02471 Machine Learning for Signal Processing

Solution

Exercise 2: Parameter Estimation

2.1 Linear Models

Exercise 2.1.2

The book uses column vectors, so we get the following dimensions for the vectors: $\boldsymbol{y} \in \mathbb{R}^{N\times 1}$, $\boldsymbol{\theta} \in \mathbb{R}^{(l+1)\times 1}$ where l is the number of dimensions in the input data. X then neeeds to be a $\mathbb{R}^{N\times(l+1)}$ matrix. We define these as

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \qquad X = \begin{bmatrix} x_{1,1} & \cdots & x_{1,l} & 1 \\ x_{2,1} & \cdots & x_{2,l} & 1 \\ \vdots & \ddots & \vdots & \vdots \\ x_{N,1} & \cdots & x_{N,l} & 1 \end{bmatrix}$$

If we write the sum for the first data-point we get:

$$\begin{aligned} J(\boldsymbol{\theta}) &= \sum_{n=1}^{N=1} (y_n - \boldsymbol{\theta}_b^T \boldsymbol{x}_n - \theta_0)^2 \\ &= (y_1 - \theta_1 x_{1,1} - \cdots - \theta_l x_{1,l} - \theta_0)^2 \end{aligned}$$

Similarly, the first component of the vector $(\boldsymbol{y} - X\boldsymbol{\theta})$ is

$$y_1 - \theta_1 x_{1,1} - \cdots - \theta_l x_{1,l} - \theta_0 \cdot 1$$

Since the inner product of a vector is the sum of all the components squared, we have shown the relation.

Exercise 2.1.3

We first make the following rewrites:

$$\begin{aligned} J(\boldsymbol{\theta}) &= (\boldsymbol{y} - X\boldsymbol{\theta})^T (\boldsymbol{y} - X\boldsymbol{\theta}) \\ &= (\boldsymbol{y}^T - (X\boldsymbol{\theta})^T)(\boldsymbol{y} - X\boldsymbol{\theta}) \\ &= \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T X\boldsymbol{\theta} - (X\boldsymbol{\theta})^T\boldsymbol{y} + (X\boldsymbol{\theta})^T X\boldsymbol{\theta} \\ &= \boldsymbol{y}^T\boldsymbol{y} - (X\boldsymbol{\theta})^T\boldsymbol{y} - (X\boldsymbol{\theta})^T\boldsymbol{y} + \boldsymbol{\theta}^T X^T X\boldsymbol{\theta} \\ &= \boldsymbol{y}^T\boldsymbol{y} - 2(X\boldsymbol{\theta})^T\boldsymbol{y} + \boldsymbol{\theta}^T X^T X\boldsymbol{\theta} \end{aligned}$$

We can now use the following rules from appendix A

$$\frac{\partial \boldsymbol{a}^T\boldsymbol{x}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{x}^T\boldsymbol{a}}{\partial \boldsymbol{x}} = \boldsymbol{a}$$

$$\frac{\partial \boldsymbol{x}^T A\boldsymbol{x}}{\partial \boldsymbol{x}} = \left(A + A^T\right)\boldsymbol{x}$$

We get

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\theta}} J(\boldsymbol{\theta}) &= -\frac{\partial}{\partial \boldsymbol{\theta}} 2(X\boldsymbol{\theta})^T \boldsymbol{y} + \frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\theta}^T X^T X \boldsymbol{\theta} \\ &= -2\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\theta}^T X^T \boldsymbol{y} + (X^T X + (X^T X)^T)\boldsymbol{\theta} \\ &= -2X^T \boldsymbol{y} + 2X^T X \boldsymbol{\theta}\end{aligned}$$

Note that $X^T \boldsymbol{y}$ results in a $(l + 1 \times 1)$ sized vector, so $\boldsymbol{\theta}^T X^T \boldsymbol{y}$ is indeed an inner product between two vectors.

2.2 Estimation

Exercise 2.2.1

Assume the model $\mathrm{y} = g(\mathbf{x}) + \eta$, then we have (since η is zero-mean, and $\boldsymbol{x}$ is observed, that is $\mathbf{x} = \boldsymbol{x}$), then

$$\begin{aligned}\mathbb{E}[\mathrm{y}|\boldsymbol{x}] &= \mathbb{E}[g(\boldsymbol{x}) + \eta] \\ &= g(\boldsymbol{x}) + \mathbb{E}[\eta] \\ &= g(\boldsymbol{x}) \\ \mathrm{MSE} &= \mathbb{E}\big[(\mathrm{y} - \mathbb{E}[\mathrm{y}|\boldsymbol{x}])^2\big] \\ &= \mathbb{E}\big[(g(\boldsymbol{x}) + \eta - \mathbb{E}[\mathrm{y}|\boldsymbol{x}])^2\big] \\ &= \mathbb{E}\big[(g(\boldsymbol{x}) + \eta - g(\boldsymbol{x})^2\big] \\ &= \mathbb{E}\big[\eta^2\big]\end{aligned}$$

The variance is defined as $\mathrm{var}[\eta] = \mathbb{E}[(\eta - \mathbb{E}[\eta])^2]$, so for a zero mean variable we have $\mathrm{var}[\eta] = \mathbb{E}[\eta^2]$. Hence we get the result

$$\mathrm{MSE} = \mathbb{E}\big[\eta^2\big] = \sigma_\eta^2$$

Exercise 2.2.2

Since we are dealing with unbiased estimator, we know that:

$$\mathbb{E}[\hat{\theta}_i] = \boldsymbol{\theta}$$

We also know that estimators are uncorrelated and all have the variance:

$$\sigma^2 = \mathbb{E}\big[(\theta_i - \boldsymbol{\theta}_o)^T(\theta_i - \boldsymbol{\theta}_o)\big]$$

$$\hat{\theta} = \frac{1}{m}\sum_{i=1}^{m} \hat{\theta}_i$$

Putting this together yields:

$$\mathbb{E}[\hat{\theta}] = \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}[\hat{\theta}_i] = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\theta} = \boldsymbol{\theta}$$

Next, assuming that estimators are uncorrelated, meaning:

$$\mathbb{E}\Big[(\hat{\theta}_i - \boldsymbol{\theta}_o)^T(\hat{\theta}_j - \boldsymbol{\theta}_o)\Big] = \sigma^2\delta_{ij}, \text{ where } \delta_{ij} = 1 \text{ when } i = j \text{ and zero otherwise.}$$

Use substitution to obtain:

$$\begin{aligned}
\sigma_c^2 &= \mathbb{E}\Big[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)^T(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)\Big] \\
&= \mathbb{E}\left[\left(\frac{1}{m}\sum_{i=1}^{m}\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_o\right)^T\left(\frac{1}{m}\sum_{j=1}^{m}\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_o\right)\right] \\
&= \mathbb{E}\left[\frac{1}{m^2}\left(\sum_{i=1}^{m}\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_o\right)^T\left(\sum_{j=1}^{m}\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_o\right)\right] \\
&= \mathbb{E}\left[\frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\Big(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_o\Big)^T\Big(\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_o\Big)\right] \\
&= \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\mathbb{E}\Big[(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_o)^T(\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_o)\Big] \\
&= \frac{1}{m^2}m\sigma^2 \\
&= \frac{1}{m}\sigma^2
\end{aligned}$$

Exercise 2.2.3

We know that $\hat{\theta}_u$ is an unbiased estimator, so that $\mathbb{E}[\hat{\theta}_u] = \theta_0$. For this exercise, the biased estimator is defined as $\hat{\theta}_b := (1+\alpha)\hat{\theta}_u$. Further, we assume that $\text{MSE}(\hat{\theta}_b) > 0$ (this is only zero if we have zero noise and a perfect fit) and that $\theta_o > 0$ (since $\theta_o = 0$ is the trivial case of a null-fit).

First calculate the MSE of biased estimator:

$$\begin{aligned}
\text{MSE}(\hat{\theta}_b) &= \mathbb{E}\Big[(\hat{\theta}_b - \theta_o)^2\Big] \\
&= \mathbb{E}\Big[((1+\alpha)\hat{\theta}_u - \theta_o)^2\Big]
\end{aligned}$$

The trick is to add and subtract term $\alpha\theta_o$:

$$\begin{aligned}
\text{MSE}(\hat{\theta}_b) &= \mathbb{E}\Big[((1+\alpha)\hat{\theta}_u + \alpha\theta_o - \alpha\theta_o - \theta_o)^2\Big] \\
&= \mathbb{E}\Big[((1+\alpha)\hat{\theta}_u + \alpha\theta_o - \theta_o(\alpha+1))^2\Big] \\
&= \mathbb{E}\Big[((1+\alpha)(\hat{\theta}_u - \theta_o) + \alpha\theta_o)^2\Big] \\
&= \mathbb{E}\Big[(1+\alpha)^2(\hat{\theta}_u - \theta_o)^2 + \alpha^2\theta_o^2 + 2(1+\alpha)(\hat{\theta}_u - \theta_o)\alpha\theta_o\Big]
\end{aligned}$$

Since α and θ_o are deterministic, we can narrow the scope of the expectations (expectations only needs to be taken wrt random variables):

$$\text{MSE}(\hat{\theta}_b) = (1+\alpha)^2\mathbb{E}\Big[(\hat{\theta}_u - \theta_o)^2\Big] + \alpha^2\theta_o^2 + 2\alpha(1+\alpha)(\mathbb{E}[\hat{\theta}_u] - \theta_o)\theta_o$$

Taking into account that $\mathbb{E}\Big[(\hat{\theta}_u - \theta_o)^2\Big] = \text{MSE}(\hat{\theta}_u)$ and $\mathbb{E}[\hat{\theta}_u] = \theta_o$, we end up with:

$$\begin{aligned}
\text{MSE}(\hat{\theta}_b) &= (1+\alpha)^2\text{MSE}(\hat{\theta}_u) + \alpha^2\theta_o^2 + 2\alpha(1+\alpha)(\theta_o - \theta_o)\theta_o \\
&= (1+\alpha)^2\text{MSE}(\hat{\theta}_u) + \alpha^2\theta_o^2
\end{aligned}$$

Now we have an expression for $\text{MSE}(\hat{\theta}_b)$. Next we seek the solution for α, so that

$$\text{MSE}(\hat{\theta}_b) < \text{MSE}(\hat{\theta}_u)$$

By substitution, we get:

$$\begin{aligned}
(1+\alpha)^2\text{MSE}(\hat{\theta}_u) + \alpha^2\theta_o^2 &< \text{MSE}(\hat{\theta}_u) \\
\Rightarrow (1+\alpha^2+2\alpha)\text{MSE}(\hat{\theta}_u) + \alpha^2\theta_o^2 &< \text{MSE}(\hat{\theta}_u) \\
\Rightarrow \text{MSE}(\hat{\theta}_u) + \alpha^2\text{MSE}(\hat{\theta}_u) + 2\alpha\text{MSE}(\hat{\theta}_u) + \alpha^2\theta_o^2 &< \text{MSE}(\hat{\theta}_u) \\
\Rightarrow \alpha^2\text{MSE}(\hat{\theta}_u) + 2\alpha\text{MSE}(\hat{\theta}_u) + \alpha^2\theta_o^2 &< 0 \\
\Rightarrow \alpha\Big(\alpha\text{MSE}(\hat{\theta}_u) + 2\text{MSE}(\hat{\theta}_u) + \alpha\theta_o^2\Big) &< 0
\end{aligned}$$

If we multiply both sides by $\frac{1}{\theta_o^2+\text{MSE}(\hat{\theta}_u)}$ (which is a positive quantity), we get:

$$\begin{aligned}
\Rightarrow \frac{\alpha\Big(\alpha\text{MSE}(\hat{\theta}_u) + 2\text{MSE}(\hat{\theta}_u) + \alpha\theta_o^2\Big)}{\theta_o^2 + \text{MSE}(\hat{\theta}_u)} &< 0 \\
\Rightarrow \alpha\left(\frac{2\text{MSE}(\hat{\theta}_u)}{\theta_o^2 + \text{MSE}(\hat{\theta}_u)} + \alpha\frac{\theta_o^2 + \text{MSE}(\hat{\theta}_u)}{\theta_o^2 + \text{MSE}(\hat{\theta}_u)}\right) &< 0 \\
\Rightarrow \alpha\left(\frac{2\text{MSE}(\hat{\theta}_u)}{\theta_o^2 + \text{MSE}(\hat{\theta}_u)} + \alpha\right) &< 0
\end{aligned}$$

To proceed, we need to handle the three cases of α: $\alpha < 0$, $\alpha > 0$, and $\alpha = 0$. The case where $\alpha = 0$, does not really make sense to consider, since this means our biased estimator will be defined as our unbiased estimator and no bias is then induced.

For the case $\alpha > 0$: here the biased estimator "expands" $\hat{\theta}_u$ (by noting the definition of $\hat{\theta}_b$), and we get:

$$\begin{aligned}
\Rightarrow \alpha\left(\frac{2\text{MSE}(\hat{\theta}_u)}{\theta_o^2 + \text{MSE}(\hat{\theta}_u)} + \alpha\right) &< 0 \\
\Rightarrow \frac{2\text{MSE}(\hat{\theta}_u)}{\theta_o^2 + \text{MSE}(\hat{\theta}_u)} + \alpha &< 0
\end{aligned}$$

Since the terms $\text{MSE}(\hat{\theta}_u)$ and θ_o^2 are both positive $\text{MSE}(\hat{\theta}_u)$, and α assumed positive, this inequality cannot hold, since there is no route to make the result on the left hand side below 0. Hence, $\alpha > 0$ will not result in a reduction in MSE.

For the case $\alpha < 0$: here the biased estimator "shrinks" $\hat{\theta}_u$, and we get

$$\begin{aligned}
\Rightarrow \alpha\left(\frac{2\text{MSE}(\hat{\theta}_u)}{\theta_o^2 + \text{MSE}(\hat{\theta}_u)} + \alpha\right) &< 0 \\
\Rightarrow \frac{2\text{MSE}(\hat{\theta}_u)}{\theta_o^2 + \text{MSE}(\hat{\theta}_u)} + \alpha &> 0
\end{aligned}$$

The term on the left hand side, is only positive if

$$\frac{2\text{MSE}(\hat{\theta}_u)}{\text{MSE}(\hat{\theta}_u) + \theta_o^2} > -\alpha \quad \Leftrightarrow \quad -\frac{2\text{MSE}(\hat{\theta}_u)}{\text{MSE}(\hat{\theta}_u) + \theta_o^2} < \alpha$$

Additionally, we can establish a lower bound for α by analyzing $\frac{\text{MSE}(\hat{\theta}_u)}{\text{MSE}(\hat{\theta}_u)+\theta_o^2}$. Since all the individual terms in $\frac{\text{MSE}(\hat{\theta}_u)}{\text{MSE}(\hat{\theta}_u)+\theta_o^2}$ are assumed positive, we get:

$$0 < \frac{\text{MSE}(\hat{\theta}_u)}{\text{MSE}(\hat{\theta}_u) + \theta_o^2} < 1$$

That means:

$$-2 < \alpha < 0 \Rightarrow -1 < 1 + \alpha < 1 \Rightarrow |1 + \alpha| < 1$$

and additionally

$$|\hat{\theta}_b| = |(1 + \alpha)\hat{\theta}_u| = |1 + \alpha||\hat{\theta}_u| < |\hat{\theta}_u|$$

2.3 Bias-variance trade-off

Exercise 2.3.1

Reusing the rewrites we made in exercise 2.1, we can readily see we can write the cost-function as

$$J(\boldsymbol{\theta}) = (\boldsymbol{y} - X\boldsymbol{\theta})^T(\boldsymbol{y} - X\boldsymbol{\theta}) + \lambda\boldsymbol{\theta}^T\boldsymbol{\theta}$$

And using the same procedure for differentation we get

$$\frac{\partial}{\partial\boldsymbol{\theta}}J(\boldsymbol{\theta}) = -2X^T\boldsymbol{y} + 2X^TX\boldsymbol{\theta} + 2\lambda I\boldsymbol{\theta}$$

Equating the derivative to 0 and re-arrange yields

$$\begin{aligned} -2X^T\boldsymbol{y} + 2X^TX\boldsymbol{\theta} + 2\lambda I\boldsymbol{\theta} = \mathbf{0} \quad &\Leftrightarrow \\ (2X^TX + 2\lambda I)\boldsymbol{\theta} = 2X^T\boldsymbol{y} \quad &\Leftrightarrow \\ (X^TX + \lambda I)\boldsymbol{\theta} = X^T\boldsymbol{y} \end{aligned}$$

02471 Machine Learning for Signal Processing

Solution

Exercise 3: Stochastic Processes and Linear Filtering

3.1 Correlation functions

Exercise 3.1.1

Now we consider the following signal:

$$\mathrm{u}_n = c_1 \mathrm{x}_n + c_2 \mathrm{y}_{n-d}$$

Using the expression for auto-correlation we obtain

$$\begin{aligned} r_u(k) &= \mathbb{E}[\mathrm{u}_n \mathrm{u}_{n-k}] \\ &= \mathbb{E}[(c_1 \mathrm{x}_n + c_2 \mathrm{y}_{n-d})(c_1 \mathrm{x}_{n-k} + c_2 \mathrm{y}_{n-d-k})] \\ &= c_1^2 \mathbb{E}[\mathrm{x}_n \mathrm{x}_{n-k}] + c_1 c_2 \mathbb{E}\big[\mathrm{x}_n \mathrm{y}_{n-(d+k)}\big] + \\ &\qquad c_1 c_2 \mathbb{E}[\mathrm{y}_{n-d} \mathrm{x}_{n-k}] + c_2^2 \mathbb{E}[\mathrm{y}_{n-d} \mathrm{y}_{n-d-k}] \end{aligned}$$

Where we used the linear properties of expectation. Using the fact that x_n and y_n are WSS (we are free to shift) we obtain

$$\begin{aligned} r_u(k) &= c_1^2 \mathbb{E}[\mathrm{x}_n \mathrm{x}_{n-k}] + c_1 c_2 \mathbb{E}\big[\mathrm{x}_n \mathrm{y}_{n-(d+k)}\big] + \\ &\qquad c_1 c_2 \mathbb{E}\big[\mathrm{y}_n \mathrm{x}_{n-(k-d)}\big] + c_2^2 \mathbb{E}[\mathrm{y}_n \mathrm{y}_{n-k}] \\ &= c_1^2 r_x(k) + c_1 c_2 r_{xy}(d+k) + c_1 c_2 r_{xy}(d-k) + c_2^2 r_y(k) \end{aligned}$$

Exercise 3.1.4

A 1st order AR process is represented as: $\mathrm{u}_n = a\mathrm{u}_{n-1} + \eta_n$, where $a \in \mathbb{R}$ and η_n is a white noise sequence, which has the properties $\mathbb{E}[\eta_n] = 0$ and $r_\eta(k) = \sigma_\eta^2 \delta(k)$.

