- For every question, start your answer **on a new page**.
- Do not hand in your scratch paper. Please do hand in this exam sheet, or leave it behind on your table.
- You are not allowed to use books nor printed or handwritten formula sheets. See the final page for supplied formulas.
- **For 5MB20, SKIP question 3!! For 5SSB0, SKIP question 2!!**

---

1. For each of the following sub-questions, provide a *short but essential* answer.

a. (2 points). The joint distribution for feature vector $\mathbf{x}$ and kth class $\mathcal{C}_k$ is given by

$$p(\mathbf{x}, \mathcal{C}_k|\boldsymbol{\theta}) = \pi_k \cdot \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$$

Write down an expression for the posterior class probability $p(\mathcal{C}_k|x)$ (No derivations are needed, just a proper expression)?

$$p(\mathcal{C}_k|x) = \frac{\pi_k \cdot \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma})}{\sum_j \pi_j \cdot \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma})}$$

b. (1 point). Why does maximum likelihood estimation become a better approximation to Bayesian learning as you collect more data?

The likelihood tends to get 'narrower' and more informative than the prior, since the latter does not change with more data.

c. (2 points). Given is a model

$$p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|\mathbf{W}\mathbf{z}, \Psi)$$
$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|0, I)$$

Work out an expression for the marginal distribution $p(\mathbf{x})$.

$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|0, \mathbf{W}\mathbf{W}^T + \Psi)$

d. (1 point). Why is (probabilistic) principal component analysis more popular than factor analysis in the signal processing community?

In contrast to FA, PCA assumes the same variance for all input components, which is appropriate if the input components are shifted versions of each other. This occurs in the processing of time sequences.

e. (2 points). Which of the following statements are justified? Just pick the correct statements; no explanation needed.
1: Discriminative classification is more similar to regression than to density estimation.
2: Density estimation is more similar to generative classification than to discriminative classification.
3: A hidden Markov model is more similar to factor-analysis-over-time than to a Gaussian-mixture-model-over-time.
4: Clustering is more similar to supervised classification than to unsupervised classification.

1 and 2 are correct.

f. (2 points). Explain shortly how Bayes rule relates to machine learning in the context of an observed data set D and a model M with parameters θ. Your answer must contain the expression for Bayes rule.

> Machine learning is inference over models (hypotheses, parameters, etc.) from a given data set. Bayes rule makes this statement precise. Let $\theta \in \Theta$ and D represent a model parameter vector and the given data set, respectively. Then, Bayes rule,
>
> $$p(\theta|D, M) = \frac{p(D|\theta, M)}{p(D|M)} p(\theta|M)$$
>
> relates the information that we have about θ before we saw the data (i.e., the distribution $p(\theta)$) to what we know after having seen the data, $p(\theta|D)$.

2. Skip this question if you make 5SSB0!
Consider an IID data set $D = \{(x_1, y_1), \ldots, (x_N, y_N)\}$. We will model this data set by a model $y_n = \theta^T f(x_n) + e_n$, where $f(x_n)$ is an M-dimensional feature vector of input x_n; y_n is a scalar output and $e_n \sim \mathcal{N}(0, \sigma^2)$. (Note the list of formula's at the final page of this exam).

a. (1 point). Rewrite the model in matrix form by lumping input features in a matrix $F = [f(x_1), \ldots, f(x_N)]^T$, outputs and noise in the vectors $y = [y_1, \ldots, y_N]^T$ and $e = [e_1, \ldots, e_N]^T$, respectively.

> $y = F\theta + e$

b. (2 points). Now derive an expression for the log-likelihood $\log p(D|\theta, \sigma^2)$.

> $$\begin{aligned} \log p(D|\theta, \sigma^2) &= \log \mathcal{N}(y|\, F\theta, \sigma^2) \\ &\propto -\frac{1}{2\sigma^2} (y - F\theta)^T (y - F\theta) \end{aligned}$$

c. (2 points). Proof that the maximum likelihood parameter estimate is given by

$$\hat{\theta}_{ml} = (F^T F)^{-1} F^T y \,.$$

> Taking the derivative to θ
>
> $$\nabla_\theta \log p(D|\theta) = \frac{1}{\sigma^2} F^T (y - F\theta)$$
>
> Set derivative to zero for maximum likelihood estimate
>
> $$\hat{\theta} = (F^T F)^{-1} F^T y$$

d. (1 point). Consider a classification problem with Gaussian class-conditional sampling distributions $p(x_n|\mathcal{C}_1)$ and $p(x_n|\mathcal{C}_2)$ for the feature observations x_n and class priors $p(\mathcal{C}_1) = \pi$, $p(\mathcal{C}_2) = 1 - \pi$. Under what condition (for the Gaussian class conditional distributions) is the discrimination boundary between the two classes a hyperplane?

