Machine Learning

Manifold learning and dimensionality reduction

Emanuele Rodolà
rodola@di.uniroma1.it

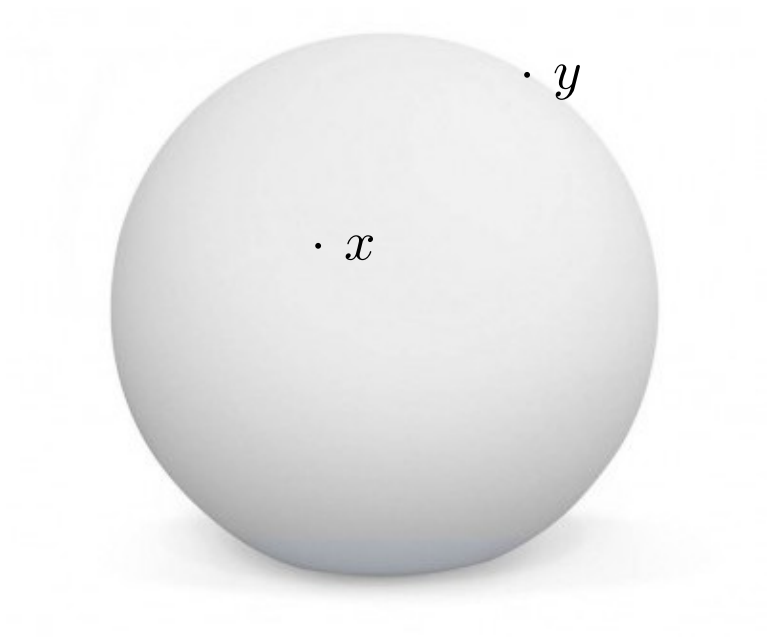
SAPIENZA
UNIVERSITÀ DI ROMA

2nd semester a.y. 2024/2025 · April 28, 2025

Metric spaces

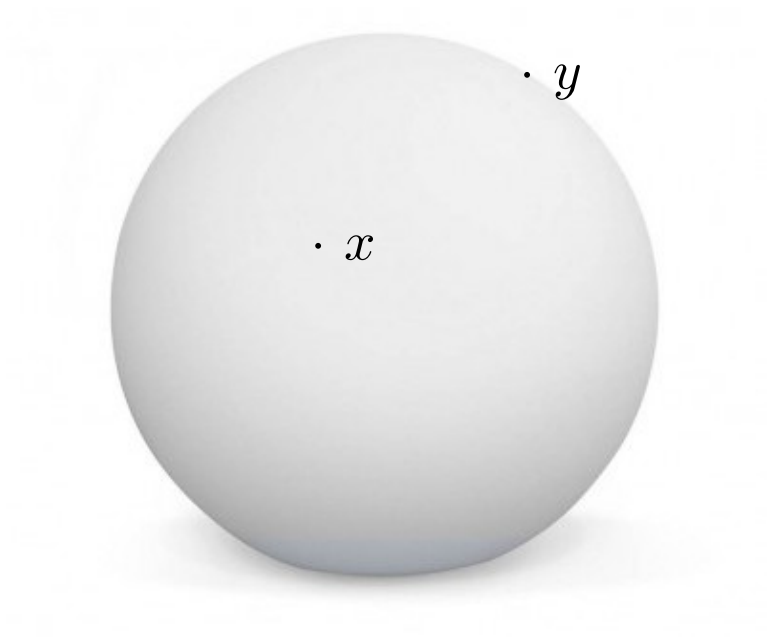
Measuring distance

What's the distance between x and y in this picture?



Measuring distance

What's the distance between x and y in this picture?



There is not a unique way!

- You can pass through the sphere with a straight line (Euclidean)
- You can walk on the surface in a "straight" path (non-Euclidean)

Metric spaces

A set $\mathcal{M}$ with a metric (or distance) function $d_{\mathcal{M}} : \mathcal{M} \times \mathcal{M} \to \mathbb{R}_+$ such that for every $x, y \in \mathcal{M}$:

Metric spaces

A set $\mathcal{M}$ with a metric (or distance) function $d_{\mathcal{M}} : \mathcal{M} \times \mathcal{M} \to \mathbb{R}_+$ such that for every $x, y \in \mathcal{M}$:

- $d_{\mathcal{M}}(x, y) = 0 \Leftrightarrow x = y$ (identity of indiscernibles)
- $d_{\mathcal{M}}(x, y) = d_{\mathcal{M}}(y, x)$ (symmetry)
- $d_{\mathcal{M}}(x, y) \leq d_{\mathcal{M}}(y, z) + d_{\mathcal{M}}(z, x)$ for any $x, y, z \in \mathcal{M}$ (triangle inequality)

Metric spaces

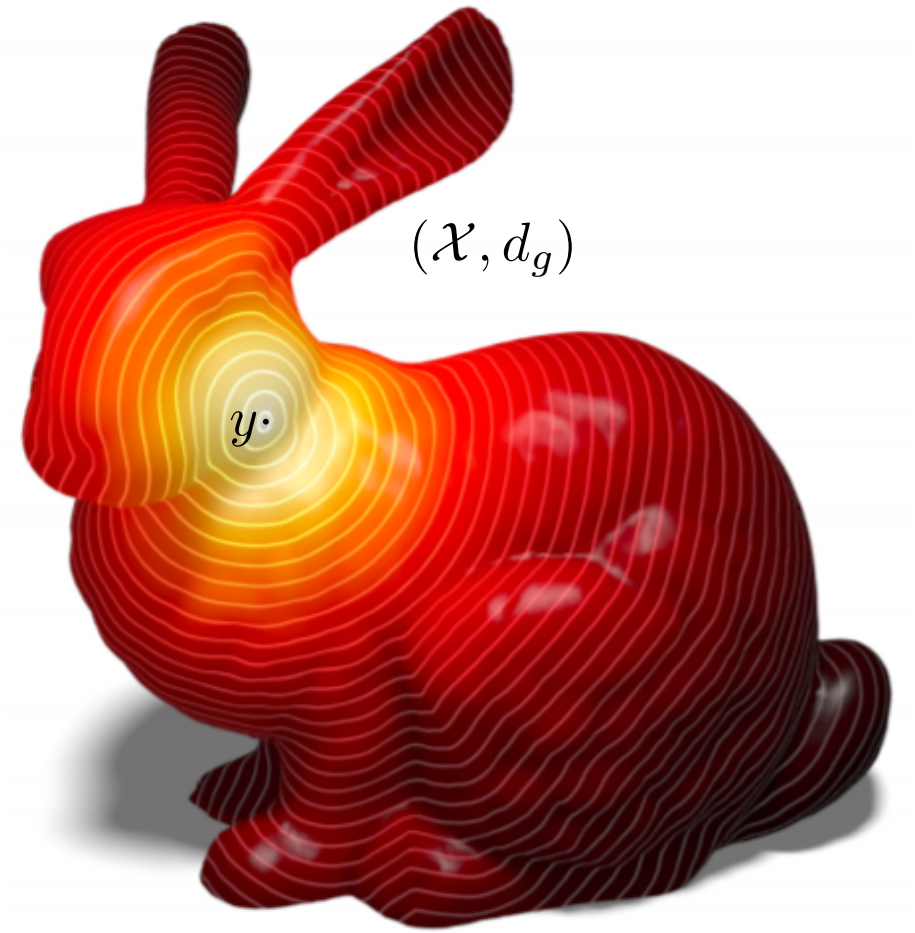
A set $\mathcal{M}$ with a metric (or distance) function $d_{\mathcal{M}} : \mathcal{M} \times \mathcal{M} \to \mathbb{R}_+$ such that for every $x, y \in \mathcal{M}$:

- $d_{\mathcal{M}}(x, y) = 0 \Leftrightarrow x = y$ (identity of indiscernibles)
- $d_{\mathcal{M}}(x, y) = d_{\mathcal{M}}(y, x)$ (symmetry)
- $d_{\mathcal{M}}(x, y) \leq d_{\mathcal{M}}(y, z) + d_{\mathcal{M}}(z, x)$ for any $x, y, z \in \mathcal{M}$ (triangle inequality)

Examples:

- The sphere with Euclidean distance is $(\mathbb{S}^2, d_{L_2})$
- The sphere with geodesic distance is $(\mathbb{S}^2, d_g)$

Example: Geodesic isolines



Each isoline identifies a set of points $x \in \mathcal{X}$ at the same distance (according to d_g) from some reference $y \in \mathcal{X}$

Examples: Metric spaces

- $\mathcal{X} = \mathbb{R}\,, \quad d_{\mathcal{X}}(x, y) = |x - y|$

Examples: Metric spaces

- $\mathcal{X} = \mathbb{R}\,, \quad d_{\mathcal{X}}(x, y) = |x - y|$
- $\mathcal{X} = \mathcal{A} \subset \mathbb{R}^k\,, \quad d_{\mathcal{X}}(x, y) = \|x - y\|_2$

Examples: Metric spaces

- $\mathcal{X} = \mathbb{R}\,, \quad d_{\mathcal{X}}(x, y) = |x - y|$
- $\mathcal{X} = \mathcal{A} \subset \mathbb{R}^k\,, \quad d_{\mathcal{X}}(x, y) = \|x - y\|_2$
- $\mathcal{X} = \mathbb{R}\,, \quad d_{\mathcal{X}}(x, y) = \log(|x - y| + 1)$

Examples: Metric spaces

- $\mathcal{X} = \mathbb{R}\,, \quad d_{\mathcal{X}}(x, y) = |x - y|$
- $\mathcal{X} = \mathcal{A} \subset \mathbb{R}^k\,, \quad d_{\mathcal{X}}(x, y) = \|x - y\|_2$
- $\mathcal{X} = \mathbb{R}\,, \quad d_{\mathcal{X}}(x, y) = \log(|x - y| + 1)$
- $\mathcal{X} = \text{any set}\,, \quad d_{\mathcal{X}}(x, y) = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}$

Examples: Metric spaces

- $\mathcal{X} = \mathbb{R}\,, \quad d_{\mathcal{X}}(x, y) = |x - y|$
- $\mathcal{X} = \mathcal{A} \subset \mathbb{R}^k\,, \quad d_{\mathcal{X}}(x, y) = \|x - y\|_2$
- $\mathcal{X} = \mathbb{R}\,, \quad d_{\mathcal{X}}(x, y) = \log(|x - y| + 1)$
- $\mathcal{X} = \text{any set}\,, \quad d_{\mathcal{X}}(x, y) = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}$
- $\mathcal{X} = \mathcal{A} \times \mathcal{B}\,, \quad d_{\mathcal{X}}((a_1, b_1), (a_2, b_2)) = \sqrt{d_{\mathcal{A}}(a_1, a_2)^2 + d_{\mathcal{B}}(b_1, b_2)^2}$

Ambient space and restriction

If $\mathcal{A}$ is a metric space and $\mathcal{X} \subset \mathcal{A}$, then $\mathcal{A}$ is called ambient space for $\mathcal{X}$.

$$\mathcal{A} = (\mathbb{R}^3, d_{L_2})$$

Ambient space and restriction

If $\mathcal{A}$ is a metric space and $\mathcal{X} \subset \mathcal{A}$, then $\mathcal{A}$ is called ambient space for $\mathcal{X}$.

$$\mathcal{A} = (\mathbb{R}^3, d_{L_2})$$

A metric on $\mathcal{X}$ can be obtained by the restriction $d_{\mathcal{X}} = d_{\mathcal{A}|\mathcal{X}}$, such that:

$$d_{\mathcal{X}}(x, y) = d_{\mathcal{A}}(x, y)$$

for all $x, y \in \mathcal{X}$

Isometric embeddings

Isometries

Let $(\mathcal{M}, d_{\mathcal{M}})$ and $(\mathcal{N}, d_{\mathcal{N}})$ be two metric spaces.

A bijective map $f : \mathcal{M} \to \mathcal{N}$ is called an isometry if:

$$d_{\mathcal{M}}(x, y) = d_{\mathcal{N}}(f(x), f(y))$$

for any $x, y \in \mathcal{M}$.

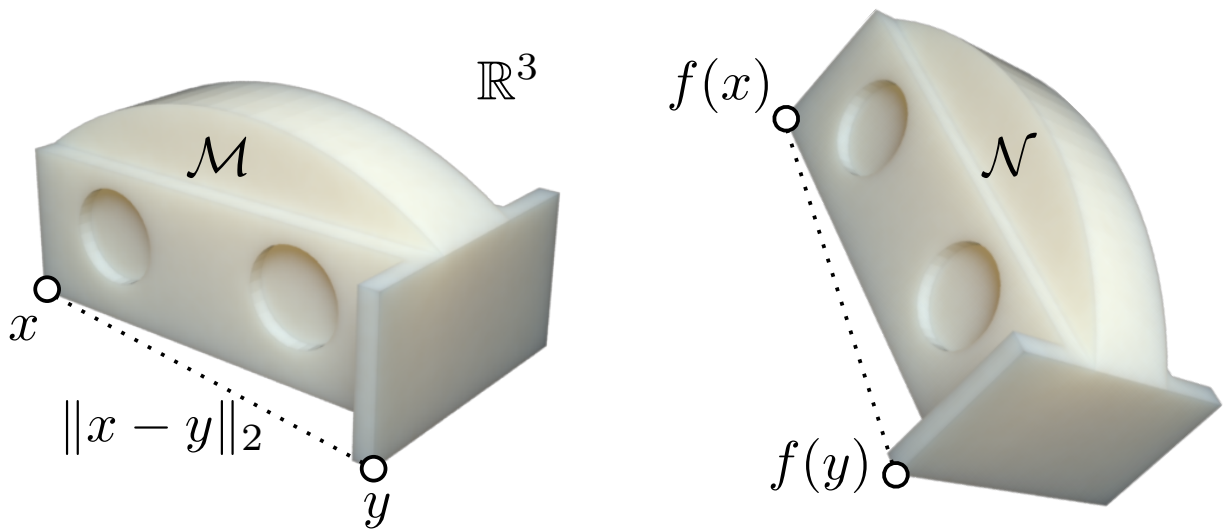
Isometries

Let $(\mathcal{M}, d_{\mathcal{M}})$ and $(\mathcal{N}, d_{\mathcal{N}})$ be two metric spaces.

A bijective map $f : \mathcal{M} \to \mathcal{N}$ is called an isometry if:

$$d_{\mathcal{M}}(x, y) = d_{\mathcal{N}}(f(x), f(y))$$

for any $x, y \in \mathcal{M}$.



