Exercises

1	2	3	4	5	6	7	8	9

Surname, First name

____

Modelling, Uncertainty and Data for Engineers EXAM (CEGM1000)
Exam Q2
25 Jan, 2024, 13:30-16:30

○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

○a ⊗b ○c ○d ○e ○f → b
○a ○b ⊗c ○d ○e ○f → c
● ○b ○c ⊗d ○e ○f → a

□a ☒b □c ☒d □e □f → b, d
□a □b ☒c □d □e □f → c
■ □b □c ■ ☒e □f → a, d

Answer multiple-choice questions as shown in the example. Circular checkboxes have one correct answer. Square checkboxes may have multiple correct answers.

Do not open the exam or turn to the back page until given permission by the instructor!
(you may write your name and student ID)

Before you start the exam, a few remarks:

- Write down your first and last name in the field on the top left corner of this page
- Fill in your student number on the top right corner of this pages. Fill in the number in the boxes on top, and mark the corresponding number. Fill the corresponding circle as indicated above.
- You may use pen or pencil and a scientific (non-graphing) calculator. Any other tools and sources of information are not allowed.
- Exam should remain stapled.
- Scratch paper is available to use during the exam, but will not be collected or graded.
- 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.
- In case you want to erase and rewrite your answer, ask an invigilator for white stickers to cover your incorrect answer.
- The answer space size is not an indicator of how long we expect your answers to be! Shorter is generally better.
- In case you want to correct your answer for a multiple choice question, follow the instructions above. If you mess up, put an arrow to the answer you think is correct. If you mess up again, add a comment.
- A summary of points and questions is provided on the last page, as well as examples of how to correct your multiple choice answers.

Good luck!

Exercise 1: Programming

4p **1a** Review the code below, that has one part cut out (YOUR_CODE_HERE), then choose the answer that would generate the output provided (note the length of the blank space is irrelevant):

```
bands = [’foo fighters’, ’kenny rogers’, ’acda en de munnik’, ’boston’, ’ac/dc’]

print(’My favorite bands are:’)

for YOUR_CODE_HERE:
  print(f’{i+1}. {j}’
```

The output found was:

```
My favorite bands are:
1. foo fighters
2. kenny rogers
3. acda en munnik
4. boston
5. ac/dc
```

What is the missing piece in the code?

- ○ a i, j in range(bands)
- ○ b j in range(bands), i in bands
- ○ c i in range(bands), j in bands
- ○ d i, j in zip(bands)
- ○ e j, i in zip(bands)
- ○ f i, j in enumerate(bands)

2p **1b** Which of the following would be useful to include in a .gitignore file in a MUDE repository? (more than one answer is possible)

- □ *.csv
- □ .csv
- □ *.md
- □ *.ipynb_checkpoints

2p **1c** You open a Jupyter notebook for Week 2.9, and see that a DataFrame is defined as df and used to store the data imported from a csv file. Choose the piece of code that provides summary information about the data, including relevant statistics

- ⓐ df.head()
- ⓑ df.describe()
- ⓒ df.summarize()
- ⓓ df.tail()
- ⓔ df.sanitize()

Exercise 2: Heat Equation with Finite Element

Consider the heat equation on a 1-D domain Ω

$\frac{\partial T}{\partial t} = \alpha \frac{\partial^2 T}{\partial x^2}$

where $T(x,t)$ is the temperature as a function of space and time and α the thermal diffusivity, with Dirichlet boundary conditions (prescribing temperature) on one end of the domain and Neumann boundary conditions (prescribing the heat flux) on the other end.

The strong form PDE can be rewritten as a weak form and then discretized in space with finite elements to arrive at a semi-discrete system of equations:

$\mathbf{A}_1 \frac{\partial \mathbf{v}}{\partial t} + \mathbf{A}_2 \mathbf{v} = \mathbf{b}$

where $\mathbf{A}_1$ and $\mathbf{A}_2$ are matrices and $\mathbf{b}$ and $\mathbf{v}$ are vectors.

Let $\mathbf{N}$, $\mathbf{B}$ and $\mathbf{C}$ be row vectors containing the shape functions, shape function derivatives and second derivatives of shape functions respectively, i.e. $\mathbf{B} = \frac{\partial \mathbf{N}}{\partial x}$ and $\mathbf{C} = \frac{\partial^2 \mathbf{N}}{\partial x^2}$.

3p **2a** What is the meaning of $\mathbf{v}$?

