INTRODUCTION TO THE THEORY OF GEODESICS

Applications to Bioinformatics

Sorin V. SABAU

Tokai University, Department of Biological Sciences, Sapporo Campus

The final version of this presentation can be downloaded from
http://sorindb.sap.u-tokai.ac.jp/pub/Sabau_LN.pdf

Introduction

Abstract

In this lecture we will show that starting from the notions of metric or distance and length of curves we are naturally led to the very basic notions of modern geometry.

The theory can be applied in bioinformatics, computer science and other fields of natural sciences.

Introduction

Bioinformatics

Applications of methods from computer science, mathematics, engineering to the management and analysis of biological data.

- Genomics
- Proteomics

Biological Sequences Comparison

- evolution of life on Earth; tree of life
- infer functions of genes and proteins

Metric spaces

Metric spaces

Definition 1 : Let X be an arbitrary set of points, then the function $d : X \times X \to [0, \infty)$ is called a *metric* on X if the following conditions are satisfied for any $x, y, z \in X$

1. Positiveness: $d(x, y) > 0$ if $x \neq y$ and $d(x, x) = 0$;
2. Symmetry: $d(x, y) = d(y, x)$;
3. Triangle inequality: $d(x, z) \leq d(x, y) + d(y, z)$.

A metric space is denoted by (X, d), and $d(x, y)$ is also called the *distance* between the points x and y, sometimes denoted by $||xy||$.

Example

On an arbitrary set X we can define the metric

$$\|xy\| := \begin{cases} 0, & \text{if } x = y \\ 1, & \text{if } x \neq y. \end{cases}$$

This is called the **discrete metric**.

Example

The real line $\mathbb{R}$ equipped with the distance $\|xy\| := |x - y|$ is a metric space. On $\mathbb{R}$ we can induce many metric function, for example:

$$d_{\log}(x, y) := \log(|x - y| + 1).$$

Example

The Euclidean plane $\mathbb{R}^2$ with the **standard Euclidean distance**

$$\|xy\| = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2},$$

for any $x = (x_1, x_2), y = (y_1, y_2) \in \mathbb{R}^2$.

Example

Recall that $\mathbb{R}^n = \{(x_1, \ldots, x_n) : x_i \in \mathbb{R}\}$ is a vector space with a operations

$$x + y = (x_1, \ldots, x_n) + (y_1, \ldots, y_n) = (x_1 + y_1, \ldots, x_n + y_n)$$

$$k \cdot x = k \cdot (x_1, \ldots, x_n) = (kx_1, \ldots, ks_n)$$

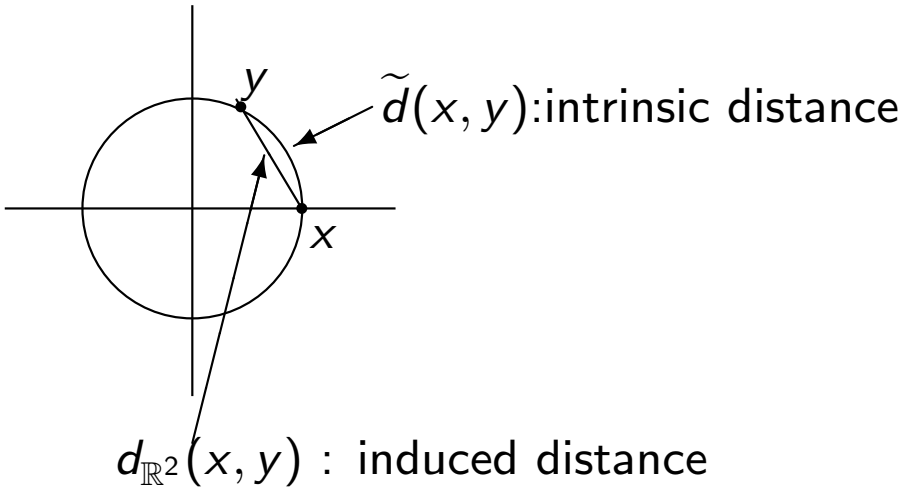
for any $x, y \in \mathbb{R}^n, k \in \mathbb{R}$. Here $\mathbb{R}^n := \underbrace{\mathbb{R} \times \cdots \times \mathbb{R}}_{n \text{ times}}$. Then $\mathbb{R}^n$ can be endowed with the standard Euclidean distance

$$\|xy\| = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}.$$

Example

Consider the unit circle $\mathbb{S}^1 = \{x \in \mathbb{R}^2 : d(0,x) = 1\}$. Then we have two kinds of distance:

1. the **induced distance** $(\mathbb{S}^1, d_{\mathbb{S}^1}) \subset (\mathbb{R}^2, d_{\mathbb{R}^2})$, where $d_{\mathbb{S}^1} := d_{\mathbb{R}^2}|_{\mathbb{S}^1}$, i.e. the Euclidean distance in plane restricted to $\mathbb{S}^1$.
2. **the intrinsic distance** $(\mathbb{S}^1, \widetilde{d})$, where $\widetilde{d}(x,y) =$ the arc length between x and y. Obviously, for $\forall x, y \in \mathbb{S}^1, x \neq y$, we have $\widetilde{d}(x,y) > d(x,y)$.



Namely, $d_{\mathbb{S}^1}(x,y) = \sqrt{(x_1-y_1)^2+(x_2-y_2)^2}$ and $\widetilde{d}(x,y) = \measuredangle(x,y)$.

Normed space

Normed space

Definition 2 : Let $(V, +, \cdot)$ be a vector space. A function $||\cdot|| : V \to [0, \infty)$ is called a *norm* on V if the following conditions are satisfied for any $v, w \in V$, $k \in \mathbb{R}$

1. Positiveness: $||v|| > 0$ if $v \neq 0$, $||0|| = 0$;
2. Absolute homogeneity: $||k \cdot v|| = |k| \cdot ||v||$;
3. Triangle inequality: $||v + w|| \leq ||v|| + ||w||$.

   The pair $(V, ||\cdot||)$ is called a *normed space*. Finite dimensional normed spaces are also called (absolute homogeneous) **Minkowski spaces**.

Remark

If $(V, \|\cdot\|)$ is a normed space, then $d(v, w) := \|v - w\|$, for $\forall v, w \in V$, is a distance function, or a metric on V.

Example

The Euclidean space $\mathbb{R}^n$ has the norm

$$\|x\| = \|(x^1, \cdots, x^n)\| = \sqrt{(x^1)^2 + \cdots + (x^n)^2}, \text{ for } \forall x = (x^1, \ldots, x^n) \in \mathbb{R}^n.$$

Example

$\mathbb{R}^n_1$ is the normed space $(\mathbb{R}^n, \|\cdot\|_1)$ with the norm

$$\|(x^1, \ldots, x^n)\|_1 = |x_1| + \cdots + |x_n|.$$

Example

Similarly, $\mathbb{R}^n_\infty$ is the normed space $(\mathbb{R}^n, \|\cdot\|_\infty)$ with the norm $\|(x^1, \ldots, x^n)\|_\infty = \max\{|x^1|, \ldots, |x^n|\}$.

Example

Let X be any set. The space $l_\infty(x)$ is the space of all bounded functions $f : X \to \mathbb{R}$, i.e. $l_\infty(x) = \{f : X \to \mathbb{R}, |f(x)| < M\}$. The standard norm $\|\cdot\|_\infty$ on $l_\infty(x)$ is

$$\|f\|_\infty = \sup_{x \in X} |f(x)|.$$

Euclidean Spaces

Definition

Let $(X, +, \cdot)$ be a vector space.

1. A **bilinear form** on X is a map $F : X \times X \to \mathbb{R}$ linear in both arguments, i.e.

$$F(ax_1 + bx_2, y) = a \cdot F(x_1, y) + b \cdot F(x_2, y)$$
$$F(x, ay_1 + by_2) = a \cdot F(x, y_1) + b \cdot F(x, y_2),$$

   for $\forall x, x_1, x_2, y, y_1, y_2 \in X, \ a, b \in \mathbb{R}$.

2. A bilinear form is called **symmetric** if $F(x, y) = F(y, x)$, for $\forall x, y \in X$.

Remark

1. If F is a symmetric bilinear form, then $Q(x) := Q_F(x) = F(x, x)$ is the **associated quadratic form**.
2. If Q is a quadratic form on X, then

$$F(x, y) = \frac{1}{4}[Q(x + y) - Q(x - y)]$$

   is a symmetric bilinear form.

Definition

A **scalar product** is a symmetric bilinear form F whose associate quadratic form is positive definite, i.e. $F(x,x) > 0$ for $\forall x \in X \setminus \{0\}$.

Definition

A **norm** associated with a scalar product $< \cdot, \cdot >$ is

$$\|v\| = \sqrt{< v, v >}, \forall v \in X.$$

Example

The **standard norm in** $\mathbb{R}^n$, i.e. $\|x\| = \sqrt{(x^1)^2 + \cdots + (x^n)^2}$, is associated with the **standard scalar product**

$$< x, y >= \sum_{i=1}^{n} x^i y^i,$$

for $\forall x = (x^1, \ldots, x^n)$, $y = (y^1, \ldots, y^n) \in \mathbb{R}^n$.

