Exercises

1	2	3	4	5	6	7	8	9	10
11	12								

1	1	1	1	1	1	1
2	2	2	2	2	2	2
3	3	3	3	3	3	3
4	4	4	4	4	4	4
5	5	5	5	5	5	5
6	6	6	6	6	6	6
7	7	7	7	7	7	7
8	8	8	8	8	8	8
9	9	9	9	9	9	9
0	0	0	0	0	0	0

Surname, First name

____

Modelling, Uncertainty and Data for Engineers (CEGM1000)
Resit 22/23 Q2

Do not open the exam until given permission by the instructor!
(You can write your name and student ID on this page)

The exam is 180 minutes. The table below gives an overview. On the following pages, some questions have a specific box for you to answer: anything written outside the boxes will not be graded. Note that we have provided a lot of space for answers. The answer space size is not an indicator of how long we expect your answers to be! (shorter is generally better). Scratch paper is available to use during the exam, but will not be collected or graded. You may use pen or pencil and a scientific calculator. Any required equations are generally provided in the question description.
Don't forget to write your student ID and fill in the bubbles on the top right of this page. Good luck!

No.	Question	Sub-Q	Type	Points
1	Coding	a, b, c, d	MC	4
2	Finite Difference Methods	a, b	MC	6
2		c	Open	3
3	Finite Element Methods	a, b	MC	5
4	Finite Element Methods	a, b, c	MC	7.5
5	Optimization	a, b, c	Open	15
6	Signal Processing	a, b, c	Open	15
7	Time Series	a, b, c, d	Open	15
8	Machine Learning		Open	8
9	Machine Learning		MC	2
10	Machine Learning		MC	4
11	Risk and Reliability	a, c, d	Open	9
		b	MC	1.5
12	Risk and Reliability	a	Open	3
		b	MC	1.5
		Total:		99.5

Part 1: Coding

1p **1a** Who benefits from applying the FAIR principles?

- ○a Primarily system admins and data stewards, so as to keep the data well organized
- ○b Researchers, publishers, and software builders, among other relevant stakeholders
- ○c FAIR is a principle that guides data storage and retrieval, so it is only relevant for the algorithm in charge of storage and retrieval
- ○d Only corporations profit from data. FAIR data is private data and, therefore, behind a paywall

1p **1b** What does it mean that data is Findable?

- ○a It means that data is easily accessed by the users. To this end, users need an authentication process
- ○b It means that data is easy to locate for its retrieval. To this end, data must be appropriately labeled
- ○c It means that data can be integrated with other data. To this end, data must be standardized
- ○d It means that data can be used and reused by diverse users. To this end, it is important to make clear its origins and conditions of (re)use

1p **1c** Can private data be FAIR?

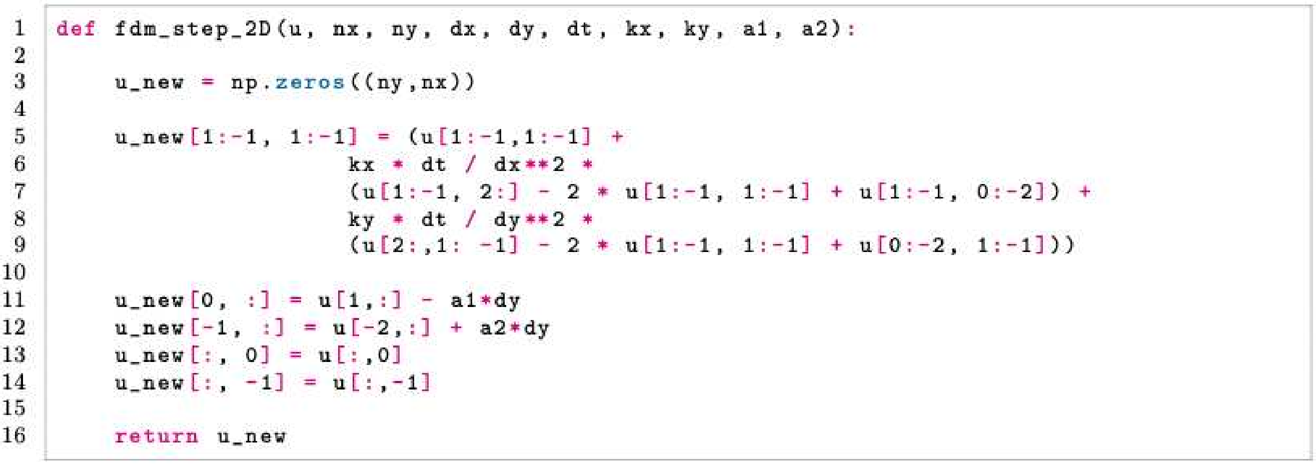
- ○a Yes. Accessible data does not mean "open data". FAIR data might only be accessible by the relevant stakeholders (for instance, physicians have access to medical records which are not open to the public)
- ○b All FAIR data is private data. So yes, private data can be FAIR
- ○c All FAIR data is open data. So no, private data cannot be FAIR