Insert into the expression for auto-correlation:

$$\begin{aligned} r_u(k) &= \mathbb{E}[\mathrm{u}_n \mathrm{u}_{n-k}] \\ &= \mathbb{E}[(a\mathrm{u}_{n-1} + \eta_n)\mathrm{u}_{n-k}] \\ &= \mathbb{E}[a\mathrm{u}_{n-1}\mathrm{u}_{n-k} + \eta_n \mathrm{u}_{n-k}] \\ &= a\mathbb{E}\big[\mathrm{u}_n \mathrm{u}_{n-(k-1)}\big] + \mathbb{E}[\eta_n \mathrm{u}_{n-k}] \end{aligned}$$

The latter term can easily be reduced if η_n and u_{n-k} is uncorrelated. If $k > 0$ this will be the case since u_{n-k} cannot be correlated with the future η_n noise sequence. Then we get $\mathbb{E}[\eta_n \mathrm{u}_{n-k}] = \mathbb{E}[\eta_n]\mathbb{E}[\mathrm{u}_{n-k}] = 0$, since $\mathbb{E}[\eta_n] = 0$.

So we get

$$\begin{aligned} r_u(k) &= a\mathbb{E}\big[\mathrm{u}_n \mathrm{u}_{n-(k-1)}\big] \\ &= a r_u(k-1) \end{aligned}$$

So, for $k > 0$ we have recursion.

Let us consider $k = 0$.

$$\begin{aligned} r_u(0) &= \mathbb{E}\big[\mathrm{u}_n^2\big] \\ &= \mathbb{E}\big[(a\mathrm{u}_{n-1} + \eta_n)^2\big] \\ &= \mathbb{E}\big[a^2\mathrm{u}_{n-1}\mathrm{u}_{n-1}\big] + \mathbb{E}\big[\eta_n^2\big] + 2a\mathbb{E}[\mathrm{u}_{n-1}\eta_n] \\ &= \mathbb{E}\big[a^2\mathrm{u}_{n-1}\mathrm{u}_{n-1}\big] + \sigma_\eta^2 \end{aligned}$$

where the latter term vanish since u_{n-1} and η_n are uncorrelated. Using the fact that u_n WSS, we shift the signal to obtain

$$\begin{aligned} r_u(0) &= \mathbb{E}\big[a^2\mathrm{u}_{n-1}\mathrm{u}_{n-1}\big] + \sigma_\eta^2 \\ &= a^2\mathbb{E}[\mathrm{u}_n\mathrm{u}_n] + \sigma_\eta^2 \\ &= a^2 r_u(0) + \sigma_\eta^2 \end{aligned}$$

Isolate for $r_u(0)$ to obtain

$$r_u(0) = \frac{\sigma_\eta^2}{1 - a^2}$$

Let's have a look on the term we derived a few steps before: $r_u(k) = ar_u(k-1)$. We see the recursive pattern:

$$\begin{aligned} r_u(k) &= ar_u(k-1) \\ r_u(1) &= ar_u(0) \\ r_u(2) &= ar_u(1) = a^2 \cdot r_u(0) \end{aligned}$$

So, we see for $k > 0$, we have $r_u(k) = ar_u(k-1)$. From properties of the auto-correlation sequence, equation 2.113 from the book, we know that $r_u(k) = r_u(-k)$, so for all k we have:

$$r_u(k) = a^{|k|} \cdot r_u(0) = \frac{a^{|k|}}{1 - a^2}\sigma_\eta^2$$

3.2 Wiener filter

Exercise 3.2.1

From the description we have $\mathrm{u}_n = \mathrm{s}_n + \epsilon_n$ where s_n is an AR(1) process and ϵ_n is a white noise sequence. Using the expression for auto-correlation we obtian

$$\begin{aligned} r_u(k) &= \mathbb{E}[\mathrm{u}_n\mathrm{u}_{n-k}] \\ &= \mathbb{E}[(\mathrm{s}_n + \epsilon_n)(\mathrm{s}_{n-k} + \epsilon_{n-k})] \\ &= \mathbb{E}[\mathrm{s}_n\mathrm{s}_{n-k}] + \mathbb{E}[\mathrm{s}_n\epsilon_{n-k}] + \mathbb{E}[\epsilon_n\mathrm{s}_{n-k}] + \mathbb{E}[\epsilon_n\epsilon_{n-k}] \end{aligned}$$

Since signal s_n and ϵ_n are uncorrelated and ϵ_n is a white noise sequence the cross-terms $\mathbb{E}[\mathrm{s}_n\epsilon_{n-k}]$ and $\mathbb{E}[\epsilon_n\mathrm{s}_{n-k}]$ vanish and we obtain

$$\begin{aligned} r_u(k) &= \mathbb{E}[\mathrm{s}_n\mathrm{s}_{n-k}] + \mathbb{E}[\epsilon_n\epsilon_{n-k}] \\ &= r_s(k) + r_\epsilon(k) \end{aligned}$$

From 3.1.4 we know that

$$r_s(k) = \frac{a^{|k|}}{1-a^2}\sigma_\eta^2$$

For $r_\epsilon(k)$ we have (since ϵ_n is a white noise sequence) $r_\epsilon(k) = \delta(k)\sigma_w^2$. So, we obtain by substitution:

$$r_u(k) = \frac{a^{|k|}}{1-a^2}\sigma_v^2 + \delta(k)\sigma_\epsilon^2$$

Exercise 3.2.2

The setup specifies that $\mathrm{d}_n = \mathrm{s}_n$ so using the expression for cross-correlation we obtain

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

Where, in the last line, we used ϵ_n is a white noise sequence.

Exercise 3.2.3

Using equation (4.43) in the book, we get

$$\left(\begin{bmatrix} r_s(0) & r_s(1) & r_s(2) \\ r_s(1) & r_s(0) & r_s(1) \\ r_s(2) & r_s(1) & r_s(0) \end{bmatrix} + \begin{bmatrix} \sigma_\epsilon^2 & 0 & 0 \\ 0 & \sigma_\epsilon^2 & 0 \\ 0 & 0 & \sigma_\epsilon^2 \end{bmatrix} \right) \boldsymbol{w} = \begin{bmatrix} r_s(0) \\ r_s(1) \\ r_s(2) \end{bmatrix}$$

Exercise 3.2.4

For $\sigma_\epsilon^2 \to 0$:

$$\boldsymbol{w} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

Thus filtering by $\boldsymbol{w}$ does not alter the signal (try and write the convolution for the filter, to see we get $\hat{\mathrm{d}}_n = \mathrm{u}_n$). This is a reasonable solution, since there is no noise to filter out.

Exercise 3.2.5

For $\sigma_\epsilon^2 \gg \sigma_\eta^2$:

$$\boldsymbol{w} = \frac{1}{\sigma_\epsilon^2} \begin{bmatrix} r_s(0) \\ r_s(1) \\ r_s(2) \end{bmatrix}$$

$\boldsymbol{w}$ tends to 0. This is a sensible solution, since there is only (unpredictable) noise in the signal. That means the energy output of the filter will be smaller as the noise dominates.

02471 Machine Learning for Signal Processing

Solution

Exercise 4: Adaptive Linear Filtering with LMS

4.1 Wiener Filter review

Exercise 4.1.1

Using the formulas (4.5)-(4.6) and (4.43) we get

$$\begin{bmatrix} r_u(0) & r_u(1) \\ r_u(1) & r_u(0) \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} r_{du}(0) \\ r_{du}(1) \end{bmatrix}$$

Exercise 4.1.2

We insert into the expression for $r_x(k)$

$$\begin{aligned} r_u(k) &= \mathbb{E}[\mathrm{u}_n \mathrm{u}_{n-k}] \\ &= \mathbb{E}[(\mathrm{s}_n + \mathrm{v}_n)(\mathrm{s}_{n-k} + \mathrm{v}_{n-k})] \\ &= \mathbb{E}[\mathrm{s}_n \mathrm{s}_{n-k}] + \mathbb{E}[\mathrm{v}_n \mathrm{v}_{n-k}] \\ &= r_s(k) + r_v(k) \end{aligned}$$

where we have used that s_n and v_n are uncorrelated and v_n is a zero-mean noise signal (the cross-terms vanish).

For the cross-correlation between the desired signal and input signal, we get (setting the desired signal to s_n)

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

Exercise 4.1.3

Since we are using a filter of length 2, we only need to compute two coefficients. They are computed using the formula $\boldsymbol{w} = \Sigma_u^{-1}\boldsymbol{p}$:

$$\begin{aligned} \boldsymbol{w} &= \begin{bmatrix} r_u(0) & r_u(1) \\ r_u(1) & r_u(0) \end{bmatrix}^{-1} \begin{bmatrix} r_{du}(0) \\ r_{du}(1) \end{bmatrix} \\ &= \begin{bmatrix} r_s(0) + r_v(0) & r_s(1) + r_v(1) \\ r_s(1) + r_v(1) & r_s(0) + r_v(0) \end{bmatrix}^{-1} \begin{bmatrix} r_s(0) \\ r_s(1) \end{bmatrix} \\ &= \begin{bmatrix} 1.2 & -0.4 \\ -0.4 & 1.2 \end{bmatrix}^{-1} \begin{bmatrix} 0.6 \\ -0.3 \end{bmatrix} \\ &= \begin{bmatrix} 0.469 \\ -0.094 \end{bmatrix} \end{aligned}$$

Exercise 4.1.4

Using formula (4.9) from the book we get the MMSE for the filter to

$$\begin{aligned}\text{MMSE}_2 &= \sigma_d^2 - \boldsymbol{p}^T \boldsymbol{w} \\ \text{MMSE}_2 &= 0.9 - \begin{bmatrix} 0.6 \\ -0.3 \end{bmatrix}^T \begin{bmatrix} 0.469 \\ -0.094 \end{bmatrix} \\ &= 0.591\end{aligned}$$

4.2 System identification using a Wiener filter

Exercise 4.2.1

The filter length is l which is directly read from the Model. The input sequence is u_n, the desired sequence is $d_n = y_n + \eta_n = H * u_n + \eta_n$. The error sequence is e_n, and the output from the model is $\hat{d} = \boldsymbol{w} * u_n$. The symbols are italic case in the figure, since the figure assumes realizations of the random processes.

Exercise 4.2.2

Using the definition of convolution (eq. 2.126), we get

$$\begin{aligned}\mathbb{E}[(\mathbf{u}_n * H)\mathrm{u}_{n-k}] &= \mathbb{E}\left[\mathrm{u}_{n-k} \sum_{i=0}^{l-1} u_{n-i} h_i\right] \\ &= \sum_{i=0}^{l-1} \mathbb{E}[\mathrm{u}_{n-k}\mathrm{u}_{n-i}] h_i\end{aligned}$$

Since u_n is a white noise sequence, we have $\mathbb{E}[\mathrm{u}_n\mathrm{u}_{n-k}] = \delta_k \sigma_u^2$, hence we have

$$\mathbb{E}[\mathrm{u}_{n-k}\mathrm{u}_{n-i}] = \begin{cases} \sigma_u^2 & k = i \\ 0 & k \neq i \end{cases}$$

Using this result we obtain

$$\mathbb{E}[(\mathbf{u}_n * H)\mathrm{u}_{n-k}] = \sigma_{\mathrm{u}}^2 h_k$$

Exercise 4.2.3

First we determine the input correlation matrix Σ_u. Since the input sequence is a white noise sequence, $r_u(k) = \sigma_u^2$ for $k = 0$ and $r_u(k) = 0$ for $k \neq 0$, therefore the input covariance matrix is:

$$\Sigma_u = \sigma_u^2 I$$

Next we identify the cross-correlation vector $\boldsymbol{p}$ whose elements are $p_k = r_{du}(k)$. We have for d_u

$$\begin{aligned}\mathrm{d}_n &= \mathrm{y}_n + \boldsymbol{\eta}_n \\ &= H * \mathbf{u}_n + \boldsymbol{\eta}_n\end{aligned}$$

For the cross-correlation we then get

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

Since η_n is uncorrelated with u_n, and $\mathbb{E}[\mathrm{u}_n] = 0$ we get:

$$r_{du}(k) = \mathbb{E}[(H * \mathbf{u}_n)\mathrm{u}_{n-k}]$$

Using the result from 4.2.2 we get:

$$\begin{aligned} r_{du}(k) = h_k \sigma_u^2 \quad &\Rightarrow \\ \boldsymbol{p} = H\sigma_u^2 \end{aligned}$$

From the normal equation we finally get:

$$\begin{aligned} \boldsymbol{w}_* &= \Sigma_u^{-1} \boldsymbol{p} \\ &= (\sigma_u^2 I)^{-1} H \sigma_u^2 \\ &= H \end{aligned}$$

Exercise 4.2.4

First we determine an expression for the error sequence

$$\begin{aligned} \mathrm{e}_n &= \mathrm{d}_n - \hat{\mathrm{d}}_n \\ &= H * \mathbf{u}_n + \eta_n - \boldsymbol{w} * \mathbf{u}_n \\ &= (H - \boldsymbol{w}) * \mathbf{u}_n + \eta_n \\ &= \boldsymbol{g}^T \mathbf{u}_n + \eta_n \end{aligned}$$

Where we have used that the convolution is distributive, ie. $f * (g + h) = f * g + f * h$, and defined the vector $\boldsymbol{g} = H - \boldsymbol{w}$. We now get the mean squared error to

$$\begin{aligned} \mathbb{E}\big[\mathrm{e}_n^2\big] &= \mathbb{E}\big[(\boldsymbol{g}^T \mathbf{u}_n + \eta_n)^2\big] \\ &= \mathbb{E}\big[\boldsymbol{g}^T \mathbf{u}_n \mathbf{u}_n^T \boldsymbol{g} + 2\boldsymbol{g}^T \mathbf{u}_n \eta_n + \eta_n^2\big] \\ &= \boldsymbol{g}^T \mathbb{E}\big[\mathbf{u}_n \mathbf{u}_n^T\big] \boldsymbol{g} + 2\boldsymbol{g}^T \mathbb{E}[\mathbf{u}_n \eta_n] + \mathbb{E}\big[\eta_n^2\big] \end{aligned}$$

where we have used that $\boldsymbol{g}$ is a deterministic vector.

From exercise 4.2.2 we found $\mathbb{E}[\mathrm{u}_n \mathrm{u}_{n-k}] = \delta_k \sigma_u^2$, hence $\mathbb{E}\big[\mathbf{u}_n \mathbf{u}_n^T\big] = \sigma_u^2 I$.

