Solutions To Problems of Chapter 2

2.1. Derive the mean and variance for the binomial distribution.

Solution: For the mean value we have that,

$$\begin{aligned}
\mathbb{E}[\mathrm{x}] &= \sum_{k=0}^{n} \frac{kn!}{(n-k)!k!} p^k (1-p)^{n-k} \\
&= \sum_{k=1}^{n} \frac{n!}{(n-k)!(k-1)!} p^k (1-p)^{n-k} \\
&= np \sum_{k=1}^{n} \frac{(n-1)!}{\big((n-1)-(k-1)\big)!(k-1)!} p^{k-1} (1-p)^{(n-1)-(k-1)} \\
&= np \sum_{l=0}^{n-1} \frac{(n-1)!}{\big((n-1)-l\big)!l!} p^{l} (1-p)^{(n-1)-l} \\
&= np(p+1-p)^{n-1} = np. \qquad (1)
\end{aligned}$$

where the formula for the binomial expansion has been employed. For the variance we have,

$$\begin{aligned}
\sigma_x^2 &= \sum_{k=0}^{n} (k-np)^2 \frac{n!}{(n-k)!k!} p^k (1-p)^{n-k} \\
&= \sum_{k=0}^{n} k^2 \frac{n!}{(n-k)!k!} p^k (1-p)^{n-k} + \\
&\quad \sum_{k=0}^{n} (np)^2 \frac{n!}{(n-k)!k!} p^k (1-p)^{n-k} - \\
&\quad 2np \sum_{k=0}^{n} k \frac{n!}{(n-k)!k!} p^k (1-p)^{n-k}, \qquad (2)
\end{aligned}$$

or

$$\sigma_x^2 = \sum_{k=0}^{n} k^2 \frac{n!}{(n-k)!k!} p^k (1-p)^{n-k} + (np)^2 - 2(np)^2, \qquad (3)$$

However,

$$\begin{aligned}
&\sum_{k=0}^{n} k^2 \frac{n!}{(n-k)!k!} p^k (1-p)^{n-k} = \\
&np \sum_{k=1}^{n} k \frac{(n-1)!}{\big((n-1)-(k-1)\big)!(k-1)!} p^{k-1} (1-p)^{(n-1)-(k-1)} = \\
&np \sum_{l=0}^{n-1} (l+1) \frac{(n-1)!}{\big((n-1)-l\big)!l!} p^l (1-p)^{(n-1)-l} = \\
&np + np(n-1)p, \qquad (4)
\end{aligned}$$

which finally proves the result.

2.2. Derive the mean and the variance for the uniform distribution.

Solution: For the mean we have

$$\begin{aligned}
\mu = \mathbb{E}[\mathrm{x}] &= \int_a^b \frac{1}{b-a} x dx \\
&= \frac{1}{b-a} \frac{b}{2}\Big|_a^b = \frac{b+a}{2}. \qquad (5)
\end{aligned}$$

For the variance, we have

$$\begin{aligned}
\sigma_x^2 &= \frac{1}{b-a} \int_a^b (x-\mu)^2 dx = \frac{1}{b-a} \int_{a-\mu}^{b-\mu} y^2 dy \\
&= \frac{1}{b-a} \frac{y^3}{3}\Big|_{a-\mu}^{b-\mu} \\
&= \frac{1}{12}(b-a)^2. \qquad (6)
\end{aligned}$$

2.3. Derive the mean and covariance matrix of the multivariate Gaussian.

Solution: Without harming generality, we assume that $\boldsymbol{\mu} = \mathbf{0}$, in order to simplify the discussion. We have that

$$\frac{1}{(2\pi)^{l/2}|\Sigma|^{1/2}} \int_{-\infty}^{+\infty} \boldsymbol{x} \exp\Big(-\frac{1}{2}\boldsymbol{x}^T \Sigma^{-1} \boldsymbol{x}\Big) d\boldsymbol{x}, \qquad (7)$$

which due to the symmetry of the exponential results in $\mathbb{E}[\mathbf{x}] = \mathbf{0}$.

For the covariance we have that

$$\int_{-\infty}^{+\infty} \exp\Big(-\frac{1}{2}\boldsymbol{x}^T \Sigma^{-1} \boldsymbol{x}\Big) d\boldsymbol{x} = (2\pi)^{l/2}|\Sigma|^{1/2}. \qquad (8)$$

Following similar arguments as for the univariate case given in the text, we are going to take the derivative on both sides with respect to matrix Σ. Recall from linear algebra the following formulas.

$$\frac{\partial \text{trace}\{AX^{-1}B\}}{\partial X} = -(X^{-1}BAX^{-1})^T, \quad \frac{\partial |X^k|}{\partial X} = k|X^k|X^{-T}.$$

Hence, taking the derivatives of both sides in (8) with respect to Σ we obtain,

$$\frac{1}{2}\int_{-\infty}^{+\infty} \exp\Big(-\frac{1}{2}\boldsymbol{x}^T\Sigma^{-1}\boldsymbol{x}\Big)\big(\Sigma^{-1}\boldsymbol{x}\boldsymbol{x}^T\Sigma^{-1}\big)^T d\boldsymbol{x} = \frac{1}{2}(2\pi)^{l/2}|\Sigma|^{1/2}\Sigma^{-T}, \tag{9}$$

which then readily gives the result.

2.4. Show that the mean and variance of the beta distribution with parameters a and b are given by

$$\mathbb{E}[\mathrm{x}] = \frac{a}{a+b},$$

and

$$\sigma_x^2 = \frac{ab}{(a+b)^2(a+b+1)}.$$

Hint: Use the property $\Gamma(a+1) = a\Gamma(a)$.

Solution: We know that

$$\text{Beta}(x|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}.$$

Hence

$$\mathbb{E}[\mathrm{x}] = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_0^1 x x^{a-1}(1-x)^{b-1}dx = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+1+b)},$$

which, using the property $\Gamma(a+1) = a\Gamma(a)$, results in

$$\mathbb{E}[\mathrm{x}] = \frac{a}{a+b}. \tag{10}$$

For the variance we have

$$\mathbb{E}[(\mathrm{x}-\mathbb{E}[\mathrm{x}])^2] = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_0^1 \left(x - \frac{a}{a+b}\right)^2 x^{a-1}(1-x)^{b-1}dx, \tag{11}$$

or

$$\begin{aligned}\sigma_x^2 = &\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_0^1 x^{a+1}(1-x)^{b-1}dx \\ &+ \frac{a^2}{(a+b)^2}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_0^1 x^{a-1}(1-x)^{b-1}dx \\ &- 2\frac{a}{a+b}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_0^1 x^a(1-x)^{b-1}dx,\end{aligned} \tag{12}$$

and following a similar path as the one adopted for the mean, it is a matter of simple algebra to show that

$$\sigma_x^2 = \frac{ab}{(a+b)^2(a+b+1)}.$$

2.5. Show that the normalizing constant in the beta distribution with parameters a, b is given by

$$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}.$$

Solution: The beta distribution is given by

$$\text{Beta}(x|a,b) = Cx^{a-1}(1-x)^{b-1},\ 0 \leq x \leq 1. \tag{13}$$

Hence

$$C^{-1} = \int_0^1 x^{a-1}(1-x)^{b-1}dx. \tag{14}$$

Let

$$x = \sin^2\theta \Rightarrow dx = 2\sin\theta\cos\theta d\theta. \tag{15}$$

Hence

$$C^{-1} = 2\int_0^{\frac{\pi}{2}} (\sin\theta)^{2a-1}(\cos\theta)^{2b-1}d\theta. \tag{16}$$

Recall the definition of the gamma function

$$\Gamma(a) = \int_0^\infty x^{a-1}e^{-x}dx,$$

and set

$$x = y^2 \Rightarrow dx = 2ydy,$$

hence

$$\Gamma(a) = 2\int_0^\infty y^{2a-1}e^{-y^2}dy. \tag{17}$$

Thus

$$\Gamma(a)\Gamma(b) = 4\int_0^\infty \int_0^\infty x^{2a-1}y^{2b-1}e^{-(x^2+y^2)}dxdy. \tag{18}$$

Let

$$x = r\sin\theta, y = r\cos\theta \Rightarrow dxdy = rdrd\theta.$$

Hence

$$\Gamma(a)\Gamma(b) = 4\int_0^{\frac{\pi}{2}} \int_0^\infty r^{2(a+b)-1}e^{-r^2}(\sin\theta)^{2a-1}(\cos\theta)^{2a-1}drd\theta. \tag{19}$$

where integration over θ is in the interval $\left[0, \frac{\pi}{2}\right]$ to guarantee that x remains non-negative. From (19) we have

$$\Gamma(a)\Gamma(b) = \left(2\int_0^{\infty} r^{2(a+b)-1}e^{-r^2}dr\right)\left(2\int_0^{\frac{\pi}{2}} (\sin\theta)^{2a-1}(\cos\theta)^{2b-1}d\theta\right)$$
$$= \Gamma(a+b)C^{-1},$$

which proves the claim.

2.6. Show that the mean and variance of the gamma pdf

$$\text{Gamma}(x|a,b) = \frac{b^a}{\Gamma(a)}x^{a-1}e^{-bx}, \ a, b, x > 0.$$

are given by

$$\mathbb{E}[\mathrm{x}] = \frac{a}{b},$$
$$\sigma_x^2 = \frac{a}{b^2}.$$

Solution: We have that

$$\mathbb{E}[\mathrm{x}] = \frac{b^a}{\Gamma(a)}\int_0^{\infty} x^a e^{-bx}dx.$$

Set $bx = y$. Then

$$\mathbb{E}[\mathrm{x}] = \frac{b^a}{\Gamma(a)}\frac{1}{b^{a+1}}\int_0^{\infty} y^a e^{-y}dy$$
$$= \frac{1}{b\Gamma(a)}\Gamma(a+1) = \frac{a\Gamma(a)}{b\Gamma(a)} = \frac{a}{b}.$$

For the variance, the following is valid

$$\sigma_x^2 = \mathbb{E}[(\mathrm{x} - \frac{a}{b})^2] = \frac{b^a}{\Gamma(a)}\Big\{\int_0^{\infty} x^{a+1}e^{-bx}dx$$
$$+ \frac{a^2}{b^2}\int_0^{\infty} x^{a-1}e^{-bx}dx - 2\frac{a}{b}\int_0^{\infty} x^a e^{-bx}dx\Big\},$$

and following a similar path as before we obtain

$$\sigma_x^2 = \frac{a}{b^2}.$$

2.7. Show that the mean and variance of a Dirichlet pdf with K variables, $\mathrm{x}_k, \ k = 1, 2, \ldots, K$ and parameters $a_k, \ k = 1, 2, \ldots, K$, are given by

$$\mathbb{E}[\mathrm{x}_k] = \frac{a_k}{\overline{a}}, \ k = 1, 2, \ldots, K$$
$$\sigma_k^2 = \frac{a_k(\overline{a} - a_k)}{\overline{a}^2(1+\overline{a})}, \ k = 1, 2, \ldots, K,$$
$$\text{cov}[\mathrm{x}_i\mathrm{x}_j] = -\frac{a_i a_j}{\overline{a}^2(1+\overline{a})}, \ i \neq j,$$

where $\overline{a} = \sum_{k=1}^{K} a_k$.

Solution: Without harm of generality, we will derive the mean for x_K. The others are derived similarly. To this end, we have

$$p(x_1, x_2, \ldots, x_{K-1}) = C \prod_{k=1}^{K-1} x_k^{a_k - 1} \left(1 - \sum_{k=1}^{K-1} x_k\right)^{a_K - 1}$$

where

$$C = \frac{\Gamma(a_1 + a_2 + \ldots + a_K)}{\Gamma(a_1)\Gamma(a_2)\ldots\Gamma(a_K)}.$$

$$\begin{aligned}
\mathbb{E}[\mathrm{x}_K] &= C \int_0^1 \cdots \int_0^1 \left[\int_0^{1-\sum_{k=1}^{K-1} x_k} x_K p(x_1, \ldots, x_{K-1}, x_K) dx_K\right] dx_{K-1} \ldots dx_1 \\
&= C \int_0^1 \cdots \int_0^1 \left[\int_0^{1-\sum_{k=1}^{K-1} x_k} x_K \prod_{k=1}^{K-1} x_k^{a_k - 1} \left(1 - \sum_{k=1}^{K-1} x_k\right)^{a_K - 1} dx_K\right] dx_{K-1} \ldots dx_1 \\
&= C \int_0^1 \cdots \int_0^1 \prod_{k=1}^{K-1} x_k^{a_k - 1} \left(1 - \sum_{k=1}^{K-1} x_k\right)^{a_K} \left[\int_0^{1-\sum_{k=1}^{K-1} x_k} dx_K\right] dx_{K-1} \ldots dx_1,
\end{aligned}$$

or

$$\begin{aligned}
\mathbb{E}[\mathrm{x}_K] &= C \int_0^1 \cdots \int_0^1 \prod_{k=1}^{K-1} x_k^{a_k - 1} \left(1 - \sum_{k=1}^{K-1} x_k\right)^{a_K} dx_{K-1} \ldots dx_1 \\
&= C \frac{\Gamma(a_1) \ldots \Gamma(a_K + 1)}{\Gamma(a_1 + a_2 + \ldots + a_K + 1)} \\
&= C \frac{a_K \Gamma(a_1) \ldots \Gamma(a_K)}{(a_1 + a_2 + \ldots + a_K)\Gamma(a_1 + a_2 + \ldots + a_K)} \\
&= \frac{a_K}{\overline{a}}.
\end{aligned}$$

In the sequel, we will show that

$$\mathbb{E}[\mathrm{x}_i \mathrm{x}_j] = -\frac{a_i a_j}{\overline{a}^2(\overline{a} + 1)}, \ i \neq j.$$

We derive it for the variables x_K and x_{K-1}, since any of the variables can be taken in place of x_K and x_{K-1}. Hence,

$$\begin{aligned}
\mathbb{E}[\mathrm{x}_{K-1}\mathrm{x}_K] &= C\int_0^1 \cdots \int_0^1 \left[\int_0^{1-\sum_{k=1}^{K-1} x_k} \left(\prod_{k=1}^{K-2} x_k^{a_k-1}\right) x_{K-1}^{a_{K-1}} x_K^{a_K}\, dx_K\right] dx_{K-1}\dots dx_1 \\
&= C\int_0^1 \cdots \int_0^1 \prod_{k=1}^{K-2} x_k^{a_k-1} x_{K-1}^{a_{K-1}} \left[\int_0^{1-\sum_{k=1}^{K-1} x_k} x_K^{a_K}\, dx_K\right] dx_{K-1}\dots dx_1 \\
&= \frac{C}{a_K+1}\int_0^1 \cdots \int_0^1 \prod_{k=1}^{K-2} x_k^{a_k-1} x_{K-1}^{a_{K-1}} \left(1-\sum_{k=1}^{K-1} x_k\right)^{a_K+1} dx_{K-1}\dots dx_1 \\
&= \frac{C}{a_K+1}\,\frac{\Gamma(a_1)\dots\Gamma(a_{K-2})\Gamma(a_{K-1}+1)\Gamma(a_K+2)}{\Gamma(a_1+\dots+a_{K-2}+a_{K-1}+a_K+2)} \\
&= \frac{C}{a_K+1}\,\frac{a_K a_{K-1}\Gamma(a_1)\dots\Gamma(a_K)(a_K+1)}{(1+a_1+\dots+a_K)(a_1+\dots+a_K)\Gamma(a_1+\dots+a_K)}
\end{aligned}$$

or

$$\mathbb{E}[\mathrm{x}_{K-1}\mathrm{x}_K] = \frac{a_K a_{K-1}}{\overline{a}(1+\overline{a})}.$$

Thus in general,

$$\mathbb{E}[\mathrm{x}_i\mathrm{x}_j] = \frac{a_i a_j}{\overline{a}(1+\overline{a})}.$$

For the covariance, we have

$$\begin{aligned}
\mathrm{cov}[\mathrm{x}_i\mathrm{x}_j] &= \mathbb{E}\left[\mathrm{x}_i - \mathbb{E}[\mathrm{x}_i]\right]\mathbb{E}\left[\mathrm{x}_j - \mathbb{E}[\mathrm{x}_j]\right] \\
&= \mathbb{E}[\mathrm{x}_i\mathrm{x}_j] - \mathbb{E}[\mathrm{x}_i]\,\mathbb{E}[\mathrm{x}_j],
\end{aligned}$$

or

$$\begin{aligned}
\mathrm{cov}[\mathrm{x}_i\mathrm{x}_j] &= \frac{a_i a_j}{\overline{a}(1+\overline{a})} - \frac{a_i a_j}{\overline{a}^2} \\
&= \frac{a_i a_j\overline{a} - a_i a_j(1+\overline{a})}{\overline{a}^2(1+\overline{a})} = -\frac{a_i a_j}{\overline{a}^2(1+\overline{a})}.
\end{aligned}$$

2.8. Show that the sample mean, using N i.i.d drawn samples, is an unbiased estimator with variance that tends to zero asymptotically, as $N \longrightarrow \infty$.

Solution: From the definition of the sample mean we have

$$\mathbb{E}[\hat{\mu}_N] = \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[\mathrm{x}_n] = \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[\mathrm{x}] = \mathbb{E}[\mathrm{x}]. \tag{20}$$

For the variance we have,

$$\begin{aligned}\sigma^2_{\hat{\mu}_N} &= \mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^{N}\mathrm{x}_i - \mu\right)\left(\frac{1}{N}\sum_{j=1}^{N}\mathrm{x}_j - \mu\right)\right] \\ &= \mathbb{E}\left[\frac{1}{N^2}\left(\sum_{i=1}^{N}(\mathrm{x}_i - \mu)\sum_{j=1}^{N}(\mathrm{x}_j - \mu)\right)\right] \qquad (21) \\ &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}\left[(\mathrm{x}_i - \mu)(\mathrm{x}_j - \mu)\right]. \qquad (22)\end{aligned}$$

However, since the samples are i.i.d. drawn, the expected value of the product is equal to the product of the mean values, hence it is zero except for $i = j$, which then results in

$$\sigma^2_{\hat{\mu}_N} = \frac{1}{N}\sigma^2_x,$$

which proves the claim.

2.9. Show that for WSS processes

$$r(0) \geq |r(k)|, \forall k \in \mathbb{Z},$$

and that for jointly WSS processes,

$$r_u(0)r_v(0) \geq |r_{uv}(k)|^2, \ \forall k \in \mathbb{Z}.$$

Solution: Both properties are shown in a similar way. So, we are going to focus on the first one. Consider the obvious inequality,

$$\mathbb{E}[|\mathrm{u}_n - \lambda \mathrm{u}_{n-k}|^2] \geq 0,$$

or

$$\mathbb{E}[|\mathrm{u}_n|^2] + |\lambda|^2\,\mathbb{E}[|\mathrm{u}_{n-k}|^2] \geq \lambda^* r(k) + \lambda r^*(k),$$

or

$$r(0) + |\lambda|^2 r(0) \geq \lambda^* r(k) + \lambda r^*(k).$$

This is true for any λ, thus it will be true for $\lambda = \frac{r(k)}{r(0)}$. Substituting, we obtain

$$r(0) \geq \frac{|r(k)|^2}{r(0)},$$

which proves the claim.

Similar steps are adopted in order to prove the property for the cross-correlation.

2.10. Show that the autocorrelation of the output of a linear system, with impulse response, $w_n,\ n \in \mathbb{Z}$, is related to the autocorrelation of the input process, via,

$$r_d(k) = r_u(k) * w_k * w_{-k}^*.$$

Solution: We have that

$$\begin{aligned} r_d(k) &= \mathbb{E}[\mathrm{d}_n \mathrm{d}_{n-k}^*] = \mathbb{E}\left[\sum_i w_i^* \mathrm{u}_{n-i} \sum_j w_j \mathrm{u}_{n-k-j}^*\right] \\ &= \sum_i \sum_j w_i^* w_j \, \mathbb{E}[\mathrm{u}_{n-i}\mathrm{u}_{n-k-j}^*] \\ &= \sum_j w_j \sum_i w_i^* r_u(k+j-i). \end{aligned} \tag{23}$$

Set

$$h(n) := w_n * r_u(n). \tag{24}$$

Then we can write,

$$\begin{aligned} r_d(k) &= \sum_j w_j h(k+j) = \sum_j w_j h\big(-((-k)-j)\big) = w_{-k}^* * h(-(-k)) \\ &= w_{-k}^* * w_k * r_u(k), \end{aligned}$$

which proves the claim.

2.11. Show that

$$\ln x \leq x - 1.$$

Solution: Define the function

$$f(x) = x - 1 - \ln x.$$

then

$$f'(x) = 1 - \frac{1}{x}, \quad \text{and} \quad f''(x) = \frac{1}{x^2}.$$

Thus $x = 1$ is a minimum, i.e.,

$$f(x) \geq f(1) = 1 - 1 - 0 = 0.$$

or

$$\ln x \leq x - 1.$$

2.12. Show that

$$I(\mathrm{x};\mathrm{y}) \geq 0.$$

Hint: Use the inequality of Problem 2.11.

Solution: By the respective definition, we have that

$$\begin{aligned} -I(\mathrm{x};\mathrm{y}) &= -\sum_x \sum_y P(x,y) \log \frac{P(x|y)}{P(x)} \\ &= \log e \sum_x \sum_y P(x,y) \ln \frac{P(x)}{P(x|y)}, \end{aligned}$$

where we have used only terms where $P(x,y) \neq 0$. Taking into account the inequality, we have that

$$\begin{aligned} -I(\mathrm{x};\mathrm{y}) &\leq \log e \sum_x \sum_y P(x,y) \left\{ \frac{P(x)}{P(x|y)} - 1 \right\} = \\ & \log e \sum_x \sum_y \left\{ P(x)P(y) - P(x,y) \right\}. \end{aligned}$$

Note that the summation over the terms in the brackets is equal to zero, which proves the claim.

Note that if the random variables are independent, then $P(x) = P(x|y)$ and $I(\mathrm{x};\mathrm{y}) = 0$.

2.13. Show that if $a_i,\ b_i,\ i = 1, 2, \ldots, M$ are positive numbers, such as

$$\sum_{i=1}^{M} a_i = 1, \text{ and } \sum_{i=1}^{M} b_i \leq 1,$$

then

$$-\sum_{i=1}^{M} a_i \ln a_i \leq -\sum_{i=1}^{M} a_i \ln b_i.$$

Solution: Recalling the inequality from Problem 2.11, that

$$\ln \frac{b_i}{a_i} \leq \frac{b_i}{a_i} - 1,$$

or

$$\sum_{i=1}^{M} a_i \ln \frac{b_i}{a_i} \leq \sum_{i=1}^{M} (b_i - a_i) \leq 0,$$

which proves the claim and where the assumptions concerning a_i and b_i have been taken into account.

2.14. Show that the maximum value of the entropy of a random variable occurs if all possible outcomes are equiprobable.

Solution: Let p_i, $i = 1, 2, \ldots, M$, be the corresponding probabilities of the M possible events. According to the inequality in Problem 2.13 for $b_i = 1/M$, we have,

$$-\sum_{i=1}^{M} p_i \ln p_i \leq \sum_{i=1}^{M} p_i \ln M,$$

or

$$-\sum_{i=1}^{M} p_i \ln p_i \leq \ln M.$$

Thus the maximum value of the entropy is $\ln M$, which is achieved if all probabilities are equal to $1/M$.

2.15. Show that from all the pdfs which describe a random variable in an interval $[a, b]$ the uniform one maximizes the entropy.

Solution: The Lagrangian of the constrained optimization task is

$$L(p(\cdot), \lambda) = -\int_{-\infty}^{+\infty} p(x) \ln p(x) dx + \lambda \Big(\int_{-\infty}^{+\infty} p(x) dx - 1 \Big).$$

According to the calculus of variations (for the unfamiliar reader, treat $p(x)$ as a variable and take derivatives under the integrals as usual) we take the derivative and set it equal to zero, resulting in

$$\ln p(x) = \lambda - 1.$$

Plugging it in the constrain equation, and performing the integration results in

$$p(x) = \frac{1}{b-a},$$

which proves the claim.

Solutions To Problems of Chapter 3

3.1. Prove the least squares optimal solution for the linear regression case given in Eq. (3.13).

Solution: The cost function is

$$\begin{aligned} J(\boldsymbol{\theta}) &= \sum_{n=1}^{N} (y_n - \boldsymbol{\theta}^T \boldsymbol{x}_n)^2 \\ &= \sum_{n=1}^{N} (y_n - \boldsymbol{\theta}^T \boldsymbol{x}_n)(y_n - \boldsymbol{x}_n^T \boldsymbol{\theta}), \end{aligned}$$

or

$$J(\boldsymbol{\theta}) = \sum_{n=1}^{N} y_n^2 - 2\boldsymbol{\theta}^T \left(\sum_{n=1}^{N} \boldsymbol{x}_n y_n \right) + \boldsymbol{\theta}^T \left(\sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^T \right) \boldsymbol{\theta}.$$

Taking the gradient of the cost with respect to $\boldsymbol{\theta}$ and equating to zero, we finally get,

$$\left(\sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^T \right) \boldsymbol{\theta} = \sum_{n=1}^{N} \boldsymbol{x}_n y_n$$

3.2. Let $\hat{\theta}_i$, $i = 1, 2, \ldots, m$, be unbiased estimators of a parameter vector $\boldsymbol{\theta}$, i.e., $\mathbb{E}[\hat{\theta}_i] = \boldsymbol{\theta}$, $i = 1, \ldots, m$. Moreover, assume that the respective estimators are uncorrelated to each other and that all have the same (total) variance, $\sigma^2 = \mathbb{E}[(\theta_i - \boldsymbol{\theta})^T (\theta_i - \boldsymbol{\theta})]$. Show that by averaging the estimates, e.g.,

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

the new estimator has total variance $\sigma_c^2 := \mathbb{E}[(\hat{\theta} - \boldsymbol{\theta})^T (\hat{\theta} - \boldsymbol{\theta})] = \frac{1}{m}\sigma^2$.

Solution: First, it is easily checked out that the new estimator is also unbiased. By the definition of the total variance (which is the trace of the respective covariance matrix), we have

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

since $\mathbb{E}[(\hat{\theta}_i - \boldsymbol{\theta})^T (\hat{\theta}_j - \boldsymbol{\theta})] = \delta_{ij}\sigma^2$.

3.3. Let a random variable x being described by a uniform pdf in the interval $[0, \frac{1}{\theta}]$, $\theta > 0$. Assume a function[1] g, which defines an estimator $\hat{\theta} := g(\mathrm{x})$ of θ. Then, for such an estimator to be unbiased, the following must hold:

$$\int_0^{\frac{1}{\theta}} g(x)dx = 1.$$

However, such a function g does not exist.

Solution: Necessarily, the pdf of x must be

$$p(x) = \begin{cases} \theta, & x \in [0, \frac{1}{\theta}], \\ 0, & \text{otherwise.} \end{cases}$$

For the estimator to be unbiased,

$$\begin{aligned} \mathbb{E}[\hat{\theta}] &= \int_{-\infty}^{\infty} g(x)p(x)dx \\ &= \int_0^{\frac{1}{\theta}} g(x)\theta dx = \theta, \quad \forall \theta > 0. \end{aligned}$$

Hence,

$$G(\theta) := \int_0^{\frac{1}{\theta}} g(x)dx = 1, \quad \forall \theta > 0. \tag{1}$$

However, such a function g cannot exist. Indeed, one can easily verify by the basic integral theory that $\lim_{\theta \to \infty} G(\theta) = 0$, and $\lim_{\theta \to 0} G(\theta) = 1$, by (1). Clearly, these results contradict each other.

3.4. A family $\{p(\mathcal{D}; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathcal{A}\}$ is called *complete* if, for any vector function $\boldsymbol{h}(\mathcal{D})$ such that $\mathbb{E}_{\mathcal{D}}[\boldsymbol{h}(\mathcal{D})] = \mathbf{0}$, $\forall \boldsymbol{\theta}$, then $\boldsymbol{h} = \mathbf{0}$.

Show that if $\{p(\mathcal{D}; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathcal{A}\}$ is complete, and there exists an MVU estimator, then this estimator is unique.

Solution: Assume two MVU estimators $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$. Then, $\mathbb{E}_{\mathcal{D}}[\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2] = \boldsymbol{\theta} - \boldsymbol{\theta} = \mathbf{0}$. Hence, by the definition of completeness, $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$.

