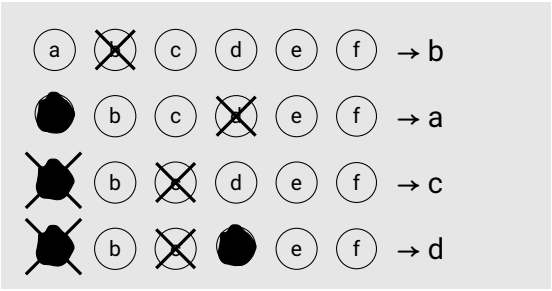
Surname, First name

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

Modelling, Uncertainty and Data for Engineers EXAM (CEGM1000)
Exam Q2 R



Answer multiple-choice questions as shown in the example.

===== DO NOT OPEN THIS EXAM OR TURN IT OVER UNTIL INSTRUCTED TO DO SO =====

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.
- Write **student number** on the **top right** corner as the **number and filled in circle**.
- **New for this exam:** several questions ask require an answer on a figure. In each case, two identical images are provided in case you must correct your work. Cross over the image you do NOT want to submit; if it is not clear which image you want to submit, we will grade the bottom one.
- Place your student ID card face up on your desk.
- The duration of the exam is 3 hours (if entitled to an extension place paper on desk, face up).
- You are only allowed a pen/pencil and approved calculator; nothing else is allowed.
- Note in the multiple choice answer examples above that filled in circles/squares take precedence over an "X." **It is your responsibility to make sure that there is no ambiguity in your final answer**s (for example, add an extra note "final answer = x, y" etc). If in doubt, ask an instructor prior to submitting your exam. Examples are provided on the back page of this exam.
- Write all answers in the space provided (scrap paper is not graded), and do not remove the staple (nietje). In case you must erase, ask an invigilator for a white sticker to cover your incorrect answer. If you use any extra space to write an answer, indicate so in the original answer field.
- A summary of all questions is provided on the back page.

Good luck!

Finite Volume Method

3p **1a** Which of the following statements correctly distinguishes the Finite **Difference** Method (FDM) from the Finite **Volume** Method (FVM)?

(a) FDM is more suitable for solving conservation laws, whereas FVM is better for problems involving high-order derivatives.

(b) FDM is based on approximating differential equations at discrete points, while FVM enforces conservation laws over control volumes.

(c) In FDM, fluxes are explicitly conserved over a control volume, while in FVM, derivatives are computed using Taylor series expansions.

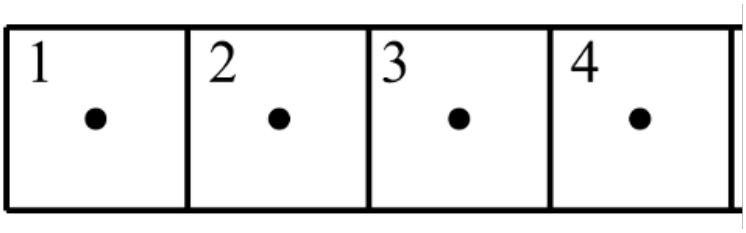
(d) FVM requires a structured grid, while FDM can be applied to both structured and unstructured grids.

(e) In FVM, integration over control volumes leads to the direct calculation of pointwise derivatives, similar to FDM.

For a purely advective problem in 1D:

$$\frac{\partial \phi}{\partial t} + c\frac{\partial \phi}{\partial x} = 0$$

After schematizing the problem with the following mesh:



And applying the Finite Volume Method, we obtain:

$$\frac{\partial \phi}{\partial t} = \frac{c(\phi_e - \phi_w)}{\Delta x}$$

2p **1b** What is the physical interpretation of the variable c, in one sentence?

4p **1c** The values of ϕ at the centre of the cell/volume are known at time $t = 0$ and vary across the mesh.

- Apply the **Forward Euler** method to discretize the time derivative.
- Approximate the surface values ϕ_e andϕ_w using **central differences**.

Write the discrete expression for ϕ_2^1, which represents the value at cell 2 after one time step.

(Here, the subscript refers to the cell number and the superscript to the time step.)

3p **1d** Consider a case where the mesh is very large (order of 1000000000 cells) and the propagation of the quantity ϕ is in the order of 10-100 cells. What will occur with this solution after many time steps? Limit your answer to 2-3 sentences.