Additionally, since $\mathbf{u}_n$ and η_n are two uncorrelated white noise sequences we have $\mathbb{E}[\mathbf{u}_n \eta_n] = \mathbb{E}[\mathbf{u}_n]\mathbb{E}[\eta_n] = 0$ (since $\mathbb{E}[\mathbf{u}_n] = \mathbb{E}[\eta_n] = 0$).

Using these results, our expression reduce to

$$\mathbb{E}\big[\mathrm{e}_n^2\big] = \boldsymbol{g}^T \sigma_u^2 I \boldsymbol{g} + \sigma_\eta^2$$

And from linear algebra we know that: $\boldsymbol{g}^T \boldsymbol{g} = \|\boldsymbol{g}\|_2^2$, hence:

$$\mathbb{E}\big[e_n^2\big] = \sigma_u^2 \|H - \boldsymbol{w}\|_2^2 + \sigma_\eta^2$$

So the lowest MSE possible is σ_η^2, and σ_u^2 is adding to the MSE proportionally to the filter differences.

02471 Machine Learning for Signal Processing

Solution

Exercise 5: Adaptive Linear Filtering with RLS

5.1 Derivation of the RLS algorithm

Exercise 5.1.1

For $\beta = 1$.

Exercise 5.1.2

If $0 < \beta < 1$, the regularization term decreases as we observe more data. The motivation is that for large amounts of data, the risk of overfitting is reduced, so the regularization term is not required.

Exercise 5.1.3

The cost function is defined as:

$$J(\boldsymbol{\theta}, \beta, \lambda) = \sum_{i=0}^{n} \beta^{n-i}(y_i - \boldsymbol{\theta}^T \boldsymbol{x}_i)^2 + \lambda\beta^{n+1}||\boldsymbol{\theta}||^2$$

Taking the derivative with respect to $\boldsymbol{\theta}$ gives:

$$\begin{aligned}
\frac{\partial J(\boldsymbol{\theta}, \beta, \lambda)}{\partial \boldsymbol{\theta}} &= \sum_{i=0}^{n} \beta^{n-i} \frac{\partial}{\partial \boldsymbol{\theta}} \Big((y_i - \boldsymbol{\theta}^T \boldsymbol{x}_i)^2 \Big) + \lambda\beta^{n+1} \frac{\partial}{\partial \boldsymbol{\theta}} \big(||\boldsymbol{\theta}||^2 \big) \\
&= -2 \sum_{i=0}^{n} \beta^{n-i}(y_i - \boldsymbol{\theta}^T \boldsymbol{x}_i)\boldsymbol{x}_i + 2\lambda\beta^{n+1}\boldsymbol{\theta} \\
&= -2 \sum_{i=0}^{n} \beta^{n-i} y_i \boldsymbol{x}_i + 2 \sum_{i=0}^{n} \beta^{n-i} \underbrace{(\boldsymbol{\theta}^T \boldsymbol{x}_i)\boldsymbol{x}_i}_{\substack{=\boldsymbol{x}_i(\boldsymbol{\theta}^T \boldsymbol{x}_i) \\ =\boldsymbol{x}_i(\boldsymbol{x}_i^T \boldsymbol{\theta}) \\ =(\boldsymbol{x}_i \boldsymbol{x}_i^T)\boldsymbol{\theta}}} + 2\lambda\beta^{n+1}\boldsymbol{\theta} \\
&= -2 \underbrace{\sum_{i=0}^{n} \beta^{n-i} y_i \boldsymbol{x}_i}_{=\boldsymbol{p}_n} + 2 \Bigg(\underbrace{\sum_{i=0}^{n} \beta^{n-i} \boldsymbol{x}_i \boldsymbol{x}_i^T + \lambda\beta^{n+1} I}_{=\Phi_n} \Bigg) \boldsymbol{\theta}
\end{aligned}$$

Setting the derivative to 0 on $\boldsymbol{\theta}_n$ gives:

$$\Phi_n \boldsymbol{\theta}_n = \boldsymbol{p}_n$$

Exercise 5.1.4

Let's start from the formula for Φ_n and make Φ_{n-1} appear:

$$\begin{aligned}
\Phi_n &= \sum_{i=0}^{n} \beta^{n-i} \boldsymbol{x}_i \boldsymbol{x}_i^T + \lambda \beta^{n+1} I \\
&= \sum_{i=0}^{n-1} \beta^{n-i} \boldsymbol{x}_i \boldsymbol{x}_i^T + \beta^{n-n} \boldsymbol{x}_n \boldsymbol{x}_n^T + \lambda \beta^{n+1} I \\
&= \sum_{i=0}^{n-1} \beta \beta^{n-1-i} \boldsymbol{x}_i \boldsymbol{x}_i^T + \boldsymbol{x}_n \boldsymbol{x}_n^T + \lambda \beta \beta^{n} I \\
&= \beta \left(\sum_{i=0}^{n-1} \beta^{n-1-i} \boldsymbol{x}_i \boldsymbol{x}_i^T + \lambda \beta^{n} I \right) + \boldsymbol{x}_n \boldsymbol{x}_n^T \\
&= \beta \Phi_{n-1} + \boldsymbol{x}_n \boldsymbol{x}_n^T
\end{aligned}$$

Similarly for $\boldsymbol{p}_n$:

$$\begin{aligned}
\boldsymbol{p}_n &= \sum_{i=0}^{n} \beta^{n-i} y_i \boldsymbol{x}_i \\
&= \sum_{i=0}^{n-1} \beta^{n-i} y_i \boldsymbol{x}_i + \beta^{n-n} y_n \boldsymbol{x}_n \\
&= \beta \sum_{i=0}^{n-1} \beta^{n-1-i} y_i \boldsymbol{x}_i + y_n \boldsymbol{x}_n \\
&= \beta \boldsymbol{p}_{n-1} + y_n \boldsymbol{x}_n
\end{aligned}$$

Exercise 5.1.5

From Appendix 1, we have:

$$(A + BD^{-1}C)^{-1} = A^{-1} - A^{-1}B(D + CA^{-1}B)^{-1}CA^{-1}$$

Let's use $A = \beta\Phi_{n-1}$, $B = \boldsymbol{x}_n$, $C = \boldsymbol{x}_n^T$ and $D = 1$. The left term becomes:

$$(A + BD^{-1}C)^{-1} = (\beta\Phi_{n-1} + \boldsymbol{x}_n \boldsymbol{x}_n^T)^{-1}$$

which is exactly Φ_n^{-1} according to the previous question. As for the right term, it becomes:

$$\begin{aligned}
A^{-1} - A^{-1}B(D + CA^{-1}B)^{-1}CA^{-1} &= \beta^{-1}\Phi_{n-1}^{-1} - \beta^{-1}\Phi_{n-1}^{-1}\boldsymbol{x}_n \left(1 + \boldsymbol{x}_n^T \beta^{-1} \Phi_{n-1}^{-1} \boldsymbol{x}_n\right)^{-1} \boldsymbol{x}_n^T \beta^{-1} \Phi_{n-1}^{-1} \\
&= \beta^{-1}\Phi_{n-1}^{-1} - \beta^{-1} \frac{\beta^{-1}\Phi_{n-1}^{-1}\boldsymbol{x}_n}{1 + \boldsymbol{x}_n^T \beta^{-1} \Phi_{n-1}^{-1} \boldsymbol{x}_n} \boldsymbol{x}_n^T \Phi_{n-1}^{-1} \\
&= \beta^{-1}\Phi_{n-1}^{-1} - \beta^{-1} \boldsymbol{k}_n \boldsymbol{x}_n^T \Phi_{n-1}^{-1}
\end{aligned}$$

where we introduced $\boldsymbol{k}_n = \dfrac{\beta^{-1}\Phi_{n-1}^{-1}\boldsymbol{x}_n}{1 + \beta^{-1}\boldsymbol{x}_n^T\Phi_{n-1}^{-1}\boldsymbol{x}_n}$.

Equating the left and right terms thus gives the desired result:

$$\Phi_n^{-1} = \beta^{-1}\Phi_{n-1}^{-1} - \beta^{-1} \boldsymbol{k}_n \boldsymbol{x}_n^T \Phi_{n-1}^{-1}$$

Exercise 5.1.6

From 5.1.3 we have:

$$\boldsymbol{\theta}_n = \Phi_n^{-1}\boldsymbol{p}_n$$

Plugging the recursive formula for $\boldsymbol{p}_n$ found in 5.1.4 gives:

$$\begin{aligned}\boldsymbol{\theta}_n &= \Phi_n^{-1}\left(\beta\boldsymbol{p}_{n-1} + \boldsymbol{x}_n y_n\right)\\ &= \beta\Phi_n^{-1}\boldsymbol{p}_{n-1} + \Phi_n^{-1}\boldsymbol{x}_n y_n\end{aligned}$$

Let's now plug the expression for Φ_n^{-1} found in the previous question, only for the left term of the sum for now. This makes $\boldsymbol{\theta}_{n-1}$ appear:

$$\begin{aligned}\boldsymbol{\theta}_n &= \beta\left(\beta^{-1}\Phi_{n-1}^{-1} - \beta^{-1}\boldsymbol{k}_n\boldsymbol{x}_n^T\Phi_{n-1}^{-1}\right)\boldsymbol{p}_{n-1} + \Phi_n^{-1}\boldsymbol{x}_n y_n\\ &= \Phi_{n-1}^{-1}\boldsymbol{p}_{n-1} - \boldsymbol{k}_n\boldsymbol{x}_n^T\Phi_{n-1}^{-1}\boldsymbol{p}_{n-1} + \Phi_n^{-1}\boldsymbol{x}_n y_n\\ &= \boldsymbol{\theta}_{n-1} - \boldsymbol{k}_n\boldsymbol{x}_n^T\boldsymbol{\theta}_{n-1} + \Phi_n^{-1}\boldsymbol{x}_n y_n\end{aligned}$$

At this point we just need to prove that $\Phi_n^{-1}\boldsymbol{x}_n = \boldsymbol{k}_n$, since we would then be able to factorize and make e_n appear. This can be done by manipulating the expression for $\boldsymbol{k}_n$:

$$\begin{aligned}\boldsymbol{k}_n &= \frac{\beta^{-1}\Phi_{n-1}^{-1}\boldsymbol{x}_n}{1+\beta^{-1}\boldsymbol{x}_n^T\Phi_{n-1}^{-1}\boldsymbol{x}_n}\\ \Rightarrow \boldsymbol{k}_n(1+\beta^{-1}\boldsymbol{x}_n^T\Phi_{n-1}^{-1}\boldsymbol{x}_n) &= \beta^{-1}\Phi_{n-1}^{-1}\boldsymbol{x}_n\\ \Rightarrow \boldsymbol{k}_n &= \beta^{-1}\Phi_{n-1}^{-1}\boldsymbol{x}_n - \beta^{-1}\boldsymbol{k}_n\boldsymbol{x}_n^T\Phi_{n-1}^{-1}\boldsymbol{x}_n\\ &= (\beta^{-1}\Phi_{n-1}^{-1} - \beta^{-1}\boldsymbol{k}_n\boldsymbol{x}_n^T\Phi_{n-1}^{-1})\boldsymbol{x}_n\\ &= \Phi_n^{-1}\boldsymbol{x}_n\end{aligned}$$

where we have used once again the expression of Φ_n^{-1} found in the previous question.

We thus obtain the desired result:

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

which is the weight update for RLS.

02471 Machine Learning for Signal Processing

Solution

Exercise 6: Sparsity-aware learning with ℓ_1

6.1 Norms

Exercise 6.1.1

Solution: We consider the case $p = 0$ and $0 < p < 1$ seperately.

For $p = 0$ we can come with the following counter-example:

Consider the vector $\boldsymbol{x} = [1, 0, \dots 0]^T$, and choose any non-zero $\alpha \in \mathbb{R}$. Then we get (left hand side of the second property) gives

$$\|\alpha \boldsymbol{x}\|_0 = \|[\alpha, 0, \dots 0]^T\|_0 = 1$$

But IF the property holds, the result should have been

$$\|\alpha \boldsymbol{x}\|_0 = |\alpha| \|\boldsymbol{x}\|_0 = |\alpha| \ 1 = |\alpha|$$

So clearly the second property is violated, thus $p = 0$ is not a norm.

For $0 < p < 1$ we us show that property three is violated. Consider the two vectors in the l -dimensional space

$$\boldsymbol{x} = [1, 0, \dots, 0]^T, \quad \boldsymbol{y} = [0, 0, \dots, 1]^T$$

We will show that for these two vectors the triangle inequality is violated for $p < 1$. Indeed, we have (assuming now the triangle inequality holds)

$$\|\boldsymbol{x} + \boldsymbol{y}\|_p = \left(\sum_{i=1}^{l} |x_i + y_i|^p \right)^{1/p} = (1^p + 1^p)^{\frac{1}{p}} = (2 \cdot 1^p)^{\frac{1}{p}} = 2^{\frac{1}{p}} \leq \|x\|_p + \|y\|_p = 1 + 1 = 2$$

which is violated for $0 < p < 1$.

6.2 The regularized least-squares solution

Exercise 6.2.1

This is solved by direct substitution. If $X^T X = I$ we get

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

For Ridge regression we get

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

Exercise 6.2.2

This is shown in the book, equation (9.11)–9.13).

Exercise 6.2.3

Direct application of the formula will give the following result

$$\hat{\boldsymbol{\theta}}_R = \frac{1}{1+\lambda}[0.7, -0.3, 0.1, -2]^T = \frac{1}{2}[0.7, -0.3, 0.1, -2]^T = [0.35, -0.15, 0.05, -1]^T$$

For the ℓ_1 norm we get

$$\hat{\boldsymbol{\theta}}_1 = \begin{bmatrix} \operatorname{sgn}(0.7)\left(|0.7| - \frac{1}{2}\right)_+ \\ \operatorname{sgn}(-0.3)\left(|-0.3| - \frac{1}{2}\right)_+ \\ \operatorname{sgn}(0.1)\left(|0.1| - \frac{1}{2}\right)_+ \\ \operatorname{sgn}(-2)\left(|-2| - \frac{1}{2}\right)_+ \end{bmatrix} = \begin{bmatrix} 1\,(0.2)_+ \\ -1\,(-0.2)_+ \\ 1\,(-0.4)_+ \\ -1\,(1.5)_+ \end{bmatrix} = \begin{bmatrix} 1 \cdot 0.2 \\ -1 \cdot 0 \\ 1 \cdot 0 \\ -1 \cdot 1.5 \end{bmatrix} = \begin{bmatrix} 0.2 \\ 0 \\ 0 \\ -1.5 \end{bmatrix}$$

Exercise 6.2.4

Ridge regression requires $\lambda \to \infty$ to result in the null vector. For LASSO, $\lambda = 4$.

02471 Machine Learning for Signal Processing

Solution

Exercise 7: Sparsity analysis models and time-frequency analysis

7.2 Iterative Shrinkage/thresholding (IST)

Exercise 7.2.1

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^l}\left\{ L(\boldsymbol{\theta},\lambda) = \frac{1}{2}\|\boldsymbol{y}-X\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1 = J(\boldsymbol{\theta}) + \lambda\|\boldsymbol{\theta}\|_1 \right\}$$

From section 5.2, formula 5.3, we have the gradient decent update defined as

$$\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^{(i-1)} - \mu_i J'\big(\boldsymbol{\theta}^{(i-1)}\big)$$

And using a constant step size μ, and taking the derivative of the means squared cost function, we get

$$\begin{aligned}\boldsymbol{\theta}^{(i)} &= \boldsymbol{\theta}^{(i-1)} - \mu J'\big(\boldsymbol{\theta}^{(i-1)}\big)\\ &= \boldsymbol{\theta}^{(i-1)} - \mu X^T\big(X\boldsymbol{\theta}^{(i-1)} - \boldsymbol{y}\big)\\ &= \boldsymbol{\theta}^{(i-1)} + \mu X^T\boldsymbol{e}^{(i-1)}\end{aligned}$$

where $\boldsymbol{e}^{(i-1)} = \boldsymbol{y} - X\boldsymbol{\theta}^{(i-1)}$.

