The 2-hour written exam for *CS4102 Geometric Foundations of Data Analysis I* will consist of four questions: three from Graham and one from Emil. Students will be required to attempt all questions.

The 2-hour written exam for *CS4103 Geometric Foundations in Data Analysis II* will consist of four questions: all four from Graham. Students will be required to attempt all questions.

Students registered for *MA500 Geometric Foundations of Data Analysis* will take both the CS4102 paper and the CS4103 paper.

The following are examples of the kinds of things that Graham and Emil could ask.

CS4102

1 Least Squares Fitting

1. Find the best least squares straight line fit to the following measurements, and sketch your solution:

   $y = 2$ at $t = -1$,

   $y = 0$ at $t = 0$,

   $y = -3$ at $t = 1$,

   $y = -5$ at $t = 2$.

2. A middle-aged man was stretched on a rack to lengths $L = 5$, 6, and 7 feet under applied forces of $F = 1$, 2 and 4 tones. Assuming Hooke's Law $L = a + bF$, find his normal length a by least squares.

3. Let
$$y = b_0 + b_1 x_1 + \cdots + b_{p-1} x_{p-1} \tag{1}$$
denote the hyperplane in $\mathbb{R}^p$ that is the best least square hyperplane fit to a given collection of data points $(y_k, x_{k,1}, \ldots, x_{k,p-1}) \in \mathbb{R}^p$, $1 \leq k \leq n$. Derive the norm equations, in matrix notation, for determining this hyperplane. (That is:

   Either

   (a) Describe, in terms of partial derivatives, the normal equations that determine the constants $b_0, \ldots, b_n$.
   (b) Then use matrix notation to express these normal equations, making sure to define those matrices involved.

   Or, more easily and more preferably,

   Observe that on letting $||\ ||$ denote the Euclidean norm and writing
$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix},\ A = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,p-1} \\ & \vdots & & \\ 1 & x_{n,1} & \cdots & x_{n,p-1} \end{pmatrix},\ b = \begin{pmatrix} b_0 \\ \vdots \\ b_p \end{pmatrix},$$
   our least squares requirement is simply that b should be chosen so that $||Ab - y||$ is as small as possible. In other words, the vector $Ab - y$ should be perpendicular to the plane spanned by the vectors of the form Au with $u \in \mathbb{R}^p$. In other words, for an arbitrary vector $u \in \mathbb{R}^p$ we need $0 = u^t A^t (Ab - y)$ or, since u is arbitrary,
$$0 = A^t (Ab - y) \tag{2}$$
   Convince yourself that (2) is the (correct way to think of the) system of normal equations.)

4. Let $y = b_0 + b_1 x$ denote the best least squares straight line fit to given data points $(y_1, x_1), ..., (y_n, x_n)$.

   (a) Define the *fitted values* $\hat{y}_i$, *residuals* e_i, *sample mean* $\overline{y}$, *total sum of squares* $SSTO$, *error sum of squares* SSE, *regression sum of squares* SSR, and *coefficient of determination* R^2.

   (b) Prove that $\sum_{i=1}^{n} e_i = 0$.

   (c) Prove that $\sum_{i=1}^{n} \hat{y}_i e_i = 0$.

   (d) Prove that $SSTO = SSE + SSR$.

   (e) Prove that $1 \leq R^2 \leq 1$.

The theory of statistical inference can be applied to the output from a least squares fit. This topic is outside the main focus of this module. Nevertheless, the following two questions touch on the topic. These kinds of questions won't appear on the exam.

1. The theory of statistical inference can be applied to the output from a least squares fitting.

   Suppose given a random variable

   $$y_i = \beta_0 + \beta_1 x_{i,1} + \cdots \beta_{p-1} x_{i,p-1} + \epsilon_i$$

   where

   - $i = 1, 2, \ldots, n$;
   - $x_{i,1}, \ldots, x_{i,p-1}$ are known constants;
   - $\beta_0, \ldots, \beta p - 1$ are unkown fixed parameters;
   - ϵ_i are independent random variables with common normal distribution $N(0, \sigma^2)$.

   Describe a criterion for choosing between the two hypotheses

   $\mathcal{C}_1 : \quad \beta_1 = \beta_2 = \cdots \beta_{p-1} = 0$

   $\mathcal{C}_2 : \quad \beta_i \neq 0$ for at least one $1 \leq i \leq p - 1$

   that controls Type I errors at level α.