Finite Element Method

Consider the homogeneous steady-state advection/diffusion equation in 1D with given velocity $v\,(x)$ and diffusion coefficient k:

$$v\frac{\partial u}{\partial x} - \frac{\partial}{\partial x}\left(k\frac{\partial u}{\partial x}\right) = 0$$

To arrive at the weak form, integration by parts is not applied on the advection term. What is then the form of the stiffness matrix contributions...

3p **2a** With v?

(a) $\int_\Omega \mathbf{N}^T v \mathbf{N} d\Omega$

(b) $\int_\Omega \mathbf{N}^T v \mathbf{B} d\Omega$

(c) $\int_\Omega \mathbf{B}^T v \mathbf{N} d\Omega$

(d) $\int_\Omega \mathbf{B}^T v \mathbf{B} d\Omega$

3p **2b** With k?

(a) $\int_\Omega \mathbf{N}^T k \mathbf{N} d\Omega$

(b) $\int_\Omega \mathbf{N}^T k \mathbf{B} d\Omega$

(c) $\int_\Omega \mathbf{B}^T k \mathbf{N} d\Omega$

(d) $\int_\Omega \mathbf{B}^T k \mathbf{B} d\Omega$

2p **2c** For a single 2-node element with the first node at $x = 0$ and the second node at $x = 10$, give a mathematical expression for $\mathbf{N}$.

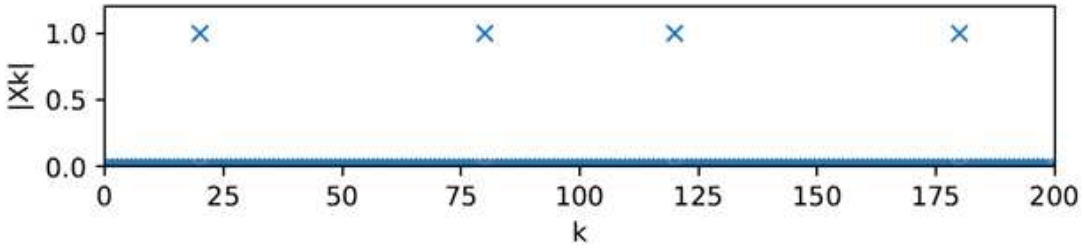
2p **2d** For the same element, give a mathematical expression for B.

Signal Processing

We start from a signal which is the sum of two cosines, both with unit amplitude and zero phase, one with a frequency of 10 Hz, and the other with 60 Hz. The signal is sampled at $f_s = 100$ Hz, for a duration of $T = 2$ seconds.

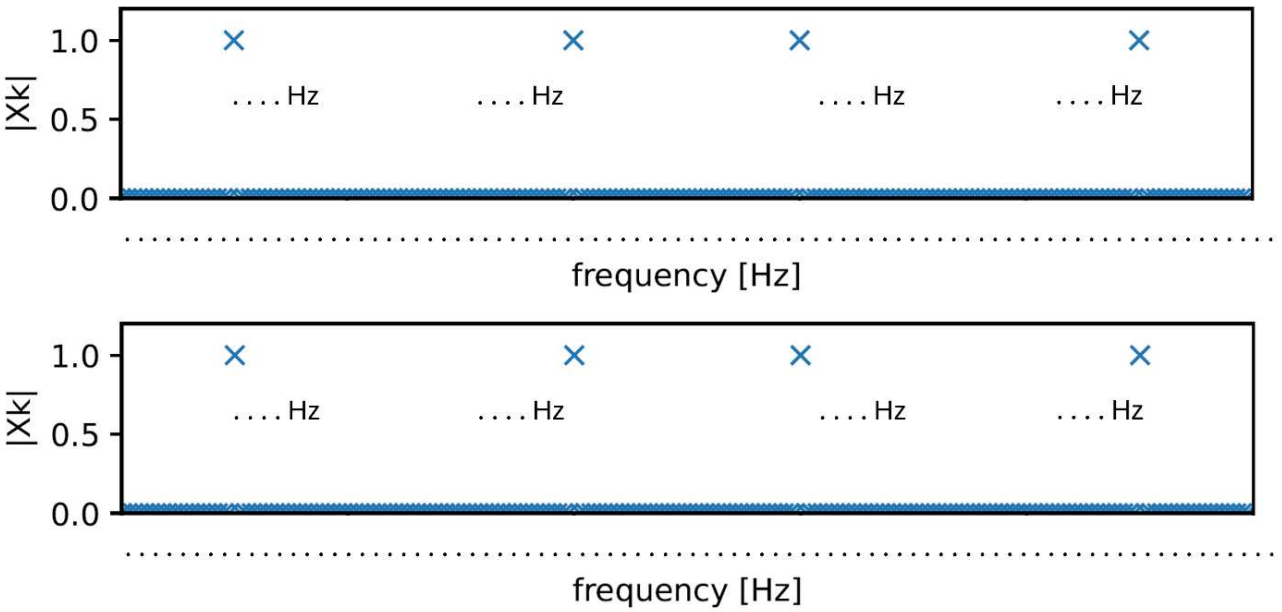
The discrete time samples are input to the Discrete Fourier Transform (DFT) and we directly plot the magnitude (modulus) of the output, hence $|Xk|$, multiplied by $\Delta t = 1/f_s$, with $k = 0, \ldots, N-1$. *(i.e. Xk=np.fft.fft(xt), and plt.plot(abs(Xk)/fs) in code),* using crosses as markers.