1p **1d** What is the value of implementing the FAIR principles?

- ○a The value of the FAIR principles is that they keep the data well-structured and easy to find
- ○b There are multiple values in the FAIR principles. Here are just some: It recognizes research by making data findable'; It democratizes access to data by allowing numerous researchers to access the same qualitative data regardless of their institution or origin; It facilitates the reproducibility of research, paving the way to more reliable scientific research
- ○c The value of the FAIR principles is economic: institutions, governments, and companies save millions of dollars annually, well-structuring the data to be findable, accessible, interoperable, and reusable
- ○d There is no value in the FAIR principles other than technical value: data can be found quickly and effectively, data can be shared using similar formats, and data can be reused by the same researchers.

Part 2: Finite Difference Method

Consider the following implementation of a finite difference time stepper:



```
1   def fdm_step_2D(u, nx, ny, dx, dy, dt, kx, ky, a1, a2):
2
3       u_new = np.zeros((ny,nx))
4
5       u_new[1:-1, 1:-1] = (u[1:-1,1:-1] +
6                       kx * dt / dx**2 *
7                       (u[1:-1, 2:] - 2 * u[1:-1, 1:-1] + u[1:-1, 0:-2]) +
8                       ky * dt / dy**2 *
9                       (u[2:,1: -1] - 2 * u[1:-1, 1:-1] + u[0:-2, 1:-1]))
10
11      u_new[0, :] = u[1,:] - a1*dy
12      u_new[-1, :] = u[-2,:] + a2*dy
13      u_new[:, 0] = u[:,0]
14      u_new[:, -1] = u[:,-1]
15
16      return u_new
```

3p **2a** What kind of approximation of the time derivative is used in the given implementation?

- ⓐ Forward difference
- ⓑ Central difference
- ⓒ Backward difference

3p **2b** Running this code in a loop over time steps gives an approximate solution for a PDE. This PDE has a term with k_x (kx in the code). To which PDE term does kx-string correspond?

- ⓐ $k_x x$
- ⓑ $k_x \frac{\partial u}{\partial x}$
- ⓒ $k_x \frac{\partial^2 u}{\partial x^2}$
- ⓓ $k_x \frac{\partial^4 u}{\partial x^4}$

3p **2c** On which line(s) in the code block are Neumann boundary conditions applied?

Part 3: Finite Element Method 1

2.5p **3a** Two-point Gauss integration on a 1D subdomain running from $\xi \in [-1,1]$ is performed with $\xi_{ip} = \pm 1/\sqrt{3}$. For a general element defined on the domain x ∈ [a, b], what are the x positions of the integration points:

- ⓐ $x_{ip} = \frac{a+b}{2} \pm \frac{b-a}{2\sqrt{3}}$
- ⓑ $x_{ip} = \frac{b-a}{2} \pm \frac{a+b}{2\sqrt{3}}$
- ⓒ $x_{ip} = \frac{a+b}{2} \pm \frac{b-a}{\sqrt{3}}$
- ⓓ $x_{ip} = \frac{b-a}{2} \pm \frac{a+b}{\sqrt{3}}$

2.5p **3b** What is the order of polynomial that can be integrated exactly with the two-point Gauss integration scheme?