2. In the context of the previous question, suppose that $q \leq p$ of the parameters β_k need to be estimated jointly. Describe the Bonferroni confidence intervals with family coefficient $1 - \alpha$ for these q parametrers.

2 Principal Component Analysis

1. Explain what is meant by *Principal Component Analysis*. Your explanation should include explanations of the terms: *geometric information*; *covariance matrix*; *orthogonal transformation*; *Spectral Theorem* and describe how the technique can be used to reduce dimensionality while retaining much geometric information.

2. Prove that any real symmetric matrix has at least one real eigenvector.

3. Use the fact that any real symmetric $n \times n$ matrix A has at least one eigenvector to prove that it has n linearly independent real eigenvectors.

4. Determine the maximum value of the function $f(x, y) = x^2 + 4xy + 4y^2$ on the unit sphere $\mathbb{S}^1 = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}$. Also, find a linear homomorphism $\phi : \mathbb{R}^2 \to \mathbb{R}^2, (x, y) \mapsto (x', y')$ such that $f\phi(x, y) = \lambda_1 x'^2 + \lambda_2 y'^2$ for some $\lambda_1, \lambda_2 \in \mathbb{R}$.

5. A data set $S \subset \mathbb{R}^p$ consists of vectors whose first component is the number of kilometers that a salesperson has travelled during the last month. A principal component analysis is performed on S, the set S is then projected onto the three principal components with largest eigenvalues, and the projected points are visualized in $\mathbb{R}^3$. Would this visualization be any different if distance had been measured in miles? Justify your answer.

6. Suppose given a finite set S of data points in $\mathbb{R}^3$ and that a visual inspection suggests that all points look to lie close to some 2-dimensional plane containing the origin. We could construct a plane by regarding the first coordinate as a dependent variable and taking a least squares fit. Alternatively, we could construct a plane using Principal Component Analysis and taking the span of the eigenvectors corresponding to the two larger eignevalues. In general, would the two constructed planes differ? If so, in what way?

3 Clustering and Persistence

1. Describe an algorithm that inputs an $n \times n$ distance (or dissimilarity) matrix for n items, applies single-linkage hierarchical clustering, and returns the corrsponding barcode. Determine a worst-case time estimate for the algorithm.

2. Describe the Smith-Waterman algorithm for determining the optimal score of a local alignment of two sequences of letters. Include an explanation of the terms *local alignment* and *optimal score*.

3. Perform the Smith-Waterman algorithm to find an optimal local alignment for the sequences

$$X = TGCATA\,, \qquad Y = ATCTGAT$$

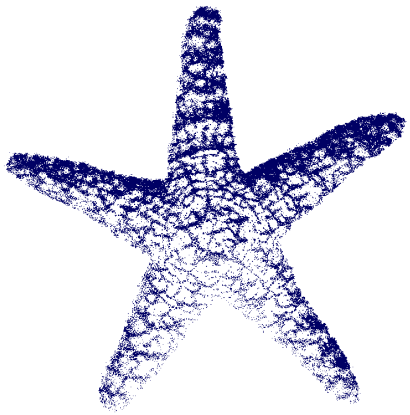
using column scores

$$\delta\begin{pmatrix} x \\ x \end{pmatrix} = 1,\ \delta\begin{pmatrix} x \\ y \end{pmatrix} = -1 \text{ if } x \neq y,\ \delta\begin{pmatrix} x \\ - \end{pmatrix} = -1.$$

4. Explain how cluster analysis and barcodes can be used to estimate the number of objects in the digital photograph.

Explain how cluster analysis can also be used to estimate the number of objects with holes.

5. Explain how cluster analysis and barcodes can be used to estimate the number of 'limbs' of an object such as a starfish from a digital image of the object.



6. A property $I(S)$ of a set $S \subset \mathbb{R}^n$ is said to be a *geometric invariant* if its value does not change when S undergoes an orthogonal tranaformation. Explain why the barcodes in the preceding questions (3) and (4) are geometric invariants.

4 Nearest Neighbours

1. Describe a real-life situation where a k-th nearest neighbour algorithm could be used to make decisions.
2. Describe a method, based on Voronoi regions, for finding the point in a finite subset S of low-dimensional Euclidean space $\mathbb{R}^d$ that is closest to some given point $v \in \mathbb{R}^n$. Your description should define the terms *Voronoi region*, *facet*, *neighbour* and describe an algorithm which assumes Voronoi regions have been computed for all points in S.
3. State the Johnson-Lindenstrauss Theorem.
4. State the Norm Preservation Proposition.
5. Use the Norm Preservation Proposition to prove the Johnson-Lindenstrauss Theorem. (This wasn't covered this year due to lack of time.)