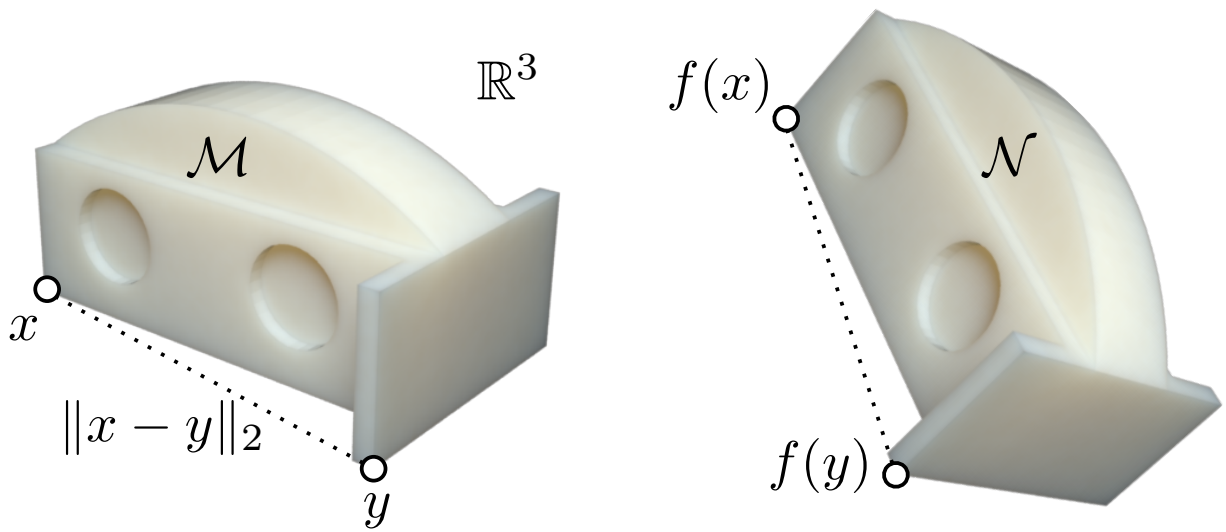
Isometries

Let $(\mathcal{M}, d_{\mathcal{M}})$ and $(\mathcal{N}, d_{\mathcal{N}})$ be two metric spaces.

A bijective map $f : \mathcal{M} \to \mathcal{N}$ is called an isometry if:

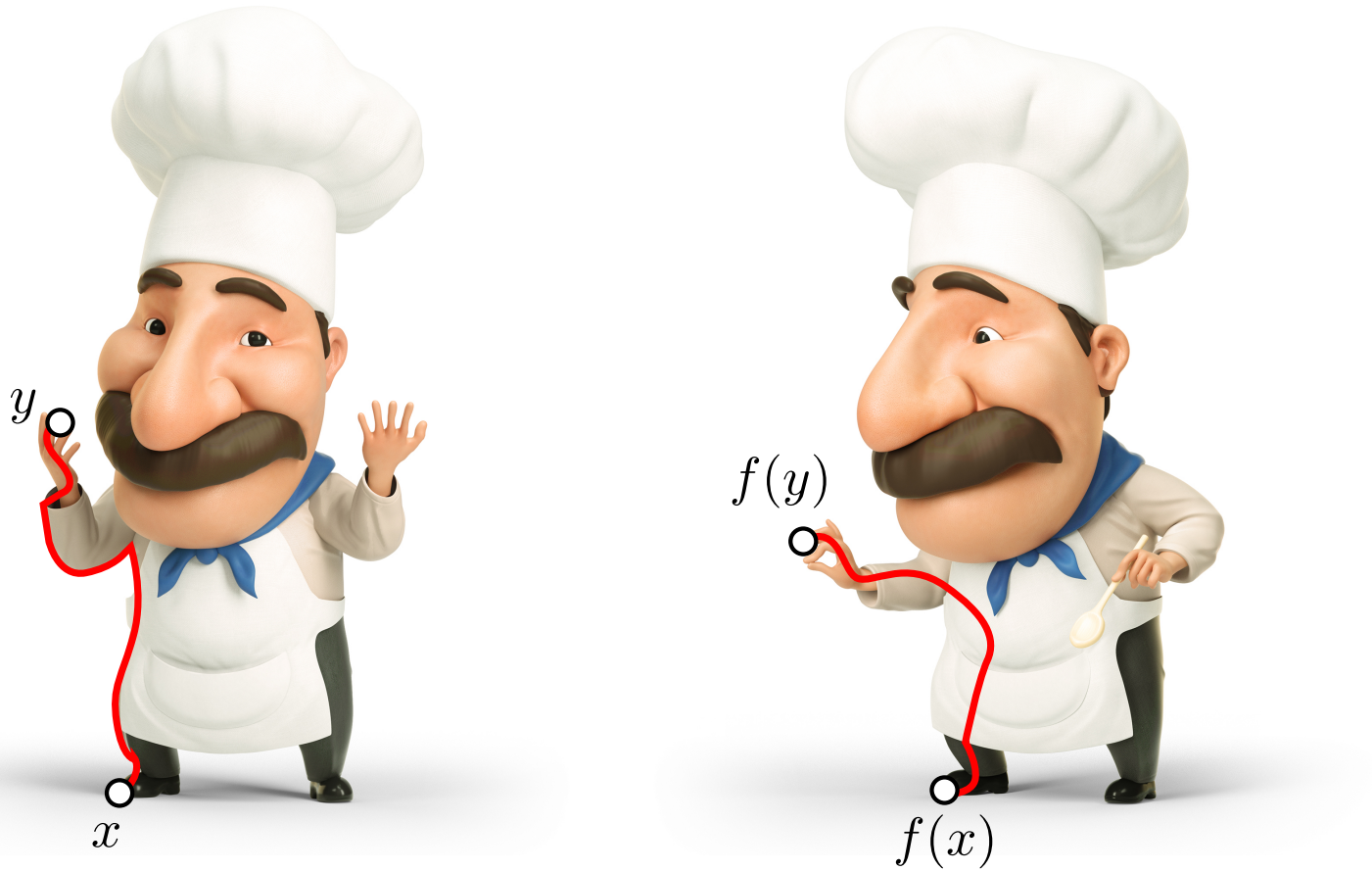
$$d_{\mathcal{M}}(x, y) = d_{\mathcal{N}}(f(x), f(y))$$

for any $x, y \in \mathcal{M}$.

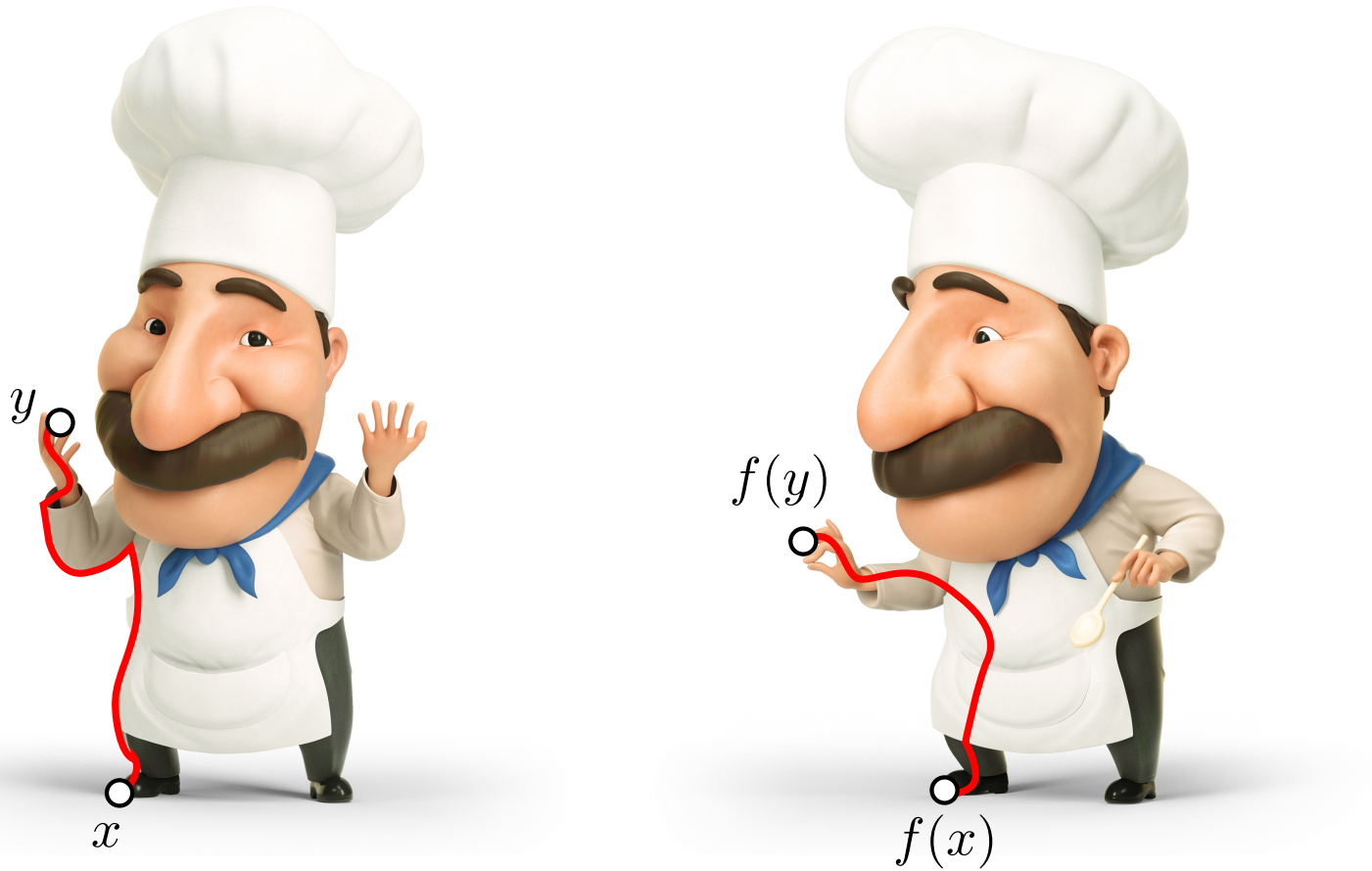


If $d_{\mathcal{M}} = \|\cdot\|_2$ and $d_{\mathcal{N}} = \|\cdot\|_2$ we say "rigid isometry"

Example: Non-rigid "quasi"-isometries



Example: Non-rigid "quasi"-isometries



$$d_{\mathcal{M}}(x, y) \approx d_{\mathcal{N}}(f(x), f(y))$$

(here $d_{\mathcal{M}}, d_{\mathcal{N}}$ are geodesic distance functions)

Isometry as equivalence

"Being isometric" is an equivalence relation, since it is:

- reflective $(a = a)$
- symmetric $(a = b \Rightarrow b = a)$
- transitive $(a = b \wedge b = c \Rightarrow a = c)$

Isometry as equivalence

"Being isometric" is an equivalence relation, since it is:

- reflective $(a = a)$
- symmetric $(a = b \Rightarrow b = a)$
- transitive $(a = b \wedge b = c \Rightarrow a = c)$

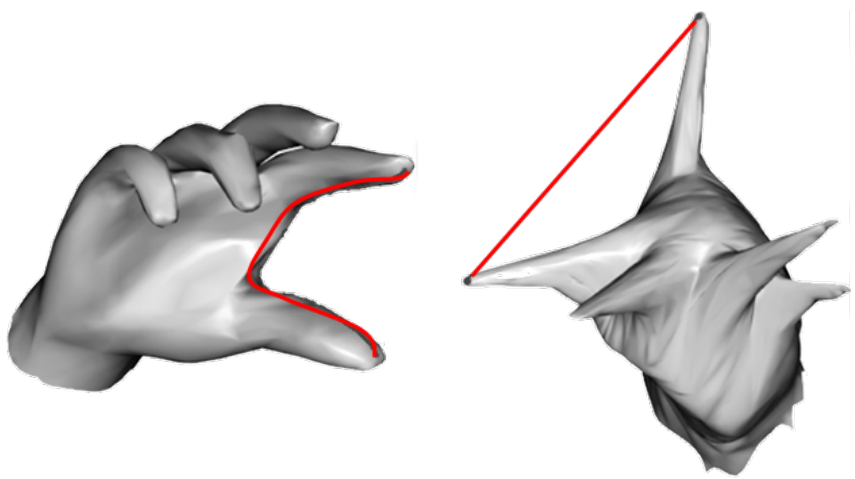
In this sense, we think of isometric shapes as being the same shape:

Isometric embeddings

An isometric embedding is a transformation $f : \mathcal{X} \to \mathcal{Z}$ which preserves the metric for all pairs $x, y \in \mathcal{X}$, i.e.

$$d_{\mathcal{Z}}(f(x), f(y)) = d_{\mathcal{X}}(x, y)$$

For example, take $d_{\mathcal{X}} = d_g$ and $d_{\mathcal{Z}} = \|\cdot\|_2$:

$(\mathcal{X}, d_g)$  $(\mathbb{R}^3, \|\cdot\|_2)$

A cartographer's problem

Is it always possible?

Consider the following:

A cartographer's problem

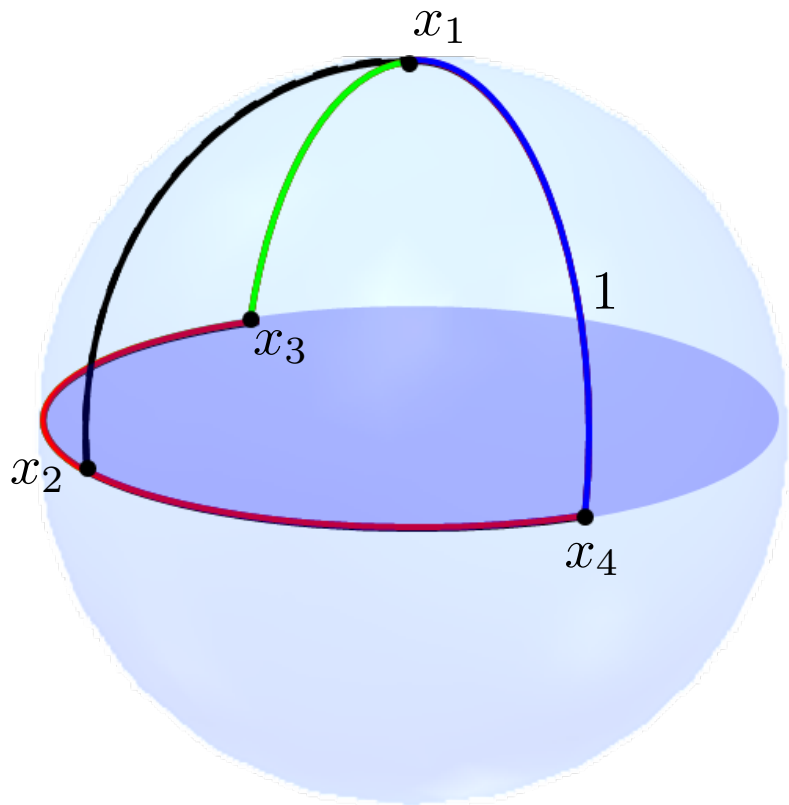
Is it always possible?

Consider the following:

An isometric embedding of $\mathbb{S}^2$ into $\mathbb{R}^2$ is not possible!

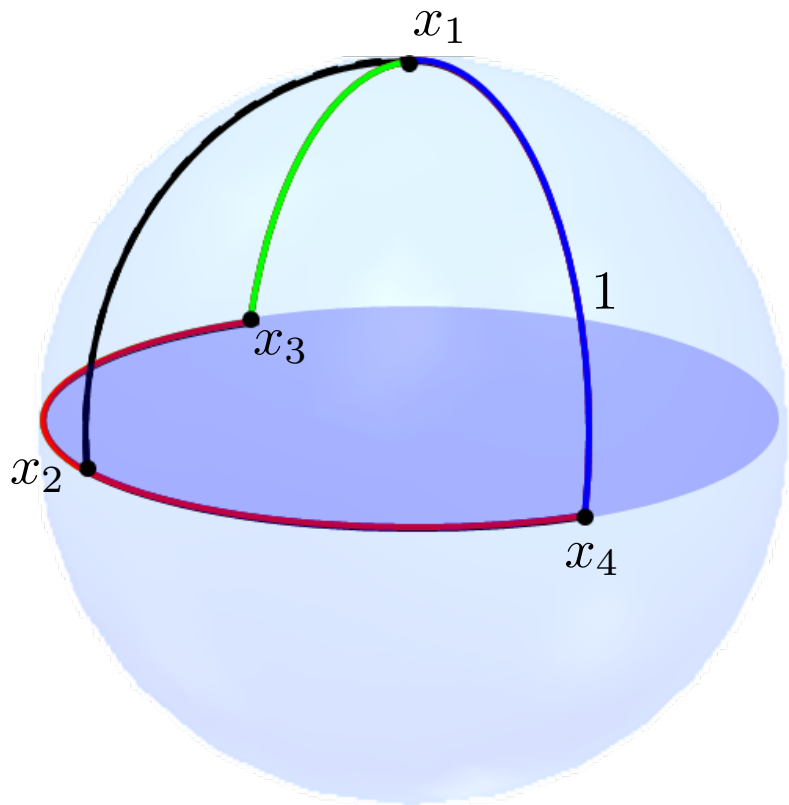
Any approximate solution introduces metric distortion.

