02471 Machine Learning for Signal Processing

Written examination: December 7, 2021.

Course name: Machine Learning for Signal Processing.

Course number: 02471.

Aids allowed: All aids permitted.

Exam duration: 4 hours.

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

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 was used from the curriculum.

Multiple choice: The following problems are multiple choice:

Problem 4.2

If the problem is answered correctly, you are given 5 points. For a wrong answer, you are subtracted 1.25 points.

Problem 1 Correlation functions (15% total weighting)

Problem 1.1 (5% weighting)

Consider the following sequence starting at $n = 0$:

$$y_n = \{5, 3, 6, 2\}$$

Calculate the biased auto-correlation value for $r_y(0)$, and $r_y(3)$.

Problem 1.2 (5% weighting)

Consider the following signal

$$\mathrm{x}_n = \alpha \mathrm{y}_{n-d_1} + \beta \mathrm{z}_{n-d_2}$$

where α and β are real constants, and d_1 and d_2 are positive integers. The signals y_n and z_n are wide-sense stationary processes. Assume that the three correlation functions $r_y(k)$, $r_z(k)$, and $r_{yz}(k)$ are known.

Determine an analytic expression of the auto-correlation function $r_x(k)$.

Problem 1.3 (5% weighting)

We consider again x_n from the previous problem. Assume that y_n now follows an AR(1) auto-regressive process, i.e. $\mathrm{y}_n = a_1 \mathrm{y}_{n-1} + \eta_n$. z_n is now a random variable taking *one* value for the entire realization of x_n (such that $z_n = z_{n+1}$). z_n is drawn from a uniform distribution in the interval [0;1]. You can assume that there are collected sufficient samples such that $\mathbb{E}[\mathrm{y}_n] = 0$.

Determine if x_n is a second order ergodic process.

Problem 2 Parameter estimation (25% total weighting)

In this problem we will consider linear regression using the mean squared error as the loss function. You have the following observations

$$y_n = \{0.5, 1.75, 4.3, 6.1\}, \quad \text{for} \quad n = \{0, 1, 2, 3\}$$

as visualized on Figure 1, page 3. To model the response, we will use a linear expression on the form $f_n = \beta + \theta n$.

Problem 2.1 (5% weighting)

Determine the least squares estimate of the coefficients b and θ. Write up the matrices and vectors used to calculate the estimate.

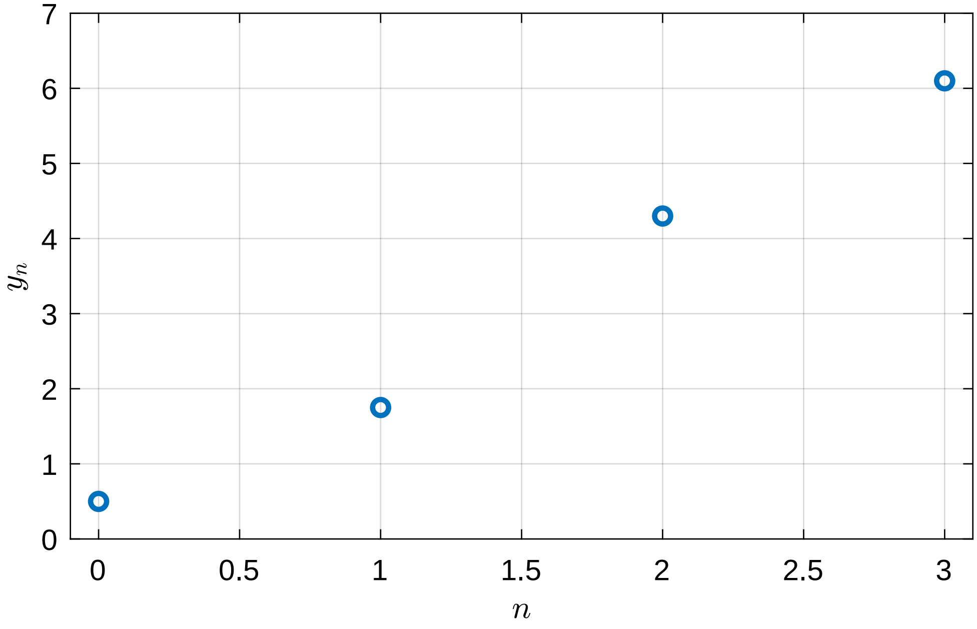

Figure 1: Problem 2.