3.5. Let $\hat{\theta}_u$ be an unbiased estimator, i.e., $\mathbb{E}[\hat{\theta}_u] = \theta_0$. Define a biased one by $\hat{\theta}_b = (1 + \alpha)\hat{\theta}_u$. Show that the range of α where the MSE of $\hat{\theta}_b$ is smaller than that of $\hat{\theta}_u$ is

$$-2 < -\frac{2\mathrm{MSE}(\hat{\theta}_u)}{\mathrm{MSE}(\hat{\theta}_u) + \theta_0^2} < \alpha < 0.$$

[1] To avoid any confusion, let g be Lebesgue integrable on intervals of $\mathbb{R}$.

Solution: The MSE for the new estimator is

$$\begin{aligned}\mathbb{E}\left[(\hat{\theta}_b-\theta_0)^2\right] &= \mathbb{E}\left[\left((1+\alpha)\hat{\theta}_u-\theta_0\right)^2\right]\\ &= \mathbb{E}\left[\left((1+\alpha)(\hat{\theta}_u-\theta_0)+\alpha\theta_0\right)^2\right]\\ &= (1+\alpha)^2\mathrm{MSE}(\hat{\theta}_u)+\alpha^2\theta_0^2.\end{aligned}$$

To obtain smaller MSE for the unbiased estimator we must have

$$(1+\alpha)^2\mathrm{MSE}(\hat{\theta}_u)+\alpha^2\theta_0^2<\mathrm{MSE}(\hat{\theta}_u),$$

or, after using elementary algebra,

$$\alpha\left[\alpha+\frac{2\,\mathrm{var}(\hat{\theta}_u)}{\theta_0^2+\mathrm{var}(\hat{\theta}_u)}\right]<0,$$

where var(·) denotes the variance, and clearly $\mathrm{MSE}(\hat{\theta}_u)=\mathrm{var}(\hat{\theta}_u)$. The solution of the previous inequality results to the desired interval:

$$-\frac{2\,\mathrm{var}(\hat{\theta}_u)}{\theta_0^2+\mathrm{var}(\hat{\theta}_u)}<\alpha<0.$$

3.6. Show that for the setting of the Problem 3.5, the optimal value of α is equal to

$$\alpha_*=-\frac{1}{1+\frac{\theta_0^2}{\mathrm{var}(\hat{\theta}_u)}},$$

where, of course, the variance of the unbiased estimator is equal to the corresponding MSE.

Solution: The minimum value of

$$\mathrm{MSE}(\hat{\theta}_b)=\mathbb{E}\left[(\hat{\theta}_b-\theta_0)^2\right]=(1+\alpha)^2\mathrm{MSE}(\hat{\theta}_u)+\alpha^2\theta_0^2,$$

with respect to α occurs when the derivative becomes zero, that is when

$$2(1+\alpha)\,\mathrm{var}(\hat{\theta}_u)+2\alpha\theta_0^2=0,$$

or, equivalently, when

$$\alpha_*=-\frac{\mathrm{var}(\hat{\theta}_u)}{\theta_0^2+\mathrm{var}(\hat{\theta}_u)}=-\frac{1}{1+\frac{\theta_0^2}{\mathrm{var}(\hat{\theta}_u)}}.$$

3.7. Show that the regularity condition for the Cramér-Rao bound holds true if the order of integration and differentiation can be interchanged.

Solution: By the definition of the expectation we have

$$\begin{aligned}\mathbb{E}\left[\frac{\partial \ln p(\mathcal{X};\theta)}{\partial\theta}\right] &= \int\cdots\int p(\mathcal{X};\theta)\frac{\partial \ln p(\mathcal{X};\theta)}{\partial\theta}\mathrm{d}\boldsymbol{x}_1\cdots\mathrm{d}\boldsymbol{x}_N\\ &= \int\cdots\int \frac{\partial p(\mathcal{X};\theta)}{\partial\theta}\mathrm{d}\boldsymbol{x}_1\cdots\mathrm{d}\boldsymbol{x}_N\\ &= \frac{\partial}{\partial\theta}\int\cdots\int p(\mathcal{X};\theta)\mathrm{d}\boldsymbol{x}_1\cdots\mathrm{d}\boldsymbol{x}_N\\ &= \frac{\partial 1}{\partial\theta}=0.\end{aligned}$$

This is in general true, unless the domain where the pdf is nonzero depends on the unknown parameter θ.

3.8. Derive the Cramér-Rao bound for the LS estimator, when the training data result from the linear model

$$y_n=\theta x_n+\eta_n,\quad n=1,2,\ldots,$$

where x_n and η_n are observations of i.i.d random variables, drawn from a zero mean random process, with variance σ_x^2, and a Gaussian white noise process, with zero mean and variance σ_η^2, respectively. Assume, also, that x and η are independent. Then, show that the LS estimator achieves the CR bound only asymptotically.

Solution: First, notice that in this case $\mathcal{X}=\{(x_n,y_n)\}_{n=1}^N$. That is, both y_n as well as x_n change as we change the training set. Define here the quantities $\boldsymbol{x}_N := [x_1,x_2,\ldots,x_N]^T$, $\boldsymbol{y}_N := [y_1,y_2,\ldots,y_N]^T$, and recall, also, the elementary relations

$$\begin{aligned}p(\mathcal{X};\theta)&=p(\boldsymbol{x}_N,\boldsymbol{y}_N;\theta)\\ &=p(\boldsymbol{y}_N|\boldsymbol{x}_N;\theta)p(\boldsymbol{x}_N;\theta)\\ &=p(\boldsymbol{y}_N|\boldsymbol{x}_N;\theta)p(\boldsymbol{x}_N).\end{aligned}$$

Hence, $\ln p(\mathcal{X};\theta)=\ln p(\boldsymbol{y}_N|\boldsymbol{x}_N;\theta)+\ln p(\boldsymbol{x}_N)$, and eventually,

$$\frac{\partial \ln p(\mathcal{X};\theta)}{\partial\theta}=\frac{\partial \ln p(\boldsymbol{y}_N|\boldsymbol{x}_N;\theta)}{\partial\theta}. \tag{2}$$

Now, notice that by our original assumptions on the data model,

$$p(\boldsymbol{y}_N|\boldsymbol{x}_N;\theta)=\frac{1}{\left(2\pi\sigma_\eta^2\right)^{N/2}}\exp\left(-\frac{1}{2\sigma_\eta^2}\sum_{n=1}^N(y_n-\theta x_n)^2\right),$$

or

$$\ln p(\boldsymbol{y}_N|\boldsymbol{x}_N;\theta)=-\frac{N}{2}\ln\left(2\pi\sigma_\eta^2\right)-\frac{1}{2\sigma_\eta^2}\sum_{n=1}^N(y_n-\theta x_n)^2.$$

Thus, by (2),

$$\frac{\partial \ln p(\mathcal{X};\theta)}{\partial \theta} = \frac{1}{\sigma_\eta^2}\sum_{n=1}^{N}(y_n - \theta x_n)\, x_n = \frac{1}{\sigma_\eta^2}\sum_{n=1}^{N}\eta_n x_n. \tag{3}$$

It can be readily verified by (3) that the regularity condition of the Cramér-Rao Theorem is satisfied.

Now,

$$\frac{\partial^2 \ln p(\mathcal{X};\theta)}{\partial \theta^2} = -\frac{1}{\sigma_\eta^2}\sum_{n=1}^{N} x_n^2.$$

Therefore

$$\mathbb{E}\left[\frac{\partial^2 \ln p(\mathcal{X};\theta)}{\partial \theta^2}\right] = -N\frac{\sigma_x^2}{\sigma_\eta^2},$$

and the Cramér-Rao bound is given by

$$\mathrm{var}(\hat{\theta}) \geq \frac{1}{N}\frac{\sigma_\eta^2}{\sigma_x^2}.$$

We will now show that this bound cannot be achieved by any unbiased estimator. The necessary and sufficient condition for the existence of an MVU estimator that achieves the Cramér-Rao bound translates, for this case, to the existence of a function $g(\mathcal{X})$ such that

$$\frac{\partial \ln p(\mathcal{X};\theta)}{\partial \theta} = N\frac{\sigma_x^2}{\sigma_\eta^2}(g(\mathcal{X}) - \theta),$$

However, looking at (3), it becomes apparent that such a factorization is not possible.

Let us now rewrite (3) as

$$\begin{aligned}\frac{\partial \ln p(\mathcal{X};\theta)}{\partial \theta} &= N\frac{\sigma_x^2}{\sigma_\eta^2}\Big(\frac{1}{N\sigma_x^2}\sum_{n=1}^{N}\left(y_n x_n - \theta x_n^2\right)\Big)\\ & N\frac{\sigma_x^2}{\sigma_\eta^2}\Big(\frac{1}{N\sigma_x^2}\sum_{n=1}^{N} y_n x_n - \theta(\frac{1}{N\sigma_x^2}\sum_{n=1}^{N} x_n^2)\Big)\\ & N\frac{\sigma_x^2}{\sigma_\eta^2}\Big(g(\mathcal{X}) - \theta(\frac{1}{N\sigma_x^2}\sum_{n=1}^{N} x_n^2)\Big) \end{aligned} \tag{4}$$

where

$$g(\mathcal{X}) := \frac{1}{N\sigma_x^2}\sum_{n=1}^{N} y_n x_n. \tag{5}$$

For a large number N of the training data set, we assume the following approximation: $\sum_{n=1}^{N} x_n^2 \approx N\sigma_x^2$. By embedding this into (4) we obtain a form that allows for an unbiased estimator to attain the Cramér-Rao bound, and the corresponding estimate is given by

$$\hat{\theta} = g(\mathcal{X}) = \frac{1}{N\sigma_x^2}\sum_{n=1}^{N} y_n x_n. \tag{6}$$

However, (6) is the LS estimator for large values of N. Indeed, by the definition of the LS estimator we have that

$$\frac{1}{N}\left(\sum_{n=1}^{N} x_n^2\right)\hat{\theta} = \frac{1}{N}\sum_{n=1}^{N} x_n y_n, \tag{7}$$

which results in (6). It is easy to verify that (7) corresponds to an unbiased estimator.

Let us do it for the sake of an exercise. First of all, let us examine if the LS estimator for this more general case is unbiased. We have

$$\begin{aligned}
\mathbb{E}[\hat{\theta}] &= \mathbb{E}\left[\frac{1}{\sum_n \mathrm{x}_n^2}\sum_n \mathrm{x}_n \mathrm{y}_n\right] = \mathbb{E}\left[\frac{1}{\sum_n \mathrm{x}_n^2}\sum_n \mathrm{x}_n(\theta \mathrm{x}_n + \eta_n)\right] \\
&= \mathbb{E}\left[\frac{1}{\sum_n \mathrm{x}_n^2}\left(\theta\sum_n \mathrm{x}_n^2 + \sum_n \mathrm{x}_n\eta_n\right)\right] = \theta + \mathbb{E}\left[\frac{1}{\sum_n \mathrm{x}_n^2}\sum_n \mathrm{x}_n\eta_n\right] \\
&= \theta + \mathbb{E}_x\left[\frac{1}{\sum_n \mathrm{x}_n^2}\,\mathbb{E}_{\eta|x}\left[\sum_n \mathrm{x}_n\eta_n\right]\right] \\
&= \theta + 0 = \theta.
\end{aligned}$$

In other words, the LS estimator is unbiased even for this case, where both output as well as input samples change in the training set and this is true independent of the number of measurements. The corresponding variance is given by

$$\begin{aligned}
\mathbb{E}\left[(\hat{\theta}-\theta)^2\right] &= \mathbb{E}\left[\frac{1}{(\sum_n \mathrm{x}_n^2)^2}\left(\sum_n \mathrm{x}_n\eta_n\right)^2\right] \\
&= \mathbb{E}_x\left[\frac{1}{(\sum_n \mathrm{x}_n^2)^2}\,\mathbb{E}_{\eta|x}\left[\sum_n \mathrm{x}_n^2\eta_n^2 + \sum_i\sum_{j\neq i}\mathrm{x}_i\mathrm{x}_j\eta_i\eta_j\right]\right] \\
&= \sigma_\eta^2\,\mathbb{E}_x\left[\frac{\sum_n \mathrm{x}_n^2}{(\sum_n \mathrm{x}_n^2)^2}\right] = \sigma_\eta^2\,\mathbb{E}\left[\frac{1}{\sum_n \mathrm{x}_n^2}\right].
\end{aligned}$$

Asymptotically, this provides the bound that we have previously derived. However, for finite N, this is different.

3.9. Let us consider the regression model

$$y_n = \boldsymbol{\theta}^T\boldsymbol{x}_n + \eta_n, \quad n = 1, 2, \ldots, N,$$

where the noise vector $\boldsymbol{\eta} := [\eta_1, \dots, \eta_N]^T$ comprises samples from zero mean Gaussian random variable, with covariance matrix Σ_η. If $X := [\boldsymbol{x}_1, \dots, \boldsymbol{x}_N]^T$ stands for the input matrix, and $\boldsymbol{y} = [y_1, \dots, y_N]^T$, the vector of the observations, then show that the corresponding estimator,

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

is an efficient one.

Notice, here, that the previous estimator coincides with the Maximum Likelihood (ML) one. Moreover, bear in mind that in the case where the noise process is considered to be white, i.e., $\Sigma_\eta = \sigma^2 I_N$, then the ML estimate becomes equal to the LS one.

Solution: In the case where the parameter $\boldsymbol{\theta}$ becomes a k-dimensional vector, the Cramér-Rao bound takes a more general form than the one we have met previously, i.e., the case where the parameter θ is a scalar. For any unbiased estimator $g(\mathcal{X})$ of the unknown parameter vector $\boldsymbol{\theta}$, the Cramér-Rao bound becomes as follows:

$$\mathbb{E}\left[(g(\mathcal{X}) - \boldsymbol{\theta})(g(\mathcal{X}) - \boldsymbol{\theta})^T\right] \succeq I^{-1}(\boldsymbol{\theta}), \quad \forall \boldsymbol{\theta},$$

where $I(\boldsymbol{\theta})$ is the *Fisher information matrix* defined as

$$I(\boldsymbol{\theta}) := -\mathbb{E}\left[\frac{\partial^2 \ln p(\mathcal{X}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}\right],$$

and which is known to be a positive semidefinite matrix. Given any matrices A, B, of the same dimensions, the inequality $A \succeq B$ means that the matrix $A - B$ is positive semidefinite. A necessary and sufficient condition for g to be efficient is for the equation

$$\frac{\partial \ln p(\mathcal{X}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = I(\boldsymbol{\theta}) \left(g(\mathcal{X}) - \boldsymbol{\theta}\right). \tag{8}$$

For the present model, we have that $\mathcal{X} = \boldsymbol{y}$ and

$$p(\boldsymbol{y}; \boldsymbol{\theta}) = \frac{1}{(2\pi)^{\frac{N}{2}} (\det \Sigma_\eta)^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\boldsymbol{y} - X\boldsymbol{\theta})^T \Sigma_\eta^{-1} (\boldsymbol{y} - X\boldsymbol{\theta})\right).$$

Hence, $\ln p(\boldsymbol{y}; \boldsymbol{\theta}) = -\frac{1}{2} (\boldsymbol{y} - X\boldsymbol{\theta})^T \Sigma_\eta^{-1} (\boldsymbol{y} - X\boldsymbol{\theta}) + \text{constant}$, and

$$\begin{aligned} \frac{\partial \ln p(\boldsymbol{y}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= X^T \Sigma_\eta^{-1} \boldsymbol{y} - X^T \Sigma_\eta^{-1} X \boldsymbol{\theta} \\ &= X^T \Sigma_\eta^{-1} X \left((X^T \Sigma_\eta^{-1} X)^{-1} X^T \Sigma_\eta^{-1} \boldsymbol{y} - \boldsymbol{\theta}\right). \end{aligned} \tag{9}$$

The second derivative is equal to

$$\frac{\partial^2 \ln p(\boldsymbol{y}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} = -X^T \Sigma_\eta^{-1} X,$$

so that the Fisher information matrix becomes $I(\boldsymbol{\theta}) = X^T \Sigma_\eta^{-1} X$. By this, and by a simple inspection of (9), we can readily verify that (8) is satisfied, if $g(\boldsymbol{y}) = \left(X^T \Sigma_\eta^{-1} X\right)^{-1} X^T \Sigma_\eta^{-1} \boldsymbol{y}$. This establishes the claim.

3.10. Assume a set of i.i.d $\mathcal{X} := \{x_1, x_2, \ldots, x_N\}$ Gaussian random variables, with mean μ and variance σ^2. Define also the quantities

$$S_\mu := \frac{1}{N} \sum_{n=1}^{N} x_n, \quad S_{\sigma^2} := \frac{1}{N} \sum_{n=1}^{N} (x_n - S_\mu)^2,$$
$$\bar{S}_{\sigma^2} := \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu)^2.$$

Show that if μ is considered to be known, a sufficient statistic for σ^2 is $\bar{S}_{\sigma^2}$. Moreover, in the case where both (μ, σ^2) are unknown, then a sufficient statistic is the pair (S_μ, S_{σ^2}).

Solution: The joint pdf of $\mathcal{X}$ is obviously

$$p(\mathcal{X}) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left(-\frac{1}{2\sigma^2} \sum_n (x_n - \mu)^2\right).$$

If only σ^2 is considered to be unknown, then notice that

$$p(\mathcal{X}; \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left(-\frac{N}{2\sigma^2} \bar{S}_{\sigma^2}\right) \equiv g_1(\bar{S}_{\sigma^2}, \sigma^2),$$

where g_1 is a function that depends only on $\bar{S}_{\sigma^2}$ and the unknown σ^2. According to the Fisher-Neyman factorization theorem, the statistic $\bar{S}_{\sigma^2}$ is sufficient.

Assume now the case where both (μ, σ^2) are unknown. Notice that by

$$\begin{aligned}\sum_{n=1}^{N} (x_n - \mu)^2 &= \sum_{n=1}^{N} (x_n - S_\mu)^2 + N(S_\mu - \mu)^2 \\ &= N S_{\sigma^2} + N(S_\mu - \mu)^2,\end{aligned}$$

the previous joint pdf becomes

$$\begin{aligned}p(\mathcal{X}; \mu, \sigma^2) &= \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left(-\frac{1}{2\sigma^2} \left(N S_{\sigma^2} + N(S_\mu - \mu)^2\right)\right) \\ &:= g_2\left((S_\mu, S_{\sigma^2}), (\mu, \sigma^2)\right).\end{aligned}$$

It can be readily verified that g_2 depends only on the statistic (S_μ, S_{σ^2}) and the unknowns (μ, σ^2). Hence, once again, by the Fisher-Neyman factorization theorem, the statistic (S_μ, S_{σ^2}) is sufficient.

3.11. Show that solving the task

$$\text{minimize} \qquad L(\boldsymbol{\theta}, \lambda) = \sum_{n=1}^{N} \left(y_n - \theta_0 - \sum_{i=1}^{l} \theta_i x_{ni} \right)^2 + \lambda \sum_{i=1}^{l} |\theta_i|^2,$$

is equivalent with minimizing

$$\text{minimize} \qquad L(\boldsymbol{\theta}, \lambda) = \sum_{n=1}^{N} \left((y_n - \bar{y}) - \sum_{i=1}^{l} \theta_i (x_{ni} - \bar{x}_i) \right)^2 + \lambda \sum_{i=1}^{l} |\theta_i|^2,$$

and the estimate of θ_0 is given by

$$\hat{\theta}_0 = \bar{y} - \sum_{i=1}^{l} \hat{\theta}_i \bar{x}_i.$$

Solution: We have that

$$L(\theta_0, \theta_{1:l}) = \sum_{n=1}^{N} \Big(y_n - \theta_0 - \sum_{i=1}^{l} \theta_i x_{ni} \Big)^2 + \sum_{i=1}^{l} \theta_i^2.$$

Taking first the derivative with respect to θ_0 and setting it equal to zero we obtain

$$\frac{\partial L}{\partial \theta_0} = \sum_{n=1}^{N} \Big(-2(y_n - \theta_0) + 2 \sum_{i=1}^{l} \theta_i x_{ni} \Big) = 0$$

or

$$N\theta_0 = \sum_{n=1}^{N} y_n - \sum_{i=1}^{l} \theta_i \sum_{n=1}^{N} x_{ni},$$

which results in

$$\theta_0 = \bar{y} - \sum_{i=1}^{l} \theta_i \bar{x}_i.$$

That is, the optimum value for θ_0, is given in terms of the rest components. Thus, optimizing with respect to θ_i, $i = 1, 2, \ldots, l$, this has to be taken into account. Substituting the above in the Lagrangian, we get

$$L(\hat{\theta}_0, \theta_{1:l}) = \sum_{n=1}^{N} \Big(y_n - \bar{y} - \sum_{i=1}^{l} \theta_i (x_{ni} - \bar{x}_i) \Big)^2 + \sum_{i=1}^{l} \theta_i^2,$$

which proves the claim.

Note that the exact form of the regularizer does not enter into the game, since does not depend on θ_0. Hence, this technique of centering the data is also applicable to other forms of regularization.

3.12. This problem refers to Example 3.4, where a linear regression task with a real valued unknown parameter θ is considered. Show that $\mathrm{MSE}(\hat{\theta}_b(\lambda)) < \mathrm{MSE}(\hat{\theta}_{\mathrm{MVU}})$, i.e., the ridge regression estimate shows a lower MSE performance than the one for the MVU estimate, if

$$\begin{cases} \lambda \in (0, \infty), & \theta^2 \leq \frac{\sigma_\eta^2}{N}, \\ \lambda \in \left(0, \frac{2\sigma_\eta^2}{\theta^2 - \frac{\sigma_\eta^2}{N}}\right), & \theta^2 > \frac{\sigma_\eta^2}{N}. \end{cases}$$

Moreover, the minimum MSE performance for the ridge regression estimate is attained at $\lambda_* = \sigma_\eta^2/\theta^2$.

Solution: Theory suggests that our estimate $\hat{\theta}_b$ is the solution of the task of minimizing the following loss function with respect to $\theta \in \mathbb{R}$:

$$L(\theta, \lambda) = \sum_{n=1}^{N} (y_n - \theta)^2 + \lambda\theta^2, \quad \lambda \geq 0.$$

The minimizer $\hat{\theta}_b$ will be obtained if we set the gradient $dL(\theta, \lambda)/d\theta$ equal to zero, or equivalently,

$$\hat{\theta}_b(\lambda) = \frac{N}{N+\lambda} \frac{1}{N} \sum_{n=1}^{N} y_n := \frac{N}{N+\lambda} \hat{\theta}_{\mathrm{MVU}},$$

where we used the notation $\hat{\theta}_b(\lambda)$ in order to highlight the dependence of the estimate $\hat{\theta}_b$ on the parameter λ. Notice here that $\mathbb{E}[\hat{\theta}_b(\lambda)] = \frac{N}{N+\lambda}\theta_0$, where θ_0 is the estimandum.

Elementary calculus helps us to express $\mathrm{MSE}\left(\hat{\theta}_b(\lambda)\right)$ as

$$\begin{aligned} \mathrm{MSE}\left(\hat{\theta}_b(\lambda)\right) &= \mathbb{E}\left[\left(\hat{\theta}_b(\lambda) - \mathbb{E}[\hat{\theta}_b(\lambda)]\right)^2\right] + \left(\mathbb{E}[\hat{\theta}_b(\lambda)] - \theta_0\right)^2 \\ &= \frac{N^2 \mathrm{MSE}\left(\hat{\theta}_{\mathrm{MVU}}\right) + \lambda^2\theta_0^2}{(N+\lambda)^2}, \end{aligned} \tag{10}$$

and

$$\frac{d}{d\lambda}\mathrm{MSE}\left(\hat{\theta}_b(\lambda)\right) = \frac{2\theta_0^2\lambda(N+\lambda)^2 - 2\left(N^2\mathrm{MSE}\left(\hat{\theta}_{\mathrm{MVU}}\right) + \lambda^2\theta_0^2\right)(N+\lambda)}{(N+\lambda)^4}. \tag{11}$$

Let us first examine the range of values of $\lambda > 0$ which guarantee that $\mathrm{MSE}\left(\hat{\theta}_b(\lambda)\right) < \mathrm{MSE}\left(\hat{\theta}_{\mathrm{MVU}}\right)$. The case of $\lambda = 0$ is excluded from the discussion, since in such a case, $\hat{\theta}_b(0) = \hat{\theta}_{\mathrm{MVU}}$. It is easy to verify by (10) that

$$\begin{aligned}
&\text{MSE}\left(\hat{\theta}_b(\lambda)\right) < \text{MSE}\left(\hat{\theta}_{\text{MVU}}\right)\\
\Leftrightarrow \lambda\left(\theta_0^2 - \text{MSE}\left(\hat{\theta}_{\text{MVU}}\right)\right) &< 2N\text{MSE}\left(\hat{\theta}_{\text{MVU}}\right).
\end{aligned}$$

In the case where $\theta_0^2 > \text{MSE}\left(\hat{\theta}_{\text{MVU}}\right)$, then the desired λ belongs to the interval $(0, 2\sigma_\eta^2/(\theta_0^2 - \sigma_\eta^2/N))$, where we have used the fact that

$$\text{MSE}\left(\hat{\theta}_{\text{MVU}}\right) = \frac{\sigma_\eta^2}{N}.$$

In the case where $\theta_0^2 \leq \text{MSE}\left(\hat{\theta}_{\text{MVU}}\right)$, notice that $\forall\lambda > 0$, we have that

$$\lambda\left(\theta_0^2 - \text{MSE}\left(\hat{\theta}_{\text{MVU}}\right)\right) \leq 0 < 2N\text{MSE}\left(\hat{\theta}_{\text{MVU}}\right),$$

i.e., the desired λ belongs to the interval $(0, \infty)$.

It is also easy to verify by equating the numerator of (11) to zero that the λ_* which minimizes $\text{MSE}\left(\hat{\theta}_b(\lambda)\right)$ becomes equal to σ_η^2/θ_0^2. To leave no place for ambiguity, we remark here that in the case where $\theta_0^2 > \text{MSE}\left(\hat{\theta}_{\text{MVU}}\right) = \sigma_\eta^2/N$, this λ_* belongs to the interval $(0, 2\sigma_\eta^2/(\theta_0^2 - \sigma_\eta^2/N))$, since

$$0 < \lambda_* = \frac{\sigma_\eta^2}{\theta_0^2} < 2\frac{\sigma_\eta^2}{\theta_0^2} < \frac{2\sigma_\eta^2}{\theta_0^2 - \frac{\sigma_\eta^2}{N}}.$$

3.13. Consider, once more, the same regression model as that of Problem 3.9, but with $\Sigma_\eta = I_N$. Compute the MSE of the predictions $\mathbb{E}[(\text{y} - \hat{\text{y}})^2]$, where y is the true response and $\hat{\text{y}}$ is the predicted value, given a test point $\boldsymbol{x}$ and using the LS estimator,

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

The LS estimator has been obtained via a set of N measurements, collected in the input matrix X and $\mathbf{y}$, where the notation has been introduced previously in this chapter. The expectation $\mathbb{E}[\cdot]$ is taken with respect to to y, the training data, $\mathcal{D}$ and the test points $\mathbf{x}$. Observe the dependence of the MSE on the dimensionality of the space.

Hint: Consider, first, the MSE, given the value of a test point $\boldsymbol{x}$, and then take the average over all the test points.