Example

Consider the sphere $\mathbb{S}^n := \{x \in \mathbb{R}^{n+1} : \|x\| = 1\} \subset \mathbb{R}^{n+1}$. Then

1. the angular metric on $\mathbb{S}^n$ is

$$d(x, y) = \arccos < x, y >,$$

   i.e. the Euclidean angle of vectors x, y.

2. another metric on $\mathbb{S}^n$ is

$$d(x, y) = 2 \cdot \arcsin \frac{\|xy\|}{2}.$$

Length structures

Question.

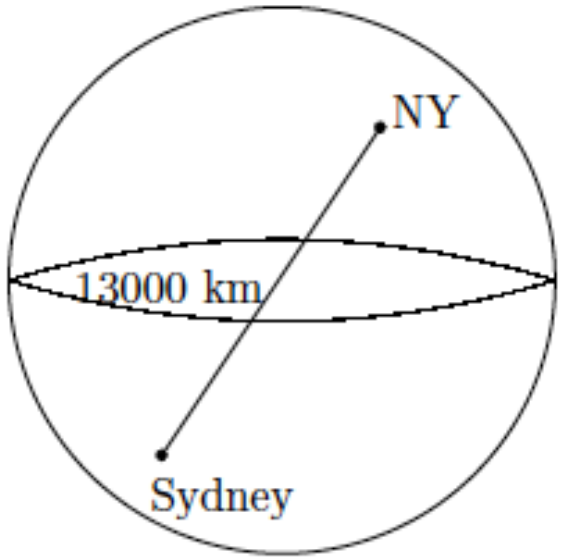
How to measure the distance between two peaks of mountain?

A possible answer : 10 km in straight line. This answer is also useless for an alpinist because this distance is a "crow fly" and no human can travel this way.

Question

"What is the distance between New-York and Sydney?"

A possible answer : About 8000 miles (=13000 km) in a straight line. However, this answer is useless because we cannot dig a tunnel in the Earth to go from NY to Sydney. This distance is obtained by using the induced metric from $\mathbb{R}^3$ on the Earth modeled as $\mathbb{S}^2 \subset \mathbb{R}^3$.



Mathematical conclusion : It is better to start with the **length of paths** and only after that define the distances.

Remark

For example, in $\mathbb{R}^2$, we can define

1. the Euclidean distance, i.e. **length of straight lines**;
2. the distance measured along the **shortest path** between two points exists, this path is called **a geodesic** of the space.

Idea

We need to define a **class of admissible paths** for which we can effectively measure lengths. (For example : the airplane route from NY to Sydney, or the traveler's path between peaks of two mountains.)

Length function

Definition (Length function)

A *path* in a space X is a continuous mapping $\gamma : [a, b] \to X$. Let A be the class of admissible paths in X, that is , the subset of all continuous paths in X, such that each $\gamma \in A$ is measurable.

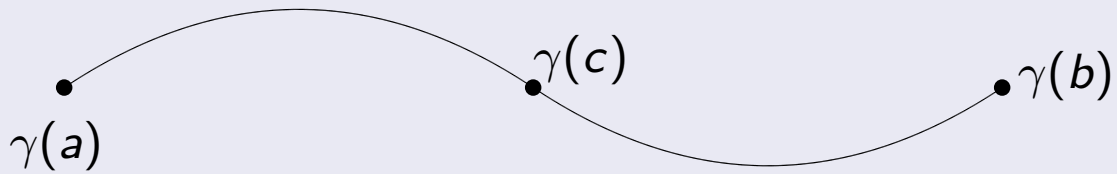
A function $L : A \to [0, \infty$ is called a *length* function if it satisfies the properties

1. Additivity: $L(\gamma|_{[a,b]}) = L(\gamma_{[a,c]}) + L(\gamma_{[c,b]})$ for any point $c \in (a, b)$.
2. Continuity: $L(\gamma; a, t) := L(\gamma|_{a,t})$ is continuous function for any $t \in [a, b]$.
3. Invariance under reparametrization: $L(\gamma \cdot \varphi) = L(\gamma)$, where $\varphi : [a, b] \to [a, b]$ is a linear homeomorphism i.e. φ continuous, bijective, φ^{-1} continuous.

Definition

A **path** γ in a space X is a continuous map $\gamma : [a, b] \to X$. Let A be the class of admissible paths on X, i.e. A is subset of all continuous paths in X. A function $\mathcal{L} : A \to \mathbb{R}^+$ is called a **length function** if it satisfies the properties:

1. Additivity : $\mathcal{L}(\gamma|_{[a,b]}) = \mathcal{L}(\gamma|_{[a,c]}) + \mathcal{L}(\gamma|_{[c,b]})$ for $\forall c \in [a, b]$.



2. Continuity : $\mathcal{L}(\gamma; a, t) := \mathcal{L}(\gamma_{[a,t]})$ is continuous function for $\forall t \in [a, b]$.
3. Invariance under re-parametrization $\mathcal{L}(\gamma \circ \varphi) = \mathcal{L}(\gamma)$ where $\varphi : [a, b] \to [a, b]$ is a linear homeomorphism.

Length spaces

Definition

Let $(X, \mathcal{L})$ be a length space. A length $\mathcal{L}$ induces **a distance** $d_{\mathcal{L}}(x, y) := \inf\{\mathcal{L}(\gamma) | \gamma : [a, b] \to X, \gamma \in A, \gamma(a) = x, \gamma(b) = y\}$.

Proposition

$(X, d_{\mathcal{L}})$ defined above is a metric space.

Definition

A metric d obtained as the distance function associated to a length structure is called **intrinsic metric**, or **length metric**.

Remark

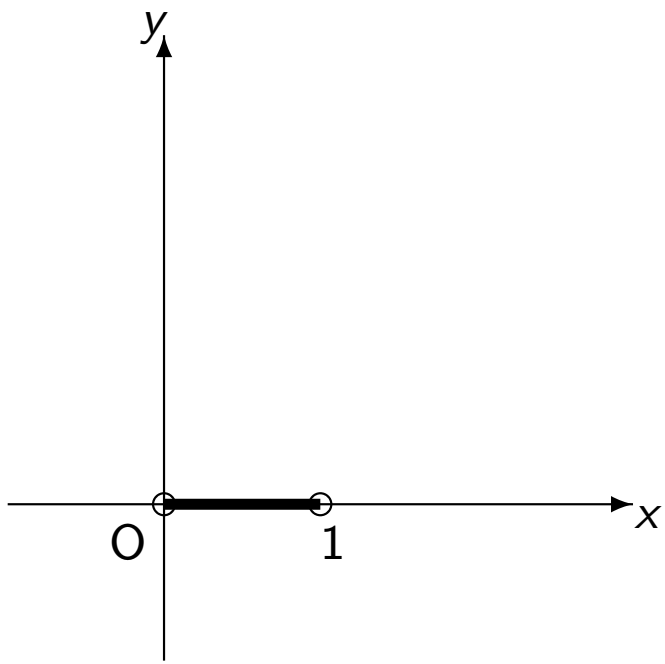
We use inf instead of min in definition of $d_{\mathcal{L}}$ because there may be no shortest path between two points.

Now, $\lambda = \inf E$ defined by

1. $\lambda \leq x$, for $\forall x \in E$.
2. there exists $x \in E$ such that $x < \lambda + \varepsilon$, for $\forall \varepsilon > 0$.

Example

Let us consider $X = \mathbb{R}^2 \setminus (0,1)$, i.e. $\mathbb{R}^2$ with the interval $(0,1)$ on the Ox axis deleted. Then there is no shortest path between the points $x = (0,0)$, $y = (1,0)$ because the shortest path was just removed. However, we can still approximate the distance between x and y. In reality, in many cases there is no shortest path between points.



Definition

A length space $(X, \mathcal{L})$ is called **complete** if for $\forall x, y \in X$ there exists an admissible path joining x to y and whose length is equal to $d_{\mathcal{L}}(x, y)$. In other words, a length structure is complete if there exists a shortest path between every two points. This shortest path is called **a geodesic** of $(X, \mathcal{L})$.