> Same variance: $p(x_n|\mathcal{C}_1) = \mathcal{N}(\mu_1, \Sigma)$ and $p(x_n|\mathcal{C}_2) = \mathcal{N}(\mu_2, \Sigma)$

e. (2 points). We model a given set of observations $D = \{x_1, x_2, \ldots, x_n\}$ by a Gaussian Mixture model

$$p(x_n) = \sum_k \pi_k \mathcal{N}(x_n|\, \mu_k, \Sigma_k) \,.$$

(1) Derive an expression for the log-likelihood $p(D|\theta)$ where θ relates to the parameter set in the model.
(2) Explain why it is generally considered easier to find (maximum likelihood) parameter estimates through the EM algorithm than by gradient-based methods. (It is easiest to explain this by considering your log-likelihood expression).

1.
$$p(D|\theta) = \sum_n \log \left(\sum_k \pi_k \mathcal{N}(x_n|\, \mu_k, \Sigma_k) \right)$$
2. The gradients of the log-likelihood w.r.t. the parameters are coupled.

f. (2 points). Consider a set of IID observations $D = \{x_1, \ldots, x_N\}$ and proposed model

$$\begin{aligned} x_n &= W z_n + e_n \\ z_n &\sim \mathcal{N}(0, I) \\ e_n &\sim \mathcal{N}(0, \Psi)\,. \end{aligned}$$

Rewrite the model in terms of $p(x_n|z_n, W, \Psi)$ and $p(z_n)$

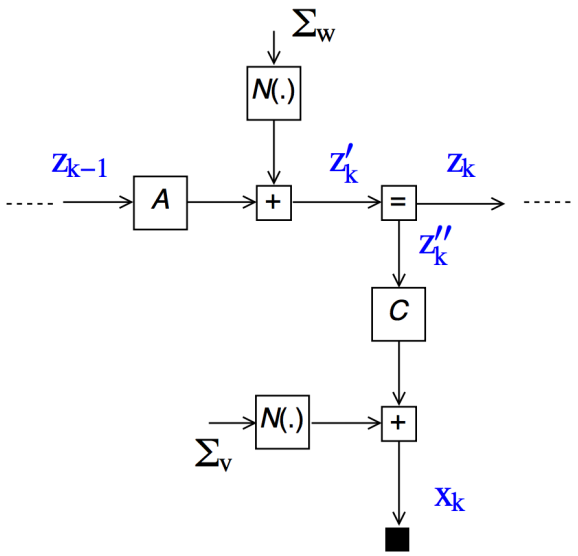
$$\begin{aligned} p(x_n|z_n, W, \Psi) &= \mathcal{N}(x_n|Wz_n, \Psi) \\ p(z_n) &= \mathcal{N}(z_n|0, I) \end{aligned}$$

3. Skip this question if you make 5MB20!

Consider the following state-space model:

$$z_k = A z_{k-1} + w_k \tag{1a}$$
$$x_k = C z_k + v_k \tag{1b}$$
$$w_k \sim \mathcal{N}(0, \Sigma_w) \tag{1c}$$
$$v_k \sim \mathcal{N}(0, \Sigma_v) \tag{1d}$$
$$z_0 \sim \mathcal{N}(0, \Sigma_0) \tag{1e}$$

where $k = 1, 2, \ldots, n$ is the time step counter; z_k is an *unobserved* state sequence; x_k is an *observed* sequence; w_k and v_k are (unobserved) state and observation noise sequences respectively; A, C, Σ_v,Σ_w and Σ_0 are known parameters. The Forney-style factor graph (FFG) for one time step is depicted here:



a. (2 points). Rewrite the state-space equations as a set of conditional probability distributions:

$$p(z_k|z_{k-1}, A, \Sigma_w) = \dots$$
$$p(x_k|z_k, C, \Sigma_v) = \dots$$
$$p(z_0|\Sigma_0) = \dots$$

