# 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 $\hat{\boldsymbol{\theta}}$.

**Solution:** From ML p 73, eq 3.17, we have

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

Defining $X$ and $\boldsymbol{y}$ as

$$X = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} \quad \boldsymbol{y} = \begin{bmatrix} -0.4 \\ 0.3 \\ 1.8 \end{bmatrix}$$

Gives

$$\hat{\boldsymbol{\theta}} = \begin{bmatrix} -0.53 \\ 1.10 \end{bmatrix}$$

**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 $\lambda$, and determine which value for $\lambda$ was used with 2 significant digits.

[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.

**Solution:** Assume again the same observations as in the previous problem. We will now The solution for Ridge regression is defined in ML eq (6.29)

$$\begin{aligned}\hat{\boldsymbol{\theta}}_R &= (X^TX+\lambda I)^{-1}X^T\boldsymbol{y} \qquad \Rightarrow\\ (X^TX+\lambda I)\hat{\boldsymbol{\theta}}_R &= X^T\boldsymbol{y} \qquad \Rightarrow\\ X^TX\hat{\boldsymbol{\theta}}_R+\lambda\hat{\boldsymbol{\theta}}_R &= X^T\boldsymbol{y} \qquad \Rightarrow\\ \lambda\hat{\boldsymbol{\theta}}_R &= X^T\boldsymbol{y}-X^TX\hat{\boldsymbol{\theta}}_R \qquad \Rightarrow\\ \lambda\hat{\boldsymbol{\theta}}_R^T\hat{\boldsymbol{\theta}}_R &= \hat{\boldsymbol{\theta}}_R^T(X^T\boldsymbol{y}-X^TX\hat{\boldsymbol{\theta}}_R) \qquad \Rightarrow\\ \lambda\frac{\hat{\boldsymbol{\theta}}_R^T\hat{\boldsymbol{\theta}}_R}{\hat{\boldsymbol{\theta}}_R^T\hat{\boldsymbol{\theta}}_R} &= \frac{\hat{\boldsymbol{\theta}}_R^T(X^T\boldsymbol{y}-X^TX\hat{\boldsymbol{\theta}}_R)}{\hat{\boldsymbol{\theta}}_R^T\hat{\boldsymbol{\theta}}_R} \qquad \Rightarrow\\ \lambda &= \frac{\hat{\boldsymbol{\theta}}_R^T(X^T\boldsymbol{y}-X^TX\hat{\boldsymbol{\theta}}_R)}{\hat{\boldsymbol{\theta}}_R^T\hat{\boldsymbol{\theta}}_R}\end{aligned}$$

Inserting the values gives $\lambda = 0.14$

**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 $\hat{\boldsymbol{\theta}}_R$.

**Solution:** From ML sec 9.3, eq 9.9 we almost have the solution. However this assumes $X^TX = I$, so we need to change it slightly. We solve by direct substitution. If $X^TX = 2I$ we get

$$\hat{\boldsymbol{\theta}}_{\text{LS}} = \left(X^TX\right)^{-1}X^T\boldsymbol{y} = (2I)^{-1}X^T\boldsymbol{y} = \frac{1}{2}X^T\boldsymbol{y}$$

For Ridge regression we get

$$\begin{aligned}\hat{\boldsymbol{\theta}}_R &= \left(X^TX+\lambda I\right)^{-1}X^T\boldsymbol{y}\\ &= (2I+\lambda I)^{-1}X^T\boldsymbol{y}\\ &= (I(2+\lambda))^{-1}X^T\boldsymbol{y}\\ &= \frac{1}{2+\lambda}X^T\boldsymbol{y}\\ &= \frac{2}{2+\lambda}\frac{1}{2}X^T\boldsymbol{y}\\ &= \frac{2}{2+\lambda}\hat{\boldsymbol{\theta}}_{\text{LS}}\end{aligned}$$

Inserting the given numbers, we get

$$\hat{\boldsymbol{\theta}}_R = \begin{bmatrix} -0.6 \\ 1.04 \end{bmatrix}$$

**Problem 1.4** (5% weighting)
Suppose now we apply $\ell_1$ 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:

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

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 $\mu$ that results in exactly one component of $\boldsymbol{\theta}^{(i)}$ being zero.