The gradient decent update that solves the MSE error can identically be written as the solution to the following optimization problem:

$$\boldsymbol{\theta}^{(i)} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^l}\left\{ J\big(\boldsymbol{\theta}^{(i-1)}\big) + \big(\boldsymbol{\theta}-\boldsymbol{\theta}^{(i-1)}\big)^T J'\big(\boldsymbol{\theta}^{(i-1)}\big) + \frac{1}{2\mu}\big\|\boldsymbol{\theta}-\boldsymbol{\theta}^{(i-1)}\big\|_2^2 \right\}$$

This can easiest be shown by taking the derivative to the expression we are minimizing w.r.t $\boldsymbol{\theta}$, set equal to zero and then solve for $\boldsymbol{\theta}$.

$$\begin{aligned}&\frac{\partial}{\partial\boldsymbol{\theta}}\left(J\big(\boldsymbol{\theta}^{(i-1)}\big) + \big(\boldsymbol{\theta}-\boldsymbol{\theta}^{(i-1)}\big)^T J'\big(\boldsymbol{\theta}^{(i-1)}\big) + \frac{1}{2\mu}\big\|\boldsymbol{\theta}-\boldsymbol{\theta}^{(i-1)}\big\|_2^2 \right) =\\ &\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{\theta}^T J'\big(\boldsymbol{\theta}^{(i-1)}\big) + \frac{1}{2\mu}\frac{\partial}{\partial\boldsymbol{\theta}}\big\|\boldsymbol{\theta}-\boldsymbol{\theta}^{(i-1)}\big\|_2^2 =\\ &J'\big(\boldsymbol{\theta}^{(i-1)}\big) + \frac{1}{2\mu}\big(2\boldsymbol{\theta} - 2\boldsymbol{\theta}^{(i-1)}\big)\end{aligned}$$

Equation to zero, we get

$$\begin{aligned}0 &= J'\big(\boldsymbol{\theta}^{(i-1)}\big) + \frac{1}{2\mu}\big(2\boldsymbol{\theta}^{(i)} - 2\boldsymbol{\theta}^{(i-1)}\big)\\ &= \mu J'\big(\boldsymbol{\theta}^{(i-1)}\big) + \boldsymbol{\theta}^{(i)} - \boldsymbol{\theta}^{(i-1)} \Leftrightarrow\\ \boldsymbol{\theta}^{(i)} &= \boldsymbol{\theta}^{(i-1)} - \mu J'\big(\boldsymbol{\theta}^{(i-1)}\big)\end{aligned}$$

Thus we have validated the claim. Hence, by substitution, we rewrite our LASSO problem to

$$\boldsymbol{\theta}^{(i)} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^l}\left\{ J\big(\boldsymbol{\theta}^{(i-1)}\big) + \big(\boldsymbol{\theta}-\boldsymbol{\theta}^{(i-1)}\big)^T \frac{\partial \boldsymbol{J}\big(\boldsymbol{\theta}^{(i-1)}\big)}{\partial\boldsymbol{\theta}} + \frac{1}{2\mu}\big\|\boldsymbol{\theta}-\boldsymbol{\theta}^{(i-1)}\big\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1 \right\}$$

To shorten the notation burden, define $\boldsymbol{\theta}' := \boldsymbol{\theta}^{(i-1)}$

$$\begin{aligned}\boldsymbol{\theta}^{(i)} &= \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^l}\left\{ J\big(\boldsymbol{\theta}^{(i-1)}\big) + \big(\boldsymbol{\theta} - \boldsymbol{\theta}^{(i-1)}\big)^T \frac{\partial \boldsymbol{J}\big(\boldsymbol{\theta}^{(i-1)}\big)}{\partial \boldsymbol{\theta}} + \frac{1}{2\mu}\big\|\boldsymbol{\theta} - \boldsymbol{\theta}^{(i-1)}\big\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1 \right\} \\ &= \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^l}\left\{ J(\boldsymbol{\theta}') + (\boldsymbol{\theta} - \boldsymbol{\theta}')^T J'(\boldsymbol{\theta}') + \frac{1}{2\mu}\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1 \right\}\end{aligned}$$

Scaling a function (multiplication) or adding a constant to a function does not change the value of the minimizer, i.e. $\arg\min_{\boldsymbol{\theta}\in\mathbb{R}^l}\{f(\boldsymbol{\theta})\} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^l}\{\alpha f(\boldsymbol{\theta}) + k\}, \alpha \in \mathbb{R}, k \in \mathbb{R}$. Using this, we can further rewrite the optimization problem

$$\begin{aligned}\boldsymbol{\theta}^{(i)} &= \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^l}\left\{ J(\boldsymbol{\theta}') + (\boldsymbol{\theta} - \boldsymbol{\theta}')^T J'(\boldsymbol{\theta}') + \frac{1}{2\mu}\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1 \right\} \\ &= \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^l}\left\{ \boldsymbol{\theta}^T J'(\boldsymbol{\theta}') + \frac{1}{2\mu}\big(\boldsymbol{\theta}^T\boldsymbol{\theta} - 2\boldsymbol{\theta}^T\boldsymbol{\theta}'\big) + \lambda\|\boldsymbol{\theta}\|_1 \right\} \\ &= \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^l}\left\{ \mu\boldsymbol{\theta}^T J'(\boldsymbol{\theta}') + \frac{1}{2}\big(\boldsymbol{\theta}^T\boldsymbol{\theta} - 2\boldsymbol{\theta}^T\boldsymbol{\theta}'\big) + \lambda\mu\|\boldsymbol{\theta}\|_1 \right\} \\ &= \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^l}\left\{ -\boldsymbol{\theta}^T(\boldsymbol{\theta}' - \mu J'(\boldsymbol{\theta}')) + \frac{1}{2}\big(\boldsymbol{\theta}^T\boldsymbol{\theta}\big) + \lambda\mu\|\boldsymbol{\theta}\|_1 \right\}\end{aligned}$$

Define $\tilde{\boldsymbol{\theta}} := \boldsymbol{\theta}' - \mu J'(\boldsymbol{\theta}')$ to obtain

$$\boldsymbol{\theta}^{(i)} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^l}\left\{ \frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{\theta} - \boldsymbol{\theta}^T\tilde{\boldsymbol{\theta}} + \lambda\mu\|\boldsymbol{\theta}\|_1 \right\}$$

To find the minimum of this function, we take the derivative if the function to minimize w.r.t. $\boldsymbol{\theta}$ to obtain

$$\frac{\partial}{\partial\boldsymbol{\theta}}\frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{\theta} - \boldsymbol{\theta}^T\tilde{\boldsymbol{\theta}} + \lambda\mu\|\boldsymbol{\theta}\|_1 = \boldsymbol{\theta} - \tilde{\boldsymbol{\theta}} + \lambda\mu\partial\|\boldsymbol{\theta}\|_1$$

where ∂ is the subdifferential set. The minimizer must satisfy

$$\mathbf{0} \in \boldsymbol{\theta}^{(i)} - \tilde{\boldsymbol{\theta}} + \lambda\mu\partial\|\boldsymbol{\theta}^{(i)}\|_1$$

What is very useful about this result, as compared to last week, is that no requirements was enforced on X. Last week we assumed that $X^TX = I$. All operations are only applied components wise, and we can therefore write, for the j'th component

$$\begin{cases} \theta_j^{(i)} - \tilde{\theta}_j + \lambda\mu = 0 \Leftrightarrow \theta_j^{(i)} = \tilde{\theta}_j - \lambda\mu & \text{if } \theta_j > 0 \\ \theta_j^{(i)} - \tilde{\theta}_j - \lambda\mu = 0 \Leftrightarrow \theta_j^{(i)} = \tilde{\theta}_j + \lambda\mu & \text{if } \theta_j < 0 \end{cases}$$

These equations are only true for $\tilde{\theta}_j > \lambda\mu$ and $\tilde{\theta}_j < \lambda\mu$ respectively. For $-\lambda\mu \leq \tilde{\theta}_j \leq \lambda\mu$ we get $\tilde{\theta}_j = 0$ (see the solution from last week for further details). These three conditions can be combined into one rule

$$\theta_j^{(i)} = \text{sign}\Big(\tilde{\theta}_j^{(i)}\Big)\max\Big(\Big|\tilde{\theta}_j\Big| - \lambda\mu, 0\Big)$$

where $\text{sign}(x)$ is

$$\text{sign}(x) = \begin{cases} 1 & \text{if } x > 0 \\ -1 & \text{if } x < 0 \\ 0 & \text{if } x = 0 \end{cases}$$

Putting this together, we get the following update for the i'th iteration, where we initialize $\boldsymbol{\theta}^{(0)} = \mathbf{0}$:

$$\begin{aligned}\boldsymbol{e}^{(i-1)} &= \boldsymbol{y} - X\boldsymbol{\theta}^{(i-1)} \\ \tilde{\boldsymbol{\theta}} &= \boldsymbol{\theta}^{(i-1)} + \mu X^T \boldsymbol{e}^{(i-1)} \\ \boldsymbol{\theta}^{(i)} &= \operatorname{sign}\Big(\tilde{\boldsymbol{\theta}}\Big) \max\Big(\Big|\tilde{\boldsymbol{\theta}}\Big| - \lambda\mu, 0\Big)\end{aligned}$$

7.3 Signal representation using the Discrete Fourier Transform (DFT)

Exercise 7.3.1

First we define $W_N := e^{-i2\pi/N}$ to rewrite the N-point DFT to

$$\tilde{x}_k = \sum_{n=0}^{N-1} x_n (W_N)^{kn}$$

This sum can be written as an inner product. If define a vector $\boldsymbol{w}^k$ whose elements is $w_i^k = (W_N)^{ki}$, we get

$$\tilde{x}_k = \boldsymbol{x}^T \boldsymbol{w}^k$$

If we want to calculate all the desired $\tilde{x}_k$ components, we can do this using the matrix product

$$\tilde{\boldsymbol{x}} = \Phi^H \boldsymbol{x}$$

Where the matrix Φ^H has the following elements, $\Phi^H_{(i,j)} = (W_N)^{ij}$.

02471 Machine Learning for Signal Processing

Solution

Exercise 8: Dictionary learning and source separation

8.1 ICA and Gaussian signals

Exercise 8.1.1

We get (since A is orthogonal)

$$\begin{aligned} p_{\mathbf{x}}(\mathbf{x}) &= \frac{p_{\mathbf{s}}(\mathbf{s})}{|\det(J(\mathbf{x},\mathbf{s}))|} \\ \mathbf{x} &= A\mathbf{s} \\ \mathbf{s} &= A^{-1}\mathbf{x} = A^T\mathbf{x} \end{aligned}$$

Next we need to compute the Jacobian. This is easiest done if we operate on each component of $\mathbf{x}$. From our knowledge of matrix multiplication, we know that the i'th component of x is $\mathrm{x}_i = A_{i,:}\mathbf{s}$, where $A_{i,:}$ denotes the i'th row of A, i.e

$$\mathbf{x} = \begin{bmatrix} \mathrm{x}_1 \\ \mathrm{x}_2 \\ \vdots \\ \mathrm{x}_l \end{bmatrix} = \begin{bmatrix} A_{1,1}\mathrm{s}_1 + A_{1,2}\mathrm{s}_2 + \cdots + A_{1,l}\mathrm{s}_l \\ A_{2,1}\mathrm{s}_1 + A_{2,2}\mathrm{s}_2 + \cdots + A_{2,l}\mathrm{s}_l \\ \vdots \\ A_{l,1}\mathrm{s}_1 + A_{l,2}\mathrm{s}_2 + \cdots + A_{l,l}\mathrm{s}_l \end{bmatrix}$$

So, we have for example for x_1:

$$\mathrm{x}_1 = A_{1,1}\mathrm{s}_1 + A_{1,2}\mathrm{s}_2 + \cdots + A_{1,l}\mathrm{s}_l$$

Taking the derivative with respect to s_1 yields

$$\begin{aligned} \frac{\partial}{\partial \mathrm{s}_1}\mathrm{x}_1 &= \frac{\partial}{\partial \mathrm{s}_1}(A_{1,1}\mathrm{s}_1 + A_{1,2}\mathrm{s}_2 + \cdots + A_{1,l}\mathrm{s}_l) \\ &= \frac{\partial}{\partial \mathrm{s}_1}A_{1,1}\mathrm{s}_1 + \frac{\partial}{\partial \mathrm{s}_1}A_{1,2}\mathrm{s}_2 + \cdots + \frac{\partial}{\partial \mathrm{s}_1}A_{1,l}\mathrm{s}_l \\ &= A_{1,1}\frac{\partial}{\partial \mathrm{s}_1}\mathrm{s}_1 + A_{1,2}\frac{\partial}{\partial \mathrm{s}_1}\mathrm{s}_2 + \cdots + A_{1,l}\frac{\partial}{\partial \mathrm{s}_1}\mathrm{s}_l \\ &= A_{1,1}\cdot 1 + A_{1,2}\cdot 0 + \cdots + A_{1,l}\cdot 0 \\ &= A_{1,1} \end{aligned}$$

Similarly, taking the derivative with respect to x_2 yields $A_{1,2}$, and so on.

$$\begin{aligned} \frac{\partial \mathrm{x}_1}{\partial \mathrm{s}_1} &= A_{1,1} \\ \frac{\partial \mathrm{x}_1}{\partial \mathrm{s}_2} &= A_{1,2} \\ &\vdots \\ \frac{\partial \mathrm{x}_1}{\partial \mathrm{s}_l} &= A_{1,l} \end{aligned}$$

Since the Jacobian matrix is defined as

$$J(\mathbf{x},\mathbf{s}) = \begin{bmatrix} \frac{\partial x_1}{\partial s_1} & \frac{\partial x_1}{\partial s_2} & \cdots & \frac{\partial x_1}{\partial s_l} \\ \frac{\partial x_2}{\partial s_1} & \frac{\partial x_2}{\partial s_2} & \cdots & \frac{\partial x_2}{\partial s_l} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial x_l}{\partial s_1} & \frac{\partial x_l}{\partial s_2} & \cdots & \frac{\partial x_l}{\partial s_l} \end{bmatrix}$$

The first row of the Jacobian will become (by substituting all the partial derivatives)

$$J(\mathrm{x}_1,\mathbf{s}) = \begin{bmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,l} \end{bmatrix}$$

and so on. Hence, we get

$$\begin{aligned} J(\mathbf{x},\mathbf{s}) &= \begin{bmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,l} \\ \vdots & \vdots & \ddots & \vdots \\ A_{l,1} & A_{l,2} & \cdots & A_{l,l} \end{bmatrix} \\ &= A \end{aligned}$$

Exercise 8.1.2

We know that: $\det(A^{-1}) = \frac{1}{\det(A)}$, and $p_{\mathbf{s}}(\mathbf{s}) = \frac{1}{(2\pi)^{l/2}} \exp\left(-\frac{\|\mathbf{s}\|^2}{2}\right)$, and $\mathbf{s} = A^T\mathbf{x}$ hence by substitution we get

$$\begin{aligned} p_{\mathbf{x}}(\mathbf{x}) &= \frac{p_{\mathbf{s}}(\mathbf{s})}{|\det(J(\mathbf{x},\mathbf{s}))|} \\ &= \frac{p_{\mathbf{s}}(\mathbf{s})}{|\det(A)|} \\ &= p_{\mathbf{s}}(\mathbf{s})|\det(A^{-1})| \\ &= p_{\mathbf{s}}(\mathbf{s})|\det(A^T)| \\ &= \frac{1}{(2\pi)^{l/2}} \exp\left(-\frac{\|\mathbf{s}\|^2}{2}\right)|\det(A^T)| \\ &= \frac{1}{(2\pi)^{l/2}} \exp\left(-\frac{\|A^T\mathbf{x}\|^2}{2}\right)|\det(A^T)| \end{aligned}$$

Since A is orthogonal we have $\det(A^T) = \pm 1 \Rightarrow |\det(A^T)| = 1$. Additionally, we have

$$\|A^T\mathbf{x}\|^2 = (A^T\mathbf{x})^T A^T\mathbf{x} = \mathbf{x}^T A A^T\mathbf{x} = \mathbf{x}^T\mathbf{x} = \|\mathbf{x}\|^2$$

Using these two results, we get

$$p_{\mathbf{x}}(\mathbf{x}) = \frac{1}{(2\pi)^{l/2}} \exp\left(-\frac{\|\mathbf{x}\|^2}{2}\right)$$

8.2 Derivation of ICA based on mutual information

Exercise 8.2.1

From section 2.5 (equation 2.158):

$$\begin{aligned} I(\mathrm{x},\mathrm{y}) &:= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x,y)\ln\frac{p(x,y)}{p(x)p(y)}dxdy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x,y)(\ln p(x,y) - \ln p(x) - \ln p(y))dxdy \end{aligned}$$

If we handle the terms individually, we get

$$\begin{aligned}
\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x,y)\ln p(x)dxdy &= \int_{-\infty}^{\infty}\ln p(x)\left(\int_{-\infty}^{\infty} p(x,y)dy\right)dx \\
&= \int_{-\infty}^{\infty}\ln p(x)p(x)dx \\
&= -H(\mathrm{x}) \\
\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x,y)\ln p(y)dxdy &= -H(\mathrm{y}) \\
\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x,y)\ln p(x,y)dxdy &= -H(\mathrm{x},\mathrm{y})
\end{aligned}$$

Combining yields

$$\begin{aligned}
I(\mathrm{x},\mathrm{y}) &:= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x,y)(\ln p(x,y) - \ln p(x) - \ln p(y))dxdy \\
&= -H(\mathrm{x},\mathrm{y}) + H(\mathrm{x}) + H(\mathrm{y})
\end{aligned}$$

This can

Which, in the general case (l–dimensional) leads to

$$I(\mathbf{z}) = -H(\mathbf{z}) + \sum_{i=1}^{l} H(\mathrm{z}_i)$$

E.g. if we set $l = 2$, we have $\mathbf{z} = [\mathrm{x}\ \mathrm{y}]^T$

Exercise 8.2.2

Using the result from the change of variables in 8.1 we have $p_{\mathbf{z}}(\mathbf{z}) = p_{\mathbf{x}}(\mathbf{x})/|\det(W)|$.

$$\begin{aligned}
p_{\mathbf{z}}(\mathbf{z}) &= p_{\mathbf{x}}(\mathbf{x})/|\det(W)| \ \Rightarrow \\
\ln p_{\mathbf{z}}(\mathbf{z}) &= \ln(p_{\mathbf{x}}(\mathbf{x})/|\det(W)|) \\
&= \ln p_{\mathbf{x}}(\mathbf{x}) - \ln|\det(W)| \ \Rightarrow \\
\mathbb{E}[\ln p_{\mathbf{z}}(\mathbf{z})] &= \mathbb{E}[\ln p_{\mathbf{x}}(\mathbf{x})] - \mathbb{E}[\ln|\det(W)|] \\
\mathbb{E}[\ln p_{\mathbf{z}}(\mathbf{z})] &= \mathbb{E}[\ln p_{\mathbf{x}}(\mathbf{x})] - \ln|\det(W)|
\end{aligned}$$

The last rewrite is made since W is deterministic.