- ⓐ First order (linear function)
- ⓑ Second order (quadratic function)
- ⓒ Third order (cubic function)
- ⓓ None, numerical integration always results in an approximation

Part 4: Finite Element Method 2

Consider a PDE of the type $-\alpha\frac{\partial^2 u}{\partial x^2} + \beta u - \gamma = 0$ where $u(x)$ is the primary unknown and $\alpha(x)$, $\beta(x)$, and $\gamma(x)$ are given and independent of u. Following the conventional notation, let the $\mathbf{N}$-matrix contain shape functions, the $\mathbf{B}$-matrix contain shape function derivatives. In the discretized form with the finite element method, terms with the parameters α, β, and γ appear. What form do these terms have:

2.5p **4a** with α?

- ⓐ $\int \mathbf{N}^T \alpha dx$
- ⓑ $\int \mathbf{B}^T \alpha dx$
- ⓒ $\int \mathbf{N}^T \alpha \mathbf{N} dx$
- ⓓ $\int \mathbf{B}^T \alpha \mathbf{B} dx$

2.5p **4b** with β?

- ⓐ $\int \mathbf{N}^T \beta dx$
- ⓑ $\int \mathbf{B}^T \beta dx$
- ⓒ $\int \mathbf{N}^T \beta \mathbf{N} dx$
- ⓓ $\int \mathbf{B}^T \beta \mathbf{B} dx$

2.5p **4c** with γ?

- ⓐ $\int \mathbf{N}^T \gamma dx$
- ⓑ $\int \mathbf{B}^T \gamma dx$
- ⓒ $\int \mathbf{N}^T \gamma \mathbf{N} dx$
- ⓓ $\int \mathbf{B}^T \gamma \mathbf{B} dx$

Part 5: Optimization

6p **5a** Consider the example of an animal feed producer that intends to produce a special type of product with specific characteristics. This feed is constituted of two types of cereals and one binding component. 10 kg of this feed has to have a minimum of 900g of nutrient A, 500g of nutrient B, and 30 g of C so that it can be sold to farms. Knowing that 1) the amount of cereal 2 cannot exceed double the amount of cereal 1 and the binder material together; 2) there must be always at least 1 kg of binding material per 500 g of cereal 1 and cereal 2, and that 3) the cereals have the following nutrients and costs:

	Nutrient A	Nutrient B	Nutrient C	Cost (m.u./kg)
Cereal 1	100 g/Kg	80 g/Kg	10 g/Kg	40
Cereal 2	200 g/Kg	150 g/Kg	-	60
Binder	-	-	-	30

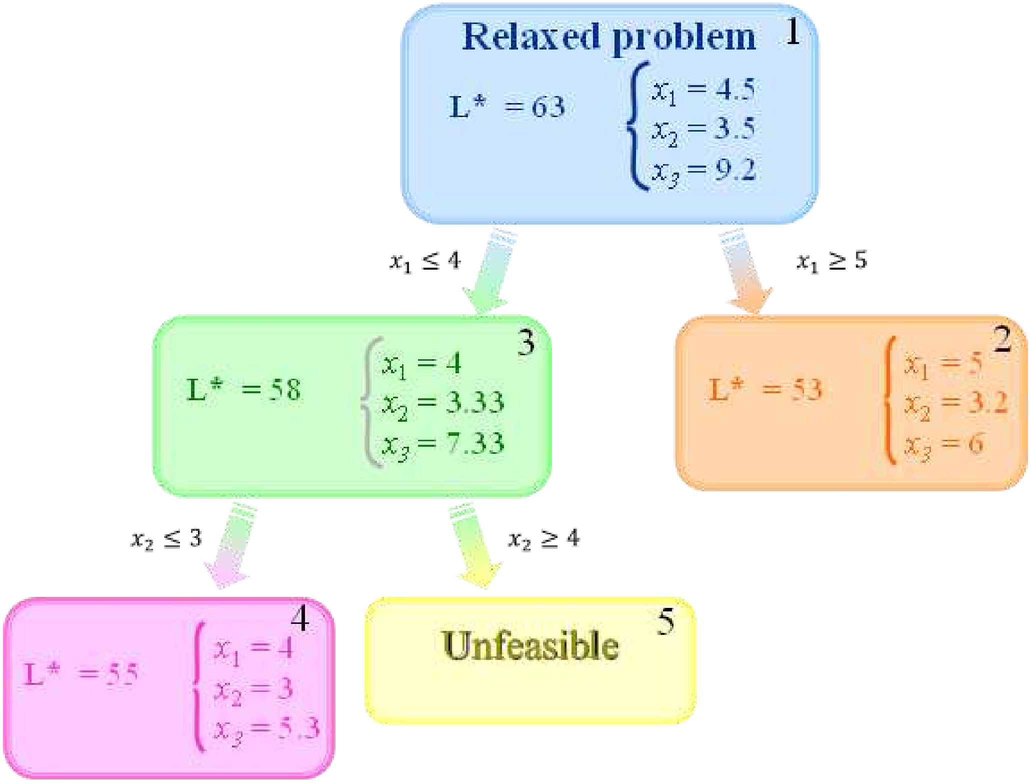
Formulate the problem that would decide on the amount of Cereal 1, Cereal 2, and Binder that this producer should select per 10 kg of animal feed. **Do not solve the problem!** Only the formulation is required.