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

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

We complete the iteration as

$$\begin{aligned}\boldsymbol{\theta}^{(i)} &= S_{\lambda\mu}\left(\boldsymbol{\theta}^{(i-1)} + \mu X^T\boldsymbol{e}^{(i-1)}\right) \\ &= S_{\lambda\mu}\left(\begin{bmatrix} -0.75 \\ 1.3 \end{bmatrix} + \mu\begin{bmatrix} 0.25 \\ -0.35 \end{bmatrix}\right)\end{aligned}$$

Since we need to find the minimum value for $\mu$ that gives exactly one component being zero, we get

$$\begin{aligned}|-0.75 + \mu 0.25| &= \lambda\mu \\ |1.3 - \mu 0.35| &= \lambda\mu\end{aligned}$$

For $\lambda = 0.5$, we get

$$\begin{aligned}|-0.75 + \mu 0.25| = 0.5\mu \quad &\Rightarrow \mu = \frac{0.75}{0.25 + 0.5} = 1 \\ |1.3 - \mu 0.35| = 0.5\mu \quad &\Rightarrow \mu = \frac{1.3}{0.35 + 0.5} = 1.53\end{aligned}$$

Hence, the lowest $\mu$ is $\mu = 1$.

**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 $\boldsymbol{\theta}$.

**Solution:** From ML, sec 12.2.2, eq 12.12, we have

$$\boldsymbol{\theta}_{MAP} = (\lambda I + \Phi^T \Phi)^{-1} \Phi^T \boldsymbol{y}$$

where, if $\lambda = \frac{\sigma_\eta^2}{\sigma_{\boldsymbol{\theta}}^2}$, and $\Phi = X$, the solution is equivalent to Ridge regression. As stated in ML sec 12.2.2, eq (12.8), the distribution of $\boldsymbol{\theta}$ is assumed to follow a normal distribution with zero-mean and variance $\sigma_{\boldsymbol{\theta}}^2$.

That means there is a direct relation between $\lambda$, the measurement noise $\sigma_\eta^2$ and the variance on the prior $\sigma_{\boldsymbol{\theta}}^2$.

Since the measurement noise is 0.3, we have $\sigma_{\boldsymbol{\theta}}^2 = \frac{\sigma_\eta^2}{\lambda} = 0.6$.

## Problem 2 Linear filtering (30% total weighting)

In this problem we are considering the Linear filtering situation as depicted in Figure 1, page 5, where we have a FIR filter. We will use a filter with 3 coefficients ($l = 3$).

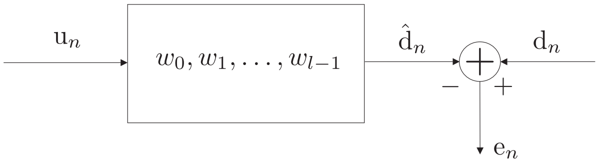

Figure 1: Problem 2 filter setup.

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 $n = 3$.

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

$$\begin{aligned}
\hat{d}_n &= \sum_{i=0}^{l-1} w_i u_{n-i} \qquad \Rightarrow \\
\hat{d}_3 &= \sum_{i=0}^{2} w_i u_{3-i} \\
&= w_0 u_3 + w_1 u_2 + w_2 u_1 \\
&= 0.6 \cdot 5 + 1.5 \cdot 1 - 1 \cdot 2 \\
&= 2.5
\end{aligned}$$