Non-embeddability of the sphere



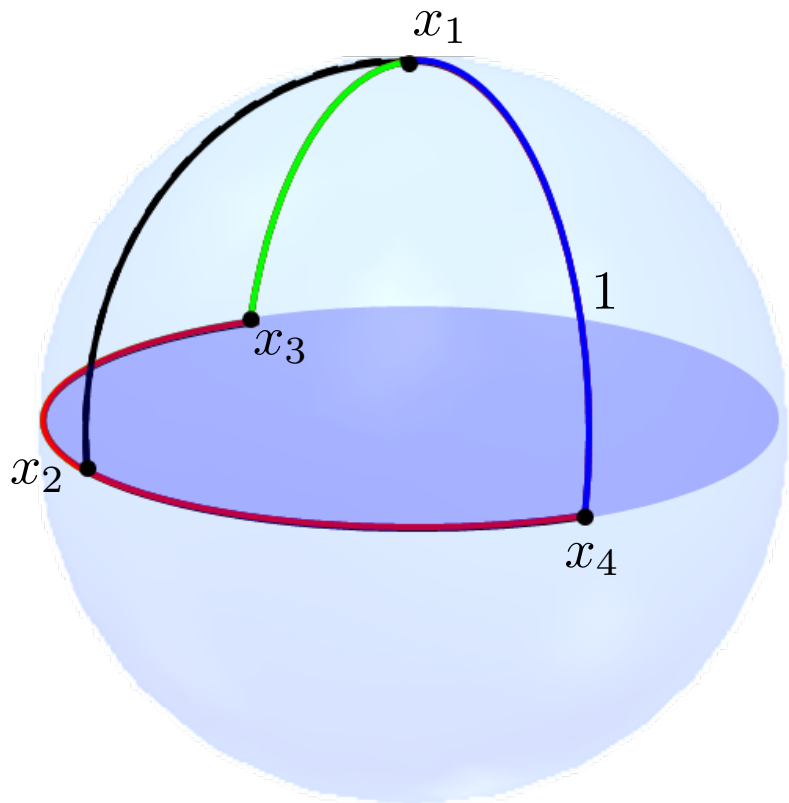
- Consider the triangle $\Delta(x_1, x_3, x_4)$

Non-embeddability of the sphere



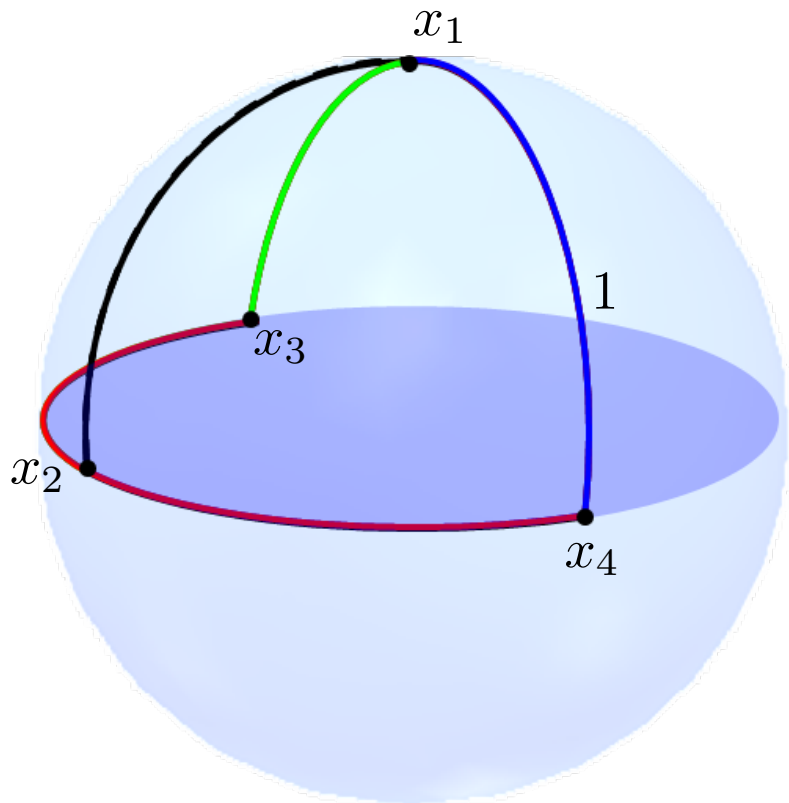
- Consider the triangle $\Delta(x_1, x_3, x_4) \Rightarrow$ collinear!

Non-embeddability of the sphere



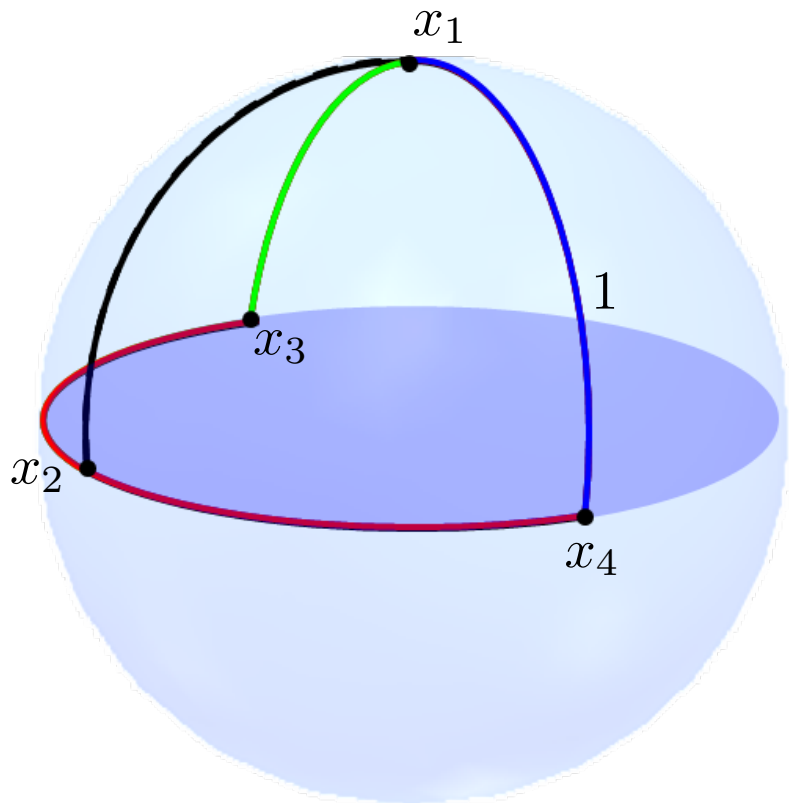
- Consider the triangle $\Delta(x_1, x_3, x_4) \Rightarrow$ collinear!
- Consider the triangle $\Delta(x_2, x_3, x_4)$

Non-embeddability of the sphere



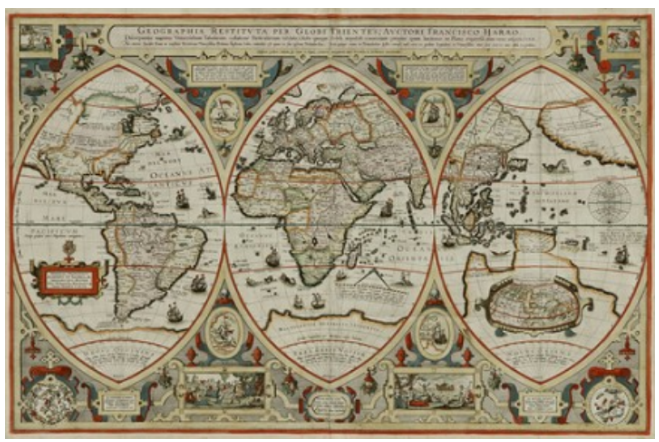
- Consider the triangle $\Delta(x_1, x_3, x_4) \Rightarrow$ collinear!
- Consider the triangle $\Delta(x_2, x_3, x_4) \Rightarrow$ collinear!

Non-embeddability of the sphere



- Consider the triangle $\Delta(x_1, x_3, x_4) \Rightarrow$ collinear!
- Consider the triangle $\Delta(x_2, x_3, x_4) \Rightarrow$ collinear!
- Then $x_1 = x_2$, which contradicts $d_g(x_1, x_2) = 1$
  $\Rightarrow$ This metric space cannot be embedded into $\mathbb{R}^k$ for any k

A cartographer's solution



Teaser exercise: Matrix calculus

Let matrix $\mathbf{X} \in \mathbb{R}^{n \times 3}$ contain the 3D coordinates of points x_i as its rows.

Consider the following expression:

$$n \sum_{i=1}^{n} \langle x_i, x_i \rangle - \sum_{i,j} \langle x_i, x_j \rangle$$

How do you write the expression above in matrix notation?

Hint: Use the trace operation, defined as $\mathrm{tr}(\mathbf{X}) = \sum_{i=1}^{n} x_{ii}$

Metric distortion

Embeddings

We seek an f such that:

$$d_{\mathcal{M}}(x, y) = d_{\mathcal{N}}(f(x), f(y))$$

But we know that the sphere cannot be embedded into $\mathbb{R}^k$ for any k

Embeddings

We seek an f such that:

$$d_{\mathcal{M}}(x, y) = d_{\mathcal{N}}(f(x), f(y))$$

But we know that the sphere cannot be embedded into $\mathbb{R}^k$ for any k

Q1: Can we embed it anyway, with distortion?

Embeddings

We seek an f such that:

$$d_{\mathcal{M}}(x, y) = d_{\mathcal{N}}(f(x), f(y))$$

But we know that the sphere cannot be embedded into $\mathbb{R}^k$ for any k

Q1: Can we embed it anyway, with distortion?

Further, suppose we don't know the positions of x, y, but only their pairwise distances.

Embeddings

We seek an f such that:

$$d_{\mathcal{M}}(x, y) = d_{\mathcal{N}}(f(x), f(y))$$

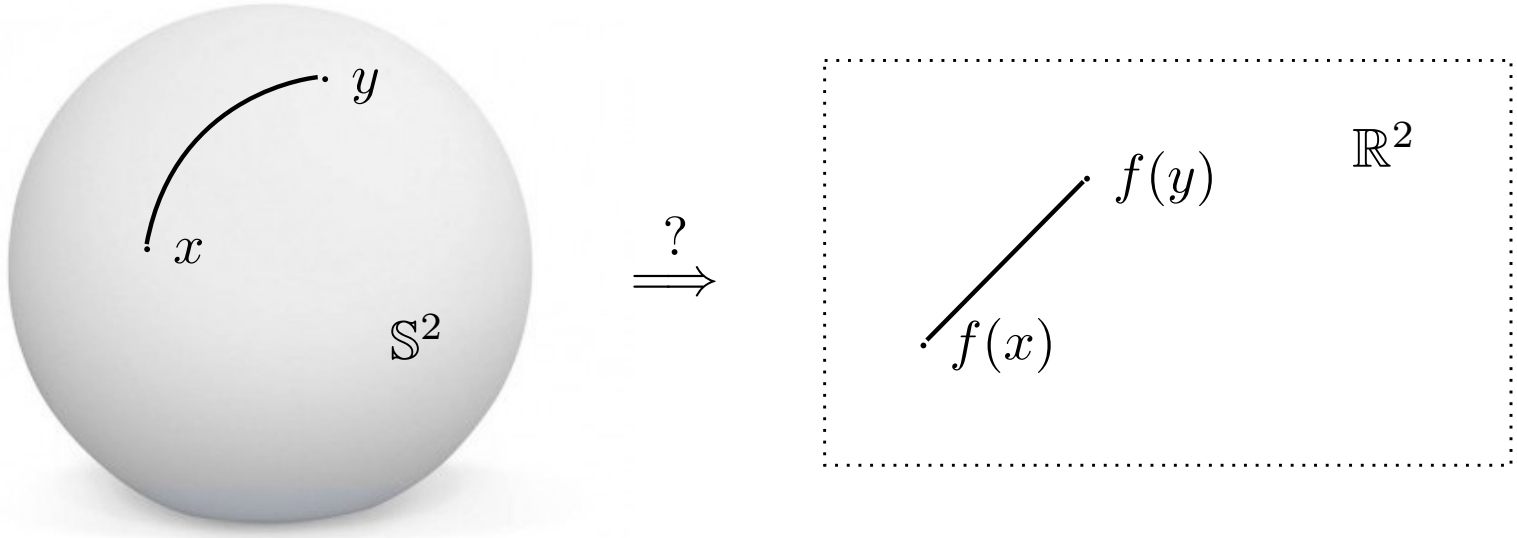
But we know that the sphere cannot be embedded into $\mathbb{R}^k$ for any k

Q1: Can we embed it anyway, with distortion?

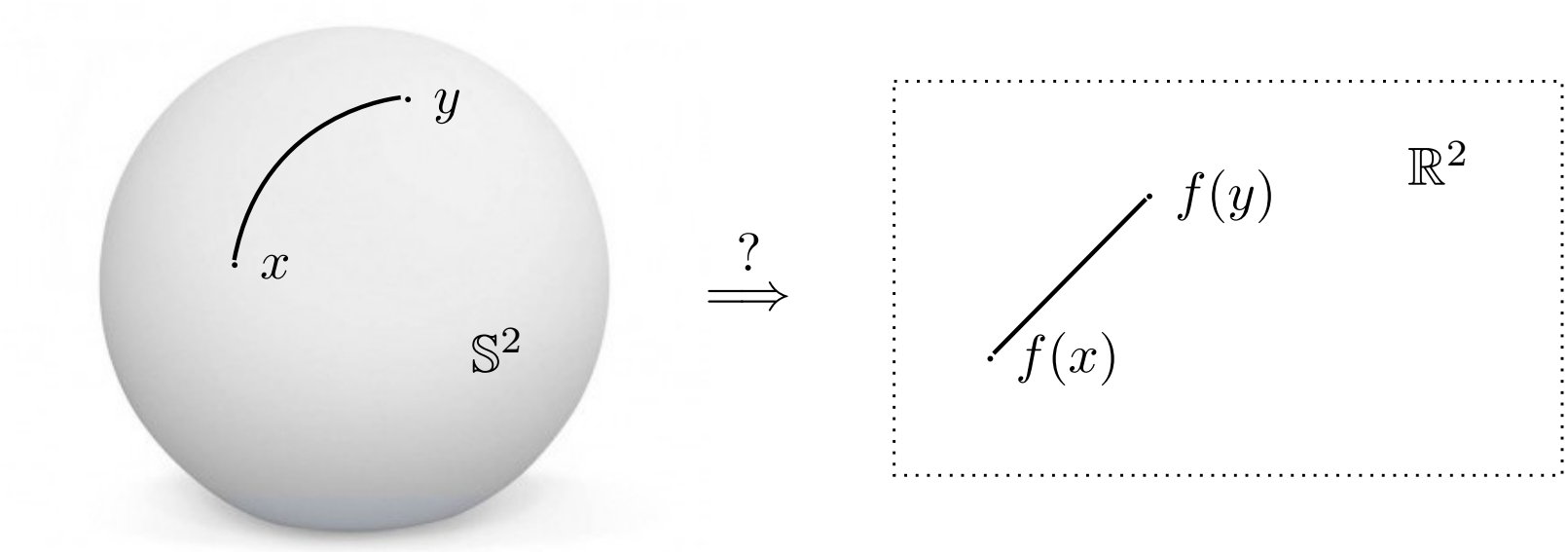
Further, suppose we don't know the positions of x, y, but only their pairwise distances.