Solution: From the theory, we have that given a point $\boldsymbol{x}$, the LS estimator of the output is given by

$$\hat{\text{y}} = \mathbf{y}^T X (X^T X)^{-1} \boldsymbol{x}.$$

Moreover,

$$\begin{aligned}
\mathrm{MSE}(\hat{\theta}(\boldsymbol{x})) &= \mathbb{E}\left[(\mathrm{y}-\hat{\mathrm{y}})^2\right] = \mathbb{E}\left[(\boldsymbol{\theta}^T\boldsymbol{x}+\eta-\hat{\theta}^T\boldsymbol{x})^2\right] \\
&= \sigma_\eta^2 + \mathbb{E}_{\mathcal{D}}\left[\boldsymbol{x}^T(\boldsymbol{\theta}-\hat{\theta})(\boldsymbol{\theta}-\hat{\theta})^T\boldsymbol{x}\right] + 2\,\mathbb{E}_{\mathcal{D}|\eta}\left[(\boldsymbol{\theta}-\hat{\theta})^T\boldsymbol{x}\,\mathbb{E}_\eta[\eta]\right] \\
&= \sigma_\eta^2 + \mathbb{E}_{\mathcal{D}}\left[\boldsymbol{x}^T(\boldsymbol{\theta}-\hat{\theta})(\boldsymbol{\theta}-\hat{\theta})^T\boldsymbol{x}\right] \\
&= \sigma_\eta^2 + \left[\boldsymbol{x}^T\Sigma_{\hat{\theta}}\boldsymbol{x}\right] = \sigma_\eta^2 + \sigma_\eta^2\left[\boldsymbol{x}^T(X^TX)^{-1}\boldsymbol{x}\right],
\end{aligned}$$

where in place of the covariance matrix of the LS estimator we used

$$\Sigma_{\hat{\theta}} = \sigma_\eta^2(X^TX)^{-1}.$$

Indeed, we have that

$$\begin{aligned}
\hat{\theta} &= (X^TX)^{-1}X^T\mathbf{y}, \\
&= (X^TX)^{-1}X^T(X\boldsymbol{\theta}+\boldsymbol{\eta}) = \boldsymbol{\theta} + (X^TX)^{-1}X^T\boldsymbol{\eta}, \qquad (12)
\end{aligned}$$

or

$$\hat{\theta}-\boldsymbol{\theta} = (X^TX)^{-1}X^T\boldsymbol{\eta}.$$

Hence,

$$\begin{aligned}
\Sigma_{\hat{\theta}} &= \mathbb{E}_{\mathcal{D}}\left[(X^TX)^{-1}X^T\boldsymbol{\eta}\boldsymbol{\eta}^TX(X^TX)^{-1}\right] \\
&= (X^TX)^{-1}X^T\,\mathbb{E}_\eta[\boldsymbol{\eta}\boldsymbol{\eta}^T]X(X^TX)^{-1} \\
&= \sigma_\eta^2(X^TX)^{-1}, \qquad (13)
\end{aligned}$$

where we used the fact that the variation in the data sets is solely due to noise variation (thus only to the output variables) and not to the input, X, which is a major assumption underlying the LS method.

We can now make the following approximation, for large values of N:

$$\Sigma := \mathbb{E}_x\left[\mathbf{x}\mathbf{x}^T\right] \approx \frac{1}{N}X^TX,$$

where Σ is the covariance matrix of the (zero mean) input vectors. Then, we have

$$\begin{aligned}
\mathrm{MSE}(\hat{\theta}) &\approx \sigma_\eta^2 + \frac{\sigma_\eta^2}{N}\,\mathbb{E}_x\left[\mathbf{x}^T\Sigma^{-1}\mathbf{x}\right] \\
&= \frac{\sigma_\eta^2}{N}\,\mathbb{E}_x\left[\mathrm{trace}\left\{\Sigma^{-1}\mathbf{x}\mathbf{x}^T\right\}\right] \\
&= \frac{\sigma_\eta^2}{N}\,\mathrm{trace}\left\{\Sigma^{-1}\,\mathbb{E}_x\left[\mathbf{x}\mathbf{x}^T\right]\right\} \\
&= \frac{\sigma_\eta^2}{N}\,\mathrm{trace}\left\{\Sigma^{-1}\Sigma\right\} = \frac{\sigma_\eta^2}{N}l.
\end{aligned}$$

In other words, the MSE is proportional to the dimensionality of the space as well as the variance of the noise (which for the case of the problem is taken equal to one), and inversely proportional to the number of data points. That is, for given number of points and noise variance, the error depends on the dimensionality, which is a manifestation of the curse of dimensionality.

3.14. Assume that the model that generates the data is

$$y_n = A \sin\left(\frac{2\pi}{N} kn + \phi\right) + \eta_n, \tag{14}$$

where $A > 0$, and $k \in \{1, 2, \dots, N-1\}$. Assume that η_n are samples from a Gaussian white noise, of variance σ_η^2. Show that there is no unbiased estimator for the phase, ϕ, based on N measurement points, y_n, $n = 0, 1, \dots N-1$, that attains the Cramér-Rao bound.

Solution: The joint pdf of the measurements $\boldsymbol{y} := [y_0, y_1, \dots, y_{N-1}]^T$ is given by

$$p(\boldsymbol{y}; \phi) = \frac{1}{(2\pi\sigma_\epsilon^2)^{N/2}} \exp\left(-\frac{1}{2\sigma_\epsilon^2} \sum_{n=0}^{N-1} \left(y_n - A\sin\left(\frac{2\pi}{N}kn + \phi\right)\right)^2\right).$$

The two derivatives of the ln of this function can be easily shown to be

$$\frac{\partial \ln p(\boldsymbol{y}; \phi)}{\partial \phi} = \frac{A}{\sigma_\epsilon^2} \sum_{n=0}^{N-1} \left(y_n \cos\left(\frac{2\pi}{N}kn + \phi\right) - \frac{A}{2}\sin\left(\frac{4\pi}{N}kn + 2\phi\right)\right), \tag{15}$$

and

$$\frac{\partial^2 \ln p(\boldsymbol{y}; \phi)}{\partial \phi^2} = -\frac{A}{\sigma_\epsilon^2} \sum_{n=0}^{N-1} \left(y_n \sin\left(\frac{2\pi}{N}kn + \phi\right) + A\cos\left(\frac{4\pi}{N}kn + 2\phi\right)\right),$$

and by substituting the value of y_n from (14), the mean value becomes

$$\begin{aligned}\mathbb{E}\left[\frac{\partial^2 \ln p(\mathbf{y}; \phi)}{\partial \phi^2}\right] &= -\frac{A^2}{\sigma_\epsilon^2} \sum_{n=0}^{N-1} \left(\frac{1}{2} + \frac{1}{2}\cos\left(\frac{4\pi}{N}kn + 2\phi\right)\right) \\ &= -\frac{NA^2}{2\sigma_\epsilon^2}.\end{aligned}$$

To derive the above, we used the trigonometric formula $\sin^2 \alpha = (1 -$

$\cos(2\alpha))/2$, and also the fact

$$\begin{aligned}\sum_{n=0}^{N-1}\cos\left(\frac{4\pi}{N}kn+2\phi\right) &= \frac{1}{2}\sum_{n=0}^{N-1}\left(e^{j\left(\frac{4\pi}{N}kn+2\phi\right)}+e^{-j\left(\frac{4\pi}{N}kn+2\phi\right)}\right)\\ &= \frac{1}{2}e^{2j\phi}\sum_{n=0}^{N-1}e^{j\frac{4\pi}{N}kn}+\frac{1}{2}e^{-2j\phi}\sum_{n=0}^{N-1}e^{-j\frac{4\pi}{N}kn}\\ &= \frac{1}{2}e^{2j\phi}\frac{1-e^{j\frac{4\pi}{N}kN}}{1-e^{j\frac{4\pi}{N}k}}+\frac{1}{2}e^{-2j\phi}\frac{1-e^{-j\frac{4\pi}{N}kN}}{1-e^{-j\frac{4\pi}{N}k}}\\ &= 0,\end{aligned}$$

since $e^{-j\frac{4\pi}{N}kN}=1$, where $j:=\sqrt{-1}$.

Hence, if $\hat{\phi}$ stands for an unbiased estimator of ϕ, then

$$\mathrm{var}(\hat{\phi}) \geq \frac{2\sigma_\epsilon^2}{NA^2}.$$

However, looking back at (15), we can verify that there does not exist a function g such that $\forall \boldsymbol{y}\in\mathbb{R}^N$,

$$g(\boldsymbol{y})-\phi=\frac{2}{NA}\sum_{n=0}^{N-1}\left(y_n\cos\left(\frac{2\pi}{N}kn+\phi\right)-\frac{A}{2}\sin\left(\frac{4\pi}{N}kn+2\phi\right)\right).$$

Thus, even if an unbiased estimator exists, this cannot achieve the Cramér-Rao bound.

3.15. Show that if $(\mathbf{y},\mathbf{x})$ are two jointly distributed random vectors, with values in $\mathbb{R}^k\times\mathbb{R}^l$, then the MSE optimal estimator of $\mathbf{y}$ given the value $\mathbf{x}=\boldsymbol{x}$ is the regression of $\mathbf{y}$ conditioned on $\boldsymbol{x}$, i.e., $\mathbb{E}[\mathbf{y}|\boldsymbol{x}]$.

Solution: The proof follows a similar line as the scalar case. Let

$$\boldsymbol{f}(\boldsymbol{x}):=[f_1(\boldsymbol{x}),\dots,f_k(\boldsymbol{x})]^T$$

be the vector estimator. Then the MSE optimal one should minimize the sum of square errors per component, i.e.,

$$\mathbb{E}\left[\sum_{i=1}^{k}(\mathrm{y}_i-f_i(\boldsymbol{x}))^2\right]=\sum_{i=1}^{k}\mathbb{E}\left[(\mathrm{y}_i-f_i(\boldsymbol{x}))^2\right].$$

This is equivalent with minimizing k scalar terms individually, which can be carried out as in the text in the Chapter. The result of the ith problem is that the respective ith component of the MSE optimal estimator is given by,

$$\hat{g}_i(\boldsymbol{x})=\mathbb{E}[\mathrm{y}_i|\boldsymbol{x}],$$

or

$$\hat{\boldsymbol{g}}(\boldsymbol{x})=\mathbb{E}[\mathbf{y}|\boldsymbol{x}].$$

Note that minimizing the sum of square errors per component is equivalent with minimizing the trace of the error covariance,

$$\mathbb{E}\left[\mathbf{e}\mathbf{e}^T\right] = \mathbb{E}\left[(\mathbf{y} - \boldsymbol{f}(\boldsymbol{x}))(\mathbf{y} - \boldsymbol{f}(\boldsymbol{x}))^T\right].$$

3.16. Assume that $\mathbf{x}$, $\mathbf{y}$ are jointly Gaussian random vectors, with covariance matrix

$$\Sigma := \mathbb{E}\left[\begin{bmatrix}\mathbf{x} - \boldsymbol{\mu}_x \\ \mathbf{y} - \boldsymbol{\mu}_y\end{bmatrix}\left[(\mathbf{x} - \boldsymbol{\mu}_x)^T, (\mathbf{y} - \boldsymbol{\mu}_y)^T\right]\right] = \begin{bmatrix}\Sigma_x & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_y\end{bmatrix}.$$

Assuming also that the matrices Σ_x and $\bar{\Sigma} := \Sigma_y - \Sigma_{yx}\Sigma_x^{-1}\Sigma_{xy}$ are non-singular, then show that the optimal MSE estimator $\mathbb{E}[\mathbf{y}|\boldsymbol{x}]$ takes the following form,

$$\mathbb{E}[\mathbf{y}|\boldsymbol{x}] = \mathbb{E}[\mathbf{y}] + \Sigma_{yx}\Sigma_x^{-1}(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}]).$$

Notice that $\mathbb{E}[\mathbf{y}|\boldsymbol{x}]$ is an affine function of $\boldsymbol{x}$. In other words, for the case where $\mathbf{x}$ and $\mathbf{y}$ are jointly Gaussian, the optimal estimator of $\mathbf{y}$, in the MSE sense, which is in general a non-linear function, becomes an affine function of $\boldsymbol{x}$.

In the special case where x, y are scalar random variables, then

$$\mathbb{E}[\mathrm{y}|x] = \mu_y + \frac{\alpha\sigma_y}{\sigma_x}(x - \mu_x),$$

where α stands for the *correlation coefficient*, defined as

$$\alpha := \frac{\mathbb{E}\left[(\mathrm{x} - \mu_x)(\mathrm{y} - \mu_y)\right]}{\sigma_x\sigma_y},$$

with $|\alpha| \leq 1$. Notice, also, that the previous assumption on the non-singularity of Σ_x and $\bar{\Sigma}$ translates, in this special case, to $\sigma_x \neq 0 \neq \sigma_y$, and $|\alpha| < 1$.

Solution: First, it is easy to verify that $\Sigma_{yx} = \Sigma_{xy}^T$. Moreover, since Σ_x and $\bar{\Sigma}$ are assumed to be non-singular, then it can be verified, e.g., [Magn 99], that the determinant $\det\Sigma = \det\Sigma_x \det\bar{\Sigma}$, and that

$$\Sigma^{-1} = \begin{bmatrix}\Sigma_x^{-1} + \Sigma_x^{-1}\Sigma_{xy}\bar{\Sigma}^{-1}\Sigma_{yx}\Sigma_x^{-1} & -\Sigma_x^{-1}\Sigma_{xy}\bar{\Sigma}^{-1} \\ -\bar{\Sigma}^{-1}\Sigma_{yx}\Sigma_x^{-1} & \bar{\Sigma}^{-1}\end{bmatrix}.$$

Observe that $\bar{\Sigma}$ is the Schur complement of $\Sigma_{\boldsymbol{x}}$ in Σ. Also, the previous formula is the matrix inversion formula in terms of the Schur complement, as provided in the Appendix A of the book. To save space, let $\bar{\boldsymbol{x}} := \boldsymbol{x} - \boldsymbol{\mu}_x$

and $\bar{\boldsymbol{y}} := \boldsymbol{y} - \boldsymbol{\mu}_y$. Then, the joint pdf of $\mathbf{x}$ and $\mathbf{y}$ becomes

$$\begin{aligned}
p(\boldsymbol{x}, \boldsymbol{y}) &= \frac{1}{(2\pi)^l (\det \Sigma)^{\frac{1}{2}}} \exp\left(-\frac{1}{2} \left[\bar{\boldsymbol{x}}^T, \bar{\boldsymbol{y}}^T\right] \Sigma^{-1} \begin{bmatrix} \bar{\boldsymbol{x}} \\ \bar{\boldsymbol{y}} \end{bmatrix} \right) \\
&= \frac{1}{(2\pi)^l (\det \Sigma)^{\frac{1}{2}}} \exp\left(-\frac{1}{2} \bar{\boldsymbol{x}}^T \Sigma_x^{-1} \bar{\boldsymbol{x}} - \frac{1}{2} \bar{\boldsymbol{x}}^T \Sigma_x^{-1} \Sigma_{xy} \bar{\Sigma}^{-1} \Sigma_{yx} \Sigma_x^{-1} \bar{\boldsymbol{x}} \right. \\
&\quad \left. + \bar{\boldsymbol{x}}^T \Sigma_x^{-1} \Sigma_{xy} \bar{\Sigma}^{-1} \bar{\boldsymbol{y}} - \frac{1}{2} \bar{\boldsymbol{y}}^T \bar{\Sigma}^{-1} \bar{\boldsymbol{y}} \right) \\
&= \frac{1}{(2\pi)^l (\det \Sigma_x)^{\frac{1}{2}} (\det \bar{\Sigma})^{\frac{1}{2}}} \exp\left(-\frac{1}{2} \bar{\boldsymbol{x}}^T \Sigma_x^{-1} \bar{\boldsymbol{x}} \right. \\
&\quad \left. -\frac{1}{2} \left(\bar{\boldsymbol{y}} - \Sigma_{yx} \Sigma_x^{-1} \bar{\boldsymbol{x}} \right)^T \bar{\Sigma}^{-1} \left(\bar{\boldsymbol{y}} - \Sigma_{yx} \Sigma_x^{-1} \bar{\boldsymbol{x}} \right) \right).
\end{aligned}$$

As a result, the marginal pdf $p(\boldsymbol{x})$ becomes

$$\begin{aligned}
p(\boldsymbol{x}) = \int p(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{y} &= \frac{1}{(2\pi)^{l/2} (\det \Sigma_x)^{\frac{1}{2}}} \exp\left(-\frac{1}{2} \bar{\boldsymbol{x}}^T \Sigma_x^{-1} \bar{\boldsymbol{x}} \right) \\
&\quad \times \frac{1}{(2\pi)^{l/2} (\det \bar{\Sigma})^{\frac{1}{2}}} \int \exp\left(-\frac{1}{2} \left(\boldsymbol{y} - \boldsymbol{\mu}_y - \Sigma_{yx} \Sigma_x^{-1} \bar{\boldsymbol{x}} \right)^T \right. \\
&\quad \left. \times\ \bar{\Sigma}^{-1} \left(\boldsymbol{y} - \boldsymbol{\mu}_y - \Sigma_{yx} \Sigma_x^{-1} \bar{\boldsymbol{x}} \right) \right) d\boldsymbol{y} \\
&= \frac{1}{(2\pi)^{l/2} (\det \Sigma_x)^{\frac{1}{2}}} \exp\left(-\frac{1}{2} \bar{\boldsymbol{x}}^T \Sigma_x^{-1} \bar{\boldsymbol{x}} \right).
\end{aligned}$$

Using the previous relations, we can easily see that

$$\begin{aligned}
p(\boldsymbol{y}|\boldsymbol{x}) &= \frac{p(\boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{x})} \\
&= \frac{1}{(2\pi)^{\frac{l}{2}} (\det \bar{\Sigma})^{\frac{1}{2}}} \exp\left(-\frac{\left(\bar{\boldsymbol{y}} - \Sigma_{yx} \Sigma_x^{-1} \bar{\boldsymbol{x}} \right)^T \bar{\Sigma}^{-1} \left(\bar{\boldsymbol{y}} - \Sigma_{yx} \Sigma_x^{-1} \bar{\boldsymbol{x}} \right)}{2} \right).
\end{aligned}$$

A simple inspection of this relation shows that the conditional pdf $p(\boldsymbol{y}|\boldsymbol{x})$ is Gaussian with covariance matrix $\bar{\Sigma}$ and conditional mean $\mathbb{E}[\boldsymbol{y}|\boldsymbol{x}] = \boldsymbol{\mu}_y + \Sigma_{yx} \Sigma_x^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_x)$.

3.17. Assume a number l of jointly Gaussian random variables $\{\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_l\}$, and a non-singular matrix $A \in \mathbb{R}^{l \times l}$. If $\mathbf{x} := [\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_l]^T$, then show that the components of the vector $\mathbf{y}$, obtained by $\mathbf{y} = A\mathbf{x}$, are also jointly Gaussian random variables.

A direct consequence of this result is that any linear combination of jointly Gaussian variables is also Gaussian.

Solution: The Jacobian matrix of a linear transform $\mathbf{y} = A\mathbf{x}$ is easily shown to be

$$J := J(\mathbf{y}; \mathbf{x}) = A.$$

Also, since A is non-singular, we have that $\mathbf{x} = A^{-1}\mathbf{y}$. Without any loss of generality, assume that $\mathbb{E}[\mathbf{x}] = \mathbf{0}$, which results into $\mathbb{E}[\mathbf{y}] = \mathbf{0}$. Hence,

$$\Sigma_y = \mathbb{E}\left[\mathbf{y}\mathbf{y}^T\right] = \mathbb{E}\left[A\mathbf{x}\mathbf{x}^T A^T\right] = A\Sigma_x A^T.$$

Clearly, $\det \Sigma_y = (\det A)^2 \det \Sigma_x$. Then, by the theorem of transformation for random variables, e.g., [Papo 02], we have the following:

$$\begin{aligned} p(\boldsymbol{y}) &= \frac{p(\boldsymbol{x})}{|\det J|} = \frac{p(A^{-1}\boldsymbol{y})}{|\det A|} \\ &= \frac{1}{(2\pi)^{l/2}|\det A|(\det \Sigma_x)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\boldsymbol{y}^T A^{-T}\Sigma_x^{-1}A^{-1}\boldsymbol{y}\right) \\ &= \frac{1}{(2\pi)^{l/2}(\det \Sigma_y)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\boldsymbol{y}^T\Sigma_y^{-1}\boldsymbol{y}\right), \end{aligned}$$

which establishes the first claim.

For the second claim, assume a non-zero vector $\boldsymbol{a} \in \mathbb{R}^l$, and define the linear combination of $\{\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_l\}$ as $\mathrm{y} = \boldsymbol{a}^T\mathbf{x}$. Elementary linear algebra guarantees that there always exists a set of non-zero l-dimensional vectors $\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{l-1}\}$ such that the collection $\{\boldsymbol{a}, \boldsymbol{a}_1, \ldots, \boldsymbol{a}_{l-1}\}$ constitutes a basis of $\mathbb{R}^l$ [Magn 99]. Thus, the matrix $A := [\boldsymbol{a}, \boldsymbol{a}_1, \ldots, \boldsymbol{a}_{l-1}]^T \in \mathbb{R}^{l\times l}$ is non-singular, and the first component of the vector $\mathbf{y} = A\mathbf{x}$ is the quantity $\mathrm{y} = \boldsymbol{a}^T\mathbf{x}$. We have already seen by the first claim that the components of $\mathbf{y}$ are jointly Gaussian random variables. Moreover, a classical result states that if a number of random variables are jointly Gaussian, then each one of them, and thus y, is also Gaussian (the opposite is not always true) [Papo 02]. This establishes the second claim of the problem.

3.18. Let $\mathbf{x} \in \mathbb{R}^l$ be a vector of jointly Gaussian random variables, of covariance matrix Σ_x. Consider the general linear regression model

$$\mathbf{y} = \Theta\mathbf{x} + \boldsymbol{\eta},$$

where $\Theta \in \mathbb{R}^{k\times l}$ is a parameter matrix and $\boldsymbol{\eta}$ is the vector of noise samples, which are considered to be Gaussian, with zero mean, and with covariance matrix Σ_η, independent of $\mathbf{x}$. Then show that $\mathbf{y}$ and $\mathbf{x}$ are jointly Gaussian, with covariance matrix given by

$$\Sigma = \begin{bmatrix} \Theta\Sigma_x\Theta^T + \Sigma_{eta} & \Theta\Sigma_x \\ \Sigma_x\Theta^T & \Sigma_x \end{bmatrix}.$$

Solution: The combined vector is given by

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{x} \end{bmatrix} = \begin{bmatrix} \Theta\mathbf{x} + \boldsymbol{\eta} \\ \mathbf{x} \end{bmatrix} = \underbrace{\begin{bmatrix} \Theta & I_k \\ I_l & 0_{l\times k} \end{bmatrix}}_{A} \begin{bmatrix} \mathbf{x} \\ \boldsymbol{\eta} \end{bmatrix}.$$

However, since $\mathbf{x}$ and $\boldsymbol{\eta}$ are both Gaussian vector variables, and mutually independent, then they are also jointly Gaussian. Notice also that the matrix A is non-singular; indeed, a simple permutation of the columns of A leads to the matrix $\begin{bmatrix} I_k & \Theta \\ 0_{l\times k} & I_l \end{bmatrix}$, whose determinant can be easily seen to be equal to 1.

Therefore, according to Problem 3.17, $[\mathbf{y}^T, \mathbf{x}^T]^T$ is also jointly Gaussian. The covariance matrix is a straightforward result following the definitions of the involved variables.

3.19. Show that a linear combination of Gaussian independent variables is also Gaussian.

Solution: This is a direct consequence of Problem 3.17, since independent Gaussian variables can be readily checked out that they are also jointly Gaussian.

3.20. Show that if a sufficient statistic $T(\mathcal{X})$ for a parameter estimation problem exists, then $T(\mathcal{X})$ suffices to express the respective ML estimate.

Solution: This is direct consequence of the Fisher-Neyman factorization theorem. Indeed, recall that $T(\mathcal{X})$ is sufficient iff the respective joint pdf can be factored as: $p(\mathcal{X};\boldsymbol{\theta}) = h(\mathcal{X})g(T(\mathcal{X}),\boldsymbol{\theta})$, where h and g are appropriate functions. Hence, by the definition of the ML estimate,

$$\hat{\boldsymbol{\theta}}_{\text{ML}} \equiv \arg\max_{\boldsymbol{\theta}} p(\mathcal{X};\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} g(T(\mathcal{X}),\boldsymbol{\theta}).$$

In other words, $T(\mathcal{X})$ is sufficient, via g, to obtain the ML estimate.

3.21. Show that if an efficient estimator exists then it is also optimal in the ML sense.

Solution: Assume the existence of an efficient estimator, i.e., a function g which achieves the Cramér-Rao bound. A necessary and sufficient condition for g to be efficient, is for (8) to hold true for all values of $\boldsymbol{\theta}$. Since (8) holds for all values of $\boldsymbol{\theta}$, then it holds true for $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_{\text{ML}}$. However, for this value, the left-hand-side of (8) becomes zero, and since $I(\hat{\boldsymbol{\theta}}_{\text{ML}})$ is non-singular, we obtain that $g(\mathcal{X}) = \hat{\boldsymbol{\theta}}_{\text{ML}}$. This establishes the claim.

3.22. Let the observations resulting from an experiment be x_n, $n = 1, 2, \ldots, N$. Assume that they are independent and that they originate from a Gaussian PDF $\mathcal{N}(\mu, \sigma^2)$. Both, the mean and the variance, are unknown. Prove that the ML estimates of these quantities are given by

$$\hat{\mu}_{\text{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n, \quad \hat{\sigma}^2_{\text{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \hat{\mu}_{\text{ML}})^2.$$

Solution: The log-likelihood function is given by

$$L(\mu, \sigma^2) = -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2.$$

Taking the gradient with respect to μ, σ^2, and equating it to zero we obtain the following system of equations

$$\begin{aligned} \frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n-\mu) &= 0 \\ -\frac{N}{2}\frac{1}{\sigma^2}+\frac{1}{2\sigma^4}\sum_{n=1}^{N}(x_n-\mu)^2 &= 0. \end{aligned}$$

The solution of this system leads trivially to the required result.

3.23. Let the observations $x_n,\ n = 1, 2, \ldots, N$, come from the uniform distribution

$$p(x;\theta) = \begin{cases} \frac{1}{\theta}, & 0 < x \leq \theta, \\ 0, & \text{otherwise.} \end{cases}$$

Obtain the ML estimate of θ.

Solution: The likelihood function is given by

$$L(\boldsymbol{x};\theta) = \prod_{n=1}^{N}\frac{1}{\theta} = \frac{1}{\theta^N}.$$

We know that $\theta \geq x_n,\ n = 1, \ldots, N$, or equivalently, $\theta \geq \max_{n=1,\ldots,N} x_n$. Hence, the likelihood function is maximized by taking the minimum value of θ, which is

$$\hat{\theta}_{\text{ML}} = \max\{x_1, x_2, \ldots, x_N\}.$$

3.24. Obtain the ML estimate of the parameter $\lambda > 0$ of the exponential distribution

$$p(x) = \begin{cases} \lambda\exp(-\lambda x), & x \geq 0, \\ 0, & x < 0, \end{cases}$$

based on a set of measurements, $x_n,\ n = 1, 2, \ldots, N$.

Solution: The log-likelihood function is

$$L(\boldsymbol{x};\lambda) = N\ln\lambda - \lambda\sum_{n=1}^{N}x_n.$$

Taking the derivative and equating to zero we obtain

$$\frac{N}{\lambda} - \sum_{n=1}^{N}x_n = 0,$$

which leads to

$$\hat{\lambda}_{\text{ML}} = \frac{N}{\sum_{n=1}^{N}x_n}.$$

3.25. Assume an $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$, and a stochastic process $\{x_n\}_{n=-\infty}^{\infty}$, consisting of i.i.d random variables, such that $p(x_n|\mu) = \mathcal{N}(\mu, \sigma^2)$. Consider N observations so that $\mathcal{X} \equiv \{x_1, x_2, \ldots, x_N\}$, and prove that the posterior $p(x|\mathcal{X})$, of any $x = x_{n_0}$ conditioned on $\mathcal{X}$, turns out to be Gaussian with mean μ_N and variance $\sigma^2 + \sigma_N^2$, where

$$\mu_N \equiv \frac{N\sigma_0^2\bar{x} + \sigma^2\mu_0}{N\sigma_0^2 + \sigma^2}, \quad \sigma_N^2 \equiv \frac{\sigma^2\sigma_0^2}{N\sigma_0^2 + \sigma^2}.$$

Solution: From basic theory we have that

$$p(\mu|\mathcal{X}) = \frac{p(\mathcal{X}|\mu)p(\mu)}{\int p(\mathcal{X}|\mu)p(\mu)d\mu} = \alpha p(\mu)\prod_{k=1}^{N} p(x_k|\mu),$$

or

$$\begin{aligned}
p(\mu|\mathcal{X}) &= \frac{\alpha}{\sqrt{2\pi}\sigma_0}\exp\left(-\frac{1}{2}\frac{(\mu-\mu_0)^2}{\sigma_0^2}\right)\prod_{k=1}^{N}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\frac{(x_k-\mu)^2}{\sigma^2}\right) \\
&= \alpha_1 \exp\left(-\frac{1}{2}\left(\sum_{k=1}^{N}\left(\frac{\mu - x_k}{\sigma}\right)^2 + \left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right)\right) \\
&= \alpha_2 \exp\left(-\frac{1}{2}\left(\left(\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}\right)\mu^2 - 2\left(\frac{1}{\sigma^2}\sum_{k=1}^{N} x_k + \frac{\mu_0}{\sigma_0^2}\right)\mu\right)\right) \\
&= \alpha_2 \exp\left(-\frac{1}{2}\left(\mu^2\frac{N\sigma_0^2+\sigma^2}{\sigma^2\sigma_0^2} - 2\mu\frac{N\sigma_0^2\bar{x}+\sigma^2\mu_0}{\sigma^2\sigma_0^2}\right)\right) \\
&= \alpha_2 \exp\left(-\frac{N\sigma_0^2+\sigma^2}{2\sigma^2\sigma_0^2}\left(\mu^2 - 2\mu\frac{N\sigma_0^2\bar{x}+\sigma^2\mu_0}{N\sigma_0^2+\sigma^2}\right)\right) \\
&= \alpha_3 \exp\left(-\frac{(\mu-\mu_N)^2}{2\sigma_N^2}\right),
\end{aligned}$$

where $\alpha, \alpha_1, \alpha_2, \alpha_3$ are factors independent of μ, and

$$\begin{aligned}
\bar{x} &:= \frac{1}{N}\sum_{k=1}^{N} x_k, \\
\mu_N &:= \frac{N\sigma_0^2\bar{x}+\sigma^2\mu_0}{N\sigma_0^2+\sigma^2}, \\
\sigma_N^2 &:= \frac{\sigma^2\sigma_0^2}{N\sigma_0^2+\sigma^2}.
\end{aligned}$$

Since $p(\mu|\mathcal{X})$ is a PDF, then necessarily

$$\alpha_3 = \frac{1}{\sqrt{2\pi}\sigma_N}.$$

Hence, $\lim_{N\to\infty}\sigma_N^2 = 0$, and for large N, $p(\mu|\mathcal{X})$ behaves like a δ function centered around μ_N. Thus,

$$p(x|\mathcal{X}) = \int p(x|\mu)p(\mu|\mathcal{X})d\mu \simeq p(x|\mu_N) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu_N)^2}{2\sigma^2}\right).$$

Therefore, $p(x|\mathcal{X})$ tends to a Gaussian pdf with mean μ_N and variance σ^2. Furthermore, $\lim_{N\to\infty}\mu_N = \bar{x}$.