Remark

Let $(X, \mathcal{L})$ be a length space. For a fixed point $p \in X$ we define

$$C(p) = \{q \in X | \text{ there exists at least 2 geodesics of equal length from } p \text{ to } q\} \subset X,$$

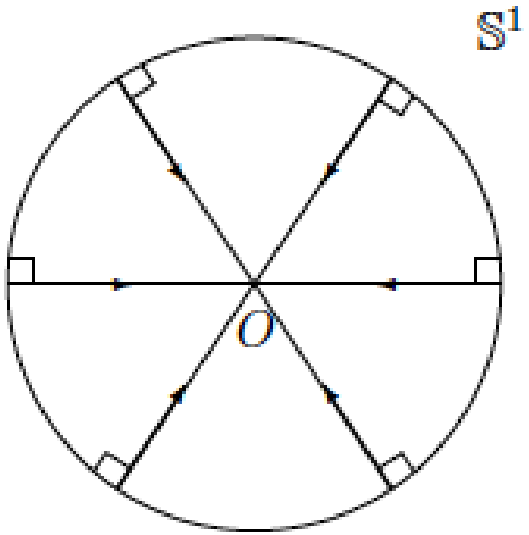
This set is called **the cut locus of** p.
For a closed subset $Y \subset X$, we define

$$C(Y) = \{q \in X | \text{there exists at least 2 geodesics of equal length from } Y \text{ to } q\} \subset X.$$

This set is called the **cut locus of** Y.

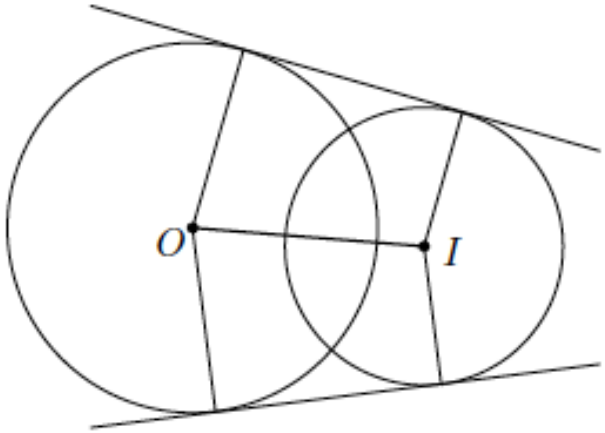
Example

Consider $\mathbb{S}^1 \subset \mathbb{R}^2$ endowed with the standard norm. Then the inner geodesic from $\mathbb{S}^1$ are inner normals that meet in the center of $\mathbb{S}^1$. It means that $C(\mathbb{S}^1) = \{0\}$. The same is true for $\mathbb{S}^n \subset \mathbb{R}^{n+1}$, namely $C(\mathbb{S}^n) = \{0\}$.

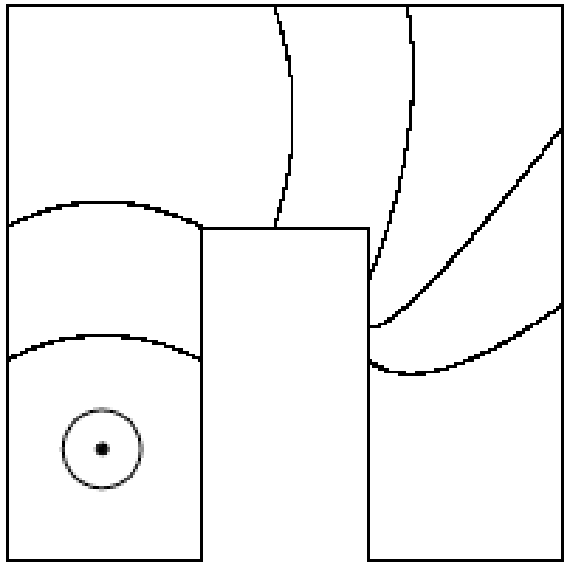


Example

Consider the following closed curve $\mathbb{S}$ in $\mathbb{R}^2$. The inner geodesics from $\mathbb{S}$ gather on the segment OI such that $C(\mathbb{S}) = OI$. The same is true in any dimension.



Different examples



A metric on an island

Consider a concave region in $\mathbb{R}^2$ of this shape with the induced Euclidean length (an island).
The set of admissible paths

$$A = \text{piecewise smooth paths contained in this region.}$$

Metric balls in a region like this one are like in the picture above.

Example

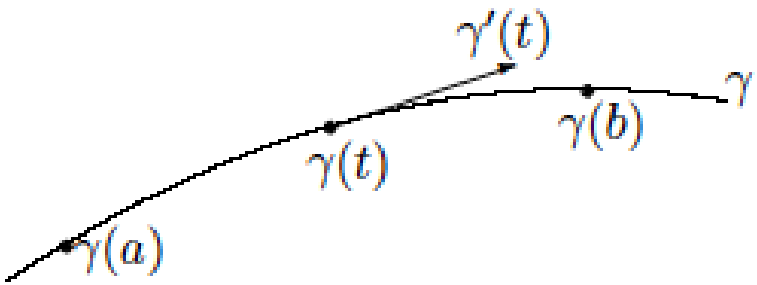
Crossing a swamp (a marsh), or a mountain trail (conformal length). Let $X = R^2$, and the set of admissible paths

$$A = \text{all smooth piecewise paths in plane.}$$

Consider $f : \mathbb{R}^2 \to \mathbb{R}^+$ a positively-valued continuous function and define the length of path $\gamma \in A, \gamma : [a, b] \to \mathbb{R}^2$ by

$$\mathcal{L}_R(\gamma) = \int_a^b f(\gamma(t)) \cdot \|\gamma'(t)\| \cdot dt,$$

where $\|\gamma(t)\|$ is the Euclidean norm of the tangent vector to γ, i.e. $\|\gamma'(t)\| = \sqrt{(\dot{\gamma}_1)^2 + (\dot{\gamma}_2)^2}$. This is a **weighted Euclidean distance**.



Remark

Imagine a traveler who measure the length (=time needed to cover) of a route.

A large f for a path difficult to traverse (e.g. a swamp or a mountain trail). This is an example of Riemannian metric structure (conformal flat).

Finslerian length

Remark

The previous example supposes that the difficulty of traversing a region is the same in each direction. A more real situation is when we assume that **the difficulty of traversing a region depends that not only on the region itself, but also on the direction of displacement**. Imagine a windy landscape. Obviously the traveling speed depends on the direction because of the wind. The same when moving on a slope.

Definition

To incorporate this new condition, we have to define a new type of length function of an admissible path $\gamma : [a, b] \to M$

$$\mathcal{L}_F(\gamma) = \int_a^b F(\gamma(t), \gamma'(t))dt,$$

when F is a smooth function of 2 variables : a point x and a vector v such that

1. $F(x, v) > 0$,
2. $F(x, \lambda v) = \lambda \cdot F(x, v)$, for $\forall \lambda > 0$,
3. The unit spheres $F^{-1}(1)$ are strongly convex.

A length function like this one is called **Finslerian length**.

Here, convex function is a function which satisfies the following condition

$$f((1 - t\lambda)x + \lambda y) \leq (1 - \lambda)f(x) + \lambda f(y), \ \forall x, y, \ \lambda \in [0, 1].$$

Definition

The Finslerian length induces the **Finslerian distance function**, or **Finslerian metric function** $d : X \times X \to \mathbb{R}$,

$$d_F(x, y) = \inf\{\mathcal{L}_F(\gamma) | \gamma : [a, b] \to X, \gamma \in A, \gamma(a) = x, \gamma(b) = y\}$$

for any $x, y \in X$.

Remark

It is important to observe that the Finslerian induced distance is not symmetric i.e.

$$d_F(x, y) \neq d_F(y, x).$$

This type of distance is called a **quasi-distance**, or **quasi-metric**.

Summary on lengths

	Length	motion type
Euclidean Length	$\mathcal{L}_E(\gamma) := \int_a^b \sqrt{(\dot{\gamma_1})^2 + (\dot{\gamma_2})^2} dt$	ideal motion example : motion in vacuum
Riemanian Length	$\mathcal{L}_R(\gamma) = \int_a^b f(\gamma(t)) \cdot \sqrt{(\dot{\gamma_1})^2 + (\dot{\gamma_2})^2} dt$	motion only depends on point (weighted Euclidean length) example : swamp
Finslerian Length	$\mathcal{L}_F(\gamma) = \int_a^b F(\gamma(t), \gamma'(t)) dt$	motion which depends on point and direction example : navigation, cruise

Distances

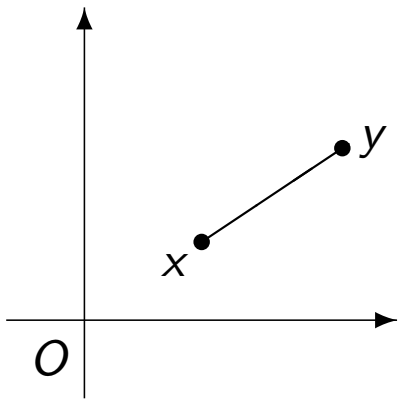
Distance(metric) on a set X

$d : X \times X \to [0, \infty)$ such that

1. Positiveness: $x \neq y \Rightarrow d(x, y) > 0$, and $x = y \Rightarrow d(x, y) = 0$, $\forall x, y \in X$
2. Symmetry: $d(x, y) = d(y, x)$
3. Triangle inequality: $d(x, z) \leq d(x, y) + d(y, z)$, $\forall x, y, z \in X$

Example: Euclidean distance in plane

- $\left|xy\right| = \sqrt{\sum_{i=1}^{2}(x_i - y_i)^2}$



Quasi-distances

Let X be an arbitrary set.

