{\partial^2 f}{\partial x_n \partial x_2}(p) & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n}(p)
\end{pmatrix}
\]

Remark 3.5. 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.

\[
\nabla f(x_1, \ldots, x^n) = \begin{pmatrix}
\frac{\partial f}{\partial x_1}(x_1, \ldots, x^n) \\
\vdots \\
\frac{\partial f}{\partial x_n}(x_1, \ldots, x^n)
\end{pmatrix}
\]

\[
D(\nabla f) = \begin{pmatrix}
\frac{\partial \nabla f}{\partial x_1}(x_1, \ldots, x^n) & \frac{\partial \nabla f}{\partial x_2}(x_1, \ldots, x^n) & \cdots & \frac{\partial \nabla f}{\partial x_n}(x_1, \ldots, x^n)
\end{pmatrix} = H_f
\]

Remark 3.6. If multivariable function \( f : \mathbb{R}^n \to \mathbb{R}, \nabla f(x) = 0 \), \( H_f(x) \) is positive (resp. negative) definite, then \( x \) is the isolated local minimum (resp. maximum).

Remark 3.7. Relationship between convexity and positive-definiteness:

\( f \) is convex \iff \( \forall p, H_f(p) \) is positive semi–definite

\( f \) is strictly convex \iff \( \forall p, H_f(p) \) is positive definite

Definition 3.14. **Laplacian** of a differentiable, multivariable function \( f : \mathbb{R}^n \to \mathbb{R} \) at \( p \) is defined by

\[
\Delta f_p = \mathrm{tr}(H_f_p) = \sum_{i=1}^{n} \frac{\partial^2 f}{\partial x_i^2} p.
\]

Definition 3.15. **Jacobian matrix** (differential) of a differentiable, vector-valued function \( f : \mathbb{R}^n \to \mathbb{R}^m \) at \( p \) is defined by

\[
Df_p = \begin{pmatrix}
\frac{\partial f_1^1}{\partial x_1}(p) & \frac{\partial f_1^2}{\partial x_1}(p) & \cdots & \frac{\partial f_1^m}{\partial x_1}(p) \\
\frac{\partial f_2^1}{\partial x_1}(p) & \frac{\partial f_2^2}{\partial x_1}(p) & \cdots & \frac{\partial f_2^m}{\partial x_1}(p) \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial f_m^1}{\partial x_1}(p) & \frac{\partial f_m^2}{\partial x_1}(p) & \cdots & \frac{\partial f_m^m}{\partial x_1}(p)
\end{pmatrix} = \begin{pmatrix}
(\nabla f_1^1)^T \\
(\nabla f_2^1)^T \\
\vdots \\
(\nabla f_m^1)^T
\end{pmatrix},
\]
where

\[ f(x^1, \ldots, x^n) = \begin{pmatrix} f^1(x^1, \ldots, x^n) \\ \vdots \\ f^m(x^1, \ldots, x^n) \end{pmatrix}. \]

**Definition 3.16.** Divergence of a differentiable, vector-valued function \( f : \mathbb{R}^n \to \mathbb{R}^m \) at \( p \) is defined by

\[ \nabla \cdot f_p = \text{tr}(Df_p) = \sum_{i=1}^{n} \partial_i f^i_p. \]

**Remark 3.8.** The Jacobian of \( \nabla f \), where \( f : \mathbb{R}^n \to \mathbb{R} \), is the Laplacian of \( f \).

**Remark 3.9.** As we all know, a matrix can be thought of as a linear transformation. Here as well, we can think of \( Df_p \) as a linear function \( Df_p : T_p\mathbb{R}^n \to T_p\mathbb{R}^m \). In other words, \( Df_p \) maps a vector in the tangent space at the source point \( p \) to a vector in the tangent space at the target point \( f(p) \). More formally, we have

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

![Figure 8: Jacobian](image)

**Remark 3.10.** 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 15.** 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 16.** 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
\[ \int_a^b (2x^3 + 1)^7(x^2)dx = \frac{1}{6} \int_a^b f(\varphi(x))\varphi'(x)dx \]
\[ = \frac{1}{6} \int_{\varphi(a)}^{\varphi(b)} f(\varphi(x))d\varphi(x) \]
\[ = \frac{1}{6} \int_{\varphi(a)}^{\varphi(b)} f(y)dy \]
\[ = \frac{1}{6} \int_{\varphi(a)}^{\varphi(b)} y^7dy \]
\[ = \frac{1}{48}[y^8]_{\varphi(a)}^{\varphi(b)} \]

Definition 3.17. [Vector field] \( \mathcal{X}(M) \) is a function on a manifold \( M \) that smoothly assigns to each point \( p \in M \) a tangent vector \( v \in T_pM \).

Figure 9: Tangent Space

Definition 3.18. [Riemannian metric] on a differential manifold \( M \) is a smooth function that assigns to each point \( p \in M \) an inner product \( \langle \cdot, \cdot \rangle \) on the tangent space \( T_pM \).

Remark 3.11. Assuming \( A \) is the metric at point \( p \in M \), and \( v, \lambda \) are the unit eigenvector and its square root corresponding eigenvalue of \( A \).
\[ Av = \lambda v \]
\[ \|v\|^2 = v^TAv = v^T\lambda v = \lambda v^Tv = \lambda \]
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.

**Definition 3.19.** [Metric tensor][17] is a function which tells how to compute the distance between any two points in a given space.

**Example 17.** Typically, we calculate the arc length as below

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

![Figure 10: Base vectors on tangent space](image)

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

\[
\left( \frac{d\tilde{R}}{d\lambda} \right)^2 = \frac{d\tilde{R}}{du} \cdot \frac{d\tilde{R}}{du} + \frac{d\tilde{R}}{dv} \cdot \frac{d\tilde{R}}{dv} = \left( \begin{array}{cc}
\frac{du}{d\lambda} & \frac{dv}{d\lambda} \\
\frac{dv}{d\lambda} & \frac{du}{d\lambda}
\end{array} \right)
\]

\[
\begin{aligned}
&= \left( \frac{du}{d\lambda} \right)^2 \left( \frac{d\tilde{R}}{du} \cdot \frac{d\tilde{R}}{du} \right) + \left( \frac{dv}{d\lambda} \right)^2 \left( \frac{d\tilde{R}}{dv} \cdot \frac{d\tilde{R}}{dv} \right) \\
&\quad + \frac{du}{d\lambda} \frac{dv}{d\lambda} \left( \frac{d\tilde{R}}{du} \cdot \frac{d\tilde{R}}{dv} + \frac{d\tilde{R}}{dv} \cdot \frac{d\tilde{R}}{du} \right)
\end{aligned}
\]

\[
= \left( \frac{du}{d\lambda} \frac{dv}{d\lambda} \right) \begin{pmatrix}
\frac{d\tilde{R}}{du} \cdot \frac{d\tilde{R}}{du} & \frac{d\tilde{R}}{du} \cdot \frac{d\tilde{R}}{dv} \\
\frac{d\tilde{R}}{dv} \cdot \frac{d\tilde{R}}{du} & \frac{d\tilde{R}}{dv} \cdot \frac{d\tilde{R}}{dv}
\end{pmatrix}
\]

\[
= \left( \begin{array}{cc}
\tilde{e}_u \cdot \tilde{e}_u & \tilde{e}_u \cdot \tilde{e}_v \\
\tilde{e}_v \cdot \tilde{e}_u & \tilde{e}_v \cdot \tilde{e}_v
\end{array} \right)
\]

\[
\begin{pmatrix}
\frac{du}{d\lambda} \\
\frac{dv}{d\lambda}
\end{pmatrix}
\]

**Remark 3.12.** 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:

\[ \vec{R} = [X, Y, Z]^T \]

where

\[ X = \cos(v) \sin(u) = \cos(\lambda) \sin(\lambda) \]

\[ Y = \sin(v) \sin(u) = \sin(\lambda) \sin(\lambda) \]

\[ Z = \cos(u) = \cos(\lambda) \]

when \( u = \lambda, v = \lambda \).