$$p(z_k|z_{k-1}, A, \Sigma_w) = \mathcal{N}(z_k|Az_{k-1}, \Sigma_w)$$
$$p(x_k|z_k, C, \Sigma_v) = \mathcal{N}(x_k|Cz_k, \Sigma_v)$$
$$p(z_0|\Sigma_0) = \mathcal{N}(z_0|0, \Sigma_0)$$

b. (1 point). Define $z^n \triangleq (z_0, z_1, \dots, z_n)$, $x^n \triangleq (x_1, \dots, x_n)$ and $\theta = \{A, C, \Sigma_w, \Sigma_v\}$. Now write out the generative model $p(x^n, z^n|\theta)$ as a product of factors.

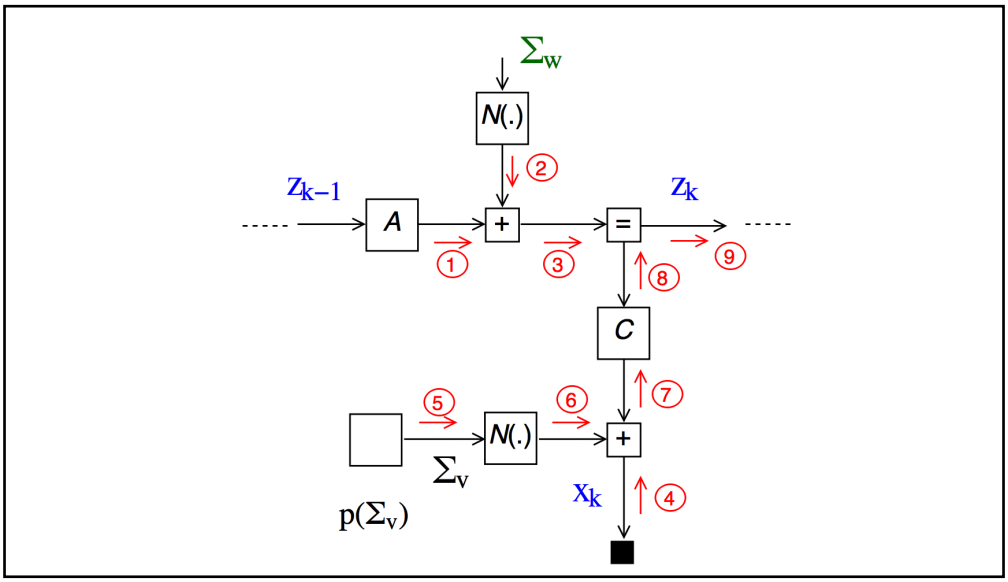
$$\begin{aligned} p(x^n, z^n|\theta) &= p(z_0|\Sigma_0) \prod_{k=1}^{n} p(x_k|z_k, C, \Sigma_v)\, p(z_k|z_{k-1}, A, \Sigma_w) \\ &= \mathcal{N}(z_0|0, \Sigma_0) \prod_{k=1}^{n} \mathcal{N}(x_k|Cz_k, \Sigma_v)\, \mathcal{N}(z_k|Az_{k-1}, \Sigma_w) \end{aligned}$$

c. (1 point). We are interested in estimating z_k from a given estimate for z_{k-1} and the current observation x_k, i.e., we are interested in computing $p(z_k|z_{k-1}, x_k, \theta)$. Can $p(z_k|z_{k-1}, x_k, \theta)$ be expressed as a Gaussian distribution? Explain why or why not in one sentence.

Yes, since the generative model $p(x^n, z^n|\theta)$ is (one big) Gaussian.

d. (2 points). Copy the graph onto your exam paper and draw the message passing schedule for computing $p(z_k|z_{k-1}, x_k, \theta)$ by drawing arrows in the factor graph. Indicate the order of the messages by assigning numbers to the arrows (① for the first message; ② for the second message and so on).

e. (1 point). Now assume that our belief about parameter Σ_v is instead given by a distribution $p(\Sigma_v)$ (rather than a known value). Adapt the factor graph drawing of the previous answer to reflects our belief about Σ_v.