The resulting graph is shown below:



5p **3** Provide an adequate and correct labelling of the horizontal axis, in terms of frequency expressed in [Hz], such as the minimum and maximum bounds, **and** annotate the four peaks.

Please note that we have provided two images for you to use in case you need to start over. **Please cross over the image you do NOT want to submit.** If it is unclear which image you want to submit, we will only grade the bottom one.



Please fill in the missing values.

Time Series Analysis

The magnitude of the vertical stress in a rock mass has been observed at times $t = 0, 1, \ldots, 4$ months and is assumed to follow a linear trend; after detrending the data, the following time series is obtained:

$$S := \hat{\epsilon} = \mathrm{y} - \mathrm{A}\hat{\mathrm{x}} = \begin{bmatrix} 1.8 \\ 10.4 \\ -1.1 \\ 5.4 \\ -8.7 \end{bmatrix} MPa$$

The estimated linear trend parameters are $\hat{\mathrm{x}} = [\hat{x_0}, \hat{v}]^T = [3.0, 0.8]^T$ ($\hat{x_0}$ [MPa] is the estimated intercept, $\hat{v}$ the estimated slope [MPa/month]).

The stochastic process underlying time series S is modelled as a **zero-mean AR(1) process**.

Below you can find important formulas that you may need to use.

Time Series Analysis formulas:

AR(p) process:

$$S_t = \sum_{i=1}^{p} \phi_i S_{\{t-i\}} + e_t$$

Autocovariance function for AR(1) process:

$$c_{\{\tau\}} = \sigma^2 \phi_1^\tau$$

Normalized autocovariance function:

$$\rho_\tau = c_\tau / c_0$$

Least-squares estimate of autocovariance:

$$\hat{c}_\tau = \frac{1}{m-\tau} \sum_{t=1}^{m-\tau} (S_t - \mu)(S_{t+\tau} - \mu) \text{ with}$$

$$\mu = \mathbb{E}(S)$$

Prediction formula:

$$\hat{\mathrm{y}}_p = \mathrm{A}_p \hat{\mathrm{x}} + \hat{\epsilon}_p$$

4p **4a** Estimate the normalized autocovariance function ρ_1 for the AR(1) process based on the stationary time series S.

4p **4b** Make a plot (no calculation needed) of the ACF for the AR(1) process (if you did not find a solution for question a, you may use $\hat{\rho}_1 = -0.1$).

4p **4c** What is the predicted value of the vertical stress at $t = 5$ months?

Optimization

7p **5a** Solve the following mathematical programming problem using the graphical solution method. Do not forget to state clearly the solution and the value of the objective function in your answer, as well as how you obtain such a solution.

