02471 Machine Learning for Signal Processing

Written examination: December 6, 2023.

Course name: Machine Learning for Signal Processing.

Course number: 02471.

Aids allowed: All DTU allowed aids are permitted.

Exam duration: 4 hours.

Weighting: The weighting is indicated in parentheses for each sub-problem.

This exam has 7 problems with a total of 18 questions, for a total of 100% weighting.

Hand-in: Hand-in on paper and/or upload a PDF file. Do not hand in duplicate information.

All answers must include relevant considerations and/or calculations/derivations.

It should be clear what theories and formulas were used from the curriculum.

Multiple choice: The following problems are multiple choice:

Problem 6.1.

If a multiple choice problem is answered correctly, you are given full points. For a wrong answer, you are subtracted 1/4 of the full point value. E.g, if a problem gives 5 points for a correct answer, a wrong answer will result in subtracting 1.25 points.

Problem 1 Parameter Estimation (15% total weighting)

In this problem we will consider the following data

n	0	1	2	3
y_n	0.5	-1.2	2.0	4.1

Problem 1.1 (5% weighting)

Consider polynomial regression using the squared error as the loss function and assume i.i.d noise. For modeling the response y_n, we will use a polynomial model of the form $f_n = \theta_0 + \theta_1 \cdot n + \theta_2 \cdot n^2$.

Determine the parameters $\hat{\boldsymbol{\theta}} = [\theta_0\ \theta_1\ \theta_2]^T$ of the model that minimizes the error. Describe the process for calculating this estimate, including the relevant matrices and vectors.

Solution: Given the model $f_n = \theta_0 + \theta_1 \cdot n + \theta_2 \cdot n^2$ and the provided data, we can set up the normal equation to find the parameters $\hat{\boldsymbol{\theta}}$.

The design matrix X and response vector $\boldsymbol{y}$ are defined as:

$$X = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} 0.5 \\ -1.2 \\ 2.0 \\ 4.1 \end{bmatrix}$$

According to ML p 73, eq (3.17), the solution is given by

$$\hat{\boldsymbol{\theta}} = (X^T X)^{-1} X^T \boldsymbol{y}$$

Plugging in, we get

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \begin{bmatrix} 4 & 6 & 14 \\ 6 & 14 & 36 \\ 14 & 36 & 98 \end{bmatrix}^{-1} \begin{bmatrix} 5.40 \\ 15.10 \\ 43.70 \end{bmatrix} \\ &= \begin{bmatrix} 0.20 \\ -1.45 \\ 0.95 \end{bmatrix} \end{aligned}$$

Thus, the estimated polynomial regression model is $f_n = 0.2 - 1.45n + 0.95n^2$.

Problem 1.2 (5% weighting)

Using the same data, we consider a linear model on the form $f_n = \theta_0 + \theta_1 \cdot n$, and we assume non-white Gaussian noise, where the variance of the noise is 1, and the covariance of the noise

between two successive samples is 0.2.

Determine the parameters $\hat{\boldsymbol{\theta}} = [\theta_0 \ \theta_1]^T$ of the model that minimizes the squared error using this noise assumption. Write the relevant matrices and vectors used to estimate $\hat{\boldsymbol{\theta}}$.

Solution: From ML pp 101–102, we have the non-white Gaussian case, and have, from eq (3.61)

$$\hat{\boldsymbol{\theta}} = \left(X^T \Sigma_\eta^{-1} X\right)^{-1} X^T \Sigma_\eta^{-1} \boldsymbol{y}$$

As matrices we have

$$X = \begin{bmatrix} 1.0 & 0.0 \\ 1.0 & 1.0 \\ 1.0 & 2.0 \\ 1.0 & 3.0 \end{bmatrix}, \qquad \Sigma_\eta = \begin{bmatrix} 1.0 & 0.2 & 0.0 & 0.0 \\ 0.2 & 1.0 & 0.2 & 0.0 \\ 0.0 & 0.2 & 1.0 & 0.2 \\ 0.0 & 0.0 & 0.2 & 1.0 \end{bmatrix}$$

and we get

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \begin{bmatrix} 3.1 & 4.7 \\ 4.7 & 11.6 \end{bmatrix}^{-1} \begin{bmatrix} 4.5 \\ 12.8 \end{bmatrix} \\ &= \begin{bmatrix} -0.52 \\ 1.31 \end{bmatrix} \end{aligned}$$

Problem 1.3 (5% weighting)

We now consider the case of general parameter estimation. Assume that your parameter is estimated using 10 different datasets (realizations). For each dataset, we have an unbiased estimator, denoted $\hat{\theta}_i$, where the variance of these individual estimators are $\sigma^2 = 1$. We aggregate the estimates using:

$$\hat{\theta}_{\text{agg}} = \frac{1}{10} \sum_{i=1}^{10} \hat{\theta}_i$$

Compute the variance of the estimator $\hat{\theta}_{\text{agg}}$.

Specify formulas, the numerical value, and assumptions made.