For the general case of any value of N, and not only the case of large N, we have

$$\begin{aligned} p(x|\mathcal{X}) &= \int p(x|\mu)p(\mu|\mathcal{X})d\mu \\ &= \frac{1}{2\pi\sigma\sigma_N}\int \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\exp\left(-\frac{(\mu-\mu_N)^2}{2\sigma_N^2}\right)d\mu. \end{aligned} \tag{16}$$

Hence, in order to obtain $p(x|\mathcal{X})$, the previous integration has to take place. Here we will follow another path, which avoids any direct integration. Assume a random variable y defined as $\mathrm{y} := \xi + \nu$, where $\xi \sim \mathcal{N}(0,\sigma^2)$ and $\nu \sim \mathcal{N}(\mu_N,\sigma_N^2)$, independent of each other. It is well-known [Papo 02] that the PDF of y is given by the joint PDF of ξ and ν as follows:

$$p(y) = \int p_{\xi\nu}(y-\nu,\nu)d\nu.$$

However, since ξ and ν are assumed to be independent, then $p_{\xi\nu}(\xi,\nu) = p_\xi(\xi)p_\nu(\nu)$, and

$$\begin{aligned} p(y) &= \int p_\xi(y-\nu)p_\nu(\nu)d\nu \\ &= \frac{1}{2\pi\sigma\sigma_N}\int \exp\left(-\frac{(y-\nu)^2}{2\sigma^2}\right)\exp\left(-\frac{(\nu-\mu_N)^2}{2\sigma_N^2}\right)d\nu, \end{aligned}$$

which is identical to (16). However, recall from basic statistics [Papo 02] that y, being the sum of two independent Gaussians is also Gaussian (see, also, Problem 3.19), with mean the sum of the mean values and variance the sum of the variances. Therefore, (16) becomes

$$p(x|\mathcal{X}) = \frac{1}{\sqrt{2\pi(\sigma^2+\sigma_N^2)}}\exp\left(-\frac{(x-\mu_N)^2}{2(\sigma^2+\sigma_N^2)}\right).$$

3.26. Show that for the linear regression model,

$$\boldsymbol{y} = X\boldsymbol{\theta} + \boldsymbol{\eta},$$

the a-posteriori probability $p(\boldsymbol{\theta}|\boldsymbol{y})$ is a Gaussian one, if the prior distribution probability is given by $p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta},\Sigma_0)$, and the noise samples follow the multivariate Gaussian distribution $p(\boldsymbol{\eta}) = \mathcal{N}(\mathbf{0}, \varSigma_\eta)$. Compute the mean vector and the covariance matrix of the posterior distribution.

Solution: It can be easily checked that $p(\boldsymbol{\theta}|\boldsymbol{y}) = \text{const} \times \exp\left(-\frac{1}{2}\Psi\right)$, where

$$\begin{aligned}\Psi &= (\boldsymbol{y} - X\boldsymbol{\theta})^T \varSigma_\eta^{-1} (\boldsymbol{y} - X\boldsymbol{\theta}) + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \varSigma_0^{-1} (\boldsymbol{\theta} - \boldsymbol{\theta}_0) \\ &= \boldsymbol{y}^T \varSigma_\eta^{-1} \boldsymbol{y} - 2\boldsymbol{y}^T \varSigma_\eta^{-1} X\boldsymbol{\theta} + \boldsymbol{\theta}^T X^T \varSigma_\eta^{-1} X\boldsymbol{\theta} + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \varSigma_0^{-1} (\boldsymbol{\theta} - \boldsymbol{\theta}_0).\end{aligned}$$

From now on, all terms that will be independent of $\boldsymbol{\theta}$ will be collected in constant terms. Hence

$$\begin{aligned}\Psi &= \alpha_1 - 2\boldsymbol{y}^T \varSigma_\eta^{-1} X\boldsymbol{\theta} + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \varSigma_0^{-1} (\boldsymbol{\theta} - \boldsymbol{\theta}_0) \\ &\quad + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T X^T \varSigma_\eta^{-1} X(\boldsymbol{\theta} - \boldsymbol{\theta}_0) - \boldsymbol{\theta}_0^T X^T \varSigma_\eta^{-1} X\boldsymbol{\theta}_0 + 2\boldsymbol{\theta}_0^T X^T \varSigma_\eta^{-1} X\boldsymbol{\theta}.\end{aligned}$$

As a result,

$$\begin{aligned}\Psi &= \alpha_2 + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \left(\varSigma_0^{-1} + X^T \varSigma_\eta^{-1} X\right) (\boldsymbol{\theta} - \boldsymbol{\theta}_0) \\ &\quad + 2\boldsymbol{\theta}_0^T X^T \varSigma_\eta^{-1} X\boldsymbol{\theta} - 2\boldsymbol{y}^T \varSigma_\eta^{-1} X\boldsymbol{\theta} \\ &= \alpha_3 + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \left(\varSigma_0^{-1} + X^T \varSigma_\eta^{-1} X\right) (\boldsymbol{\theta} - \boldsymbol{\theta}_0) \\ &\quad - 2\left(\boldsymbol{y} - X\boldsymbol{\theta}_0\right)^T \varSigma_\eta^{-1} X \left(\boldsymbol{\theta} - \boldsymbol{\theta}_0\right). \end{aligned} \tag{17}$$

In the sequel, we will follow a standard trick that we do in situations like that. We introduce an auxiliary variable $\bar{\boldsymbol{\theta}}$, whose value is to be determined so that to make the following to be true,

$$\begin{aligned}\Psi &= \alpha_4 + (\boldsymbol{\theta} - \boldsymbol{\theta}_0 - \bar{\boldsymbol{\theta}})^T \left(\varSigma_0^{-1} + X^T \varSigma_\eta^{-1} X\right) (\boldsymbol{\theta} - \boldsymbol{\theta}_0 - \bar{\boldsymbol{\theta}}) \\ &= \alpha_4 + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \left(\varSigma_0^{-1} + X^T \varSigma_\eta^{-1} X\right) (\boldsymbol{\theta} - \boldsymbol{\theta}_0) \\ &\quad + \bar{\boldsymbol{\theta}}^T \left(\varSigma_0^{-1} + X^T \varSigma_\eta^{-1} X\right) \bar{\boldsymbol{\theta}} \\ &\quad - 2\bar{\boldsymbol{\theta}}^T \left(\Sigma_0^{-1} + X^T \varSigma_\eta^{-1} X\right) (\boldsymbol{\theta} - \boldsymbol{\theta}_0). \end{aligned} \tag{18}$$

Inspection of (17) and (18) indicates that this can happen if we choose

$$\bar{\boldsymbol{\theta}} = \left(\varSigma_0^{-1} + X^T \varSigma_\eta^{-1} X\right)^{-1} X^T \varSigma_\eta^{-1} \left(\boldsymbol{y} - X\boldsymbol{\theta}_0\right).$$

Then, we can finally write that

$$\Psi = \alpha_4 + \left(\boldsymbol{\theta} - \mathbb{E}[\theta|\boldsymbol{y}]\right)^T \varSigma_{\boldsymbol{\theta}|\boldsymbol{y}}^{-1} \left(\boldsymbol{\theta} - \mathbb{E}[\theta|\boldsymbol{y}]\right),$$

where

$$\mathbb{E}[\theta|\boldsymbol{y}] = \boldsymbol{\theta}_0 + \left(\varSigma_0^{-1} + X^T \varSigma_\eta^{-1} X\right)^{-1} X^T \varSigma_\eta^{-1} \left(\boldsymbol{y} - X\boldsymbol{\theta}_0\right),$$

and

$$\varSigma_{\theta|y} = \left(\varSigma_0^{-1} + X^T \varSigma_\eta^{-1} X\right)^{-1}.$$

3.27. Assume that x_n, $n = 1, 2 \ldots, N$, are i.i.d observations from a Gaussian $\mathcal{N}(\mu, \sigma^2)$. Obtain the MAP estimate of μ, if the prior follows the exponential distribution

$$p(\mu) = \lambda \exp\left(-\lambda\mu\right), \quad \lambda > 0,\ \mu \geq 0.$$

Solution: Upon defining $\mathcal{X} := \{x_1, x_2, \ldots, x_N\}$, the posterior distribution is given by

$$p(\mu|\mathcal{X}) \propto p(\mathcal{X}|\mu)p(\mu) = \frac{\lambda \exp\left(-\lambda\mu\right)}{(2\pi)^{N/2}\sigma^N} \prod_{n=1}^{N} \exp\left(-\frac{(x_n - \mu)^2}{2\sigma^2}\right).$$

Taking the ln, differentiating with respect to μ, and equating to zero we obtain

$$\frac{\partial\left(-\lambda\mu - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2\right)}{\partial\mu} = 0,$$

or

$$-\lambda + \frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n - \mu) = 0.$$

Finally,

$$\hat{\mu}_{\text{MAP}} = \frac{\sum_{n=1}^{N} x_n - \lambda\sigma^2}{N},$$

for nonnegative values of the numerator.

Bibliography

[Magn 99] Magnus, J. R., and Neudecker, H. *Matrix Differential Calculus with Applications in Statistics and Econometrics.* John Wiley & Sons, revised Ed., 1999.

[Papo 02] Papoulis, A., and Unnikrishna, P. *Probability, Random Variables, and Stochastic Processes.* McGraw Hill, 4th Ed., 2002.

Solutions To Problems of Chapter 4

4.1. Show that the set of equations

$$\Sigma\boldsymbol{\theta} = \boldsymbol{p}$$

has a unique solution if $\Sigma > 0$ and infinite many if Σ is singular.

Solution: a) Let $\Sigma > 0$. Then the linear system of equations has a unique solution. The converse is also true. Let the linear system has a unique solution, $\boldsymbol{\theta}_*$. Then $\Sigma > 0$. Indeed, if this was not the case, then there exists a nonzero vector, $\boldsymbol{a} \neq \boldsymbol{0}$, such that

$$\Sigma\boldsymbol{a} = \boldsymbol{0}.$$

Then $\boldsymbol{\theta} = \boldsymbol{\theta}_* + \boldsymbol{a}$ is also a solution, which contradicts the uniqueness assumption.

b) Let Σ is singular and let $\boldsymbol{\theta}_*$ be a solution. Then, any vector $\boldsymbol{a} \in \mathcal{N}(\Sigma)$, i.e., the null subspace, will also be a solution. For the converse, since there are infinite many solutions, let two of them, $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$. Since, $\Sigma\boldsymbol{\theta}_1 \neq \Sigma\boldsymbol{\theta}_2 \Rightarrow \Sigma(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2) = \boldsymbol{0}$, hence Σ is singular.

We still have to show that there always exists a solution. This is proved in the next problem.

4.2. Show that the set of equations

$$\Sigma\boldsymbol{\theta} = \boldsymbol{p}$$

has always a solution.

Solution: The existence of solution when $\Sigma > 0$ is obvious. Let Σ be singular. Then, in order to guarantee a solution, we have to show that $\boldsymbol{p}$ lies in the range space of Σ. For this, it suffices to show that $\boldsymbol{p} \perp \boldsymbol{a}$, $\forall \boldsymbol{a} \in \mathcal{N}(\Sigma)$. Since the range spaces and null spaces are orthogonal to each other and $\mathcal{R}(\Sigma) \bigoplus \mathcal{N}(\Sigma) = \mathbb{R}^l$, then if $\boldsymbol{p}$ is orthogonal to $\mathcal{N}(\Sigma)$, it will necessary lie in the column space (range) of matrix Σ.

Let us now assume that $\exists\ \boldsymbol{a} \in \mathcal{N}(\Sigma)$, such as

$$\boldsymbol{a}^T\boldsymbol{p} \neq 0.$$

This means that,

$$\Sigma\boldsymbol{a} = \boldsymbol{0} \Rightarrow \boldsymbol{a}^T\Sigma\boldsymbol{a} = 0 \Rightarrow \boldsymbol{a}^T\,\mathbb{E}[\mathbf{x}\mathbf{x}^T]\boldsymbol{a} = 0 \text{ or}$$
$$\mathbb{E}[\left(\boldsymbol{a}^T\mathbf{x}\right)^2] = 0 \Rightarrow \boldsymbol{a}^T\mathbf{x} = 0.$$

Hence,

$$\boldsymbol{a}^T\boldsymbol{p} = \boldsymbol{a}^T\,\mathbb{E}[\mathbf{x}\mathrm{y}] = \mathbb{E}[(\boldsymbol{a}^T\mathbf{x})\mathrm{y}] = 0,$$

which contradicts the claim $\boldsymbol{a}^T\boldsymbol{p} \neq 0$.

4.3. Show that the shape of the isovalue contours of the mean-square error ($J(\boldsymbol{\theta})$) surface,

$$J(\boldsymbol{\theta}) = J(\boldsymbol{\theta}_*) + (\boldsymbol{\theta} - \boldsymbol{\theta}_*)^T \Sigma (\boldsymbol{\theta} - \boldsymbol{\theta}_*),$$

are ellipses whose axes depend on the eigenstructure of Σ.

Hint: Assume that Σ has discrete eigenvalues.

Solution: Since Σ is symmetric, we know that it can be diagonalized,

$$\Sigma = Q \Lambda Q^T,$$

where Λ is the diagonal matrix with eigenvalues along the main diagonal and Q the unitary matrix ($QQ^T = I$) comprising the respective eigenvectors as its columns. Then we have

$$\begin{aligned} J(\boldsymbol{\theta}) &= J(\boldsymbol{\theta}_*) + (\boldsymbol{\theta} - \boldsymbol{\theta}_*)^T Q \Lambda Q^T (\boldsymbol{\theta} - \boldsymbol{\theta}_*) \\ &= J(\boldsymbol{\theta}_*) + \boldsymbol{v}^T \Lambda \boldsymbol{v} \\ &= J(\boldsymbol{\theta}_*) + \sum_{i=0}^{l-1} \lambda_i v_i^2, \end{aligned} \tag{1}$$

where

$$\boldsymbol{v} = Q^T (\boldsymbol{\theta} - \boldsymbol{\theta}_*).$$

For a fixed value of $J(\boldsymbol{\theta})$, Eq. (1) corresponds to an (hyper)ellipse centered at $\boldsymbol{v} = \mathbf{0}$, (or equivalently at $\boldsymbol{\theta} = \boldsymbol{\theta}_*$). Its principle axes are parallel to those obtained by rotating the original coordinates ($[1, 0, ..., 0]^T$, $[0, 1, ..., 0]^T$, $[0, 0, ..., 1]^T$) by Q; that is, parallel to the eigenvectors of Σ.

4.4. Prove that if the true relation between the input $\mathbf{x}$ and the true output y is linear, i.e.,

$$\mathrm{y} = \boldsymbol{\theta}_o^T \mathbf{x} + \mathrm{v}, \ \boldsymbol{\theta}_o \in \mathbb{R}^l$$

where v is independent of x, then the optimal MSE $\boldsymbol{\theta}_*$ satisfies

$$\boldsymbol{\theta}_* = \boldsymbol{\theta}_o.$$

Solution: The optimal parameter vector is given by

$$\begin{aligned} \Sigma \boldsymbol{\theta}_* &= \mathbb{E}[\mathbf{x}\mathrm{y}] = \mathbb{E}[\mathbf{x}\big(\mathbf{x}^T \boldsymbol{\theta}_o + \mathrm{v}\big)] \\ &= \Sigma \boldsymbol{\theta}_o, \end{aligned}$$

or

$$\boldsymbol{\theta}_* = \boldsymbol{\theta}_o.$$

The noise variance is equal to

$$\begin{aligned} \sigma_v^2 &= \mathbb{E}[\big(\mathrm{y} - \boldsymbol{\theta}_o^T \mathbf{x}\big)\big(\mathrm{y} - \mathbf{x}^T \boldsymbol{\theta}_o\big)] \\ &= \sigma_y^2 - 2\boldsymbol{\theta}_o^T \, \mathbb{E}[\mathbf{x}y] + \boldsymbol{\theta}_o^T \Sigma \boldsymbol{\theta}_o \\ &= \sigma_y^2 - \boldsymbol{\theta}_o^T \Sigma \boldsymbol{\theta}_o = MSE(\boldsymbol{\theta}_*). \end{aligned}$$

4.5. Show that if

$$\mathrm{y} = \boldsymbol{\theta}_o^T \mathbf{x} + \mathrm{v}, \ \ \boldsymbol{\theta}_o \in \mathbb{R}^k$$

where v is independent of $\mathbf{x}$, then the optimal MSE $\boldsymbol{\theta}_* \in \mathbb{R}^l, \ l < k$ is equal to the top l components of $\boldsymbol{\theta}_o$, if the components of $\mathbf{x}$ are uncorrelated.

Solution: Let

$$\boldsymbol{\theta}_o = \begin{bmatrix} \boldsymbol{\theta}_o^1 \\ \boldsymbol{\theta}_o^2 \end{bmatrix}, \ \boldsymbol{\theta}_o^1 \in \mathbb{R}^l, \ \boldsymbol{\theta}_o^2 \in \mathbb{R}^{k-l}.$$

Then

$$\mathrm{y} = \begin{bmatrix} \boldsymbol{\theta}_o^{1\,T} & \boldsymbol{\theta}_o^{2\,T} \end{bmatrix} \begin{bmatrix} \mathbf{x}_l \\ \boldsymbol{\phi} \end{bmatrix},$$

$$\boldsymbol{\phi} := \begin{bmatrix} \mathrm{x}(l+1) \\ \vdots \\ \mathrm{x}(k) \end{bmatrix}.$$

The optimal $\boldsymbol{\theta}_* \in \mathbb{R}^l$ is given by

$$\begin{aligned} \Sigma \boldsymbol{\theta}_* &= \mathbb{E}[\mathbf{x}_l \mathrm{y}] \\ &= \mathbb{E}\left[\mathbf{x}_l \big(\boldsymbol{\theta}_o^{1^T} \mathbf{x}_l + \boldsymbol{\theta}_o^{2^T} \boldsymbol{\phi} + \mathrm{v}\big)\right] \\ &= \Sigma \boldsymbol{\theta}_o^1 + \mathbb{E}[\mathbf{x}_l \boldsymbol{\phi}^T] \boldsymbol{\theta}_o^2. \end{aligned}$$

If the input variables are mutually uncorrelated, $\boldsymbol{\theta}_* = \boldsymbol{\theta}_o^1$. If not

$$\boldsymbol{\theta}_* = \boldsymbol{\theta}_o^1 + \Sigma^{-1}\, \mathbb{E}[\mathbf{x}_l \boldsymbol{\phi}^T] \boldsymbol{\theta}_o^2,$$

that is, it is equal to $\boldsymbol{\theta}_o^1$ plus a weighted combination of the components of $\boldsymbol{\theta}_o^2$.

4.6. Derive the normal equations by minimizing the cost in (4.15).
Hint: Express the cost in terms of the real part $\boldsymbol{\theta}_r$ and its imaginary part $\boldsymbol{\theta}_i$ of $\boldsymbol{\theta}$ and optimize with respect to $\boldsymbol{\theta}_r, \boldsymbol{\theta}_i$.

Solution: The cost function is

$$J(\boldsymbol{\theta}) := \mathbb{E}\left[|\mathrm{y} - \boldsymbol{\theta}^H \mathbf{x}|^2\right].$$

Let

$$\mathbf{x} = \mathbf{x}_r + j\mathbf{x}_i$$

and

$$\boldsymbol{\theta} = \boldsymbol{\theta}_r + j\boldsymbol{\theta}_i.$$

Then we obtain

$$\begin{aligned} J(\boldsymbol{\theta}) &\equiv J(\boldsymbol{\theta}_r, \boldsymbol{\theta}_i) = \mathbb{E}[|(\mathrm{y}_r + j\mathrm{y}_i) - (\boldsymbol{\theta}_r^T - j\boldsymbol{\theta}_i^T)(\mathbf{x}_r + j\mathbf{x}_i)|^2] \\ &= \mathbb{E}[|\big(\mathrm{y}_r - \boldsymbol{\theta}_r^T \mathbf{x}_r - \boldsymbol{\theta}_i^T \mathbf{x}_i\big) + j\big(\mathrm{y}_i - \boldsymbol{\theta}_r^T \mathbf{x}_i + \boldsymbol{\theta}_i^T \mathbf{x}_r\big)|^2] \\ &= \mathbb{E}[\big(\mathrm{y}_r - \boldsymbol{\theta}_\epsilon^T \mathbf{x}_\epsilon\big)^2 + \big(\mathrm{y}_i - \boldsymbol{\theta}_\epsilon^T \tilde{\mathbf{x}}_\epsilon\big)^2], \qquad (2) \end{aligned}$$

where

$$\boldsymbol{\theta}_\epsilon = \begin{bmatrix} \boldsymbol{\theta}_r \\ \boldsymbol{\theta}_i \end{bmatrix}, \ \mathbf{x}_\epsilon = \begin{bmatrix} \mathbf{x}_r \\ \mathbf{x}_i \end{bmatrix}, \ \tilde{\mathbf{x}}_\epsilon = \begin{bmatrix} \mathbf{x}_i \\ -\mathbf{x}_r \end{bmatrix}. \tag{3}$$

The cost in (2) can now be written as

$$\begin{aligned} J(\boldsymbol{\theta}_r, \boldsymbol{\theta}_i) \ &= \ \mathbb{E}[\mathrm{y}_r^2] - 2\,\mathbb{E}[\mathbf{x}_\epsilon^T \mathrm{y}_r]\boldsymbol{\theta}_\epsilon + \boldsymbol{\theta}_\epsilon^T\,\mathbb{E}[\mathbf{x}_\epsilon \mathbf{x}_\epsilon^T]\boldsymbol{\theta}_\epsilon + \\ &\quad + \mathbb{E}[\mathrm{y}_i^2] - 2\,\mathbb{E}[\tilde{\mathbf{x}}_\epsilon^T \mathrm{y}_i]\boldsymbol{\theta}_\epsilon + \boldsymbol{\theta}_\epsilon^T\,\mathbb{E}[\tilde{\mathbf{x}}_\epsilon \tilde{\mathbf{x}}_\epsilon^T]\boldsymbol{\theta}_\epsilon. \end{aligned}$$

Taking the derivative with respect to $\boldsymbol{\theta}_\epsilon$ we obtain

$$\left[\,\mathbb{E}[\mathbf{x}_\epsilon \mathbf{x}_\epsilon^T] + \mathbb{E}[\tilde{\mathbf{x}}_\epsilon \tilde{\mathbf{x}}_\epsilon^T]\right] \begin{bmatrix} \boldsymbol{\theta}_r \\ \boldsymbol{\theta}_i \end{bmatrix} = \mathbb{E}[\mathbf{x}_\epsilon \mathrm{y}_r] + \mathbb{E}[\tilde{\mathbf{x}}_\epsilon \mathrm{y}_i]$$

and after some algebra we obtain

$$\tilde{\Theta} \begin{bmatrix} \boldsymbol{\theta}_r \\ \boldsymbol{\theta}_i \end{bmatrix} = \begin{bmatrix} \mathbb{E}[\mathbf{x}_r \mathrm{y}_r] + \mathbb{E}[\mathbf{x}_i \mathrm{y}_i] \\ \mathbb{E}[\mathbf{x}_i \mathrm{y}_r] - \mathbb{E}[\mathbf{x}_r \mathrm{y}_i] \end{bmatrix},$$

where

$$\tilde{\Theta} := \begin{bmatrix} \mathbb{E}[\mathbf{x}_r \mathbf{x}_r^T] + \mathbb{E}[\mathbf{x}_i \mathbf{x}_i^T] & \mathbb{E}[\mathbf{x}_r \mathbf{x}_i^T] - \mathbb{E}[\mathbf{x}_i \mathbf{x}_r^T] \\ \mathbb{E}[\mathbf{x}_i \mathbf{x}_r^T] - \mathbb{E}[\mathbf{x}_r \mathbf{x}_i^T] & \mathbb{E}[\mathbf{x}_i \mathbf{x}_i^T] + \mathbb{E}[\mathbf{x}_r \mathbf{x}_r^T] \end{bmatrix}.$$

Verify now that this is the same set of equations as the normal equations, after equating the real and imaginary parts on both sides, in $\varSigma\boldsymbol{\theta} = \boldsymbol{p}$.

4.7. Consider the multichannel filtering task

$$\hat{\mathbf{y}} = \begin{bmatrix} \hat{\mathrm{y}}_r \\ \hat{\mathrm{y}}_i \end{bmatrix} = \Theta \begin{bmatrix} \mathbf{x}_r \\ \mathbf{x}_i \end{bmatrix}.$$

Estimate Θ so that to minimize the error norm:

$$\mathbb{E}[||\mathbf{y} - \hat{\mathbf{y}}||^2].$$

Solution: The error norm is equal to

$$\begin{aligned} J(\Theta) \ &= \ \mathbb{E}\left[\left(\mathrm{y}_r - \hat{\mathrm{y}}_r\right)^2\right] + \mathbb{E}\left[\left(\mathrm{y}_i - \hat{\mathrm{y}}_i\right)^2\right] \\ &= \ \mathbb{E}\left[\left(\mathrm{y}_r - \boldsymbol{\theta}_{11}^T\mathbf{x}_r - \boldsymbol{\theta}_{12}^T\mathbf{x}_i\right)^2\right] + \mathbb{E}\left[\left(\mathrm{y}_i - \boldsymbol{\theta}_{21}^T\mathbf{x}_r - \boldsymbol{\theta}_{22}^T\mathbf{x}_i\right)^2\right], \end{aligned}$$

where

$$\Theta := \begin{bmatrix} \boldsymbol{\theta}_{11}^T & \boldsymbol{\theta}_{12}^T \\ \boldsymbol{\theta}_{21}^T & \boldsymbol{\theta}_{22}^T \end{bmatrix}.$$

Hence, one can separately optimize the two terms in the cost with respect to $(\boldsymbol{\theta}_{11}, \boldsymbol{\theta}_{12})$ and $(\boldsymbol{\theta}_{21}, \boldsymbol{\theta}_{22})$, respectively.

From the first term, working similarly as in Problem 4.6 we obtain

$$\begin{bmatrix} \mathbb{E}[\mathbf{x}_r \mathbf{x}_r^T] & \mathbb{E}[\mathbf{x}_r \mathbf{x}_i^T] \\ \mathbb{E}[\mathbf{x}_i \mathbf{x}_r^T] & \mathbb{E}[\mathbf{x}_i \mathbf{x}_i^T] \end{bmatrix} \begin{bmatrix} \boldsymbol{\theta}_{11} \\ \boldsymbol{\theta}_{12} \end{bmatrix} = \begin{bmatrix} \mathbb{E}[\mathbf{x}_r \mathrm{y}_r] \\ \mathbb{E}[\mathbf{x}_i \mathrm{y}_r] \end{bmatrix}$$

and a similar set of equations, as in theory, results from the second term.