The entropy can be written as $H(\mathbf{z}) = -\mathbb{E}[\ln p(\mathbf{z})]$, so we obtain

$$\begin{aligned}
\mathbb{E}[\ln p_{\mathbf{z}}(\mathbf{z})] &= \mathbb{E}[\ln p_{\mathbf{x}}(\mathbf{x})] - \ln|\det(W)| \ \Rightarrow \\
-H(\mathbf{z}) &= -H(\mathbf{x}) - \ln|\det(W)|
\end{aligned}$$

Using derivations from previous exercise, we get

$$\begin{aligned}
I(\mathbf{z}) &= -H(\mathbf{z}) + \sum_{i=1}^{l} H(\mathrm{z}_i) \\
&= -H(\mathbf{x}) - \ln|\det(W)| + \sum_{i=1}^{l} H(\mathrm{z}_i)
\end{aligned}$$

Exercise 8.2.3

From the definition we get

$$\begin{aligned}\hat{W} &= \arg\min_W \ I(\mathbf{z}) \\ &= \arg\min_W \ -H(\mathbf{x}) - \ln|\det(W)| + \sum_{i=1}^{l} H(\mathrm{z}_i) \\ &= \arg\min_W \ -H(\mathbf{x}) - \ln|\det(W)| - \sum_{i=1}^{l} \mathbb{E}[\ln p_i(\mathrm{z}_i)])\end{aligned}$$

where we in the last line used $H(\mathbf{z}) = -\mathbb{E}[\ln p(\mathbf{z})]$.

Since $H(\mathbf{x})$ is not a function of W, we can discard that in our optimization problem, and then change the minimization problem to a maximization problem by changing signs

$$\arg\min_W \ I(\mathbf{z}) = \arg\max_W \ \ln|\det(W)| + \mathbb{E}\left[\sum_{i=1}^{l} \ln p_i(\mathrm{z}_i)\right]$$

Exercise 8.2.4

We know that (the rule is given in the exercise text):

$$\frac{d}{dW}\det(W) = W^{-T}\det(W), \quad W^{-T} := (W^{-1})^T$$

and also that:

$$\frac{d}{dx}\ln x = \frac{1}{x}, x > 0$$

Using the chain rule, and assuming that $\det(W) > 0$, we get, since $|\det(W)| = \det(W)$:

$$\begin{aligned}\frac{d}{dW}\ln|\det(W)| &= \frac{\partial \ln\det(W)}{\partial \det(W)} \cdot \frac{d\det(W)}{dW} \\ &= \frac{1}{\det(W)} \cdot W^{-T}\det(W) \\ &= W^{-T}\end{aligned}$$

For $\det(W) < 0$, that is, $|\det(W)| = -\det(W)$, we get:

$$\begin{aligned}\frac{d}{dW}\ln|\det(W)| &= \frac{\partial \ln(-\det(W))}{\partial(-\det(W))} \cdot \frac{d(-\det(W))}{dW} \\ &= \frac{1}{-\det(W)} \cdot (-1) \cdot W^{-T}\det(W) \\ &= W^{-T}\end{aligned}$$

Hence, $\frac{d}{dW}\ln|\det(W)| = W^{-T}$ for $\det(W) \neq 0$.

Exercise 8.2.5

Since we are taking the logarithm to a distribution, we know for sure it will be non-negative. We can also assume that the probability will be greater than zero (albeit infinitely small) since we

are searching for signals that we have a probability to observe, hence we can assume $p_i(\mathrm{z}_i) > 0$ and then $\ln p_i(\mathrm{z}_i)$ is integrable. Since $\log p_z(\mathbf{z}_i)$ is integrable we carry out the interchange of expectation and derivative.

$$\begin{aligned}\frac{d}{dW}\sum_{i=1}^{l}\ln p_i(\mathrm{z}_i) &= \sum_{i=1}^{l}\frac{\partial \ln p_i(\mathrm{z}_i)}{\partial p_i(\mathrm{z}_i)}\frac{dp_i(\mathrm{z}_i)}{dW}\\ &= \sum_{i=1}^{l}\frac{\partial \ln p_i(\mathrm{z}_i)}{\partial p_i(\mathrm{z}_i)}\frac{\partial p_i(\mathrm{z}_i)}{\partial \mathrm{z}_i}\frac{d\mathrm{z}_i}{dW}\\ &= \sum_{i=1}^{l}\frac{1}{p_i(\mathrm{z}_i)}\frac{\partial p_i(\mathrm{z}_i)}{\partial \mathrm{z}_i}\frac{d\mathrm{z}_i}{dW}\end{aligned}$$

Exercise 8.2.6

We first observe from the definition of $\boldsymbol{\phi}(\mathbf{z})$, that we have a vector with the i'th element

$$\boldsymbol{\phi}(\mathbf{z})_i = \frac{1}{p(\mathrm{z}_i)}\frac{\partial p_i(\mathrm{z}_i)}{\partial \mathrm{z}_i}$$

Thus we can rewrite

$$\begin{aligned}\frac{d}{dW}\sum_{i=1}^{l}\ln p_i(\mathrm{z}_i) &= \sum_{i=1}^{l}\frac{1}{p_i(\mathrm{z}_i)}\frac{\partial p_i(\mathrm{z}_i)}{\partial \mathrm{z}_i}\frac{d\mathrm{z}_i}{dW}\\ &= \sum_{i=1}^{l}\boldsymbol{\phi}(\mathbf{z})_i\frac{d\mathrm{z}_i}{dW}\end{aligned}$$

At this point it is easiest to consider the component-wise derivative wrt W. A scalar-by-matrix (of size $k \times m$) derivative is defined as

$$\frac{dx}{dA} = \begin{bmatrix} \frac{dx}{dA_{1,1}} & \frac{dx}{dA_{1,2}} & \cdots & \frac{dx}{dA_{1,m}} \\ \frac{dx}{dA_{2,1}} & \frac{dx}{dA_{2,2}} & \cdots & \frac{dx}{dA_{2,m}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{dx}{dA_{k,1}} & \frac{dx}{dA_{k,2}} & \cdots & \frac{dx}{dA_{k,m}} \end{bmatrix}$$

Hence if we have an expression for $\frac{d\mathrm{z}_i}{dW_{k,m}}$ we also have $\frac{d\mathrm{z}_i}{dW}$. To get an expression for this derivative, we consider the i'th z_i, where we have (from the ICA model):

$$\begin{aligned}\mathrm{z}_i = W_{i,:}^T\mathbf{x} &= \sum_{j=1}^{l}W_{i,j}\mathrm{x}_j\\ &= W_{i,1}\mathrm{x}_1 + W_{i,2}\mathrm{x}_2 + \cdots + W_{i,l}\mathrm{x}_l\end{aligned}$$

We immediately get

$$\begin{aligned}\frac{d\mathrm{z}_i}{dW_{k,m}} &= \frac{d}{dW_{k,m}}(W_{i,1}\mathrm{x}_1 + W_{i,2}\mathrm{x}_2 + \cdots + W_{i,l}\mathrm{x}_l)\\ &= \mathrm{x}_m \quad \text{only if i = k, otherwise the derivative vanish}\end{aligned}$$

Which we can use to create the component-wise derivative

$$\begin{aligned}\frac{d}{dW_{k,m}}\sum_{i=1}^{l}\ln p_i(\mathrm{z}_i) &= \sum_{i=1}^{l}\boldsymbol{\phi}(\mathbf{z})_i\frac{d\mathrm{z}_i}{dW_{k,m}}\\ &= \boldsymbol{\phi}(\mathbf{z})_k\frac{d\mathrm{z}_k}{dW_{k,m}}\\ &= \boldsymbol{\phi}(\mathbf{z})_k\mathrm{x}_m\end{aligned}$$

Writing the full matrix we get

$$\begin{aligned}\frac{d}{dW}\sum_{i=1}^{l}\ln p_i(\mathrm{z}_i) &= \begin{bmatrix}\boldsymbol{\phi}(\mathbf{z})_1\mathrm{x}_1 & \boldsymbol{\phi}(\mathbf{z})_1\mathrm{x}_2 & \cdots & \boldsymbol{\phi}(\mathbf{z})_1\mathrm{x}_m\\ \boldsymbol{\phi}(\mathbf{z})_2\mathrm{x}_1 & \boldsymbol{\phi}(\mathbf{z})_2\mathrm{x}_2 & \cdots & \boldsymbol{\phi}(\mathbf{z})_2\mathrm{x}_m\\ \vdots & \vdots & \ddots & \vdots\\ \boldsymbol{\phi}(\mathbf{z})_k\mathrm{x}_1 & \boldsymbol{\phi}(\mathbf{z})_k\mathrm{x}_2 & \cdots & \boldsymbol{\phi}(\mathbf{z})_k\mathrm{x}_m\end{bmatrix}\\ &= \boldsymbol{\phi}(\mathbf{z})\mathbf{x}^T\end{aligned}$$

Exercise 8.2.7

Gradient descent (equation 5.3 from the book) is written as:

$$\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^{(i-1)} - \mu_i\nabla J(\boldsymbol{\theta}^{(i-1)})$$

In our case, the parameter vector $\boldsymbol{\theta}$ we are optimizing is W, so plugging in we get

$$W^{(i)} = W^{(i-1)} - \mu_i\left(\left(\left(W^{(i-1)}\right)^{-1}\right)^T + \mathbb{E}\left[\boldsymbol{\phi}(\mathbf{z})\mathbf{x}^T\right]\right)$$

Denote $(W^{-1})^T := W^{-T}$. Our model for $\mathbf{x}$ can now be rewritten as

$$\mathbf{x} = W^{-1}\mathbf{z} \Rightarrow \mathbf{x}^T = \mathbf{z}^T W^{-T}$$

Substituting $\mathbf{x}^T$ then yields the final result:

$$\begin{aligned}W^{(i)} &= W^{(i-1)} - \mu_i\left(W^{(i-1)}\right)^{-T} + \mathbb{E}\left[\boldsymbol{\phi}(\mathbf{z})\mathbf{z}^T\left(W^{(i-1)}\right)^{-T}\right]\\ &= W^{(i-1)} - \mu_i\left(W^{(i-1)}\right)^{-T} + \mathbb{E}\left[\boldsymbol{\phi}(\mathbf{z})\mathbf{z}^T\right]\left(W^{(i-1)}\right)^{-T}\\ &= W^{(i-1)} - \mu_i\left(I + \mathbb{E}\left[\boldsymbol{\phi}(\mathbf{z})\mathbf{z}^T\right]\right)\left(W^{(i-1)}\right)^{-T}\end{aligned}$$

02471 Machine Learning for Signal Processing

Solution

Exercise 9: Bayesian inference and the EM algorithm

9.1 Cost functions, Maximum Likelihood and Bayesian Inference

Exercise 9.1.1

The multivariate normal distribution (or multivariate Gaussian distribution) is

$$p(\boldsymbol{y}|\boldsymbol{\theta};\boldsymbol{\mu_y},\Sigma_{\boldsymbol{y}}) = \frac{1}{(2\pi)^{N/2}|\Sigma_{\boldsymbol{y}}|^{1/2}}\exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu_y})^T\Sigma_{\boldsymbol{y}}^{-1}(\boldsymbol{x}-\boldsymbol{\mu_y})\right)$$

The log to this expression, using the rules $\ln ab = \ln a + \ln b$ and $\ln a^b = b\ln a$ becomes

$$\begin{aligned}\ln p(\boldsymbol{y}|\boldsymbol{\theta};\boldsymbol{\mu_y},\Sigma_{\boldsymbol{y}}) &= \ln(2\pi)^{-N/2} + \ln|\Sigma_{\boldsymbol{y}}|^{-1/2} - \frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu_y})^T\Sigma^{-1}(\boldsymbol{y}-\boldsymbol{\mu_y}) \\ &= -\frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma_{\boldsymbol{y}}| - \frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu_y})^T\Sigma_{\boldsymbol{y}}^{-1}(\boldsymbol{y}-\boldsymbol{\mu_y})\end{aligned}$$

We need to find the expression for $\boldsymbol{\mu_y}$ which is

$$\begin{aligned}\boldsymbol{\mu_y} &= \mathbb{E}[\boldsymbol{y}] \\ &= \mathbb{E}[f(X,\boldsymbol{\theta}) + \eta] \\ &= f(X,\boldsymbol{\theta}) + \mathbb{E}[\eta]\end{aligned}$$

If we assume zero-mean noise, $\mathbb{E}[\eta] = 0$, we have $\mathbb{E}[\boldsymbol{y}] = f(X,\boldsymbol{\theta})$. Additionally we need to find the expression for $\Sigma_{\boldsymbol{y}}$:

$$\begin{aligned}\Sigma_{\boldsymbol{y}} &= \mathbb{E}\Big[(\boldsymbol{y}-\mathbb{E}[\boldsymbol{y}])(\boldsymbol{y}-\mathbb{E}[\boldsymbol{y}])^T\Big] \\ &= \mathbb{E}\Big[(f(X,\boldsymbol{\theta})+\eta-f(X,\boldsymbol{\theta}))(f(X,\boldsymbol{\theta})+\eta-f(X,\boldsymbol{\theta}))^T\Big] \\ &= \mathbb{E}\big[\eta\eta^T\big] \\ &= \Sigma_\eta\end{aligned}$$

By substitution we now obtain

$$\begin{aligned}\ln p(\boldsymbol{y}|\boldsymbol{\theta};\boldsymbol{\mu_y},\Sigma_{\boldsymbol{y}}) &= -\frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma_{\boldsymbol{y}}| - \frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu_y})^T\Sigma_{\boldsymbol{y}}^{-1}(\boldsymbol{y}-\boldsymbol{\mu_y}) \\ &= -\frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma_\eta| - \frac{1}{2}(\boldsymbol{y}-f(X,\boldsymbol{\theta}))^T\Sigma_\eta^{-1}(\boldsymbol{y}-f(X,\boldsymbol{\theta}))\end{aligned}$$

Exercise 9.1.2

If we assume that we have noise that is statistically independent sample to sample (e.g white noise), and assume $\Sigma_\eta = \sigma^2 I$. In that case, we have $|\Sigma_\eta| = |\sigma^2 I| = \sigma^{2N}$, and $\Sigma_\eta^{-1} = (\sigma^2 I)^{-1} =$

$\frac{1}{\sigma^2}I$. Thus we can rewrite

$$\begin{aligned}
\ln p(\boldsymbol{y}|\boldsymbol{\theta};\boldsymbol{\mu_y},\Sigma_{\boldsymbol{y}}) &= -\frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma_\eta| - \frac{1}{2}(\boldsymbol{y}-f(X,\boldsymbol{\theta}))^T\Sigma_\eta^{-1}(\boldsymbol{y}-f(X,\boldsymbol{\theta})) \\
&= -\frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln\sigma^{2N} - \frac{1}{2}(\boldsymbol{y}-f(X,\boldsymbol{\theta}))^T\frac{1}{\sigma^2}I(\boldsymbol{y}-f(X,\boldsymbol{\theta})) \\
&= -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(\boldsymbol{y}-f(X,\boldsymbol{\theta}))^T(\boldsymbol{y}-f(X,\boldsymbol{\theta})) \\
&= -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\|\boldsymbol{y}-f(X,\boldsymbol{\theta})\|^2
\end{aligned}$$