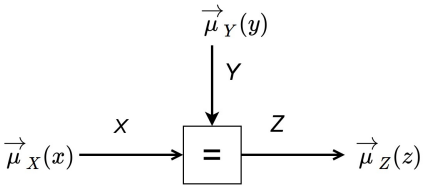
See drawing in previous answer.

f. (1 point). The FFG contains an equality node with edges z_k'', z_k' and z_k. Write down the equation for the factor in the equality node:

$$f(z_k'', z_k', z_k) = \ldots$$

$$f(z_k'', z_k', z_k) = \delta(z_k - z_k')(z_k - z_k'')$$

g. (2 points). Consider a general equality node with edges X, Y and Z as given in the following graph:



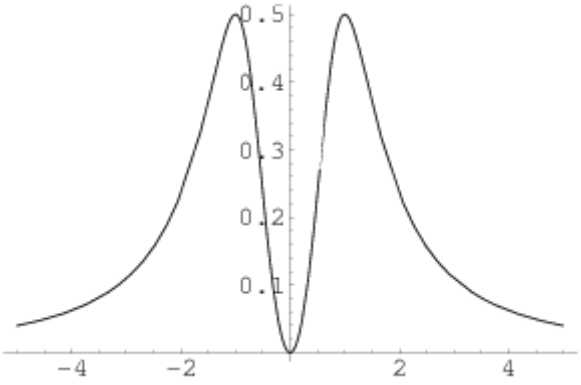
Using the sum-product rule, proof that

$$\overrightarrow{\mu}_Z(z) = \overrightarrow{\mu}_X(z)\overrightarrow{\mu}_Y(z)$$

$$\begin{aligned}\overrightarrow{\mu}_Z(z) &= \int \overrightarrow{\mu}_X(x)\overrightarrow{\mu}_Y(y)\delta(z-x)(z-y)\mathrm{d}x\mathrm{d}y \quad \text{(sum-product rule)} \\ &= \overrightarrow{\mu}_X(z)\int \overrightarrow{\mu}_Y(y)\delta(z-y)\mathrm{d}y \\ &= \overrightarrow{\mu}_X(z)\overrightarrow{\mu}_Y(z)\end{aligned}$$

4. The function F is defined on the real numbers and

$$f(x) = \frac{x^2}{x^4+1}.$$



The function $f(x)$

The partial derivatives are given as

$$f'(x) = -\frac{2x(x^4-1)}{(x^4+1)^2},$$
$$f''(x) = \frac{6x^8 - 24x^4 + 2}{(x^4+1)^3}.$$

a. Approximate

$$Z_f = \int_{-\infty}^{\infty} f(x)\,dx,$$

using the Laplace approximation.
Give the detailed derivation, not just the answer.

From the graph and $f'(x)$ we see that f has a maximum at $x = 1$ (and the same maximum at $x = -1$).

$$\begin{aligned}\frac{\partial^2}{\partial x^2}\ln f(x) &= \frac{f(x)f''(x) - (f'(x))^2}{f^2(x)} \\ &= \frac{(6x^8 - 24x^4 + 2)x^2 - (2x(x^4-1))^2}{x^4(x^4+1)^2} \\ &= \frac{6x^8 - 24x^4 + 2 - 4(x^4-1))^2}{x^2(x^4+1)^2} \\ A &= -\frac{\partial^2}{\partial x^2}\ln f(x)\Big|_{x=1} = 4.\end{aligned}$$

So, the approximation is

$$g(x) = f(1)e^{-\frac{1}{2}A(x-1)^2} = \frac{1}{2}e^{-2(x-1)^2}.$$

The integral of $g(x)$ is

$$\int_{-\infty}^{\infty} g(x)\,dx = f(1)\sqrt{\frac{2\pi}{A}} = \frac{1}{2}\sqrt{\frac{\pi}{2}}.$$

The *Bayesian Information Criterion* is results in

$$\underbrace{\log\frac{p(\mathcal{M}_1|x^N)}{p(\mathcal{M}_2|x^N)}}_{(*1)} \approx \underbrace{\log\frac{p(\mathcal{M}_1)}{p(\mathcal{M}_2)}}_{(*2)} + \underbrace{\log\frac{p(x^N|\mathcal{M}_1,\hat{\underline{\theta}}_1)}{p(x^N|\mathcal{M}_2,\hat{\underline{\theta}}_2)}}_{(*3)} + \underbrace{\frac{1}{2}(k_1-k_2)\log N}_{(*4)}.$$

Here x^N is a binary data sequence of length N, k_1 and k_2 are the number of free parameters in respectively model $\mathcal{M}_1$ and $\mathcal{M}_2$, and $\underline{\hat{\theta}}_1$ and $\underline{\hat{\theta}}_2$ are the estimated (ML) parameter vectors.

b. Explain the four terms marked by (*1), (*2), (*3), and (*4).