4.8. Show that (4.34) is the same as (4.25).

Solution: By the respective definitions, we have

$$\begin{aligned}\hat{\mathrm{y}}_r + j\hat{\mathrm{y}}_i &= (\boldsymbol{\theta}_r^T - j\boldsymbol{\theta}_i^T)(\mathbf{x}_r + j\mathbf{x}_i) + \\ &\quad (\mathbf{v}_r^T - j\mathbf{v}_i^T)(\mathbf{x}_r - j\mathbf{x}_i),\end{aligned}$$

or after some trivial algebra

$$\begin{bmatrix}\hat{\mathrm{y}}_r \\ \hat{\mathrm{y}}_i\end{bmatrix} = \begin{bmatrix}\boldsymbol{\theta}_r^T + \boldsymbol{v}_r^T & \boldsymbol{\theta}_i^T - \boldsymbol{v}_i^T \\ -\boldsymbol{\theta}_i^T - \boldsymbol{v}_i^T & \boldsymbol{\theta}_r^T - \boldsymbol{v}_r^T\end{bmatrix}\begin{bmatrix}\mathbf{x}_r \\ \mathbf{x}_i\end{bmatrix}, \tag{4}$$

which results in (4.25) after substituting in (4) the definitions in (4.32) and (4.33).

4.9. Show that the MSE achieved by a linear complex-valued estimator is always larger than that obtained by a widely linear one. Equality is achieved only under the circularity conditions.

Solution: The minimum MSE for the linear filter is

$$\begin{aligned}MSE_l &= \mathbb{E}[(\mathrm{y} - \boldsymbol{\theta}_*^H\mathbf{x})(\mathrm{y}^* - \mathbf{x}^H\boldsymbol{\theta}_*)] \\ &= \mathbb{E}[\mathrm{y}^2] + \boldsymbol{\theta}_*^H\,\mathbb{E}[\mathbf{x}\mathbf{x}^H]\boldsymbol{\theta}_* - \boldsymbol{\theta}_*^H\,\mathbb{E}[\mathbf{x}\mathrm{y}^*] - \mathbb{E}[\mathrm{y}\mathbf{x}^H]\boldsymbol{\theta}_* \\ &= \mathbb{E}[\mathrm{y}^2] - \mathbb{E}[\mathbf{x}^H\mathrm{y}]\boldsymbol{\theta}_* = \mathbb{E}[\mathrm{y}^2] - \boldsymbol{p}^H\Sigma_x^{-1}\boldsymbol{p},\end{aligned}$$

where

$$\Sigma_x = \mathbb{E}[\mathbf{x}\mathbf{x}^H],$$
$$\boldsymbol{p} = \mathbb{E}[\mathbf{x}\mathrm{y}^*].$$

Similarly we can show that

$$MSE_{wl} = \mathbb{E}[\mathrm{y}^2] - \tilde{\boldsymbol{p}}^H\tilde{\Sigma}_x^{-1}\tilde{\boldsymbol{p}},$$

where

$$\tilde{\boldsymbol{p}} = \begin{bmatrix}\boldsymbol{p} \\ \boldsymbol{q}^*\end{bmatrix},\ \tilde{\Sigma}_x = \begin{bmatrix}\Sigma_x & P \\ P^* & \Sigma_x^*\end{bmatrix},$$

with

$$P = \mathbb{E}[\mathbf{x}\mathbf{x}^T],\ \boldsymbol{q} = \mathbb{E}[\mathbf{x}\mathrm{y}].$$

Using the matrix inversion lemma and after a bit of algebra, we obtain

$$MSE_l - MSE_{wl} = (\boldsymbol{q}^* - P^*\Sigma_x^{-1}\boldsymbol{q})^H(\Sigma_x^* - P^*\Sigma_x^{-1}P)^{-1}(\boldsymbol{q}^* - P^*\Sigma_x^{-1}\boldsymbol{q}).$$

However,

$$\Sigma_x^* - P^*\Sigma_x^{-1}P$$

is the Schur compliment of $\tilde{\Sigma}_x$ and since we know that is positive definite, the Schur compliment is also positive definite (as well as Σ_x). This is a well known result from linear algebra. Thus $MSE_l - MSE_{wl} \geq 0$. Note that under circularity conditions, $P = 0$ and $\boldsymbol{q} = 0$, leading to

$$MSE_l = MSE_{wl}.$$

4.10. Show that under the second order circularity assumption, the conditions in (4.39) hold true.

Solution: By the second order circularity condition we have

$$\mathbb{E}[\mathbf{x}\mathbf{x}^T] = 0,$$

or

$$\mathbb{E}[(\mathbf{x}_r + j\mathbf{x}_i)(\mathbf{x}_r^T + j\mathbf{x}_i^T)] = 0,$$

or

$$\mathbb{E}[\mathbf{x}_r\mathbf{x}_r^T] - \mathbb{E}[\mathbf{x}_i\mathbf{x}_i^T] = 0 \Rightarrow \varSigma_r = \varSigma_i,$$

and

$$\mathbb{E}[\mathbf{x}_r\mathbf{x}_i^T] + \mathbb{E}[\mathbf{x}_i\mathbf{x}_r^T] = 0 \Rightarrow \varSigma_{ri} = -\varSigma_{ir}.$$

The rest are shown in a similar way.

4.11. Show that if

$$f : \mathbb{C} \longrightarrow \mathbb{R},$$

then the Cauchy-Riemann conditions are violated.

Proof: Let

$$f(x + jy) = u(x, y) \in \mathbb{R}.$$

Then by assumption, the imaginary part $v(x, y)$ is identically zero. Hence the Cauchy-Riemann conditions, i.e.,

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}, \ \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x},$$

are violated, since this would only hold true if

$$\frac{\partial u}{\partial x} = \frac{\partial u}{\partial y} = 0,$$

which is the trivial case of a constant function.

4.12. Derive the optimality condition in (4.45).

Solution: We will show that any other filter, $h_i, \ i \in \mathbb{Z}$, results in a larger MSE compared to the filter $w_i, \ i \in \mathbb{Z}$, which satisfies the condition. Indeed, we have that

$$A := \mathbb{E}\left[(d_n - \sum_i h_i u_{n-i})^2\right] = \mathbb{E}\left[(d_n - \sum_i (h_i - w_i + w_i) u_{n-i})^2\right].$$

After expanding the term in squares, we obtain that

$$\begin{aligned} A \quad = \quad & \mathbb{E}\left[(d_n - \sum_i w_i u_{n-i})^2\right] + \mathbb{E}\left[\big(\sum_i (h_i - w_i) u_{n-i}\big)^2\right] - \\ & 2\sum_j (h_j - w_j)\, \mathbb{E}\left[(d_n - \sum_i w_i u_{n-i}) u_{n-j}\right]. \end{aligned}$$

However, by assumption the last term is zero. Hence, the error associated with h_i is always larger than that of the filter associated with the optimality condition.

4.13. Show Equations (4.50) and (4.51).

Solution:
a) Eq. (4.51): We know from Chapter 2 that the power spectrum of the output process, $\mathrm{y}(n,m)$ is related to that of the input, $\mathrm{d}(n,m)$, by

$$S_y(\omega_1,\omega_2) = |H(\omega_1,\omega_2)|^2 S_d(\omega_1,\omega_2).$$

Let now, $\mathrm{u}(n,m) = \mathrm{y}(n,m) + \eta(n,m)$, where $\eta(n,m)$ is independent of $\mathrm{y}(n,m)$. The autocorrelation of u is now given as

$$r_u(i,j) = \mathbb{E}\left[\big(\mathrm{y}(n,m)+\eta(n,m)\big)\big(\mathrm{y}(n-i,m-i)+\eta(n-i,m-j)\big)\right] = r_y(i,j)+r_\eta(i,j),$$

and taking the Fourier transform the claim is proved.

b) Eq. (4.50): By the respective definition we have that,

$$r_{du}(k,l) = \mathbb{E}\left[\mathrm{d}(n,m)\mathrm{u}(n-k,m-l)\right],$$

and also

$$\mathrm{u}(n,m) = \sum_{i=-\infty}^{+\infty}\sum_{j=-\infty}^{+\infty} h(i,j)\mathrm{d}(n-i,m-j).$$

Combining the previous two equations, we obtain

$$r_{du}(k,l) = \sum_{i=-\infty}^{+\infty}\sum_{j=-\infty}^{+\infty} h(i,j)\,\mathbb{E}\left[\mathrm{d}(n,m)\mathrm{d}(n-i-k,m-j-l)\right],$$

or

$$r_{du}(k,l) = \sum_{i=-\infty}^{+\infty}\sum_{j=-\infty}^{+\infty} h(i,j)r_d(i+k,j+l).$$

Setting

$$i' = i+k, \quad j' = j+l,$$

we obtain

$$r_{du}(k,l) = \sum_{i'=-\infty}^{+\infty}\sum_{j'=-\infty}^{+\infty} r_d(i',j')h(i'-k,j'-l)$$

or

$$r_{du}(k,l) = \sum_{i'=-\infty}^{+\infty}\sum_{j'=-\infty}^{+\infty} r_d(i',j')h(-(k-i'),-(l-j')).$$

Taking into account that if H is the Fourier transform of $h(n,m)$ then the sequence $h(-n,-m)$ has as Fourier transform H^*, then we finally obtain that

$$S_{du}(\omega_1,\omega_2) = H^*(\omega_1,\omega_2)S_d(\omega_1,\omega_2).$$

4.14. Derive the normal equations for Example 4.2.

Solution: We have

$$\begin{aligned}
\mathbb{E}[\mathrm{u}_n\mathrm{u}_n] &= \mathbb{E}\left[\big(0.5\mathrm{s}_n + \mathrm{s}_{n-1} + \eta_n\big)\big(0.5\mathrm{s}_n + \mathrm{s}_{n-1}\right. \\
&\quad \left. +\eta_n\big)\right] \\
&= 0.25r_s(0) + r_s(0) + r_\eta(0) \\
&= 1.25\sigma_s^2 + \sigma_\eta^2, \\
\mathbb{E}[\mathrm{u}_n\mathrm{u}_{n-1}] &= \mathbb{E}\left[\big(0.5\mathrm{s}_n + \mathrm{s}_{n-1} + \eta_n\big)\big(0.5\mathrm{s}_{n-1} + \mathrm{s}_{n-2}\right. \\
&\quad \left. +\eta_{n-1}\big)\right] \\
&= 0.5\sigma_s^2, \\
\mathbb{E}\left[\mathrm{u}_n\mathrm{u}_{n-2}\right] &= \mathbb{E}[\big(0.5\mathrm{s}_n + \mathrm{s}_{n-1} + \eta_n\big)\big(0.5\mathrm{s}_{n-2} + \mathrm{s}_{n-3} \\
&\quad +\eta_{n-2}\big)] \\
&= 0.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathbb{E}[\mathrm{u}_n\mathrm{d}_n] &= \mathbb{E}\left[\mathrm{u}_n\mathrm{s}_{n-1}\right] \\
&= \mathbb{E}\left[\,(0.5\mathrm{s}_n + \mathrm{s}_{n-1} + \eta_n)\,\mathrm{s}_{n-1}\right] = \sigma_s^2, \\
\mathbb{E}\left[\mathrm{u}_{n-1}\mathrm{d}_n\right] &= \mathbb{E}\left[\mathrm{u}_{n-1}\mathrm{s}_{n-1}\right] \\
&= \mathbb{E}\left[\,(0.5\mathrm{s}_{n-1} + \mathrm{s}_{n-2} + \eta_{n-1})\,\mathrm{s}_{n-1}\right] = 0.5\sigma_s^2, \\
\mathbb{E}\left[\mathrm{u}_{n-2}\mathrm{d}_n\right] &= \mathbb{E}\left[\mathrm{u}_{n-2}\mathrm{s}_{n-1}\right] = 0.
\end{aligned}$$

4.15. The input to the channel is a white noise sequence s_n of variance σ_s^2. The output of the channel is the AR processes

$$\mathrm{y}_n = a_1\mathrm{y}_{n-1} + \mathrm{s}_n. \tag{5}$$

The channel also adds white noise η_n of variance σ_η^2. Design an optimal equalizer of order two, which at its output recovers an approximation of s_{n-L}. Sometimes, this equalization task is also known as whitening, since in this case the action of the equalizer is to "whiten" the AR process.

Solution: The input to the equalizer is

$$\mathrm{u}_n = \mathrm{y}_n + \eta_n.$$

The desired response is

$$\mathrm{d}_n = \mathrm{s}_{n-L}.$$

Thus, the elements of the input covariance/autocorrelation matrix are given by

- $$\begin{aligned} r_u(0) =& \mathbb{E}[\mathrm{u}_n\mathrm{u}_n] &&= \mathbb{E}[(\mathrm{y}_n + \eta_n)(\mathrm{y}_n + \eta_n)] \\ =& r_y(0) + \sigma_\eta^2 &&= \frac{\sigma_s^2}{1-a_1^2} + \sigma_\eta^2 \end{aligned}$$
- $$\begin{aligned} r_u(1) =& \mathbb{E}[(\mathrm{y}_n + \eta_n)(\mathrm{y}_{n-1} + \eta_{n-1})] \\ =& r_y(1) = \frac{a_1\sigma_s^2}{1-\alpha_1^2}. \end{aligned}$$

  For the cross-correlation, we have
- $p(0) := \mathbb{E}[\mathrm{u}_n\mathrm{s}_{n-L}] = \mathbb{E}[\big(\mathrm{y}_n + \eta_n\big)\mathrm{s}_{n-L}] = \mathbb{E}[\mathrm{y}_n\mathrm{s}_{n-L}]$.

  Applying the recursion in (5), we have that
- $$\begin{aligned} p(0) :=& \mathbb{E}[\big(\alpha_1^L\mathrm{y}_{n-L} + \sum_{j=0}^{L-1}\alpha_1^j\mathrm{s}_{n-j}\big)\mathrm{s}_{n-L}] \\ =& \alpha_1^L\,\mathbb{E}[\mathrm{y}_{n-L}\mathrm{s}_{n-L}],\ L \geq 1. \end{aligned}$$

  Due to stationarity, it suffices to compute the following:
- $\mathbb{E}[\mathrm{y}_{n-L}\mathrm{s}_{n-L}] = \mathbb{E}[\mathrm{y}_n\mathrm{s}_n] = \mathbb{E}[\big(a_1\mathrm{y}_{n-1} + \mathrm{s}_n\big)\mathrm{s}_n] = \sigma_s^2$. Hence,

  $$p(0) = a_1^L\sigma_s^2.$$

  Note that the last equation is also valid for $L = 0$, as a direct consequence of (5).
- $$\begin{aligned} p(1) :=& \mathbb{E}[\mathrm{u}_{n-1}\mathrm{s}_{n-L}] \\ =& \mathbb{E}[\big(\alpha_1^{L-1}\mathrm{y}_{n-L} + \sum_{j=0}^{L-2}\alpha^j\mathrm{s}_{n-1-j}\big)\mathrm{s}(n-L)] = \alpha_1^{L-1}\sigma_s^2,\ L \geq 2. \end{aligned}$$

The latter one is also valid for $L = 1$. However, for $L = 0$ it is readily seen from (5), that $p(1) = 0$. Hence the equalizer is obtained via the following normal equations

$$\begin{bmatrix} \frac{\sigma_s^2}{1-a_1^2} + \sigma_\eta^2 & \frac{a_1\sigma_s^2}{1-a_1^2} \\ \frac{a_1\sigma_s^2}{1-a_1^2} & \frac{\sigma_s^2}{1-a_1^2} + \sigma_\eta^2 \end{bmatrix} \begin{bmatrix} w_*(0) \\ w_*(1) \end{bmatrix} = \begin{bmatrix} \alpha_1^L\sigma_s^2 \\ \alpha_1^{L-1}\sigma_s^2 \end{bmatrix}$$

4.16. Show that the forward and backward MSE optimal predictors are conjugate reverse of each other.

Solution: By the respective definitions we have

$$\boldsymbol{a}_m = \Sigma_m^{-1}\boldsymbol{r}_m^*$$

and

$$\boldsymbol{b}_m = \Sigma_m^{-1}J_m\boldsymbol{r}_m.$$

Since Σ_m is a Hermitian Toeplitz matrix, it is easily checked out that

$$J_m \Sigma_m J_m = \Sigma_m^*$$

or

$$J_m \Sigma_m^{-1} J_m = (\Sigma_m^{-1})^*.$$

Hence, since $J_m J_m = I_m$,

$$\begin{aligned} \boldsymbol{b}_m = & \quad J_m J_m \Sigma_m^{-1} J_m \boldsymbol{r}_m = J_m (\Sigma_m^{-1})^* \boldsymbol{r}_m \text{ or} \\ & \boldsymbol{b}_m^* = J_m \boldsymbol{a}_m. \end{aligned}$$

Hence for the minimum MSE backward error we have

$$\begin{aligned} \alpha_m^b &= r(0) - \boldsymbol{r}_m^H J_m \Sigma_m^{-1} J_m \boldsymbol{r}_m = r(0) - \boldsymbol{r}_m^H (\Sigma_m^{-1})^* \boldsymbol{r}_m \\ &= r(0) - \boldsymbol{r}_m^T \Sigma_m^{-1} \boldsymbol{r}_m^* = \alpha_m^f. \end{aligned}$$

4.17. Show that the MSE prediction errors ($\alpha_m^f = \alpha_m^b$) are updated according to the recursion

$$\alpha_m^b = \alpha_{m-1}^b (1 - |\kappa_{m-1}|^2).$$

Solution: By the respective definition we have

$$\begin{aligned} \alpha_m^b &= r(0) - \boldsymbol{r}_m^H J_m \Sigma_m^{-1} J_m \boldsymbol{r}_m = r(0) - \boldsymbol{r}_m^H J_m \boldsymbol{b}_m = r(0) - \boldsymbol{r}_m^H \boldsymbol{a}_m^* \\ &= r(0) - \begin{bmatrix} \boldsymbol{r}_{m-1}^H & r^*(m) \end{bmatrix} \left\{ \begin{bmatrix} \boldsymbol{a}_{m-1}^* \\ 0 \end{bmatrix} + \begin{bmatrix} -\boldsymbol{b}_{m-1}^* \\ 1 \end{bmatrix} \kappa_{m-1}^* \right\} \\ &= r(0) - \Big(\boldsymbol{r}_{m-1}^H \boldsymbol{a}_{m-1}^* + \big(r^*(m) - \boldsymbol{r}_{m-1}^H \boldsymbol{b}_{m-1}^* \big) \kappa_{m-1}^* \Big) \\ &= r(0) - \boldsymbol{r}_{m-1}^H \boldsymbol{a}_{m-1}^* - |\kappa_{m-1}|^2 \alpha_{m-1}^b \\ &= r(0) - \boldsymbol{r}_{m-1}^H J \boldsymbol{b}_{m-1} - |\kappa_{m-1}|^2 \alpha_{m-1}^b \\ &= \alpha_{m-1}^b (1 - |\kappa_{m-1}|^2). \end{aligned}$$

4.18. Derive the BLUE in the Gauss-Markov theorem.

Solution: The optimization task is

$$\begin{aligned} H_* \quad &:= \quad \arg\min\nolimits_H \operatorname{trace}\{H \Sigma_\eta H^T\}, \\ \text{s.t.} \quad & \qquad H X = I. \end{aligned}$$

Let

$$H := \begin{bmatrix} \boldsymbol{h}_1^T \\ \boldsymbol{h}_2^T \\ \vdots \\ \boldsymbol{h}_l^T \end{bmatrix}, \ X = \begin{bmatrix} \boldsymbol{x}_1, \ \boldsymbol{x}_2, \cdots, \boldsymbol{x}_l \end{bmatrix}.$$

Observe that

$$\text{trace}\{H\Sigma_\eta H^T\} = \sum_{i=1}^{l} \boldsymbol{h}_i^T \Sigma_\eta \boldsymbol{h}_i. \tag{6}$$

Also, $HX = I$ is equivalent with

$$\boldsymbol{h}_i^T \boldsymbol{x}_i = 1, \tag{7}$$

and

$$\boldsymbol{h}_i^T \boldsymbol{x}_j = 0,\ j \neq i. \tag{8}$$

Since Σ_η is positive definite, our optimization task is equivalent with l constrained minimization problems, one for each $\boldsymbol{h}_i,\ i = 1, 2, ..., l$. Using Lagrange multipliers, the Lagrangian for each one of the tasks is written as

$$\min_{\boldsymbol{h}_i}\ \boldsymbol{h}_i^T \Sigma_\eta \boldsymbol{h}_i - \lambda_i^{(i)}(\boldsymbol{h}_i^T \boldsymbol{x}_i - 1) - \sum_{j\neq i}^{l} \lambda_j^{(i)} \boldsymbol{h}_i^T \boldsymbol{x}_j.$$

Taking the gradient we obtain,

$$2\Sigma_\eta \boldsymbol{h}_i - \sum_{j=1}^{l} \lambda_j^{(i)} \boldsymbol{x}_j = \boldsymbol{0} \Rightarrow$$

$$\Sigma_\eta \boldsymbol{h}_i = \frac{1}{2} X \boldsymbol{\lambda}^{(i)} := \frac{1}{2} X \begin{bmatrix} \lambda_1^{(i)} \\ \lambda_2^{(i)} \\ \vdots \\ \lambda_l^{(i)} \end{bmatrix},$$

or

$$\begin{aligned} \boldsymbol{h}_i &= \frac{1}{2}\Sigma_\eta^{-1} X \boldsymbol{\lambda}^{(i)} \Rightarrow \\ \boldsymbol{h}_i^T &= \frac{1}{2}{\boldsymbol{\lambda}^{(i)}}^T X^T \Sigma_\eta^{-1},\ i = 1, 2, .., l, \end{aligned}$$

or

$$H = \Lambda X^T \Sigma_\eta^{-1},$$

where

$$\Lambda = \frac{1}{2} \begin{bmatrix} {\boldsymbol{\lambda}^{(1)}}^T \\ {\boldsymbol{\lambda}^{(2)}}^T \\ \vdots \\ {\boldsymbol{\lambda}^{(l)}}^T \end{bmatrix},$$

is computed from the constraints, i.e.,

$$\begin{aligned} \Lambda X^T \Sigma_\eta^{-1} X = I \quad &\Rightarrow \\ \Lambda = (X^T \Sigma_\eta^{-1} X)^{-1} & \end{aligned}$$

or

$$H_* = (X^T \Sigma_\eta^{-1} X)^{-1} X^T \Sigma_\eta^{-1}.$$

4.19. Show that the mean-square error (which in this case coincides with the variance of the estimator) of any linear unbiased estimator is higher than that associated with the BLUE.

Solution: The mean-square error of any H is given by

$$MSE(H) = \text{trace}\{H\Sigma_\eta H^T\}.$$

However,

$$(H - H_*)\Sigma_\eta(H - H_*)^T = H\Sigma_\eta H^T - H\Sigma_\eta H_*^T - H_*\Sigma_\eta H^T + H_*\Sigma_\eta H_*^T,$$

or

$$H\Sigma_\eta H^T = (H - H_*)\Sigma_\eta(H - H_*)^T + H\Sigma_\eta H_*^T + H_*\Sigma_\eta H^T - H_*\Sigma_\eta H_*^T.$$

Recall that

$$H_* = (X^T\Sigma_\eta^{-1}X)^{-1}X^T\Sigma_\eta^{-1}$$

and also

$$HX = H_*X = I.$$

Hence,

$$H\Sigma_\eta H_*^T = H\Sigma_\eta\Sigma_\eta^{-1}X(X^T\Sigma_\eta^{-1}X)^{-1} = (X^T\Sigma_\eta^{-1}X)^{-1}.$$

Similarly,

$$H_*\Sigma_\eta H^T = (X^T\Sigma_\eta^{-1}X)^{-1}$$

and

$$\begin{aligned} H_*\Sigma_\eta H_*^T &= (X^T\Sigma_\eta^{-1}X)^{-1}X^T\Sigma_\eta^{-1}\Sigma_\eta\Sigma_\eta^{-1}X(X^T\Sigma_\eta^{-1}X)^{-1} \\ &= (X^T\Sigma_\eta^{-1}X)^{-1}. \end{aligned}$$

Thus,

$$H\Sigma_\eta H^T = (H - H_*)\Sigma_\eta(H - H_*)^T + H_*\Sigma_\eta H_*^T.$$

Hence,

$$\begin{aligned} \text{trace}\{H\Sigma_\eta H^T\} \quad &= \quad \sum_i \boldsymbol{h}_i^T\Sigma_\eta\boldsymbol{h}_i = \\ \sum_i(\boldsymbol{h}_i^T - \boldsymbol{h}_{*,i}^T)\Sigma_\eta(\boldsymbol{h}_i - \boldsymbol{h}_{*,i}) \quad &+ \quad \text{trace}\{H_*\Sigma_\eta H_*^T\}. \end{aligned}$$

Since Σ_η is positive definite, all terms are positive and the minimum occurs if $\boldsymbol{h}_i = \boldsymbol{h}_{*,i}$. Moreover, the minimum MSE is

$$MSE(H_*) = \text{trace}\{H_*\Sigma_\eta H_*^T\}.$$

However,

$$H_*\Sigma_\eta H_*^T = (X^T\Sigma_\eta^{-1}X)^{-1}$$

or

$$MSE(H_*) = \text{trace}\{(X^T\Sigma_\eta^{-1}X)^{-1}\}.$$

4.20. Show that if Σ_η is positive definite, then $X^T \Sigma_\eta^{-1} X$ is also positive definite if X is full rank.

Solution: Recall from linear algebra that if X is full rank, then $X^T X$ is positive definite and vice versa. Indeed, assume that this is not the case. Then, there will be $\boldsymbol{a} \neq \mathbf{0}$, such that

$$X^T X \boldsymbol{a} = \mathbf{0},$$

or

$$||X\boldsymbol{a}||^2 = \boldsymbol{\alpha}^T X^T X \boldsymbol{a} = 0 \Leftrightarrow X\boldsymbol{a} = \mathbf{0},$$

which contradicts the assumption that X is full rank.

For our case, we know that Σ_η is symmetric and positive definite. Thus,

$$\Sigma_\eta = U \Lambda U^T,$$

where Λ is the diagonal matrix with the eigenvalues of Σ_η and U the unitary matrix with the orthonormal eigenvectors as its columns. Hence,

$$X^T \Sigma_\eta^{-1} X = X (U \Lambda U^T)^{-1} X = X U \Lambda^{-1} U^T X = B^T B,$$

where

$$B = \Lambda^{-1/2} U^T X.$$

Note that if X is full rank, then B is also full rank, hence $B^T B$ is positive definite.