Exercise 9.1.3

If we consider this as an optimization problem we have

$$\begin{aligned}
\hat{\boldsymbol{\theta}} &= \arg\max_{\boldsymbol{\theta}} \ln p(\boldsymbol{\theta}|\boldsymbol{y};\boldsymbol{\mu_y},\Sigma_{\boldsymbol{y}}) \\
&= \arg\max_{\boldsymbol{\theta}} \ln p(\boldsymbol{y}|\boldsymbol{\theta};\boldsymbol{\mu_y},\Sigma_{\boldsymbol{y}}) + \ln p(\boldsymbol{\theta}) \\
&= \arg\max_{\boldsymbol{\theta}} -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\|\boldsymbol{y}-f(X,\boldsymbol{\theta})\|^2 + \ln p(\boldsymbol{\theta}) \\
&= \arg\max_{\boldsymbol{\theta}} -\frac{1}{2\sigma^2}\|\boldsymbol{y}-f(X,\boldsymbol{\theta})\|^2 + \ln p(\boldsymbol{\theta})
\end{aligned}$$

If we consider the prior as constant we can remove that from the optimization problem, thus we get

$$\begin{aligned}
\hat{\boldsymbol{\theta}} &= \arg\max_{\boldsymbol{\theta}} -\frac{1}{2\sigma^2}\|\boldsymbol{y}-f(X,\boldsymbol{\theta})\|^2 \\
&= \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{y}-f(X,\boldsymbol{\theta})\|^2
\end{aligned}$$

Exercise 9.1.4

Reusing the expression from the previous exercise we get

$$\begin{aligned}
\ln p(\boldsymbol{\theta};\boldsymbol{0},\sigma_\theta^2 I) &= -\frac{K}{2}\ln(2\pi) - \frac{K}{2}\ln\sigma_\theta^2 - \frac{1}{2\sigma_\theta^2}\|\boldsymbol{\theta}-\boldsymbol{0}\|^2 \\
&= -\frac{K}{2}\ln(2\pi) - \frac{K}{2}\ln\sigma_\theta^2 - \frac{1}{2\sigma_\theta^2}\|\boldsymbol{\theta}\|^2
\end{aligned}$$

Let us combine this result with the log-posterior we derived in the last exercise

$$\begin{aligned}
\hat{\boldsymbol{\theta}} &= \arg\max_{\boldsymbol{\theta}} -\frac{1}{2\sigma^2}\|\boldsymbol{y}-f(X,\boldsymbol{\theta})\|^2 - \frac{K}{2}\ln(2\pi) - \frac{K}{2}\ln\sigma_\theta^2 - \frac{1}{2\sigma_\theta^2}\|\boldsymbol{\theta}\|^2 \\
&= \arg\max_{\boldsymbol{\theta}} -\frac{1}{2\sigma^2}\|\boldsymbol{y}-f(X,\boldsymbol{\theta})\|^2 - \frac{1}{2\sigma_\theta^2}\|\boldsymbol{\theta}\|^2 \\
&= \arg\min_{\boldsymbol{\theta}} \frac{1}{2\sigma^2}\|\boldsymbol{y}-f(X,\boldsymbol{\theta})\|^2 + \frac{1}{2\sigma_\theta^2}\|\boldsymbol{\theta}\|^2
\end{aligned}$$

If we reparameterize with $\sigma_\theta^2 = \frac{\sigma^2}{\lambda} \Leftrightarrow \lambda = \frac{\sigma^2}{\sigma_\theta}$, we get

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \arg\min_{\boldsymbol{\theta}} \frac{1}{2\sigma^2}\|\boldsymbol{y} - f(X, \boldsymbol{\theta})\|^2 + \frac{1}{2\frac{\sigma^2}{\lambda}}\|\boldsymbol{\theta}\|^2 \\ &= \arg\min_{\boldsymbol{\theta}} \frac{1}{2\sigma^2}\|\boldsymbol{y} - f(X, \boldsymbol{\theta})\|^2 + \frac{\lambda}{2\sigma^2}\|\boldsymbol{\theta}\|^2 \\ &= \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{y} - f(X, \boldsymbol{\theta})\|^2 + \lambda\|\boldsymbol{\theta}\|^2\end{aligned}$$

Exercise 9.1.5

We consider the log to the univariate Laplacian distribution

$$\begin{aligned}\ln p(x|\mu, b) &= \ln\left(\frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right)\right) \\ &= \ln(2b)^{-1} - \frac{1}{b}|x-\mu| \\ &= -\ln 2 - \ln b - \frac{1}{b}|x-\mu|\end{aligned}$$

Exercise 9.1.6

Let is now consider a weight vector $\boldsymbol{\theta}$ of length l. If we assume each θ_k follows a zero-mean Laplacian distribution, and the individual weights are statistical independent, we get

$$\begin{aligned}\ln p(\boldsymbol{\theta}|0, b) &= \sum_{i=1}^{l} -\ln 2 - \ln b - \frac{1}{b}|\theta_i| \\ &= -l\ln 2 - l\ln b - \frac{1}{b}\sum_{i=1}^{l}|\theta_i| \\ &= -l\ln 2 - l\ln b - \frac{1}{b}\|\boldsymbol{\theta}\|_1\end{aligned}$$

Exercise 9.1.7

Combine this with the previous results, and obtain the compete log-likelihood $\boldsymbol{\theta}$

$$\begin{aligned}\ln p(\boldsymbol{\theta}, \boldsymbol{y}|X) &= \ln p(\boldsymbol{y}|\boldsymbol{\theta}; f(X, \boldsymbol{\theta}), \sigma^2 I) + \ln p(\boldsymbol{\theta}|0, b) \\ &= -\frac{N}{2}\ln 2\pi - \frac{N}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\|(\boldsymbol{y} - f(X, \boldsymbol{\theta}))\|^2 - l\ln 2 - l\ln b - \frac{1}{b}\|\boldsymbol{\theta}\|_1\end{aligned}$$

Exercise 9.1.8

From Bayes formula, we know, given a dataset X, optimizing $\ln p(\boldsymbol{\theta}, \boldsymbol{y}|X)$ is the same as optimizing $\ln p(\boldsymbol{\theta}|\boldsymbol{y}|X)$. Disregarding all terms not related to $\boldsymbol{\theta}$ we get

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

If we reparameterize with $b = 2\sigma^2/\lambda$ we get the Lasso cost function and we can see that LASSO corresponds to having normal likelihood with i.id.d samples and univariate Laplace prior on $\boldsymbol{\theta}$.

9.2 Derive EM updates for Bayesian linear regression

Exercise 9.2.1

We have already derived expressions for these in the previous exercise. Using the previous results we get:

$$\begin{aligned}\ln p(\boldsymbol{y}, \boldsymbol{\theta}|\alpha, \beta) &= \ln p(\boldsymbol{y}|\boldsymbol{\theta}; \boldsymbol{\theta}, \beta) + \ln p(\boldsymbol{\theta}; \mathbf{0}, \alpha) \\ &= -\frac{N}{2}\ln(2\pi) + \frac{N}{2}\ln\beta - \frac{\beta}{2}\|\boldsymbol{y} - \Phi\boldsymbol{\theta}\|^2 - \frac{K}{2}\ln(2\pi) + \frac{K}{2}\ln\alpha - \frac{\alpha}{2}\|\boldsymbol{\theta}\|^2 \\ &= -\frac{1}{2}(N+K)\ln(2\pi) + \frac{N}{2}\ln\beta - \frac{\beta}{2}\|\boldsymbol{y} - \Phi\boldsymbol{\theta}\|^2 + \frac{K}{2}\ln\alpha - \frac{\alpha}{2}\|\boldsymbol{\theta}\|^2\end{aligned}$$

To compute the expectation we use the following rule $A^TA = \text{trace}(AA^T)$, and use that trace is a linear operator i.e. $\text{trace}(A+B) = \text{trace}(A) + \text{trace}(B)$, and $\mathbb{E}[\text{trace}(A)] = \text{trace}(\mathbb{E}[A])$:

$$\begin{aligned}A := \mathbb{E}\big[\theta^T\theta\big] &= \mathbb{E}\big[\text{trace}(\theta\theta^T)\big] \\ &= \text{trace}\big(\mathbb{E}\big[\theta\theta^T\big]\big)\end{aligned}$$

We recognize $\mathbb{E}\big[\theta\theta^T\big]$ as the structure of the correlation matrix eq (2.33), hence we have, at step j

$$\begin{aligned}\mathbb{E}\big[\theta\theta^T\big] &= \text{Cov}(\theta) + \mathbb{E}[\theta]\mathbb{E}\big[\theta^T\big] \\ &= \Sigma_{\theta|y}^{(j)} + \boldsymbol{\mu}_{\theta|y}^{(j)}\boldsymbol{\mu}_{\theta|y}^{(j)T}\end{aligned}$$

Inserting into the trace we get $\text{trace}(A+B) = \text{trace}(A) + \text{trace}(B)$, and $\mathbb{E}[\text{trace}(A)] = \text{trace}(\mathbb{E}[A])$, we get

$$\begin{aligned}A &= \text{trace}\Big(\Sigma_{\theta|y}^{(j)} + \boldsymbol{\mu}_{\theta|y}^{(j)}\boldsymbol{\mu}_{\theta|y}^{(j)T}\Big) \\ &= \text{trace}\Big(\Sigma_{\theta|y}^{(j)}\Big) + \text{trace}\Big(\boldsymbol{\mu}_{\theta|y}^{(j)}\boldsymbol{\mu}_{\theta|y}^{(j)T}\Big) \\ &= \text{trace}\Big(\Sigma_{\theta|y}^{(j)}\Big) + \|\boldsymbol{\mu}_{\theta|y}^{(j)}\|^2\end{aligned}$$

Exercise 9.2.2

The other term we need to evaluate is $\|\boldsymbol{y} - \Phi\boldsymbol{\theta}\|^2$. To evaluate, we again use the $\text{trace}(\cdot)$ function and perform the following rewrite

$$\begin{aligned}\|\boldsymbol{y} - \Phi\boldsymbol{\theta}\|^2 &= (\boldsymbol{y} - \Phi\boldsymbol{\theta})^T(\boldsymbol{y} - \Phi\boldsymbol{\theta}) \\ &= \boldsymbol{y}^T\boldsymbol{y} - (\Phi\boldsymbol{\theta})^T\boldsymbol{y} - \boldsymbol{y}^T\Phi\boldsymbol{\theta} + (\Phi\boldsymbol{\theta})^T\Phi\boldsymbol{\theta} \\ &= \boldsymbol{y}^T\boldsymbol{y} - (\Phi\boldsymbol{\theta})^T\boldsymbol{y} - \boldsymbol{y}^T\Phi\boldsymbol{\theta} + \text{trace}\big(\Phi\boldsymbol{\theta}(\Phi\boldsymbol{\theta})^T\big) \\ &= \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\Phi\boldsymbol{\theta} - \boldsymbol{y}^T\Phi\boldsymbol{\theta} + \text{trace}\big(\Phi\boldsymbol{\theta}\boldsymbol{\theta}^T\Phi^T\big) \\ &= \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T\Phi\boldsymbol{\theta} + \text{trace}\big(\Phi\boldsymbol{\theta}\boldsymbol{\theta}^T\Phi^T\big)\end{aligned}$$

To proceed we now take the expectation to $\|\boldsymbol{y} - \Phi\boldsymbol{\theta}\|^2$, where $\boldsymbol{\theta}$ is the only random variable, and again using that $\text{trace}(\cdot)$ is a linear operator we get

$$\begin{aligned}B := \mathbb{E}\big[\|\boldsymbol{y} - \Phi\theta\|^2\big] &= \mathbb{E}\big[\boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T\Phi\theta + \text{trace}\big(\Phi\theta\theta^T\Phi^T\big)\big] \\ &= \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T\Phi\mathbb{E}[\theta] + \text{trace}\big(\Phi\mathbb{E}\big[\theta\theta^T\big]\Phi^T\big)\end{aligned}$$

Exercise 9.2.3

We have already found the expressions for $\mathbb{E}[\theta]$ and $\mathbb{E}\big[\theta\theta^T\big]$ earlier, so by substitution, and again using that trace$(\cdot)$ is a linear operator we get

$$\begin{aligned} B &= \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T\Phi\boldsymbol{\mu}_{\theta|y}^{(j)} + \text{trace}\Big(\Phi\Big(\Sigma_{\theta|y}^{(j)} + \boldsymbol{\mu}_{\theta|y}^{(j)}\boldsymbol{\mu}_{\theta|y}^{(j)T}\Big)\Phi^T\Big) \\ &= \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T\Phi\boldsymbol{\mu}_{\theta|y}^{(j)} + \text{trace}\Big(\Phi\Sigma_{\theta|y}^{(j)}\Phi^T\Big) + \text{trace}\Big(\Phi\boldsymbol{\mu}_{\theta|y}^{(j)}\boldsymbol{\mu}_{\theta|y}^{(j)T}\Phi^T\Big) \\ &= \|\boldsymbol{y} - \Phi\boldsymbol{\mu}_{\theta|y}^{(j)}\|^2 + \text{trace}\Big(\Phi\Sigma_{\theta|y}^{(j)}\Phi^T\Big) \end{aligned}$$

Exercise 9.2.4

From the book, sec 12.9.4 we have expressions for how to specify the posterior. From eq. (12.135) and eq. (12.136) we have, if

$$\begin{aligned} p(\boldsymbol{z}) &= \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_z, \Sigma_z) \\ p(\boldsymbol{t}|\boldsymbol{z}) &= \mathcal{N}(\boldsymbol{t}|\boldsymbol{z}; A\boldsymbol{z}, \Sigma_{t|z}) \end{aligned}$$

then the posterior is

$$\begin{aligned} p(\boldsymbol{z}|\boldsymbol{t}) &= \mathcal{N}(\boldsymbol{z}|\boldsymbol{t}; \boldsymbol{\mu}_{z|t}, \Sigma_{z|t}) \\ \boldsymbol{\mu}_{z|t} &= \boldsymbol{\mu}_z + \Sigma_{z|t}A^T\Sigma_{t|z}^{-1})(\boldsymbol{t} - A\boldsymbol{\mu}_z) \\ \Sigma_{z|t} &= (\Sigma_z^{-1} + A^T\Sigma_{t|z}^{-1}A)^{-1} \end{aligned}$$

In our case, we have $\boldsymbol{z} := \boldsymbol{\theta}$, $\boldsymbol{\mu}_z := \boldsymbol{0}$, $\boldsymbol{t} := \boldsymbol{y}$, $\Sigma_z^{-1} := \alpha I$, $\boldsymbol{t} := \boldsymbol{y}$, $A := \Phi$, and $\Sigma_{t|z}^{-1} := \beta I$. Then we get the following expressions

$$\begin{aligned} \boldsymbol{\mu}_{\theta|y} &= \beta\Sigma_{\theta|y}\Phi^T\boldsymbol{y} \\ \Sigma_{\theta|y} &= (\alpha I + \beta\Phi^T\Phi)^{-1} \end{aligned}$$

Exercise 9.2.5

The derivative of $\mathcal{Q}\big(\alpha, \beta; \alpha^{(j)}, \beta^{(j)}\big)$ follows the same structure, so we only show one of them.

$$\begin{aligned} \frac{\partial}{\partial\alpha}\mathcal{Q}\big(\alpha, \beta; \alpha^{(j)}, \beta^{(j)}\big) = \frac{K}{2}\frac{1}{\alpha} - \frac{1}{2}A = 0, \quad &\Leftrightarrow \\ \frac{1}{\alpha} = \frac{A}{K}, \quad &\Leftrightarrow \\ \alpha = \frac{K}{A} \end{aligned}$$

By symmetry, we get

$$\beta = \frac{N}{B}$$

Hence, the update equations will be $\alpha^{j+1} = K/B$ and $\beta^{j+1} = N/B$.

02471 Machine Learning for Signal Processing

Solution

Exercise 10: State-space models – the Hidden Markov Model

10.1 Probabilities in HMM

Exercise 10.1.1

Use marginalization (sum rule)

$$\begin{aligned} P(y_1) &= \sum_{i=1}^{K} P(y_1|x_1=i)P(x_1=i) \\ P(y_1=2) &= \sum_{i=1}^{K} P(y_1=2|x_1=i)P(x_1=i) \\ &= P(y_1=2|x_1=1)P(x_1=1) + P(y_1=2|x_1=2)P(x_1=2) \\ &= 0.6 \cdot 0.7 + 0.2 \cdot 0.3 \\ &= 0.48 \end{aligned}$$

Exercise 10.1.3

Use marginalization (sum rule)

$$\begin{aligned} P(y_1,y_2) &= \sum_{i=1}^{K}\sum_{j=1}^{K} P(y_1,y_2|x_1=i,x_2=j)P(x_1=i,x_2=j) \\ &= \sum_{i=1}^{K}\sum_{j=1}^{K} P(y_2|y_1,x_1=i,x_2=j)P(y_1|x_1=i,x_2=j)P(x_2=j|x_1=i)P(x_1=i) \\ &= \sum_{i=1}^{K}\sum_{j=1}^{K} P(y_2|x_2=j)P(y_1|x_1=i)P(x_2=j|x_1=i)P(x_1=i) \end{aligned}$$

where, in the last line, we have used the information from the HMM graph to remove unneeded conditionals.