1. Minimize:

$$Z = x_A + x_B$$

2. Subject to:

$$2x_A + x_B \leq 16 \quad \text{(Constraint 1)}$$

$$x_A + 3x_B \geq 12 \quad \text{(Constraint 2)}$$

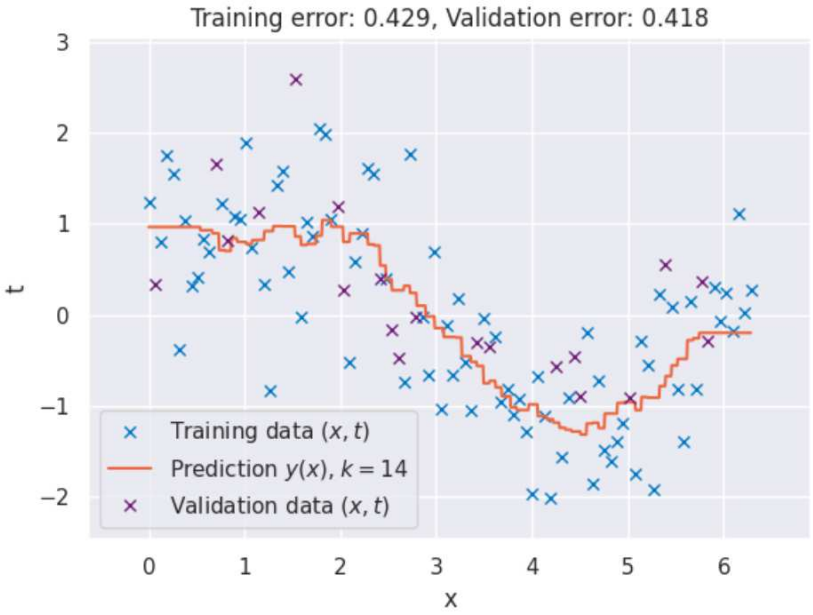
$$x_A, x_B \geq 0 \quad \text{(Non-negativity)}$$

3p **5b** If eventually you were told that the variables must be integer, what would the solution to the same problem be if the objective function became: $\max(Z) = x_A$?

Please mark this in the figure you made in **part a**. In addition to marking this on your figure, use the textbox below to explain your reasoning briefly.

Machine Learning

Consider the following kNN model trained on 80 noisy observations, with 20 observations left out for validation:



3p **6a** Regarding decision theory and model selection for kNN models, which **ONE** of the following statements is true?

(a) Decision theory dictates that our regression function should be $y(x) = \int tp(t|x)dt$. In kNN we have direct access to $p(t|x)$ but cannot integrate it exactly, so we opt to define $y(x)$ as the average of all data points located at position x

(b) For datasets containing only a single observation at each location x, kNN shows extreme overfitting for $k = 1$, with the training loss being exactly zero. By computing the loss on a validation set, we can have a more realistic idea about the performance of the model and adjust k accordingly

(c) In order to avoid overfitting when using kNN models, we use only the training set to obtain k and average over only validation samples when computing $y(x)$

(d) In kNN, k can be seen as a hyperparameter, since it controls model complexity: as we increase k we decrease model complexity, as the neighbourhood we use to compute $y(x)$ shrinks in size and, therefore, contains less and less observation noise.

3p **6b** Regarding the specific kNN model above, its losses $L_{\text{train}} = 0.429$ and $L_{vali} = 0.418$ and considering that $k = 14$ **leads to the best possible model for this dataset**, which **ONE** of the following statements is true

(a) Starting from $k = 14$, further increasing k will lead to models with higher variance (and therefore lower bias), leading to a reduction of L_{train}

(b) Regardless of the value of k, neither L_{train} nor L_{vali} can be reduced below the irreducible loss $L_{\text{noise}} = \int \int \left(\mathbb{E}\left[t\,|x|\right] - t\right)^2 p\left(x, t\right) \mathrm{d}x\mathrm{d}t$

(c) Moving away from $k = 14$ in any direction should makeL_{vali} increase. For $k < 14$ we expect L_{train} to decrease, while for $k > 14$ we expect L_{train} to increase.

(d) From the situation above, we could further regularize the kNN model by adding an L_2 regularization term of the form $\frac{\lambda}{2}\mathbf{w}^T\mathbf{w}$ to L_{train}, where $\mathbf{w}$ are the weights of the kNN model.

3p **6c** Imagine you are working on your thesis and come across a regression problem you would like to solve with a **feedforward neural network.** Regarding dataset preparation, training and model selection, which **ONE** of the following modeling steps should be **AVOIDED**?

(a) Split the dataset into 70% for training, 20% for validation and 10% for testing.

(b) Normalize only the training dataset first and then use the same resulting normalization coefficient for the other two datasets

(c) Monitor the validation loss during training and stop when its value does not decrease for a number of epochs (*early stopping*)

(d) Shuffle the full dataset before splitting, in order to remove unintended biases during data acquisition and cleaning