**Problem 2.2** (5% weighting)

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

**Solution:** According to ML p 136, we need to determine the input autocorrelation function $r_u(k) = \mathbb{E}[\mathrm{u}_n \mathrm{u}_{n-k}]$ and cross-correlation function between the input signal and desired signal, $r_{du}(k) = \mathbb{E}[\mathrm{d}_n \mathrm{u}_{n-k}]$, where $\mathrm{u}_n$ denotes the input signal and $\mathrm{d}_n$ denotes the desired signal. In this case, $\mathrm{u}_n = \mathrm{s}_n + \mathrm{x}_n$ and $\mathrm{d}_n = \mathrm{s}_n + \eta_n$. Thus we get

$$\begin{aligned}
r_u(k) &= \mathbb{E}[\mathrm{u}_n \mathrm{u}_{n-k}] \\
&= \mathbb{E}[(\mathrm{s}_n + \mathrm{x}_n)(\mathrm{s}_{n-k} + \mathrm{x}_{n-k})] \\
&= \mathbb{E}[\mathrm{s}_n \mathrm{s}_{n-k} + \mathrm{s}_n \mathrm{x}_{n-k} + \mathrm{x}_n \mathrm{s}_{n-k} + \mathrm{x}_n \mathrm{x}_{n-k}] \\
&= r_s(k) + r_{sx}(k) + r_{xs}(k) + r_x(k)
\end{aligned}$$

Since AR processes are generated using white noise (ML sec 2.4.4, eq 2.128), they can be assumed uncorrelated and thus $r_{sx}(k) = 0$ and $r_{xs}(k) = 0$, hence

$$r_u(k) = r_s(k) + r_x(k)$$

For cross-correlation, we get

$$\begin{aligned}
r_{du}(k) &= \mathbb{E}[\mathrm{d}_n \mathrm{u}_{n-k}] \\
&= \mathbb{E}[(\mathrm{s}_n + \eta_n)(\mathrm{s}_{n-k} + \mathrm{x}_{n-k})] \\
&= \mathbb{E}[\mathrm{s}_n \mathrm{s}_{n-k} + \mathrm{s}_n \mathrm{x}_{n-k} + \eta_n \mathrm{s}_{n-k} + \eta_n \mathrm{x}_{n-k}]
\end{aligned}$$

Since $\eta_n$ is a white-noise sequence with zero mean, the terms with $\eta_n$ vanish. Additionally, $r_{sx}(k) = 0$ so we get

$$r_{du}(k) = r_s(k)$$

**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 $\mathrm{u}_n$, $r_d(k)$ denotes the correlation function for the desired signal $\mathrm{d}_n$, and $r_{du}(k)$ denotes the cross-correlation function between $\mathrm{d}_n$ and $\mathrm{u}_n$

| $k$ | 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.

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

$$\begin{aligned}
\begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix} &= \begin{bmatrix} r_u(0) & r_u(1) & r_u(2) \\ r_u(1) & r_u(0) & r_u(1) \\ r_u(2) & r_u(1) & r_u(0) \end{bmatrix}^{-1} \begin{bmatrix} r_{du}(0) \\ r_{du}(1) \\ r_{du}(2) \end{bmatrix} \\
&= \begin{bmatrix} 1.1 & 0.6 & 0.2 \\ 0.6 & 1.1 & 0.6 \\ 0.2 & 0.6 & 1.1 \end{bmatrix}^{-1} \begin{bmatrix} 0.25 \\ 0.50 \\ 0.30 \end{bmatrix} \\
&= \begin{bmatrix} -0.02 \\ 0.45 \\ 0.03 \end{bmatrix}
\end{aligned}$$

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

$$\begin{aligned}
J(\boldsymbol{\theta}) &= \sigma_y^2 - \boldsymbol{p}^T \Sigma_u^{-1} \boldsymbol{p} \\
&= r_d(0) - \begin{bmatrix} 0.25 \\ 0.50 \\ 0.30 \end{bmatrix}^T \begin{bmatrix} 1.1 & 0.6 & 0.2 \\ 0.6 & 1.1 & 0.6 \\ 0.2 & 0.6 & 1.1 \end{bmatrix}^{-1} \begin{bmatrix} 0.25 \\ 0.50 \\ 0.30 \end{bmatrix} \\
&= 0.47
\end{aligned}$$