↳

6p **5b** Solve the following mathematical programming problem using the graphical solution method. Clearly state the solution and the value of the objective function in your answer as well as the intermediate steps you need to take.

$$\begin{aligned}&Minimize(F) = 6X + 5Y\\&Y \leq \tfrac{5}{6}X + 5\\&X \leq 3\\&Y \leq 5\\&Y \geq 0; X \in \mathbb{R}\end{aligned}$$

3p **5c** The following is the solving process of a mixed integer program - some variables are integers and other variables are continuous - using branch and bound with the objective of maximization. x_1 and x_2 are integer variables and x_3 is continuous. 5 nodes have been explored in the tree in the order that is shown in the upper right corner of each node. Has the optimal solution been obtained? Justify your answer with the information you see in the tree.

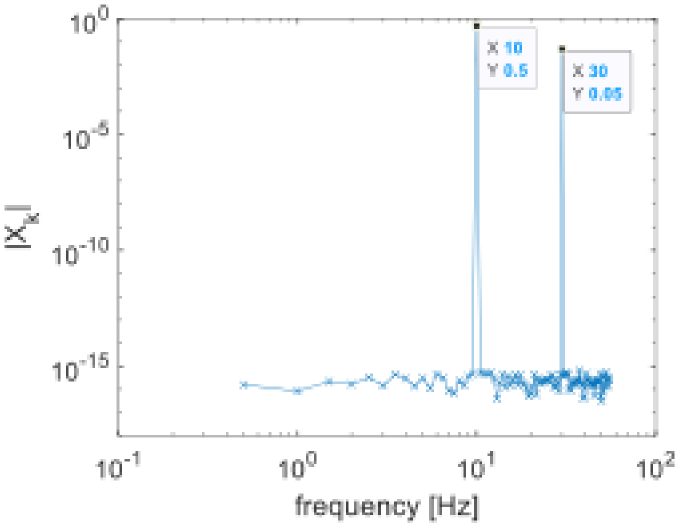


Part 6: Signal Processing

A continuous-time signal $x(t)$ is given as: $x(t) = A_1 \cos(2\pi f_1 t) + A_2 \cos(2\pi f_2 t)$, with $A_1 = 1$, $A_2 = 0.1$, $f_1 = 10$ Hz, and $f_2 = 80$ Hz. In three experiments the signal has been sampled using each time a different sampling frequency (f_s) and a different measurement duration (T_{meas}). The frequency domain plots (magnitude spectrum in logarithmic scale, as a result of the DFT) are shown below; the spectrum is double sided, but only shown for positive frequencies, and, as commonly done in practice, up to the Nyquist frequency. The values X_k, straight from the FFT, have been divided by N, the number of samples.

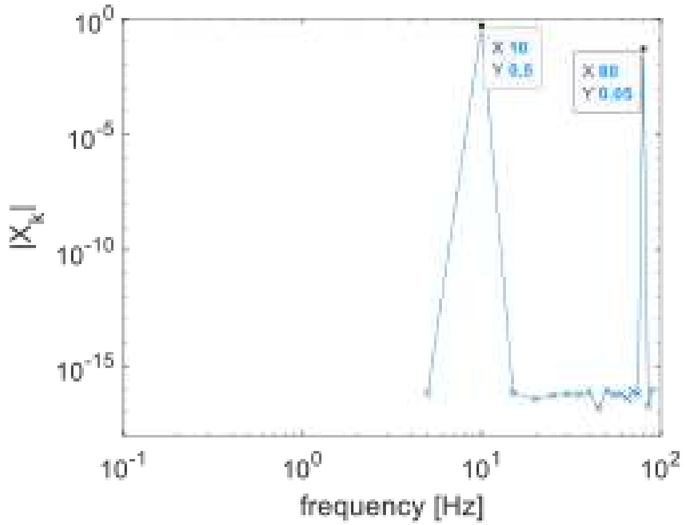
Determine, for each experiment/plot, the sampling frequency (f_s), as well as the measurement duration (T_{meas}). Only the final numerical answers are asked!