5 Python Part I

1. Suppose that you have a text file with the following format:

```
p[0] = (12.3, 4.5)
p[1] = (-1.6, 7.9)
p[2] = (11.0, 9.8)
...
```

   (a) Write Python code that reads the file and creates two lists: `xs` and `ys` with the x-values and y-values of the points. For the example file above the initial parts of the lists would be `xs = [12.3, -1.6, 11.0,...]` and `ys = [4.5, 7.9, 9.8, ...]`

   (b) Write Python code that creates a plot of the points using matplotlib.

2. (a) Give a concise explanation of the following concepts

i. *Classes* and *objects* in Python.

ii. *Abstract classes* and how they can be simulated in Python.

iii. Some of Python's special methods, namely `__init__`, `__str__` and `__call__`.

(b) In a Python program dealing with vectors, you might need to calculate the distance between two vectors using different norms. Therefore you decide to implement classes `EuclideanDist` and `InfinityDist` for calculating the distance using the Euclidean norm, and the infinity norm, respectively. Show how to do this, by completing the code below:

```
class Distance:
    def __init__(self):
        # your code here
    def __call__(self, p, q):
        # your code here
    def __str__(self):
        # your code here

class EuclideanDistance(Distance):
    def __init__(self):
        # your code here
    def __call__(self, p, q):
        # your code here
    def __str__(self):
        # your code here

class InfinityDistance(Distance):
    def __init__(self):
        # your code here
    def __call__(self, p, q):
        # your code here
    def __str__(self):
        # your code here
```

Your implementation should allow for the following usage in the program.

```
euclideanDist = EuclideanDistance()
d = euclideanDist(vec1, vec2)
print("The distance between vec1 and vec2 is",
      d, "using the", euclideanDist, "distance")
```

assuming that `vec1` and `vec2` are vectors on the appropriate form.

3. Suppose that you are given two text files named `a.txt` and `b.txt` respectively, both containing n lines with n entries each. For example, one of the files could have the following contents:

```
12  4  9 -1
 1  6  0  0
 3  4  5  7
21 15 -4 18
```

Write Python code that

(a) Reads the contents of the files into two numpy arrays called `a` and `b`.

(b) Compute a new numpy array named `c`.

(c) Write `c` to the file `c.txt` using the same file format as for `a.txt` and `b.txt`

Use *exceptions* to handle any IO errors. Your program *should not crash* even if the reading and writing to files does not succeed.

CS4103

6 Simplicial complexes, clique complexes, and Mapper clustering

1. Define what is meant by a *geometric k-simplex*, a *simplicial complex* K, and the *geometric realization* $|K|$ of a finite simplicial complex K. Illustrate your answers with a **simple** example.

2. Which of the following K are simplicial complexes? For those that are, sketch the simplicial complex, and determine the Euler characteristic

$$\chi(K) = \alpha_0 - \alpha_1 + \alpha_2 - \alpha_3 + \cdots$$

where α_k denotes the number of k-simplexes.

   (a) $V = \{1,2,3,4,5,6,7,8\}, K = \{\{1\},\{2\},\{3\},\{4\},\{5\},\{6\},\{7\},\{8\},\{1,2\},\{1,6\},\{2,3\},\{2,4\},\{2,5\},\{3,4\},\{3,5\},\{4,5\},\{4,6\},\{5,6\},\{2,3,4\},\{2,3,5\},\{2,4,5\},\{3,4,5\}\}$.

   (b) $V = \{1,2,3,4,5,6,7,8\}, K = \{\{1\},\{2\},\{3\},\{4\},\{5\},\{6\},\{7\},\{8\},\{1,2\},\{1,6\},\{2,3\},\{2,4\},\{1,5\},\{3,4\},\{3,5\},\{4,5\},\{4,6\},\{5,6\},\{2,3,4\},\{2,3,5\},\{2,4,5\},\{3,4,5\}\}$.