4.21. Derive a MSE optimal linearly constrained widely linear beamformer.

Solution: The output of the widely linear beamformer is given by

$$\begin{aligned} \hat{s}(t) &= \boldsymbol{w}^H \mathbf{u}(t) + \boldsymbol{v}^H \mathbf{u}^*(t) \\ &:= \tilde{\boldsymbol{w}}_e^H(t) \tilde{\mathbf{u}}_e(t), \end{aligned}$$

where

$$\tilde{\mathbf{u}}_e(t) = \begin{bmatrix} \mathbf{u}(t) \\ \mathbf{u}^*(t) \end{bmatrix}, \ \tilde{\boldsymbol{w}}_e = \begin{bmatrix} \boldsymbol{w} \\ \boldsymbol{v} \end{bmatrix}.$$

The constraint must guarantee the distortionless response for signals impinging on the array from the direction ϕ, associated with the steering vector $\boldsymbol{x}$. That is,

$$\hat{\mathrm{s}}(t) = \boldsymbol{w}^H \boldsymbol{x} \mathrm{s}(t) + \boldsymbol{v}^H \boldsymbol{x}^* \mathrm{s}^*(t) = \mathrm{s}(t).$$

Thus,

$$\boldsymbol{w}^H \boldsymbol{x} = 1$$

and

$$\boldsymbol{v}^H \boldsymbol{x}^* = 0.$$

The error signal is

$$\begin{aligned}\mathrm{s}(t) - \hat{\mathrm{s}}(t) &= s(t) - \boldsymbol{w}^H \boldsymbol{x}\mathrm{s}(t) - \boldsymbol{w}^H \boldsymbol{\eta}(t) - \boldsymbol{v}^H \boldsymbol{x}^* \mathrm{s}^*(t) - \boldsymbol{v}^H \boldsymbol{\eta}^*(t) \\ &= -\boldsymbol{w}^H \boldsymbol{\eta}(t) - \boldsymbol{v}^H \boldsymbol{\eta}^*(t) = -\tilde{\boldsymbol{w}}_e^H \tilde{\boldsymbol{\eta}}(t),\end{aligned}$$

where

$$\tilde{\boldsymbol{\eta}}(t) = \begin{bmatrix} \boldsymbol{\eta}(t) \\ \boldsymbol{\eta}^*(t) \end{bmatrix},$$

and the constraints have been taken into account. Thus, the error variance

$$\tilde{\boldsymbol{w}}_e^H \Sigma_{\tilde{\eta}} \tilde{\boldsymbol{w}}_e$$

must be minimized, where

$$\Sigma_{\tilde{\eta}} = \begin{bmatrix} \Sigma_\eta & P_\eta \\ P_\eta^* & \Sigma_\eta^* \end{bmatrix} = \mathbb{E} \begin{bmatrix} \boldsymbol{\eta}(t)\boldsymbol{\eta}^H(t) & \boldsymbol{\eta}(t)\boldsymbol{\eta}^T(t) \\ \boldsymbol{\eta}^*(t)\boldsymbol{\eta}^H(t) & \boldsymbol{\eta}^*(t)\boldsymbol{\eta}^T(t) \end{bmatrix}.$$

Let us write the two constraints in a more compact form to facilitate the optimization. To this end, we have that

$$X^H \tilde{\boldsymbol{w}}_e = \begin{bmatrix} 1 \\ 0 \end{bmatrix},$$

where

$$X := \begin{bmatrix} \boldsymbol{x} & 0 \\ 0 & \boldsymbol{x}^* \end{bmatrix} = \begin{bmatrix} \tilde{\boldsymbol{x}}_1 \; \tilde{\boldsymbol{x}}_2 \end{bmatrix}, \; \tilde{\boldsymbol{x}}_1 := \begin{bmatrix} \boldsymbol{x} \\ 0 \end{bmatrix}, \; \tilde{\boldsymbol{x}}_2 := \begin{bmatrix} 0 \\ \boldsymbol{x}^* \end{bmatrix}.$$

Thus, our task now becomes

$$\begin{aligned} \text{minimize w.r.t. } \tilde{\boldsymbol{w}}_e \quad & \tilde{\boldsymbol{w}}_e^H \Sigma_{\tilde{\eta}} \tilde{\boldsymbol{w}}_e \\ \text{s.t.} \quad & \tilde{\boldsymbol{w}}_e^H X = [1, 0], \end{aligned}$$

or using the Lagrange multipliers,

$$L(\tilde{\boldsymbol{w}}_e) := \tilde{\boldsymbol{w}}_e^H \Sigma_{\tilde{\eta}} \tilde{\boldsymbol{w}}_e - \lambda_1(\tilde{\boldsymbol{w}}_e^H \tilde{\boldsymbol{x}}_1 - 1) - \lambda_2 \tilde{\boldsymbol{w}}_e^H \tilde{\boldsymbol{x}}_2.$$

Minimization of the Lagrangian results in

$$\Sigma_{\tilde{\eta}} \tilde{\boldsymbol{w}}_e = \frac{1}{2} X \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix}.$$

Hence, plugging the previous into the constraint, we get

$$\begin{aligned} \tilde{\boldsymbol{w}}_e^H X = [1, 0] \Rightarrow X^H \tilde{\boldsymbol{w}}_e &= \begin{bmatrix} 1 \\ 0 \end{bmatrix} \Rightarrow \\ X^H \Sigma_{\tilde{\eta}}^{-1} X \begin{bmatrix} \lambda_1/2 \\ \lambda_2/2 \end{bmatrix} &= \begin{bmatrix} 1 \\ 0 \end{bmatrix} \end{aligned}$$

or

$$\frac{1}{2} \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} = (X^H \Sigma_{\tilde{\eta}}^{-1} X)^{-1} \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

and finally

$$\tilde{\boldsymbol{w}}_e = \Sigma_{\tilde{\eta}}^{-1} X (X^H \Sigma_{\tilde{\eta}}^{-1} X)^{-1} \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

Note that, if the noise vector is circular, i.e., $P_\eta = 0$, the results become identical to the linear beamformer.

4.22. Prove that the Kalman gain that minimizes the error variance matrix

$$P_{n|n} = \mathbb{E}[(\mathbf{x}_n - \hat{\mathbf{x}}_{n|n})(\mathbf{x}_n - \hat{\mathbf{x}}_{n|n})^T],$$

is given by

$$K_n = P_{n|n-1} H_n^H (R_n + H_n P_{n|n-1} H_n^T)^{-1}.$$

Hint: Use the following formulas

$$\begin{aligned} \frac{\partial \operatorname{trace}\{AB\}}{\partial A} &= B^T \ (AB \text{ a square matrix}) \\ \frac{\partial \operatorname{trace}\{ACA^T\}}{\partial A} &= 2AC, \ (C = C^T). \end{aligned}$$

Solution: We know that

$$\hat{\mathbf{x}}_{n|n} = \hat{\mathbf{x}}_{n|n-1} + K_n(\mathbf{y}_n - H_n \hat{\mathbf{x}}_{n|n-1}).$$

Thus,

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

where

$$\mathbf{e}_n = \mathbf{y}_n - H_n \hat{\mathbf{x}}_{n|n-1}.$$

Hence,

$$P_{n|n} = P_{n|n-1} - K_n \, \mathbb{E}[\mathbf{e}_n \mathbf{e}_{n|n-1}^T] - \mathbb{E}[\mathbf{e}_{n|n-1} \mathbf{e}_n^T] K_n^T + K_n \, \mathbb{E}[\mathbf{e}_n \mathbf{e}_n^T] K_n^T.$$

However,

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

Similarly,

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

Moreover,

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

Hence, we have that

$$P_{n|n} = P_{n|n-1} - K_n H_n P_{n|n-1} - P_{n|n-1} H_n^T K_n^T + K_n (R_n + H_n P_{n|n-1} H_n^T) K_n^T, \tag{9}$$

or

$$\begin{aligned} \text{trace}\{P_{n|n}\} &= \text{trace}\{P_{n|n-1}\} - 2\,\text{trace}\{K_n H_n P_{n|n-1}\} + \\ &\quad \text{trace}\left\{K_n (R_n + H_n P_{n|n-1} H_n^T) K_n^T\right\}. \end{aligned} \tag{10}$$

Taking the derivative (gradient) with respect to K_n, we obtain

$$\frac{\partial\,\text{trace}\{P_{n|n}\}}{\partial K_n} = -2 P_{n|n-1} H_n^T + 2 K_n (R_n + H_n P_{n|n-1} H_n^T) = 0$$

or

$$K_n = P_{n|n-1} H_n^T (R_n + H_n P_{n|n-1} H_n^T)^{-1}.$$

4.23. Show that in Kalman filtering, the prior and posterior error covariance matrices are related as

$$P_{n|n} = P_{n|n-1} - K_n H_n P_{n|n-1}.$$

Solution: Recall from the solution of the Problem 4.22 that

$$P_{n|n} = P_{n|n-1} - K_n H_n P_{n|n-1} - P_{n|n-1} H_n^T K_n^T + K_n (R_n + H_n P_{n|n-1} H_n^T) K_n^T.$$

Plug in the above the optimum Kalman gain

$$K_n = P_{n|n-1} H_n^T (R_n + H_n P_{n|n-1} H_n^T)^{-1},$$

which results in the desired update.

4.24. Derive the Kalman algorithm in terms of the inverse state-error covariance matrices, $P_{n|n}^{-1}$. In statistics, the inverse error covariance matrix is related to Fisher's information matrix, hence the name of the scheme.

Solution: To build the Kalman algorithm around the inverse state-error covariance matrices $P_{n|n}^{-1}$, $P_{n|n-1}^{-1}$, we need to apply the following matrix inversion Lemmas,

$$\begin{aligned} PB^T (R + BPB^T)^{-1} &= (P^{-1} + B^T R^{-1} B)^{-1} B^T R^{-1} \\ (A + BD^{-1}C)^{-1} &= A^{-1} - A^{-1} B (D + CA^{-1} B)^{-1} CA^{-1}. \end{aligned} \tag{11}$$

According to the text we have:

- $S_n^{-1} = (R_n + H_n P_{n|n-1} H_n^T)^{-1} = R_n^{-1} - R_n^{-1} H_n (P_{n|n-1}^{-1} + H_n^T R_n^{-1} H_n)^{-1} H_n^T R_n^{-1}$,
- $K_n = P_{n|n-1} H_n^T (R_n + H_n P_{n|n-1} H_n^T)^{-1} = (P_{n|n-1}^{-1} + H_n^T R_n^{-1} H_n)^{-1} H_n^T R_n^{-1}$,
- $P_{n|n} = P_{n|n-1} - P_{n|n-1} H_n^T (R_n + H_n P_{n|n-1} H_n^T)^{-1} H_n P_{n|n-1}$,

which gives

- $P_{n|n}^{-1} = (P_{n|n-1}^{-1} + H_n^T R_n^{-1} H_n)^{-1}$,
- $P_{n+1|n} = F_{n+1} P_{n|n} F_{n+1}^T + Q_{n+1}$,

and the last one results in

$$P_{n+1|n}^{-1} = Q_{n+1}^{-1} - Q_{n+1}^{-1} F_{n+1} (P_{n|n}^{-1} + F_{n+1}^T Q_{n+1}^{-1} F_{n+1})^{-1} F_{n+1}^T Q_{n+1}^{-1}.$$

Solutions To Problems of Chapter 5

5.1. Show that the gradient vector is perpendicular to the tangent at a point of an isovalue curve.

Solution: The differential of the cost function, $J(\boldsymbol{\theta})$, at a point $\boldsymbol{\theta}^{(i)}$, is given by

$$\begin{aligned} dJ(\boldsymbol{\theta}^{(i)}) &= \frac{\partial J(\boldsymbol{\theta}^{(i)})}{\partial \theta_1^{(i)}} d\theta_1^{(i)} + \ldots + \frac{\partial J(\boldsymbol{\theta}^{(i)})}{\partial \theta_l^{(i)}} d\theta_l^{(i)} \\ &= [d\theta_1^{(i)}, \ldots, d\theta_l^{(i)}]^T \nabla J(\boldsymbol{\theta}^{(i)}) = d\boldsymbol{\theta}^{(i)T} \nabla J(\boldsymbol{\theta}^{(i)}). \end{aligned}$$

However, moving along the isovalue curve crossing $\boldsymbol{\theta}^{(i)}$ makes the differential value zero. Hence, the gradient vector is perpendicular to $d\boldsymbol{\theta}^{(i)}$, whose direction defines the respective tangent plane.

5.2. Prove that if

$$\sum_{i=1}^{\infty} \mu_i^2 < \infty, \quad \sum_{i=1}^{\infty} \mu_i = \infty,$$

the steepest descent scheme, for the MSE loss function and for the iteration-dependent step size case, converges to the optimal solution.

Solution: The basic iteration update is given by

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

or

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

where $\boldsymbol{\theta}_*$ is the optimal point. Thus,

$$\begin{aligned} (\boldsymbol{\theta}^{(i)} - \boldsymbol{\theta}_*)^T(\boldsymbol{\theta}^{(i)} - \boldsymbol{\theta}_*) &= (\boldsymbol{\theta}^{(i-1)} - \boldsymbol{\theta}_*)^T(\boldsymbol{\theta}^{(i-1)} - \boldsymbol{\theta}_*) + \\ &\quad \mu_i^2 \nabla^T J(\boldsymbol{\theta}^{(i-1)}) \nabla J(\boldsymbol{\theta}^{(i-1)}) - \\ &\quad 2\mu_i (\boldsymbol{\theta}^{(i-1)} - \boldsymbol{\theta}_*)^T \nabla J(\boldsymbol{\theta}^{(i-1)}), \end{aligned}$$

and applying the above recursively, we get

$$\begin{aligned} (\boldsymbol{\theta}^{(i)} - \boldsymbol{\theta}_*)^T(\boldsymbol{\theta}^{(i)} - \boldsymbol{\theta}_*) &= (\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}_*)^T(\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}_*) + \\ &\quad \sum_{k=1}^{i} \mu_k^2 \nabla^T J(\boldsymbol{\theta}^{(k-1)}) \nabla J(\boldsymbol{\theta}^{(k-1)}) - \\ &\quad 2\sum_{k=1}^{i} \mu_k (\boldsymbol{\theta}^{(k-1)} - \boldsymbol{\theta}_*)^T \nabla J(\boldsymbol{\theta}^{(k-1)}). \end{aligned}$$

However,

$$\nabla^T J(\boldsymbol{\theta}^{(k-1)}) \nabla J(\boldsymbol{\theta}^{(k-1)}) \geq 0.$$

Also, since the MSE loss function is a convex one (hence the current proof carries on to any other convex loss function) we have that[1]

$$J(\boldsymbol{\theta}_*) \geq J(\boldsymbol{\theta}^{(k)}) + \nabla^T J(\boldsymbol{\theta}^{(k)})(\boldsymbol{\theta}_* - \boldsymbol{\theta}^{(k)}), \ \forall k,$$

from which we have that

$$(\boldsymbol{\theta}^{(k-1)} - \boldsymbol{\theta}_*)^T \nabla J(\boldsymbol{\theta}^{(k-1)}) \geq 0.$$

Hence, assuming that the gradient is *bounded* at every point, then we can write,

$$(\boldsymbol{\theta}^{(i)} - \boldsymbol{\theta}_*)^T(\boldsymbol{\theta}^{(i)} - \boldsymbol{\theta}_*) - (\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}_*)^T(\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}_*) \leq A\sum_{k=1}^{i} \mu_k^2 - 2\sum_{k=1}^{i} \mu_k (\boldsymbol{\theta}^{(k-1)} - \boldsymbol{\theta}_*)^T \nabla J(\boldsymbol{\theta}^{(k-1)}),$$

for some $A > 0$. However, the term on the left hand side is bounded from below, since the first term is always positive and if it diverges it will go to $+\infty$. Also, $\sum_{k=1}^{i} \mu_k^2$ tends to zero, by assumption. Thus, the only way, if we assume that $\sum_{k=1}^{i} \mu_k$ diverges, that the right hand side does not diverge to $-\infty$, is when

$$\boldsymbol{\theta}^{(i)} \longrightarrow \boldsymbol{\theta}_*.$$

Otherwise the left hand side becomes unbounded form below, which is a contradiction. Hence the claim has been proved.

5.3. Derive the steepest gradient descent direction for the complex-valued case.

Solution: We know from the text that

$$\begin{aligned} J(\boldsymbol{\theta} + \Delta\boldsymbol{\theta}) &= J(\boldsymbol{\theta}_r, \boldsymbol{\theta}_i) + \Delta\boldsymbol{\theta}_r^T \nabla_r J(\boldsymbol{\theta}_r, \boldsymbol{\theta}_i) + \\ & \quad \Delta\boldsymbol{\theta}_i^T \nabla_i J(\boldsymbol{\theta}_r, \boldsymbol{\theta}_i). \end{aligned}$$

Simplifying a bit the notation and taking into account that the real and imaginary parts can be expressed in terms of the complex number and its conjugate we obtain,

$$\begin{aligned} J(\boldsymbol{\theta} + \Delta\boldsymbol{\theta}) &= J(\boldsymbol{\theta}) + \frac{1}{2}\{\Delta\boldsymbol{\theta}^T + \Delta\boldsymbol{\theta}^H\} \nabla_r J(\boldsymbol{\theta}) + \\ & \quad \frac{1}{2j}\{\Delta\boldsymbol{\theta}^T - \Delta\boldsymbol{\theta}^H\} \nabla_i J(\boldsymbol{\theta}), \end{aligned}$$

or

$$\begin{aligned} J(\boldsymbol{\theta} + \Delta\boldsymbol{\theta}) &= J(\boldsymbol{\theta}) + \frac{1}{2}\{\Delta\boldsymbol{\theta}^T (\nabla_r J(\boldsymbol{\theta}) - j\nabla_i J(\boldsymbol{\theta}))\} + \\ & \quad \frac{1}{2}\{\Delta\boldsymbol{\theta}^H (\nabla_r J(\boldsymbol{\theta}) + j\nabla_i J(\boldsymbol{\theta}))\}, \end{aligned}$$

[1] This is the definition of convexity, and it is valid for *any* points in place of $\boldsymbol{\theta}_*$ and $\boldsymbol{\theta}^{(k)}$.

and finally, taking into account that $J(\boldsymbol{\theta})$ is a real function,

$$J(\boldsymbol{\theta} + \Delta\boldsymbol{\theta}) = J(\boldsymbol{\theta}) + \frac{1}{2}\big\{\Delta\boldsymbol{\theta}^T(\nabla_{\boldsymbol{\theta}^*}J(\boldsymbol{\theta}))^* + \Delta\boldsymbol{\theta}^H\nabla_{\boldsymbol{\theta}^*}J(\boldsymbol{\theta})\big\},$$

which proves the claim, since the addition of a number with its complex conjugate gives its real part.

5.4. Let θ, x be two jointly distributed random variables. Let also the function (regression)

$$f(\theta) = \mathbb{E}[\mathrm{x}|\theta],$$

assumed to be an increasing one. Show that under the conditions in (5.29) the recursion

$$\theta_n = \theta_{n-1} - \mu_n x_n$$

converges in probability to the root of $f(\theta)$.

Solution: Let us divide x_n into two parts, i.e.,

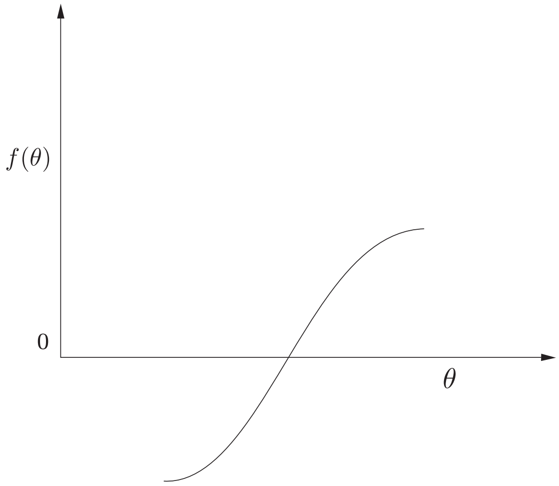

Figure 1: Figure for Problem 5.4.

$$\mathrm{x}_n = f(\theta_n) + \eta_n.$$

Then,

$$\mathbb{E}[\eta_n|\theta_n] = \mathbb{E}[\mathrm{x}_n|\theta_n] - f(\theta_n) = 0.$$

Moreover, we will assume that the respective variance is finite,

$$\mathbb{E}[\eta_n^2] = \sigma^2 < \infty,$$

and also we assume that $\upeta_n$ and $\uptheta_n$ are independent. Let θ_* be a root. Then, we form

$$\uptheta_n - \theta_* = \uptheta_{n-1} - \theta_* - \mu_n f(\uptheta_{n-1}) - \mu_n \upeta_n.$$

Taking the expectation of the square we get

$$\begin{aligned} \mathbb{E}[(\uptheta_n - \theta_*)^2] \;&=\; \mathbb{E}[(\uptheta_{n-1} - \theta_*)^2] + \mu_n^2\, \mathbb{E}[f^2(\uptheta_{n-1})] + \\ &\quad\; \mu_n^2\, \mathbb{E}[\upeta_n^2] - 2\mu_n\, \mathbb{E}[(\uptheta_n - \theta_*) f(\uptheta_{n-1})], \end{aligned}$$

and applying the previous one iteratively we obtain

$$\begin{aligned} \mathbb{E}[(\uptheta_n - \theta_*)^2] - \mathbb{E}[(\uptheta_0 - \theta_*)^2] \;&=\; \sum_{i=1}^{n} \mu_i^2 \left\{ \mathbb{E}[f^2(\uptheta_i)] + \mathbb{E}[\upeta_i^2] \right\} - \\ &\quad\; 2 \sum_{i=1}^{n} \mu_i\, \mathbb{E}[(\uptheta_i - \theta_*) f(\uptheta_i)]. \end{aligned}$$

Assume now that

$$\mathbb{E}[f^2(\uptheta_{n-1})] \le b < \infty.$$

Then we obtain the following,

$$\mathbb{E}[(\uptheta_n - \theta_*)^2] - \mathbb{E}[(\uptheta_0 - \theta_*)^2 \le (b + \sigma^2) \sum_{i=1}^{n} \mu_i^2 - \\ 2 \sum_{i=1}^{n} \mu_i\, \mathbb{E}[(\uptheta_i - \theta_*) f(\uptheta_i)]. \tag{1}$$

However, from Figure 1, it is readily seen that

$$\begin{aligned} f(\theta) > 0 &\ \text{ if } \ (\theta - \theta_*) > 0 \\ f(\theta) < 0 &\ \text{ if } \ (\theta - \theta_*) < 0 \\ f(\theta) = 0 &\ \text{ if } \ (\theta - \theta_*) = 0. \end{aligned}$$

As a matter of fact, the above are the crucial conditions for convergence. Note that even if a function is decreasing, one can consider its negative form and still find the root. Thus, these conditions are quite general, provided a single root exists.

Thus,

$$f(\theta)(\theta - \theta_*) \ge 0, \tag{2}$$

and hence,

$$\mathbb{E}[(\uptheta_i - \theta_*) f(\uptheta_i)] \ge 0.$$

If both conditions for μ_n hold true, then the first term on the right hand side of (1) tends to zero. Also, the left hand side is bounded from below.

The only way that the second term remains bounded from below, and having assumed that $\sum_i \mu_i$ diverges to ∞ is the following one to be true,

$$\lim_{n\longrightarrow\infty} \mathbb{E}[(\theta_{n-1} - \theta_*)f(\theta_n)] = 0.$$

However, since (2) is valid for all θ, the last condition is equivalent to

$$\lim_{n\longrightarrow\infty} \Pr\left\{\theta_n = \theta_*\right\} = 1,$$

which proves the claim.

5.5. Show that the LMS algorithm is a nonlinear estimator.

Solution: Consider the basic LMS recursion

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

Applying it recursively and assuming that we start from $\boldsymbol{\theta}_{-1} = \mathbf{0}$, we obtain

$$\boldsymbol{\theta}_n = \mu \sum_{i=0}^{n} \boldsymbol{x}_i e_i.$$

Note that e_i depends on $\boldsymbol{x}_i$, and $\boldsymbol{\theta}_{i-1}$ depends on $\boldsymbol{x}_{i-1}$, which shares common elements with $\boldsymbol{x}_i$, which are multiplied. Moreover, the output of the filter

$$\hat{y}_n = \boldsymbol{\theta}_n^T \boldsymbol{x}_n = \mu \sum_{i=0}^{n} e_i \boldsymbol{x}_i^T \boldsymbol{x}_n,$$

which also verifies the nonlinearity.

5.6. Show equation (5.42).

Solution: Our starting point is

$$\begin{aligned} \Sigma_{c,n} &= \Sigma_{c,n-1} - \mu \Sigma_x \Sigma_{c,n-1} - \mu \Sigma_{c,n-1} \Sigma_x \\ &+ 2\mu^2 \Sigma_x \Sigma_{c,n-1} \Sigma_x + \mu^2 \Sigma_x \text{trace}\{\Sigma_x \Sigma_{c,n-1}\} + \mu^2 \sigma_\eta^2 \Sigma_x \qquad (3) \end{aligned}$$

Take into account that $QQ^T = Q^T Q = I$ and also that $Q^T \text{trace}\{A\} Q = \text{trace}\{Q^T A Q\} I$, for any matrix A. Hence,

$$\begin{aligned} Q^T \Sigma_{c,n} Q &= Q^T \Sigma_{c,n-1} Q - \mu Q^T \Sigma_x Q Q^T \Sigma_{c,n-1} Q \\ &- \mu Q^T \Sigma_{c,n-1} Q Q^T \Sigma_x Q \\ &+ 2\mu^2 Q^T \Sigma_x Q Q^T \Sigma_{c,n-1} Q Q^T \Sigma_x Q \\ &+ \mu^2 Q^T \Sigma_x Q Q^T \text{trace}\{\Sigma_x Q Q^T \Sigma_{c,n-1}\} Q \\ &+ \mu^2 \sigma_\eta^2 Q^T \Sigma_x Q. \qquad (4) \end{aligned}$$

which then proves the claim.

5.7. Derive the bound in (5.45).
Hint: Use the well known property from linear algebra, that the eigenvalues of a matrix, $A \in \mathbb{R}^{l \times l}$, satisfy the following bound,

$$\max_{1 \leq i \leq l} |\lambda_i| \leq \max_{1 \leq i \leq l} \sum_{j=1}^{l} |a_{ij}| := \|A\|_1.$$

Solution: The elements (entries) of the matrix

$$A := (I - \mu\Lambda)^2 + \mu^2\Lambda^2 + \mu^2 \boldsymbol{\lambda}\boldsymbol{\lambda}^T$$

are:

$$A_{ij} = \begin{cases} (1 - \mu\lambda_i)^2 + \mu^2\lambda_i^2 + \mu^2\lambda_i^2 \ , & i = j \\ \mu^2 \lambda_i \lambda_j \ , & i \neq j. \end{cases}$$

Note that all the entries are positive, since Σ_x is positive definite and has positive eigenvalues. Hence, it is required that

$$\max_i \left\{ (1 - \mu\lambda_i)^2 + \mu^2\lambda_i^2 + \mu^2\lambda_i \sum_{j=1}^{l} \lambda_j \right\} < 1.$$

This is a sufficient condition for stability and guarantees that all eigenvalues of A have magnitude less than one, or

$$1 - 2\mu\lambda_i + 2\mu^2\lambda_i^2 + \mu^2\lambda_i \sum_{j=1}^{l} \lambda_j < 1 \ \text{ or}$$

$$\mu \left(2\lambda_i + \sum_{j=1}^{l} \lambda_j \right) < 2,$$

or,

$$\begin{aligned} 0 < \mu < \frac{2}{2\lambda_i + \sum_{j=1}^{l} \lambda_j} &\leq \frac{2}{2\lambda_{\min} + \sum_{j=1}^{l} \lambda_j} \\ &< \frac{2}{\sum_{j=1}^{l} \lambda_j} = \frac{2}{\text{trace}\{\Sigma_x\}}, \end{aligned}$$

which proves the claim.

5.8. *Gershgorin circle theorem.* Let A be an $l \times l$ matrix, with entries a_{ij}, $i, j = 1, 2, \ldots, l$. Let $R_i := \sum_{\substack{j=1 \\ j \neq i}}^{l} |a_{ij}|$, be the sum of absolute values of the non-diagonal entries in row i. Then show that if λ is an eigenvalue of A, then there exists at least one row i, such that the following is true,

$$|\lambda - a_{ii}| \leq R_i.$$

The last bound defines a circle, which contains the eigenvalue λ.

Solution: By the definition of an eigenvalue we have that

$$A\boldsymbol{x} = \lambda \boldsymbol{x}.$$

Let $|x_i|$ be the maximum entry, i.e.,

$$i: \quad |x_i| \geq |x_j|, \quad j \neq i.$$

Then, we can write that

$$\sum_{j=1}^{l} a_{ij} x_j = \lambda x_i$$

or

$$-a_{ii} x_i + \lambda x_i = \sum_{j \neq i}^{l} a_{ij} x_j$$

or

$$|\lambda - a_{ii}| = \frac{|\sum_{j \neq i}^{l} a_{ij} x_j|}{|x_i|} \leq \sum_{j \neq i}^{l} |a_{ij}| \frac{|x_j|}{|x_i|} \leq \sum_{j \neq i}^{l} |a_{ij}| := R_i,$$

which proves the claim. The same is true if we add the elements column-wise. It suffices to apply the theorem on A^T.

5.9. Apply the Gershgorin circle theorem to prove the bound in (5.45).

Solution: First note that the matrix

$$A = (I - \mu\Lambda)^2 + \mu^2 \Lambda^2 + \mu^2 \boldsymbol{\lambda}\boldsymbol{\lambda}^T$$

is non-negative definite. Indeed $\forall \boldsymbol{x} \in \mathbb{R}^l$, we have that

$$\boldsymbol{x}^T (I - \mu\Lambda)^T (I - \mu\Lambda) \boldsymbol{x} + \mu^2 \boldsymbol{x}^T \Lambda^T \Lambda \boldsymbol{x} + \mu^2 \boldsymbol{x}^T \boldsymbol{\lambda}\boldsymbol{\lambda}^T \boldsymbol{x} \geq 0.$$

Hence all its eigenvalues are non-negative; recall that the eigenvalues are real, since $A = A^T$. Hence, there must be a row, i, such that

$$|\lambda_{\max}| - |a_{ii}| \leq |\lambda_{\max} - a_{ii}| \leq \sum_{j \neq i}^{l} |a_{ij}| = \mu^2 \lambda_i \sum_{j \neq i}^{l} \lambda_j,$$

since the off-diagonal elements are $\mu^2 \lambda_i \lambda_j$. Hence,

$$\begin{aligned} 0 \leq \lambda_{\max} \quad &\leq \quad a_{ii} + \mu^2 \lambda_i \sum_{j=1}^{l} \lambda_j \\ &= \quad (1 - \mu\lambda_i)^2 + \mu^2 \lambda_i^2 + \mu^2 \lambda_i \sum_{j=1}^{l} \lambda_j \end{aligned}$$

which then proves the claims following the steps in Problem 5.7.