**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.

**Solution:** According to ML sec 5.5, algorithm 5.1, we have

$$\begin{aligned} e_n &= y_n - \boldsymbol{\theta}_{n-1}^T \boldsymbol{x}_n \\ \boldsymbol{\theta}_n &= \boldsymbol{\theta}_{n-1} + \mu e_n \boldsymbol{x}_n \end{aligned}$$

In this setup, we have

$$\begin{aligned} y_n &= 3 \\ \boldsymbol{\theta}_{n-1} &= \begin{bmatrix} 0.6 & 1.5 & -1 \end{bmatrix}^T \\ \boldsymbol{x}_n &= \begin{bmatrix} 5 & 1 & 2 \end{bmatrix}^T \end{aligned}$$

Pluggin into the equation yields

$$\begin{aligned} e_n &= 3 - \begin{bmatrix} 0.6 & 1.5 & -1 \end{bmatrix} \begin{bmatrix} 5 & 1 & 2 \end{bmatrix}^T \\ &= 0.5 \\ \boldsymbol{\theta}_n &= \begin{bmatrix} 0.6 \\ 1.5 \\ -1 \end{bmatrix} + 0.1 \cdot 0.5 \cdot \begin{bmatrix} 5 \\ 1 \\ 2 \end{bmatrix} \\ &= \begin{bmatrix} 0.85 \\ 1.55 \\ -0.90 \end{bmatrix} \end{aligned}$$

**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 $\beta$ value that results in the lowest steady-state excess MSE.

**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

$$\begin{aligned} J_{MSE} &= \frac{1}{2}(1-\beta)\sigma_\eta^2 l + \frac{1}{2}(1-\beta)^{-1}\text{trace}(\Sigma_\omega \Sigma_u) \\ &= \frac{1}{2}(1-0.95)\cdot 0.75 \cdot 2 + \frac{1}{2}(1-0.95)^{-1}\text{trace}\left(\begin{bmatrix} 0.02 & 0.00 \\ 0.00 & 0.03 \end{bmatrix}\begin{bmatrix} 1.1 & 0.6 \\ 0.6 & 1.1 \end{bmatrix}\right) \\ &= 0.59 \end{aligned}$$

To determine the optimal value for $\beta$, we use the formula given on page 277,

$$\begin{aligned} \beta_{opt} &= 1 - \sqrt{\frac{\text{trace}(\Sigma_\omega \Sigma_u)}{\sigma_\eta^2 l}} \\ &= 0.81 \end{aligned}$$

As a check (not part of the problem), this gives a $J_{MSE} = 0.29$ which is smaller than the $J_{MSE}$ for $\beta = 0.95$.

## 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 10.

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

| **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 $i$.

**Solution:** ICA assumes linear mixing, this we have

$$\mathbf{x} = A\mathbf{s}$$

According to the description, the signals arrive at the microphone with an amplitude 1%-point per meter of travel. microphone 1 only picks up source 1, and microphone 3 only picks up microphone 2. The other microphones pick up everything, thus we get

$$A = \begin{bmatrix} 0.99 & 0 & 0 & 0 \\ 0.99 & 0.89 & 0.90 & 0.89 \\ 0 & 0.9 & 0 & 0 \\ 0.95 & 0.86 & 0.89 & 0.90 \end{bmatrix}$$

**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 $\mathrm{z}_i$.
**F**: Don't know.

**Solution:** Correct answer is D.

$\ln|\det(W)|$ is not introduced as regularizer, but is introduced as part of the change of variables.

## 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.

**Solution:** We make estimates based on the frequencies. The chain is in state 1 for 30 counts (denoted $N_1 = 30$), and state 2 for 6 counts (denoted $N_2 = 6$).

| **state** | **counts** | **probability** |
|---|---|---|
| $1 \to 1$ | 27 | $27/30 = 0.9$ |
| $1 \to 2$ | 3 | $3/30 = 0.1$ |
| $2 \to 1$ | 3 | $3/6 = 0.5$ |
| $2 \to 2$ | 3 | $3/6 = 0.5$ |

**Problem 4.2** (7% weighting)

Now consider the information displayed in Table 4.2, page 12.

