- 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)?

b. (1 point). Why does maximum likelihood estimation become a better approximation to Bayesian learning as you collect 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 $p(\mathbf{x})$.

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

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.

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.

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.

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

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

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

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?

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).

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

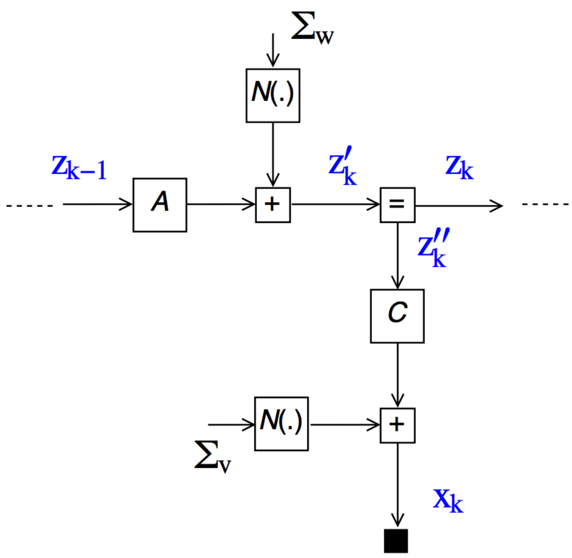
$$\begin{aligned} x_n &= W z_n + e_n \\ z_n &\sim \mathcal{N}(0, I) \\ e_n &\sim \mathcal{N}(0, \Psi)\,. \end{aligned}$$

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

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:

$$
\begin{aligned}
p(z_k|z_{k-1}, A, \Sigma_w) &= \ldots \\
p(x_k|z_k, C, \Sigma_v) &= \ldots \\
p(z_0|\Sigma_0) &= \ldots
\end{aligned}
$$

b. (1 point). Define $z^n \triangleq (z_0, z_1, \ldots, z_n)$, $x^n \triangleq (x_1, \ldots, x_n)$ 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.

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.

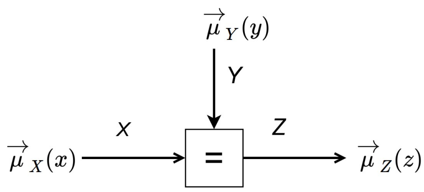
d. (2 points). Copy the graph onto your exam paper and draw the message passing schedule for computing $p(z_k|z_{k-1}, x_k, \theta)$ by drawing arrows in the factor graph. Indicate the order of the messages by assigning numbers to the arrows (1 for the first message; 2 for the second message and so on).

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.

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:

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

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

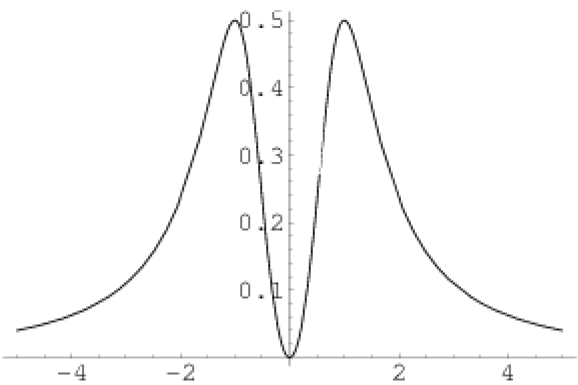


Using the sum-product rule, proof that

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

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.

The *Bayesian Information Criterion* is results in

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

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).

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:

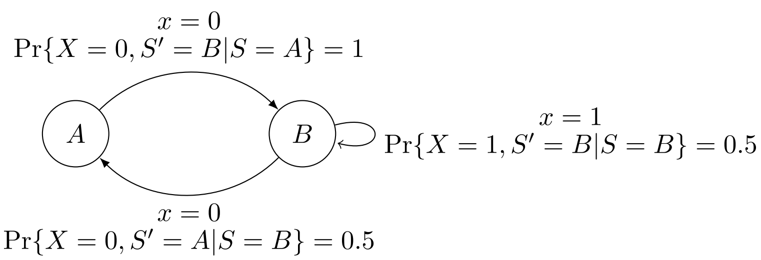
$$p(\theta|\mathcal{M}) = \frac{1}{\pi}\frac{1}{\sqrt{\theta(1-\theta)}}.$$

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

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.

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

The following figure depicts this model.



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

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

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

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

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**.

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.