5.10. Derive the misadjustment formula given in (5.52).

Solution: Taking into account that in steady state, $\boldsymbol{s}_n = \boldsymbol{s}_{n-1} := \boldsymbol{s}$ in (5.42), for the ith element we can write

$$s_i = (1 - 2\mu\lambda_i + 2\mu^2\lambda_i^2)s_i + \mu^2\lambda_i \sum_j \lambda_j s_j + \mu^2\sigma_\eta^2\lambda_i.$$

It is common to neglect the third term in the parenthesis on the right hand side, for very small values of μ, hence,

$$s_i \simeq (1 - 2\mu\lambda_i)s_i + \mu^2\lambda_i \sum_j \lambda_j s_j + \mu^2\sigma_\eta^2\lambda_i.$$

Summing up both sides we obtain,

$$\sum_i s_i \simeq \sum_i s_i - 2\mu \sum_i \lambda_i s_i + \sum_i \mu^2\lambda_i \sum_j \lambda_j s_j + \mu^2\sigma_\eta^2 \sum_i \lambda_i$$

and using (5.48), we get

$$J_{\text{exc},\infty}\left(2 - \mu \sum_i \lambda_i\right) = \mu\sigma_\eta^2 \sum_i \lambda_i.$$

Note that the sum of the eigenvalues is equal to the trace of a matrix. This completes the proof.

5.11. Derive the APA iteration scheme.

Solution: The optimization task is

$$\begin{aligned} &\boldsymbol{\theta}_n = \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\|^2 \\ &\text{s.t.}\;\; \boldsymbol{x}_{n-i}^T\boldsymbol{\theta} = y_{n-i},\; i = 0, \ldots, q-1. \end{aligned}$$

Using Lagrange multipliers, the Lagrangian becomes

$$L(\boldsymbol{\theta}, \boldsymbol{\lambda}) = (\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1})^T(\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}) + \sum_{i=0}^{q-1} \lambda_i (y_{n-i} - \boldsymbol{\theta}^T\boldsymbol{x}_{n-i}),$$

Taking the derivatives and equating to zero, we get

$$\begin{aligned} \boldsymbol{\theta} &= \boldsymbol{\theta}_{n-1} + \frac{1}{2}\sum_{i=0}^{q-1} \lambda_i \boldsymbol{x}_{n-i} \\ &= \boldsymbol{\theta}_{n-1} + \frac{1}{2}X_n^T\boldsymbol{\lambda}, \end{aligned} \qquad (5)$$

where

$$X_n = \begin{bmatrix} \boldsymbol{x}_n^T \\ \vdots \\ \boldsymbol{x}_{n-q+1}^T \end{bmatrix}, \quad X_n^T = [\boldsymbol{x}_n, \dots, \boldsymbol{x}_{n-q+1}].$$

Substituting (5) into the constraints, written as

$$\boldsymbol{y}_n = X_n \boldsymbol{\theta}$$

or

$$\frac{1}{2}\boldsymbol{\lambda} = \left(X_n X_n^T\right)^{-1} (\boldsymbol{y}_n - X_n \boldsymbol{\theta}_{n-1}).$$

Hence

$$\begin{aligned} \boldsymbol{\theta}_n &= \boldsymbol{\theta}_{n-1} + X_n^T \left(X_n X_n^T\right)^{-1} \boldsymbol{e}_n \\ \boldsymbol{e}_n &= \boldsymbol{y}_n - X_n \boldsymbol{\theta}_{n-1}, \end{aligned}$$

5.12. Given a value $\boldsymbol{x}$, define the hyperplane comprising all values of $\boldsymbol{\theta}$ such as

$$\boldsymbol{x}^T \boldsymbol{\theta} - y = 0.$$

Then $\boldsymbol{x}$ is perpendicular to the hyperplane .

Solution: Let two points that belong to the hyperplane, $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. Then by definition we have

$$\begin{aligned} \boldsymbol{x}^T \boldsymbol{\theta}_1 - y &= \boldsymbol{x}^T \boldsymbol{\theta}_2 - y \Rightarrow \\ \boldsymbol{x}^T (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2) &= 0. \end{aligned}$$

Since this is true for any $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$ in the hyperplane, $\boldsymbol{x}$ is perpendicular to it.

5.13. Derive the recursions for the widely-linear APA.

Solution: Define

$$e_n = y_n - \boldsymbol{\phi}_{n-1}^H \tilde{\boldsymbol{x}}_n,$$

where

$$\tilde{\boldsymbol{x}}_n = \begin{bmatrix} \boldsymbol{x}_n \\ \boldsymbol{x}_n^* \end{bmatrix}.$$

Let also

$$\tilde{X}_n = \begin{bmatrix} \tilde{\boldsymbol{x}}_n^H \\ \vdots \\ \tilde{\boldsymbol{x}}_{n-q+1}^H \end{bmatrix}.$$

For the APA, we require

$$\boldsymbol{\varphi}^H \tilde{\boldsymbol{x}}_{n-i} = y_{n-i}, \ i = 0, \dots, q-1.$$

or

$$(\boldsymbol{\varphi}^*)^T \tilde{\boldsymbol{x}}_{n-i} = y_{n-i},$$

or

$$\tilde{\boldsymbol{x}}_{n-i}^H \boldsymbol{\varphi} = y_{n-i}^*, \;\; i = 0, 1, \ldots, q-1,$$

or

$$\tilde{X}_n \boldsymbol{\varphi} = \boldsymbol{y}_n^*, \tag{6}$$

with

$$\boldsymbol{y}_n = [y_n, \ldots, y_{n-q+1}]^T.$$

The Lagrangian becomes

$$(\boldsymbol{\varphi} - \boldsymbol{\varphi}_{n-1})^H (\boldsymbol{\varphi} - \boldsymbol{\varphi}_{n-1}) + \sum_{i=0}^{q-1} \lambda_i (y_{n-i}^* - \tilde{\boldsymbol{x}}_{n-i}^H \boldsymbol{\varphi})$$

or

$$\boldsymbol{\varphi}^H \boldsymbol{\varphi} + \boldsymbol{\phi}_{n-1}^H \boldsymbol{\varphi}_{n-1} - \boldsymbol{\varphi}_{n-1}^H \boldsymbol{\varphi} - \boldsymbol{\varphi}^H \boldsymbol{\varphi}_{n-1} + \sum_{i=0}^{q-1} \lambda_i (y_{n-1}^* - \tilde{\boldsymbol{x}}_{n-i}^H \boldsymbol{\varphi}).$$

Taking the gradient with respect to $\boldsymbol{\varphi}$ and treating $\boldsymbol{\varphi}^*$ as a constant and equating to $\mathbf{0}$ we obtain

$$\boldsymbol{\varphi}^* = \boldsymbol{\varphi}_{n-1}^* + \tilde{X}_n^T \boldsymbol{\lambda}, \tag{7}$$

or

$$\boldsymbol{\varphi} = \boldsymbol{\varphi}_{n-1} + \tilde{X}_n^H \boldsymbol{\lambda}^*. \tag{8}$$

Plugging (8) into (6) we obtain

$$\boldsymbol{\lambda}^* = (\tilde{X}_n \tilde{X}_n^H)^{-1} \boldsymbol{e}_n^*$$

where

$$e_n^* = \boldsymbol{y}_n^* - \tilde{X}_n \boldsymbol{\varphi}_{n-1}$$

and the widely linear APA has been obtained.

5.14. Show that a similarity transformation of a square matrix, via a unitary matrix, does not affect the eigenvalues.

Solution: Let Σ be a matrix and let

$$\acute{\Sigma} = T^H \Sigma T.$$

Let λ be an eigenvalue of $\acute{\Sigma}$. Then

$$\begin{aligned} \acute{\Sigma} \boldsymbol{q} &= \lambda \boldsymbol{q} \Rightarrow \\ T^H \Sigma T \boldsymbol{q} &= \lambda \boldsymbol{q} \Rightarrow \\ \Sigma T \boldsymbol{q} &= \lambda T \boldsymbol{q} \end{aligned}$$

Hence λ is also an eigenvalue of Σ. This is true for all eigenvalues of $\acute{\Sigma}$. Thus, both matrices share the same eigenvalues.

5.15. Show that if $\mathbf{x} \in \mathbb{R}^l$ is a Gaussian random vector, then

$$F := \mathbb{E}[\mathbf{x}\mathbf{x}^T S \mathbf{x}\mathbf{x}^T] = \mathit{\Sigma}_x \mathrm{trace}\{S \mathit{\Sigma}_x\} + 2\mathit{\Sigma}_x S \mathit{\Sigma}_x$$

and if $\mathbf{x} \in \mathbb{C}^l$,

$$F := \mathbb{E}[\mathbf{x}\mathbf{x}^H S \mathbf{x}\mathbf{x}^H] = \mathit{\Sigma}_x \mathrm{trace}\{S \mathit{\Sigma}_x\} + \mathit{\Sigma}_x S \mathit{\Sigma}_x$$

Solution: Let

$$Q^H \mathit{\Sigma}_x Q = \Lambda = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_l \end{pmatrix},$$

where we know that $\lambda_i \geq 0$, $i = 1, 2, \dots, l$. Define

$$\boldsymbol{x}' = Q^H \boldsymbol{x}.$$

Then

$$\mathit{\Sigma}_{\boldsymbol{x}'} = \mathbb{E}[\boldsymbol{x}'\boldsymbol{x}'^H] = Q^H \mathit{\Sigma}_x Q = \Lambda.$$

Thus the entries of $\boldsymbol{x}'$ are uncorrelated, hence independent since they are Gaussians. Define

$$\begin{aligned} F' = Q^H F Q &= \mathbb{E}[Q^H \mathbf{x}\mathbf{x}^H Q Q^H S Q Q^H \mathbf{x}\mathbf{x}^H Q] \\ &= \mathbb{E}[\mathbf{x}'\mathbf{x}'^H S' \mathbf{x}'\mathbf{x}'^H], \end{aligned}$$

where

$$S' = Q^H S Q.$$

- For real data, the matrix under the expectation will comprise entries of the form
$$\mathbb{E}[(x'_{k_1} x'_{k_2} x'_{k_3} x'_{k_4}) S'_{ij}].$$
Since the variables are independent, the only non zero entries will be those such as $k_1 = k_2 = k_3 = k_4$, or $k_1 = k_2$ and $k_3 = k_4$, $k_1 = k_4$ and $k_2 = k_3$ or $k_1 = k_3$ and $k_2 = k_4$. Then it can be checked out (use e.g., a 3×3 matrix example) that
$$F' = \mathit{\Sigma}'_x \mathrm{trace}\{S' \mathit{\Sigma}'_x\} + 2\mathit{\Sigma}'_x S' \mathit{\Sigma}'_x$$
which results in
$$F = \mathit{\Sigma}_x \mathrm{trace}\{S \mathit{\Sigma}_x\} + 2\mathit{\Sigma}_x S \mathit{\Sigma}_x.$$
- For complex valued data, we assume circular symmetric variable, thus $\mathbb{E}[x_k^2] = \mathbb{E}[x_k'^2] = 0$, which results in
$$F = \mathit{\Sigma}_x \mathrm{Trace}\{S \mathit{\Sigma}_x\} + \mathit{\Sigma}_x S \mathit{\Sigma}_x.$$

5.16. Show that if a $l \times l$ matrix C is right stochastic, then all its eigenvalues satisfy

$$|\lambda_i| \leq 1, \quad i = 1, 2, \ldots, l.$$

The same holds true for left and doubly stochastic matrices.

Solution: By the definition of a stochastic matrix, its elements are nonnegative and each row adds to 1. Recalling the Gershgorin circle theorem, we know that for any matrix, all the eigenvalues satisfy

$$\exists i: \ |\lambda - c_{ii}| \leq \sum_{\substack{j=1 \\ j \neq i}}^{l} |c_{ij}| = \sum_{\substack{j=1 \\ j \neq i}}^{l} c_{ij}$$

for at least one i. That is

$$\exists i: \ -\sum_{\substack{j=1 \\ j \neq i}}^{l} c_{ij} \leq \lambda - c_{ii} \leq \sum_{\substack{j=1 \\ j \neq i}}^{l} c_{ij}$$

or

$$-1 \leq \lambda \leq 1$$

or

$$|\lambda| \leq 1.$$

Moreover note that $\lambda = 1$ is an eigenvalue, since

$$C\mathbf{1} = \mathbf{1}.$$

For the left and doubly stochastic, use the fact that C, C^T share the same eigenvalues.

5.17. Prove Theorem 5.2

Solution: Collect all estimates from all nodes at iteration, i, together in a common vector,

$$\underline{\boldsymbol{\theta}}^{(i)} = [\boldsymbol{\theta}_1^{T(i)}, \boldsymbol{\theta}_2^{T(i)}, \ldots, \boldsymbol{\theta}_K^{T(i)}]^T; \ \boldsymbol{\theta}_k \in \mathbb{R}^l; \ k = 1, 2, \ldots, K.$$

Then the consensus iteration can be compactly written (check it) as,

$$\underline{\boldsymbol{\theta}}^{(i)} = \mathcal{A}^T \underline{\boldsymbol{\theta}}^{(i-1)}, \tag{9}$$

where $\mathcal{A}^T$ is defined as the Kronecker product

$$\mathcal{A}^T = A^T \otimes I_l \tag{10}$$

with

$$\underline{\boldsymbol{\theta}}^{(0)} = [\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_k^T]^T, \ \boldsymbol{x}_K \in \mathbb{R}^l, \ k = 1, 2, \ldots, K.$$

Let

$$\boldsymbol{x} := \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{x}_k.$$

Then, by the definition of the Kronecker product it can be written as

$$\boldsymbol{x} = \frac{1}{K}\left(\mathbf{1}^T \otimes I_l\right)\underline{\boldsymbol{\theta}}^{(0)}.$$

Define

$$\boldsymbol{c}_k^i := \boldsymbol{\theta}_k^{(i)} - \boldsymbol{x},$$

that is, the error vector. Then, we can write that,

$$\underline{\boldsymbol{c}}^{(i)} := [\boldsymbol{c}_1^{T(i)}, \ldots, \boldsymbol{c}_k^{T(i)}]^T = \underline{\boldsymbol{\theta}}^{(i)} - (\mathbf{1} \otimes I_l)\boldsymbol{x}.$$

- *Sufficiency.* Iterating (9) we obtain

  $$\underline{\boldsymbol{\theta}}^{(i)} = \left(\mathcal{A}^T\right)^i \underline{\boldsymbol{\theta}}^{(0)};$$

  or

  $$\begin{aligned}\underline{\boldsymbol{c}}^{(i)} &= \left(\mathcal{A}^T\right)^i \underline{\boldsymbol{\theta}}^{(0)} - (\mathbf{1} \otimes I_l)\boldsymbol{x} \\ &= \left(\mathcal{A}^T\right)^i \underline{\boldsymbol{\theta}}^{(0)} - \frac{1}{K}(\mathbf{1} \otimes I_l)(\mathbf{1}^T \otimes I_l)\underline{\boldsymbol{\theta}}^{(0)} \\ &= \left(\left(\mathcal{A}^T\right)^i - \frac{1}{K}(\mathbf{1} \otimes I_l)(\mathbf{1}^T \otimes I_l)\right)\underline{\boldsymbol{\theta}}^{(0)}.\end{aligned}$$

  Using the identity

  $$(B \otimes C)(D \otimes E) = (BD \otimes CE),$$

  we obtain

  $$\begin{aligned}\underline{\boldsymbol{c}}^{(i)} &= \left(\left(\mathcal{A}^T\right)^i - \frac{1}{K}\left(\mathbf{1}\mathbf{1}^T \otimes I_l I_l\right)\right)\underline{\boldsymbol{\theta}}^{(0)} \\ &= \left(\left(\mathcal{A}^T\right)^i - \frac{1}{K}\left(\mathbf{1}\mathbf{1}^T \otimes I_l\right)\right)\underline{\boldsymbol{\theta}}^{(0)}.\end{aligned}$$

  Recalling the definition of $\mathcal{A}$ in (10) and the properties

  $$(B + C) \otimes D = B \otimes D + C \otimes D,$$

  $$(B \otimes C)^i = B^i \otimes C^i,$$

  we obtain

  $$\underline{\boldsymbol{c}}^{(i)} = \left(A^T)^i - \frac{1}{K}\mathbf{1}\mathbf{1}^T\right) \otimes I_l \underline{\boldsymbol{\theta}}^{(0)}$$

  However, note that

  $$\left(A^T - \frac{1}{K}\mathbf{1}\mathbf{1}^T\right)^i = (A^T)^i - \frac{1}{K}\mathbf{1}\mathbf{1}^T. \tag{11}$$

This is because of the doubly stochastic assumption on A. Indeed

$$\begin{aligned}\left(A^T - \frac{1}{K}\mathbf{1}\mathbf{1}^T\right)^2 &= (A^T)^2 - \frac{1}{K}\mathbf{1}\mathbf{1}^T A^T - \frac{1}{K}A^T\mathbf{1}\mathbf{1}^T \\ &+ \frac{1}{K^2}\mathbf{1}\mathbf{1}^T\mathbf{1}\mathbf{1}^T,\end{aligned}$$

and since

$$A^T\mathbf{1} = \mathbf{1}, \text{ and } A\mathbf{1} = \mathbf{1}, \text{ and } \mathbf{1}^T\mathbf{1} = K$$

we get

$$\left(A^T - \frac{1}{K}\mathbf{1}\mathbf{1}^T\right)^2 = (A^T)^2 - \frac{1}{K}\mathbf{1}\mathbf{1}^T,$$

which generalizes to higher powers by induction. Hence

$$\begin{aligned}\underline{\boldsymbol{c}}^{(i)} &= \left(\left(A^T - \frac{1}{K}\mathbf{1}\mathbf{1}^T\right)^i \otimes I_l\right)\underline{\boldsymbol{\theta}}^{(0)} \\ &= \left(\left(A^T - \frac{1}{K}\mathbf{1}\mathbf{1}^T\right) \otimes I_l\right)^i \underline{\boldsymbol{\theta}}^{(0)},\end{aligned}$$

due to the property

$$(B \otimes C)^i = B^i \otimes C^i.$$

It is known from the linear algebra that the eigenvalues of the Kronecker product $B \otimes C$ are given by all combinations of the products of the eigenvalues of B and C. Since A is a consensus matrix, this guarantees that all eigenvalues are less than one and the iteration converges to zero.

- *Necessary*. Assume that

$$\underline{\boldsymbol{\theta}}^{(i)} = \left(\mathcal{A}^T\right)^i \underline{\boldsymbol{\theta}}^{(0)}$$

converges, i.e.,

$$\begin{aligned}\lim_{i\to\infty}\left(\mathcal{A}^T\right)^i \boldsymbol{\theta}^{(0)} &= (\mathbf{1} \otimes I_l)\boldsymbol{x} \\ &= \frac{1}{K}(\mathbf{1} \otimes I_l)(\mathbf{1}^T \otimes I_l)\boldsymbol{\theta}^{(0)}\end{aligned}$$

for any $\boldsymbol{\theta}^{(0)}$. Hence, this implies that

$$\begin{aligned}\lim_{i\to\infty}\left(\mathcal{A}^T\right)^i &= \frac{1}{K}(\mathbf{1} \otimes I_l)(\mathbf{1}^T \otimes I_l) \\ &= \frac{1}{K}(\mathbf{1}\mathbf{1}^T \otimes I_l).\end{aligned}$$

However,

$$\left(\mathcal{A}^T\right)^i = (A^T \otimes I_l)^i = \left(A^T\right)^i \otimes I_l.$$

Hence, combining

$$\lim_{i\to\infty} \left(A^T\right)^i = \frac{1}{K}\mathbf{1}\mathbf{1}^T. \tag{12}$$

This in turn implies that

$$\begin{aligned}\lim_{i\to\infty} A^T \left(A^T\right)^i &= A^T \frac{1}{K}\mathbf{1}\mathbf{1}^T \\ &= \lim_{i\to\infty} \left(A^T\right)^{i+1}.\end{aligned}$$

Thus

$$\frac{1}{K}A^T\mathbf{1}\mathbf{1}^T = \frac{1}{K}\mathbf{1}\mathbf{1}^T,$$

or

$$A^T\mathbf{1} = \mathbf{1},$$

that is, it is left stochastic. By considering

$$\lim_{i\to\infty} \left(A^T\right)^i A^T = \frac{1}{K}\mathbf{1}\mathbf{1}^T A^T.$$

we can prove that it is right stochastic as well,

$$A\mathbf{1} = \mathbf{1}.$$

Since we have proved that A is doubly stochastic, then (11) holds true. Thus, from (12), we get

$$\lim_{i\to\infty} \left(\left(A^T\right)^i - \frac{1}{K}\mathbf{1}\mathbf{1}^T\right) = \lim_{i\to\infty} \left(\left(A^T\right) - \frac{1}{K}\mathbf{1}\mathbf{1}^T\right)^i = 0,$$

which can only be true if the respective eigenvalues are less than one. This concludes the proof.

Solutions To Problems of Chapter 6

6.1. Show that if $A \in \mathbb{C}^{m \times m}$ is nonnegative definite, its trace is nonnegative.

Solution: By the definition of a positive semidefinite matrix, $\forall \boldsymbol{x} \in \mathbb{C}^m$,

$$\boldsymbol{x}^H A \boldsymbol{x} \geq 0.$$

Hence, this will also be true for $\boldsymbol{x} = [1, 0, \ldots, 0]$. Thus, $[A]_{11} \geq 0$. In the same way, we can prove that any element in the diagonal is a nonnegative real number.

6.2. Show that under a) the independence assumption of successive observation vectors and b) the presence of white noise independent of the input, then the LS estimator is asymptotically distributed according to the normal distribution, i.e.,

$$\sqrt{N}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \longrightarrow \mathcal{N}(\mathbf{0}, \sigma^2 \Sigma_x^{-1}),$$

where σ^2 is the noise variance and Σ_x the covariance matrix of the input observation vectors, assuming that it is invertible.

Solution: Recall that according to the law of large numbers, we have

$$\frac{1}{N} \sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^T \longrightarrow \Sigma_x,$$

where the limit indicates convergence in probability. Also,

$$\mathbb{E}[\mathbf{x}_n \eta_n] = \mathbb{E}[\mathbf{x}_n]\, \mathbb{E}[\eta_n] = \mathbf{0}.$$

Moreover, we have that,

$$\mathrm{Cov}(\mathbf{x}_n \eta_n) = \mathbb{E}[\eta_n^2 \mathbf{x}_n \mathbf{x}_n^T] = \mathbb{E}_x[\mathbb{E}[\eta_n^2 | \mathbf{x}_n] \mathbf{x}_n \mathbf{x}_n^T] = \sigma^2 \Sigma_x.$$

Thus, the covariance matrix of the sum of independent terms in

$$\frac{1}{\sqrt{N}} \sum_{n=1}^{N} \mathbf{x}_n \eta_n \tag{1}$$

is given as,

$$\Sigma_{x\eta} = \sigma^2 \Sigma_x.$$

Also, due to the independence of the summands and from the central limit theorem, we have that

$$\frac{1}{\sqrt{N}} \sum_{n=1}^{N} \mathbf{x}_n \eta_n \longrightarrow \mathcal{N}(\mathbf{0}, \sigma^2 \Sigma_x).$$

Combining the previous findings and using the result from the text, we get

$$\begin{aligned}\sqrt{N}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) &= \left(\frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^T\right)^{-1} \left(\frac{1}{\sqrt{N}}\sum_{n=1}^{N} \mathbf{x}_n \eta_n\right) \\ &\longrightarrow \Sigma_x^{-1}\mathcal{N}(\mathbf{0}, \sigma^2 \Sigma_x) = \mathcal{N}(\mathbf{0}, \sigma^2 \Sigma_x^{-1}), \qquad (2)\end{aligned}$$

where the limit is meant to be convergent in distribution. To prove the above we have made use of the so called Slutsky's theorem, which states the following:

If $\mathbf{x}_n \longrightarrow \mathbf{x}$, in distribution and $\mathrm{Y}_n \longrightarrow Y$, in probability, where Y is a constant, then

$$\mathrm{Y}_n^{-1}\mathbf{x}_n \longrightarrow Y^{-1}\mathbf{x},$$

where the last limit is in distribution.

6.3. Let $X \in \mathbb{C}^{m \times l}$. Then show that the two matrices,

$$XX^H \text{ and } X^H X,$$

have the same nonzero eigenvalues.

Solution: Let λ_i be an eigenvalue of $X^H X$. Then

$$\begin{aligned}(X^H X)\boldsymbol{v}_i &= \lambda_i \boldsymbol{v}_i \Rightarrow \\ X(X^H X)\boldsymbol{v}_i &= \lambda_i X \boldsymbol{v}_i \Rightarrow \\ (XX^H)X\boldsymbol{v}_i &= \lambda_i X\boldsymbol{v}_i,\end{aligned}$$

or

$$(XX^H)\boldsymbol{q}_i = \lambda_i \boldsymbol{q}_i.$$

That is, λ_i is also an eigenvalue of XX^H.

6.4. Show that if $X \in \mathbb{C}^{m \times l}$, then the eigenvalues of XX^H $(X^H X)$ are real and nonnegative. Moreover, show that if $\lambda_i \neq \lambda_j$, $\boldsymbol{v}_i \perp \boldsymbol{v}_j$.

Solution: By the definition we have,

$$\begin{aligned}XX^H\boldsymbol{v}_i &= \lambda_i \boldsymbol{v}_i \Rightarrow \\ \boldsymbol{v}_i^H XX^H \boldsymbol{v}_i &= \lambda_i \|\boldsymbol{v}_i\|^2 \Rightarrow \\ \|X^H \boldsymbol{v}_i\|^2 &= \lambda_i \|\boldsymbol{v}_i\|^2 \Rightarrow \\ \lambda_i &\geq 0.\end{aligned}$$

Let now $\lambda_i \neq \lambda_j$. Then we have,

$$\begin{aligned}XX^H\boldsymbol{v}_i &= \lambda_i \boldsymbol{v}_i \Rightarrow \\ \boldsymbol{v}_j^H XX^H \boldsymbol{v}_i &= \lambda_i \boldsymbol{v}_j^H \boldsymbol{v}_i,\end{aligned}$$

and since XX^H is Hermitian and λ_i real,

$$\boldsymbol{v}_i^H XX^H \boldsymbol{v}_j = \lambda_i \boldsymbol{v}_i^H \boldsymbol{v}_j. \tag{3}$$

Similarly,

$$\begin{aligned} XX^H \boldsymbol{v}_j &= \lambda_j \boldsymbol{v}_j \Rightarrow \\ \boldsymbol{v}_i^H XX^H \boldsymbol{v}_j &= \lambda_j \boldsymbol{v}_i^H \boldsymbol{v}_j \end{aligned} \tag{4}$$

Subtracting (3) from (4) we obtain

$$(\lambda_i - \lambda_j)(\boldsymbol{v}_i^H \boldsymbol{v}_j) = 0 \Rightarrow \boldsymbol{v}_i \perp \boldsymbol{v}_j.$$

Note that it can be shown that for Hermitian matrices, if $\lambda_i = \lambda_j$, still one can construct orthogonal eigenvectors.

6.5. Let $X \in \mathbb{C}^{m \times l}$. Then show that if $\boldsymbol{v}_i$ is the normalized eigenvector of $X^H X$, corresponding to $\lambda_i \neq 0$, then the corresponding normalized eigenvector $\boldsymbol{u}_i$ of XX^H is given by,

$$\boldsymbol{u}_i = \frac{1}{\sqrt{\lambda_i}} X \boldsymbol{v}_i$$

Solution: From Problem 6.3, we know that the eigenvectors XX^H and $X^H X$ corresponding to λ_i are related as,

$$\boldsymbol{q}_i = X \boldsymbol{v}_i.$$

By the respective definition we have

$$\begin{aligned} XX^H \boldsymbol{v}_i &= \lambda_i \boldsymbol{v}_i \Rightarrow \\ \boldsymbol{v}_i^H XX^H \boldsymbol{v}_i &= \lambda_i \|\boldsymbol{v}_i\|^2 = \lambda_i \end{aligned}$$

for normalized $\boldsymbol{v}_i$, or

$$\|X\boldsymbol{v}_i\|^2 = \|\boldsymbol{q}_i\|^2 = \lambda_i.$$

Thus the normalized $\boldsymbol{u}_i$ will be

$$\boldsymbol{u}_i = \frac{1}{\sqrt{\lambda_i}} \boldsymbol{q}_i = \frac{1}{\sqrt{\lambda_i}} X \boldsymbol{v}_i$$

6.6. Show Eq. (6.19).

Solution: From the respective definitions, we get

$$\begin{aligned} \hat{\boldsymbol{y}} &= X(X^T X)^{-1} X^T \boldsymbol{y} = U_l D V_l^T (V_l D U_l^T U_l D V_l^T)^{-1} V_l D U_l^T \boldsymbol{y} \\ &= U_l D V_l^T (V_l D^2 V_l^T)^{-1} V_l D U_l^T \boldsymbol{y} \\ &= U_l D V_l^T (V_l D^{-2} V_l^T) V_l D U_l^T \boldsymbol{y} \\ &= U_l U_l^T \boldsymbol{y}, \end{aligned}$$

where we have used the definition of the unitary $(VV^T = V^T V = I)$.