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 $x_n = 1$.

**Solution:** We can solve this either by Bayes theorem, or according to ML eq (16.49), combined with eq (16.47) and eq (16.43). We have (using the information from ML sec 16.5)

$$\begin{aligned} P(\boldsymbol{x}_n|\boldsymbol{y}_{[1:N]}) &= \frac{\alpha(\boldsymbol{x}_n)\beta(\boldsymbol{x}_n)}{P(\boldsymbol{y}_{[1:N]})} \\ &= \frac{\alpha(\boldsymbol{x}_n)P(\boldsymbol{y}_{[n+1:N]}|\boldsymbol{x}_n)}{P(\boldsymbol{y}_{[1:N]})} \\ &= \frac{P(\boldsymbol{x}_n|\boldsymbol{y}_{[1:n]})P(\boldsymbol{y}_{[1:n]})P(\boldsymbol{y}_{[n+1:N]}|\boldsymbol{x}_n)}{P(\boldsymbol{y}_{[1:N]})} \end{aligned}$$

Plugging the numbers from the table into the equation, we get

$$\begin{aligned} P(x_2 = 1|\boldsymbol{y}_{[1:N]}) &= \frac{P(x_2 = 1|\boldsymbol{y}_{[1:2]})P(\boldsymbol{y}_{[1:2]})P(\boldsymbol{y}_{[3:5]}|x_2 = 1)}{P(\boldsymbol{y}_{[1:5]})} \\ &= \frac{0.87 \cdot 0.55 \cdot 0.12}{0.06} \\ &= 0.96 \end{aligned}$$

## 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 $\xi_n$ is a white-noise sequence.

**Problem 5.1** (5% weighting)

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

We wish to estimate $\mathrm{s}_n$ 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.

**Solution:** This problem is readily solved using the example 4.4 in ML p 171. For a state-space model we have

$$\begin{aligned} \mathbf{x}_n &= F\mathbf{x}_{n-1} + \boldsymbol{\eta}_n \\ \mathbf{y}_n &= H\mathbf{x}_n + \mathbf{v}_n \end{aligned}$$

In our case, we have $\mathbf{x}_n$ as a 3-dimensional vector with the elements from time instant $n$ to $n-2$, and $\mathbf{y}_n$ is a scalar instead of a vector. So we get

$$\mathbf{x}_n = \begin{bmatrix} \mathrm{s}_n \\ \mathrm{s}_{n-1} \\ \mathrm{s}_{n-2} \end{bmatrix}, \qquad F = \begin{bmatrix} a_1 & a_2 & a_3 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \qquad \boldsymbol{\eta}_n = \begin{bmatrix} \xi_n \\ 0 \\ 0 \end{bmatrix}$$

$$H = \begin{bmatrix} 0.5 & 0 & 0 \end{bmatrix}, \qquad \mathbf{v}_n = \epsilon_n$$

## 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 $\kappa_p$ and $\kappa_l$ 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 14.

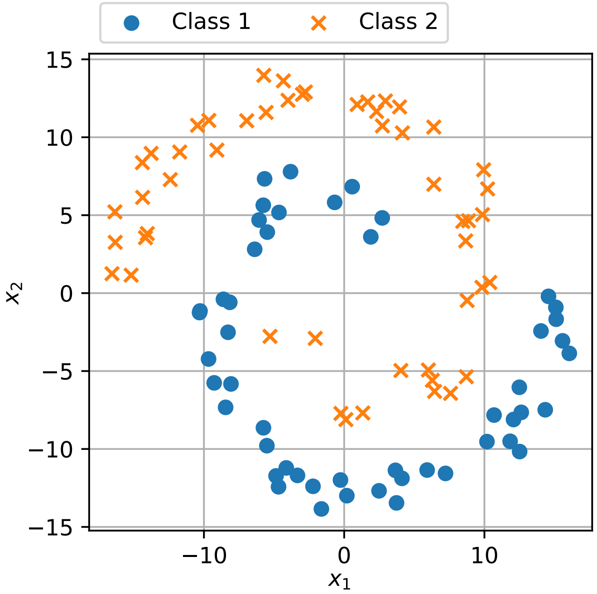

Figure 2: Problem 6.1.