5p



6a

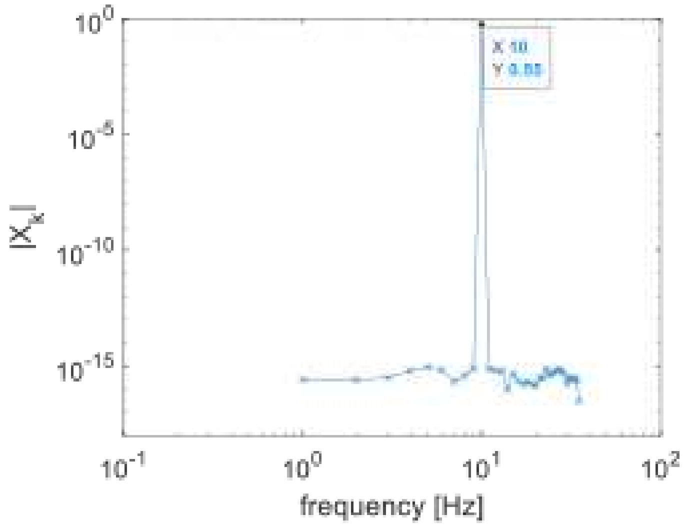
5p



$|X_k|$
10^{0}
10^{-5}
10^{-10}
10^{-15}
X 10
Y 0.5
X 80
Y 0.05
10^{-1}
10^{0}
10^{1}
10^{2}
frequency [Hz]

6b

5p



6c

Part 7: Time Series Analysis

The deformation pattern (in mm) of the East component of a global navigation satellite system (GNSS) station is expressed as the following equation:
$y(t) = 15 + 5t + 3\cos 2\pi t + 4\sin 2\pi t + \cos 6\pi t + 2\sin 6\pi t$
where t is 'time' in the unit of year. In order to verify all the coefficients and frequencies of this deformation pattern, we have measured 2-year daily positions of this East component. The time series is then $y = [y(t_1), ..., y(t_{730})]^T$, with $t_1 = 1/365, t_2 = 2/365$ and $t_{730} = 2$ (years). Further we assume that the measurements consist of ARMA(2,0) noise, with a standard deviation of $\sigma = 5$ mm. You are required to answer the following questions:

2p **7a** Assuming the expression for the deformation pattern given above, compute the amplitude and initial phase of the periodic signals of this time series ($A =?, \theta =?$).
You may make use of the following formulae:
$y(t) = a\cos(\omega t) + b\sin(\omega t) = A\sin(\omega t + \theta)$
$A = \sqrt{a^2 + b^2}, \theta = \arctan\frac{a}{b}$.

4p **7b** Assume that we applied the least-squares harmonic estimation (LS-HE) to compute the power spectral density (PSD) for this time series. Sketch (rough drawing) to illustrate its PSD, and its critical value in a given confidence level. Label the horizontal and vertical axes, and indicate the values on the horizontal axis corresponding to the locations of the PSD peaks.

5p **7c** If we assume that the frequencies are given, but the coefficients of the periodic, linear, and bias terms are unknown, what would be your suggestion to estimate such unknown parameters? Write in particular the first and last rows of the design matrix A for the given time series.

4p **7d** In order to implement prediction, the noise characteristics of time series should be determined using the normalized auto-covariance function (ACF) and partial ACF (PACF). Sketch the ACF and PACF for the given time series.

Part 8: Machine Learning 1

Devise a suitable machine learning approach for the problem described below concerning the prediction of water salinity in a river.

- How would you frame this problem from a machine learning perspective, e.g., type(s) of learning and task(s)?
- Which techniques would you employ and why?
- What are the main steps needed to implement them correctly?

