AMAT 342 Lecture 11. 10/1/19

Today: Metric Spaces, Continued
- more examples.
- Open sets and continuity.

<u>Recall</u>: A metric space is a pair (S,d), where S is a set and $d: S\times S \to [0,\infty)$ is a function such that
1) $d(x,y)=0$ iff $x=y$,
2) $d(x,y)=d(x,y)$
3) $d(x,z) \leq d(x,y)+d(y,z)$.

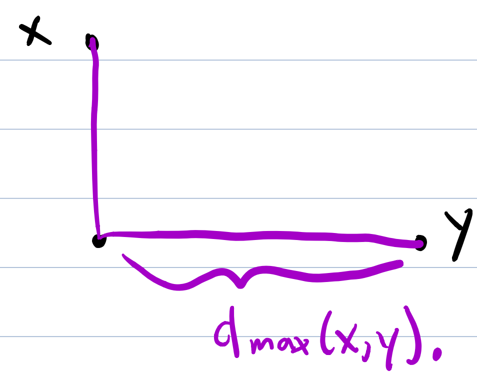
<u>Examples from last time</u>:

- The (usual) Euclidean metric d_2 on $\mathbb{R}^n$
$$d_2(x,y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$$
d_2 is sometimes called the ℓ^2-metric

- The "taxicab metric" d_1 on $\mathbb{R}^n$
$$d_1(x,y)=\sum_{i=1}^{n}|x_i-y_i|$$
(a.k.a. the ℓ^1-metric)

<u>Example</u>: $S=\mathbb{R}^n$, $d_{max}:\mathbb{R}^n\times\mathbb{R}^n \to [0,\infty)$,

$$d_{max}(x,y)=\max\left(|x_1-y|,|x_2-y_2|,\ldots,|x_n-y_n|\right)$$



This is also a metric.

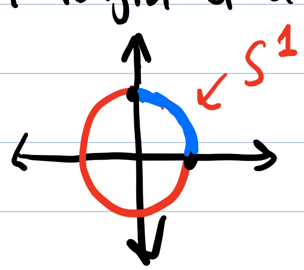
<u>Fact</u>: If (M, d^M) is a metric space, $S \subset M$, and $d^S : S \times S \to [0, \infty)$ is the restriction of d^M to $S \times S$ (i.e., $d^S(x,y) = d^M(x,y)\ \forall x, y \in S$), then (S, d^S) is a metric space.

That is, subspaces of metric spaces inherit the structure of a metric space in the obvious way.

<u>Note</u>: In applications, the subsets are often finite

But in many cases, there is another construction of a metric on a subspace, the <u>intrinsic</u> <u>metric</u>.

<u>Example</u>: Define a metric d on S^1 by $d(x,y)$ = minimum length of an arc in S^1 connecting x and y.



This is called the <u>intrinsic</u> <u>metric</u> on S^1.

e.g. $d((1,0),(0,1)) = \frac{\pi}{2}$ because

minimum length of an arc from $(1,0)$ to $(0,1)$ is $\frac{1}{4}(\text{circumference of } S^1) = \frac{2\pi}{4} = \frac{\pi}{2}$.

By comparison $d_2((1,0),(0,1)) = \sqrt{1^2+1^2} = \sqrt{2}$.

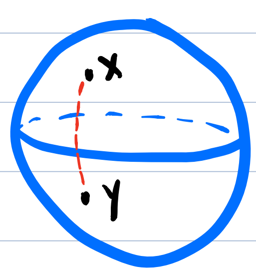
length of the straight line connecting $(1,0)$ and $(0,1)$

More generally, the intrinsic metric d can be defined on a very large class of subsets $S \subset \mathbb{R}^n$ as follows:

$d(x,y)$ = minimum length of a differentiable path $\gamma : I \to S$ from x to y. (Since codomain of γ is S, $\operatorname{im}(\gamma)$ is required to lie in S.)

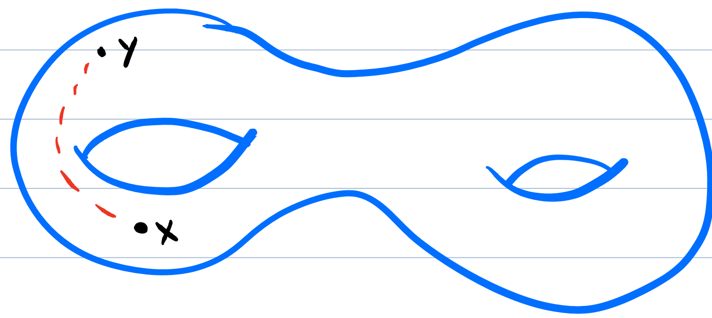
As in calculus, $\operatorname{length}(\gamma) := \int_0^1 |\gamma'(t)| \, dt$.