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 $\theta_n$ in the Representer theorem be?)

**Solution:** The Laplace kernel can solve this problem much better than the polynomial kernel, as this kernel measures the distance between points. An example of the kernel matrix for the two kernel functions is depicted on Figure 3, page 15.

Relating this result to the Representer theorem;

$$f(\boldsymbol{x}_{test}) = \sum_{n=1}^{N} \theta_n \kappa(\boldsymbol{x}_{test}, \boldsymbol{x}_n)$$

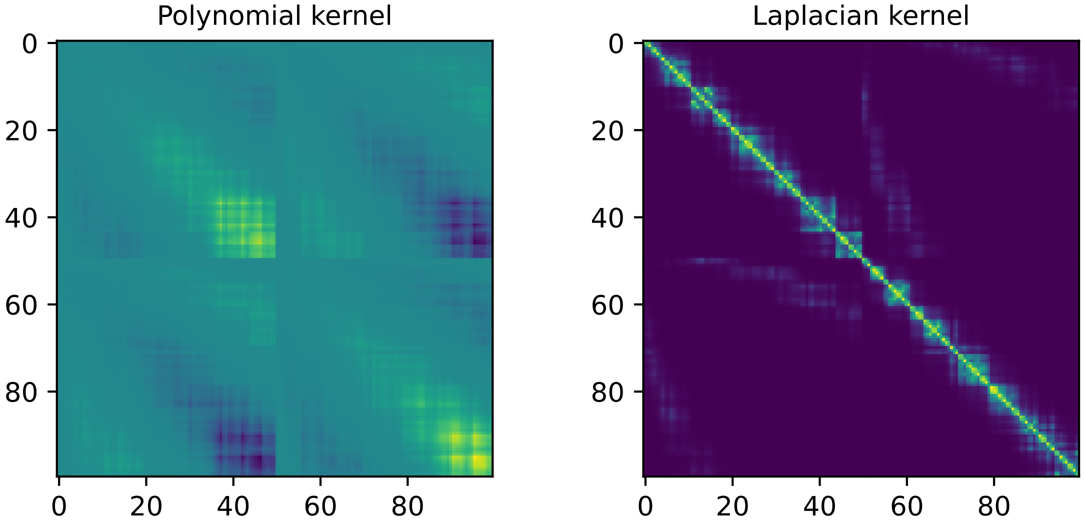

Figure 3: Problem 6.1 solution. The points are first sorted according to the class label, and there after sorted according to the angle relative to the x-axis. This makes the kernel matrix easy to interpret in terms of the block-structure.