Q2: Can we recover the point coordinates?

Embeddings: Q1 example



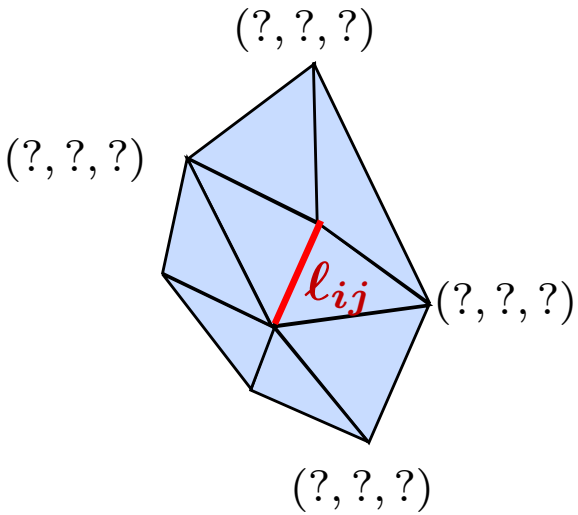
Embeddings: Q1 example



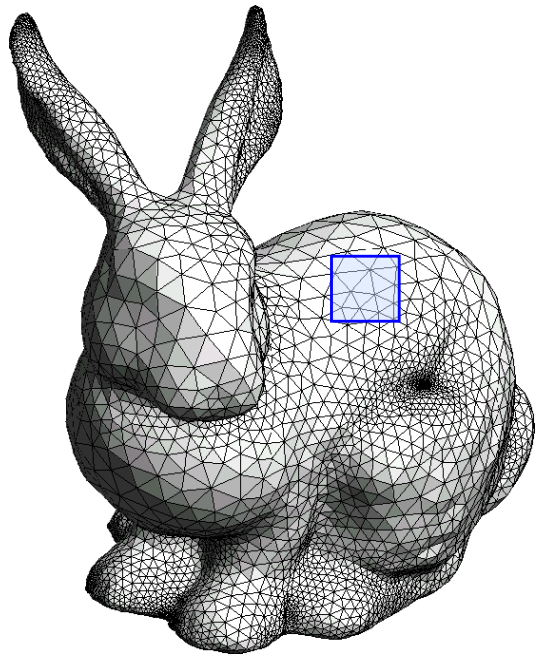
Given $d_g(x, y)$, find $f(x), f(y) \in \mathbb{R}^2$ such that :

$$d_g(x, y) \approx \|f(x) - f(y)\|_2 \text{ for all } x, y \in \mathbb{S}^2$$

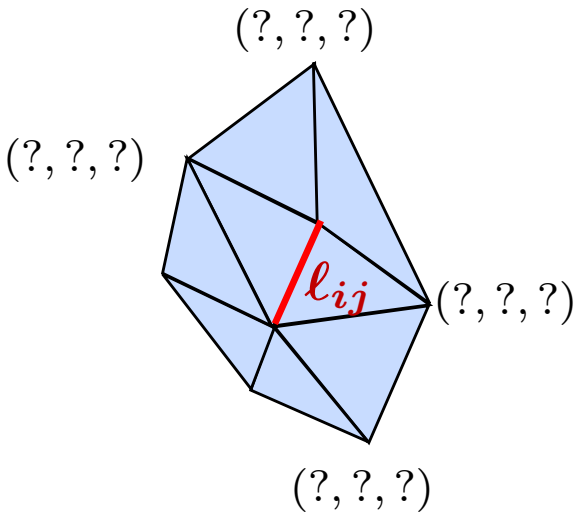
Embeddings: Q2 example



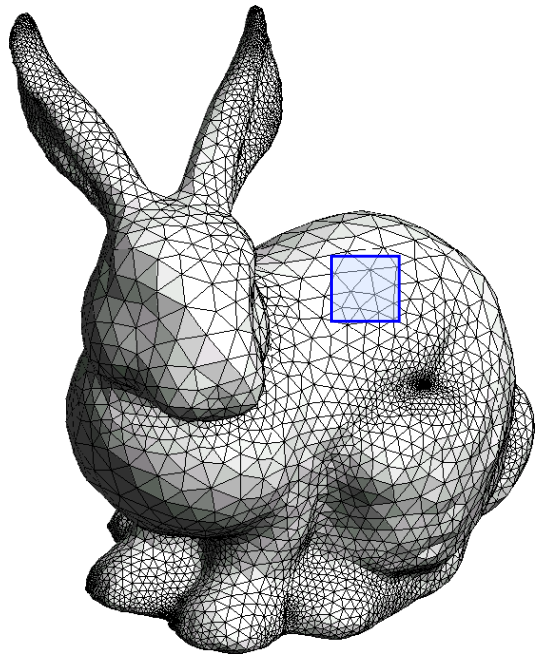
$\overset{?}{\Longrightarrow}$



Embeddings: Q2 example

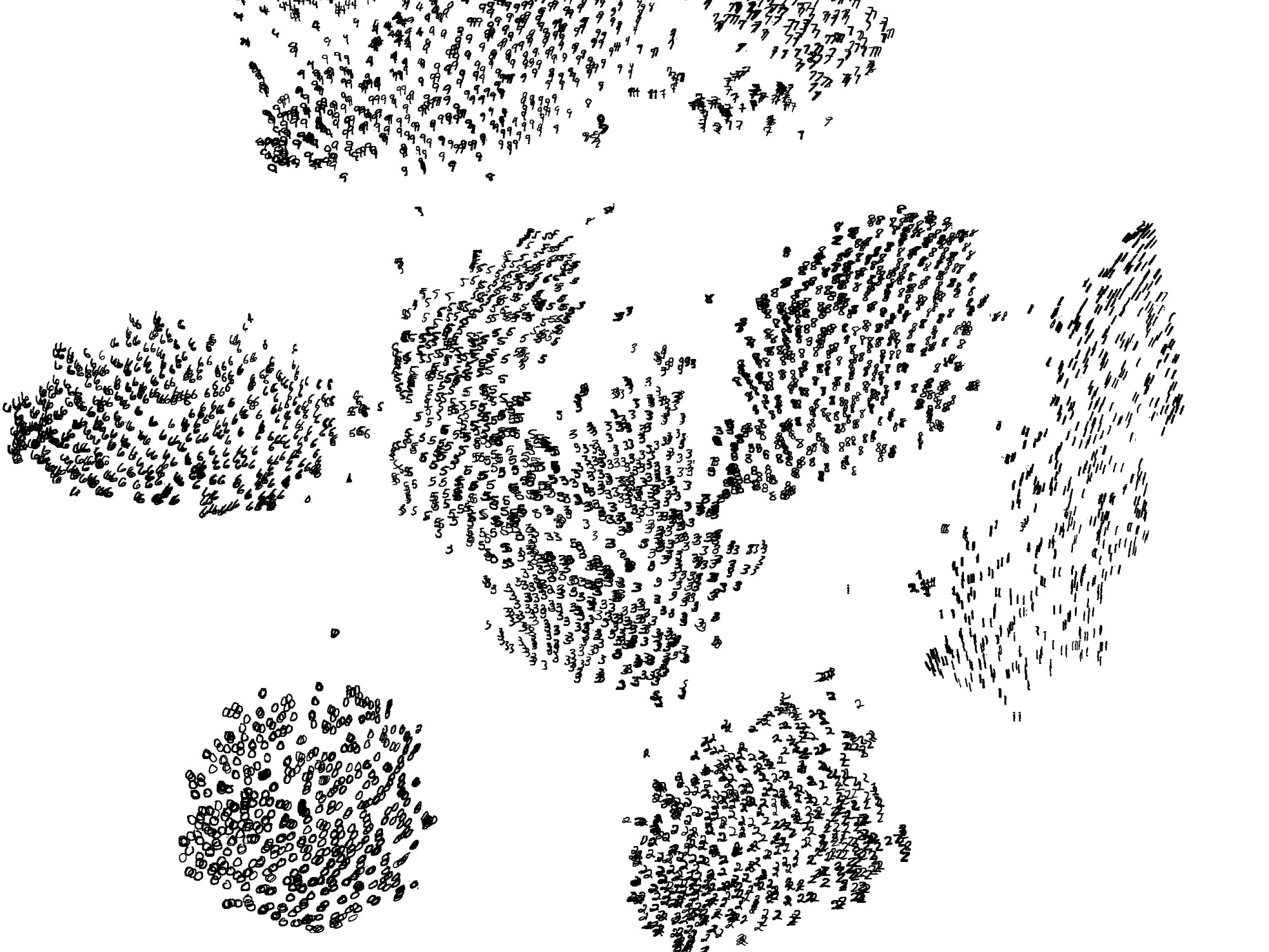


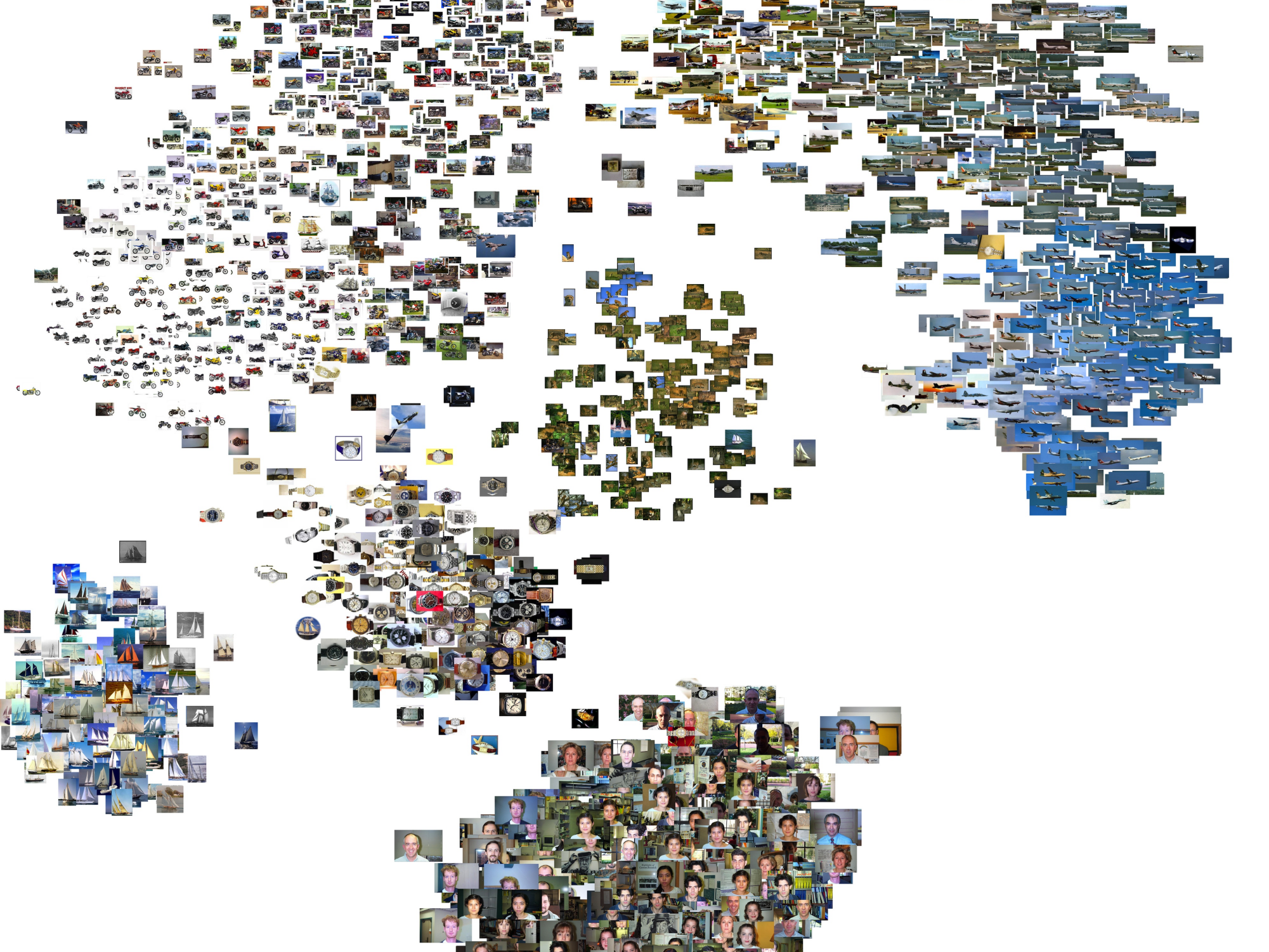
$\overset{?}{\Longrightarrow}$



Given ℓ_{ij}, find $x_i, x_j \in \mathbb{R}^3$ such that:

$$\ell_{ij} \approx \|x_i - x_j\|_2 \text{ for all } i, j$$


i
ii



Whenever we deal with data that we suspect having a structure, and we want to visualize it

Metric distortion

The distortion induced by f on the metric can be quantified as:

- The relative change between metrics ("dilation"):

$$\frac{d_{\mathcal{Y}}(f(x_1), f(x_2))}{d_{\mathcal{X}}(x_1, x_2)} \approx 1 \quad \text{for all } x_1, x_2 \in \mathcal{X}$$

Metric distortion

The distortion induced by f on the metric can be quantified as:

- The relative change between metrics ("dilation"):

$$\frac{d_{\mathcal{Y}}(f(x_1), f(x_2))}{d_{\mathcal{X}}(x_1, x_2)} \approx 1 \quad \text{for all } x_1, x_2 \in \mathcal{X}$$

- The absolute distortion of the metric:

$$|d_{\mathcal{Y}}(f(x_1), f(x_2)) - d_{\mathcal{X}}(x_1, x_2)| \approx 0 \quad \text{for all } x_1, x_2 \in \mathcal{X}$$

Metric distortion

The distortion induced by f on the metric can be quantified as:

- The relative change between metrics ("dilation"):

$$\frac{d_{\mathcal{Y}}(f(x_1), f(x_2))}{d_{\mathcal{X}}(x_1, x_2)} \approx 1 \quad \text{for all } x_1, x_2 \in \mathcal{X}$$

- The absolute distortion of the metric:

$$|d_{\mathcal{Y}}(f(x_1), f(x_2)) - d_{\mathcal{X}}(x_1, x_2)| \approx 0 \quad \text{for all } x_1, x_2 \in \mathcal{X}$$

A minimum-distortion embedding is the f minimizing absolute or relative distortion for all $x_1, x_2 \in \mathcal{X}$

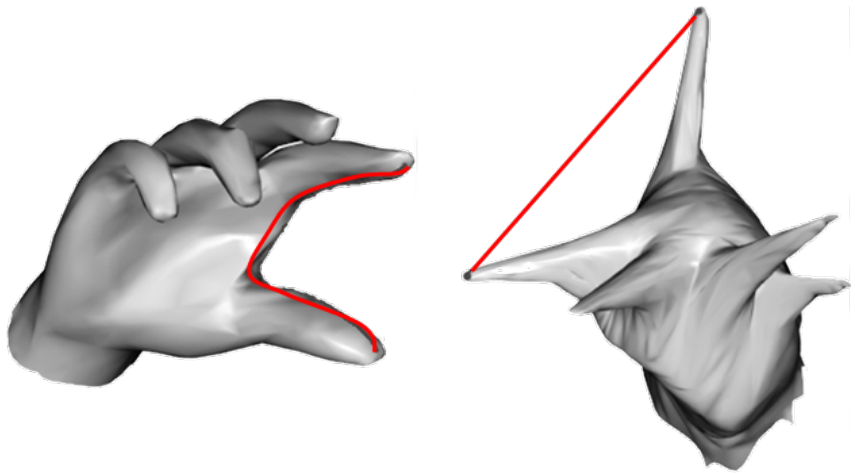
Canonical forms

If $d_{\mathcal{Y}} = \|\cdot\|_2$ we call the embedded shape (the image under f) the canonical form of $\mathcal{X}$.