For example, we can take S to be a sphere in $\mathbb{R}^3$



← $d(x,y)$ is the length of the shortest curve connecting x and y.

or any other surface in $\mathbb{R}^3$.



<u>Fact</u>: On $S^1 \subset \mathbb{R}^2$, the intrinsic metric given by the general definition is equal to the version for S^1 defined earlier.

This fact is proven, in more generality, in a course on differential geometry

Example of a metric space from biology

<u>Background</u>: A DNA molecule consists of of two chains of subunits.

- the subunits are called <u>nucleotides</u>
- there are four nucleotides:
  - Cytosine [denoted C]
  - Adenine [A]
  - Guanine [G]
  - Thymine [T]

The two chains are bound together (by weak hydrogen bonds).

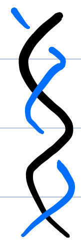
- The i^{th} nucleotides in the two chains are bonded
- G binds to C, A bonds to T. Thus one chain determines the other!

Illustration: C – G – A – A – T – A – C
(schematic)
¦ ¦ ¦ ¦ ¦ ¦ ¦
G – C – T – T – A – T – G

Solid lines = covalent bonds (strong)
Dashed lines = hydrogen bonds (weak)

Can represent this more compactly as: $\boxed{\text{CGAATAC}}$ ← called a "DNA sequence"
(bottom chain is determined by the top).

The two chains wind around each other, forming a "double helix"



Fundamental question: How do we quantify the similarity between two DNA sequences?

- This is relevant to the study of evolution:
  - close relatives should have similar DNA
  - distant relatives should have dissimilar DNA

Classical solution: Use the **edit metric**.

Before giving the definition, let's motivate it with examples.

Ex: Consider the two DNA sequences

CGATTGC

CAATTGT

These differ in two spots, so we'd like to say their distance is 2.

Ex: CGATTGC

CGCATTGC

These differ by the insertion of one element, so we'd like to say that the distance is 1.

Let S denote the set consisting of sequences of the letters A, C, G, T (of any length ≥ 0).

For $x \in S$, an <u>elementary operation</u> on x is any one of the following operations:

- replace one letter in the sequence by a different one,
- remove one letter from any one position in the sequence,
- add one letter at any one position in the sequence.

<u>Definition of the edit distance</u>:

define $d_{edit}: S \times S \to [0, \infty)$ by

$d_{edit}(x,y)$ = minimum number of elementary operations need to transform x into y.

[Lecture ended here.]

<u>Let's verify that this is a metric:</u>

- Property 1) is clearly satisfied
- An elementary operation can always be undone by an elementary operation, so $d_{edit}(x,y) = d_{edit}(y,x)$.
- If α is a sequence of m elt. ops, transforming x into y, and β is a sequence of n elt. ops. transforming y into z, then α followed by β is a sequence of $m+n$ elt. ops. transforming x into z. We can choose α, β s.t. $m = d(x,y)$ and $n = d(y,z)$. Then α followed by β is a sequence of $d(x,y) + d(y,z)$ elt. ops transforming x into z. It now follows that $d(x,z) \leq d(x,y) + d(x,z)$. ∎

let's denote d_{edit} as d.

<u>Examples:</u>

$x = AAAA$ $\quad d_{edit}(x,y) = 4$

$y = TTTT$ (at most one T can be created per el. op.)

$x = ACTG$ $\quad d_{edit}(x,y) = 2$

$y = GACT$ $\quad ACTG \to ACT \to GACT$

Remark: The definition of edit distance generalizes to any set of symbols. For example, the set of symbols could be the entire alphabet. Then, the problem of spell-checking a string of letters x can be formalized (very naively) as the problem of fining a word in the dictionary closest in edit distance to x.

Remark: Note that d_{edit} is integer-valued.

Another example of a metric space from biology

Background: The primary function of DNA is to serve as a blue-print from which proteins are constructed.

Simplified definition of a protein:
A protein is a string of subunits called amino acids connected by covalent bonds.

There are 20 different amino acids, with names like "arginine," "lysine," and "tryptophan."

Protein fold into complex 3-D structures , with essential biological function (e.g. enzymes, neurotransmitters)

DNA sequences called <u>Genes</u> specify the amino acid sequence of protein.

<u>Rough explanation</u>: Three nucleotides specify one amino acid.

<u>Ex</u>: CGA TTT ACC

⇓

Alanine ~ Lysine ~ Tryptophan

Determining the amino acid sequence from the DNA sequence is very easy.

But accurately determining the 3-D structure of the protein from the amino acid sequence is challenging.

This is called the "protein structure prediction problem."

- one of the fundamental problems of computational biology
- applications to drug discovery
- anual competions on this problem.
- lots of software available.

Note: In favorable cases, the structure can be determined by experiment, e.g., by a technique called x-ray cystalography. But this is expensive, time consuming, and requires a lot of skill.

Question: Suppose I know the folded structure of