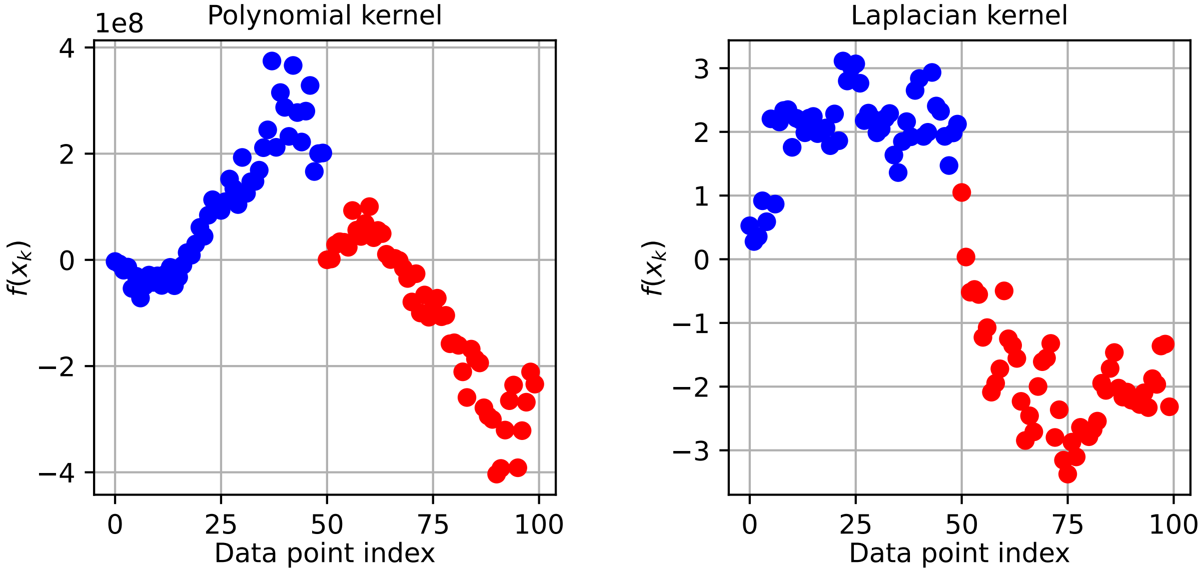

Figure 4: Problem 6.1 solution. Plots of the Representer theorem value for the data points computed using leave-out-out.

we can e.g. have all $\theta_n$ associated with class one have a value of 1, and $\theta_n$ associated with class two have the value -1. Then the class decision boundary could naturally be $f(\boldsymbol{x}_{test}) > 0$ classifies as class one, and $f(\boldsymbol{x}_{test}) < 0$ classifies as class two. This is illustrated on figure Figure 4, page 15

**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 |
| $\theta_n$ | -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 $\theta_n$ 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.

**Solution:** To determine the values that should be included in the calculation, we solve for $\|x - y\|$

$$\begin{aligned}
\kappa(x, y) &= \exp(-a\|x - y\|) \qquad \Rightarrow \\
\ln \kappa(x, y) &= -a\|x - y\| \qquad \Rightarrow \\
\|x - y\| &= \frac{\ln \kappa(x, y)}{-a} \\
&= \frac{\ln 0.1}{-0.5} \\
&= 4.61
\end{aligned}$$

Since we are calculating for $t = 1$, we can ignore all points above $t > 5.61$.

According to ML sec 11.7, p 553, we have

$$\begin{aligned}
\hat{y}_1 &= \sum_n \hat{\theta}_n \kappa(1, n) \\
&= \theta_1 \kappa(1, 0) + \theta_2 \kappa(1, 2) + \theta_3 \kappa(1, 3) \\
\kappa(1, 0) &= \exp(-0.5\|1\|) = 0.61 \\
\kappa(1, 2) &= \exp(-0.5\|1\|) = 0.61 \\
\kappa(1, 3) &= \exp(-0.5\|2\|) = 0.37 \\
\hat{y}_1 &= -0.11\kappa(1, 0) + 2.82\kappa(1, 2) + -0.28\kappa(1, 3) \\
&= 1.54
\end{aligned}$$

**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 5, page 18 is fitted using Support Vector Regression with the $\epsilon$-insensitive loss with $C = 1$. The support vectors are indicated by a cross.

Identify the $\epsilon$ used to create the data-fit.