Canonical forms

If $d_{\mathcal{Y}} = \|\cdot\|_2$ we call the embedded shape (the image under f) the canonical form of $\mathcal{X}$.

Example:

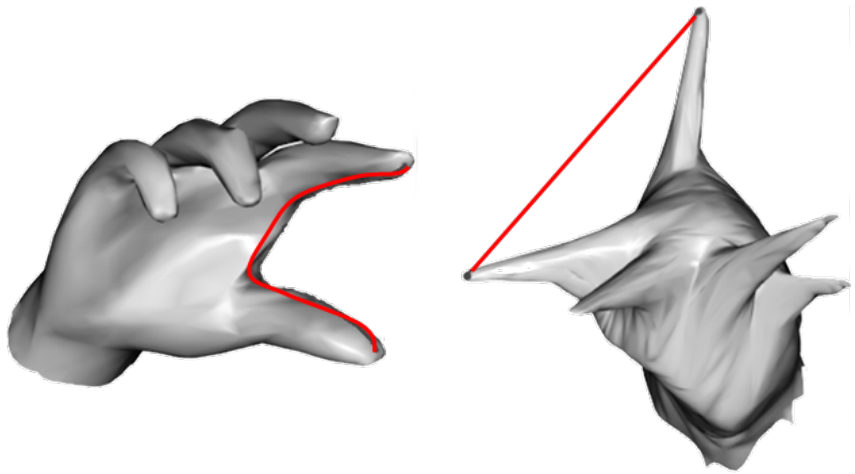


$(\mathcal{X}, d_{\mathcal{X}})$ $(f(\mathcal{X}), \|\cdot\|_2)$

Canonical forms

If $d_{\mathcal{Y}} = \|\cdot\|_2$ we call the embedded shape (the image under f) the canonical form of $\mathcal{X}$.

Example:

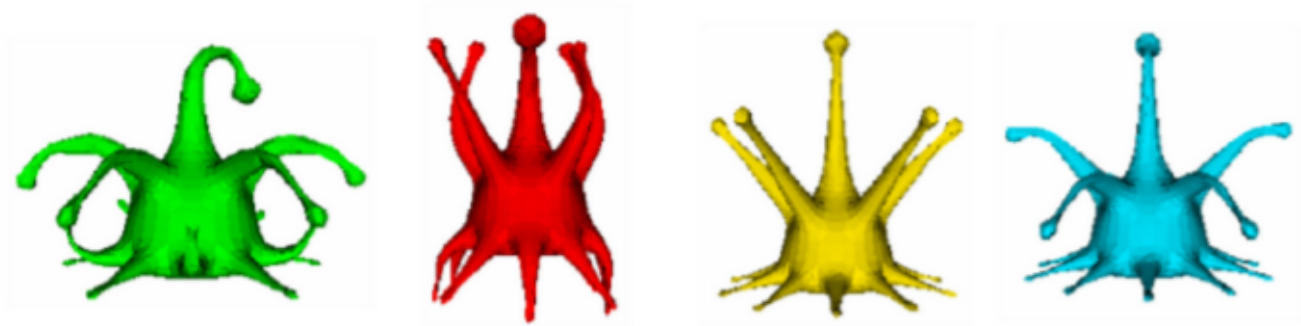


$(\mathcal{X}, d_{\mathcal{X}})$ $(f(\mathcal{X}), \|\cdot\|_2)$

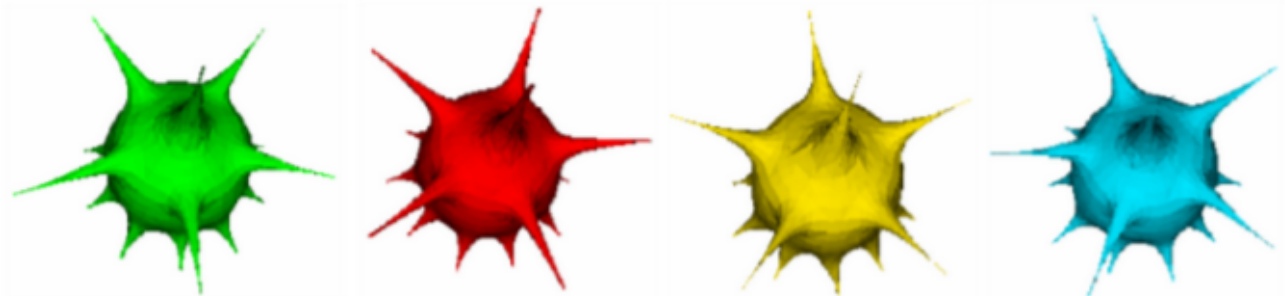
The canonical form $f(\mathcal{X})$ is an approximation, since zero-distortion is not always achievable (Q: when is it achievable?)

Canonical forms

A canonical form defines an isometry class (equivalence class up to an isometry) in $\mathbb{R}^k$.



near-isometric deformations of a shape



canonical forms

We are reducing intrinsic isometries into extrinsic isometries.

Multi-Dimensional Scaling (MDS)

Stress minimization

As a global measure of distortion, consider the quadratic stress

$$f = \arg \min_{f:\mathcal{X}\to\mathbb{R}^k} \sum_{i>j} |d_{\mathcal{X}}(x_i, x_j) - \|f(x_i) - f(x_j)\|_2|^2$$

Stress minimization

As a global measure of distortion, consider the quadratic stress

$$f = \arg \min_{f:\mathcal{X}\to\mathbb{R}^k} \sum_{i>j} |d_{\mathcal{X}}(x_i, x_j) - \|f(x_i) - f(x_j)\|_2|^2$$

Define $\mathbf{z}_i = f(x_i)$ and arrange these vectors into a $n \times k$ matrix $\mathbf{Z}$.

We get the equivalent problem:

$$\mathbf{Z}^* = \arg \min_{\mathbf{Z}\in\mathbb{R}^{n\times k}} \sum_{i>j} |d_{\mathcal{X}}(x_i, x_j) - d_{ij}(\mathbf{Z})|^2$$

where $d_{ij}(\mathbf{Z}) = \|\mathbf{z}_i - \mathbf{z}_j\|_2$

Stress minimization

As a global measure of distortion, consider the quadratic stress

$$f = \arg \min_{f:\mathcal{X}\to\mathbb{R}^k} \sum_{i>j} |d_{\mathcal{X}}(x_i, x_j) - \|f(x_i) - f(x_j)\|_2|^2$$

Define $\mathbf{z}_i = f(x_i)$ and arrange these vectors into a $n \times k$ matrix $\mathbf{Z}$.

We get the equivalent problem:

$$\mathbf{Z}^* = \arg \min_{\mathbf{Z}\in\mathbb{R}^{n\times k}} \sum_{i>j} |d_{\mathcal{X}}(x_i, x_j) - d_{ij}(\mathbf{Z})|^2$$

where $d_{ij}(\mathbf{Z}) = \|\mathbf{z}_i - \mathbf{z}_j\|_2$

Given $\mathbf{Z}$, the stress measures how well that configuration matches the data. We look for the configuration of minimum stress.

Stress minimization

As a global measure of distortion, consider the quadratic stress

$$f = \arg \min_{f:\mathcal{X}\to\mathbb{R}^k} \sum_{i>j} |d_{\mathcal{X}}(x_i, x_j) - \|f(x_i) - f(x_j)\|_2|^2$$

Define $\mathbf{z}_i = f(x_i)$ and arrange these vectors into a $n \times k$ matrix $\mathbf{Z}$.

We get the equivalent problem:

$$\mathbf{Z}^* = \arg \min_{\mathbf{Z}\in\mathbb{R}^{n\times k}} \sum_{i>j} |d_{\mathcal{X}}(x_i, x_j) - d_{ij}(\mathbf{Z})|^2$$

where $d_{ij}(\mathbf{Z}) = \|\mathbf{z}_i - \mathbf{z}_j\|_2$

Given $\mathbf{Z}$, the stress measures how well that configuration matches the data. We look for the configuration of minimum stress.

No unique solution!

Any Euclidean isometry of the optimizer $\mathbf{Z}^*$ does not change the stress.

Multidimensional scaling

Why "stress"?

Empirical procedures of several diverse kinds have this in common: they start with a fixed set of entities and determine, for every pair of these, a number reflecting how closely the two entities are related psychologically. The nature of the psychological relation depends upon the nature of the entities. If the entities are all stimuli or all responses, we are inclined to think of the relation as one of similarity. A somewhat more objective (though less intuitive) characterization of such a relation, perhaps, is that of substitutability. The statement that stimulus A is more similar to B than to C, for example, could be interpreted to say that the psychological (or behavioral) consequences are greater when C, rather than B, is substituted for A. From this standpoint a natural procedure for determining similarities of stimuli or responses is by recording substitution errors during identification learning [2, 7, 12, 14, 17, 18]. In addition, though, disjunctive reaction time and sorting time have also been proposed as measures of psychological similarity [20]. Finally, of course, individuals have sometimes been instructed simply to rate each pair of stimuli, directly, on a scale of apparent similarity [1, 6]. The notion of similarity is not necessarily restricted to stimuli or responses (in the narrow sense of these words), however. Serviceable measures of similarity may also be found for concepts, attitudes, personality structures, or even social institutions, political systems, and the like.

Shepard, "The analysis of proximities: Multidimensional scaling with an unknown distance function". Psychometrika 27(2), 1962

Quadratic stress in matrix form

$$\sigma(\mathbf{Z}) = \sum_{i>j} |d_{\mathcal{X}}(x_i, x_j) - d_{ij}(\mathbf{Z})|^2$$

Quadratic stress in matrix form

$$\sigma(\mathbf{Z}) = \sum_{i>j} |d_{\mathcal{X}}(x_i, x_j) - d_{ij}(\mathbf{Z})|^2$$

This can be rewritten as:

$$\begin{aligned}\sigma(\mathbf{Z}) &= \sum_{i>j} |d_{\mathcal{X}}(x_i, x_j) - d_{ij}(\mathbf{Z})|^2 \\ &= \sum_{i>j} \underbrace{d_{ij}^2(\mathbf{Z})}_{\text{Term 1}} - \underbrace{2d_{ij}(\mathbf{Z})d_{\mathcal{X}}(x_i, x_j)}_{\text{Term 2}} + d_{\mathcal{X}}^2(x_i, x_j)\end{aligned}$$

Quadratic stress: Term 1

$$\sum_{i>j} d_{ij}^2(\mathbf{Z}) = \sum_{i>j} \|\mathbf{z}_i - \mathbf{z}_j\|_2^2$$

Quadratic stress: Term 1

$$\sum_{i>j} d_{ij}^2(\mathbf{Z}) = \sum_{i>j} \|\mathbf{z}_i - \mathbf{z}_j\|_2^2$$

$$= \sum_{i>j} \sum_{c=1}^{k} (\mathbf{z}_i^c - \mathbf{z}_j^c)^2$$

Quadratic stress: Term 1

$$\begin{aligned}\sum_{i>j} d_{ij}^2(\mathbf{Z}) &= \sum_{i>j} \|\mathbf{z}_i - \mathbf{z}_j\|_2^2 \\ &= \sum_{i>j} \sum_{c=1}^{k} (\mathbf{z}_i^c - \mathbf{z}_j^c)^2 \\ &= \sum_{i>j} \sum_{c=1}^{k} (\mathbf{z}_i^c)^2 - 2\mathbf{z}_i^c \mathbf{z}_j^c + (\mathbf{z}_j^c)^2\end{aligned}$$

Quadratic stress: Term 1

$$\begin{aligned}
\sum_{i>j} d_{ij}^2(\mathbf{Z}) &= \sum_{i>j} \|\mathbf{z}_i - \mathbf{z}_j\|_2^2 \\
&= \sum_{i>j} \sum_{c=1}^{k} (\mathbf{z}_i^c - \mathbf{z}_j^c)^2 \\
&= \sum_{i>j} \sum_{c=1}^{k} (\mathbf{z}_i^c)^2 - 2\mathbf{z}_i^c \mathbf{z}_j^c + (\mathbf{z}_j^c)^2 \\
&= \sum_{i>j} \langle \mathbf{z}_j, \mathbf{z}_j \rangle + \langle \mathbf{z}_i, \mathbf{z}_i \rangle - 2\langle \mathbf{z}_i, \mathbf{z}_j \rangle
\end{aligned}$$

Quadratic stress: Term 1

$$\begin{aligned}
\sum_{i>j} d_{ij}^2(\mathbf{Z}) &= \sum_{i>j} \|\mathbf{z}_i - \mathbf{z}_j\|_2^2 \\
&= \sum_{i>j} \sum_{c=1}^{k} (\mathbf{z}_i^c - \mathbf{z}_j^c)^2 \\
&= \sum_{i>j} \sum_{c=1}^{k} (\mathbf{z}_i^c)^2 - 2\mathbf{z}_i^c \mathbf{z}_j^c + (\mathbf{z}_j^c)^2 \\
&= \sum_{i>j} \langle \mathbf{z}_j, \mathbf{z}_j \rangle + \langle \mathbf{z}_i, \mathbf{z}_i \rangle - 2\langle \mathbf{z}_i, \mathbf{z}_j \rangle \\
&= \sum_{i>j} \langle \mathbf{z}_j, \mathbf{z}_j \rangle + \langle \mathbf{z}_i, \mathbf{z}_i \rangle - 2 \sum_{i>j} \langle \mathbf{z}_i, \mathbf{z}_j \rangle
\end{aligned}$$

Quadratic stress: Term 1

$$
\begin{aligned}
\sum_{i>j} d_{ij}^2(\mathbf{Z}) &= \sum_{i>j} \|\mathbf{z}_i - \mathbf{z}_j\|_2^2 \\
&= \sum_{i>j} \sum_{c=1}^{k} (\mathbf{z}_i^c - \mathbf{z}_j^c)^2 \\
&= \sum_{i>j} \sum_{c=1}^{k} (\mathbf{z}_i^c)^2 - 2\mathbf{z}_i^c \mathbf{z}_j^c + (\mathbf{z}_j^c)^2 \\
&= \sum_{i>j} \langle \mathbf{z}_j, \mathbf{z}_j \rangle + \langle \mathbf{z}_i, \mathbf{z}_i \rangle - 2\langle \mathbf{z}_i, \mathbf{z}_j \rangle \\
&= \sum_{i>j} \langle \mathbf{z}_j, \mathbf{z}_j \rangle + \langle \mathbf{z}_i, \mathbf{z}_i \rangle - 2 \sum_{i>j} \langle \mathbf{z}_i, \mathbf{z}_j \rangle \\
&= (n-1) \sum_{i=1}^{n} \langle \mathbf{z}_i, \mathbf{z}_i \rangle - \left(\sum_{i,j} \langle \mathbf{z}_i, \mathbf{z}_j \rangle - \sum_{i=1}^{n} \langle \mathbf{z}_i, \mathbf{z}_i \rangle \right)
\end{aligned}
$$