Solution: From ML p 81, they write, that if the estimators are mutually uncorrelated, and have the same variance, the new estimator will have a variance of

$$\sigma^2_{\hat{\theta}_{\text{agg}}} = \frac{\sigma^2}{10} = \frac{1}{10} = 0.1$$

Problem 2 Sparse learning (10% total weighting)

In this problem, we consider sparse learning and apply ℓ_1 regularization and linear regression:

$$\arg\min_{\boldsymbol{\theta}} \left\{ \|\boldsymbol{y} - X^T\boldsymbol{\theta}\|^2 + \lambda\|\boldsymbol{\theta}\|_1 \right\}$$

Problem 2.1 (5% weighting)

For training, we will use the "iterative shrinkage/thresholding" scheme:

$$\boldsymbol{\theta}^{(i)} = S_{\lambda\mu}\left(\boldsymbol{\theta}^{(i-1)} + \mu X^T \boldsymbol{e}^{(i-1)}\right)$$

where $S_{\lambda\mu}(\cdot)$ denotes the shrinkage/thresholding function.

We run one iteration of the algorithm. Assume the following:

$$\boldsymbol{\theta}^{(i-1)} = \begin{bmatrix} 1 & 0.9 \end{bmatrix}^T$$
$$X^T\boldsymbol{e}^{(i-1)} = \begin{bmatrix} -0.5 & -0.3 \end{bmatrix}^T$$

Determine the smallest value of λ that results in both components in $\boldsymbol{\theta}^{(i)}$ being zero with $\mu = 0.1$.

Solution: The shrinkage / thresholding function is defined in ML sec 10.2.2, p 481, as

$$S_{\lambda\mu}(\theta) = \text{sgn}(\theta)\left(|\theta| - \lambda\mu\right)_+$$

We complete the iteration as (eq. 10.7)

$$\begin{aligned}\boldsymbol{\theta}^{(i)} &= S_{\lambda\mu}\left(\boldsymbol{\theta}^{(i-1)} + \mu X^T \boldsymbol{e}^{(i-1)}\right) \\ &= S_{\lambda\mu}\left(\begin{bmatrix} 1 \\ 0.9 \end{bmatrix} + \mu \begin{bmatrix} -0.5 \\ -0.3 \end{bmatrix}\right)\end{aligned}$$

Since we need to find the minimum value for λ that gives both component being zero, we get

$$|1 - \mu 0.5| = \lambda\mu$$
$$|0.9 - \mu 0.3| = \lambda\mu$$

For $\mu = 0.1$, we get

$$|1 - 0.1 \cdot 0.5| = 0.1\lambda \quad \Rightarrow \lambda = 10 \cdot (1 - 0.1 \cdot 0.5) = 9.5$$
$$|0.9 - 0.1 \cdot 0.3| = 0.1\lambda \quad \Rightarrow \lambda = 10 \cdot (0.9 - 0.1 \cdot 0.3) = 8.7$$

Hence, the lowest possible value is $\lambda = 9.5$.

Problem 2.2 (5% weighting)
Assume we have chosen $\lambda = 1$, and we obtain an reliable estimate of the noise of the data, denoted $\sigma^2 = 1$. How does the original regression problem relate to the prior distribution of $\boldsymbol{\theta}$? Determine the parameters of the prior distribution of $\boldsymbol{\theta}$.

Solution: We know from exercise 9.1 that the LASSO regression is equivalent to the MAP estimate, if using a normal likelihood and a univariate zero-mean Laplace distribution on each weight component in $\boldsymbol{\theta}$. We have

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \frac{1}{2\sigma^2}\|(\boldsymbol{y} - f(X, \boldsymbol{\theta}))\|^2 + \frac{1}{b}\|\boldsymbol{\theta}\|_1$$

If we multiply this expression with $2\sigma^2$ we get the Lasso cost function

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \arg\min_{\boldsymbol{\theta}} \|(\boldsymbol{y} - f(X, \boldsymbol{\theta}))\|^2 + \frac{2\sigma^2}{b}\|\boldsymbol{\theta}\|_1 \\ &= \arg\min_{\boldsymbol{\theta}} \|(\boldsymbol{y} - f(X, \boldsymbol{\theta}))\|^2 + \lambda\|\boldsymbol{\theta}\|_1\end{aligned}$$

where $\lambda = 2\frac{\sigma^2}{b}$. Inserting the values, we get

$$b = 2\frac{\sigma^2}{\lambda} = 2\frac{1}{1} = 2$$

and we see that for this choice of λ and assumed measurement noise σ_2, we have $b = 2$ as the parameter to the Laplace distribution, defined as

$$p(x|\mu, b) = \frac{1}{2b}\exp\left(-\frac{|x - \mu|}{b}\right)$$

Problem 3 Linear filtering (30% total weighting)

In this problem we are considering the echo canceling setup, where we assume the echo canceller is a linear FIR filter, with l coefficients. The setup is depicted in Figure 1, page 6.