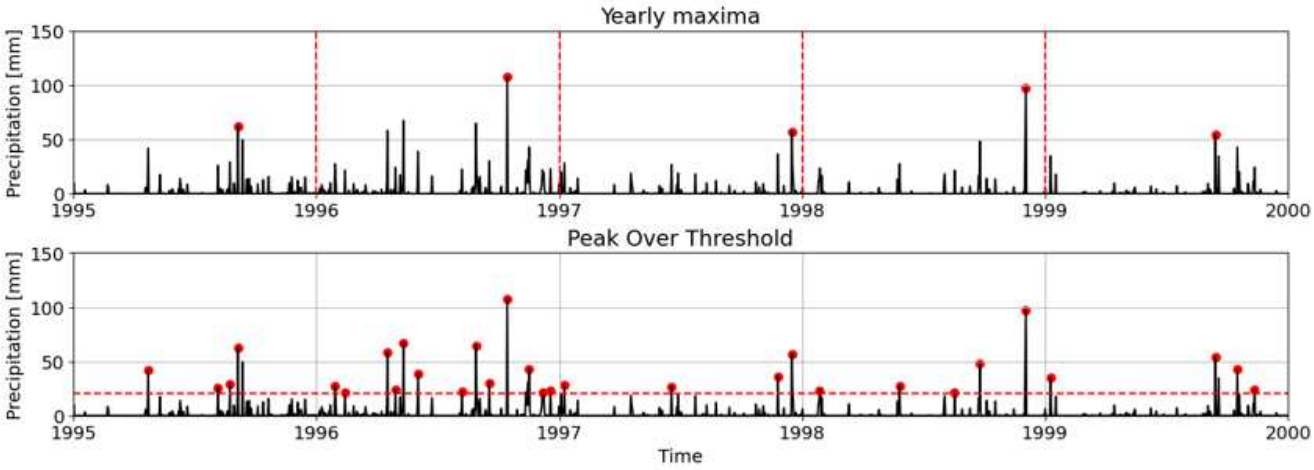
(e) Pick a set of network architectures (layer sizes, number of neurons, activation functions), train them all separately and then pick the one with the lowest error on the test dataset

(f) Include an L_2 regularization term to the loss function in order to minimize the risk of overfitting and calibrate λ using the validation dataset

Extreme Value Analysis

Your team is investigating the efficiency and performance of drainage systems in cities along the South of Europe. Drainage systems collect rainfall water during extreme rainfall events to help control the flooding in the city.

One of your colleagues is analyzing the extreme precipitations used to design the drainage system in Barcelona (Spain). They obtained a timeseries of 5 years of precipitation and have performed both Yearly Maxima and Peak Over Threshold to select the extremes in the timeseries. They want your advice on which method to use for further analysis. Consider **ONLY** the two approaches shown in the figure below.



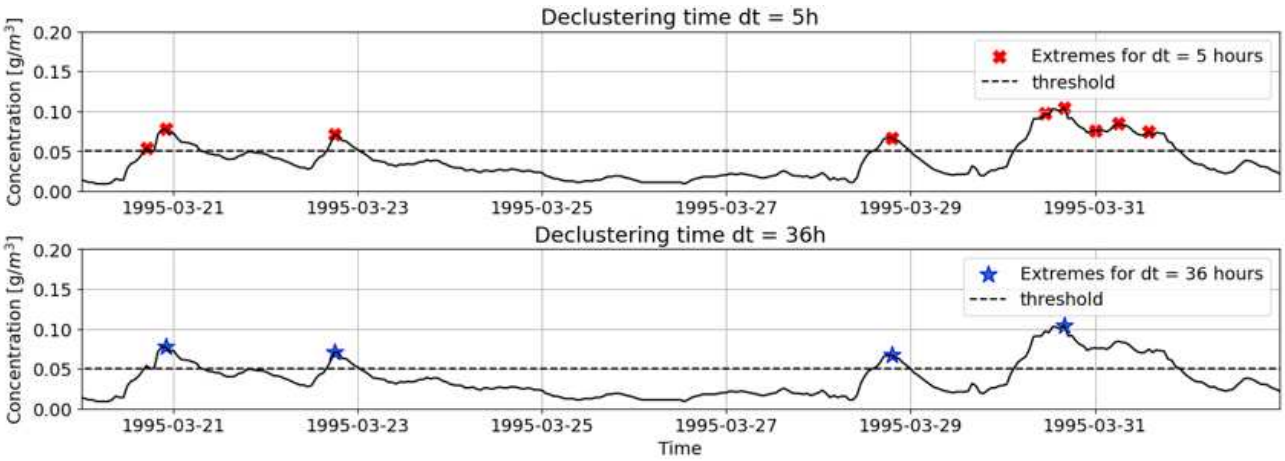
1p **7a** Which option would you recommend to your colleague?

(a) Yearly Maxima

(b) Peak Over Threshold

2p **7b** Justify your answer briefly.