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

\[ \vec{e}_u = \frac{d\vec{R}}{du} = \frac{\partial \vec{R}}{\partial X} \frac{\partial X}{\partial u} + \frac{\partial \vec{R}}{\partial Y} \frac{\partial Y}{\partial u} + \frac{\partial \vec{R}}{\partial Z} \frac{\partial Z}{\partial u} \]

\[ = \cos(v) \cos(u) \frac{\partial \vec{R}}{\partial X} + \sin(v) \cos(u) \frac{\partial \vec{R}}{\partial Y} - \sin(u) \frac{\partial \vec{R}}{\partial Z} \]

\[ = \cos(v) \cos(u) \vec{e}_X + \sin(v) \cos(u) \vec{e}_Y - \sin(u) \vec{e}_Z \]

\[ \vec{e}_v = \frac{d\vec{R}}{dv} = \frac{\partial \vec{R}}{\partial X} \frac{\partial X}{\partial v} + \frac{\partial \vec{R}}{\partial Y} \frac{\partial Y}{\partial v} + \frac{\partial \vec{R}}{\partial Z} \frac{\partial Z}{\partial v} \]

\[ = -\sin(v) \sin(u) \frac{\partial \vec{R}}{\partial X} + \cos(v) \sin(u) \frac{\partial \vec{R}}{\partial Y} \]

\[ = -\sin(v) \sin(u) \vec{e}_X + \cos(v) \sin(u) \vec{e}_Y \]

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

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

\[
= \left( \frac{du}{d\lambda} \right)^2 + \sin^2(u) \left( \frac{dv}{d\lambda} \right)^2
\]

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

\[
\left\| \frac{d\vec{R}}{d\lambda} \right\|^2 = \left( \frac{du}{d\lambda} \right)^2 + \sin^2(u) \left( \frac{dv}{d\lambda} \right)^2
\]

\[
= 0^2 + \sin^2 \left( \frac{\pi}{4} \right) \cdot 1^2 = \frac{1}{2}
\]

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

- 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

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

Figure 12: $u = \frac{\pi}{2}, v = \lambda$

**Remark 3.13.** So the figure below is a good illustration of the role metric tensor plays:

Figure 13: 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 3.14.** 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.

![Visualization of metric tensor](image1.png)

Figure 14: Visualization of metric tensor

![Visualization of sphere metric field](image2.png)

Figure 15: Visualization of sphere metric field

**Definition 3.20.** A **Riemannian manifold** $(M, g)$ is a differentiable (smooth) manifold $M$ equipped with a Riemannian metric $g$.

**Example 18.** 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 $v, w \in T_p M$, 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 \(v^T g w\). Also, for a vector \(v\), we can compute the length of the vector as \(\langle v, v \rangle^{1/2}\), which is the \(L^2\) norm. Sometimes, people utilize the inverse of the diffusion tensor, \(D^{-1}\), to define a local cost function as

\[
\langle v, w \rangle = v^T D^{-1} w,
\]

where \(v, w \in T_p M\). 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 3.21.** Runge–Kutta methods (RK4)

Let an initial value problem be specified as follows: The only things we know are the slope of the tangent line to the curve \(y\) at any point

\[
\frac{dy}{dt} = f(t, y)
\]

and the initial condition \((t_0, y_0)\), namely

\[
y(t_0) = y_0.
\]

We want to approximate the original function \(y\), which equals to \(\int f(t, y)\). For the \(n + 1^{th}\) iteration, we have the approximation

\[
y_{n+1} = y_n + \frac{1}{6} h (k_1^n + 2k_2^n + 2k_3^n + k_4^n),
\]

\[
t_{n+1} = t_n + h,
\]

where \(h\) is the step size and

\[
k_1^n = f(t_n, y_n)
\]

\(\triangleright\) The slope at the beginning of the interval

\[
k_2^n = f \left( t_n + \frac{h}{2}, y_n + \frac{h}{2} k_1^n \right)
\]

\(\triangleright\) The slope at the midpoint of the interval, using \(y\) and \(k_1^n\)

\[
k_3^n = f \left( t_n + \frac{h}{2}, y_n + \frac{h}{2} k_2^n \right)
\]

\(\triangleright\) The slope at the midpoint of the interval, using \(y\) and \(k_2^n\)

\[
k_4^n = f(t_n + h, y_n + hk_3^n)
\]

\(\triangleright\) The slope at the end of the interval, using \(y\) and \(k_3^n\)

**Remark 3.15.** When \(k_4^n = k_3^n = k_2^n = k_1^n = f(t_n, y_n)\), it’s the simplest Runge–Kutta method, namely the Euler method.

**Remark 3.16.** In an ideal world, we can integrate a function by integration rules, like \(\int e^x = e^x + c\). However, in the real world, we can only use a numerical procedure to integrate a random function. That’s why we call the result of RK4 or Euler methods as integral curve, which is simply the result after integration.
Definition 3.22. Integral curve is a parametric curve that represents a specific solution to an ordinary differential equation or system of equations. If the differential equation is represented as a vector field or slope field, then the corresponding integral curves are tangent to the field at each point.

Definition 3.23. 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 $v, w \in T_xM$ is given by $\langle v, w \rangle = v^T g(x) w$, where $g(x)$ is the Riemannian metric at point $x$.

**Remark 3.17.** 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 19.** 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

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

**Definition 3.24.** [Geodesic Equation] guarantees the acceleration vector normal to the surface

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

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

Velocity Vector:

$$
\frac{d\bar{R}}{dt} = \frac{\partial \bar{R}}{\partial u} \frac{du}{dt} + \frac{\partial \bar{R}}{\partial v} \frac{dv}{dt}
$$

Acceleration Vector:

$$
\frac{d^2 \bar{R}}{dt^2} = \frac{d}{dt} \left( \frac{\partial \bar{R}}{\partial u} \frac{du}{dt} + \frac{\partial \bar{R}}{\partial v} \frac{dv}{dt} \right)
$$
Expend the expression of acceleration vector, we have

\[
\frac{d^2 \vec{R}}{dt^2} = \frac{d}{dt} \left( \frac{\partial \vec{R}}{\partial u} \frac{du}{dt} + \frac{\partial \vec{R}}{\partial v} \frac{dv}{dt} \right)
\]

\[
= \frac{\partial \vec{R}}{\partial u} \frac{d^2 u}{dt^2} + \frac{du}{dt} \left( \frac{d}{dt} \frac{\partial \vec{R}}{\partial u} \right) + \frac{\partial \vec{R}}{\partial v} \frac{d^2 v}{dt^2} + \frac{dv}{dt} \left( \frac{d}{dt} \frac{\partial \vec{R}}{\partial v} \right)
\]

\[
= \frac{\partial \vec{R}}{\partial u} \frac{d^2 u}{dt^2} + \frac{du}{dt} \left[ \frac{\partial}{\partial u} \left( \frac{\partial \vec{R}}{\partial u} \frac{du}{dt} + \frac{\partial \vec{R}}{\partial v} \frac{dv}{dt} \right) \right]
\]

\[
+ \frac{\partial \vec{R}}{\partial v} \frac{d^2 v}{dt^2} + \frac{dv}{dt} \left[ \frac{\partial}{\partial v} \left( \frac{\partial \vec{R}}{\partial u} \frac{du}{dt} + \frac{\partial \vec{R}}{\partial v} \frac{dv}{dt} \right) \right]
\]

\[
= \frac{\partial \vec{R}}{\partial u} \frac{d^2 u}{dt^2} + \frac{\partial^2 \vec{R}}{\partial u \partial v} \left( \frac{du}{dt} \right)^2 \frac{\partial}{\partial u} \frac{\partial \vec{R}}{\partial v} \frac{dv}{dt} + \frac{\partial^2 \vec{R}}{\partial u \partial v} \frac{du}{dt} \frac{dv}{dt}
\]

\[
+ \frac{\partial \vec{R}}{\partial v} \frac{d^2 v}{dt^2} + \frac{\partial^2 \vec{R}}{\partial u \partial v} \frac{du}{dt} \frac{dv}{dt} + \frac{\partial^2 \vec{R}}{\partial v^2} \left( \frac{dv}{dt} \right)^2
\]

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

\[
\frac{d^2 \vec{R}}{dt^2} = \frac{d^2 u^i}{dt^2} \frac{\partial \vec{R}}{\partial u^j} + \frac{du^i}{dt} \frac{du^j}{dt} \frac{\partial^2 \vec{R}}{\partial u^i \partial u^j} \tag{1}
\]