Problem 2.2 (5% weighting)

For the remainder of this problem, β is set to zero such that only θ is estimated. We want to explore biased estimation, and calculate the error of our estimator $\hat{\theta}$ compared to the optimal value θ^{opt}. Additionally, the following holds

- The variance of the minimum variance unbiased estimator of θ is one, $\text{var}[\hat{\theta}_{\text{MVU}}] = 1$.
- The optimal value of θ is known to be $\theta^{\text{opt}} = 2$.

The biased estimator, $\hat{\theta}_b$, is constructed as $\hat{\theta}_b = (1+\alpha)\hat{\theta}_{\text{MVU}}$.

Calculate the values of α where $\hat{\theta}_b$ have a lower mean squared error than $\hat{\theta}_{\text{MVU}}$.

Problem 2.3 (5% weighting)

We will now disregard the observations from Figure 1, page 3, and assume a completely new set of observations and parameter value for θ (β is still assumed to be zero). Suppose now we apply ℓ_1 regularization to the regression problem, with λ denoting the regularization strength. We will use the “iterative shrinkage/thresholding” scheme:

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

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

We run one iteration of the algorithm. Set the step-size to 0.5, assume that $\theta^{(i-1)} = 4$, and the error vector is $\boldsymbol{e}^{(i-1)} = \begin{bmatrix} 1 & -1 & -3 & 2 \end{bmatrix}^T$.

Determine the value of λ that results in $\theta^{(i)} = 3$.

Problem 2.4 (5% weighting)

Suppose now we apply ℓ_2 regularization to our problem, with λ denoting the regularization strength of the ℓ_2. Suppose that we choose $\lambda = 2$, and we interpret Ridge regression from a Bayesian viewpoint. In this viewpoint, we choose a distribution on our θ with a variance set to 0.5. Answer the two questions:

- What does these parameters indicate about our assumption of the statistical properties of measurement noise?
- What is the assumed value of the measurement noise?

Problem 2.5 (5% weighting)

Now assume that we take a Bayesian approach, and have a likelihood that is Gaussian with mean $X\theta$ and covariance $\sigma_\eta^2 I$, and a prior distribution on θ that is Gaussian with zero mean and variance denoted as σ_θ^2.

The measurement noise can now be estimated using the Expectation-Maximization algorithm by specifying the complete log-likelihood

$$\ln p(\boldsymbol{y}, \theta; \boldsymbol{\xi}), \quad \boldsymbol{\xi} = \{\sigma_\eta^2, \sigma_\theta^2\}$$

Determine an expression for $\ln p(\boldsymbol{y}, \theta; \boldsymbol{\xi})$ as a function of $\sigma_\eta^2, \sigma_\theta^2$. Define the meaning of other terms that is in your expression.

Problem 3 Linear adaptive filtering (15% total weighting)

We now consider a linear filtering setup, using a FIR filter of length $l = 3$, as illustrated on Figure 2, page 5

Problem 3.1 (7% weighting)

You are now informed of the following correlation function values, where d denotes the desired signal, and x denotes the input signal.

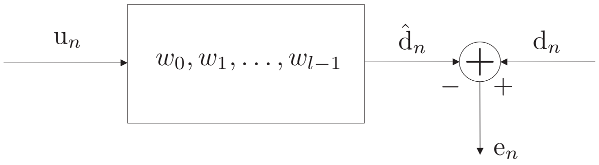

Figure 2: Problem 3.

$r_u(0)$	$r_u(1)$	$r_u(2)$	$r_u(3)$
1.7	0.7	0.4	0.2

$r_{du}(0)$	$r_{du}(1)$	$r_{du}(2)$	$r_{du}(3)$
1.2	0.6	0.35	0.05

$r_d(0)$	$r_d(1)$	$r_d(2)$	$r_d(3)$
1.4	1.0	0.7	0.3

Determine the filter coefficient values $\boldsymbol{w}$ and the minimum mean squared error (MMSE) as achieved by the filter. You should specify both the exact formulas and the numerical values.

Problem 3.2 (8% weighting)

We consider again the filter with the same correlation values as the previous problem. Assume now that the filter is deployed in a time-varying environment and reduced to a filter size of 2. The time-varying component of the system is adequately modeled by the following system:

$$
\begin{aligned}
\mathrm{y}_n &= \theta_{o,n-1}^T \mathrm{u}_n + \eta_n \\
\theta_{o,n} &= \theta_{o,n-1} + \omega_n \\
\mathbb{E}\left[\omega_n \omega_n^T\right] &= \begin{bmatrix} 0.1 & 0.1 \\ 0.3 & 0.2 \end{bmatrix}
\end{aligned}
$$

where $\mathbb{E}[\eta_n] = 0$, $\mathbb{E}[\omega_n] = \mathbf{0}$, and $\sigma_\eta^2 = 0.5$.