- ⓐ Heat fluxes between the elements
- ⓑ Heat fluxes at the integration points
- ⓒ Temperature values at the nodes
- ⓓ Temperature values at the integration points

3p **2b** What are the contents of $\mathbf{A}_1$?

- ⓐ $\int_\Omega \mathbf{N}^T \mathbf{B} d\Omega$
- ⓑ $\int_\Omega \mathbf{N}^T \mathbf{N} d\Omega$
- ⓒ $\int_\Omega \mathbf{B}^T \mathbf{B} d\Omega$
- ⓓ $\int_\Omega \mathbf{B}^T \mathbf{N} d\Omega$

3p **2c** What are the contents of $\mathbf{A}_2$?

- ⓐ $\int_\Omega \mathbf{N}^T \alpha \mathbf{N} d\Omega$
- ⓑ $\int_\Omega \mathbf{B}^T \alpha \mathbf{B} d\Omega$
- ⓒ $\int_\Omega \mathbf{N}^T \alpha \mathbf{C} d\Omega$
- ⓓ $\int_\Omega \mathbf{C}^T \alpha \mathbf{N} d\Omega$

3p **2d** When would we get $\mathbf{b} = \mathbf{0}$?

ⓐ Always because there is no source term in the strong form equation
ⓑ When Dirichlet boundary conditions are equal to zero
ⓒ When Neumann boundary conditions are equal to zero

Exercise 3: Finite _______ Methods

3p **3** For different modelling approaches, continuous problems are discretized in different ways. Connect the three methods below to the type of discretization they are most strongly related to.

Use three straight lines to connect the term on the left that best matches the term on the right, below. If you make mistakes or need to correct your choice, use the space below to write a note that clarifies your final answer. If multiple answers are provided, you will not receive any credit.

Finite Element Method	Derivatives are discretized
Finite Difference Method	Conservation is discretized
Finite Volume Method	The solution is discretized

Exercise 4: Signal Processing

10p **4** We start from a cosine signal with a frequency of $f = 3$Hz, and sample it as $f_s = 8$Hz, for a duration of $T = 2$s. The $N = 16$ discrete time samples are input to the Discrete Fourier Transform (DFT) and we directly plot the magnitude (modulus) of the output, hence $|X_k|$ with $k = 0, ..., N - 1$. Create the resulting plot.

Exercise 5: Time-Series Analysis

We intend to simulate a time series of 200 samples at 1-day intervals (so $m = 200$ and time unit is day), using a first-order auto-regressive $AR(1)$ random process s_t as follows:

$$s_t = \beta s_{t-1} + e_t$$

where $t = 1, ..., m$ and $\beta = -0.9$ is the given $AR(1)$ parameter. We further assume $E(s_t) = 0$ and $D(s_t) = \sigma^2 = 2$. The series is simulated using a normal distribution where it is initialized using $s_1 = s(t = 1) = randn(1)$ as the first point. Further the standard deviation of e_t can be obtained from $\sigma_e^2 = \sigma^2(1 - \beta^2)$.

Given the information above, answer the following True/False or multiple choice questions.

2p **5a** The time series is non-stationary.

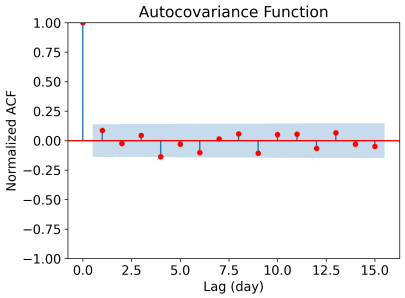
a True　　b False

2p **5b** The time series is characterized by colored noise.

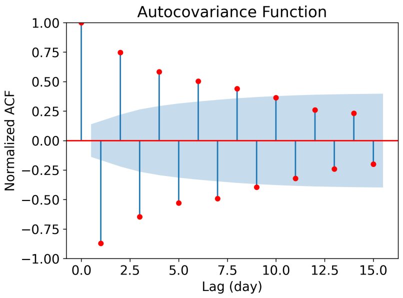
a True　　b False

4p **5c** Select which of the four plots below shows the correct autocovariance function for the time series