Definition

A function $d : X \times X \to \mathbb{R}$ is called a **quasi-metric** on X if the following conditions are satisfied for $\forall x, y, z \in X$:

1. Positiveness : $d(x, y) > 0$ if $x \neq y, d(x, x) = 0$.
2. Triangle inequality : $d(x, z) \leq d(x, y) + d(y, z)$.
3. Separation axiom : $d(x, y) = d(y, x) = 0 \Rightarrow x = y$.

A quasi-metric d that satisfies the condition :

Symmetry : $d(x, y) = d(y, x)$ for $\forall x, y \in X$ is a metric on X.

Definition

If d is a quasi-metric, then the pair (X, d) is called a **quasi-metric space**, and the quantity $d(x, y)$ is called a **quasi-distance** on X.

If (X, d) is a quasi-metric space, then

$$\overline{d}(x, y) := d(y, x), \text{for } \forall x, y \in X,$$

is called **the dual**, or **the conjugate**, quasi-metric of d, and

$$d^*(x, y) := \max\{d(x, y), d(y, x)\}$$

is called the **associated metric** of d. The associated metric is the smallest metric majoring d.

Remark

A quasi-metric is a metric if and only if it coincides to its dual.

Definition

If (X, d) is a quasi-metric space then the function $\theta : X \times X \to \mathbb{R}^+,\ \theta(x, y) := \frac{1}{2}[d(x, y) + d(y, x)]$, for $\forall x, y \in X$, is called the **symmetrization** of d.

Remark

If (X, d) is a quasi-metric space and θ its symmetrization, then (X, θ) is a metric space.

Definition

Let (X, d) be a quasi-metric space, $x \in X$, $A, B \subseteq X$, $\varepsilon > 0$. We denote for later use :

- $diam(A) := \sup\{d(x, y) : x, y \in A\}$ the **diameter** of the set A;
- $\mathbb{B}^f_\varepsilon(x) := \{y \in X : d(x, y) < \varepsilon\}$ **the forward open ball** of radius ε centered at x;
- $\mathbb{B}^b_\varepsilon(x) := \{y \in X : d(y, x) < \varepsilon\}$ **the backward open ball** of radius ε centered at x.
- $d(x, A) := \inf\{d(x, y) : y \in A\}$ the **forward distance** from x to A;
- $d(A, x) := \inf\{d(y, x) : y \in A\}$ the **backward distance from** x to A;
- $A^f_\varepsilon := \{x \in X : d(A, x) < \varepsilon\}$ the **forward ε-neighborhood** of A;
- $A^b_\varepsilon := \{x \in X : d(x, A) < \varepsilon\}$ the **backward ε-neighborhood** of A;
- $d(A, B) := \inf\{d(x, y) : x \in A, y \in B\}$ the **distance between sets** A and B.

Example

The **left quasi-metric** $u^L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is defined by

$$u^L(x, y) := \max\{x - y, 0\}.$$

Similarly, we define the **right quasi-metric** $u^R : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ by

$$u^R(x, y) := \max\{y - x, 0\}.$$

Obviously u^L and u^R are dual quasi-metrics each other. The associated metric

$$u := \max\{u^L, u^d\}$$

is **the canonical absolute value** metric on $\mathbb{R}$ defined by

$$u(x, y) = |x - y|.$$

Example

Another quasi-metric on $\mathbb{R}^+$ is

$$d(x,y) = \begin{cases} \min(1, y-x), & \text{if } x \leq y \\ 1, & \text{otherwise.} \end{cases}$$

The associated metric d^* is the distance metric.

Example

Let (X,d) be **an extended quasi-metric**, i.e. a quasi-metric that takes values in $\mathbb{R}^+ \cup \{\infty\}$. Then $\rho : X \times X \to \mathbb{R}^+$, defined by

$$\rho(x,y) = \min\{1, d(x,y)\}$$

is a quasi-metric. In this way any unbounded quasi-metric can be easily converted to a bounded quasi-metric.

Example

Let $(X_i, d_i), i = 1, 2, \cdots, n$ be different quasi-metrics and denote $X = X_1 \times X_2 \times \cdots \times X_n$, i.e. for $\forall x \in X$, we have $x = (x_1, \ldots, x_n)$, where $x_i \in X_i$, for $\forall i = 1, \ldots, n$. We define the function $d : X \times X \to \mathbb{R}$ by

$$d(x, y) := \sum_{i=1}^{n} d_i(x_i, y_i).$$

It can be easily seen that (X, d) is a quasi-metric space. The pair (X, d) is called the l_1-**type quasi-metric space**.

Example

(The Hamming distance)

Let $X = X_1 \times \cdots \times X_n$ be an l_1-type product space as above. In particular, we think

$$X = \mathbb{R}^n = \underbrace{\mathbb{R} \times \cdots \times \mathbb{R}}_{n \text{ times}}.$$

The **Hamming metric** is the metric obtained by setting each d_i above to be the distance metric. In other words $d(x, y) = |\{i : x_i \neq y_i\}|$, i.e. number of positions that differs in $x, y \in X$. For instance, if $X = \mathbb{R}^6$, we consider the points

$$\begin{array}{c} x{=}\text{ABCDEF} \\ \quad\;\; | \; | \; | \; | \\ y{=}\text{FBCDEA} \end{array}$$

it follows that the Hamming distance between x and y is $d(x, y) = 2$.

Quasi-normed spaces

Definition

A **semigroup** $(X, *)$ is a set X with a binary operation $*$ satisfying

1. Closure : $x * y \in X$, for $\forall x, y \in X$.
2. Associativity : $x * (y * z) = (x * y) * z$, for $\forall x, y, z \in X$.

A **monoid** is a semigroup with identity, i.e. there exists a unique neutral element $e \in X$ such that $x * e = e * x = x$, for $\forall x \in X$.

A **group** is an inversable monoid, i.e. for $\forall x \in X$, there exists $x^{-1} \in X$ such that $x * x^{-1} = x^{-1} * x = e$.

A group $(X, *)$ is called **abelian** (or **commutative**) if $x * y = y * x$, for $\forall x, y \in X$.

Definition

A **homomorphism** from a semigroup $(X, *)$ to a semigroup $(Y, *)$ is a map $\varphi : X \to Y$ such that $\varphi(x * y) = \varphi(x) * \varphi(y)$, for $\forall x, y \in X$.
An **isomorphism** is a bijective homomorphism $\varphi : X \to Y$ such that its inverse function $\varphi^{-1} : Y \to X$ is also a homomorphism.

Definition

A **semilinear(or semi vector) space** on $\mathbb{R}^+$ is a triple $(X, +, \cdot)$ such that

1. $(X, +)$ is abelian semigroup with neutral element $O \in X$;
2. $(X, \cdot)$ satisfies the conditions:
   - 2-1 $a \cdot (b \cdot x) = (ab) \cdot x$
   - 2-2 $(a + b) \cdot x = a \cdot x + b \cdot x$
   - 2-3 $a \cdot (x + y) = a \cdot x + a \cdot y$
   - 2-4 $1 \cdot x = x$

   for $\forall x, y \in X, a, b \in \mathbb{R}^+$.

Remark

1. If an element $x \in X$ has an inverse, then it can be proved to be unique.
2. If we replace the word "semigroup" with "group" and $\mathbb{R}^+$ with $\mathbb{R}$ in the definition above we obtain an ordinary **linear (or vector) space**.

Definition

Let $(E, +, \cdot)$ be a linear space over $\mathbb{R}$, and let $e \in E$ be the neutral element of $(E, +)$. A **quasi-norm** on E is a function $\|\cdot\| : E \to \mathbb{R}^+$ such that

1. $\|x\| = \|-x\| = 0 \Leftrightarrow x = e$,
2. $\|a \cdot x\| = a \cdot \|x\|$,
3. $\|x + y\| \leq \|x\| + \|y\|$, for $\forall x, y \in E, a \in \mathbb{R}^+$.

The pair $(E, \|\cdot\|)$ is called a **quasi-normed** space.

Remark

1. One can easily verify that the function $\|\cdot\|^* : E \to \mathbb{R}^+$ defined by

$$\|x\|^* := \max\{\|x\|, \|-x\|\}$$

   is a norm on E.
2. A quasi-norm $\|\cdot\|$ induces a quasi-distance $d_{\|\cdot\|}$ in a natural way.