Quadratic stress: Term 1

$$
\begin{aligned}
\sum_{i>j} d_{ij}^2(\mathbf{Z}) &= \sum_{i>j} \|\mathbf{z}_i - \mathbf{z}_j\|_2^2 \\
&= \sum_{i>j} \sum_{c=1}^{k} (\mathbf{z}_i^c - \mathbf{z}_j^c)^2 \\
&= \sum_{i>j} \sum_{c=1}^{k} (\mathbf{z}_i^c)^2 - 2\mathbf{z}_i^c \mathbf{z}_j^c + (\mathbf{z}_j^c)^2 \\
&= \sum_{i>j} \langle \mathbf{z}_j, \mathbf{z}_j \rangle + \langle \mathbf{z}_i, \mathbf{z}_i \rangle - 2\langle \mathbf{z}_i, \mathbf{z}_j \rangle \\
&= \sum_{i>j} \langle \mathbf{z}_j, \mathbf{z}_j \rangle + \langle \mathbf{z}_i, \mathbf{z}_i \rangle - 2 \sum_{i>j} \langle \mathbf{z}_i, \mathbf{z}_j \rangle \\
&= (n-1) \sum_{i=1}^{n} \langle \mathbf{z}_i, \mathbf{z}_i \rangle - \left(\sum_{i,j} \langle \mathbf{z}_i, \mathbf{z}_j \rangle - \sum_{i=1}^{n} \langle \mathbf{z}_i, \mathbf{z}_i \rangle \right) \\
&= n \sum_{i=1}^{n} \langle \mathbf{z}_i, \mathbf{z}_i \rangle - \sum_{i,j} \langle \mathbf{z}_i, \mathbf{z}_j \rangle
\end{aligned}
$$

Quadratic stress: Term 1

$$\sum_{i>j} d_{ij}^2(\mathbf{Z}) = n \sum_{i=1}^{n} \langle \mathbf{z}_i, \mathbf{z}_i \rangle - \sum_{i,j} \langle \mathbf{z}_i, \mathbf{z}_j \rangle$$

Quadratic stress: Term 1

$$\sum_{i>j} d_{ij}^2(\mathbf{Z}) = n \sum_{i=1}^{n} \langle \mathbf{z}_i, \mathbf{z}_i \rangle - \sum_{i,j} \langle \mathbf{z}_i, \mathbf{z}_j \rangle$$

$$\overset{\text{(a)}}{=} n\mathrm{tr}(\mathbf{Z}\mathbf{Z}^\top) - \mathrm{tr}(\mathbf{1}\mathbf{1}^\top \mathbf{Z}\mathbf{Z}^\top)$$

(a) See teaser exercise from a few slides back

Quadratic stress: Term 1

$$\sum_{i>j} d_{ij}^2(\mathbf{Z}) = n \sum_{i=1}^{n} \langle \mathbf{z}_i, \mathbf{z}_i \rangle - \sum_{i,j} \langle \mathbf{z}_i, \mathbf{z}_j \rangle$$

$$\overset{\text{(a)}}{=} n\text{tr}(\mathbf{Z}\mathbf{Z}^\top) - \text{tr}(\mathbf{1}\mathbf{1}^\top \mathbf{Z}\mathbf{Z}^\top)$$

$$\overset{\text{(b)}}{=} \text{tr}(\mathbf{V}\mathbf{Z}\mathbf{Z}^\top)$$

where $\mathbf{V} = n\mathbf{I} - \mathbf{1}\mathbf{1}^\top$.

(a) See teaser exercise from a few slides back

(b) By linearity of the trace

Quadratic stress: Term 1

$$\begin{aligned}\sum_{i>j} d_{ij}^2(\mathbf{Z}) &= n\sum_{i=1}^{n}\langle \mathbf{z}_i, \mathbf{z}_i\rangle - \sum_{i,j}\langle \mathbf{z}_i, \mathbf{z}_j\rangle \\ &\overset{(a)}{=} n\mathrm{tr}(\mathbf{Z}\mathbf{Z}^\top) - \mathrm{tr}(\mathbf{1}\mathbf{1}^\top\mathbf{Z}\mathbf{Z}^\top) \\ &\overset{(b)}{=} \mathrm{tr}(\mathbf{V}\mathbf{Z}\mathbf{Z}^\top) \\ &\overset{(c)}{=} \mathrm{tr}(\mathbf{Z}^\top\mathbf{V}\mathbf{Z})\end{aligned}$$

where $\mathbf{V} = n\mathbf{I} - \mathbf{1}\mathbf{1}^\top$.

(a) See teaser exercise from a few slides back

(b) By linearity of the trace

(c) Because $\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA})$

Quadratic stress: Term 2

$$\begin{aligned}\sigma(\mathbf{Z}) &= \sum_{i>j} |d_{\mathcal{X}}(x_i, x_j) - d_{ij}(\mathbf{Z})|^2 \\ &= \sum_{i>j} \underbrace{d_{ij}^2(\mathbf{Z})}_{\mathrm{tr}(\mathbf{Z}^\top \mathbf{V}\mathbf{Z})} \underbrace{- 2 d_{ij}(\mathbf{Z}) d_{\mathcal{X}}(x_i, x_j)}_{\text{Term 2}} + d_{\mathcal{X}}^2(x_i, x_j)\end{aligned}$$

Quadratic stress: Term 2

$$\begin{aligned}\sigma(\mathbf{Z}) &= \sum_{i>j} |d_{\mathcal{X}}(x_i, x_j) - d_{ij}(\mathbf{Z})|^2 \\ &= \sum_{i>j} \underbrace{d_{ij}^2(\mathbf{Z})}_{\mathrm{tr}(\mathbf{Z}^\top \mathbf{V} \mathbf{Z})} \underbrace{- 2 d_{ij}(\mathbf{Z}) d_{\mathcal{X}}(x_i, x_j)}_{\text{Term 2}} + d_{\mathcal{X}}^2(x_i, x_j)\end{aligned}$$

$$\sum_{i>j} d_{ij}(\mathbf{Z}) d_{\mathcal{X}}(x_i, x_j) = \sum_{i>j} d_{\mathcal{X}}(x_i, x_j) d_{ij}^{-1}(\mathbf{Z}) d_{ij}^2(\mathbf{Z})$$

Quadratic stress: Term 2

$$\begin{aligned}\sigma(\mathbf{Z}) &= \sum_{i>j} |d_{\mathcal{X}}(x_i, x_j) - d_{ij}(\mathbf{Z})|^2 \\ &= \sum_{i>j} \underbrace{d_{ij}^2(\mathbf{Z})}_{\mathrm{tr}(\mathbf{Z}^\top \mathbf{V}\mathbf{Z})} \underbrace{- 2 d_{ij}(\mathbf{Z}) d_{\mathcal{X}}(x_i, x_j)}_{\text{Term 2}} + d_{\mathcal{X}}^2(x_i, x_j)\end{aligned}$$

$$\begin{aligned}\sum_{i>j} d_{ij}(\mathbf{Z}) d_{\mathcal{X}}(x_i, x_j) &= \sum_{i>j} d_{\mathcal{X}}(x_i, x_j) d_{ij}^{-1}(\mathbf{Z}) d_{ij}^2(\mathbf{Z}) \\ &= \sum_{i>j} d_{\mathcal{X}}(x_i, x_j) d_{ij}^{-1}(\mathbf{Z}) \sum_{d=1}^{k} (\mathbf{z}_i^d - \mathbf{z}_j^d)^2\end{aligned}$$

Quadratic stress: Term 2

$$\begin{aligned}\sigma(\mathbf{Z}) &= \sum_{i>j} |d_{\mathcal{X}}(x_i, x_j) - d_{ij}(\mathbf{Z})|^2 \\ &= \sum_{i>j} \underbrace{d_{ij}^2(\mathbf{Z})}_{\mathrm{tr}(\mathbf{Z}^\top \mathbf{V}\mathbf{Z})} \underbrace{- 2 d_{ij}(\mathbf{Z}) d_{\mathcal{X}}(x_i, x_j)}_{\text{Term 2}} + d_{\mathcal{X}}^2(x_i, x_j)\end{aligned}$$

$$\begin{aligned}\sum_{i>j} d_{ij}(\mathbf{Z}) d_{\mathcal{X}}(x_i, x_j) &= \sum_{i>j} d_{\mathcal{X}}(x_i, x_j) d_{ij}^{-1}(\mathbf{Z}) d_{ij}^2(\mathbf{Z}) \\ &= \sum_{i>j} d_{\mathcal{X}}(x_i, x_j) d_{ij}^{-1}(\mathbf{Z}) \sum_{d=1}^{k} (\mathbf{z}_i^d - \mathbf{z}_j^d)^2 \\ &= \sum_{i>j} \underbrace{d_{\mathcal{X}}(x_i, x_j) d_{ij}^{-1}(\mathbf{Z})}_{a_{ij} = a_{ji}} (\langle \mathbf{z}_i, \mathbf{z}_i \rangle + \langle \mathbf{z}_j, \mathbf{z}_j \rangle - 2\langle \mathbf{z}_i, \mathbf{z}_j \rangle)\end{aligned}$$

Quadratic stress: Term 2

$$\sum_{i>j} d_{ij}(\mathbf{Z}) d_{\mathcal{X}}(x_i, x_j) = \sum_{i>j} a_{ij}(\langle \mathbf{z}_i, \mathbf{z}_i \rangle + \langle \mathbf{z}_j, \mathbf{z}_j \rangle - 2\langle \mathbf{z}_i, \mathbf{z}_j \rangle)$$

Quadratic stress: Term 2

$$\sum_{i>j} d_{ij}(\mathbf{Z}) d_{\mathcal{X}}(x_i, x_j) = \sum_{i>j} a_{ij}(\langle \mathbf{z}_i, \mathbf{z}_i \rangle + \langle \mathbf{z}_j, \mathbf{z}_j \rangle - 2\langle \mathbf{z}_i, \mathbf{z}_j \rangle)$$
$$= \sum_{i,j} a_{ij}(\langle \mathbf{z}_i, \mathbf{z}_i \rangle - \langle \mathbf{z}_i, \mathbf{z}_j \rangle)$$

Quadratic stress: Term 2

$$\begin{aligned}\sum_{i>j} d_{ij}(\mathbf{Z}) d_{\mathcal{X}}(x_i, x_j) &= \sum_{i>j} a_{ij}(\langle \mathbf{z}_i, \mathbf{z}_i \rangle + \langle \mathbf{z}_j, \mathbf{z}_j \rangle - 2\langle \mathbf{z}_i, \mathbf{z}_j \rangle) \\ &= \sum_{i,j} a_{ij}(\langle \mathbf{z}_i, \mathbf{z}_i \rangle - \langle \mathbf{z}_i, \mathbf{z}_j \rangle) \\ &= \operatorname{tr}(\mathbf{B}\mathbf{Z}\mathbf{Z}^\top) \\ &= \operatorname{tr}(\mathbf{Z}^\top \mathbf{B}\mathbf{Z})\end{aligned}$$

where $b_{ij} = \begin{cases} -a_{ij} & \text{if } i \neq j \\ -\sum_{\ell \neq i} b_{i\ell} & \text{if } i = j \end{cases}$

Quadratic stress: Term 2

$$\begin{aligned}\sum_{i>j} d_{ij}(\mathbf{Z}) d_{\mathcal{X}}(x_i, x_j) &= \sum_{i>j} a_{ij}(\langle \mathbf{z}_i, \mathbf{z}_i \rangle + \langle \mathbf{z}_j, \mathbf{z}_j \rangle - 2\langle \mathbf{z}_i, \mathbf{z}_j \rangle) \\ &= \sum_{i,j} a_{ij}(\langle \mathbf{z}_i, \mathbf{z}_i \rangle - \langle \mathbf{z}_i, \mathbf{z}_j \rangle) \\ &= \operatorname{tr}(\mathbf{B}\mathbf{Z}\mathbf{Z}^\top) \\ &= \operatorname{tr}(\mathbf{Z}^\top \mathbf{B}\mathbf{Z})\end{aligned}$$

where $b_{ij} = \begin{cases} -d_{\mathcal{X}}(x_i, x_j) d_{ij}^{-1}(\mathbf{Z}) & \text{if } i \neq j \\ -\sum_{\ell \neq i} b_{i\ell} & \text{if } i = j \end{cases}$

Quadratic stress: Term 2

$$\begin{aligned}\sum_{i>j} d_{ij}(\mathbf{Z}) d_{\mathcal{X}}(x_i, x_j) &= \sum_{i>j} a_{ij}(\langle \mathbf{z}_i, \mathbf{z}_i \rangle + \langle \mathbf{z}_j, \mathbf{z}_j \rangle - 2\langle \mathbf{z}_i, \mathbf{z}_j \rangle) \\ &= \sum_{i,j} a_{ij}(\langle \mathbf{z}_i, \mathbf{z}_i \rangle - \langle \mathbf{z}_i, \mathbf{z}_j \rangle) \\ &= \operatorname{tr}(\mathbf{B}\mathbf{Z}\mathbf{Z}^\top) \\ &= \operatorname{tr}(\mathbf{Z}^\top \mathbf{B}\mathbf{Z})\end{aligned}$$

where $\mathbf{B} = -\mathbf{D}_{\mathcal{X}} \oslash \mathbf{D}_{\mathbf{Z}} + \operatorname{diag}((\mathbf{D}_{\mathcal{X}} \oslash \mathbf{D}_{\mathbf{Z}})\mathbf{1})$ and $\oslash$ denotes element-wise division

Note that $\mathbf{B}$ directly depends on the unknown $\mathbf{Z}$, so we will write $\mathbf{B}(\mathbf{Z})$

Quadratic stress in matrix form

$$\sigma(\mathbf{Z}) = \sum_{i>j} |d_{\mathcal{X}}(x_i, x_j) - d_{ij}(\mathbf{Z})|^2$$

$$\Downarrow$$

$$\sigma(\mathbf{Z}) = \mathrm{tr}(\mathbf{Z}^\top \mathbf{V}\mathbf{Z}) - 2\mathrm{tr}(\mathbf{Z}^\top \mathbf{B}(\mathbf{Z})\mathbf{Z}) + \sum_{i>j} d^2_{\mathcal{X}}(x_i, x_j)$$

where

$$\mathbf{V} = n\mathbf{I} - \mathbf{1}\mathbf{1}^\top$$
$$\mathbf{B}(\mathbf{Z}) = -\mathbf{D}_{\mathcal{X}} \oslash \mathbf{D}_{\mathbf{Z}} + \mathrm{diag}((\mathbf{D}_{\mathcal{X}} \oslash \mathbf{D}_{\mathbf{Z}})\mathbf{1})$$

Quadratic stress in matrix form

$$\sigma(\mathbf{Z}) = \sum_{i>j} |d_{\mathcal{X}}(x_i, x_j) - d_{ij}(\mathbf{Z})|^2$$

$$\Downarrow$$

$$\sigma(\mathbf{Z}) = \mathrm{tr}(\mathbf{Z}^\top \mathbf{V}\mathbf{Z}) - 2\mathrm{tr}(\mathbf{Z}^\top \mathbf{B}(\mathbf{Z})\mathbf{Z}) + \sum_{i>j} d^2_{\mathcal{X}}(x_i, x_j)$$

where

$$\mathbf{V} = n\mathbf{I} - \mathbf{1}\mathbf{1}^\top$$
$$\mathbf{B}(\mathbf{Z}) = -\mathbf{D}_{\mathcal{X}} \oslash \mathbf{D}_{\mathbf{Z}} + \mathrm{diag}((\mathbf{D}_{\mathcal{X}} \oslash \mathbf{D}_{\mathbf{Z}})\mathbf{1})$$

Our task is to solve the minimization problem:

$$\min_{\mathbf{Z} \in \mathbb{R}^{n \times k}} \sigma(\mathbf{Z})$$

We will use gradient descent.