Another colleague is studying the water quality of the rainfall water that is received by the drainage system during extreme events. During extreme rainfall events, the water collected by the drainage system “cleans” the streets. They are working with a timeseries of the concentration of nitrogen in the water. They have applied Peak Over Threshold using a threshold value of $0.05 g/m^3$. However, they are not sure about the declustering time (dt) they should use. In the figure below, you can see a small part of the timeseries and the extremes extracted using two declustering times, $dt = 5h$ and $dt = 36$.



1p **7c** Which dt would you recommend to your colleague to use?

(a) dt = 5h

(b) dt = 36h

2p **7d** Justify your answer briefly.

6p **7e** A third colleague asked you for your help as probability expert. They are studying the extreme discharges in a drainage system. They have sampled 60 extremes using Peak Over Threshold with a threshold 50 l/s on 6 years of observations of discharges (l/s) and have fitted a Generalized Pareto distribution (GPD) so $Q \sim GPD\,(\sigma = 20, \xi = 0.25)$.
Compute the discharge associated with a return period of 10 years. The GPD probability density function is given by

$$h(x) = \begin{cases} \left(1 + \frac{\xi x}{\sigma_{th}}\right)^{-\frac{\xi+1}{\xi}} & \text{for } \xi \neq 0 \\ \exp\left(-\frac{x}{\sigma_{th}}\right) & \text{for } \xi = 0 \end{cases}$$

And the GPD cumulative distribution function is given by

$$H(x) = \begin{cases} 1 - \left(1 + \frac{\xi x}{\sigma_{th}}\right)^{-\frac{1}{\xi}} & \text{for } \xi \neq 0 \\ 1 - \exp\left(-\frac{x}{\sigma_{th}}\right) & \text{for } \xi = 0 \end{cases}$$

Risk and Reliability
Your are asked to perform a risk analysis of a chemical spill into a groundwater aquifer which can occur in two ways:

A) 'A small spill'
B) 'A big spill'

Spill A has a 10% probability of occuring each year, which would cause 10 fatalities.
Spill B has a 1% chance of ocurring, which would cause 100 fatalities.

Assume that the regulating authority is risk averse ($\alpha = 2$) and that the tolerable risk limit is governed using a constant of $C = 1$.
Recall that the risk limit line can formulated as $1 - F_N(n) = C^\alpha$

2p **8a** Compute the expected fatalities caused by the system

2p **8b** If the chemical industry improved the safety of its systems, what would be the **maximum allowable number of fatalities** that would be considered acceptable due to a small spill?

2p **8c** What is the **maximum value for the probability** of a big spill that would be considered acceptable by the regulating authority?

4p **8d** Create an FN curve for this situation, including the limit line. Be sure to clearly label the axes and the values of key points on your diagram.

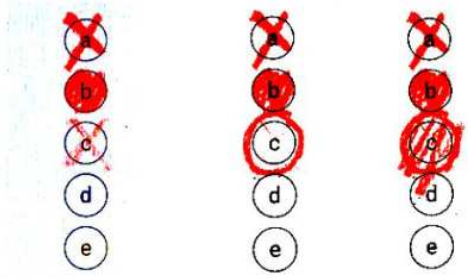
Exam overview

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

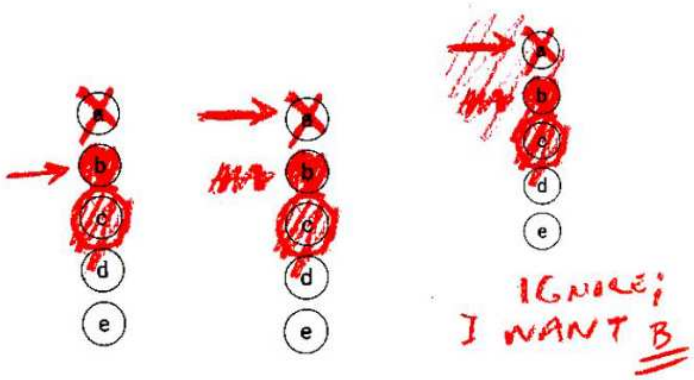
No.	Topic	Number of Sub-parts	Points
-	-	-	-
1	Finite Volume Method	4	12
2	Finite Element Method	4	10
3	Signal Processing	1	5
4	Time-series Analysis	3	12
5	Optimization	2	10
6	Machine Learning	3	9
7	Extreme Value Analysis	5	12
8	Risk & Reliability	4	10
Total			**80**

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:



Examples of CLEAR multiple choice response:



Answer: B Answer: A Answer: B

Multiple choice examples.