3. Explain how, for each $\epsilon \geq 0$, one can associate a clique complex K_ϵ to a symmetric $n \times n$ matrix of distances between n items. For $\epsilon = 2.5$ and for the following 6×6 matrix of distances

$$\begin{pmatrix} 0 & 1 & 2 & 3 & 3 & 3 \\ 1 & 0 & 1 & 3 & 3 & 3 \\ 2 & 1 & 0 & 1 & 3 & 3 \\ 3 & 3 & 1 & 0 & 1 & 2 \\ 3 & 3 & 3 & 1 & 0 & 1 \\ 3 & 3 & 3 & 2 & 1 & 0 \end{pmatrix}$$

   determine the simplicial complex K_ϵ; then sketch the geometric realization $|K_\epsilon|$ and calculate the Euler characteristic $\chi(K_\epsilon)$.

4. Consider the collection $\mathcal{U} = \{\{1,2,3\},\{2,3,4\},\{3,4,5\},\{4.6\},\{1,6\}\}$ of sets. Sketch the nerve $N\mathcal{U}$ and calculate its Euler characteristic $\chi(N\mathcal{U})$.

5. Define what is means for two maps to be *homotopic* and for two spaces to be *homotopy equivalent*.

6. Let $Y \subset \mathbb{E}^n$ be an arbitrary convex subset of Euclidean space and let X be an arbirary topological space. Prove that any two continuous maps $f, g\colon X \to Y$ are homotopy equivalent.

7. Prove that any convex subspace $Y \subset \mathbb{E}^n$ is homotopy equivalent to the space consisting of a single point.

8. Prove that homotopy equivalence of maps $f \simeq g$ is an equivalence relation on the set of continuous maps $X \to Y$ from a given space X to a given space Y.

9. Give a careful proof that $\mathbb{S}^1 = \{z \in \mathbb{C} : |z| = 1\}$ is homotopy equivalent to $\mathbb{C} \setminus \{0\}$.

10. State Leray's theorem about the nerve $N\mathcal{U}$ of an open cover $\mathcal{U}$ of a space X. Illustrate the theorem by using an appropriate open cover of the annulus $X = \{z \in \mathbb{C} : 1 \leq |z| \leq 2\}$. Calculate the Euler characteristic $N\mathcal{U}$ in your illustration.

11. Give an account of the Mapper clustering algorithm. In your account, illustrate the algorithm on a set of points chosen from an annulus.

12. Give an example of a finite data set $S \subset \mathbb{R}^3$ for which 'topological information' will be lost under any linear projection $\mathbb{R}^3 \to \mathbb{R}^2$, but for which this information might be preserved under a map $S \to K$ to a 2-dimensional simplicial complex K produced from the Mapper clustering procedure.

7 Homology and Persistent Homology

1. Consider the mod-2 chain complex C_*K of the simplicial complex with $V = \{1, 2, 3, 4\}$, $K = \{\{1\}, \{2\}, \{3\}, \{4\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{2, 4\}, \{3, 4\}, \{1, 2, 3\}\}$.

   (a) Determine the matrices for the linear homomorphisms $d_2 \colon C_1K \to C_0K$ and $d_1 \colon C_2K \to C_1K$.

   (b) Use these matrices to compute $H_0(C_*K)$ and $H_1(C_*K)$.

2. Consider the simplicial complex with $V = \{1, 2, 3, 4, 5, 6\}$, and with K consisting of the six 2-simplices $\{1, 2, 3\}, \{2, 3, 4\}, \{3, 4, 5\}, \{4, 5, 6\}, \{1, 5, 6\}, \{1, 2, 6\}$ together with all non-empty subsets of these 2-simplices. Sketch the geometric realization K. Then exhibit a simplicial subcomplex L of K for which the inclusion $|L| \hookrightarrow |K|$ is a homotopy equivalence. (Don't prove the homotopy equivalence.) Hence determine the mod-2 homology $H_n(K)$ for all $n \geq 0$.

3. Explain what is meant by:

   (a) a *simplicial complex* K,

   (b) a *filtered* simplicial complex K,

   (c) the degree n *persistent Betti numbers* $\beta_n^{s,t}$ of a filtered simplicial complex K.

4. The degree 0 and degree 1 persistent Betti numbers of a certain filtered simplicial complex K are given by the matrices

$$\beta_0^{s,t} = \begin{pmatrix} 3 & 0 & 0 & 0 \\ 0 & 3 & 1 & 0 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 3 \end{pmatrix} \quad \beta_1^{s,t} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 3 & 2 & 2 \\ 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 \end{pmatrix}.$$

   How many 'persistent connected components' and how many 'persistent 1-dimensional holes' does K have?