Proposition

Let $(E, \|\cdot\|)$ be a quasi-normed space. Then the function $d_{\|\cdot\|} : E \times E \to \mathbb{R}^+$ defined by

$$d_{\|\cdot\|}(x, y) = \|y - x\|$$

is a quasi-metric whose conjugate $\overline{d}_{\|\cdot\|}$ is given by

$$\overline{d}_{\|\cdot\|}(x, y) = \|x - y\|.$$

Example

A quasi-norm in $\mathbb{R}$ is given by

$$\|x\| = \max\{x, 0\}, \text{ for } \forall x \in \mathbb{R}.$$

Its induced distance coincides with

$$u^b(x, y) := \max\{y - x, 0\}.$$

Remark

It is important to remark that for the quasi-norm in Example above $\|a \cdot x\| = a \cdot \|x\|$ for $a > 0$, but this is not true any more for $a < 0$. Indeed, for example,

$$\|2 \cdot 3\| = \max\{2 \cdot 3, 0\} = 2 \cdot 3 = 2 \cdot \|3\|,$$

but

$$\|(-2) \cdot 3\| = \max\{-6, 0\} = 0.$$

Weighted quasi-distance

One important class of quasi-metrics are the weighted quasi-metrics because they are strictly related to the geometry of sequence comparison.

Definition

Let (X, d) be a quasi-metric space. The quasi-metric d is called a **weightable quasi-metric** if there exists a function $w : X \to \mathbb{R}^+$ called the **weight function**, or simply the **weight**, that satisfies

$$d(x, y) + w(x) = d(y, x) + w(y), \text{ for } \forall x, y \in X.$$

Remark

In the case when (X, d) is a metric space, i.e.
$d(x, y) = d(y, x)$ for $\forall x, y \in X$, the triple $(X, d, w = \text{constant})$ is a quasi-metric space.

Definition

A quasi-metric d is called **co-weightable** if its conjugate quasi-metric $\overline{d}$ is weightable. The weight function w by which $\overline{d}$ is weightable is called the **co-weight** of d.

Definition

A triple (X, d, w), where (X, d) is a quasi-metric space and $w : X \to \mathbb{R}^+$ a function, is called **weighted quasi-metric space** if (X, d) is weightable by w. The triple (X, d, w) is called a **co-weighted quasi-metric space** if (X, d) is co-weightable by w.

In the definitions above, if the weight function w take values in $\mathbb{R}$ instead of $\mathbb{R}^+$, the (X, d, w) is called a **generalized weighted quasi space** or a **generalized co-weighted quasi-space**, respectively.

Remark

Given a weighted quasi-metric space (X, d, w) the symmetrization $\theta : X \times X \to \mathbb{R}^+$ of (X, d, w) can be written as

$$\theta(x, y) = d(x, y) + \frac{1}{2}[w(x) - w(y)].$$

It can be seen that

$$\left|\frac{1}{2}[w(x) - w(y)]\right| \leq \theta(x, y), \text{ for } \forall x, y \in X.$$

Remark

If (X, d, w) is a weighted quasi-metric space, then w' is another weight function for (X, d) if and only if $w - w'$ is constant. Hence if a given quasi-metric space (X, d) admits a weight w, then its conjugate space $(X, \overline{d})$ is weightable if and only if w is bounded. This remark shows that not any weighted quasi-space is a co-weighted quasi-space.

Proposition

A quasi-metric space (X, d) admits a generalized weight w if and only if

$$d(x,y) + d(y,z) + d(z,x) = d(x,z) + d(z,y) + d(y,x), \ \textit{for} \ \forall x, y, z \in X.$$

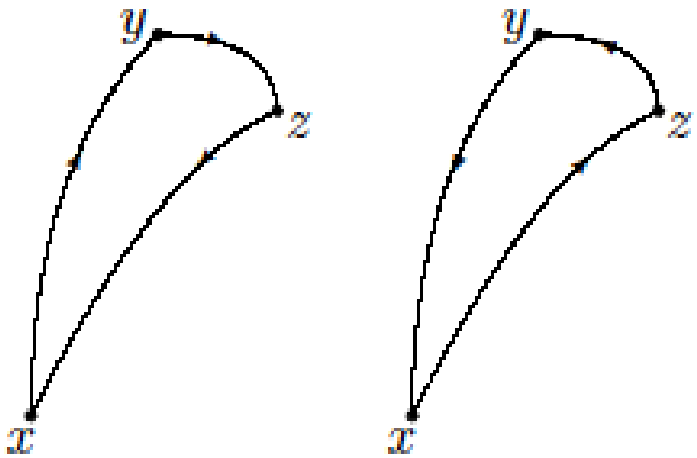
Indeed, let $a \in X$ be a fixed reference point. Then if one defines

$$w_a(x) := d(a,x) - d(x,a)$$

for $\forall x \in X$. Then the function w_a is a generalized weight for (X, d), i.e.

$$d(x,y) + w_a(x) = d(y,x) + w_a(y),$$

for $\forall x, y \in X$.



The perimeter (with respect to a weighted quasi-distance) of the triangle Δxyz is independent of the orientation.

Example

Let $X = \mathbb{R}^+$ and set $d := u^R|_{\mathbb{R}^+}$ the restriction of the right quasi-metric $u^R(x, y) := \max\{y - x, 0\}$ to positive reals, i.e. for $\forall x, y \in \mathbb{R}^+$,

$$d(x, y) = \begin{cases} y - x, & \text{if } x \leq y \\ 0, & \text{if } y < x. \end{cases}$$

Set $w(x) := x$ for $\forall x \in X$. Then one can easily see that (X, d, w) is a weighted quasi-metric space.

Lemma

Let (X, d, w) be a generalized quasi-metric space.

1. *If w is bounded below, i.e. there exists a constant m such that*

$$w(x) > m, \ for \ \forall x \in X,$$

then $(X, d, w - m)$ is a weighted quasi-metric space.

2. *If w is bounded above, i.e. there exists a constant M such that*

$$w(x) < M, \ for \ \forall x \in X,$$

then $(X, \overline{d}, M - w)$ is a weighted quasi-metric space.

3. *If $(X.\overline{d}, u)$ is generalized weighted quasi-metric space then $w + u$ is constant on X.*

Lemma

Let (X, d, w) be a weighted quasi-metric space. Then w is a right 1-Lipschitz function, i.e. $\|w(x) - w(y)\| \leq d(y, x)$, for $\forall x, y \in X$.

Relation with Finsler metrics

Definition

A Finsler norm, or metric, on a real smooth, n-dimensional manifold M is a function $F : TM \to [0, \infty)$ that is positive and smooth on $\widetilde{TM} = TM \backslash \{0\}$, has the *homogeneity property* $F(x, \lambda v) = \lambda F(x, v)$, for all $\lambda > 0$ and all $v \in T_x M$, having also the *strong convexity* property that the Hessian matrix

$$g_{ij} = \frac{1}{2} \frac{\partial^2 F^2}{\partial y^i \partial y^j}$$

is positive definite at any point $u = (x^i, y^i) \in \widetilde{TM}$.

(M, F) is called a *Finsler manifold* or *Finsler structure*.

The Finsler structure is called *absolute homogeneous* if $F(x, -y) = F(x, y)$ because this leads to the homogeneity condition $F(x, \lambda y) = |\lambda| F(x, y)$, for any $\lambda \in \mathbb{R}$. We don't need this assumption in the present talk.

Remark.

The fundamental function F of a Finsler structure (M, F) determines and it is determined by the (tangent) *indicatrix*, or the total space of the unit tangent bundle of

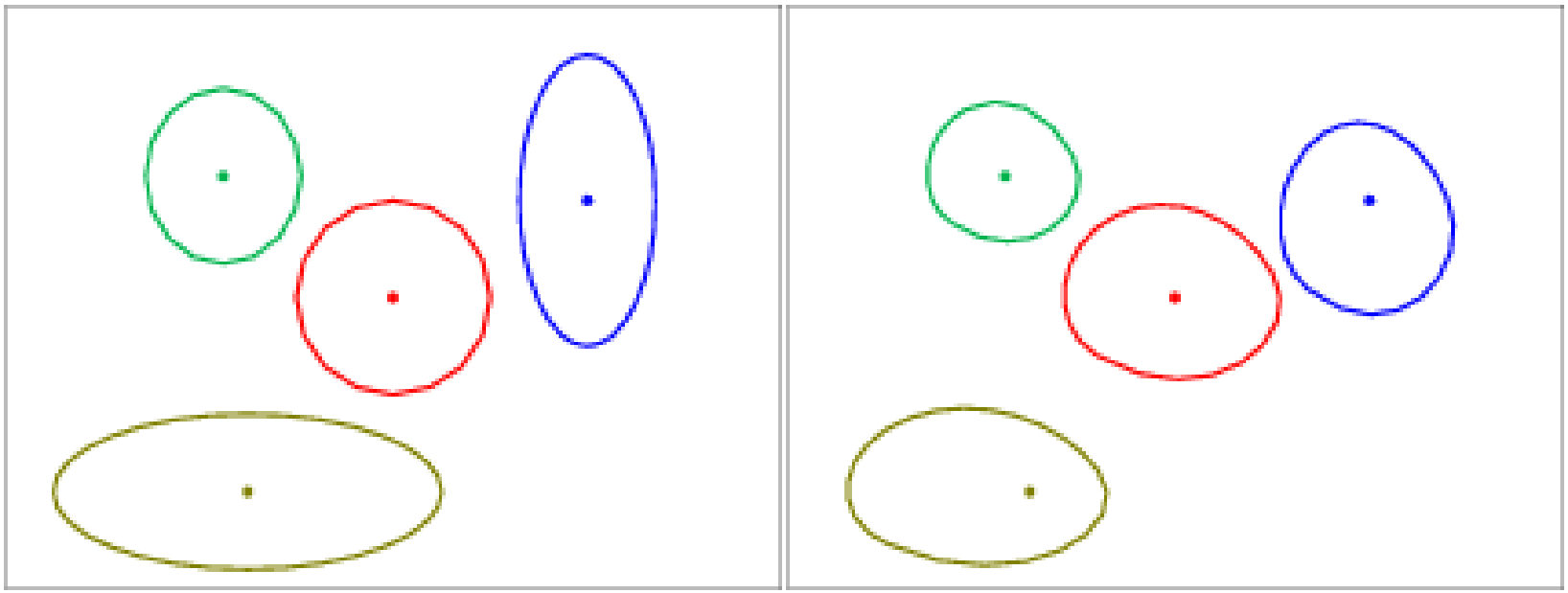
$$SM := \{u \in TM : F(u) = 1\}$$

which is a smooth hypersurface of TM.