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

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

where the Christoffel symbol \( \Gamma^k_{ij} \), 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:

\[
\frac{\partial^2 \vec{R}}{\partial u^i \partial u^j} = \Gamma^k_{ij} \frac{\partial \vec{R}}{\partial u^k} + L_{ij} \vec{n} \tag{2}
\]
Finally, by substituting Eq. (2) into Eq. (1), we can have acceleration vector as below

\[
\begin{align*}
\frac{d^2 \vec{R}}{dt^2} &= \frac{d^2 u^i}{dt^2} \cdot \frac{\partial \vec{R}}{\partial u^i} + \frac{du^i}{dt} \cdot \frac{du^j}{dt} \cdot \frac{\partial^2 \vec{R}}{\partial u^i \partial u^j} \\
&= \frac{d^2 u^k}{dt^2} \cdot \frac{\partial \vec{R}}{\partial u^k} + \frac{du^i}{dt} \cdot \frac{du^j}{dt} \cdot \left( \Gamma^k_{ij} \frac{\partial \vec{R}}{\partial u^k} + L_{ij} \vec{n} \right) \\
&= \frac{d^2 u^k}{dt^2} + \Gamma^k_{ij} \frac{du^i}{dt} \cdot \frac{du^j}{dt} \cdot \frac{\partial \vec{R}}{\partial u^k} + L_{ij} \frac{du^i}{dt} \cdot \frac{du^j}{dt} \cdot \vec{n} \\
\end{align*}
\]


\[
\begin{align*}
\text{tangential part} \quad \text{normal part}
\end{align*}
\]

That acceleration vector normal to the surface requires

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

which is the Geodesic Equation.

**Derivation of** $\Gamma^k_{ij}$ **and** $L_{ij}$  

As $\vec{n}$ is perpendicular to the tangent vectors, therefore, by multiplying $\vec{R}$ on both sides of equation above, we can yield that

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

(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)

\[
\begin{pmatrix}
\frac{\partial \vec{R}}{\partial u^1} & \frac{\partial \vec{R}}{\partial u^2} & \frac{\partial \vec{R}}{\partial u^3}
\end{pmatrix} =
\begin{pmatrix}
g_{11} & g_{12} \\
g_{21} & g_{22}
\end{pmatrix}
\]

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

\[
\begin{align*}
\Gamma^k_{ij} g_{kl} g^{lm} &= \frac{\partial^2 \vec{R}}{\partial u^i \partial u^j} \cdot \frac{\partial \vec{R}}{\partial u^l} \cdot g^{lm} \\
\Gamma^k_{ij} \delta^m_k &= \frac{\partial^2 \vec{R}}{\partial u^i \partial u^j} \cdot \frac{\partial \vec{R}}{\partial u^l} \cdot g^{lm} \\
\Gamma^m_{ij} &= \left( \frac{\partial \vec{e}_j}{\partial u^i} \cdot \vec{e}_l \right) g^{lm}
\end{align*}
\]

(4)

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

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

\[\square\]

**Definition 3.25.** Given a vector space $V$ and a functional $f : V \rightarrow \mathbb{R}$, $x, h \in V, \alpha \in \mathbb{R}$, if the limit

\[
\delta f(x) = \lim_{\alpha \rightarrow 0} \frac{1}{\alpha} \left[ f(x + \alpha h) - f(x) \right]
\]

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 $x$. 31
Example 20. 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 3.26. 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 3.18. 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 3.19. Relationship between Gateaux derivative(scalar), directional derivative(scalar) and covariant derivative(vector):
- Gateaux 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 covariant derivative of a function is directional.
- Gateaux 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.

\[
\nabla_h f(x) = h^i \nabla \left( \frac{\partial f}{\partial x^i} \right)
\]

\[
\nabla_h v = h^i \nabla \left( \frac{\partial v^j}{\partial x^i} \right)
\]

\[
= h^i \left( \frac{\partial v^j}{\partial u^i} \right) \vec{e}_j + \left( \frac{\partial v^j}{\partial u^i} \right) \vec{v}^j
\]

\[
= h^i \left( \frac{\partial v^j}{\partial u^i} \right) \vec{e}_j + \left( \frac{\partial v^j}{\partial w^i} \right)
\]

\[
= h^i \left( \frac{\partial v^j}{\partial u^i} \right) \vec{e}_j + v^j \Gamma^k_{ij} \vec{e}_k
\]

\[
= h^i \left( \frac{\partial v^k}{\partial u^i} \right) \vec{e}_k + v^j \Gamma^k_{ij} \vec{e}_k
\]

\[
= h^i \left( \frac{\partial v^k}{\partial u^i} + v^j \Gamma^k_{ij} \right) \vec{e}_k
\]
Definition 3.27. **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. $\nabla_{a\vec{w}+b\vec{v}} = a\nabla_{\vec{w}}\vec{v} + b\nabla_{\vec{v}}\vec{v}$
2. $\nabla_{\vec{w}}(\vec{v} + \vec{u}) = \nabla_{\vec{w}}\vec{v} + \nabla_{\vec{w}}\vec{u}$ \hspace{1cm} \text{Distributive Property}
3. $\nabla_{\vec{w}}(\vec{v} \cdot \vec{u}) = (\nabla_{\vec{w}}\vec{v}) \cdot \vec{u} + \vec{v} \cdot (\nabla_{\vec{w}}\vec{u})$ \hspace{1cm} \text{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. $\nabla_{\vec{w}}\vec{v} = \nabla_{\vec{v}}\vec{w}$ \hspace{1cm} \text{Commutative Property}

Remark 3.20. • 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 $\Gamma^m_{kj}$ give us different ways of “parallel transport”. If $\Gamma^m_{kj} = \frac{1}{2}g^{im}\left(\frac{\partial g_{ji}}{\partial u^k} + \frac{\partial g_{jk}}{\partial u^i} - \frac{\partial g_{ij}}{\partial u^k}\right)$, then it’s Levi-Civita connection. If $\Gamma^m_{kj} = 0$, it’s another connection.

\[
\Gamma^k_{ij} = \frac{1}{2}g^{mk}(\partial_j g_{mi} + \partial_i g_{jm} - \partial_m g_{ij})
\]
\[
\widetilde{\Gamma}^k_{ij} = 0
\]

Figure 20: Different definitions of $\Gamma^m_{kj}$ 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.

- $\Gamma^{m}_{kj}$ Covariant derivative $\nabla_{\vec{w}}\vec{v}$ is the difference between a vector field $\vec{v}$ and its parallel transport in the direction $\vec{w}$. 

33
• 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 $\gamma$.

• 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 21. 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.
The covariant derivative defined in Euclidean space is just intuitive:

\[
\nabla \frac{\partial}{\partial u^1} \vec{v} = \frac{\partial}{\partial u^1} (v^1 \vec{e}_1 + v^2 \vec{e}_2) \\
= \frac{\partial}{\partial u^1} v^1 \vec{e}_1 + \frac{\partial}{\partial u^1} v^2 \vec{e}_2 \\
= \frac{\partial v^1}{\partial u^1} \vec{e}_1 + \frac{\partial v^2}{\partial u^1} \vec{e}_2
\]

\[
\nabla \frac{\partial}{\partial u^2} \vec{v} = \frac{\partial}{\partial u^2} \vec{v} = \vec{0}
\]

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.

\[
\frac{\partial}{\partial u^i} \vec{v} = \frac{\partial v^j}{\partial u^i} \vec{e}_j + \frac{\partial}{\partial u^i} \vec{e}_j \\
= \frac{\partial}{\partial u^i} v^j \vec{e}_j + \frac{\partial}{\partial u^i} \vec{e}_j
\]

Example 22. [15] In extrinsic case, the covariant derivative is defined as below

\[
\nabla \frac{\partial}{\partial u^i} \vec{v} = \frac{\partial}{\partial u^i} \vec{v} - \vec{n} \\
= \frac{\partial}{\partial u^i} (v^1 \vec{e}_1 + v^2 \vec{e}_2) - \vec{n} \\
= \frac{\partial v^1}{\partial u^i} \vec{e}_1 + \frac{\partial v^2}{\partial u^i} \vec{e}_2 - \vec{n} \\
= \frac{\partial v^j}{\partial u^i} \vec{e}_j + \frac{\partial}{\partial u^i} \vec{e}_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} \\

\nabla \frac{\partial}{\partial u^i} \vec{e}_j
\]

where \(L_{ij}\) is the second fundamental form and \(\Gamma^k_{ij}\) is in form of Eq.(4).
Parameterize the space with tangent space basis

\[ \vec{R} = [X, Y, Z]^T \]

where \( X = \cos(u^2) \sin(u^1) \)
\( Y = \sin(u^2) \sin(u^1) \)
\( Z = \cos(u^1) \),