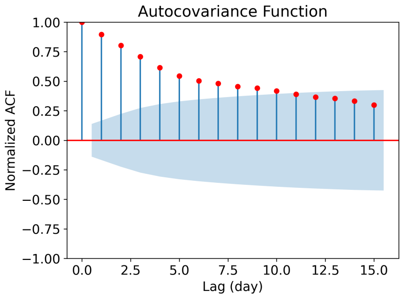
OPTION A



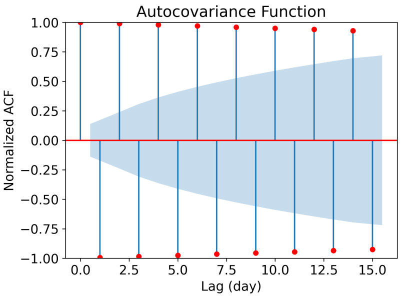
OPTION B



OPTION C



OPTION D



a A　　b B　　c C　　d D

Exercise 6: Optimization

Imagine you are managing a warehouse with limited storage space. The warehouse has different types of products, each requiring a specific amount of storage space. Your goal is to maximize the total value of the stored products while respecting the storage capacity constraints.

8p **6a** You have a list of available products, i, each associated with a value of profit, and each product has a known storage space requirement per unit. The warehouse has a fixed total storage capacity.

Your task is to formulate a model that helps decide how many units of each product to store in the warehouse to maximize the total value of the stored products, while ensuring that the total storage space used does not exceed the warehouse capacity.

You are free to define and describe your own variables and symbols in your model formulation, as long as you explain what they represent.

4p **6b** Discuss the type of variables that you proposed for the formulation - are there different options? What is that choice depending on?

Exercise 7: Machine Learning

Imagine you would like to train a machine learning model for regression: you attempt to map a set of D input features $\mathbf{x}$ to a single target variable t using a dataset with N observations $(\mathbf{x}_n, t_n)$. The task is therefore to learn a model $y(\mathbf{x})$ that is hopefully as close as possible to the observations t.
First of all, you opt for a squared error loss function to quantify how well the model does:

$$L(t, y) = (y(\mathbf{x}) - t)^2$$

and from a Decision Theory perspective, your goal is to minimize:

$$\mathbb{E}[L] = \int \int (y(\mathbf{x}) - t)^2 p(\mathbf{x}, t) \mathrm{d}\mathbf{x}\mathrm{d}t$$

2p **7a** Minimizing the expected loss leads to the classical result $y(\mathbf{x}) = \mathbb{E}_t[t|\mathbf{x}]$, the conditional expectation of t given $\mathbf{x}$. How can this important result be interpreted?

- ⓐ When making new predictions, it suffices to average t over all possible values of $\mathbf{x}$
- ⓑ Making predictions involves two steps: first fix a value of $\mathbf{x}$ and then average out the noise in t
- ⓒ The value $y(\mathbf{x})$ for a new prediction should be picked to be one of the values of t in the dataset

2p **7b** You then consider a few options for how to model $y(\mathbf{x})$. When comparing parametric and non-parametric models, which one of the following statements is true?

- ⓐ Non-parametric models can be efficient because new predictions only depend on the most recent value of $y(\mathbf{x})$ probed from the model
- ⓑ Parametric k-Nearest Neighbors models can be described by two parameters, namely k and the size of the neighborhood used to make predictions
- ⓒ Parametric models can be advantageous because their training datasets can be fully discarded after training: predictions depend exclusively on $\mathbf{x}$ and the trained parameters
- ⓓ Neural networks can be seen as non-parametric models because the choice of activation function cannot be parametrized (*i.e.* one cannot assign a numerical value to this choice)

2p **7c** You decide to go for a **feedforward neural network**. The next step is to train the model by making use of your limited dataset of N observations as efficiently as you can. Which of the following strategies makes most sense?

- ⓐ Use all N samples for training, but making sure the model $y(\mathbf{x})$ is as flexible as possible in order to avoid overfitting
- ⓑ Use 80% of your N samples to train the model, 10% for hyperparameter calibration and 10% for a final assessment of the model
- ⓒ Use all N samples and train two separate models: one as rigid as possible and the other as flexible as possible. Finally, combine the two models into one by averaging their weights
- ⓓ Use 40% of the N samples to train a first model and 40% to train a second one with the same level of flexibility, with the remaining 20% being left as validation data. When making new predictions, randomly pick one of the two models and use it

2p **7d** With the correct dataset setup, you perform data normalization in order to facilitate training. About this part of the workflow, which of the following options would work best?

