02471 Machine Learning for Signal Processing

Written examination: December 7, 2022.

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.

This exam has 6 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 3.2

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

Problem 1 Parameter estimation (28% total weighting)

In this problem we will consider various forms of parameter estimation.

Problem 1.1 (5% weighting)
Consider linear regression using the squared error as the loss function. You have the following three observations:

n	0	1	2
y_n	-0.4	0.3	1.8

To model the response y_n, we will use a linear model on the form $f_n = \theta_0 + \theta_1 \cdot n$.

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

Problem 1.2 (8% weighting)
Assume now that the parameters are determined using Ridge regression:

$$\hat{\boldsymbol{\theta}}_R = \arg\min_{\boldsymbol{\theta}} \left\{ \|\boldsymbol{y} - X^T\boldsymbol{\theta}\|^2 + \lambda\|\boldsymbol{\theta}\|^2 \right\} \tag{1}$$

The parameters are estimated to $\hat{\boldsymbol{\theta}}_R = \begin{bmatrix} -0.41 & 1 \end{bmatrix}^T$.

Derive a closed form expression[1] for λ, and determine which value for λ was used with 2 significant digits.

Problem 1.3 (7% weighting)
We now consider a different situation where we still assume a linear model and Ridge regression, but we assume that $X^TX = 2I$. Also assume that the least squares estimate (also under the assumption that $X^TX = 2I$) is $\hat{\boldsymbol{\theta}}_{LS} = \begin{bmatrix} -.75 & 1.3 \end{bmatrix}^T$. Additionally, assume $\lambda = 0.5$.

Derive an expression for $\hat{\boldsymbol{\theta}}_R$ under these circumstances, and determine the value of $\hat{\boldsymbol{\theta}}_R$.

Problem 1.4 (5% weighting)
Suppose now we apply ℓ_1 regularization to the regression problem (still using $\lambda = 0.5$):

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

We will now disregard the assumption $X^TX = 2I$ and instead use the "iterative shrinkage/thresholding" scheme:

$$\boldsymbol{\theta}^{(i)} = S_{\lambda\mu}\left(\boldsymbol{\theta}^{(i-1)} + \mu X^T\boldsymbol{e}^{(i-1)}\right) \tag{3}$$

[1] You will probably need the rule: $\lambda \cdot \boldsymbol{a} = \boldsymbol{b} \Rightarrow \lambda = \frac{\boldsymbol{a}^T\boldsymbol{b}}{\boldsymbol{a}^T\boldsymbol{a}}$, where $\boldsymbol{a}$ and $\boldsymbol{b}$ are vectors.

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

We run one iteration of the algorithm. Assume that $\boldsymbol{\theta}^{(i-1)} = \hat{\boldsymbol{\theta}}_{LS}$ from the previous problem, and that $X^T \boldsymbol{e}^{(i-1)} = \begin{bmatrix}0.25 & -0.35\end{bmatrix}^T$.

Determine the lowest value of μ that results in exactly one component of $\boldsymbol{\theta}^{(i)}$ being zero.

Problem 1.5 (3% weighting)

Assume now that we again use Ridge regression and $\lambda = 0.5$. Assume additionally that we have a reliable estimate of the variance of the measurement noise of $\sigma^2 = 0.3$.

How does this relate to the prior distribution of $\boldsymbol{\theta}$? Determine the prior variance on $\boldsymbol{\theta}$.

Problem 2 Linear filtering (30% total weighting)

In this problem we are considering the Linear filtering situation as depicted in Figure 1, page 3, where we have a FIR filter. We will use a filter with 3 coefficients ($l = 3$).

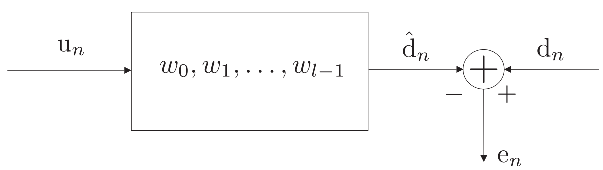

Figure 1: Problem 2 filter setup.

Assume that we have an input sequence to the filter as

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

The filter has the coefficients $\boldsymbol{w} = \begin{bmatrix}0.6 & 1.5 & -1\end{bmatrix}^T$.

Problem 2.1 (5% weighting)

Determine the output value of the filter at time instant $n = 3$.

Problem 2.2 (5% weighting)

Now assume that we use a learning-based setup to recover s_n. As input the filter takes the signal $\mathrm{u}_n = \mathrm{s}_n + \mathrm{x}_n$. Both s_n and x_n follows an AR(1) process. Additionally, we have another signal available, $\mathrm{v}_n = \mathrm{s}_n + \eta_n$, where η_n follows a white-noise sequence.