This operation scales with K^2.

Exercise 10.1.5

$$
\begin{aligned}
P(y_2, y_1) &= \sum_{x_2} P(y_1, y_2, x_2) \\
P(y_1, y_2, x_2) &= \sum_{x_1} P(y_1, y_2, x_1, x_2) \\
&= \sum_{x_1} P(y_2|y_1, x_1, x_2) P(y_1, x_1, x_2) \\
&= \sum_{x_1} P(y_2|y_1, x_1, x_2) P(x_2|y_1, x_1) P(y_1, x_1)
\end{aligned}
$$

Removing the terms that are not needed for the conditionals, we get

$$
\begin{aligned}
P(y_1, y_2, x_2) &= \sum_{x_1} P(y_2|x_2) P(x_2|x_1) P(y_1, x_1) \\
&= P(y_2|x_2) \sum_{x_1} P(x_2|x_1) P(y_1, x_1)
\end{aligned}
$$

Exercise 10.1.6

Using $\alpha(x_n) := P(y_{[1:n]}, x_n)$ we can write

$$
\begin{aligned}
P(y_1, y_2, x_2) = \alpha(x_2) &= P(y_2|x_2) \sum_{x_1} P(x_2|x_1) P(y_1, x_1) \\
&= P(y_2|x_2) \sum_{x_1} P(x_2|x_1) \alpha(x_1)
\end{aligned}
$$

Notice that this case is general, i.e. we could have used n and $n-1$ as time index instead of 2 and 1, and all derivations are still correct. By that, we get the recursive formula:

$$
\alpha(x_n) = P(y_n|x_n) \sum_{x_{n-1}} P(x_n|x_{n-1}) \alpha(x_{n-1})
$$

All of this is derived for the case where y_i is a discrete random variable, but we did not use that property in the derivation, so the result also holds for (multivariate) continuous random variables. In that case, the capital $P(\cdot)$ is replaced with $p(\cdot)$ for the expressions that go over a distribution for $\mathbf{y}$.

This formula contains $K + 1$ multiplications and K additions.

Exercise 10.1.7

We can exploit this result further using Bayes formula

$$\begin{aligned} P(X|Y) &= \frac{P(Y,X)}{P(Y)} \quad \Rightarrow \\ P(x_n|y_{[1:n]}) &= \frac{P(y_{[1:n]}, x_n)}{P(y_{[1:n]})} \\ &= \frac{\alpha(x_n)}{P(y_{[1:n]})} \end{aligned}$$

We get an even more efficient formula since, from the sum formula we have

$$\begin{aligned} P(y_{[1:n]}) &= \sum_{x_n} P(y_{[1:n]}, x_n) \\ &= \sum_{x_n} \alpha(x_n) \end{aligned}$$

Combining this yields

$$P(x_n|y_{[1:n]}) = \frac{\alpha(x_n)}{\sum_{x_n} \alpha(x_n)}$$

$P(y_{[1:n]})$ has K additions, so including the calculations for computing $\alpha(x_n)$, we get $K+n(2K+1) = \mathcal{O}(nK)$ operations, as opposed to the direct implementation that had $\mathcal{O}(K^N)$ operations.

10.2 HMM model formulation and EM updates

There are no explicit solutions for exercise 10.1.1–10.1.3. The book readily derives the expressions. We'll provide reading directions instead.

Exercise 10.2.1

Use the information from how the model parameters are setup (sec 16.5.1, page 847), and then derive eq. (16.32)–(16.35).

Exercise 10.2.2

This derivation is described in sec 16.5.2, page 852, eq (16.52).

Exercise 10.2.3

This derivation is described in sec 16.5.2, page 853, eq (16.53)–(16.55).

Exercise 10.2.4

The maximization step becomes

$$\begin{aligned}\mathcal{Q}\big(\Theta, \Theta^{(t)}\big) = &\sum_{k=1}^{K} \gamma\big(x_{1,k} = 1; \Theta^{(t)}\big) \ln P_k \\ &+ \sum_{n=2}^{N} \sum_{i=1}^{K} \sum_{j=1}^{K} \xi\big(x_{n-1,j} = 1, x_{n,i} = 1; \Theta^{(t)}\big) \ln P_{ij} \\ &+ \text{ constant}\end{aligned}$$

where the constant involves parameters independent of P_k, P_{ij}. Since P_k and P_{ij} are decoupled, they can be solved independently.

Since each row of the matrix containing P_{ij} is a discrete distribution, each row much sum to one. We have K states, hence we will have K rows and thus K constraints

$$\sum_{k=1}^{K} P_{kj} = 1, \qquad j = 1, \cdots, K$$

The Lagrangian then becomes

$$L(P_{ij}, \lambda) = \sum_{n=2}^{N} \sum_{i=1}^{K} \sum_{j=1}^{K} \xi\big(x_{n-1,j} = 1, x_{n,i} = 1; \Theta^{(t)}\big) \ln P_{ij} - \lambda \left(\sum_{k=1}^{K} P_{kj} - 1 \right)$$

Taking the derivative with respect to P_{ij} and equating to zero we get,

$$\frac{1}{\lambda} \sum_{n=2}^{N} \xi\big(x_{n-1,j} = 1, x_{n,i} = 1; \Theta^{(t)}\big) = P_{ij}$$

and plugging into the constraint, in order to compute λ, we obtain

$$\begin{aligned}&\sum_{k=1}^{K} \frac{1}{\lambda} \sum_{n=2}^{N} \xi\big(x_{n-1,j} = 1, x_{n,k} = 1; \Theta^{(t)}\big) = 1 \\ \Rightarrow &\frac{1}{\sum_{n=2}^{N} \sum_{k=1}^{K} \xi(x_{n-1,j} = 1, x_{n,k} = 1; \Theta^{(t)})} = \frac{1}{\lambda}\end{aligned}$$

Substituting $\frac{1}{\lambda}$ and adding the iteration index $(t+1)$ then yields

$$P_{ij}^{(t+1)} = \frac{\sum_{n=2}^{N} \xi\big(x_{n-1,j} = 1, x_{n,i} = 1; \Theta^{(t)}\big)}{\sum_{n=2}^{N} \sum_{k=1}^{K} \xi(x_{n-1,j} = 1, x_{n,k} = 1; \Theta^{(t)})}$$

02471 Machine Learning for Signal Processing

Solution

Exercise 11: State-space models – Kalman filtering

11.1 Derivation of the Kalman filter

Exercise 11.1.3

We have

$$P_{n|n-1} = \mathbb{E}\big[\mathbf{e}_{n|n-1}\mathbf{e}_{n|n-1}^T\big]$$

But the error can be rewritten as

$$\begin{aligned}\mathbf{e}_{n|n-1} &= \mathbf{x}_n - \hat{\mathbf{x}}_{n|n-1} \\ &= F_n\mathbf{x}_{n-1} + \boldsymbol{\eta}_n - F_n\hat{\mathbf{x}}_{n-1|n-1} \\ &= F_n(\mathbf{x}_{n-1} - \hat{\mathbf{x}}_{n-1|n-1}) + \boldsymbol{\eta}_n \\ &= F_n\mathbf{e}_{n-1|n-1} + \boldsymbol{\eta}_n\end{aligned}$$

Substituting we get

$$\begin{aligned}P_{n|n-1} &= \mathbb{E}\big[(F_n\mathbf{e}_{n-1|n-1} + \boldsymbol{\eta}_n)(F_n\mathbf{e}_{n-1|n-1} + \boldsymbol{\eta}_n)^T\big] \\ &= \mathbb{E}\big[F_n\mathbf{e}_{n-1|n-1}\mathbf{e}_{n-1|n-1}^T F_n^T + \boldsymbol{\eta}_n\boldsymbol{\eta}_n^T\big] \\ &= F_n\mathbb{E}\big[\mathbf{e}_{n-1|n-1}\mathbf{e}_{n-1|n-1}^T\big]F_n^T + \mathbb{E}\big[\boldsymbol{\eta}_n\boldsymbol{\eta}_n^T\big] \\ &= F_nP_{n-1|n-1}F_n^T + Q_n\end{aligned}$$

Where we have used that $\mathbb{E}\big[F_n\mathbf{e}_{n-1|n-1}\boldsymbol{\eta}_n\big] = F_n\mathbb{E}\big[\mathbf{e}_{n-1|n-1}\big]\mathbb{E}[\boldsymbol{\eta}_n] = 0$, since $\mathbb{E}[\boldsymbol{\eta}_n] = 0$.

Exercise 11.1.4

With these definitions we can now carry out the following rewrites (using $\hat{\mathbf{x}}_{n|n} = \hat{\mathbf{x}}_{n|n-1} - K_n\mathbf{e}_n$)

$$\begin{aligned}P_{n|n} &= \mathbb{E}\big[\mathbf{e}_{n|n}\mathbf{e}_{n|n}^T\big] \\ &= \mathbb{E}\big[(\mathbf{x}_n - \hat{\mathbf{x}}_{n|n})(\mathbf{x}_n - \hat{\mathbf{x}}_{n|n})^T\big] \\ &= \mathbb{E}\big[(\mathbf{x}_n - \hat{\mathbf{x}}_{n|n-1} - K_n\mathbf{e}_n)(\mathbf{x}_n - \hat{\mathbf{x}}_{n|n-1} - K_n\mathbf{e}_n)^T\big] \\ &= \mathbb{E}\big[(\mathbf{e}_{n|n-1} - K_n\mathbf{e}_n)(\mathbf{e}_{n|n-1} - K_n\mathbf{e}_n)^T\big] \\ &= \mathbb{E}\big[\mathbf{e}_{n|n-1}\mathbf{e}_{n|n-1}^T\big] - \mathbb{E}\big[K_n\mathbf{e}_n\mathbf{e}_{n|n-1}^T\big] - \mathbb{E}\big[\mathbf{e}_{n|n-1}\mathbf{e}_n^TK_n^T\big] + \mathbb{E}\big[K_n\mathbf{e}_n\mathbf{e}_n^TK_n^T\big] \\ &= \mathbb{E}\big[\mathbf{e}_{n|n-1}\mathbf{e}_{n|n-1}^T\big] - K_n\mathbb{E}\big[\mathbf{e}_n\mathbf{e}_{n|n-1}^T\big] - \mathbb{E}\big[\mathbf{e}_{n|n-1}\mathbf{e}_n^T\big]K_n^T + K_n\mathbb{E}\big[\mathbf{e}_n\mathbf{e}_n^T\big]K_n^T\end{aligned}$$

Let us inspect the expectations term by term

$$\begin{aligned}
\mathbb{E}\big[\mathbf{e}_n\mathbf{e}_{n|n-1}^T\big] &= \mathbb{E}\big[(\mathbf{y}_n - \hat{\mathbf{y}}_n)\mathbf{e}_{n|n-1}^T\big] \\
&= \mathbb{E}\big[(H_n\mathbf{x}_n + \mathbf{v}_n - H_n\hat{\mathbf{x}}_{n|n-1})\mathbf{e}_{n|n-1}^T\big] \\
&= \mathbb{E}\big[H_n(\mathbf{x}_n - \hat{\mathbf{x}}_{n|n-1})\mathbf{e}_{n|n-1}^T\big] + \mathbb{E}\big[\mathbf{v}_n\mathbf{e}_{n|n-1}^T\big] \\
&= H_n\mathbb{E}\big[\mathbf{e}_{n|n-1}\mathbf{e}_{n|n-1}^T\big] + \mathbb{E}\big[\mathbf{v}_n\mathbf{e}_{n|n-1}^T\big] \\
&= H_nP_{n|n-1} + \mathbb{E}\big[\mathbf{v}_n(\mathbf{x}_n - \hat{\mathbf{x}}_{n|n-1})^T\big] \\
&= H_nP_{n|n-1}
\end{aligned}$$

where the last term vanishes because we assumed $\mathbf{v}_n$ is uncorrelated with $\mathbf{x}_n$ and $\hat{\mathbf{x}}_{n|n-1}$.

Using identical derivations, we get

$$\begin{aligned}
\mathbb{E}\big[\mathbf{e}_{n|n-1}\mathbf{e}_n^T\big] &= \mathbb{E}\big[\mathbf{e}_{n|n-1}(\mathbf{y}_n - \hat{\mathbf{y}}_n)^T\big] \\
&= \mathbb{E}\big[\mathbf{e}_{n|n-1}(H_n\mathbf{x}_n + \mathbf{v}_n - H_n\hat{\mathbf{x}}_{n|n-1})^T\big] \\
&= \mathbb{E}\big[\mathbf{e}_{n|n-1}(\mathbf{x}_n - \hat{\mathbf{x}}_{n|n-1})^TH_n^T\big] + \mathbb{E}\big[\mathbf{e}_{n|n-1}\mathbf{v}_n^T\big] \\
&= \mathbb{E}\big[\mathbf{e}_{n|n-1}\mathbf{e}_{n|n-1}^T\big]H_n^T \\
&= P_{n|n-1}H_n^T
\end{aligned}$$

The last expectation gives

$$\begin{aligned}
\mathbb{E}\big[\mathbf{e}_n\mathbf{e}_n^T\big] &= \mathbb{E}\big[(\mathbf{y}_n - \hat{\mathbf{y}}_n)(\mathbf{y}_n - \hat{\mathbf{y}}_n)^T\big] \\
&= \mathbb{E}\big[(H_n\mathbf{x}_n + \mathbf{v}_n - H_n\hat{\mathbf{x}}_{n|n-1})(H_n\mathbf{x}_n + \mathbf{v}_n - H_n\hat{\mathbf{x}}_{n|n-1})^T\big] \\
&= \mathbb{E}\big[(H_n(\mathbf{x}_n - \hat{\mathbf{x}}_{n|n-1}) + \mathbf{v}_n)(H_n(\mathbf{x}_n - \hat{\mathbf{x}}_{n|n-1}) + \mathbf{v}_n)^T\big] \\
&= \mathbb{E}\big[(H_n\mathbf{e}_{n|n-1} + \mathbf{v}_n)(H_n\mathbf{e}_{n|n-1} + \mathbf{v}_n)^T\big] \\
&= H_n\mathbb{E}\big[\mathbf{e}_{n|n-1}\mathbf{e}_{n|n-1}^T\big]H_n^T + \mathbb{E}\big[\mathbf{v}_n\mathbf{v}_n^T\big] \\
&= H_nP_{n|n-1}H_n^T + R_n
\end{aligned}$$

Let $S = H_nP_{n|n-1}H_n^T + R_n$, so that $\mathbb{E}\big[\mathbf{e}_n\mathbf{e}_n^T\big] = S$, then put it all together to get

$$P_{n|n} = P_{n|n-1} - K_nH_nP_{n|n-1} - P_{n|n-1}H_n^TK_n^T + K_nSK_n^T$$

Exercise 11.1.5

We have to minimize $\text{trace}(P_{n|n})$ with respect to K_n, so we take the derivate w.r.t. K_n and set to zero. We will use that the trace is a linear operator i.e. $\text{trace}(A+B) = \text{trace}(A) + \text{trace}(B)$, and $\text{trace}(A) = \text{trace}(A^T)$, and the following two differentiation rules

$$\begin{aligned}
\frac{\partial\text{trace}(AB)}{\partial A} &= B^T \\
\frac{\partial\text{trace}(ACA^T)}{\partial A} &= 2AC
\end{aligned}$$

First we rewrite a bit

$$\begin{aligned}
\text{trace}(P_{n|n-1}H_n^TK_n^T) &= \text{trace}((P_{n|n-1}H_n^TK_n^T)^T) \\
&= \text{trace}(K_nH_nP_{n|n-1}^T)
\end{aligned}$$

Since $P_{n|n-1}^T = P_{n|n-1}$ we have $\text{trace}(P_{n|n-1}H_n^T K_n^T) = \text{trace}(K_n H_n P_{n|n-1})$ an we get

$$\text{trace}(P_{n|n}) = \text{trace}(P_{n|n-1}) - 2\text{trace}(K_n H_n P_{n|n-1}) + \text{trace}(K_n S K_n^T)$$

Using the rules of differentation specified earlier, we get

$$\begin{aligned}\frac{\partial \text{trace}(P_{n|n})}{\partial K_n} &= \frac{\partial}{\partial K_n}\text{trace}(P_{n|n-1}) - 2\frac{\partial}{\partial K_n}\text{trace}(K_n H_n P_{n|n-1}) + \frac{\partial}{\partial K_n}\text{trace}(K_n S K_n^T) \\ &= -2(H_n P_{n|n-1})^T + 2K_n S\end{aligned}$$

Setting the derivative to zero gives

$$\begin{aligned}2K_n S &= 2(H_n P_{n|n-1})^T \quad \Leftrightarrow \\ K_n &= (H_n P_{n|n-1})^T S^{-1} \\ &= P_{n|n-1}^T H_n^T S^{-1} \\ &= P_{n|n-1} H_n^T S^{-1}\end{aligned}$$

Exercise 11.1.6

We can now go back and finalize the recursion for $P_{n|n}$

$$P_{n|n} = P_{n|n-1} - K_n H_n P_{n|n-1} - P_{n|n-1}H_n^T K_n^T + K_n S K_n^T$$

The last term can be rewritten

$$\begin{aligned}K_n S K_n^T &= P_{n|n-1}H_n^T S^{-1} S K_n^T \\ &= P_{n|n-1}H_n^T K_n^T\end{aligned}$$

Using substitution we get

$$\begin{aligned}P_{n|n} &= P_{n|n-1} - K_n H_n P_{n|n-1} - P_{n|n-1}H_n^T K_n^T + P_{n|n-1}H_n^T K_n^T \\ &= P_{n|n-1} - K_n H_n P_{n|n-1}\end{aligned}$$

02471 Machine Learning for Signal Processing

Solution

Exercise 12: Kernel methods

12.1 Obtaining linear separability using Kernels

Exercise 12.1.1

The points are generated using two uniform distributions. The (x, y) coordinates for the points are generated using the real and imaginary part of the complex exponential function

$$f(\mathrm{r}, \theta) = \mathrm{r}\exp(i\theta), \quad \theta \sim \mathcal{U}[0; 2\pi]$$

For class 1 we have $\mathrm{r} \sim \mathcal{U}[0; 1]$ and for class 2 we have $\mathrm{r} \sim \mathcal{U}[1.5; 2.5]$, where $\mathcal{U}[a, b]$ denotes the uniform distribution on the interval $[a, b]$.