Gradient descent for stress minimization

$$\min_{\mathbf{Z} \in \mathbb{R}^{n \times k}} \sigma(\mathbf{Z})$$

Gradient descent for stress minimization

$$\min_{\mathbf{Z} \in \mathbb{R}^{n \times k}} \sigma(\mathbf{Z})$$

Our gradient can be computed as:

$$\begin{aligned}
\nabla \sigma(\mathbf{Z}) &= \nabla \left(\operatorname{tr}(\mathbf{Z}^\top \mathbf{V} \mathbf{Z}) - 2\operatorname{tr}(\mathbf{Z}^\top \mathbf{B}(\mathbf{Z}) \mathbf{Z}) + \sum_{i>j} d^2_{\mathcal{X}}(x_i, x_j) \right) \\
&= 2\mathbf{V}\mathbf{Z} - 2\mathbf{B}(\mathbf{Z})\mathbf{Z}
\end{aligned}$$

Gradient descent for stress minimization

$$\min_{\mathbf{Z} \in \mathbb{R}^{n \times k}} \sigma(\mathbf{Z})$$

Our gradient can be computed as:

$$\begin{aligned}\nabla \sigma(\mathbf{Z}) &= \nabla \left(\operatorname{tr}(\mathbf{Z}^\top \mathbf{V} \mathbf{Z}) - 2\operatorname{tr}(\mathbf{Z}^\top \mathbf{B}(\mathbf{Z}) \mathbf{Z}) + \sum_{i>j} d^2_{\mathcal{X}}(x_i, x_j) \right) \\ &= 2\mathbf{V}\mathbf{Z} - 2\mathbf{B}(\mathbf{Z})\mathbf{Z}\end{aligned}$$

Then we can follow the algorithm:

- Start with a random configuration of points $\mathbf{Z}^{(0)}$

Gradient descent for stress minimization

$$\min_{\mathbf{Z} \in \mathbb{R}^{n \times k}} \sigma(\mathbf{Z})$$

Our gradient can be computed as:

$$\begin{aligned}\nabla \sigma(\mathbf{Z}) &= \nabla \left(\operatorname{tr}(\mathbf{Z}^\top \mathbf{V} \mathbf{Z}) - 2\operatorname{tr}(\mathbf{Z}^\top \mathbf{B}(\mathbf{Z}) \mathbf{Z}) + \sum_{i>j} d^2_{\mathcal{X}}(x_i, x_j) \right) \\ &= 2\mathbf{V}\mathbf{Z} - 2\mathbf{B}(\mathbf{Z})\mathbf{Z}\end{aligned}$$

Then we can follow the algorithm:

- Start with a random configuration of points $\mathbf{Z}^{(0)}$
- Iterate $\mathbf{Z}^{(t+1)} = \mathbf{Z}^{(t)} - \alpha \nabla \sigma(\mathbf{Z}^{(t)})$

Gradient descent for stress minimization

$$\min_{\mathbf{Z} \in \mathbb{R}^{n \times k}} \sigma(\mathbf{Z})$$

Our gradient can be computed as:

$$\begin{aligned}\nabla \sigma(\mathbf{Z}) &= \nabla \left(\operatorname{tr}(\mathbf{Z}^\top \mathbf{V} \mathbf{Z}) - 2\operatorname{tr}(\mathbf{Z}^\top \mathbf{B}(\mathbf{Z}) \mathbf{Z}) + \sum_{i>j} d^2_{\mathcal{X}}(x_i, x_j) \right) \\ &= 2\mathbf{V}\mathbf{Z} - 2\mathbf{B}(\mathbf{Z})\mathbf{Z}\end{aligned}$$

Then we can follow the algorithm:

- Start with a random configuration of points $\mathbf{Z}^{(0)}$
- Iterate $\mathbf{Z}^{(t+1)} = \mathbf{Z}^{(t)} - \alpha \nabla \sigma(\mathbf{Z}^{(t)})$
- Terminate when $|\sigma(\mathbf{Z}^{(t+1)}) - \sigma(\mathbf{Z}^{(t)})| < \epsilon$

Stochastic Neighbor Embedding

Similarities

Instead of computing global Euclidean distances:

$$d_{ij} = \|x_i - x_j\|_2 \in \mathbb{R}$$

Consider local similarities:

$$p_{ij} = e^{-\frac{\|x_i - x_j\|^2}{2\sigma_i^2}} \in (0, 1)$$

Similarities

Instead of computing global Euclidean distances:

$$d_{ij} = \|x_i - x_j\|_2$$

Consider local similarities:

$$p_{ij} = e^{-\frac{\|x_i - x_j\|^2}{2\sigma_i^2}}, \quad p_{ij} = \frac{p_{ij}}{\sum_{k \neq i} p_{ik}}$$

Similarities

Instead of computing global Euclidean distances:

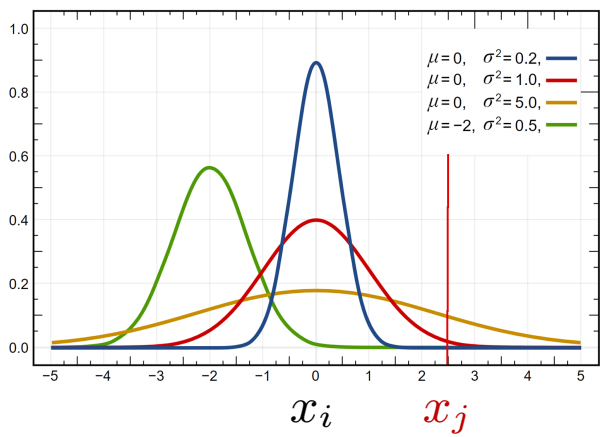
$$d_{ij} = \|x_i - x_j\|_2$$

Consider local similarities:

$$p_{ij} = e^{-\frac{\|x_i - x_j\|^2}{2\sigma_i^2}}, \quad p_{ij} = \frac{p_{ij}}{\sum_{k \neq i} p_{ik}}$$

Conditional probability that x_j came from a Gaussian centered at x_i.

"If x_i chooses a random neighbor, this is the probability that x_j is picked in proportion to the Gaussian density centered at x_i"



Stochastic Neighbor Embedding (SNE)

Instead of using Euclidean distortion:

$$\min_{z_i \in \mathbb{R}^k} \sum_{i,j} \Big(\underbrace{\|x_i - x_j\|_2}_{\substack{\text{Euclidean in} \\ \text{high dimensions}}} - \underbrace{\|z_i - z_j\|_2}_{\substack{\text{Euclidean in} \\ \text{low dimensions}}} \Big)^2$$

Stochastic Neighbor Embedding (SNE)

Instead of using Euclidean distortion:

$$\min_{z_i \in \mathbb{R}^k} \sum_{i,j} \Big(\underbrace{\|x_i - x_j\|_2}_{\substack{\text{Euclidean in} \\ \text{high dimensions}}} - \underbrace{\|z_i - z_j\|_2}_{\substack{\text{Euclidean in} \\ \text{low dimensions}}} \Big)^2$$

Use:

$$\min_{z_i \in \mathbb{R}^k} \sum_{i,j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

with

$$q_{ij} = e^{-\|z_i - z_j\|^2}, \quad q_{ij} = \frac{q_{ij}}{\sum_{k \neq i} q_{ik}}$$

Hinton & Roweis, "Stochastic neighbor embedding", 2002

Stochastic Neighbor Embedding (SNE)

Instead of using Euclidean distortion:

$$\min_{z_i \in \mathbb{R}^k} \sum_{i,j} \Big(\underbrace{\|x_i - x_j\|_2}_{\substack{\text{Euclidean in} \\ \text{high dimensions}}} - \underbrace{\|z_i - z_j\|_2}_{\substack{\text{Euclidean in} \\ \text{low dimensions}}} \Big)^2$$

Use:

$$\min_{z_i \in \mathbb{R}^k} \sum_{i,j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

with

$$q_{ij} = e^{-\|z_i - z_j\|^2}, \quad q_{ij} = \frac{q_{ij}}{\sum_{k \neq i} q_{ik}}$$

Intuitively, we want the distribution of the low-dimensional z_i to "look similar" to the distribution of the high-dimensional x_i.

Hinton & Roweis, "Stochastic neighbor embedding", 2002

KL divergence

Construct matrices $\mathbf{P} = (p_{ij})$ and $\mathbf{Q} = (q_{ij})$ with $p_{ii} = q_{ii} = 0$.

We get:

$$D_{\mathsf{KL}}(\mathbf{P}||\mathbf{Q}) = \sum_{i,j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

called the Kullback-Leibler divergence between distributions p and q.

KL divergence

Construct matrices $\mathbf{P} = (p_{ij})$ and $\mathbf{Q} = (q_{ij})$ with $p_{ii} = q_{ii} = 0$.

We get:

$$D_{\mathsf{KL}}(\mathbf{P}||\mathbf{Q}) = \sum_{i,j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

called the Kullback-Leibler divergence between distributions p and q.

KL divergence is asymmetric.

KL divergence

Construct matrices $\mathbf{P} = (p_{ij})$ and $\mathbf{Q} = (q_{ij})$ with $p_{ii} = q_{ii} = 0$.

We get:

$$D_{\mathrm{KL}}(\mathbf{P}||\mathbf{Q}) = \sum_{i,j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

called the Kullback-Leibler divergence between distributions p and q.

KL divergence is asymmetric. This emphasizes local structure!

- If p_{ij} is large and q_{ij} is small $\rightarrow$ large penalty

KL divergence

Construct matrices $\mathbf{P} = (p_{ij})$ and $\mathbf{Q} = (q_{ij})$ with $p_{ii} = q_{ii} = 0$.

We get:

$$D_{\mathsf{KL}}(\mathbf{P}||\mathbf{Q}) = \sum_{i,j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

called the Kullback-Leibler divergence between distributions p and q.

KL divergence is asymmetric. This emphasizes local structure!

- If p_{ij} is large and q_{ij} is small $\rightarrow$ large penalty
- If p_{ij} is small and q_{ij} is large $\rightarrow$ small penalty

  ...because $p_{ij} \log \frac{p_{ij}}{q_{ij}} \approx 0$ when $p_{ij} \approx 0$, no matter the value of q_{ij}

KL divergence

Construct matrices $\mathbf{P} = (p_{ij})$ and $\mathbf{Q} = (q_{ij})$ with $p_{ii} = q_{ii} = 0$.

We get:

$$D_{\mathrm{KL}}(\mathbf{P}||\mathbf{Q}) = \sum_{i,j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

called the Kullback-Leibler divergence between distributions p and q.

KL divergence is asymmetric. This emphasizes local structure!

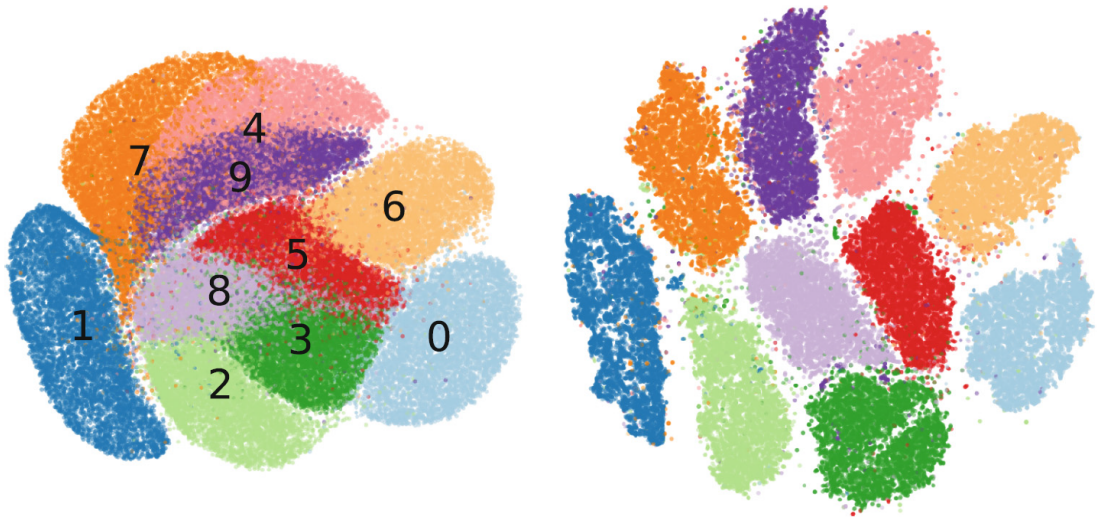
- If p_{ij} is large and q_{ij} is small $\rightarrow$ large penalty
- If p_{ij} is small and q_{ij} is large $\rightarrow$ small penalty

  ...because $p_{ij} \log \frac{p_{ij}}{q_{ij}} \approx 0$ when $p_{ij} \approx 0$, no matter the value of q_{ij}

As a result: local structures are emphasized forming clusters, but distant points tend to crowd together.