Describe your approach based on what you learned in class in **max 300 words**. Use this number to gauge the level of detail in your description. Note that extra details on the case study or knowledge of this specific topic are not needed to answer correctly. While you can approach this problem using the tools of Time Series Analysis, here we ask you to use what you have learned in Machine Learning.

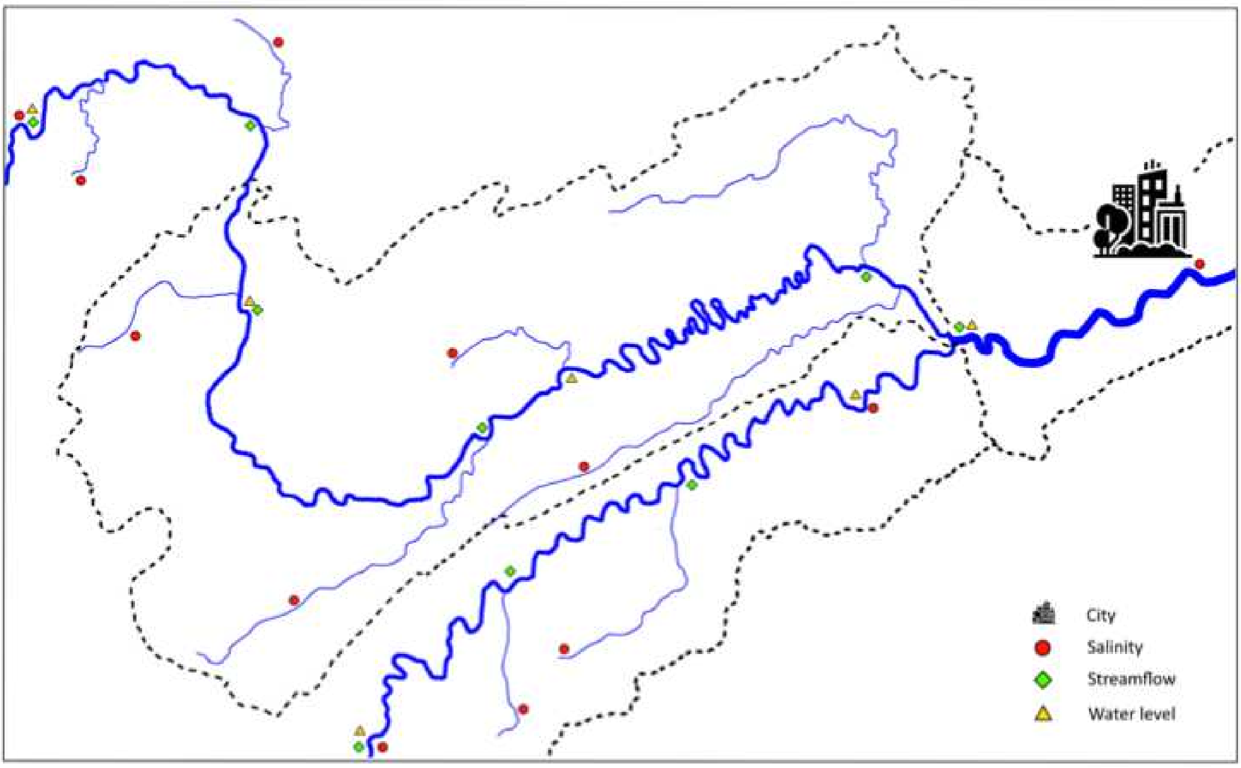
Problem description

To face increasing drinking water demands due to urbanization, the water managers of a given metropolitan area decide to draw water from a large river flowing nearby. Due to the high levels of dissolved minerals in the river's catchment (i.e., the area of land that drains into a particular river), the water pumped from the river may experience high levels of salinity. Using water with high salinity has adverse effects on both domestic and industrial users. By knowing salinity values in advance, operators can change the pumping policies so that more water is pumped during days of low salinity and less water is pumped in high salinity days. **Your task is to develop a machine learning model that forecasts the average daily salinity in the river nearby the city, one week in advance, that is at time t + 7**, where t identifies the current day.