Determine the steady-state excess MSE when using the LMS algorithm for $\mu = 0.5$. You should specify both the exact formulas and the numerical values.

Problem 4 Dictionary learning (10% total weighting)

We consider the independent component analysis model.

Problem 4.1 (5% weighting)

Suppose that we have three recorded signals, denoted as X:

$$X = \begin{bmatrix} 4 & 6 & 8 & 4 & 6 \\ 2 & 8 & 8 & 2 & 6 \\ 9 & 9 & 6 & 3 & 6 \end{bmatrix}$$

Assume that you are informed that the sources are:

$$Z = \begin{bmatrix} 3 & 2 & 4 & 3 & 3 \\ 1 & 4 & 4 & 1 & 3 \\ 3 & 3 & 2 & 1 & 2 \end{bmatrix}$$

Determine the mixing matrix that explains the observations in X. You can assume that A has a total of four non-zero elements, and all the diagonal elements in A are non-zero.

Problem 4.2 (5% weighting)

In the following we refer to the following function as "transfer function"

$$\boldsymbol{\phi}(\mathbf{z}) := \left[-\frac{p_1'(\mathrm{z}_1)}{p_1(\mathrm{z}_1)}, \ldots, -\frac{p_l'(\mathrm{z}_l)}{p_l(\mathrm{z}_l)} \right]^T \text{ (Eq 19.57) in ML}$$

Which of the following statements concerning ICA is **false**:

A: If the sources are Gaussian distributed, one cannot accurately identify the sources.

B: The choice of super-Gaussian and sub-Gaussian as transfer function may affect which solution is identified by ICA.

C: If ICA is used to recover Gaussian sources, the mixing matrix is not estimated correctly.

D: If ICA is used to recover one Gaussian distributed source and one uniform source, the Gaussian source can be estimated.

E: The ICA solution can be found by minimizing the mutual information of the mixing matrix.

F: Don't know.

Problem 5 State-space models and time-frequency analysis (20% total weighting)

In this problem we will consider activity recognition of a person. We have accelerometer data recorded for 10 minutes of a persons cell-phone, and now want to estimate whether the person is running or walking.

Problem 5.1 (5% weighting)

The first step is to analyze the data using the short-time Fourier transform. The sampling rate of the signal is 5 Hz and we use a window size of 30 seconds, and a window overlap of 50%.

Determine the number of samples per window, and the number of windows being processed by STFT.

Problem 5.2 (5% weighting)

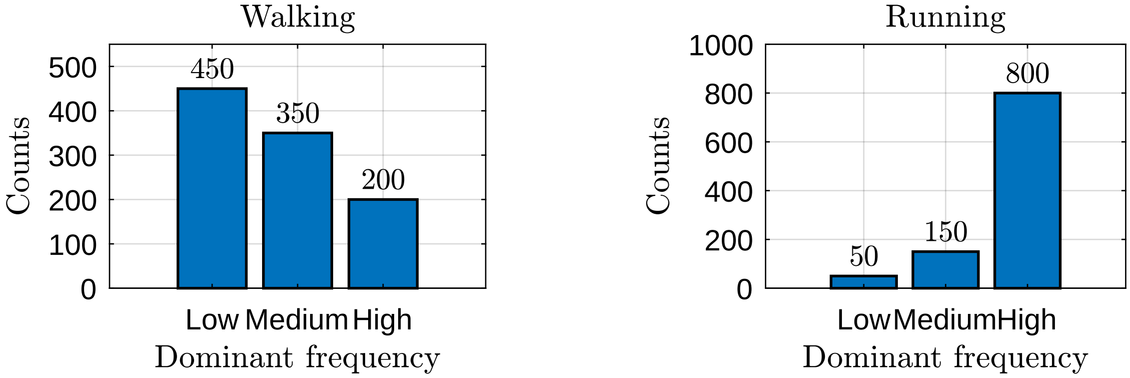

Figure 3: Problem 5.2.

We now consider a Hidden Markov model for the activity recognition of the person.

For this task, we collect five hours of walking data and five hours of running data, and segment the data into 1000 bins of each class. For each time segment, we calculate the total energy in three frequency ranges (low, medium and high frequency) and record the frequency range with the highest energy. The counts are shown on Figure 3, page 7.

Estimate as many parameter of the Hidden Markov model as possible with the information you currently have.