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

\[ \vec{e}_1 = \frac{\partial \vec{R}}{\partial u^1} = +\cos(u^2) \cos(u^1) \frac{\partial \vec{R}}{\partial X} + \sin(u^2) \cos(u^1) \frac{\partial \vec{R}}{\partial Y} - \sin(u^1) \frac{\partial \vec{R}}{\partial Z} \] (6)

\[ \vec{e}_2 = \frac{\partial \vec{R}}{\partial u^2} = -\sin(u^2) \sin(u^1) \frac{\partial \vec{R}}{\partial X} + \cos(u^2) \sin(u^1) \frac{\partial \vec{R}}{\partial Y} \] (7)

By using the chain rule, we have

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

\[ g^{ij} = \begin{pmatrix} \vec{e}_1 \cdot \vec{e}_1 & \vec{e}_1 \cdot \vec{e}_2 \\ \vec{e}_2 \cdot \vec{e}_1 & \vec{e}_2 \cdot \vec{e}_2 \end{pmatrix} = \begin{pmatrix} \frac{\partial \vec{R}}{\partial u^1} \cdot \frac{\partial \vec{R}}{\partial u^1} & \frac{\partial \vec{R}}{\partial u^1} \cdot \frac{\partial \vec{R}}{\partial u^2} \\ \frac{\partial \vec{R}}{\partial u^2} \cdot \frac{\partial \vec{R}}{\partial u^1} & \frac{\partial \vec{R}}{\partial u^2} \cdot \frac{\partial \vec{R}}{\partial u^2} \end{pmatrix}^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{\sin^2(u^1)} \end{pmatrix} \]
Substituting Eq.(6,7) into the second derivative of position vector, we get
\[
\frac{\partial \vec{e}_1}{\partial u_1} = \frac{\partial}{\partial u_1} \left( \frac{\partial \vec{R}}{\partial u_1} \right) = -\cos(u_2) \cos(u_1) \vec{e}_X - \sin(u_2) \sin(u_1) \vec{e}_Y - \cos(u_1) \vec{e}_Z
\]
\[
= -\cos(u_2) \cos(u_1) \vec{e}_X - \sin(u_2) \sin(u_1) \vec{e}_Y - \cos(u_1) \vec{e}_Z
\] (8)

\[
\frac{\partial \vec{e}_2}{\partial u_2} = \frac{\partial}{\partial u_2} \left( \frac{\partial \vec{R}}{\partial u_2} \right) = -\cos(u_2) \sin(u_1) \vec{e}_X - \sin(u_2) \sin(u_1) \vec{e}_Y
\]
\[
= -\cos(u_2) \sin(u_1) \vec{e}_X - \sin(u_2) \sin(u_1) \vec{e}_Y
\] (9)

\[
\frac{\partial \vec{e}_2}{\partial u_1} = \frac{\partial}{\partial u_1} \left( \frac{\partial \vec{R}}{\partial u_2} \right) = -\sin(u_2) \cos(u_1) \vec{e}_X + \cos(u_2) \cos(u_1) \vec{e}_Y
\]
\[
= -\sin(u_2) \cos(u_1) \vec{e}_X + \cos(u_2) \cos(u_1) \vec{e}_Y
\] (10)

Substituting $g^{ij}$ and Eq.(6,7,8,9,10) into Eq.(4), we can yield the Christoffel symbols as below
\[
\Gamma^1_{11} = 0 \quad \Gamma^1_{12} = 0 \quad \Gamma^1_{21} = 0 \quad \Gamma^1_{22} = -\frac{1}{2} \sin(2u_1)
\]
\[
\Gamma^2_{11} = 0 \quad \Gamma^2_{12} = \cot(u_1) \quad \Gamma^2_{21} = \cot(u_1) \quad \Gamma^2_{22} = 0
\]

Substituting the Christoffel symbols into Eq.(5), we finally get the extrinsic expression of covariant derivative on sphere as below
\[
\nabla_{\vec{e}_1} \vec{v} = \left( \frac{\partial v_2}{\partial u_1} + v^2 \cot(u_1) \right) \vec{e}_2
\]
\[
\nabla_{\vec{e}_2} \vec{v} = \left( \frac{\partial v_1}{\partial u_2} - \frac{1}{2} \sin(2u_1)v^2 \right) \vec{e}_1 + \left( \frac{\partial v_2}{\partial u_2} + v^1 \cot(u_1) \right) \vec{e}_2
\] (11)

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
  \[
  \vec{v} = \cos(u_2) \vec{e}_1 + \sin(u_2) \vec{e}_2 \quad \text{where} \quad u_1 = \frac{\pi}{2}, u_2 = \lambda \in [0, \frac{\pi}{2}]
  \]
Substitute the \( u^1, u^2 \) into Eq. (11), we have

\[
\nabla_{e_2} \vec{v} = \left( \frac{\partial v^1}{\partial u^2} - \frac{1}{2} \sin(2u^1)v^2 \right) e_1 + \left( \frac{\partial v^2}{\partial u^2} + v^1 \cot(u^1) \right) e_2
\]

\[
= -\sin(u^2)e_1 + \cos(u^2)e_2.
\]

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

- The second vector field along the equator is

\[
\vec{v} = 0e_1 + 1e_2
\]

where \( u^1 = \frac{\pi}{2}, u^2 = \lambda \in [0, \frac{\pi}{2}] \).

\[
\text{Figure 25: Exponential and log map}
\]

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

\[
\nabla_{e_2} \vec{v} = \left( \frac{\partial v^1}{\partial u^2} - \frac{1}{2} \sin(2u^1)v^2 \right) e_1 + \left( \frac{\partial v^2}{\partial u^2} + v^1 \cot(u^1) \right) e_2
\]

\[
= (0 - 0)e_1 + (0 + 0)e_2 = \vec{0},
\]

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 23.** \[16\] In extrinsic case, we have to subtract the normal component, however, in intrinsic case, such normal component doesn’t exist, so we have

\[
\nabla_{\frac{\partial}{\partial u^i}} \vec{v} = \frac{\partial \vec{v}}{\partial u^i}
\]

\[
= \frac{\partial}{\partial u^i} \left( v^j e_j \right)
\]

\[
= \frac{\partial v^j}{\partial u^i} e_j + v^j \frac{\partial e_j}{\partial u^i}
\]

\[
= \frac{\partial v^k}{\partial u^i} e_k + v^j \Gamma^k_{ij} e_k
\]

\[
= \left( \frac{\partial v^k}{\partial u^i} + v^j \Gamma^k_{ij} \right) e_k
\]

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:

\[ g_{ij} = \frac{\partial}{\partial u^k} \cdot (\vec{e}_i \cdot \vec{e}_j) \]

We define

\[ \Gamma^l_{ik} (\vec{e}_l \cdot \vec{e}_j) + \Gamma^l_{jk} (\vec{e}_i \cdot \vec{e}_l) \]

Similarly, we can yield other two expressions:

\[ g_{ij} = \frac{\partial}{\partial u^k} \cdot (\vec{e}_i \cdot \vec{e}_j) \]

\[ g_{ki} = \frac{\partial}{\partial u^j} \cdot (\vec{e}_k \cdot \vec{e}_i) \]

\[ g_{jk} = \frac{\partial}{\partial u^i} \cdot (\vec{e}_j \cdot \vec{e}_k) \]

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

\[ 2 \Gamma^l_{kj} g_{il} \]

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

\[ 2 \Gamma^l_{kj} g_{il} g^{im} = \frac{1}{2} \left( \frac{\partial g_{ij}}{\partial u^k} + \frac{\partial g_{ki}}{\partial u^j} - \frac{\partial g_{jk}}{\partial u^i} \right) \]

\[ \Gamma^m_{kj} = \frac{1}{2} \frac{\partial g_{ij}}{\partial u^k} + \frac{\partial g_{ki}}{\partial u^j} - \frac{\partial g_{jk}}{\partial u^i} \]

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

\[ \Gamma^m_{kj} = \frac{1}{2} \frac{\partial g_{ij}}{\partial u^k} + \frac{\partial g_{ki}}{\partial u^j} - \frac{\partial g_{jk}}{\partial u^i} \]
Definition 3.28. **First fundamental form** on manifold $M$ is the field which assigns to each $p \in M$ the bilinear map

$$g_p(v, w) : T_p M \times T_p M \rightarrow \mathbb{R}$$

$$g_p(v, w) = \langle v, w \rangle$$

where $v, w \in T_p M$.