The water utility provides you with complete daily average time series data for salinity, flow, and water levels for multiple locations in the river and its tributaries (see map). This includes salinity data at the target location, measured nearby the city. You are also provided with the following information:

- Some physical processes governing the river can be approximated as linear; others are nonlinear.
- All measured variables are continuous; they have different units of measurement and distribution.
- All measured variables, at all locations in the catchment, may help predict salinity nearby the city;
- For each variable at each location, different lagged values (e.g., variable measured at t, t − 1, t − 2, ...) may provide additional explanatory power. This increases the overall amount of variables in the dataset to a few hundreds.
- The final dataset is given in tabular format, where each column is a variable measured in a certain location at a given time lag. Given the nature of the problem and the different lags, there is likely a high linear correlation between many variables.
- You have no access to other sources of data.

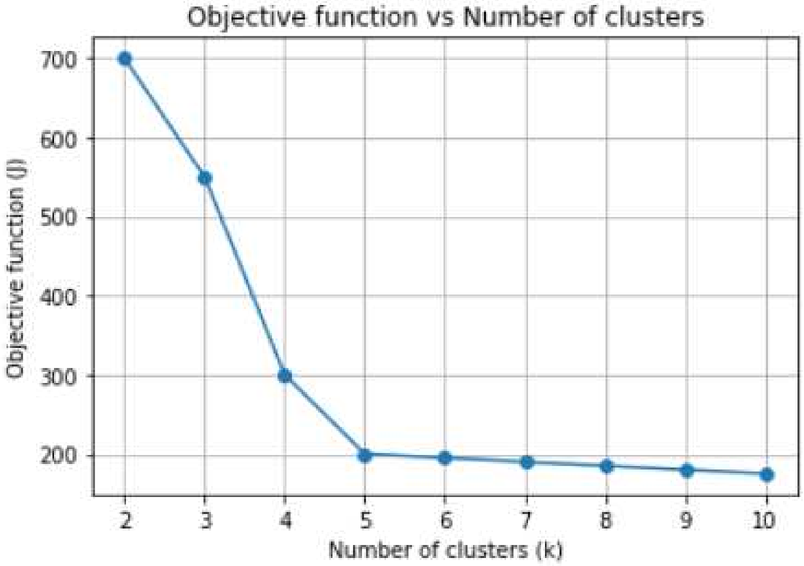
8p **8**



⤵

Part 9: Machine Learning 2

2p



9

Looking at the change of objective function when performing K-means clustering on a dataset, which number of clusters is optimal using the Elbow method?

ⓐ 3 ⓑ 4 ⓒ 5 ⓓ 10

Part 10: Machine Learning 3

4p **10** When building machine learning models for regression, a number of crucial aspects related to model selection and bias/variance tradeoff must be considered. From the list of statements below, identify all the ones that are true. Each wrong answer will result in the loss of one point, with the lowest possible score being zero (i.e., we will not subtract points from the rest of the exam).

□ Increasing the number and/or size of the layers in a neural network can make it fit arbitrarily complex functions. This in turn means neural nets can achieve an expected loss of zero for arbitrarily noisy datasets of these complex functions;

□ Fitting noisy functions using neural networks with a large number of parameters tends to lead to models with high variance, that is, with high sensitivity to specific realizations of small datasets;

□ Switching from large neural networks to smaller linear basis function models should translate to significant increases in bias, leading to models that are more robust to changes in dataset;

□ When adding an L2 regularization term $\frac{\lambda}{2}\mathbf{w}^T\mathbf{w}$ to the loss function, higher values of λ should lead to models with higher variance and lower bias;

□ Given a model with a single regularization hyperparameter λ, an appropriate model selection procedure would be to pick the value of λ that minimizes the error computed over a test dataset;

Part 11: Reliability 1

You are asked to evaluate the system reliability of an industrial facility, specifically with respect to limiting the chance that people working and living near the plant may die due to one of three different failure modes, M_i

(note: the descriptions of the failure modes and the numbers in the table are not realistic, but meant to illustrate how the system behaves)

- M_1: a toxic chemical leaks into the groundwater system, poisoning the entire community
- M_2: a toxic gas leak occurs and the detection system does not provide a warning
- M_3: an explosion occurs and the automatic system to put out the fire fails

All failure modes are mutually exclusive, and the probability of occurrence is provided in the following table, along with the expected fatalities:

	$P(M_i)$	Fatalities, N (-)
M_1	0.02	1000
M_2	0.03	50
M_3	0.05	10

3p **11a** Construct the FN curve for the industrial facility. Don't worry about the scale of your plot being precise, as long as the FN values are clearly indicated at each point.