- ⓐ Concatenate training, validation and test data into a single matrix and normalize the full dataset in one go
- ⓑ Normalize the training and validation datasets separately, since they will be used to compute different loss functions
- ⓒ First normalize only the training dataset, then use the resulting normalization coefficients for the validation and test datasets

2p **7e** Finally, you use Stochastic Gradient Descent and observe how the training and validation losses of one of your neural networks evolve with the number of epochs. Both of the losses decrease at first but then the validation loss starts to increase. What makes most sense?

- ⓐ Stop training and set the final model to be the one with the lowest historical validation loss
- ⓑ Restart training but now use more epochs and a higher learning rate
- ⓒ Add an extra hidden layer to the network, just before the output layer and with the same activation function, in order to avoid overfitting. Then resume training

Exercise 8: Extreme Value Analysis

As a scientist, you are assessing with your team the performance of a *thingamajig*. As everybody knows, *thingamajigs* are very sensitive to wind, and you need this *thingamajig* to withstand high wind speeds. Your colleague is then performing Extreme Value Analysis on the wind speeds in the location where the intervention takes place and found a time series of wind speeds of approximately 5 years.

General information about extreme value distributions is provided below. Think of it like a formula sheet. You may not need this information for all questions, and may not need all of it.

Recall that the Generalized Extreme Value distribution can be written as follows:

$$P[X < x] = exp\big(-[1 + \xi \tfrac{x-\mu}{\sigma}]^{-1/\xi}\big) \qquad (1 + \xi \tfrac{x-\mu}{\sigma}) > 0$$

whereby the random variable X is defined by the location, scale and shape parameters μ, σ and ξ, respectively. The design value x for annual (yearly) probability of non-exceedance p_y is:

$$x = \begin{cases} \mu - \dfrac{\sigma}{\xi}\Big[1 - \big[-\ln(1 - p_y)\big]^{-\xi}\Big] & \xi \neq 0 \\ \mu - \sigma \ln\big[1 - p_y\big] & \xi = 0 \end{cases}$$

with the design life probability over DL years given as:

$$p_{DL} = 1 - (1 - p_y)^{DL}$$

Recall that the Generalized Pareto distribution can be written as follows for random variable X with parameters th (threshold), shape ξ and scale σ_{th}:

$$P[X < x|X > th] = \begin{cases} 1 - \left(1 + \dfrac{\xi(x - th)}{\sigma_{th}}\right)^{-1/\xi} & \xi \neq 0 \\ 1 - exp\left(-\dfrac{x - th}{\sigma_{th}}\right) & \xi = 0 \end{cases}$$

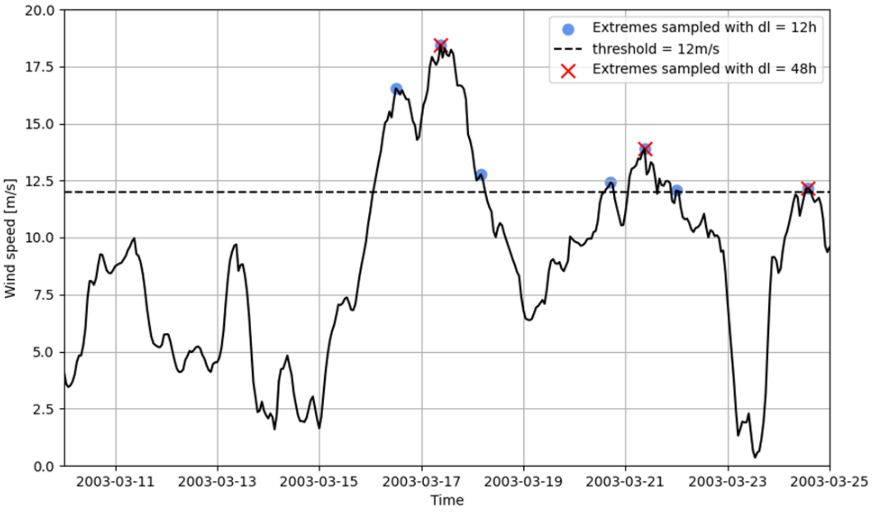
and the design value x_N for an N-years return level.

$$x_N = \begin{cases} th + \dfrac{\sigma_{th}}{\xi}\left[\left(\lambda N\right)^{\xi} - 1\right] & \xi \neq 0 \\ th + \sigma_{th} \ln\left(\lambda N\right) & \xi = 0 \end{cases}$$

where λ is the average number of excesses during the observation period, and N is the N-years return level.