(*1) This ratio of model posteriors (given the data) allows us to select the most appropriate model of the two options.

(*2) This ratio shows our initial preference of the first model relative to the second one.

(*3) This is the log-likelihood ratio of the two models after observing the data.

(*4) This is the correction term needed to compare models of different complexity.

c. The binary data $x^N = x_1, x_2, \ldots, x_N$ is generated by a Bernoulli process, i.e.

$$p(x^N|\mathcal{M}, \theta) = (1-\theta)^{n(0|x^N)}\theta^{n(1|x^N)}.$$

The parameter prior $p(\theta|\mathcal{M})$ is given by the Beta distribution:

$$p(\theta|\mathcal{M}) = \frac{1}{\pi}\frac{1}{\sqrt{\theta(1-\theta)}}.$$

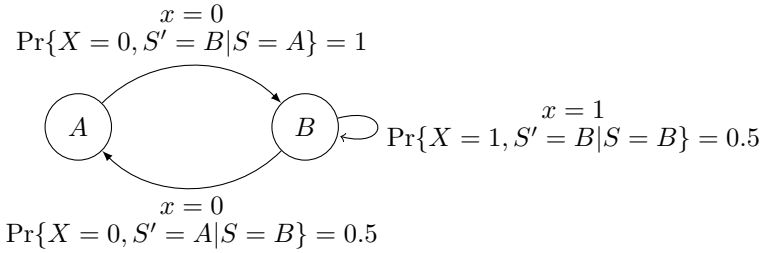
Let $N = 10$ and $x^{10} = 1001101101$.
Determine $p(x^N|\mathcal{M})$.
Give the complete derivation starting with the information given above.

$$\begin{aligned}
p(x^N|\mathcal{M}) &= \int_0^1 p(\theta|\mathcal{M})p(x^N|\mathcal{M},\theta)\,d\theta, \\
&= \int_0^1 \frac{1}{\pi}\frac{1}{\sqrt{\theta(1-\theta)}}(1-\theta)^4\theta^6\,d\theta, \\
&= \frac{\Gamma(4+\frac{1}{2})\Gamma(6+\frac{1}{2})}{\pi\Gamma(11)}, \\
&= \frac{\frac{1}{2}\frac{3}{2}\frac{5}{2}\frac{7}{2}\cdot\frac{1}{2}\frac{3}{2}\frac{5}{2}\frac{7}{2}\frac{9}{2}\frac{11}{2}}{10!}, \\
&= \frac{77}{262144} = 0.0002937.
\end{aligned}$$

5. Consider the following binary finite state model (Markov source). This model produces outputs X_t where the probability of the next output symbol depends on the current state of the source. We list all non-zero probabilities.

$$\begin{aligned}
&\Pr\{X_t = 0, S_{t+1} = B|S_t = A\} = 1, \\
&\Pr\{X_t = 0, S_{t+1} = A|S_t = B\} = 0.5, \\
&\Pr\{X_t = 1, S_{t+1} = B|S_t = B\} = 0.5,
\end{aligned}$$

The following figure depicts this model.



a. Compute the stationary probabilities $q(A)$ and $q(B)$ where

$$q(s) = \lim_{t\to\infty} \Pr\{S_t = s\} \qquad \text{for } s \in \{A, B\}.$$

$$q(A) = \frac{1}{2}q(B),$$
$$q(B) = 1 - q(A)$$

This results in $q(A) = \frac{1}{3}$ and $q(B) = \frac{2}{3}$.

b. Compute the following probabilities assuming that the model is stationary (i.e. $\Pr\{S_1 = A\} = q(A)$ and $\Pr\{S_1 = B\} = q(B)$).