Derive expressions for the correlation functions needed ($r_u(k)$ and $r_{du}(k)$) expressed in terms of $r_s(k)$, $r_x(k)$ and $r_\eta(k)$ and the associated cross-correlation functions, so that $\boldsymbol{w}$ can be estimated (estimating $\boldsymbol{w}$ is not part of this problem). Only keep the correlation functions that are non-zero in the final expressions for $r_u(k)$ and $r_{du}(k)$.

Problem 2.3 (7% weighting)

We are now informed of the following correlation function values, where $r_u(k)$ denotes the correlation function for the input signal u_n, $r_d(k)$ denotes the correlation function for the desired signal d_n, and $r_{du}(k)$ denotes the cross-correlation function between d_n and u_n

k	0	1	2	3
$r_u(k)$	1.10	0.60	0.20	0.01
$r_{du}(k)$	0.25	0.50	0.30	0.05
$r_d(k)$	0.70	0.30	0.15	0.30

Determine the filter coefficient values $\boldsymbol{w}$ (still with filter length $l = 3$) and the minimum mean squared error (MMSE) as achieved by the filter. You should specify both the exact formulas and the numerical values.

Problem 2.4 (5% weighting)

We will now setup a learning system of the filter using the LMS algorithm. Assume a step size of $\mu = 0.1$, and that the filter weights at time instant $n = 2$ are as originally specified in the problem. The input to the filter is also as originally specified. Assume that $d_3 = 3$.

Determine the new value of the filter coefficients at iteration $n = 3$ when using LMS.

Problem 2.5 (8% weighting)

As a final step, we consider the RLS algorithm, and reduce the filter to size $l = 2$. We assume that the time-varying component of the system is adequately modeled as:

$$\mathrm{d}_n = \boldsymbol{\theta}_{o,n-1}^T \mathrm{u}_n + \eta_n \tag{4a}$$

$$\boldsymbol{\theta}_{o,n} = \boldsymbol{\theta}_{o,n-1} + \omega_n \tag{4b}$$

$$\mathbb{E}\left[\omega_n \omega_n^T\right] = \begin{bmatrix} 0.03 & 0.00 \\ 0.00 & 0.02 \end{bmatrix} \tag{4c}$$

where $\mathbb{E}[\eta_n] = 0$, $\mathbb{E}[\omega_n] = \mathbf{0}$, and $\sigma_\eta^2 = 0.75$. As $r_u(k)$, we assume the same values as previously given.

Determine the steady-state excess MSE when using the RLS algorithm for $\beta = 0.95$. You should specify both the formulas and the numerical value.

Additionally, determine the β value that results in the lowest steady-state excess MSE.

Problem 3 Dictionary learning (10% total weighting)

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

Problem 3.1 (5% weighting)

Suppose that we have a room with four microphones and four sources. We assume that the four sources are stationary (not moving around) and that the assumption of instantaneous mixing is satisfied. Some microphones are equipped with ideal filters that remove all frequency information outside the desired range. The sources has energy in the following spectral range:

Source 1 below 2 kHz, source 2 above 10 kHz, and source 3 and 4 in the range between 2 kHz and 8 kHz.

The source audio is attenuated with 1%-point per meter of travel (e.g. that means 10% attenuation for microphones on a distance of 10 meters). The distances and filter setup is listed in Table 3.1, page 5.

Microphone	d_1	d_2	d_3	d_4	**Filter**
1	1	11	10	11	Lowpass (2KHz)
2	1	11	10	11	None
3	5	10	11	14	Highpass (8Khz)
4	5	14	11	10	None

Table 3.1: d_i denotes the distance the given microphone has to source i.

Write up the mixing matrix that describes how sources are mixed at the four microphones **after** the filtering is applied.

Problem 3.2 (5% weighting)

The ICA solution for mutual information optimizes the mutual information with respect to the sources

$$\hat{W} = \arg\min_{W} I(\mathbf{z}) \tag{5}$$

Where $\mathbf{z}$ is the sources and $\hat{W}$ is the estimated de-mixing matrix. When we use the definition of the mutual information (ML eq. 2.158) in the two-dimensional case, we get

$$I(\mathrm{z}_1; \mathrm{z}_2) := \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(z_1, z_2) \ln \frac{p(z_1, z_2)}{p(z_1)p(z_2)} dz_1 dz_2 \tag{6}$$

where z_i denotes the i'th component of the vector $\mathbf{z}$.