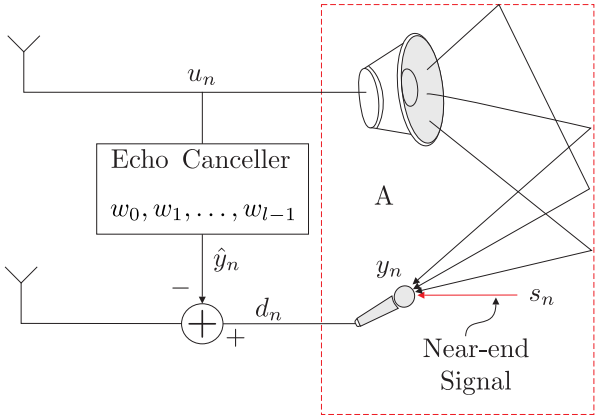

Figure 1: Problem 3 setup.

Assume that we have a u_n sequence as follows:

n	0	1	2	3	4	5	6
u_n	1	6	5	2	3	1	6

Problem 3.1 (5% weighting)

The filter has the coefficients $\boldsymbol{w} = \begin{bmatrix} 0.50 & 0.30 & w_2 \end{bmatrix}^T$.

The output of the filter at $n = 4$ is $\hat{y}_4 = 3.1$. Determine the numerical value of w_2 that was used to compute $\hat{y}_4$.

Solution: From ML sec 4.5, eq (4.41), we have

$$\begin{aligned}
\hat{d}_n &= \sum_{i=0}^{l-1} w_i u_{n-i} \qquad \Rightarrow \\
\hat{d}_4 &= \sum_{i=0}^{2} w_i u_{4-i} \\
&= w_0 u_4 + w_1 u_3 + w_2 u_2 \qquad \Leftrightarrow \\
w_2 &= -\frac{w_0 u_4 + w_1 u_3 - \hat{d}_4}{u_2} \\
w_2 &= -\frac{0.5 \cdot 3 + 0.3 \cdot 2 - 3.1}{5} \\
&= 0.2
\end{aligned}$$

Problem 3.2 (7% weighting)
We will now consider all signals as wide-sense stationary stochastic processes.

We are informed that room A has a room impulse response of $H_A = [0, h_1, h_2]$, such that the proportion of u_n that is received by the microphone y_n is H_A convolved with $\mathbf{u}_n$ (where $\mathbf{u}_n = \begin{bmatrix}\mathrm{u}_n & \mathrm{u}_{n-1} & \mathrm{u}_{n-2}\end{bmatrix}^T$).

Derive an analytical expression for the cross-correlation function $r_{yu}(k)$.

Solution: First we need to define y_n, which according to the setup is

$$\mathrm{y}_n = \mathrm{s}_n + H_A * \mathbf{u}_n$$

where $\mathbf{u}_n = \begin{bmatrix}\mathrm{u}_n & \mathrm{u}_{n-1} & \mathrm{u}_{n-2}\end{bmatrix}^T$. According to ML p 136, the cross-correlation function is defined as $r_{yu}(k) = \mathbb{E}[\mathrm{y}_n \mathrm{u}_{n-k}]$. Thus we get

$$\begin{aligned}
r_{yu}(k) &= \mathbb{E}[\mathrm{y}_n \mathrm{u}_{n-k}] \\
&= \mathbb{E}[(\mathrm{s}_n + H_A * \mathbf{u}_n)\, \mathrm{u}_{n-k}] \\
&= \mathbb{E}\left[\left(\mathrm{s}_n + \sum_{i=0}^{2} h_i \mathrm{u}_{n-i}\right) \mathrm{u}_{n-k}\right] \\
&= \mathbb{E}[(\mathrm{s}_n + h_1 \mathrm{u}_{n-1} + h_2 \mathrm{u}_{n-2})\, \mathrm{u}_{n-k}] \\
&= \mathbb{E}[\mathrm{s}_n \mathrm{u}_{n-k} + h_1 \mathrm{u}_{n-1}\mathrm{u}_{n-k} + h_2 \mathrm{u}_{n-2}\mathrm{u}_{n-k}] \\
&= \mathbb{E}\big[\mathrm{s}_n \mathrm{u}_{n-k} + h_1 \mathrm{u}_n \mathrm{u}_{n-(k-1)} + h_2 \mathrm{u}_n \mathrm{u}_{n-(k-2)}\big] \\
&= r_{su}(k) + h_1 r_u(k-1) + h_2 r_u(k-2)
\end{aligned}$$

Problem 3.3 (5% weighting)
We now obtain precise measurements of the following correlation functions, where $r_u(k)$ denotes the correlation function for the signal u_n, and $r_{du}(k)$ denotes the cross-correlation function between d_n and u_n:

k	0	1	2	3
$r_u(k)$	1.0	0.9	0.7	0.5
$r_{du}(k)$	0.45	0.55	0.30	0.20

Determine the filter coefficient values $\boldsymbol{w}$ (still with a filter length $l = 3$) based on the measured correlation values.