Definition 3.29. **Second fundamental form** on manifold $M$ is defined by

$$h_p(v, w) : T_p M \times T_p M \rightarrow T_p M$$

$$h_p(v, w) = (d\Pi(p)v)w = (d\Pi(p)w)v$$

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

Definition 3.30. **Riemannian exponential map** takes the position $p = \gamma(0) \in M$ and velocity $v = \dot{\gamma}(0) \in T_p M$ as input and returns the point at time 1 along the geodesic with these initial conditions. When $\gamma$ is defined over the interval $[0, 1]$, the Riemannian exponential map at $p$ is defined as

$$\text{Exp}_p(v) : T_p M \rightarrow M$$

$$\text{Exp}_p(v) = \text{Exp}(p, v) = \gamma(1)$$

**Example 24.** 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 $X \in g$, we have

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

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

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

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

Definition 3.31. **Riemannian log map** is the inverse of Riemannian exponential map, defined in the neighborhood $\text{Exp}_p(v)$

$$\text{Log}_p : M \rightarrow T_p M$$

$$\text{Log}_p(\gamma(1)) = v$$

**Remark 3.21.** The matrix logarithm of $M$ is defined as

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

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

40
Remark 3.22. According to the property above, we can also have

\[
\text{tr}(\log(M)) = \text{tr}(X^{-1} \log(D).X) \\
= \text{tr}(XX^{-1} \log(D)) \\
= \text{tr}(\log(D))
\]

Figure 26: Exponential and log map

Definition 3.32. **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 3.33. **Lie group** is a smooth manifold equipped with group structures, where the two group operations

\[
\begin{align*}
(x, y) &\rightarrow x \cdot y : G \times G \rightarrow G & \text{Multiplication} \\
x &\rightarrow x^{-1} : G \rightarrow G & \text{Inverse}
\end{align*}
\]

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

Definition 3.34. **Orbit** of a point \(p \in M\) is defined as

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

Example 25. If \(G = SO(2)\), the orbit of point \(p\) is a circle.

Definition 3.35. **Lie algebra** is a vector space \(\mathfrak{g}\) together with an operation called Lie bracket \([\cdot, \cdot]\), a alternating bilinear map \(\mathfrak{g} \times \mathfrak{g} \rightarrow \mathfrak{g}\). \(\forall x, y, z \in \mathfrak{g}\) 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 3.36.** Lie derivative of vector fields \([X, Y]\) is the derivative of \(Y\) along the flow generated by \(X\), and is sometimes denoted \(\mathcal{L}_X Y\) (“Lie derivative of \(Y\) along \(X\)”). This generalizes to the Lie derivative of any tensor field along the flow generated by \(X\).

\[ [X, Y](f) = X(Y(f)) - Y(X(f)) \quad \text{for all } f \in C^\infty(M). \]

**Remark 3.23.** 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.

**Example 26.** The example below shows how to calculate the flow curve, given the flow field.

Vector arrows tell you velocity at each point:

\[ \vec{w} = 1 \vec{e}_x + xe \vec{e}_y. \]

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

\[ \vec{w} = \frac{d\vec{R}}{d\lambda} = \frac{dx}{d\lambda} \frac{\partial \vec{R}}{\partial x} + \frac{dy}{d\lambda} \frac{\partial \vec{R}}{\partial y} = \frac{dx}{d\lambda} \vec{e}_x + \frac{dy}{d\lambda} \vec{e}_y \]

Figure 27: Coordinate lines

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

Figure 28: Flow field
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 27.** The Lie bracket of the vector field is defined below:

\[
[u, v] = \begin{pmatrix}
\frac{\partial}{\partial y}v^i \\
\frac{\partial}{\partial x}v^j
\end{pmatrix} - \begin{pmatrix}
\frac{\partial}{\partial y}u^i \\
\frac{\partial}{\partial x}u^j
\end{pmatrix}
\]

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

\[
\begin{align*}
\vec{u} &= 1\vec{e}_x \\
\vec{v} &= xe_y
\end{align*}
\]

**Figure 29: Two vector fields**

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

\[
\frac{d\vec{u}}{d\vec{v}} = v^i\partial_i(u^j\vec{e}_j) \\
= v^i\partial_i(u^j\partial_j) \\
= v^i[(\partial_iu^j)\partial_j + u^j(\partial_i\partial_j)] \quad \triangleright \text{product rule} \\
= v^i(\partial_iu^j)\partial_j + v^j u^i(\partial_i\partial_j) \\
= v^y(\partial_yu^x)\partial_x + v^x u^y(\partial_y\partial_x) \\
= x(\partial_y\partial_x) + x \cdot 1(\partial_y\partial_x) \\
= x(\partial_y\partial_x) \\
= x(\partial_y\vec{e}_x) \\
= x\frac{\partial}{\partial y}\vec{e}_x \\
= 0
\]
Derivative of $\vec{v}$ in the direction of $\vec{u}$ is shown below

\[
\vec{u}(\vec{v}) = u^i \partial_i (v^j \partial_j) \\
= u^i \partial_i (v^j \partial_j) \\
= u^i [(\partial_i v^j) \partial_j + v^j (\partial_i \partial_j)] \\
= u^i (\partial_i v^j) \partial_j + u^i v^j (\partial_i \partial_j) \\
= 1(\partial_x x) \partial_y + 1 \cdot x(\partial_x \partial_y) \\
= \partial_y + x(\partial_y \partial_x) \\
= \partial_y + x(\partial_y e_x) \\
= \partial_y + xe_y \\
= e_y
\]

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

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

**Definition 3.37.** Lie derivative of tensor fields $T$ w.r.t. smooth vector field $X$ is defined by

\[
(L_X T)_p = \frac{d}{dt} \bigg|_{t=0} (\phi_t^* T)_p = \lim_{t \to 0} \frac{\phi_t^* (T_{\phi_t(p)}) - T_p}{t},
\]

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 $\phi$

**Remark 3.24.** Intuitively, if you have a tensor field $T$ and a vector field $X$, then $L_X T$ 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 $X$. 

Figure 30: Separation vector
Definition 3.38. **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 \( \phi \) at \( x \) is an isomorphism of tangent space \( D\phi_x : T_x M \to T_{\phi(x)} N \)

2. \( \forall v, w \in T_x M \), the Riemannian metric preserves as \( \langle v, w \rangle = \langle D\phi_x \cdot v, D\phi_x \cdot w \rangle \)

Example 28. 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 \( X_p, Y_p \in T_p S \) and all \( p \in S \), we have

\[ \langle X_p, Y_p \rangle_g = \langle D\phi_p \cdot X_p, D\phi_p \cdot Y_p \rangle_{g'} \]

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

Definition 3.39. **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 3.40. **Conformality** \( \phi : S_1 \to S_2 \) is a function that for all \( X, Y \in T_p M \), 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 3.25. 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 3.41. **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 3.42. **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 \( \phi \) if \( \phi(x) = x \).

Definition 3.43. **Automorphism group** \( \Psi : G \to G \) is defined as

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

given \( \forall g, h \in G \).

Definition 3.44. **Inner automorphism group** \( \text{Inn}(G) \) is the collection of all inner automorphisms of the form \( \Psi_g, \forall g \in G \). \( \text{Inn}(G) \) is a Lie group and commutative.

Definition 3.45. **Dual pairing** \( (m, v) \), where \( m \in V^* \), the dual space to \( V \), and \( v \in V \)
**Definition 3.46.** [Adjoint action \( \text{Ad}_g \)] is the derivative of \( \Psi_g(h) \) with respect to \( h \) at the identity, which is

\[
\text{Ad}_g : G \times g \rightarrow g \\
\text{Ad}_g = d(\Psi_g)_e
\]

- For matrix group, \( \text{Ad}_g \) is derived as

\[
\text{Ad}_g w = \frac{\partial}{\partial \xi} \Psi_g(h) \\
= \frac{\partial}{\partial \xi} \Psi_g(h|\xi=0) \\
= \frac{\partial}{\partial \xi} (g(h|\xi=0)g^{-1}) \\
= g \left( \frac{\partial}{\partial \xi} h|\xi=0 \right) g^{-1} \\
= gwg^{-1}
\]

where \( h_\xi \) denotes the variation of \( h \) by \( \xi \) 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 \( \text{Ad}_g w = gwg^{-1} \), we have

\[
(m, \text{Ad}_g w) = (m, gwg^{-1}) \\
= (g^Tm, wg^{-1}) \\
= (g^Tmg^{-T}, w)
\]

\[
\text{Ad}_g^*m = g^Tmg^{-T},
\]

as if \( A \) is a linear operator from \( V \) to \( V \), its conjugate \( A^* : V^* \rightarrow V^* \) is defined by \( (A^*m, v) = (m, Av) \). For detail, see [5].