At each $x \in M$ we also have the *indicatrix at* x

$$S_xM := \{v \in T_xM \mid F(x, v) = 1\} = SM \cap T_xM$$

which is a smooth, closed, strictly convex hypersurface in T_xM.



Riemannian vs. Finslerian unit circles.

Let $\gamma : [a, b] \to M$ be a regular piecewise C^∞-curve in M, and let $a := t_0 < t_1 < \cdots < t_k := b$ be a partition of $[a, b]$ such that $\gamma|_{[t_{i-1}, t_i]}$ is smooth for each interval $[t_{i-1}, t_i]$, $i \in \{1, 2, \ldots, k\}$.

Definition

The *forward integral length* of γ is given by

$$\mathcal{L}_\gamma^+ := \sum_{i=1}^{k} \int_{t_{i-1}}^{t_i} F(\gamma(t), \dot{\gamma}(t))dt,$$

where $\dot{\gamma} = \dfrac{d\gamma}{dt}$ is the tangent vector along the curve $\gamma|_{[t_{i-1}, t_i]}$.

Proposition

$$
\begin{aligned}
(\mathcal{L}^+)'(0) =& g_{\dot\gamma(b)}(\gamma, U)\big|_a^b \\
&+ \sum_{i=1}^{k} \Big[g_{\dot\gamma(t_i^-)}(\dot\gamma(t_i^-), U(t_i)) - g_{\dot\gamma(t_i^+)}(\dot\gamma(t_i^+), U(t_i)) \Big] \\
&- \int_a^b g_{\dot\gamma}(D_{\dot\gamma}\dot\gamma, U) dt,
\end{aligned}
$$

where $D_{\dot\gamma}$ is the covariant derivative along γ with respect to the Chern connection and γ is arc length parametrized.

Definition

A regular piecewise C^∞-curve γ on a Finsler manifold is called a *forward geodesic* if $(\mathcal{L}^+)'(0) = 0$ for all piecewise C^∞-variations of γ that keep its ends fixed. In terms of Chern connection a constant speed geodesic is characterized by the condition $D_{\dot\gamma}\dot\gamma = 0$.

For any two points p, q on M, let us denote by $\Omega_{p,q}$ the set of all piecewise C^∞-curves $\gamma : [a, b] \to M$ such that $\gamma(a) = p$ and $\gamma(b) = q$.

Proposition

The map

$$d : M \times M \to [0, \infty), \qquad d(p, q) := \inf_{\gamma \in \Omega_{p,q}} \mathcal{L}_\gamma^+$$

gives the *Finslerian distance* on M. It can be easily seen that d is in general a quasi-distance, i.e., it has the properties

1. $d(p, q) \geq 0$, with equality if and only if $p = q$;
2. $d(p, q) \leq d(p, r) + d(r, q)$, with equality if and only if r lies on a minimal geodesic segment joining from p to q (triangle inequality).

The reverse distance $d(q, p)$ is actually the Finslerian distance induced by the backward integral length.

Remark

In the case where (M, F) is absolutely homogeneous, the symmetry condition $d(p, q) = d(q, p)$ holds and therefore (M, d) is a genuine metric space.

Recall that a Finsler metric F on a n-dimensional differential manifold M is called *with reversible geodesics* if and only if for any geodesic $\gamma : [0,1] \to M$ of F, the reverse curve $\bar{\gamma}(t) := \gamma(1-t)$ is also a geodesic of F.
We point out that even a Finsler space is with reversible geodesics, the Finslerian distance function d_F is not symmetric, except for the absolute homogeneous case.
We have (see [Masca, Sabau, Shimada 2010] and references herein)

Proposition

Let $(M, F = F_0 + \beta)$ be a Finsler space whose fundamental function is obtained by a Randers change of an absolute homogeneous Finsler metric F_0 by a 1-form β. Then (M, F) is with reversible geodesics if and only β is closed.

The intuitive meaning of the Randers change

$$F = F_0 + \beta, \quad d\beta = 0 \tag{1}$$

is that the F-geodesics coincide with the F_0-geodesics as set of points, i.e. F and F_0 are projectively equivalent. Remark that this Randers change is a special Randers change with β closed. Hereafter, when referring to this formula, we always mean *Randers change with β closed*.

Remark

1. A special case is the case of Randers metrics $F = \alpha + \beta$, where α is a Riemannian metric and β closed 1-form. It is known that a Randers metric is positive definite if and only if the Riemannian length of the vector $b_i(x)$ is less than one, i.e. $b(x) < 1$, for $\forall x \in M$.
2. It is not clear yet if all Finsler metrics with reversible geodesics are given by (1). However, in the case of (α, β)-metrics we have shown that if (M, F) is with reversible geodesics then F is given by the Randers change (1).

Theorem

Let M be an n-dimensional simply connected smooth manifold.
A Finsler metric F induces a generalized weighted quasi-distance d_F on M if and only if it is the Randers change of an absolute homogeneous Finsler space F_0 by an exact one-form β.

For later use we recall the following lemma.

Lemma

Let (M, d) be any quasi-metric space. Then d is weightable if and only if there exists $w : M \to [0, \infty)$ such that

$$d(x, y) = \rho(x, y) + \frac{1}{2}[w(y) - w(x)], \quad \forall x, y \in M,$$

where ρ is the symmetrized distance of d. Moreover, we have

$$\frac{1}{2}|w(x) - w(y)| \leq \rho(x, y), \quad \forall x, y \in M.$$

The proof is trivial from the definition of a weighted quasi-metric.

Remark

If (M, F) is a Finsler space given by the Randers change (1), then the induced quasi-metric d_F and the symmetrized metric ρ induce the same topology on M.

Remark

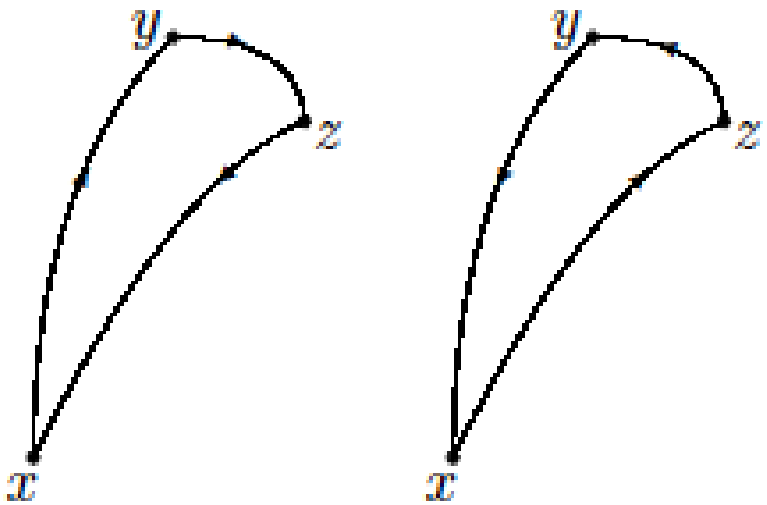
From Lemma above it can be seen that the assumption of w to be smooth is not essential. Indeed, it can be seen that if d_F is a weighted quasi-metric, the function w is 1-locally Lipschitz, that is differentiable almost everywhere on M. Therefore, the one-form β exists almost everywhere on M.

We discuss an interesting geometric property concerning the geodesic triangles.

Proposition

Let (M, F) be a Finsler metric given by the Randers change (1). Then the perimeter length of any geodesic triangle on M does not depend on the orientation, that is

$$d_F(x,y) + d_F(y,z) + d_F(z,x) = d_F(x,z) + d_F(z,y) + d_F(y,x), \quad \forall x, y, z \in M. \tag{2}$$



The perimeter (with respect to the Finsler metric) of the geodesic triangle Δxyz is independent of the orientation.