Additionally, we need a filter which has, at most, a minimum mean squared error (MMSE) of 0.1. Determine the highest tolerable value for $r_d(0)$.

You should specify both the exact formulas and the numerical values.

Solution: According to ML sec 4.5 we have for the filter coefficients

$$\begin{aligned}\begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix} &= \begin{bmatrix} r_u(0) & r_u(1) & r_u(2) \\ r_u(1) & r_u(0) & r_u(1) \\ r_u(2) & r_u(1) & r_u(0) \end{bmatrix}^{-1} \begin{bmatrix} r_{du}(0) \\ r_{du}(1) \\ r_{du}(2) \end{bmatrix} \\ &= \begin{bmatrix} 1.00 & 0.90 & 0.70 \\ 0.90 & 1.00 & 0.90 \\ 0.70 & 0.90 & 1.00 \end{bmatrix}^{-1} \begin{bmatrix} 0.45 \\ 0.55 \\ 0.30 \end{bmatrix} \\ &= \begin{bmatrix} -1.25 \\ 3.25 \\ -1.75 \end{bmatrix}\end{aligned}$$

According to ML eq 4.9, we get, for the minimum error

$$\begin{aligned}J(\boldsymbol{\theta}) &= \sigma_y^2 - \boldsymbol{p}^T \Sigma_u^{-1} \boldsymbol{p} \\ &= r_d(0) - \begin{bmatrix} 0.45 \\ 0.55 \\ 0.30 \end{bmatrix}^T \begin{bmatrix} 1.00 & 0.90 & 0.70 \\ 0.90 & 1.00 & 0.90 \\ 0.70 & 0.90 & 1.00 \end{bmatrix}^{-1} \begin{bmatrix} 0.45 \\ 0.55 \\ 0.30 \end{bmatrix} \\ &= r_d(0) - 0.7\end{aligned}$$

That means $r_d(0)$ can at most be 0.8, in order to obtain $J(\boldsymbol{\theta}) \leq 0.1$

Problem 3.4 (5% weighting)
We now have a change of the signal u_n with new correlation values, and decide to learn the filter weights using the LMS algorithm. We are given the following knowledge of the covariance matrix of u_n where u_n is assumed to be a wide-sense stationary stochastic process.

$$\Sigma_u = \begin{bmatrix} 2.00 & 1.40 & 0.70 \\ 1.40 & 2.00 & 1.40 \\ 0.70 & 1.40 & 2.00 \end{bmatrix}$$

We want to choose the step size such that we are sure that the filter converges and the weight estimator has bounded variance.

Determine the maximum value for the step-size μ.

You should specify both the exact formulas and the numerical values.

Solution: From ML pp 200–202, we have two conditions to check, both on the input covariance matrix, that is, Σ_u.

First we find the eigenvalues of Σ_u to be

$$\boldsymbol{\lambda} = \begin{bmatrix} 0.339 & 1.300 & 4.361 \end{bmatrix}^T$$

Then, the maximum μ for LMS convergence is

$$\mu < \frac{2}{\lambda_{max}} = \frac{2}{4.361} = 0.459$$

However, according to formula (5.45), the maximum μ is

$$\mu < \frac{2}{\text{Trace}(\Sigma_u)} = \frac{2}{6.000} = 0.333$$

Hence, the maximum step-size is $\mu = 0.333$

Problem 3.5 (8% weighting)
We now assume that the time-varying component of the entire system is adequately modeled using the following model:

$$\begin{aligned} \mathrm{d}_n &= \boldsymbol{w}_{o,n-1}^T \mathbf{u}_n + \eta_n \\ \boldsymbol{w}_{o,n} &= \boldsymbol{w}_{o,n-1} + \boldsymbol{\omega}_n \\ \mathbb{E}\left[\boldsymbol{\omega}_n \boldsymbol{\omega}_n^T\right] &= \begin{bmatrix} 0.03 & 0.00 & 0.00 \\ 0.00 & 0.02 & 0.00 \\ 0.00 & 0.00 & 0.01 \end{bmatrix} \end{aligned}$$

Where η_n and $\boldsymbol{\omega}_n$ are assumed to follow zero-mean normal distributions, and u_n has the same covariance matrix as the previous problem.

We will use NLMS instead of LMS, and want to determine the step-size μ that ensures that the time-varying component of the excess MSE reaches at maximum 0.1. Additionally, NLMS must be guaranteed to remain stable.

Determine the range of values for μ that fulfill these requirements.

You should specify both the formulas and the numerical values.