Plugging this expression into our optimization problem, conducting various manipulations, we can arrive at

$$\hat{W} = \arg\min_{W} I(\mathbf{z}) \tag{7a}$$

$$= \arg\max_{W} \ln|\det(W)| + \mathbb{E}\left[\sum_{i=1}^{l} \ln p_i(\mathrm{z}_i)\right] \tag{7b}$$

Which of the following operations is **NOT** carried out in the derivation going from (7a) to (7b):

A: The problem is changed from a minimization problem to a maximization problem.

B: The integral in eq. (6) is rewritten in terms of the entropy of $\mathbf{z}$.

C: A change of variables going from the distribution of the observations to $p_{\mathbf{z}}(\mathbf{z})$.

D: The $\ln|\det(W)|$ term is introduced as regularizer.

E: The expectation is introduced by rewritting the entropy of z_i.

F: Don't know.

Problem 4 Hidden Markov Models (12% total weighting)

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

Problem 4.1 (5% weighting)
Suppose that you have, for a 2-state Hidden Markov Model, the following state sequence (with $N = 37$ elements):

$$x_n = [1111211112222111121111111111111111111] \tag{8}$$

Possible values for the transition probabilities are:

$$P_{ij} \in \{0, 0.08, 0.1, 0.25, 0.5, 0.75, 0.9, 0.92, 1\} \tag{9}$$

What are the most likely transition probabilities used to generate the state sequence? Argue for your choices.

Problem 4.2 (7% weighting)
Now consider the information displayed in Table 4.2, page 6.

Compute the probability of being in state $k = 1$ at time $n = 2$ given the observations $\boldsymbol{y}_{1:5}$.

n	1	2	3	4	5
$P(\boldsymbol{y}_{1:n})$	0.74	0.55	0.12	0.09	0.06
$P(x_n = 1 \vert \boldsymbol{y}_{1:n})$	0.86	0.87	0.74	0.80	0.84
$P(\boldsymbol{y}_{n+1:5} \vert x_n = 1)$	0.09	0.12	0.58	0.77	

Table 4.2: $P(\boldsymbol{y}_{1:n})$ denotes the probability of observing the given sequence $\boldsymbol{y}$ from time 1 to n. $P(x_n = 1|\boldsymbol{y}_{1:n})$ denotes the probability of being in state 1 given the sequence $\boldsymbol{y}$ from time 1 to n. $P(\boldsymbol{y}_{n+1:5}|x_n = 1)$ denotes the probability of observing the sequence $\boldsymbol{y}$ from time $n+1$ to 5, given the chain is in state $x_n = 1$.

Problem 5 Kalman filtering (5% total weighting)

Suppose that we have an auto-regressive process of order 3 defined as

$$\mathrm{s}_n = \sum_{i=1}^{l=3} a_i \mathrm{s}_{n-i} + \xi_n \tag{10}$$

where ξ_n is a white-noise sequence.

Problem 5.1 (5% weighting)

We are not able to observe s_n directly, but only through y_n which measures an attenuated and noisy version of s_n through the relation $\mathrm{y}_n = 0.5\mathrm{s}_n + \epsilon_n$ where ϵ_n is a white noise sequence.

We wish to estimate s_n using Kalman filtering, defined as

$$\mathbf{x}_n = F\mathbf{x}_{n-1} + \boldsymbol{\eta}_n \tag{11a}$$

$$\mathbf{y}_n = H\mathbf{x}_n + \mathbf{v}_n \tag{11b}$$

Setup the required components ($\mathbf{x}_n$, F, $\boldsymbol{\eta}_n$, H, and $\mathbf{v}_n$) such that the Kalman filter algorithm can be used.

Problem 6 Kernel methods (15% total weighting)

In this problem we will consider kernel methods, and we will use two kernel functions, the homogeneous polynomial kernel with $r = 1$ and the Laplacian kernel, denoted κ_p and κ_l respectively

$$\kappa_p(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}^T \boldsymbol{y})^r \tag{12a}$$

$$\kappa_l(\boldsymbol{x}, \boldsymbol{y}) = \exp(-a\|\boldsymbol{x} - \boldsymbol{y}\|) \tag{12b}$$

Problem 6.1 (5% weighting)

Consider the classification problem with the points listed in Figure 2, page 8.