- For \( \text{Diff}(\Omega) \), \( \text{Ad}_\phi \) is derived as

\[
\text{Ad}_\phi w = \frac{\partial}{\partial \xi} (\Psi_\phi h_\xi)|\xi=0 \\
= \frac{\partial}{\partial \xi} (\phi \circ h_\xi \circ \phi^{-1})|\xi=0 \\
= D\phi|_{h_\xi \circ \phi^{-1}} w|_{\phi^{-1}} \\
= (D\phi \circ \phi^{-1})w \circ \phi^{-1}
\]

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

**Example 29.** Adjoint of the gradient is the negative divergence:

\[
\langle \nabla f, g \rangle = -\langle f, \text{div} g \rangle
\]
Definition 3.47. **Infinitesimal adjoint action** \( \text{ad} \) is the derivative of the adjoint map \( \text{Ad} \) with respect to \( g \) at identity, which is

\[
\text{ad} : g \times g \to g \\
\text{ad} = d(\text{Ad})_e
\]

- For matrix group, \( \text{ad}_g \) is derived as

\[
\text{ad}_g w = \frac{\partial}{\partial \xi} \text{Ad}_{g\xi} w |_{\xi=0} \\
= \frac{\partial}{\partial \xi} (g\xi wg_{\xi}^{-1}) |_{\xi=0} \\
= (\frac{\partial}{\partial \xi} g\xi wg_{\xi}^{-1}) |_{\xi=0} + (g\xi w \frac{\partial}{\partial \xi} g_{\xi}^{-1}) |_{\xi=0} \\
= vw - (g\xi wg_{\xi}^{-1} \frac{\partial}{\partial \xi} g_{\xi}^{-1}) |_{\xi=0} \\
= vw - wv
\]

where \( g_{\xi} \) is the variation of \( g \) by \( \xi \) with \( g_0 = e \) and \( \frac{\partial}{\partial \xi} g_{\xi} |_{\xi=0} = v \) for \( \text{Ad}_{g\xi} w = g\xi wg_{\xi}^{-1} \).

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

\[
(m, \text{ad}_v w) = (m, vw - wv) \\
= (m, vw) - (m, wv) \\
= (v^T m, w) - (mv^T, w) \\
= (v^T m - mv^T, w)
\]

\( \text{ad}_v^* m = v^T m - mv^T \),

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 [5].

- For \( \text{Diff}(\Omega) \), \( \text{ad}_v \) is derived as

\[
\text{ad}_v w = \frac{\partial}{\partial \xi} ((D\phi_\xi \circ \phi_\xi^{-1})w \circ \phi_\xi^{-1}) |_{\xi=0} \\
= \frac{\partial}{\partial \xi} ((D\phi_\xi \circ \phi_\xi^{-1})w + D\text{Id} \frac{\partial}{\partial \xi} [w \circ \phi_\xi^{-1}] |_{\xi=0} \\
= \left( \frac{\partial}{\partial \xi} D\phi_\xi |_{\xi=0} + D D\phi_\xi |_{\xi=0} D\phi_\xi^{-1} \right) w \circ \phi_\xi^{-1} + D v \frac{\partial}{\partial \xi} \phi_\xi^{-1} |_{\xi=0} \\
= (Dv + 0)w - Dvw \\
= Dvw - Dvw
\]

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

Definition 3.48. **Left/Right multiplication** is a diffeomorphism such that

\[
L_y : x \to y \cdot x \\
R_y : x \to x \cdot y
\]

where \( y \in G \), \( G \) is a Lie group.
Definition 3.49. **Left/Right-invariant** means \( \forall y \in G \), we have \( L_y X = X \) or \( R_y X = X \).

Example 30. The metric \( G^I \) has the property of right-invariance: if \( U, V \in T_\phi \text{Diff}(M) \) then

\[
G^I_\phi(U, V) = G^I_{\phi \circ \psi}(U \circ \psi, V \circ \psi) \quad \forall \psi \in \text{Diff}(M)
\]

Definition 3.50. **Inertia operator** \( L : g \to g^* \) is defined by

\[
\langle v, w \rangle = (Lv, w), \forall v, w \in g
\]

\( L \) must be invertible and

\[
(Lv, w) = (Lw, v), \forall v, w \in g
\]

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

Definition 3.51. A linear operator \( f : g \to g \) is **transposed** with respect to the inner product defined by \( L \), using the formula

\[
\langle f^t v, w \rangle = \langle v, fw \rangle, \forall v, w \in g
\]

We use this to define the adjoint-transpose action \( \text{Ad}^t : G \times g \to g \) via the transpose of \( \text{Ad}_g \) and the infinitesimal adjoint-transpose \( \text{ad}^t : g \times g \to g \) via the transpose of \( \text{ad}_v \).

Definition 3.52. **Information metric** \( G^I \) is defined by

\[
G^I_g(U, V) = -\int_M \langle \Delta u, v \rangle_g \text{vol} + \lambda \sum_{i=1}^k \left( \int_M \langle u, \xi_i \rangle_g \text{vol} \cdot \int_M \langle v, \xi_i \rangle_g \text{vol} \right),
\]

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

Definition 3.53. **Fisher-Rao metric** is the Riemannian metric on \( \text{Dens}(M) \) given by

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

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

Definition 3.54. The background metric \( g \) on manifold \( M \) is called **compatible with** \( \mu \) if \( \text{vol}_g = \mu \), for \( \mu \in \text{Dens}(M) \).

Definition 3.55. **Set of vector space isomorphisms** \( \mathcal{L}_{\text{iso}} \) is defined by

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

Definition 3.56. **Frame** of \( m \)-dimension real vector space \( V \) is the basis \( e_1, \ldots, e_m \) of \( V \).

Definition 3.57. **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 \( \mathcal{F}(M)_p = \mathcal{L}_{\text{iso}}(\mathbb{R}^m, T_p M) \) is the space of frames of tangent space at \( p \).
Definition 3.58. A smooth curve $\beta : \mathbb{R} \to F(M)$ is called a lift of smooth curve $\gamma : \mathbb{R} \to M$ if $\pi \circ \beta = \gamma$.

Example 31. The general linear group $GL(m, \mathbb{R})$ acts on this space by composition on the right via

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

Definition 3.59. A curve $\beta(t) = (\gamma(t), e(t)) \in F(M)$ is called horizontal lift of $\gamma$ if the vector field $X(t) = e(t)\xi$ along $\gamma$ is parallel for every $\xi \in \mathbb{R}^m$. Thus a horizontal lift of $\gamma$ has the form

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

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

Definition 3.60. 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 pushforward tangent vector of $v$ induced by $\phi$.

Definition 3.61. Pullback and pushforward are defined as below

Pullback(right action): $\varphi^* \rho = |D\varphi| \rho(\varphi(\cdot))$

$$= |D\varphi| \rho \circ \varphi$$

Pushforward(left action): $\varphi_* \rho = (\varphi^{-1})^* \rho$

$$= |D\varphi^{-1}| \rho(\varphi^{-1}(\cdot))$$

$$= |D\varphi^{-1}| \rho \circ \varphi^{-1}$$

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

Remark 3.26. Given a $\phi$, there can be two effects or two directions of the $\phi$. The pullback is the one from distorted to checkered, while the pushforward is the one from checkered to distorted, which is more intuitive.