Solution: The model description fits the model described in ML p 275, eq 6.53–6.54. Thus, from ML p 276, table 6.1 we have

$$J_{MSE} = \frac{1}{2}\mu\sigma_\eta^2 \,\text{trace}(\Sigma_u)\mathbb{E}\left[\frac{q}{\|x\|^2}\right] + \frac{1}{2}\mu^{-1}\,\text{trace}(\Sigma_u)\,\text{trace}(\Sigma_\omega)$$

The terms that include $\text{trace}(\Sigma_\omega)$ will be the only part that contribute to the time-varying

component , hence, we solve for the time-varying part:

$$J_{MSE,time} = \frac{1}{2}\mu^{-1}\,\mathrm{trace}(\Sigma_x)\,\mathrm{trace}(\Sigma_\omega) \qquad \Leftrightarrow$$
$$\mu = \frac{1}{2J_{MSE,time}}\,\mathrm{trace}(\Sigma_x)\,\mathrm{trace}(\Sigma_\omega)$$

Inserting the numbers gives

$$\mu = \frac{1}{2 \cdot 0.1} \cdot 6 \cdot 0.06 = 1.8$$

Increasing the step-size further, will lower the error for the time-varying part further, however, to ensure stability we must have $\mu < 2$ (ML. p 212). The allowable interval for the step-size is thus $1.8 \leq \mu < 2$.

Problem 4 Dictionary learning (10% total weighting)

This problem concerns Independent Component Analysis (ICA) using Mutual information.

Problem 4.1 (5% weighting)
Let us assume that we have two sources, s_1 and s_2, which are statistically independent, and two observable variables x_1 and x_2, defined as

$$\mathrm{x}_1 = \mathrm{s}_1 + \mathrm{s}_2$$
$$\mathrm{x}_2 = \mathrm{s}_2$$

We will now unmix the signals x_1 and x_2 using ICA uising an unmixing matrix W:

$$\begin{bmatrix}\hat{\mathrm{s}}_1 \\ \hat{\mathrm{s}}_2\end{bmatrix} = W \begin{bmatrix}\mathrm{x}_1 \\ \mathrm{x}_2\end{bmatrix}$$

Identify W such that $\hat{\mathrm{s}}_1$ and $\hat{\mathrm{s}}_2$ has an average mutual information $I(\hat{\mathrm{s}}_1, \hat{\mathrm{s}}_2) = 0$.
Show formally that $I(\hat{\mathrm{s}}_1, \hat{\mathrm{s}}_2) = 0$ for your solution.

Solution: According to the ICA model, we have $\mathbf{x} = A\mathbf{s}$, where we read off the mixing matrix A as

$$A = \begin{bmatrix}1 & 1 \\ 0 & 1\end{bmatrix}$$

An obvious unmatrix to choose is then $W = A^{-1}$:

$$W = \begin{bmatrix}1 & -1 \\ 0 & 1\end{bmatrix}$$

Applying this matrix, we see that

$$\begin{bmatrix}\hat{s_1}\\ \hat{s_2}\end{bmatrix} = \begin{bmatrix}1 & -1\\ 0 & 1\end{bmatrix}\begin{bmatrix}s_1+s_2\\ s_2\end{bmatrix} = \begin{bmatrix}s_1+s_2\\ s_2\end{bmatrix}\begin{bmatrix}s_1+s_2-s_2\\ s_2\end{bmatrix} = \begin{bmatrix}s_1\\ s_2\end{bmatrix}$$

From 2.518, we have the definition of the average mutual information

$$I(s_1,s_2) = \int\int p(s_1,s_2)\log\frac{p(s_1,s_2)}{p(s_1)p(s_2)}ds_1ds_2$$

Since s_1 and s_2 are independent, we have $p(s_1,s_2) = p(s_1)p(s_2)$. Hence, we get a $\log 1 = 0$ term in the integrand, and hence the mutual information between s_1 and s_2 becomes zero.

Problem 4.2 (5% weighting)
We know that the ICA model cannot identify the scale and direction of the identified vectors of the unmixing matrix W, and consequently in the mixing matrix A.

Write a short proof that shows the ICA model has a scaling ambiguity, that is, for a specific identified W (or A), this W (or A) can be scaled, vector-wise, and still result in sources that are the same, up to scaling factor.

Solution: This can be shown in a few ways. One approach is to start from the ICA model. We have $\mathbf{x} = A\mathbf{s}$. We can now create a new matrix, denoted C, with the same dimensions as A, as

$$\begin{bmatrix}c_1 & 0 & \cdots & 0\\ 0 & c_2 & \cdots & 0\\ \vdots & & \ddots & \\ 0 & 0 & \cdots & c_D\end{bmatrix}$$

where c_i is a non-zero scalar. In that case, C is invertible, and $CC^{-1} = I$. We can thus write the ICA model as

$$\begin{aligned}\mathbf{x} &= ACC^{-1}\mathbf{s}\\ &= A'\mathbf{s}'\end{aligned}$$

where $A' = AC$ and $\mathbf{s}' = C^{-1}\mathbf{s}$. Since the matrix C is a diagonal matrix, the mixing matrix and source matrix is only scaled, and the resulting $\mathbf{x}$ will be the same.

Problem 5 Hidden Markov Models (12% total weighting)

This problem concerns a Hidden Markov Model (HMM) where x_n denotes the state of the chain at time n, and y_n denotes the observation at time n.

