

# AMAT 583 Lecture II 10/1/19

Recall: A metric space is a pair  $(S, d)$ , where  $S$  is a set and  $d: S \times S \rightarrow [0, \infty)$  is a function such that

- 1)  $d(x, y) = 0$  iff  $x = y$ ,
- 2)  $d(x, y) = d(y, x)$
- 3)  $d(x, z) \leq d(x, y) + d(y, z)$ .

Examples from last time:

- 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<sub>2</sub> is sometimes called the l<sup>2</sup>-metric*

- The "taxicab metric"  $d_1$  on  $\mathbb{R}^n$

$$d_1(x, y) = \sum_{i=1}^n |x_i - y_i| \quad (\text{a.k.a. the } l^1\text{-metric})$$

- $S = \mathbb{R}^n$ ,  $d_{\max}: \mathbb{R}^n \times \mathbb{R}^n \rightarrow [0, \infty)$ ,

$$d_{\max}(x, y) = \max(|x_1 - y_1|, |x_2 - y_2|, \dots, |x_n - y_n|)$$

Let's check that  $d_1$  is a metric.

1) Clearly  $d_1(x, x) = 0$  for all  $x \in \mathbb{R}^n$ .

If  $x \neq y$ , then  $x_k \neq y_k$  for some  $k \in \{1, \dots, n\}$   
so  $0 < |x_k - y_k| \leq d_1(x, y)$ , so  $0 < d_1(x, y)$ .

2)  $d_1(x, y) = d_1(y, x)$  because  $|x_k - y_k| = |y_k - x_k|$   
for all  $k \in \{1, \dots, n\}$ .

3)  $d_1(x, z) \leq d_1(x, y) + d_1(y, z)$  because

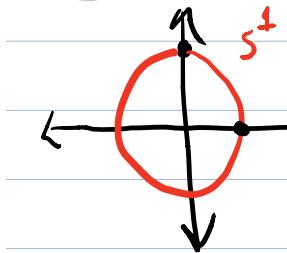
$$|x_k - z_k| \leq |x_k - y_k| + |y_k - z_k|$$

(explanation:  $|a+b| \leq |a| + |b|$ . Take  $a = x_k - y_k$ ,  $b = y_k - z_k$ .)

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

Ex:  $M = \mathbb{R}^2$ ,  $S = S^1$ ,  $d^M = d_2$

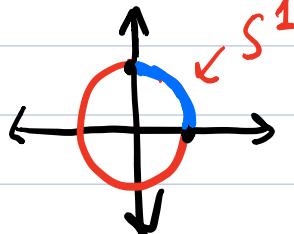


Exercise: What is  $d^S((1,0), (0,1))$ ?  
Ans:  $\sqrt{2}$ .

Note: In applications, the subsets  $S$  are often finite.

In many cases, there is another construction of a metric on a subspace, the intrinsic metric.

Example: Define a metric  $d$  on  $S^1$  by  
 $d(x,y) = \text{minimum length of an arc in } S^1 \text{ connecting } x \text{ and } y.$



This is called the intrinsic metric 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  $d_2((1,0),(0,1)) = \sqrt{1^2+1^2} = \sqrt{2}$ .

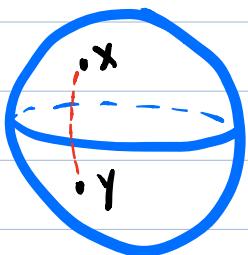
length  
of the  
straight  
line connecting  
(1,0) and (0,1)

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

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

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

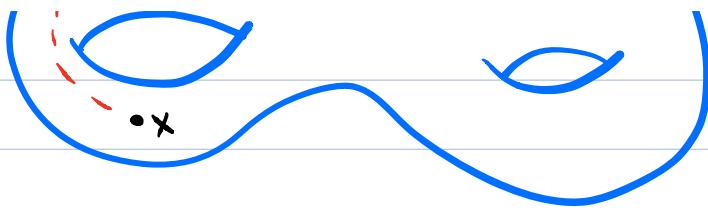
For example, we can take  $S$  to be a sphere in  $\mathbb{R}^3$



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

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





Fact: 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

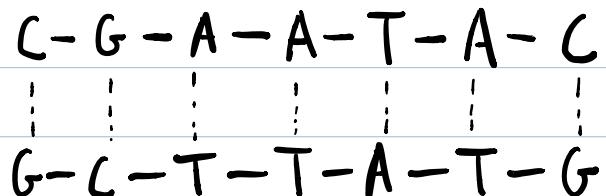
Background: A DNA molecule consists of of two chains of subunits.

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

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

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

Illustration:  
(schematic)



Solid lines = covalent bonds (strong)

Can represent this more compactly as:

Dashed lines = hydrogen bonds (weak)

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

(GATTGC) } These differ in two spots,  
so we'd like to say

(AATTGT) their distance is 2.

Ex: (GATTGC) } These differ by the insertion  
(GCATTGC) } 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  $\geq 0$ ).

For  $x \in S$ , an elementary operation 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.

Definition of the edit distance:

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

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

Let's verify that this is a metric:

- 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  $\alpha$  is a sequence of  $m$  elt. ops, transforming  $x$  into  $y$ , and  $\beta$  is a sequence of  $n$  elt. ops, transforming  $y$  into  $z$ , then  $\alpha$  followed by  $\beta$  is a sequence of  $m+n$  elt. ops. transforming  $x$  into  $z$ . We can choose  $\alpha, \beta$  s.t.

let's denote credit as  $d$ .  $m = d(x,y)$  and  $n = d(y,z)$ . Then  $\alpha$  followed by  $\beta$  is a sequence of  $d(x,y) + d(y,z)$  elt. ops transforming  $x$  into  $z$ . It now follows that  $d(x,z) \leq d(x,y) + d(y,z)$ .  $\blacksquare$

Examples:  $x = \text{AAAA}$   $d_{edit}(x,y) = 4$

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

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

$y = \text{GACT}$   $\text{ACTG} \rightarrow \text{ACT} \rightarrow \text{GACT}$

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 finding a word in the dictionary closest in edit distance to  $x$ .

Remark: Note that 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.

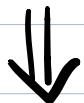
There are 20 different amino acids, with names like "arginine," "lysine," and "tryptophan."

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

DNA sequences called Genes specify the amino acid sequence of protein.

Rough explanation: Three nucleotides specify one amino acid.

Ex: CGA TTTACC



Alanine ~ Lysine ~ Tryptophan

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

accurately

But 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
- annual competitions 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 crystallography. But this is expensive, time consuming, and requires a lot of skill.

Question: Suppose I know the folded structure of a protein  $P$ . How do I measure a predicted structure  $P'$ ?

To assess the performance of a structure prediction method, we need an answer.

Standard Answer: Compute the RMSD  
(root mean squared deviation) between  $P$  and  $P'$ .