↳

1.5p **11b** Which of the following statements best describes the dependence between each failure mode?

ⓐ Statistically independent

ⓑ Strong positive dependence

ⓒ Strong negative dependence

ⓓ Not possible to determine with information provided

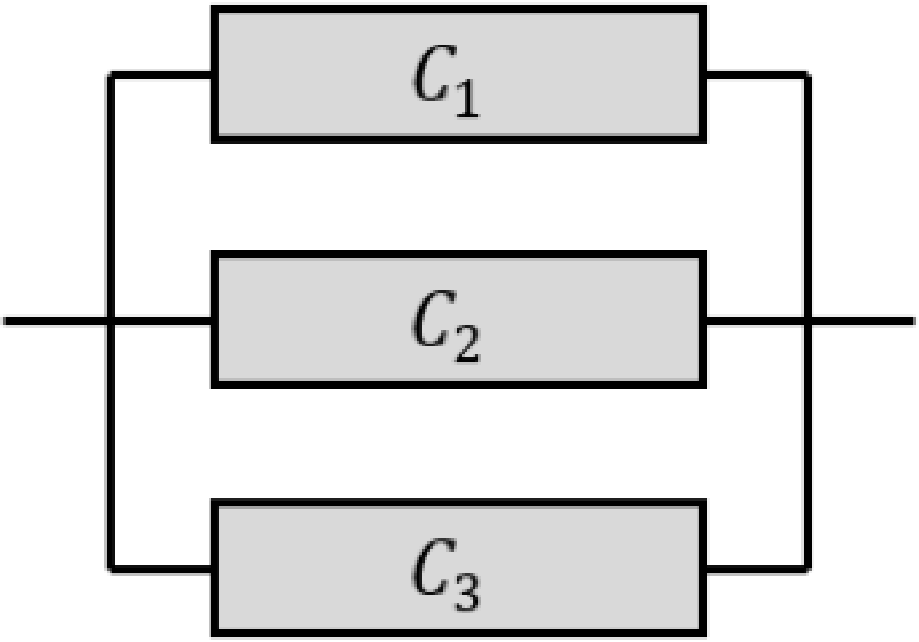
The regulatory agency in charge of the safety standards in this area have specified a limit line of the form C/n^{α}, where $C = 0.2$ and $\alpha = 0.5$, which indicates the system may not be safe.

3p **11c** Explain why the system is not safe (include a quantitative result in your explanation and show your work; you may draw the limit line on the plot from your previous answers)

3p **11d** Propose one mitigation measure that you can implement to make the system safe. Be sure to specify exactly how your measure would influence the calculations made in the previous 3 questions.

Part 12: Reliability 2

You are designing a system that can be illustrated as follows, where the failure probabilities of components 1, 2 and 3 are 0.1, 0.2 and 0.5, respectively.

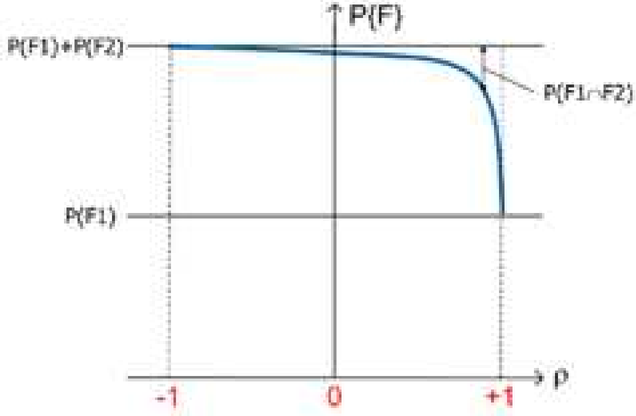


Recall that the failure probability of a series and parallel system can be determined using the following equations:

Parallel: $p_f = \prod_{i}^{n} P(F_i)$

Series: $p_f = 1 - \prod_{i}^{n}(1 - P(F_i))$

The following figure from the lecture notes may also help you answer this question:



3p **12a** What is the failure probability of the system?

1.5p **12b** How would positive dependence between the components influence the system failure probability?

ⓐ Increase

ⓑ Decrease

ⓒ No change

ⓓ Not enough information provided

This page is left blank intentionally