Problem 5.1 (5% weighting)
Suppose a 3-state HMM have the initial state probability vector

$$P_k = \begin{bmatrix} 0.25 & 0.25 & 0.50 \end{bmatrix}^T$$

and the following state transition probabilities

$$P_{ij} = \begin{bmatrix} 0.5 & 0.5 & 0.0 \\ 0.0 & 0.2 & 0.8 \\ 0.0 & 0.9 & 0.1 \end{bmatrix}$$

We will construct a state sequence of length 8, by using each of the following state sub-sequences exactly once:

$$A = \begin{bmatrix} 1 & 2 \end{bmatrix}, \quad B = \begin{bmatrix} 2 & 3 \end{bmatrix}, \quad C = \begin{bmatrix} 3 & 2 \end{bmatrix}, \quad D = \begin{bmatrix} 3 & 3 \end{bmatrix}$$

E.g., the composition 'ABCD' will give the state sequence

$$x_n = \begin{bmatrix} 1 & 2 & 2 & 3 & 3 & 2 & 3 & 3 \end{bmatrix}$$

Specify the composition that creates the most likely state sequence given the defined HMM. Justify your choices.

Solution: Firstly, since the only way to be in state 1 is by having the chain start in state 1, we get A as the first sub-sequence (since $P_{21} = 0$ and $P_{31} = 0$).

Secondly, when in state 2 and state 3, the probabilities of making a state transition to other states are higher than staying in the current state.

That means we need to find the combination of sub-sequences such that our state transition gives the maximum number of state changes.

This combination of the sub-sequences "ACDB" which has a state change at every sub-sequence shift. The breakdown is as follows (number of changes between sub-sequences)

ABCD : 1 change.
ABDC : 0 changes.
ACBD : 1 change.
ACDB : 3 changes.
ADBC : 2 changes.
ADCB : 1 change.

Problem 5.2 (7% weighting)
You are now informed that the 3-state HMM allows three possible observable actions, a_1, a_2, and a_3 with the following emission probabilities

	a_1	a_2	a_3
$P(y_n = a_i \mid x_n = 1)$	0.7	0.1	0.2
$P(y_n = a_i \mid x_n = 2)$	0.5	0.4	0.1
$P(y_n = a_i \mid x_n = 3)$	0.9	0.0	0.1

We want to compute the initial state probability vector P_k, instead of using the P_k from the original problem.

Based on observations, you additionally know that $P(y_1 = a_1) = 0.68$ and $P(y_1 = a_2) = 0.15$.

Compute the initial state probability vector P_k using the given information.

You should specify both the formulas and the numerical values.

Solution: Use marginalization (sum rule), and the structure of a HMM, we get

$$P(y_1) = \sum_{i=1}^{K} P(y_1 \mid x_1 = i)P(x_1 = i)$$

Additionally, we know that

$$\sum_{i=1}^{3} P(y_1 \mid a_i) = 1$$

which means $P(y_1 \mid a_3) = 1 - P(y_1 \mid a_1) - P(y_1 \mid a_2) = 0.17$.

We can now, using marginalization, create three equations with three unknowns. Let us denote P_i as the probability of starting in state i, and $P_{a_i,x_j} = P(y_1 = a_i \mid x_1 = j)$, we get

$$\begin{aligned}
P(y_1 = a_1) &= P_{a_1,x_1}P_1 + P_{a_1,x_2}P_2 + P_{a_1,x_3}P_3 \\
P(y_1 = a_2) &= P_{a_2,x_1}P_1 + P_{a_2,x_2}P_2 + P_{a_2,x_3}P_3 \\
P(y_1 = a_3) &= P_{a_3,x_1}P_1 + P_{a_3,x_2}P_2 + P_{a_3,x_3}P_3
\end{aligned}$$

We can now solve using normal linear algebra

$$\begin{bmatrix} P_1 \\ P_2 \\ P_3 \end{bmatrix} = \begin{bmatrix} 0.70 & 0.50 & 0.90 \\ 0.10 & 0.40 & 0.00 \\ 0.20 & 0.10 & 0.10 \end{bmatrix}^{-1} \begin{bmatrix} 0.68 \\ 0.15 \\ 0.17 \end{bmatrix} = \begin{bmatrix} 0.70 \\ 0.20 \\ 0.10 \end{bmatrix}$$

Problem 6 Kalman filtering (8% total weighting)

In this problem we consider Kalman filtering.

Problem 6.1 (8% weighting)
Consider a one-dimensional discrete-time system being tracked using a Kalman filter. The system is defined by the following state-space equations:

$$\begin{aligned} \mathrm{x}_n &= \mathrm{x}_{n-1} + \eta_n \\ \mathrm{y}_n &= \mathrm{x}_n + \mathrm{v}_n \end{aligned}$$

Assume that η_n follows a zero-mean normal distribution with $\sigma_\eta^2 = 0.01$, and v_n follows a zero-mean normal distribution with $\sigma_v^2 = 0.1$. We now observe one observation, $y_1 = 2$.