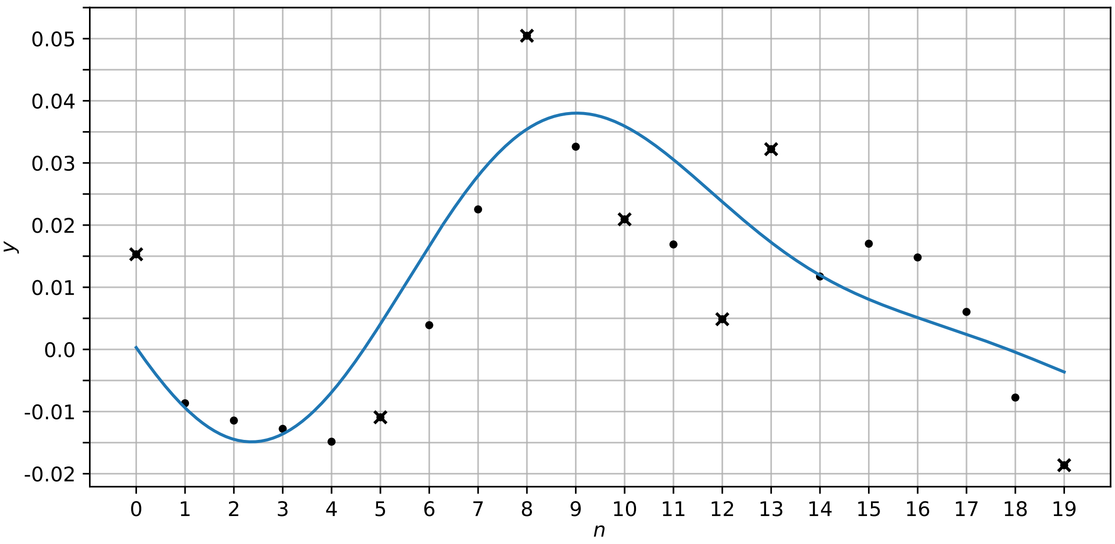

Figure 5: Problem 6.3.

**Solution:** From ML sec 11.8, p 559-560 (and in particular figure 11.11), we know that support vectors are either on the epsilon tube or outside the epsilon tube. This, the smallest error between the data fit and the support vector must be the epsilon value used. Going through the points, we can see that no support vector has an error smaller than 0.015, hence $\epsilon = 0.015$.

For illustration purpose, the solution with epsilon tube is drawn in Figure 6, page 19.

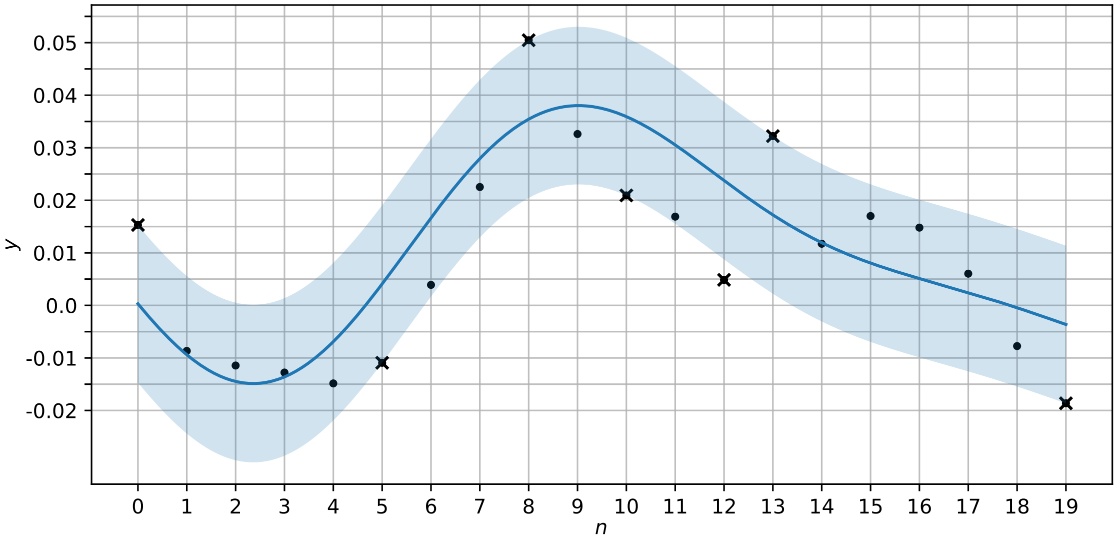

Figure 6: Problem 6.3 solution.