$$\Pr\{X_1 = 1\}$$
$$\Pr\{X_2 = 1|X_1 = 0\}$$

$$\Pr\{X_1 = 1\} = q(A) \cdot 0 + q(B) \cdot \frac{1}{2} = \frac{1}{3}.$$

For the next one we need

$$\Pr\{X^2 = 01\} = q(A) \cdot 1 \cdot \frac{1}{2} + q(B) \cdot \frac{1}{2} \cdot 0 = \frac{1}{6},$$

and we find

$$\Pr\{X_2 = 1|X_1 = 0\} = \frac{\Pr\{X^2 = 01\}}{\Pr\{X_1 = 0\}} = \frac{1/6}{2/3} = \frac{1}{4}.$$

c. Let $\mathcal{M}_0$ be the i.i.d. model with

$$\underline{\theta}_0 = (\Pr\{X_1 = 1\}).$$

Also $\mathcal{M}_1$ is the first order model with

$$\underline{\theta}_1 = (\theta_{10}, \theta_{11}) = (\Pr\{X_2 = 1|X_1 = 0\}, \Pr\{X_2 = 1|X_1 = 1\}).$$

And $\mathcal{M}_2$ is the second order model with

$$\begin{aligned}\underline{\theta}_2 &= (\theta_{200}, \theta_{201}, \theta_{210}, \theta_{211}) \\ &= (\Pr\{X_3 = 1|X_1 = 0, X_2 = 0\}, \Pr\{X_3 = 1|X_1 = 0, X_2 = 1\}, \\ &\qquad \Pr\{X_3 = 1|X_1 = 1, X_2 = 0\}, \Pr\{X_3 = 1|X_1 = 1, X_2 = 1\}).\end{aligned}$$

The Markov model produces a 'typical' sequence so

$$-\log_2 \Pr\{X^n = x^n|\mathcal{M}_i, \underline{\theta}_i\} \approx H_i(X^n),$$

where $H_i(X^n)$ is the entropy rate of the i^{th} model.
Given is that

$$H_0(X^n) = 0.9183 \cdot n$$
$$H_1(X^n) = 0.8742 \cdot n$$
$$H_2(X^n) = 0.7925 \cdot n$$

Determine for what range of n you should use $\mathcal{M}_0$. And when $\mathcal{M}_1$ and when $\mathcal{M}_2$? Use the idea of stochastic complexity and **motivate your answer**.

The remaining conditional probabilities should be computed in order to find the entropies.

Then we find the stochastic complexities

$$S.C._0 = \frac{\log_2 n}{2} + 0.9183 \cdot n$$
$$S.C._1 = \frac{2\log_2 n}{2} + 0.8742 \cdot n$$
$$S.C._2 = \frac{4\log_2 n}{2} + 0.7925 \cdot n$$

Equality of $S.C._0$ and $S.C._1$ happens at $n = 69 \sim 70$.
Equality of $S.C._1$ and $S.C._2$ happens at $n = 76 \sim 77$.

So up to $n = 70$ we use $\mathcal{M}_0$, then until $n = 77$ we use $\mathcal{M}_1$ and afterwards we use $\mathcal{M}_2$.

Appendix: formula's

$$
\begin{aligned}
|A^{-1}| &= |A|^{-1} \\
\nabla_A \log|A| &= (A^T)^{-1} = (A^{-1})^T \\
\mathrm{Tr}[ABC] &= \mathrm{Tr}[CAB] = \mathrm{Tr}[BCA] \\
\nabla_A \mathrm{Tr}[AB] &= \nabla_A \mathrm{Tr}[BA] = B^T \\
\nabla_A \mathrm{Tr}[ABA^T] &= A(B + B^T) \\
\nabla_x x^T A x &= (A + A^T)x \\
\nabla_X a^T X b &= \nabla_X \mathrm{Tr}[ba^T X] = ab^T
\end{aligned}
$$

Points that can be scored per question:

Question 1: a) 2 points; b) 1 point; c) 2 points; d) 1 point; e) 2 points; f) 2 points. Total 10 points..

Question 2: a) 1 point; b) 2 points; c) 2 points; d) 1 point; e) 2 points; f) 2 points. Total 10 points.

Question 3: a) 2 points; b) 1 point; c) 1 point; d) 2 points; e) 1 point; f) 1 point; g) 2 points. Total 10 points.

Question 4: a) 5 points; b) 2 points; c) 3 points. Total 10 points.

Question 5: a) 2 points; b) 3 points; c) 5 points. Total 10 points.

You should make either question 2 or 3. Max score that can be obtained: 40 points.
The final grade is obtained by dividing the score by 4 and rounding to the nearest integer.