Calculate the updated state estimate $\hat{x}_1$ and error covariance P_1 after receiving the first measurement. Use as initial state $x = 0$ and $P = 1$.

Select the correct statement:

A : $\hat{x}_1 = 1.818$, $P_1 = 0.101$
B : $\hat{x}_1 = 1.818$, $P_1 = 0.091$
C : $\hat{x}_1 = 1.980$, $P_1 = 0.010$
D : $\hat{x}_1 = 2.165$, $P_1 = 0.135$
E : $\hat{x}_1 = 1.523$, $P_1 = 0.241$
F : Don't know.

Solution: From ML p 169, algorithm 4.2, we get the Kalman filtering algorithm. The problem asks to calculate the estimates after correction, so that is, using the notation from the book: $\hat{\boldsymbol{x}}_{n|n}$ and $P_{n|n}$. Following the algorithm we get:

$$\begin{aligned} S_n &= R_n + H_n P_{n|n-1} H_n^T \\ &= 0.1 + 1 \cdot 1 \cdot 1 = 1.1 \\ K_n &= P_{n|n-1} H_n^T S_n^{-1} \\ &= 1 \cdot 1 \cdot 1.1^{-1} = 0.909 \\ \hat{\boldsymbol{x}}_{n|n} &= \hat{\boldsymbol{x}}_{n|n-1} + K_n(\boldsymbol{y}_n - H_n \hat{\boldsymbol{x}}_{n|n-1}) \\ &= 0 + 0.909(2 - 1 \cdot 0) = 1.818 \\ P_{n|n} &= P_{n|n-1} - K_n H_n P_{n|n-1} \\ &= 1 - 0.909 \cdot 1 \cdot 1 = 0.091 \end{aligned}$$

Thus the correct answer is B.

Problem 7 Kernel methods (15% total weighting)

In this problem we will consider kernel methods.

Problem 7.1 (5% weighting)

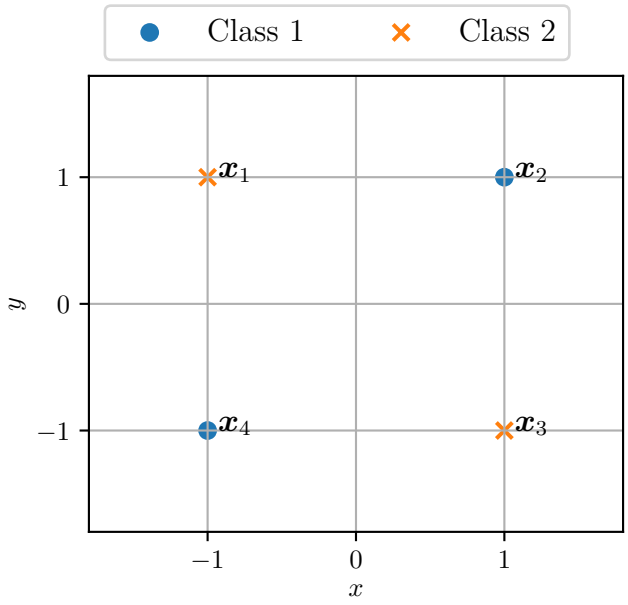

Figure 2: Problem 7.1.

In this problem we will use the Gaussian kernel

$$\kappa(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{y}\|^2}{2\sigma^2}\right)$$

We have observed the data points as depicted on Figure 2, page 15 as training data.

Write up the complete kernel matrix $\mathcal{K}$ for the data, where you have used the kernel function to evaluate the points. Use $\sigma = 2$.

You should specify both the formulas and the numerical values.

Solution: The kernel matrix $\mathcal{K}$ is defined in ML p 542, eq. (11.11), and we get

$$\mathcal{K} = \begin{bmatrix} \kappa(\boldsymbol{x}_1, \boldsymbol{x}_1) & \kappa(\boldsymbol{x}_1, \boldsymbol{x}_2) & \cdots & \kappa(\boldsymbol{x}_1, \boldsymbol{x}_4) \\ \kappa(\boldsymbol{x}_2, \boldsymbol{x}_1) & \kappa(\boldsymbol{x}_2, \boldsymbol{x}_2) & \cdots & \kappa(\boldsymbol{x}_2, \boldsymbol{x}_4) \\ \vdots & & \ddots & \\ \kappa(\boldsymbol{x}_4, \boldsymbol{x}_1) & \kappa(\boldsymbol{x}_4, \boldsymbol{x}_2) & \cdots & \kappa(\boldsymbol{x}_4, \boldsymbol{x}_4) \end{bmatrix}$$