3p **8a** State which sampling method you would recommend, along with a (short!) justification why (2 sentences max.)

Regardless of your advice, your colleague is familiar with Peak Over Threshold, so they decide to go for it. They ask you for advice to choose the parameters of the method: the threshold and the declustering time. They apply two sets of parameters: threshold = 12m/s and a declustering time (dl) of 12h and threshold = 12m/s and a declustering time (dl) of 48h. They show you a zoom of the time series with the extracted extremes (see Figure below).



3p **8b** Based only on the above figure, what declustering time (dl) would you advise your colleague to use? State a specific value and provide a (short!) justification.

4p **8c** Your colleague applied the Peak Over Threshold method and sampled 35 extremes in the time series of 5 years, to which a Generalized Pareto distribution is fit with scale parameter $\sigma_{th} = 1$ and shape parameter $\xi = 0.2$. Remember that the threshold used was 12 m/s. Find the design value for wind speed for a design lifetime of 100 years. State your assumptions and calculation steps clearly.

Exercise 9: Risk & Reliability

An event A is defined as $F_X\left(X > x_A\right)$ and event B is defined as $F_Y\left(Y > y_B\right)$, where X and Y are random variables defined by continuous parametric distributions (F is the CDF). Consider the probability of interest $P\left(A \cap B\right)$ (intersection), read the following two statements and select whether they are true or false.

2p **9a** This is a series system.

ⓐ True ⓑ False

2p **9b** If there is negative dependence between X and Y, the probability would increase.

ⓐ True ⓑ False

The next question is completely independent from those before.

3p **9c** A regulatory authority is deciding how to establish risk-based criteria to protect against fatalities during train accidents (there have been too many over the last few decades).

Sketch an FN-curve that illustrates the following: a recently completed risk analysis; a limit line; clearly labelled x and y axes; specify units, but not values or equations. The y-axis should be labelled using both words and an equation. Points will be given for correctness, not beauty.

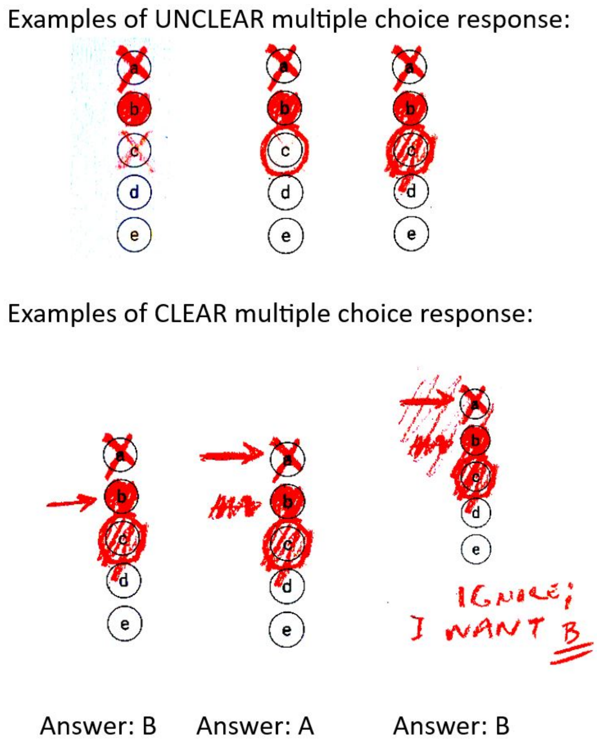
This page left intentionally blank.

Exam Overview and Multiple Choice Examples

The table below gives an overview of the questions to help you plan your time during the exam.

No.	Topic	Number of Sub-parts	Points
1	Programming	3	8
2	Finite Element Method	4	12
3	Finite ______ Analysis	1	3
4	Signal Processing	1	10
5	Time-series Analysis	3	8
6	Optimization	2	12
7	Machine Learning	5	10
8	Extreme Value Analysis	4	10
9	Risk & Reliability	3	7
Total		26	**80**

In case you want to correct your answer for a multiple choice question put an ARROW in front of your final answer. If you also make a mistake with your arrow, write a clear message on the page. Here are a few examples:


Examples of UNCLEAR multiple choice response:
a
b
c
d
e
Examples of CLEAR multiple choice response:
IGNORE;
I WANT B
Answer: B
Answer: A
Answer: B

Multiple choice examples.