In other words, even though the distance between two points x and y depends on the orientation of a minimizing geodesic joining points x and y, i.e. $d_F(x,y) \neq d_F(y,x)$, the sum of distances between three points x, y, z on M do not depend on the direction we trace out the perimeter of the geodesic triangle Δxyz. We point out that weighted quasi-metric spaces can be characterized by this property without the explicit use of the weight function. Indeed, a quasi-metric d is weightable if and only if relation (2) holds.

Proposition

Let (M, F) be a Finsler space that satisfies (2). Then F can be written as the Randers change of an absolute homogeneous Finsler metric F_0 by an exact one-form β.

Remark

It should be clear that not any quasi-metric space is weightable. In fact, it can be shown that the class of weightable quasi-metric spaces are exactly those quasi-metric spaces that satisfy relation (2).

References on metric spaces

- D. Azagra, J. Ferrera, F. Lopez-Mesas, Y. Rangel, *Smooth approximation of Lipschitz functions on Riemannian manifolds*, J. Math. Anal. Appl. **326** (2007) 1370–1378.
- D. Burago, Y. Burago, S. Ivanov, A course in Metric Geometry, GSM, vol.33, American Math-Soc, 2001.
- H. P. A. Künzi, V. Vajner, *Weighted quasi-metrics*, in: Papers on General topology and Applic., Annals New York Acad. Sci. **728**, (1994) 64–77.
- A. Stojmirović, Quasi-metrics, Similarities and searches : aspects of geometry of protein data sets, Ph. D. Thesis, 2005, arxiv : 0810.5407v1.
- P. Vitolo, *A Representation theorem for quasi-metric Spaces*, Topology and its Appl., **65** (1995), 101–104.
- P. Vitolo, *The Representation of Weighted Quasi-Metric Spaces*, Rend. Istit. Mat. Univ. Trieste, Vol. XXXI, 95-100 (1999).

References on Finsler spaces

- D. Bao, S. S. Chern, Z. Shen, An Introduction to Riemann Finsler Geometry, Springer, GTM **200**, 2000.
- I. Masca, S. V. Sabau, H. Shimada, *Reversible geodesics for* (α, β)*-metrics*, Intl. Journal Math. 21/8 (2010), 1071–1094.
- S. V. Sabau, K. Shibuya, H. Shimada, *Metric structures associated to Finsler metrics*, 2013, arXiv:1305.5880.
- Z. Shen, Lectures in Finsler Geometry, World Scientific, Singapore, 2001.

Biological motivation

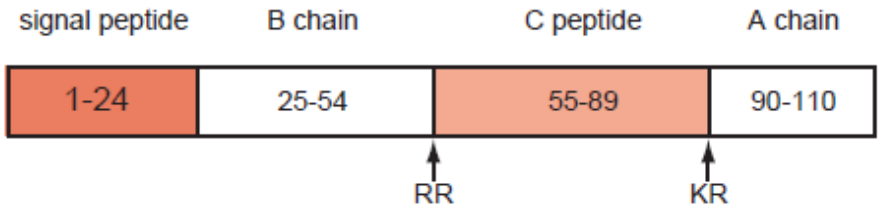
- $\sum = \{A, B, C\}$: **alphabet**
  $A, B, \ldots$: **letters of alphabet (generators)**
- $\sum^*$: **word set (free monoid),**
  $\sum^+$: **word except empty word (free semigroup).**

The macromolecules that contain the fundamental information of life can be expressed as words over a finite alphabet, such that

- **DNA molecules: words on** $\sum = \{A, C, T, G\}$,
- **RNA molecules: words on** $\sum = \{A, C, U, G\}$,
- **proteins molecules:**
  words on $\sum = \{$**20 amino proteinogenic acid**$\}$.

Insulin

Human pre-proinsulin sequence



Words: from left, $u_1^*, u_2^*, u_3^*, u_4^*$.

u_1^* =MALWMRLLPLLALLALWGPDPAAA: prefix

u_4^* =GIVEQCCTSICSLYQLENYCN: suffix

Comparison of biological sequence

Estimations of homology, evolutionary relation, etc.

similarity	distance
low	long
high	short

- short distance $\Leftrightarrow$ high similarity

Score function

Score function $s : X \times X \to \mathbb{R}$

1. $s(x,x) > 0$, $\forall x \in X$,
2. $s(x,x) \geq s(x,y)$, $\forall x, y \in X$,
3. $s(x,y) = s(x,x)$ and
   $s(y,x) = s(y,y) \Rightarrow x = y$, $\forall x, y \in X$,
4. $s(x,y) + s(y,z) \leq s(x,z) + s(y,y)$, $\forall x, y, z \in X$.

Induced distance function

If there is a score function s, then $d : X \times X \to \mathbb{R}$,

$$\boxed{d(x,y) := s(x,x) - s(x,y)}$$

is a quasi-metric (Seler, 1974).

Similarity score matrices (PAM & BLOSUM)

PAM matrix: (Dayhoff, 1978)

S_1, S_2 are at 1-PAM unit evolutionary distance if S_1 is transformed in S_2 with an average of 1 pointwise mutations per 100 residues.

BLOSUM matrix: (Henickhoff and Henickhoff, 1992)

Defined from local, ungapped alignment (blocks) of closely and distantly related proteins.

Conclusion on evolutionary distances

1. The distance induced by PAM, BLOSUM matrices is not symmetric
   1. PAM-induced metrical geometry is highly inconsistent.
   2. **BLOSUM-induced distance is a well-defined weighted quasi-metric (more than 40% sequence identity).**
2. When comparing distantly related proteins the hydrophobic residues need to be explicitely considered and studied separately.

Alignments

- **String edit distances** = the smallest number of permitted edit operations required to transform one string into another.
- **Permitted edit operations** = substitutions, insertions, and deletions.
- **Alignment** = mappings that transforms a sequence into another.

Example : $u = COMPLEXITY$, $v = FLEXIBILITY$

(a) **global alignment**

```
COMPLE - - - -XITY
.  | |          | | |
FLE X I B I L ITY
```

(b) **local alignment**

```
ITY    LEXI
| | |  | | | |
ITY    LEXI
```

Dynamic Programming Algorithms

- **Global similarity** : Needleman-Wunsch dynamic programming algorithm
- **Local similarity** : Smith-Waterman dynamic programming algorithm

Theorem 1. [Stojmirovic,Yu 2009]

- Let $s : \sum \times \sum \to \mathbb{R}$ be a score function on X, and let g, h be gap penalties. Then the formula

$$d_{\mathcal{S}}(x, y) = \mathcal{S}(x, x) - \mathcal{S}(x, y),$$

where $x, y \in \sum^{+}$ and $\mathcal{S}$ is a global similarity (given by s, g, h) defines a τ-quasi-metric $\mathcal{S}$ on $\sum^{*}$. *This is the quasi-distance induced on* $\sum^{*}$ *by the global similarity function* $\mathcal{S}$.

Theorem 2. [Stojmirovic,Y. Yu 2009]

- Let $\sum$ be a finite alphabet. Suppose $g := h : \mathbb{N}^+ \to \mathbb{R}$ is an increasing function and $s : \sum \times \sum \to \mathbb{R}$ a scoring function, and symmetric, i.e. $s(a, b) = s(b, a)$, $\forall a, b \in \sum$.
  Let $\mathcal{H}$ be the local similarity with respect to s, g, h.
  Then, a function $d : \sum^* \times \sum^* \to \mathbb{R}_+$ given by