There are three different cases. Taking $\boldsymbol{x}_1$ as the template, we get

$$\kappa(\boldsymbol{x}_1, \boldsymbol{x}_1) = \exp(0) = 1$$
$$\kappa(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp\left(-\frac{2^2}{2 \cdot 2^2}\right) = \exp\left(-\frac{1}{2}\right) = 0.61$$
$$\kappa(\boldsymbol{x}_1, \boldsymbol{x}_4) = \exp\left(-\frac{\sqrt{8}^2}{2 \cdot 2^2}\right) = \exp\left(-1\right) = 0.37$$

Putting this together, we get the kernel matrix

$$\mathcal{K} = \begin{bmatrix} 1.00 & 0.61 & 0.37 & 0.61 \\ 0.61 & 1.00 & 0.61 & 0.37 \\ 0.37 & 0.61 & 1.00 & 0.61 \\ 0.61 & 0.37 & 0.61 & 1.00 \end{bmatrix}$$

Problem 7.2 (5% weighting)
Now assume that two test points are given, $\boldsymbol{x}_1^{test}$ which belongs to class 1, and $\boldsymbol{x}_2^{test}$ which belongs to class 2. Their kernel values w.r.t. the training data (from Figure 2, page 15) are given as

	$\kappa(\cdot, \boldsymbol{x}_1)$	$\kappa(\cdot, \boldsymbol{x}_2)$	$\kappa(\cdot, \boldsymbol{x}_3)$	$\kappa(\cdot, \boldsymbol{x}_4)$
$\boldsymbol{x}_1^{test}$	0.97	0.59	0.46	0.75
$\boldsymbol{x}_2^{test}$	0.59	0.97	0.75	0.46

Use the representer theorem to determine how these two points can be correctly classified by finding suitable numerical values for θ_n, such that both are correctly classified using the same values for θ_n. Have a decision threshold as:

$$f(\boldsymbol{x}_1^{test}) > 0$$
$$f(\boldsymbol{x}_2^{test}) < 0$$

and have as many $\theta_n = 0$ as possible.

Solution: The representer theorem is written in ML p 548, and we have from eq. (11.17):

$$f(\cdot) = \sum_{n=1}^{4} \theta_n \kappa(\cdot, \boldsymbol{x}_n)$$

If we choose $\theta_1 = 1$, $\theta_2 = -1$, and the remaining $\theta_n = 0$, we get

$$f(\boldsymbol{x}_1^{test}) = 0.97 - 0.59 = 0.38$$
$$f(\boldsymbol{x}_2^{test}) = -0.97 + 0.59 = -0.38$$

Having a classification threshold where $f(\boldsymbol{x}) > 0$ gives class 1, and $f(\boldsymbol{x}) < 0$ gives class 2 solves the problem.

Problem 7.3 (5% weighting)
We consider a regression problem using support vector regression, fitted on five data points. We know that the data is fitted using $C = 1$ and $\epsilon = 0.1$, and the bias is estimated to $\hat{\theta}_0 = 0$. We have the following information:

n	1	2	3	4	5
$y_n - \hat{y}_n$	-0.14	0.04	0.11	-0.01	0.10
$\kappa(\boldsymbol{x}^{test}, \boldsymbol{x}_n)$	0.38	0.92	0.84	0.28	0.03

Computer either an exact value for $\hat{y}(\boldsymbol{x}^{test})$, or a tight bound for $\hat{y}(\boldsymbol{x}^{test})$, whichever is possible. You should specify both the formulas and the numerical values.

Solution: From ML p 557, we have the prediction formula eq (11.36) given as

$$\hat{y}(\boldsymbol{x}^{test}) = \sum_{n=1}^{N} (\tilde{\lambda}_n - \lambda_n)\kappa(\boldsymbol{x}^{test}, \boldsymbol{x}_n) + \hat{\theta}_0$$

From ML p 556 eq. (11.31-11.32), and the remarks on ML p 559, we know that $\tilde{\lambda} = C$ when the error $e_n = y_n - \hat{y}_n > \epsilon$, and $\lambda = C$ when the error $e_n = y_n - \hat{y}_n < \epsilon$ (for $n = 1$ and $n = 3$).
For $|e_n| < \epsilon$, we have $\tilde{\lambda}_n = \lambda_n = 0$ ($n = 2$ and $n = 4$).
Finally, for $\boldsymbol{x}_5$, the error is exactly equal to ϵ, so we know possible values are $0 \leq \tilde{\lambda}_5 \leq C$.
Putting all this information together, we get

$$\begin{aligned}\hat{y}(\boldsymbol{x}^{test}) &= -\lambda_1\kappa(\boldsymbol{x}^{test}, \boldsymbol{x}_1) + \tilde{\lambda}_3\kappa(\boldsymbol{x}^{test}, \boldsymbol{x}_3) + \tilde{\lambda}_5\kappa(\boldsymbol{x}^{test}, \boldsymbol{x}_5) \\ &= -0.38 + 0.84 + \tilde{\lambda}_5 0.03 \\ &= 0.46 + \tilde{\lambda}_5 0.03\end{aligned}$$

Thus we can conclude $0.46 \leq \hat{y}(\boldsymbol{x}^{test}) \leq 0.49$.