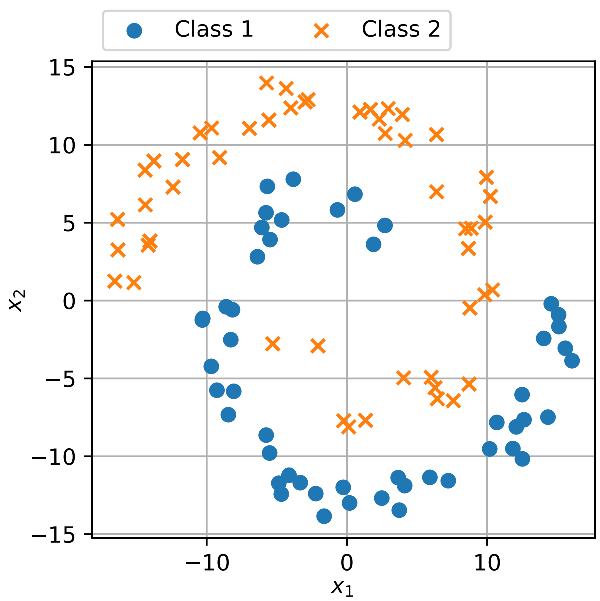

Figure 2: Problem 6.1.

Which of the two kernel functions $\kappa_p(\boldsymbol{x}, \boldsymbol{y})$ and $\kappa_l(\boldsymbol{x}, \boldsymbol{y})$ will be most appropriate?

Explain your decision for the kernel, and relate the solution to the Representer theorem (e.g. what could possible values for the θ_n in the Representer theorem be?)

Problem 6.2 (5% weighting)

Now consider the kernel Ridge regression for a one-dimensional problem. We will use the Laplacian kernel $\kappa_l(\boldsymbol{x}, \boldsymbol{y})$ with $a = 0.5$. Suppose we have five observations, and that we have fitted a Kernel ridge regression problem with $C = 0.01$. The data and data-fit is:

n	1	2	3	4	5
t	0	2	3	6	7
y_n	1	3	2	3	2
θ_n	-0.11	2.82	-0.28	2.48	0.30

where n denotes the point index, t the time, y_n is the value for point n, and θ_n is the coefficient estimated using Kernel Ridge regression associated with point n.

Compute the regression value for $\hat{y}$ at $t = 1$, disregarding all points that have a kernel value lower that 0.1.

Problem 6.3 (5% weighting)

We will now consider a completely different regression problem, with a new set of points. The regression problem listed in Figure 3, page 9 is fitted using Support Vector Regression with the ϵ-insensitive loss with $C = 1$. The support vectors are indicated by a cross.

Identify the ϵ used to create the data-fit.

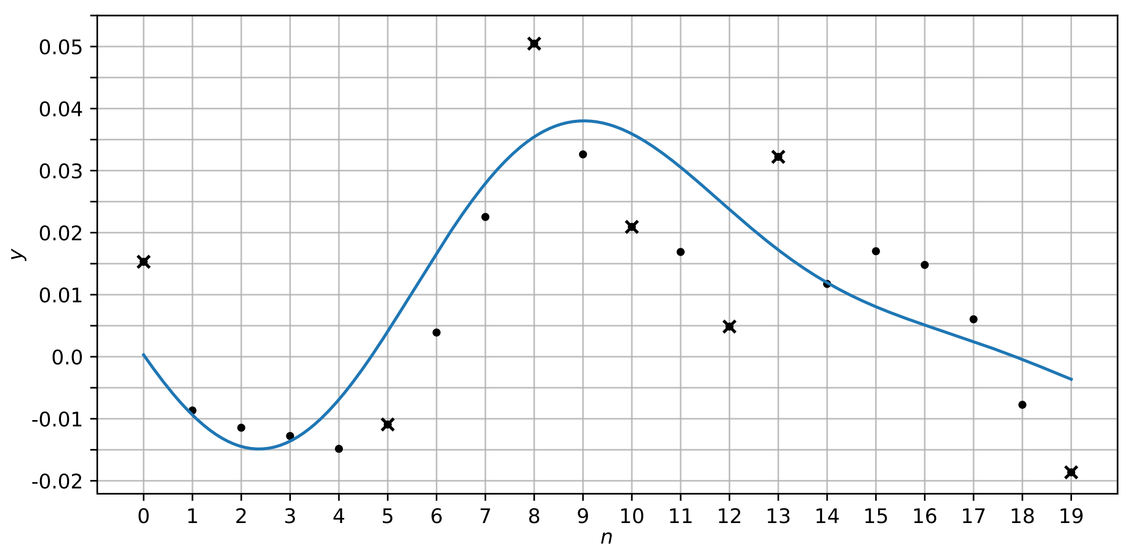

Figure 3: Problem 6.3.