Remark 3.27. 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.
Remark 3.28. **Relationship between** $I_0, I_1, \phi, \phi^{-1}$: Intuitively, we may think that $\phi$ 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 $\phi^{-1}$, instead of the more intuitively-correct $\phi$. 
Remark 3.29. “Blank Space” You may have the confusion that if we deform the image using the diffeomorphism $\varphi$ 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 $I \circ \varphi^{-1} : \mathbb{Z}^2 \rightarrow \mathbb{R}^2 \rightarrow \mathbb{R}$. 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, $I \circ \varphi^{-1} : \mathbb{Z}^2 \rightarrow \mathbb{R}$ 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 $\phi : \mathbb{Z}^2 \rightarrow \mathbb{R}^2$, 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 34: $\varphi$
4 Statistics for Images

**Definition 4.1.** **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 \( e_1, \ldots, e_d \) of \( T_\mu M \). Then, these vectors are used to form a sequence of nested subspaces \( V_k = \text{span}(\{e_1, \ldots, e_k\}) \cap U \), where \( U \subset T_\mu M \) is a neighbourhood of 0, such that projection is well-defined for all geodesic submanifolds of \( \text{Exp}_\mu(U) \).

The principal geodesic submanifolds are given by

\[
H_k = \text{Exp}_\mu(V_k)
\]

The first principal direction is now chosen to maximize the projected variance along the corresponding geodesic:

\[
e_1 = \arg \max_{\|e\|=1} \sum_{i=1}^{n} \| \log_\mu(\pi_{H}(x_i)) \|^2, \quad \text{where } H = \text{Exp}_\mu(\text{span}(\{e\}) \cap U)
\]

\[
e_k = \arg \max_{\|e\|=1} \sum_{i=1}^{n} \| \log_\mu(\pi_{H}(x_i)) \|^2, \quad \text{where } H = \text{Exp}_\mu(\text{span}(\{e_1, \ldots, e_{k-1}, e\}) \cap U)
\]

where we define the projection operator \( \pi_H : M \to H \) as

\[
\pi_H(x) = \arg \min_{y \in H} d(x, y)^2
= \arg \min_{y \in H} \| \log_\mu(y) \|^2
\approx \arg \min_{y \in H} \| \log_\mu(x) - \log_\mu(y) \|^2
\]

- **Definition 2.** The intent of principal geodesic analysis is to find an orthonormal basis \( \{e_1, \ldots, e_k\} \) of a set of points \( \{x_1, \ldots, x_n\} \in \mathbb{R}^d \), which satisfies the recursive relationship

\[
e_1 = \arg \min_{\|e\|=1} \sum_{i=1}^{n} \| x_i - \langle e, x_i \rangle e \|^2
\]

\[
e_2 = \arg \min_{\|e\|=1} \sum_{i=1}^{n} \| x_i - \langle e_1, x_i \rangle e_1 - \langle e, x_i \rangle e \|^2
\]

\[ \vdots \]

\[
e_k = \arg \min_{\|e\|=1} \sum_{i=1}^{n} \| x_i - \langle e_1, x_i \rangle e_1 - \cdots - \langle e_{k-1}, x_i \rangle e_{k-1} - \langle e, x_i \rangle e \|^2
\]

In other words, the subspace \( V_k = \text{span}(\{e_1, \ldots, e_k\}) \) is the \( k \)-dimensional subspace that minimizes the sum-of-squared distances to the data. The principal geodesic submanifolds are given by

\[
H_k = \text{Exp}_\mu(V_k)
\]
The first principal direction is now chosen to minimize the sum-of-squared distance of the data to the corresponding geodesic:

\[
e_1 = \arg \max_{\|e\|=1} \sum_{i=1}^{n} \| \log_{x_i}(\pi_H(x_i)) \|^2, \quad \text{where } H = \text{Exp}_\mu(\text{span}(\{e\}) \cap U)
\]

\[
e_k = \arg \max_{\|e\|=1} \sum_{i=1}^{n} \| \log_{x_i}(\pi_H(x_i)) \|^2, \quad \text{where } H = \text{Exp}_\mu(\text{span}(\{e_1, \cdots, e_{k-1}, e\}) \cap U)
\]

Example 32.

**Algorithm 1 Principal Geodesic Analysis**

**Inputs:** \(x_1, \cdots, x_n \in M\)

\(\mu \leftarrow \text{intrinsic mean of } \{x_i\}\)

\(u_i \leftarrow \log_{\mu}(x_i)\) \hfill \triangleright \text{Remark 4.6}

\(\Sigma \leftarrow \frac{1}{n} \sum_{i=1}^{n} u_i u_i^T\)

\(\{e_k, \lambda_k\} \leftarrow \text{eigenvectors and eigenvalues of } \Sigma\)

**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.** 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 33.** 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. [Fréchet mean and variance] are defined as

\[ p_* = \arg\min_p \phi(p), \quad \triangledown \text{Fréchet mean} \]

\[ \phi(p) = \sum_{i=1}^N d^2(p, x_i) w_i, \quad \triangledown \text{Fréchet variance} \]

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 \(\phi\), while the Fréchet means are then those points, which globally minimize \(\phi\).

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 [21], Karcher[20] shows that the gradient of \(\phi\) above is

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

If \(w_i = \frac{1}{2N}\), then we have

\[ \phi(p) = \frac{1}{2N} \sum_{i=1}^N d^2(p, x_i), \]

\[ \nabla \phi(p) = -\frac{1}{N} \sum_{i=1}^N \text{Log}_p(x_i). \]

The gradient descent algorithm takes successive steps in the negative gradient direction. Given a current estimate \(\mu_j\), 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

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

where \(\tau\) 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 \(\mu_j\), 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. [Fréchet median] is defined as

\[ p_* = \arg\min_p \psi(p), \quad \triangledown \text{Fréchet median} \]

\[ \psi(p) = \sum_{i=1}^N |d^2(p, x_i)|, \]
where \( x_1, x_2, \cdots, x_n \) are the random points in \( M \) and \( p \in M \).

**Definition 4.5.** **Weighted geometric median** is defined as

\[
p_* = \arg \min \psi(p),
\]

\[
\psi(p) = \sum_{i=1}^{N} d^2(p, x_i)w_i,
\]

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

**Definition 4.6.** **Fréchet distance** between \( A \) and \( B \) is defined as infimum over all reparameterizations \( \alpha \) and \( \beta \) of the maximum over all \( t \in [0, 1] \) of the distance between \( A(\alpha(t)) \) and \( B(\beta(t)) \).

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

where \( d \) is the distance function, e.g. Euclidean distance.

**Example 34.** 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 \( \alpha, \beta \)), which can be exhibited as follows:

\[
A(\cdot) = [A(0), A(1), A(2), \cdots],
\]

\[
B(\cdot) = [B(0), B(1), B(2), \cdots],
\]

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 \( \alpha, \beta \). And for each combination of \( \alpha, \beta \), 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 maximum distance during this travel is minimized. The satisfied moving patterns are often making the two points moving forward “simultaneously".
Figure 35: The Fréchet distance between two same shape curves is the norm of the translation vector.

Definition 4.7. Wasserstein distance is defined as

$$W_p(\mu, \nu) = \left( \inf_{\gamma \in \Gamma(\mu, \nu)} \int_{M \times M} \text{dist}(x, y)^p \, d\gamma(x, y) \right)^{1/p},$$

$$W_2(\mu, \nu) = \left( \inf_{\gamma \in \Gamma(\mu, \nu)} \int_{M \times M} \text{dist}(x, y)^2 \, d\gamma(x, y) \right)^{1/2},$$

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

In discrete case, the distance reads as

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

where $M_{\mu\nu} = [\text{dist}(x_i, y_j)^p]_{ij} \in \mathbb{R}^{m \times n}$ and $\Gamma(\mu, \nu) = \{ T \in \mathbb{R}^{m \times n} | T \mathbf{1}_m = a, T^T \mathbf{1}_n = b \}$, 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 $\text{dist}(x, y)$ and $\gamma(x, y)$ in continuous form.

Figure 36: 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 \( c(x, y) \rightarrow [0, +\infty) \) 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

\[
\int \int c(x, y) \gamma(x, y) dx\;dy,
\]

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

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

\[
W_p(\mu, \nu) = \left( \int_{-\infty}^{+\infty} |U(s) - V(s)|^p ds \right)^{1/p} = \left( \int_{0}^{1} |U^{-1}(t) - V^{-1}(t)|^p dt \right)^{1/p},
\]

where \( U \) and \( V \) are the CDFs of \( \mu, \nu \) respectively.

**Proof.** Assuming \( \{(x_1, y_1), (x_2, y_2)\} \subset (\gamma^*)^p, x_1 < x_2 \), where \( \gamma^* \) 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:

\[
|x_1 - y_2|^p + |x_2 - y_1|^p < |x_1 - y_1|^p + |x_2 - y_2|^p
\]

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 \( \gamma^* \), as it indicates that \( \gamma^* \) is no cyclically monotone.