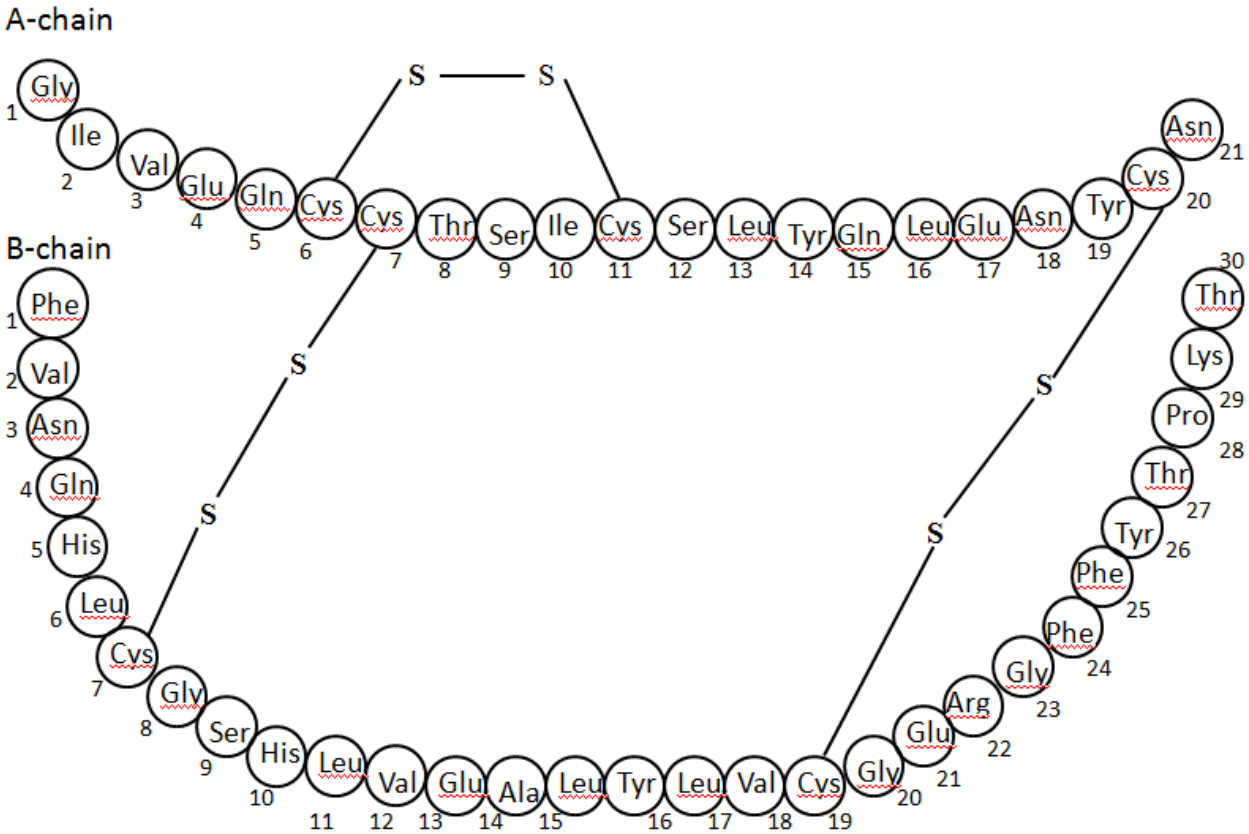
$$d_{\mathcal{H}}(x, y) = \mathcal{H}(x, x) - \mathcal{H}(x, y)$$

  is a co-weightable quasi-metric with co-weight

$$w : \sum^* \to \mathbb{R}_+,\ w(x) = \mathcal{H}(x, x).$$

  This is the co-weightable quasi-metric induced on $\sum^*$ by the local similarity function $\mathcal{H}$.

Insulin Evolution



- 91 insulin homologous sequences (NCBI) belonging to 73 different organisms

↓

- 60 non-redundant insulin homologous sequences

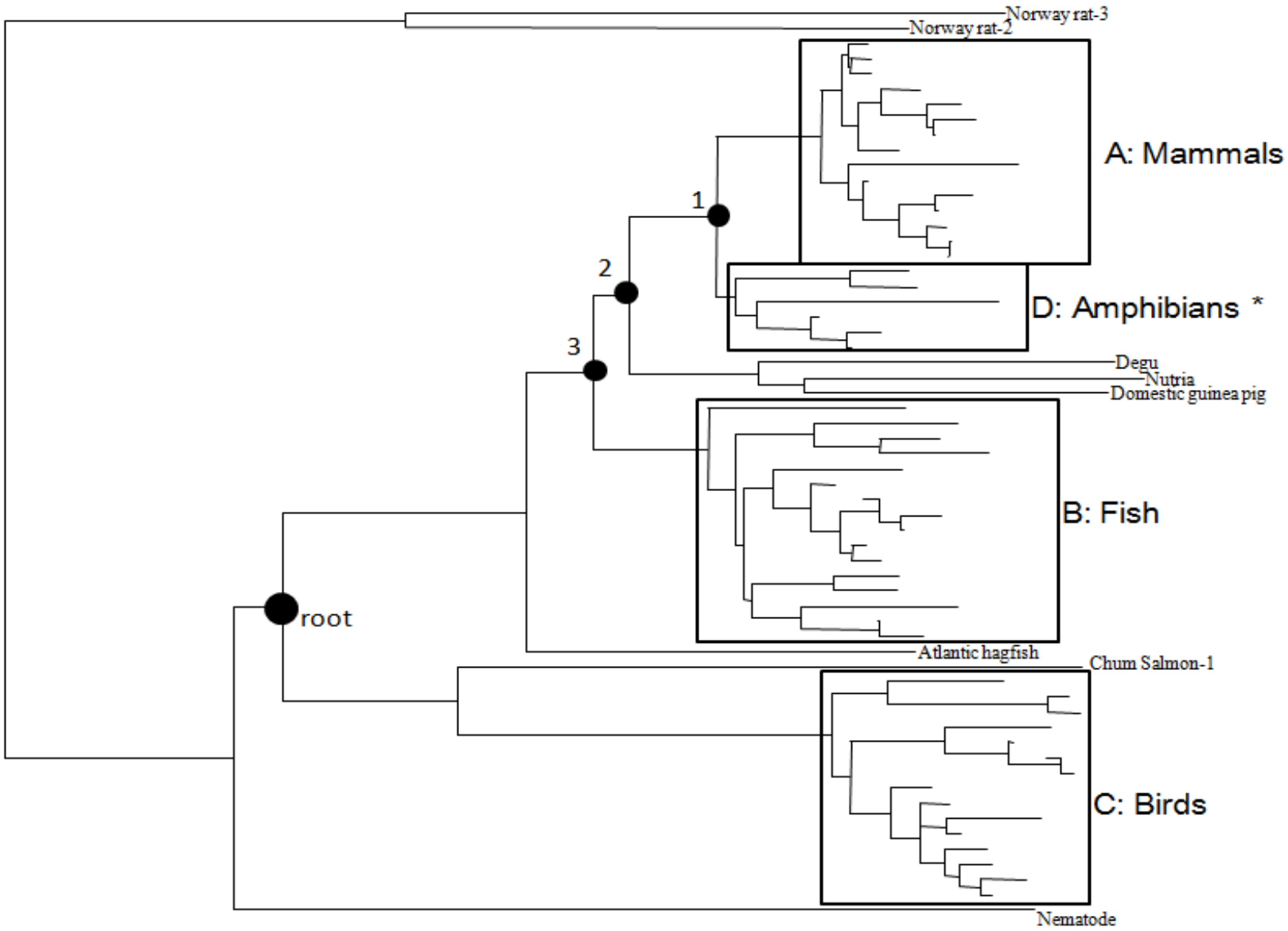
Example

- Human insulin (Acc. No. NP_000198)
- Chimpanzee (Acc. No. P30410)
- Savanna Monkey (Acc. No. CAA43405)
- Crab-eating macaque (Acc. No. AAA36849)

have identical sequence:

- B chain
  –FVNQHLCGSHLVEALYLVCGERGFFYTPKT
- A chain
  GIVEQCCTSICSLYQLENYCN——

A rooted phylogenetic tree (ClustalX)



Conclusions

Even though there are differences when using different metrics, the topology of $\sum^*$ is basically the same fact proving that there is a high degree of consistency between the geometry and topology of the insulin homologous sequences space.

The phylogenetic analysis clustered our 60 sequences in 5 clades with good biological significance.

For organisms in clade A, namely in the domain extending to 22.5, 25 and 23 evolutionary units from the human insulin sequence (Acc. No. NP_000198), with respect to the average, forward and backward distance, respectively, the insulin binding site is well preserved and therefore the function is unaltered.

The binding site structure seems to be preserved up to 40.5, 39 and 42 evolutionary units from the human insulin, but no further with respect to the average, forward and backward distance, respectively (birds in clade D).

At a distance of 103, 100 and 106 evolutionary units from the human insulin sequence, the protein structure changes dramatically, its effectiveness decreasing 5%.

References on Bioinformatics

- J. G. Henickoff, and S. Henickoff, Amino acids substitution matrices from protein blocks. Proc. Natl. Acad. Sci., 89, 10915-10919, 1992.
- Y. Kunihiro and V. S. Sabau, Quasi-metrics. Geometry of Sequence Comparison, Proceedings of Asia Symposium on Engineering and Information, 2013, p.186-196.
- J. Pevsner, Bioinformatics and Functional Genomics, Second Edition, Wiley-Liss, 2003.
- S. V. Sabau, K. Shibuya and H. Shimada, Metric structures associated to Finsler metrics, Publ. Math. Debrecen, 84 (2014), no. 1-2, 89-103.

About the reporting assignment

1. Question 1. What is the triangle inequality and why it is important in any geometrical theory. As an application, please consider the following problem.
   Let ABC be a triangle in plane with the sides length AB=5 cm, BC=7 cm and AC=12 cm. We denote by AD the height of the triangle, where D is the foot of the perpendicular from A to BC. Compute the length of the height AD.
2. Question 2. Perform an internat search about the human insulin. Summarize as many information you can get about the evolution of human insulin in relation to other mammals.
3. Question 3. What did you find interesting in this lecture?

Submit by email to Sabau
sorin@tokai.ac.jp
until 27-th February, 2018.