Crowding problem

The embeddings are compressed near the center (MNIST example):

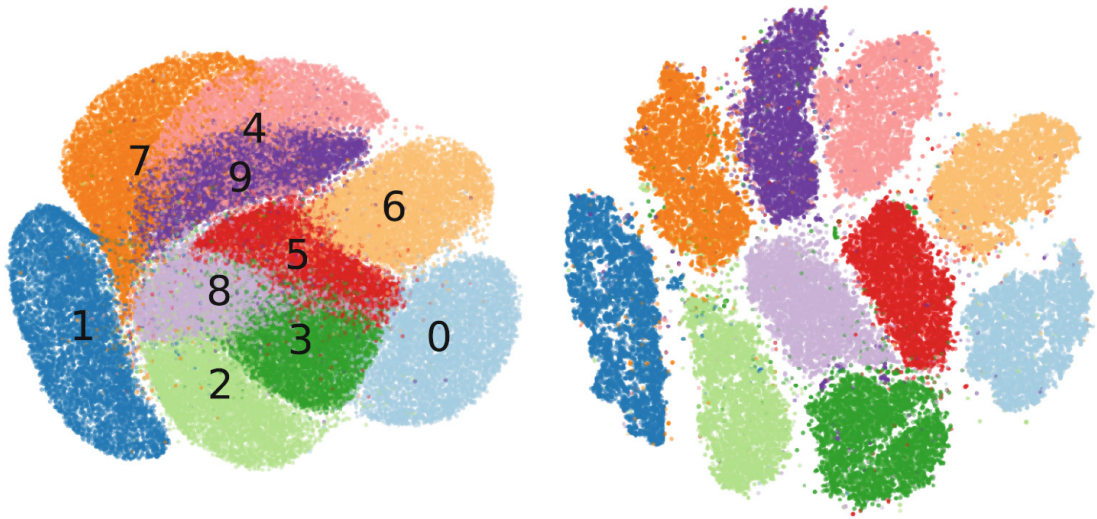


SNE

t-SNE
(next slides)

Crowding problem

The embeddings are compressed near the center (MNIST example):



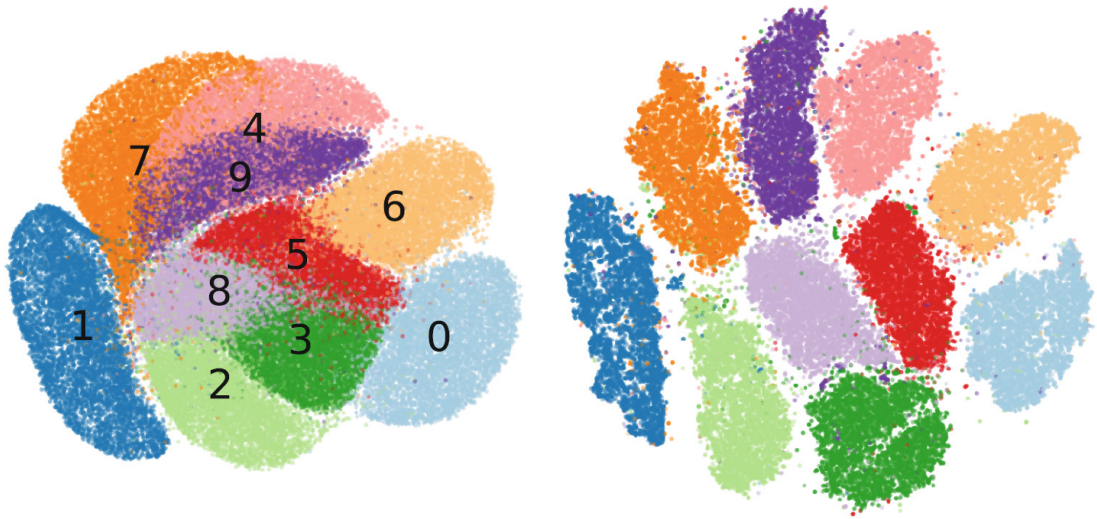
SNE

t-SNE
(next slides)

Clusters are very tight $\rightarrow$ difficult to distinguish any structure!

Crowding problem

The embeddings are compressed near the center (MNIST example):



SNE

t-SNE
(next slides)

Clusters are very tight $\rightarrow$ difficult to distinguish any structure!

Note: You choose the target dimension.

Perplexity

Recall the definition:

$$p_{ij} = e^{-\frac{\|x_i - x_j\|^2}{\sigma_i^2}}, \quad p_{ij} = \frac{p_{ij}}{\sum_{k \neq i} p_{ik}}$$

How to choose the σ_i?

Perplexity

Recall the definition:

$$p_{ij} = e^{-\frac{\|x_i - x_j\|^2}{\sigma_i^2}}, \quad p_{ij} = \frac{p_{ij}}{\sum_{k \neq i} p_{ik}}$$

How to choose the σ_i? The user chooses it indirectly.

Define perplexity: the number of effective neighbors.

Perplexity

Recall the definition:

$$p_{ij} = e^{-\frac{\|x_i - x_j\|^2}{\sigma_i^2}}, \quad p_{ij} = \frac{p_{ij}}{\sum_{k \neq i} p_{ik}}$$

How to choose the σ_i? The user chooses it indirectly.

Define perplexity: the number of effective neighbors.

Set σ_i such that the perplexity of x_i matches the user-given value for all $i = 1, \ldots, n$.

Perplexity

Recall the definition:

$$p_{ij} = e^{-\frac{\|x_i - x_j\|^2}{\sigma_i^2}}, \quad p_{ij} = \frac{p_{ij}}{\sum_{k \neq i} p_{ik}}$$

How to choose the σ_i? The user chooses it indirectly.

Define perplexity: the number of effective neighbors.

Set σ_i such that the perplexity of x_i matches the user-given value for all $i = 1, \ldots, n$.

- Small value: local structures, potentially missing broader patterns.

Perplexity

Recall the definition:

$$p_{ij} = e^{-\frac{\|x_i - x_j\|^2}{\sigma_i^2}}, \quad p_{ij} = \frac{p_{ij}}{\sum_{k \neq i} p_{ik}}$$

How to choose the σ_i? The user chooses it indirectly.

Define perplexity: the number of effective neighbors.

Set σ_i such that the perplexity of x_i matches the user-given value for all $i = 1, \ldots, n$.

- Small value: local structures, potentially missing broader patterns.
- Large value: emphasizes global structure but misses local detail.

t-SNE

Recipe to avoid crowding:

- Keep a Gaussian in the high-dimensional data space:

$$p_{ij} = e^{-\frac{\|x_i - x_j\|^2}{\sigma_i^2}}, \quad p_{ij} = \frac{p_{ij}}{\sum_{k \neq i} p_{ik}}$$

van der Maaten & Hinton, "Visualizing Data using t-SNE", 2008

t-SNE

Recipe to avoid crowding:

- Keep a Gaussian in the high-dimensional data space:

$$p_{ij} = e^{-\frac{\|x_i - x_j\|^2}{\sigma_i^2}}, \quad p_{ij} = \frac{p_{ij}}{\sum_{k \neq i} p_{ik}}$$

- Use a Cauchy distribution in the low-dimensional target space:

$$q_{ij} = (1 + \|z_i - z_j\|^2)^{-1}, \quad q_{ij} = \frac{q_{ij}}{\sum_{k \neq i} q_{ik}}$$

van der Maaten & Hinton, "Visualizing Data using t-SNE", 2008

t-SNE

Recipe to avoid crowding:

- Keep a Gaussian in the high-dimensional data space:

$$p_{ij} = e^{-\frac{\|x_i - x_j\|^2}{\sigma_i^2}}, \quad p_{ij} = \frac{p_{ij}}{\sum_{k \neq i} p_{ik}}$$

- Use a Cauchy distribution in the low-dimensional target space:

$$q_{ij} = (1 + \|z_i - z_j\|^2)^{-1}, \quad q_{ij} = \frac{q_{ij}}{\sum_{k \neq i} q_{ik}}$$

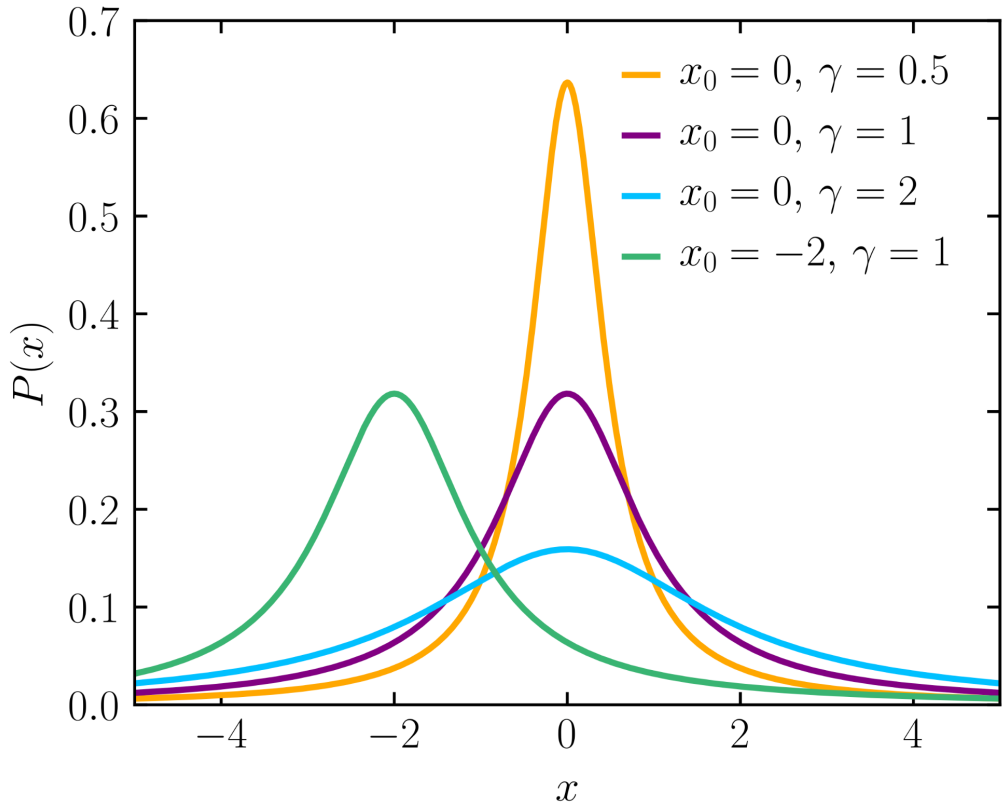
Cauchy is heavy-tailed (goes slowly to zero).

To match the Gaussian's fast decay, you must place distant pairs even farther apart in the embedding space.

van der Maaten & Hinton, "Visualizing Data using t-SNE", 2008

Cauchy distribution

Suppose $x := \|z_i - z_j\|$.



To minimize $D_{\mathrm{KL}}(p||q)$, where p is Gaussian, the Cauchy distribution q allows distant points, because it decreases more slowly with distance.

Optimization

We can optimize using gradient descent, possibly with momentum.

We want to minimize:

$$D_{\mathrm{KL}}(\mathbf{P}||\mathbf{Q}) = \sum_{i,j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

SNE and t-SNE differ in their choice of q.

Optimization

We can optimize using gradient descent, possibly with momentum.

We want to minimize:

$$D_{\mathrm{KL}}(\mathbf{P}||\mathbf{Q}) = \sum_{i,j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

SNE and t-SNE differ in their choice of q.

- Gradient for SNE (Gaussian p, Gaussian q):

$$\frac{\partial D_{\mathrm{KL}}}{\partial \mathbf{z}_i} = 2 \sum_j (p_{ij} - q_{ij} + p_{ji} - q_{ji})(\mathbf{z}_i - \mathbf{z}_j)$$

- Gradient for t-SNE (Gaussian p, Cauchy q):

$$\frac{\partial D_{\mathrm{KL}}}{\partial \mathbf{z}_i} = 4 \sum_j (p_{ij} - q_{ij})(\mathbf{z}_i - \mathbf{z}_j)(1 + \|\mathbf{z}_i - \mathbf{z}_j\|^2)^{-1}$$

Algorithm

Input:

- High-dimensional data $X = \{x_1, \ldots, x_n\}$
- Desired target dimension k
- Desired perplexity $Perp$
- Learning rate η
- Momentum $\alpha(t)$

Output:

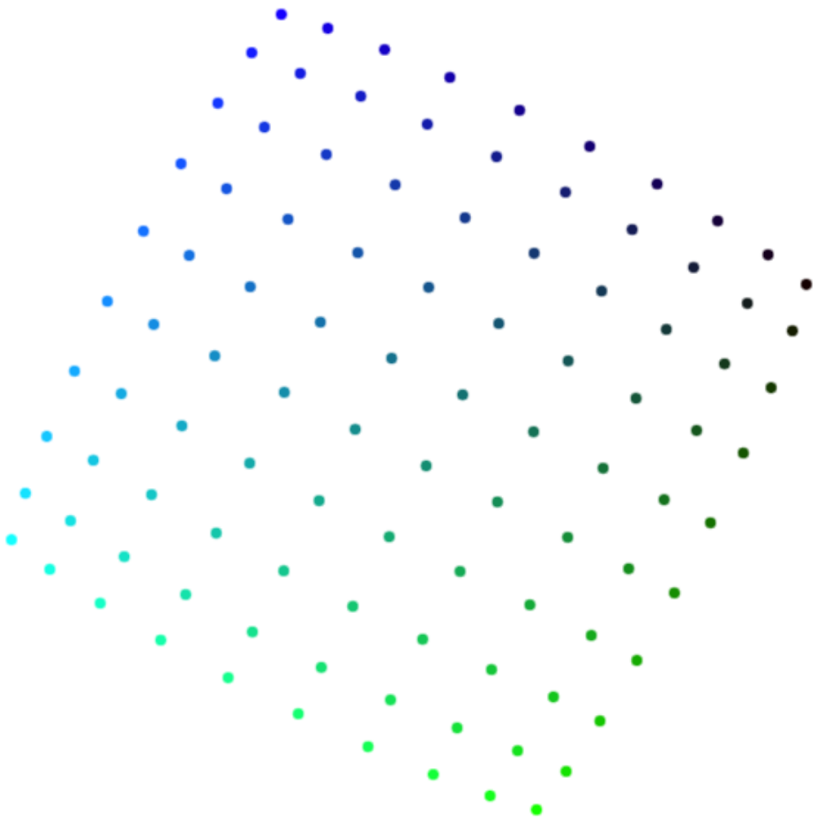
- Low-dimensional embeddings $Z = \{z_1, \ldots, z_n\}$ in $\mathbb{R}^k$

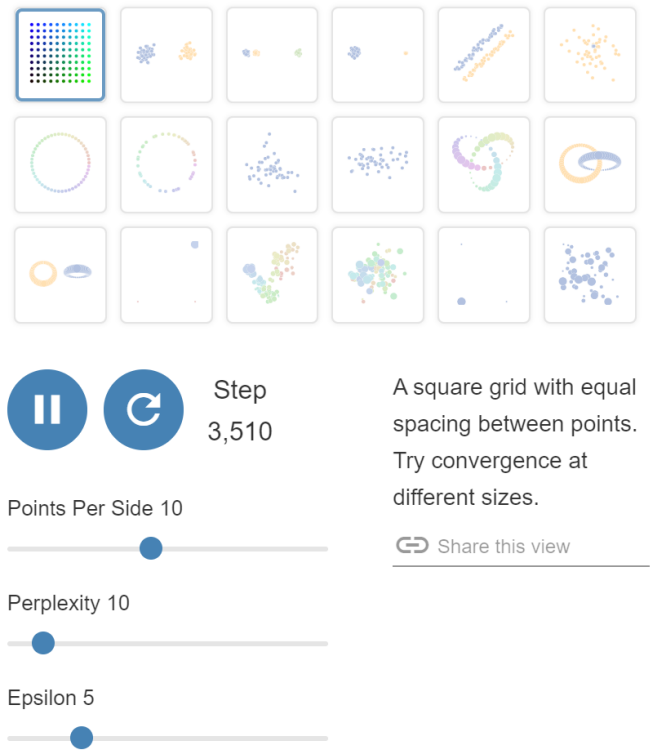
Algorithm

1. Find σ_i to get the desired $Perp$
2. Compute $\mathbf{P}$ using the data X
3. Initialize $\mathbf{Z}^{(t=0)} \sim N(0, 10^{-4})$
4. Compute $\mathbf{Q}$ using $\mathbf{Z}^{(t)}$
5. Compute the gradient $\nabla_{\mathbf{Z}} D_{\mathsf{KL}}(\mathbf{P}||\mathbf{Q})$
6. Compute $\mathbf{Z}^{(t)} = \underbrace{\mathbf{Z}^{(t-1)} - \eta \nabla_{\mathbf{Z}} D_{\mathsf{KL}}(\mathbf{P}||\mathbf{Q})}_{\text{gradient descent}} + \underbrace{\alpha(t)(\mathbf{Z}^{(t-1)} - \mathbf{Z}^{(t-2)})}_{\text{momentum}}$
7. Go back to step 4.
8. Stop when criterion is met.

How to Use t-SNE Effectively

Although extremely useful for visualizing high-dimensional data, t-SNE plots can sometimes be mysterious or misleading. By exploring how it behaves in simple cases, we can learn to use it more effectively.





"How to Use t-SNE Effectively": https://distill.pub/2016/misread-tsne/