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

\[
\gamma^*([0, \infty], (-\infty, x]) = \gamma^*((-\infty, x], (-\infty, y]) = \gamma^*((-\infty, x], \mathbb{R})
\]

In turn, the fact that \( \gamma^* \in \Gamma(\mu, \nu) \) implies that \( \gamma^*((-\infty, x], \mathbb{R}) = F(x) \) and \( \gamma^*([0, \infty], (-\infty, y]) = G(y) \). From previous relation, we conclude that

\[
W_p(\mu, \nu) = \left( \inf_{\gamma \in \Gamma(\mu, \nu)} \int \int_{M \times M} \text{dist}(x, y)^p \gamma(x, y) dx\;dy \right)^{1/p} = \left( \int_{0}^{1} |F^{-1}(t) - G^{-1}(t)|^p dt \right)^{1/p}
\]

**Example 35.** For one dimensional discrete case, to transport \( \nu \) to \( \mu \),

\footnote{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.}

\footnote{For a more detailed deviation of this inequality in the case of \( p > 1 \), refers to Appendix A in \url{https://arxiv.org/pdf/1509.02237.pdf}}
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 37: Two distribution $\mu$ and $\nu$.

Figure 38: Optimal transport measures the minimal effort required for filling $-\mu_1$ with $\mu_0$, 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**

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

![Figure 39: Top: Some random circles. Bottom left: Euclidean average of the circles. Bottom right: Wasserstein barycenter.](https://www.stat.cmu.edu/~larry/=sml/Opt.pdf)

- 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 40: 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 $P_1$.](https://people.csail.mit.edu/eddchien/presentations/Stochastic_Wasserstein_Barycenters.pdf)

- Wasserstein distance is insensitive to small wiggles.

**Definition 4.8.** 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 $\mu_i$ is the sample distribution. For computation in discrete situation, please refer to [https://arxiv.org/pdf/1310.4375.pdf](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.

---

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.

After withdrawing the common factor, we have

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

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

<table>
<thead>
<tr>
<th>Transformation</th>
<th>Matrix</th>
<th>#DoF</th>
<th>Preserves</th>
</tr>
</thead>
</table>
| Translation    | \[
\begin{pmatrix}
1 & 0 & t_1 \\
0 & 1 & t_2 \\
0 & 0 & 1 \\
s_1 & 0 & 0
\end{pmatrix}
\] | 2 orientation |
| Scaling        | \[
\begin{pmatrix}
0 & s_2 & 0 \\
0 & 0 & 1 \\
1 & s_1 & 0
\end{pmatrix}
\] | 2 orientation |
| Shearing       | \[
\begin{pmatrix}
s_2 & 1 & 0 \\
0 & 0 & 1
\end{pmatrix}
\] | 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,

\[
\begin{pmatrix}
1 & 0 & t_1 \\
0 & 1 & t_2 \\
0 & 0 & 1
\end{pmatrix} \cdot \begin{pmatrix}
\cos(\theta) & -\sin(\theta) & 0 \\
\sin(\theta) & \cos(\theta) & 0 \\
0 & 0 & 1
\end{pmatrix} \cdot \begin{pmatrix}
s_1 & 0 & 0 \\
0 & s_2 & 0 \\
0 & 0 & 1
\end{pmatrix} = \begin{pmatrix}
\cos(\theta) & -\sin(\theta) & s_1 \cos(\theta) - s_2 \sin(\theta) \\
\sin(\theta) & \cos(\theta) & s_2 \cos(\theta) + s_1 \sin(\theta) \\
0 & 0 & 1
\end{pmatrix}
\]
especially when there is a translation operation.

\[
\begin{pmatrix}
\cos(\theta) & -\sin(\theta) & 0 \\
\sin(\theta) & \cos(\theta) & 0 \\
0 & 0 & 1
\end{pmatrix}
\begin{pmatrix}
s_1 & 0 & 0 \\
0 & s_2 & 0 \\
0 & 0 & 1
\end{pmatrix}
\begin{pmatrix}
1 & 0 & t_1 \\
0 & 1 & t_2 \\
0 & 0 & 1
\end{pmatrix}
\]

\[
= \begin{pmatrix}
s_1 \cos(\theta) & -\sin(\theta) & 0 \\
s_1 \sin(\theta) & s_2 \cos(\theta) & 0 \\
0 & 0 & 1
\end{pmatrix}
\begin{pmatrix}
1 & 0 & t_1 \\
0 & 1 & t_2 \\
0 & 0 & 1
\end{pmatrix}
\]

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. **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:

\[
\frac{\partial}{\partial x}(b^T x) = \frac{\partial}{\partial x}(x^T b) = b
\]

\[
\frac{\partial}{\partial x}(x^T x) = 2x
\]

\[
\frac{\partial}{\partial x}(x^T Ax) = 2A
\]

\[
\frac{\partial}{\partial x} \|Ax - b\|_2^2 = 2A^T(Ax - b)
\]

\[
\frac{\partial}{\partial x} \|Ax - b\|_2 = \frac{A^T(Ax - b)}{\|Ax - b\|_2}
\]
For $f(X): \mathbb{R}^{n \times m} \to \mathbb{R}, f(X)' \in \mathbb{R}^{n \times m}$, we can have:

$$\frac{\partial}{\partial X}(a^T X b) = ab^T$$
$$\frac{\partial}{\partial X}(a^T X^T b) = ba^T$$
$$\frac{\partial}{\partial X} \text{tr}(A^T X B) = AB^T$$
$$\frac{\partial}{\partial X} \text{tr}(A^T X^T B) = BA^T$$
$$\frac{\partial}{\partial X} |X| = |X|(X^{-1})^T$$

6 Domain Knowledge for Medical Imaging

6.1 Medical Image Formats

Definition 6.1. **Neuroimaging Informatics Technology Initiative (NIfTI)** is an open file format commonly used to store brain imaging data obtained using Magnetic Resonance Imaging methods.

Example 36. The filename extension of NIfTI includes `.nia`, `.nii`, `.nii.gz`, `.hdr`.

Definition 6.2. **Digital Imaging and Communications in Medicine (DICOM)** the international standard to transmit, store, retrieve, print, process, and display medical imaging information.

Example 37. Below is an example snippet of DICOM metadata:
**Definition 6.3.** Nearly Raw Raster Data (NRRD) is a library and file format for the representation and processing of n-dimensional raster data. It was developed by Gordon Kindlmann from University of Utah.

**Example 38.** The filename extension of NRRD includes .nrrd, .nhdr. Below is an example of NRRD file:

```
NRRD0004
# Complete NRRD file format specification at:
# http://teem.sourceforge.net/nrrd/format.html
type: float
dimension: 3
space: left-posterior-superior
sizes: 145 174 145
```
Remark 6.1. While DICOM is very complicated, both NRRD and NIfTI are very simple. NIfTI is very explicit, with a fixed 348-byte binary header. NRRD is text based, so it is human readable. It is easy to write, as you can easily describe whatever order you want to your dimensions. The trade off is that it is harder to create a reader, as you need to juggle the image dimensions. What is elegant about NRRD is that the flexibility of the header allows you to create a tiny header that describes a complicated image in a different format.

6.2 MRI modalities

<table>
<thead>
<tr>
<th>Tissue</th>
<th>T1-Weighted</th>
<th>T2-Weighted</th>
<th>Flair</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSF</td>
<td>Dark</td>
<td>Bright</td>
<td>Dark</td>
</tr>
<tr>
<td>White Matter</td>
<td>Light</td>
<td>Dark Gray</td>
<td>Dark Gray</td>
</tr>
<tr>
<td>Cortex</td>
<td>Gray</td>
<td>Light Gray</td>
<td>Light Gray</td>
</tr>
<tr>
<td>Fat</td>
<td>Bright</td>
<td>Light</td>
<td>Light</td>
</tr>
<tr>
<td>Inflammation</td>
<td>Dark</td>
<td>Bright</td>
<td>Bright</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>T1-Weighted (short TR and TE)</th>
<th>500</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td>T2-Weighted (long TR and TE)</td>
<td>4000</td>
<td>90</td>
</tr>
<tr>
<td>Flair (very long TR and TE)</td>
<td>9000</td>
<td>114</td>
</tr>
</tbody>
</table>
References