Problem 5.3 (10% weighting)

Now assume the following

- If the person is running, there is 80% probability that the high-frequency range has the highest energy, and 20% probability that the medium-frequency range has the highest energy.
- If the person is walking, there is 20% probability that the high-frequency range has the highest energy, and 50% probability that the medium-frequency range has the highest energy.
- The person was confirmed to be running for sure in the 30–60 second interval.
- Once the person is running, there is a 90% probability that the person will keep on running, and once the person is walking there is a 90% probability that the person will keep on walking.

Calculate the probability that the person was running in the 60–90 second interval, given that only high frequency content was observed for the first 90 seconds.

Problem 6 Kernels (15% total weighting)

In this problem we consider first a kernel function for a classification problem, and then kernel regression. For all problems, we will use the Gaussian kernel defined as

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

Problem 6.1 (5% weighting)

Consider the points on Figure 4, page 9 as the training points in a dataset, where the red crosses are class one, the blue circles are class two.

Assume now that you observe a test point of class blue, $\boldsymbol{x}_{test} = (-2, 0)$.

- Apply the kernel function to the closest points of each class in the data-set using $\sigma = 1$.
- Argue whether the kernel can be useful or not for separating the two classes.

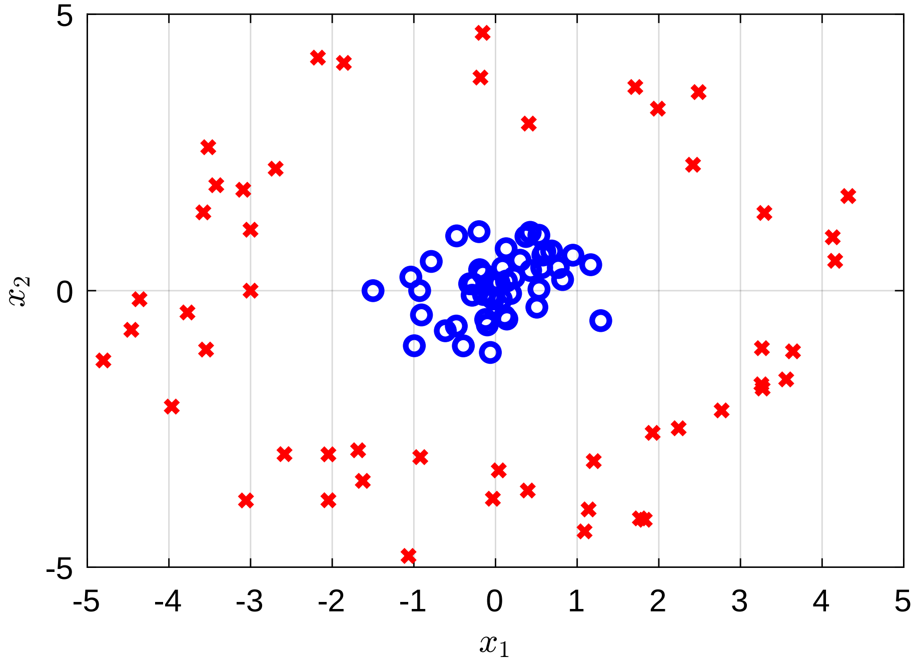

Figure 4: Problem 6.1

Problem 6.2 (5% weighting)

Now consider the kernel Ridge regression. We again use the Gaussian kernel with $\sigma = 1.5$. Suppose that we have three training points: $(x_1, y_1) = (1, 0.5)$, $(x_2, y_2) = (2, 0.8)$, and $(x_3, y_3) = (3, 0.7)$. Use $C = 0.02$.

Compute the weight vector $\boldsymbol{\theta}$, and specify both formulas used and the numerical result.

Problem 6.3 (5% weighting)

Consider again kernel Ridge regression with three observations. Assume now that a weight vector of $\theta_1 = 1$, $\theta_2 = 3$ was computed. You can assume that $\kappa(\boldsymbol{x}_{\text{test}}, \boldsymbol{x}_1) = 0.3$, $\kappa(\boldsymbol{x}_{\text{test}}, \boldsymbol{x}_2) = 0$, and $\kappa(\boldsymbol{x}_{\text{test}}, \boldsymbol{x}_3) = 0.6$.

We are also informed that the regression value at $\boldsymbol{x}_{\text{test}} = 3$ yielded a value of 3, i.e. $\hat{y}(\boldsymbol{x}_{\text{test}}) = 3$.

Determine the value of θ_3 used in the regression problem.