Exercise 12.1.3

$$\begin{aligned}
\boldsymbol{\phi}^T(\boldsymbol{x})\boldsymbol{\phi}(\boldsymbol{y}) &= \begin{bmatrix} x_1^2 & \sqrt{2}x_1x_2 & x_2^2 \end{bmatrix} \begin{bmatrix} y_1^2 \\ \sqrt{2}y_1y_2 \\ y_2^2 \end{bmatrix} \\
&= x_1^2y_1^2 + 2x_1x_2y_1y_2 + x_2^2y_2^2 \\
&= (x_1y_1 + x_2y_2)^2 \\
&= (\boldsymbol{x}^T\boldsymbol{y})^2
\end{aligned}$$

12.2 Derivation of the kernel ridge regression

Exercise 12.2.1

From def 8.15 we use $\langle a\boldsymbol{x}, \boldsymbol{y}\rangle = a\langle \boldsymbol{x}, \boldsymbol{y}\rangle$, and additionally, since we are in $\mathbb{R}$ we have $\langle \boldsymbol{x}, \boldsymbol{y}\rangle = \langle \boldsymbol{y}, \boldsymbol{x}\rangle^* = \langle \boldsymbol{y}, \boldsymbol{x}\rangle$.

Now, by inserting f in the inner product, we get

$$\langle \sum_{n=1}^{N} \theta_n \kappa(\cdot, \boldsymbol{x}_n), \sum_{m=1}^{N} \theta_m \kappa(\cdot, \boldsymbol{x}_m)\rangle = \sum_{n=1}^{N} \theta_n \sum_{m=1}^{N} \theta_m \langle \kappa(\cdot, \boldsymbol{x}_n), \kappa(\cdot, \boldsymbol{x}_m)\rangle$$

We now use the property $\langle \kappa(\cdot, \boldsymbol{y}), \kappa(\cdot, \boldsymbol{x})\rangle = \kappa(\boldsymbol{x}, \boldsymbol{y}) = \kappa(\boldsymbol{y}, \boldsymbol{x})$ (eq. (11.9) in the book) to get

$$\sum_{n=1}^{N} \theta_n \sum_{m=1}^{N} \theta_m \langle \kappa(\cdot, \boldsymbol{x}_n), \kappa(\cdot, \boldsymbol{x}_m)\rangle = \sum_{n=1}^{N} \theta_n \sum_{m=1}^{N} \theta_m \kappa(\boldsymbol{x}_n, \boldsymbol{x}_m)$$

Using $\mathcal{K} = \mathcal{K}^T$ and the definition of the kernel matrix (Eq 11.10–11.12 in the book), we obtain

$$\sum_{n=1}^{N} \theta_n \sum_{m=1}^{N} \theta_m \kappa(\boldsymbol{x}_n, \boldsymbol{x}_m) = \boldsymbol{\theta}^T\mathcal{K}^T\boldsymbol{\theta}$$

Exercise 12.2.2

If we define

$$\kappa(\cdot) := [\kappa(\cdot, \boldsymbol{x}_1), \cdots, \kappa(\cdot, \boldsymbol{x}_n)]^T \quad \Rightarrow$$
$$\sum_{n=1}^{N} \theta_n \kappa(\cdot, \boldsymbol{x}_n) = \boldsymbol{\theta}^T \kappa(\cdot)$$

This allows the following rewrite

$$\sum_{n=1}^{N} \big(y_n - \boldsymbol{\theta}^T \kappa(\cdot)\big)^2 = \|\boldsymbol{y} - \mathcal{K}\boldsymbol{\theta}\|^2 = (\boldsymbol{y} - \mathcal{K}\boldsymbol{\theta})^T(\boldsymbol{y} - \mathcal{K}\boldsymbol{\theta}) = \boldsymbol{y}\boldsymbol{y}^T - 2\boldsymbol{y}^T\mathcal{K}\boldsymbol{\theta} + \boldsymbol{\theta}^T\mathcal{K}^T\mathcal{K}\boldsymbol{\theta}$$

Now using the following properties from Appendix A:

$$\frac{\partial \boldsymbol{x}^T A \boldsymbol{x}}{\partial \boldsymbol{x}} = (A + A^T)\boldsymbol{x}$$
$$\frac{\partial A\boldsymbol{x}}{\partial \boldsymbol{x}} = A^T$$

We obtain

$$\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -2(\boldsymbol{y}^T\mathcal{K})^T + 2\mathcal{K}^T\mathcal{K}\boldsymbol{\theta} + 2C\mathcal{K}^T\boldsymbol{\theta} = 0$$

Which leads to

$$\mathcal{K}^T\boldsymbol{y} = (\mathcal{K}^T\mathcal{K} + C\mathcal{K}^T)\hat{\boldsymbol{\theta}}$$

If K^{-1} exists we then obtain

$$(\mathcal{K}^T)^{-1}\mathcal{K}\boldsymbol{y} = (\mathcal{K}^T)^{-1}\mathcal{K}^T(\mathcal{K} + CI)\hat{\boldsymbol{\theta}} \quad \Rightarrow$$
$$\boldsymbol{y} = (\mathcal{K} + CI)\hat{\boldsymbol{\theta}} \quad \Rightarrow$$
$$\hat{\boldsymbol{\theta}} = (\mathcal{K} + CI)^{-1}\boldsymbol{y}$$

02471 Machine Learning for Signal Processing

Solution

Exercise 13: Support vector regression

13.1 Support vector regression (SVR)

Exercise 13.1.1

We have the model

$$f(\boldsymbol{x}) = \boldsymbol{\theta}^T \boldsymbol{x} + \theta_0$$

$$\mathcal{L}(y, f(\boldsymbol{x})) = \begin{cases} |y - f(\boldsymbol{x})| - \epsilon, & \text{if} \quad |y - f(\boldsymbol{x})| > \epsilon \\ 0, & \text{if} \quad |y - f(\boldsymbol{x})| \leq \epsilon \end{cases}$$

Two cases: if $y_n - f(\boldsymbol{x}_n) \geq \epsilon$ then

$$\begin{aligned} y_n - f(\boldsymbol{x}_n) \geq \epsilon \quad &\Leftrightarrow \\ y_n - (\boldsymbol{\theta}^T \boldsymbol{x}_n + \theta_0) \geq \epsilon \quad &\Leftrightarrow \\ y_n - (\boldsymbol{\theta}^T \boldsymbol{x}_n + \theta_0) \leq \epsilon + \tilde{\xi}_n \end{aligned}$$

where $\tilde{\xi}_n$ is chosen big enough for the inequality to be true. The bound is then

$$\tilde{\xi}_n \geq y_n - \boldsymbol{\theta}^T \boldsymbol{x}_n + \theta_0 - \epsilon \geq 0$$

From this we see that the smallest $\tilde{\xi}_n$ we can choose is $\tilde{\xi}_n = 0$, and if this choice is made when $y_n - \boldsymbol{\theta}^T \boldsymbol{x}_n + \theta_0 \leq \epsilon$. In this case, the loss for that particular point would be 0 (because we have the case $|y - f(\boldsymbol{x})| \leq \epsilon$), so ideally, our optimization would select $\boldsymbol{\theta}$ and θ_0 so that $\tilde{\xi}_n = 0$.

The other case is $y_n - f(\boldsymbol{x}_n) \leq -\epsilon$ then

$$\begin{aligned} y_n - (\boldsymbol{\theta}^T \boldsymbol{x}_n + \theta_0) \leq -\epsilon \quad &\Leftrightarrow \\ \boldsymbol{\theta}^T \boldsymbol{x}_n - y_n - \theta_0 \geq \epsilon \\ \boldsymbol{\theta}^T \boldsymbol{x}_n - y_n - \theta_0 \leq \epsilon + \xi_n \end{aligned}$$

Following similar arguments, $\xi_n \geq 0$, and we ideally want the optimization to end up with $\xi_n = 0$.

That means we can restate the optimization problem (with an added regularization) as

$$\begin{aligned} \underset{\boldsymbol{\theta}, \boldsymbol{\xi}, \tilde{\boldsymbol{\xi}},}{\arg\min} \quad & J(\boldsymbol{\theta}, \boldsymbol{\xi}, \tilde{\boldsymbol{\xi}}) := \frac{1}{2}\|\boldsymbol{\theta}\|^2 + C\left(\sum_{n=1}^{N} \xi_n + \sum_{n=1}^{N} \tilde{\xi}_n\right) \\ \text{s.t.} \quad & y_n - f(\boldsymbol{x}_n) \leq \epsilon + \tilde{\xi}_n \\ & -(y_n - f(\boldsymbol{x}_n)) \leq \epsilon + \xi_n \\ & \tilde{\xi}_n \geq 0 \\ & \xi_n \geq 0 \end{aligned}$$

Exercise 13.1.2

We have 4 constraints that can all be written as

$$\begin{aligned}
\text{s.t.} \quad & y_n - f(\boldsymbol{x}_n) - (\epsilon + \tilde{\xi}_n) \leq 0 \quad \Leftrightarrow \quad y_n - \boldsymbol{\theta}^T \boldsymbol{x}_n - \theta_0 - \epsilon - \tilde{\xi}_n \leq 0 \\
& -(y_n - f(\boldsymbol{x}_n)) - (\epsilon + \xi_n) \leq 0 \quad \Leftrightarrow \quad \boldsymbol{\theta}^T \boldsymbol{x}_n + \theta_0 - y_n - \epsilon - \xi_n \leq 0 \\
& \tilde{\xi_n} \geq 0 \\
& \xi_n \geq 0
\end{aligned}$$

Then by using the results from C.2 as stated in the exercise, we introduce the Lagrange multipliers $\tilde{\lambda}_n$, λ_n, $\tilde{\mu_n}$, $\mu_n \geq 0$ to obtain

$$\begin{aligned}
& \tilde{\lambda}_n(y_n - \boldsymbol{\theta}^T \boldsymbol{x}_n - \theta_0 - \epsilon - \tilde{\xi_n}) = 0 \\
& \lambda_n(\boldsymbol{\theta}^T \boldsymbol{x}_n + \theta_0 - y_n - \epsilon - \xi_n) = 0 \\
& \tilde{\mu}_n \tilde{\xi_n} = 0 \\
& \mu_n \xi_n = 0
\end{aligned}$$

Using the result:

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = J(\boldsymbol{\theta}) - \sum_{i=1}^{m} \lambda_i f_i(\boldsymbol{\theta})$$

will readily give the problem stated in the exercise.

Exercise 13.1.3

We need the rules

$$\begin{aligned}
\frac{\partial a^T \boldsymbol{x}}{\partial \boldsymbol{x}} &= \frac{\partial \boldsymbol{x}^T a}{\partial \boldsymbol{x}} = a \\
\frac{\partial \boldsymbol{x}^T A \boldsymbol{x}}{\partial \boldsymbol{x}} &= (A + A^T)\boldsymbol{x}
\end{aligned}$$

Then we get (only including the terms with $\boldsymbol{\theta}$)

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{L} &= \frac{\partial}{\partial \boldsymbol{\theta}} \left(\frac{1}{2} \boldsymbol{\theta}^T \boldsymbol{\theta} - \sum_{n=1}^{N} \tilde{\lambda}_n \boldsymbol{\theta}^T \boldsymbol{x}_n + \sum_{n=1}^{N} \lambda_n \boldsymbol{\theta}^T \boldsymbol{x}_n \right) \\
&= \boldsymbol{\theta} - \sum_{n=1}^{N} \tilde{\lambda}_n \boldsymbol{x}_n + \sum_{n=1}^{N} \lambda_n \boldsymbol{x}_n
\end{aligned}$$

Setting this derivate to zero gives

$$\begin{aligned}
\hat{\boldsymbol{\theta}} &= \sum_{n=1}^{N} \tilde{\lambda}_n \boldsymbol{x}_n - \sum_{n=1}^{N} \lambda_n \boldsymbol{x}_n \\
&= \sum_{n=1}^{N} (\tilde{\lambda}_n - \lambda_n) \boldsymbol{x}_n
\end{aligned}$$

If we have $f(\boldsymbol{x}) = \boldsymbol{\theta}^T \boldsymbol{x} + \theta_0$ we get the following prediction

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

For the next derivate we get (only including the terms for ξ_n)

$$\begin{aligned}\frac{\partial}{\partial \xi_n}\mathcal{L} &= \frac{\partial}{\partial \xi_n}\left(C\sum_{n=1}^{N}\xi_n + \sum_{n=1}^{N}\lambda_n(-\xi_n) - \sum_{n=1}^{N}\mu_n\xi_n\right) \\ &= C - \lambda_n - \mu_n\end{aligned}$$

Setting this derivate to zero gives

$$\lambda_n + \mu_n = C$$

Since both $\mu_n \geq 0$ and $\lambda_n \geq 0$, we can deduce the constraint

$$0 \leq \lambda_n \leq \ C$$

And finally for θ_0 we get

$$\frac{\partial}{\partial \theta_0}\mathcal{L} = -\sum_{n=1}^{N}\tilde{\lambda}_n + \sum_{n=1}^{N}\lambda_n$$

Setting the derivative to zero gives

$$\sum_{n=1}^{N}\lambda_n = \sum_{n=1}^{N}\tilde{\lambda}_n$$

Exercise 13.1.4

For $\xi_n > 0$, the reason why this implies $\mu_n = 0$ is because we have a constraint $\xi_n\mu_n = 0$. The second implication is then due to the solution from previous exercise: $C = \lambda_n + \mu_n$. If $\mu_n = 0$, then $C = \lambda_n$.

For $\xi_n = 0$, we get

$$\begin{aligned}\lambda_n(\boldsymbol{\theta}^T\boldsymbol{x}_n + \theta_0 - y_n - \epsilon - \xi_n) = 0 \quad &\Rightarrow \\ \lambda_n(\boldsymbol{\theta}^T\boldsymbol{x}_n + \theta_0 - y_n - \epsilon) = 0 \quad &\end{aligned}$$

If the prediction is exactly $-\epsilon$ off, we have $y_n - \boldsymbol{\theta}^T\boldsymbol{x}_n + \theta_0 = -\epsilon$, which leads to

$$\begin{aligned}\lambda_n(\boldsymbol{\theta}^T\boldsymbol{x}_n + \theta_0 - y_n - \epsilon - \xi_n) = 0 \quad &\Rightarrow \\ \lambda_n(\epsilon - \epsilon) = 0 \quad &\end{aligned}$$

This implies that λ_n can be any value. By similar arguments, we see that if $y_n - \boldsymbol{\theta}^T\boldsymbol{x}_n + \theta_0 < \epsilon$, the same constraint enforces $\lambda_n = 0$.