6.7. Show that the right singular vectors, $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r$, corresponding to the r singular values of a rank-r matrix, X, solve the following iterative optimization task: compute $\boldsymbol{v}_k,\ k = 2, 3, \ldots, r$, such as,

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}||X\boldsymbol{v}||^2, \\ \text{subject to} \quad & ||\boldsymbol{v}||^2 = 1, \\ & \boldsymbol{v} \perp \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{k-1}\},\ k \neq 1, \end{aligned}$$

where $|| \cdot ||$ denotes the Euclidean norm.

Solution: We start with $k = 1$, to solve the (Rayleigh ratio) task

$$\begin{aligned} \boldsymbol{v}_1 : \quad & \min_{\boldsymbol{v}} \frac{1}{2}||X\boldsymbol{v}||^2, \\ \text{s.t.} \quad & ||\boldsymbol{v}||^2 = 1. \end{aligned}$$

The corresponding Lagrangian becomes

$$L(\boldsymbol{v}, \mu) = \frac{1}{2}\boldsymbol{v}^T X^T X\boldsymbol{v} - \mu(\boldsymbol{v}^T\boldsymbol{v} - 1),$$

which, after taking the gradient w.r. to $\boldsymbol{v}$ and equating to zero gives,

$$X^T X\boldsymbol{v} = \mu\boldsymbol{v},$$

which shows that the solution $\boldsymbol{v}_1$, is an eigenvector of $X^T X$ and in order to maximize the cost, it must necessarily correspond to the largest singular value. We now proceed to $k = 2$, to solve the task,

$$\begin{aligned} \boldsymbol{v}_k : \quad & \min_{\boldsymbol{v}} \frac{1}{2}||X\boldsymbol{v}||^2,\ k = 1, 2, \ldots, r, \\ \text{s.t.} \quad & ||\boldsymbol{v}||^2 = 1, \\ & \boldsymbol{v} \perp \boldsymbol{v}_1., \end{aligned}$$

The Lagrangian is now given by,

$$L(\boldsymbol{v}, \mu_1, \mu_2) = \frac{1}{2}\boldsymbol{v}^T X^T X\boldsymbol{v} - \mu_1(\boldsymbol{v}^T\boldsymbol{v} - 1) - \mu_2\boldsymbol{v}^T\boldsymbol{v}_1,$$

which results in

$$X^T X\boldsymbol{v} - \mu_1\boldsymbol{v} = \mu_2\boldsymbol{v}_1.$$

However, since $\boldsymbol{v} \perp \boldsymbol{v}_1$, μ_2 has to be equal to zero, which again leads to the fact that $\boldsymbol{v}_2$ must also be an eigenvector of $X^T X$. Since it cannot be equal to $\boldsymbol{v}_1$, the next eigenvector that maximises the cost is the one which corresponds to the second largest singular value. The same rationale carries on for larger values of k.

6.8. Show that projecting the rows of X onto the k-rank subspace, $V_k = \text{span}\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\}$, results in the largest variance, compared to any other k-dimensional subspace, Z_k.

Solution: The projection of a row $\boldsymbol{x}_n$ of X onto V_k is given by

$$\hat{\boldsymbol{x}}_n = \sum_{i=1}^{k} (\boldsymbol{x}_n^T \boldsymbol{v}_i)\boldsymbol{v}_i,$$

and the respective square norm ((scaled) variance of its elements) is

$$||\hat{\boldsymbol{x}}_n||^2 = \sum_{i=1}^{k} (\boldsymbol{x}_n^T \boldsymbol{v}_i)^2.$$

Then, projecting all the rows of the matrix onto V_k results in a total variance,

$$\sum_{n=1}^{N} ||\hat{\boldsymbol{x}}_n||^2 = \sum_{n=1}^{N} \sum_{i=1}^{k} (\boldsymbol{x}_n^T \boldsymbol{v}_i)^2 = \sum_{i=1}^{k} ||X\boldsymbol{v}_i||^2. \tag{5}$$

Hence, it is readily observed that, in general, the variance of the matrix elements, after the projection onto a subspace, depends on the specific subspace. Our goal here is to show that the maximum variance is attained when the subspace is V_k. To this end, let us consider any other subspace Z_k. We will prove the claim inductively.

- $k = 1$: From Problem 6.7, it is readily concluded that

  $$||X\boldsymbol{z}_1||^2 \leq ||X\boldsymbol{v}_1||^2,$$

  where, $\boldsymbol{z}_1$ is a unit vector in the one-dimensional Z_1.
- $k = 2$: In the two-dimensional subspace Z_2, define a basis around $\boldsymbol{z}_2$ so that, $\boldsymbol{z}_2 \perp \boldsymbol{v}_1$. This can always be done, by projection $\boldsymbol{v}_1$ onto Z_2 and selecting $\boldsymbol{z}_2$ to be perpendicular to the projection. Then, from Problem 6.7, we have that

  $$||X\boldsymbol{z}_1||^2 \leq ||X\boldsymbol{v}_1||^2,$$

  and also, since both $\boldsymbol{v}_2$, $\boldsymbol{z}_2$ are perpendicular to $\boldsymbol{v}_1$,

  $$||X\boldsymbol{z}_{\mathbf{2}}||^2 \leq ||X\boldsymbol{v}_2||^2.$$

  Adding by parts the two inequalities, proves the claim that V_2 is the two-dimensional subspace which maximizes the total variance, due to (5).
- Assume that V_{k-1} is the optimal $(k-1)$-dimensional subspace. Build an orthonormal basis in any Z_k, around a $\boldsymbol{z}_k$ so that

  $$\boldsymbol{z}_k \perp \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{k-1}\}.$$

This can always be done. Indeed, let a basis $\boldsymbol{e}_i,\ i = 1, 2, \ldots, k$ be a basis in Z_k. Then $\boldsymbol{z}_k$ can be found by solving the linear system of equations,

$$\boldsymbol{z}_k = \sum_{i=1}^{k} a_i \boldsymbol{e}_i : \quad \sum_{i=1}^{K} a_i \boldsymbol{e}_i^T \boldsymbol{v}_j = 0,\ j = 1, 2, \ldots k-1,$$

which is a system with $k-1$ equations and k unknowns and which, in general, has always a solution.
Then, we have

$$||X\boldsymbol{z}_k||^2 \leq ||X\boldsymbol{v}_k||^2,$$

and also by the assumption of the induction,

$$\sum_{i=1}^{k-1} ||X\boldsymbol{z}_i||^2 \leq \sum_{i=1}^{k-1} ||X\boldsymbol{v}_i||^2.$$

Adding the two inequalities by part, the claim is proved. Note that if k is equal to the rank of the matrix, r, then the total variance of the matrix, which also defines the Frobenius norm, is given by (Problem 6.13),

$$\text{trace}\{X^T X\} = \sum_{i=1}^{r} \sigma_i^2 = \sum_{i=1}^{r} ||X\boldsymbol{v}_i||^2.$$

6.9. Show that the squared Frobenius norm is equal to the sum of the squared singular values.

Solution: By the definition of the Frobenius norm,

$$\|X\|_F^2 := \sum_i \sum_j |X(i,j)|^2 = \|U_r D V_r\|_F.$$

However, it is known from linear algebra that if Q and U are orthogonal matrices then

$$\|X\|_F^2 = \|QXU\|_F^2,$$

and since U_r and V_r are orthogonal matrices, we have that

$$\|X\|_F^2 = \|D\|_F^2 = \sum_{i=1}^{r} \sigma_i^2.$$

The previously used property, concerning orthogonal matrices, is easily shown from the fact that,

$$\|X\|_F^2 = \sum_i \sum_j |X(i,j)|^2 = \text{trace}\{X^T X\}.$$

However, we know that multiplication with unitary matrices does not change the trace. Indeed,

$$\|X\|_F^2 = \text{trace}\{X^T X\} = \text{trace}\{X^T Q^T Q X\} = \|QX\|_F^2.$$

6.10. Show that the best k rank approximation of a matrix X of rank $r > k$, in the Frobenius norm sense, is given by:

$$\hat{X} = \sum_{i-1}^{k} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T,$$

where σ_i are the singular values and $\boldsymbol{v}_i,\ \boldsymbol{u}_i,\ i = 1, 2, \ldots, r$, are the right and left singular vectors of X, respectively. Then show that the approximation error is given by:

$$\sqrt{\sum_{i=k+1}^{r} \sigma_i^2}.$$

Solution: Let,

$$\hat{X} := \sum_{i=1}^{k} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T = \sum_{i=1}^{k} (X\boldsymbol{v}_i)\boldsymbol{v}_i^T. \tag{6}$$

From (6), it is readily deduced that the rows of $\hat{X}$, $\hat{\boldsymbol{x}}_n^T,\ n = 1, 2, \ldots, N$, are projections of the respective rows of X onto the subspace spanned by $V_k = \text{span}\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\}$, i.e.,

$$\hat{\boldsymbol{x}}_n^T = \sum_{i=1}^{k} (\boldsymbol{x}_n^T \boldsymbol{v}_i)\boldsymbol{v}_i^T.$$

Thus, we can write that

$$\boldsymbol{x}_n = \hat{\boldsymbol{x}}_n + \boldsymbol{e}_n,$$

where $\boldsymbol{e}_n$ is the error vector. Hence, the respective norms are related (due to the orthogonality) via the Pythagoras theorem,

$$||\boldsymbol{x}_n||^2 = ||\hat{\boldsymbol{x}}_n||^2 + ||\boldsymbol{e}_n||^2.$$

However, we know form Problem 6.8, that projecting onto the V_k results in the largest norm of $||\hat{\boldsymbol{x}}_n||^2$, compared to any other subspace; this is equivalent with the smallest error norm $||\boldsymbol{e}_n||$. Hence, the corresponding Frobenius norm,

$$||X - \hat{X}||_F^2 = \sum_{n=1}^{N} ||\boldsymbol{e}_n||^2,$$

is the smallest one. We have proved that from all possible approximations of X, obtained by projecting its rows onto any rank k subspace, the best one is $\hat{X}$.

Let us now assume that there is another matrix B, of rank k, which results in a better approximation. However, this is not possible. Let X_B be the approximation of X with rows equal to the projections of the rows of X

onto the row space of B, which is of rank k. Then, if $B \neq X_B$, then X_B will be a better approximation compared to B, i.e.,

$$||X - \hat{X}||_F^2 < ||X - X_B||_F^2 < ||X - B||_F^2,$$

which proves the claim.

For the error matrix we have that,

$$E = \sum_{i=k+1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T.$$

Thus,

$$\begin{aligned} ||E||_F^2 &= \text{trace}\{E^T E\} = \text{trace}\Big\{ \sum_{i=k+1}^{r} \sigma_i \boldsymbol{v}_i \boldsymbol{u}_i^T \sum_{j=k+1}^{r} \boldsymbol{u}_j \boldsymbol{v}_j^T \Big\} \\ &= \text{trace}\Big\{ \sum_i \sum_j \sigma_i \sigma_j \boldsymbol{v}_i \boldsymbol{u}_i^T \boldsymbol{u}_j \boldsymbol{v}_j^T \Big\} \\ &= \sum_{i=k+1}^{r} \sigma_i^2, \qquad (7) \end{aligned}$$

due to the orthonormality of the involved vectors and the properties of the trace.

6.11. Show that $\hat{X}$, as given in Problem 6.10, also minimizes the spectral norm and that,

$$||X - \hat{X}||_2 = \sigma_{k+1}.$$

Solution: We first show that,

$$||X - \hat{X}||_2 = \sigma_{k+1}.$$

By the definition of the spectral norm, it is equal to the maximum singular value of the respective matrix; thus, it suffices to compute the singular value of the error matrix.

By the definition of the maximum singular value, we know that it is the value that maximizes the square norm,

$$||(X - \hat{X})\boldsymbol{v}||^2,$$

subject to the constrain that $||\boldsymbol{v}|| = 1$. Let

$$\boldsymbol{v} = \sum_{i=1}^{l} a_i \boldsymbol{v}_i.$$

Then, we have that

$$\begin{aligned} ||(X-\hat{X})\boldsymbol{v}||^2 &= ||\sum_{i=k+1}^{r}(\sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T)\sum_{j=1}^{l} a_j \boldsymbol{v}_j||^2 \\ &= ||\sum_{i=k+1}^{r} a_i \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T \boldsymbol{v}_i||^2 \\ &= \sum_{i=k+1}^{r} a_i^2 \sigma_i^2. \end{aligned} \tag{8}$$

The maximum value of the above, taking into account that $\sum_{i=1}^{l} a_i^2 = 1$, occurs if $a_{k+1} = 1$ and the rest are zero, which proves the claim that

$$||X-\hat{X}||_2 = \sigma_{k+1}.$$

Let us now turn our focus in showing the best k-rank approximation. Let B be another matrix that results in lower error, $||X-B||_2 < \sigma_{k+1}$. Since B is a rank k matrix, its null space will be of dimension $l-k$. Thus, by basic dimension arguments,

$$S := \mathcal{N}(B) \cap \operatorname{span}\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k, \boldsymbol{v}_{k+1}\} \neq \emptyset.$$

Hence, $\exists\ \boldsymbol{z} \neq \boldsymbol{0} \in S$, $||\boldsymbol{z}|| = 1$. Then, by the definition of the spectral norm, and taking into account that $B\boldsymbol{z} = \boldsymbol{0}$, we get

$$\begin{aligned} ||X-B||_2^2 &= \max_{||\boldsymbol{v}||=1} ||(X-B)\boldsymbol{v}||^2 \geq ||(X-B)\boldsymbol{z}||^2 = ||X\boldsymbol{z}||^2 \\ &= ||\sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T \boldsymbol{z}||^2 \\ &= ||\sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T \sum_{j=1}^{k+1} (\boldsymbol{v}_j^T \boldsymbol{z})\boldsymbol{v}_j||^2 \\ &= ||\sum_{i=1}^{k+1} \sigma_i \boldsymbol{u}_i (\boldsymbol{v}_i^T \boldsymbol{z})||^2 \\ &= \sum_{i=1}^{k+1} \sigma_i^2 (\boldsymbol{v}_i^T \boldsymbol{z})^2 \geq \sigma_{k+1}^2 \sum_{i=1}^{k+1} (\boldsymbol{v}_i^T \boldsymbol{z})^2 = \sigma_{k+1}^2, \end{aligned}$$

since $||\boldsymbol{z}|| = 1$. This contradicts the assumption that B provides a smaller error norm and proves the claim.

6.12. Show that the Frobenius and spectral norms are unaffected by multiplication with unitary matrices, i.e.,

$$\|X\|_F = \|QXU\|_F$$

and

$$\|X\|_2 = \|QXU\|,$$

if $QQ^T = UU^T = I$.

Solution: The proof for the Frobenius was given in Problem 6.9. From the definition of the the spectral norm, we have that

$$\|X\|_2 = \sigma_1,$$

where λ_1 is the largest eigenvalue of XX^T (X^TX). However, multiplication by a unitary matrix does not affect the respective eigenvalues. The latter is known from linear algebra and it is trivially shown.

6.13. Show that the null and range spaces of a $m \times l$ matrix, X, of rank r are given by,

$$\begin{aligned} \mathcal{N}(X) &= \text{span}\{\boldsymbol{v}_{r+1}, \ldots, \boldsymbol{v}_l\}, \\ \mathcal{R}(X) &= \text{span}\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r\}, \end{aligned}$$

where

$$X = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m] \begin{bmatrix} D & O \\ O & O \end{bmatrix} \begin{bmatrix} \boldsymbol{v}_1^T \\ \vdots \\ \boldsymbol{v}_l^T \end{bmatrix}.$$

Solution: Recall that

$$X = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T.$$

Hence, $\forall \boldsymbol{a} \in \mathbb{R}^l$,

$$X\boldsymbol{a} = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i (\boldsymbol{v}_i^T \boldsymbol{a}),$$

and $\mathcal{R}(X) = \text{span}\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r\}$.

From linear algebra we know that if the dimensionality of the range space is r, then the dimensionality of the the null space will be $l - r$, and since

$$X\boldsymbol{v}_j = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i (\boldsymbol{v}_i^T \boldsymbol{v}_j) = \boldsymbol{0}, \quad j = r+1, \ldots, l\ ,$$

due to the orthonormality of $\boldsymbol{v}_i$, $i = 1, 2, \ldots, l$, then

$$\mathcal{N}(X) = \text{span}\{\boldsymbol{v}_{r+1}, \ldots, \boldsymbol{v}_l\}.$$

6.14. Show that for the ridge regression

$$\hat{\boldsymbol{y}} = \sum_{i=1}^{l} \frac{\sigma_i^2}{\lambda + \sigma_i^2} (\boldsymbol{u}_i^T \boldsymbol{y}) \boldsymbol{u}_i$$

Solution: We have that

$$\begin{aligned}
\hat{\boldsymbol{y}} &= U_l D V_l^T \left(V_l D U_l^T U_l D V_l^T + \lambda I \right)^{-1} V_l D U_l^T \boldsymbol{y} \\
&= U_l D V_l^T \left(V_l D^2 V_l^T + \lambda V_l V_l^T \right)^{-1} V_l D U_l^T \boldsymbol{y} \\
&= U_l D V_l^T \left[V_l (D^2 + \lambda I) V_l^T \right]^{-1} V_l D U_l^T \boldsymbol{y} \\
&= U_l D (D^2 + \lambda I)^{-1} D U_l^T \boldsymbol{y} \\
&= \sum_{i=1}^{l} \frac{\sigma_i^2}{\lambda + \sigma_i^2} (\boldsymbol{u}_i^T \boldsymbol{y}) \boldsymbol{u}_i .
\end{aligned}$$

6.15. Show that the normalized steepest descent direction of $J(\boldsymbol{\theta})$ at a point $\boldsymbol{\theta}_0$, for the quadratic norm $\|v\|_P$ is given by

$$\boldsymbol{v} = \frac{1}{||P^{-1} \nabla J(\boldsymbol{\theta}_0)||_P} P^{-1} \nabla J(\boldsymbol{\theta}_0).$$

Solution: The task is

$$\begin{aligned}
\underset{\boldsymbol{v}}{\text{minimize}} \qquad & \boldsymbol{v}^T \nabla J, \\
\text{s.t.} \qquad & \boldsymbol{v}^T P \boldsymbol{v} = 1.
\end{aligned}$$

Using Lagrangian multipliers we obtain

$$L(\boldsymbol{v}) = \boldsymbol{v}^T \nabla J + \lambda (\boldsymbol{v}^T P \boldsymbol{v} - 1)$$

or

$$\nabla_v L = \nabla J + 2\lambda P \boldsymbol{v} = \boldsymbol{0}$$

or

$$\boldsymbol{v} = -\frac{1}{2\lambda} P^{-1} \nabla J.$$

Substituting the above in the constraint we get

$$\frac{1}{4\lambda^2} (\nabla J)^T P^{-1} P P^{-1} \nabla J = 1$$

or

$$2\lambda = \left((\nabla J)^T P^{-1} \nabla J \right)^{\frac{1}{2}} .$$

6.16. Justify why the convergence of Newton's iterative minimization method is relatively insensitive on the Hessian matrix.
Hint: Let P be a positive definite matrix. Define a change of variables,

$$\tilde{\boldsymbol{\theta}} = P^{\frac{1}{2}} \boldsymbol{\theta},$$

and carry gradient descent minimization based on the new variable.

Solution: Let

$$\tilde{\boldsymbol{\theta}} = P^{\frac{1}{2}}\boldsymbol{\theta}.$$

Note that $\|\boldsymbol{\theta}\|_p^2 := \boldsymbol{\theta}^T P \boldsymbol{\theta}$. Also, define,

$$\tilde{J}(\tilde{\boldsymbol{\theta}}) := J(P^{-\frac{1}{2}}\tilde{\boldsymbol{\theta}}) = J(\boldsymbol{\theta}).$$

We have that

$$\nabla_{\tilde{\theta}}(\tilde{J}(\tilde{\boldsymbol{\theta}})) = P^{-\frac{1}{2}}\nabla_{\theta}J(\boldsymbol{\theta}). \tag{9}$$

Then the gradient descent step for $\tilde{J}(\tilde{\boldsymbol{\theta}})$ will be

$$\Delta\tilde{\boldsymbol{\theta}} = -P^{-\frac{1}{2}}\nabla_{\theta}J(\boldsymbol{\theta}).$$

However,

$$\Delta\tilde{\boldsymbol{\theta}} = P^{\frac{1}{2}}\Delta\boldsymbol{\theta}.$$

or

$$\Delta\boldsymbol{\theta} = -P^{-1}\nabla_{\theta}J(\boldsymbol{\theta}).$$

Note that the latter is the step for the steepest descent direction with respect $\|\boldsymbol{\theta}\|_p$ norm. Set now

$$P = \nabla^2 J(\boldsymbol{\theta}_*),$$

where $\boldsymbol{\theta}_*$ is the optimum, and assume that close enough to the optimum the Hessian does not vary much. Then from 9, the Hessian of $\tilde{J}(\tilde{\boldsymbol{\theta}})$ will be equal to

$$P^{-\frac{1}{2}}\nabla^2 J(\boldsymbol{\theta}_*)P^{-\frac{1}{2}} \simeq I.$$

Thus the step for the Newton's method can equivalently been seen as a gradient descent with a cost whose Hessian matrix is of low condition number.

6.17. Show that the steepest descent direction, $\boldsymbol{v}$, of $J(\boldsymbol{\theta})$ at a point, $\boldsymbol{\theta}_0$, constrained to

$$\|\boldsymbol{v}\|_1 = 1,$$

is given by $\boldsymbol{e}_k$, where $\boldsymbol{e}_k$ is the standard basis vector in the direction, k, such that

$$|(\nabla J(\boldsymbol{\theta}_0))_k| > |(\nabla J(\boldsymbol{\theta}_0))_j|, \quad k \neq j.$$

Solution: We will use two ways to prove it. The second one is lengthier, but it is an exercise for the reader to get familiar with optimizing with respect to nondifferential functions via the use of the subdifferential notion.

a) It is known from linear algebra and the properties of norms, that

$$|\boldsymbol{v}^T\boldsymbol{a}| \leq \|\boldsymbol{v}\|_1\|\boldsymbol{a}\|_\infty,$$

where $\|\boldsymbol{a}\|_\infty = \max_i |a_i|$, $i = 1, 2, \ldots, l$. Indeed,

$$\begin{aligned} |\boldsymbol{v}^T\boldsymbol{a}| = \left|\sum_{i=1}^{l} v_i a_i\right| &\leq \sum_{i=1}^{l} |v_i||a_i| \leq \left(\sum_{i=1}^{l} |v_i|\right) \|\boldsymbol{a}\|_\infty \\ &= \|\boldsymbol{v}\|_1 \|\boldsymbol{a}\|_\infty. \end{aligned}$$

Hence,

$$-\|\boldsymbol{a}\|_\infty \leq \boldsymbol{v}^T\boldsymbol{a} \leq \|\boldsymbol{a}\|_\infty,$$

since $\|\boldsymbol{v}\|_1 = 1$. Thus the minimum is achieved if

$$\boldsymbol{v} = -\operatorname{sgn}(a_i)\boldsymbol{e}_i$$

where $\boldsymbol{e}_i$ the direction corresponding to the component associated with $\|\boldsymbol{a}\|_\infty$.

b) Since the ℓ_1 norm is not differentiable, the notion of subgradient will be mobilized. We have

$$\begin{aligned} \text{minimize} \qquad & \boldsymbol{v}^T\nabla J \\ \text{s.t.} \qquad & \|\boldsymbol{v}\|_1 = 1. \end{aligned}$$

The Lagrangian is

$$L(\boldsymbol{v}) = \boldsymbol{v}^T\nabla J + \lambda(\|\boldsymbol{v}\|_1 - 1)$$

or

$$\frac{\partial L(\boldsymbol{v})}{\partial \boldsymbol{v}} = \nabla J + \lambda \frac{\partial \|\boldsymbol{v}\|_1}{\partial \boldsymbol{v}}$$

or

$$\mathbf{0} \in \quad \nabla J + \lambda \frac{\partial \|\boldsymbol{v}\|_1}{\partial \boldsymbol{v}}, \tag{10}$$

where $\frac{\partial \|\boldsymbol{v}\|_1}{\partial \boldsymbol{v}}$ is the subdifferential set for $\|\boldsymbol{v}\|_1$. Then, taking into account the subgradient of the absolute value, (10) can be written component wise as

$$(\nabla J)_j = -\lambda \begin{cases} \{1\}, & v_j > 0 \\ \{-1\}, & v_j < 0 \\ a \in [-1, 1], & v_j = 0 \end{cases} \qquad j = 1, 2, \ldots, l. \tag{11}$$

Since $\boldsymbol{v} \neq \mathbf{0}$, λ has to be chosen so as to satisfy one of these equations, otherwise there is no solution. We choose $j \to k$ corresponding to the largest component $|(\nabla J)_k| > |(\nabla J)_j|$, $j \neq k$. Thus

$$\boldsymbol{v} : v_j = 0, \ j \neq k, \text{ and } v_k > 0.$$

Then

$$\lambda = -(\nabla J)_k,$$

which guarantees that

$$\boldsymbol{v}^T \nabla J = -|(\nabla J)_k|^2,$$

which for all available choices, corresponds to the steepest descent.

Note that this choice also guarantees that all equations in (11) are satisfied, since in this case there is always a value a_j, such as

$$(\nabla J)_j = -a_j \lambda, \; j \neq k.$$

with $a_j \in [-1, 1]$.

6.18. Show that the TLS solution is given by,

$$\hat{\boldsymbol{\theta}} = \left(X^T X - \bar{\sigma}_{l+1}^2 I \right)^{-1} X^T \boldsymbol{y},$$

where $\bar{\sigma}_{l+1}$ is the smallest singular value of $[X \vdots \boldsymbol{y}]$.

Solution: By the respective definition we have that,

$$\begin{bmatrix} X^T \\ \boldsymbol{y}^T \end{bmatrix} \begin{bmatrix} X \vdots \boldsymbol{y} \end{bmatrix} \bar{\boldsymbol{v}}_{l+1} = \bar{\sigma}_{l+1}^2 \bar{\boldsymbol{v}}_{l+1},$$

or

$$\begin{bmatrix} X^T X & X^T \boldsymbol{y} \\ \boldsymbol{y}^T X & \boldsymbol{y}^T \boldsymbol{y} \end{bmatrix} \bar{\boldsymbol{v}}_{l+1} = \bar{\sigma}_{l+1}^2 \bar{\boldsymbol{v}}_{l+1},$$

and after dividing by $-\bar{v}_{l+1}^{(l+1)}$,

$$\begin{bmatrix} X^T X & X^T \boldsymbol{y} \\ \boldsymbol{y}^T X & \boldsymbol{y}^T \boldsymbol{y} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\theta}}_{TLS} \\ -1 \end{bmatrix} = \bar{\sigma}_{l+1}^2 \begin{bmatrix} \hat{\boldsymbol{\theta}}_{TLS} \\ -1 \end{bmatrix},$$

or

$$X^T X \hat{\boldsymbol{\theta}}_{TLS} - X^T \boldsymbol{y} = \bar{\sigma}_{l+1}^2 \hat{\boldsymbol{\theta}}_{TLS}$$

or

$$\left(X^T X - \bar{\sigma}_{l+1}^2 I \right) \hat{\boldsymbol{\theta}}_{TLS} = X^T \boldsymbol{y},$$

which proves the claim.

6.19. Given a set of centered data points, $(y_n, \boldsymbol{x}_n) \in \mathbb{R}^{l+1}$, derive a hyperplane

$$\boldsymbol{a}^T \boldsymbol{x} + y = 0,$$

which crosses the origin, such as the total square distance of all the points from it to be minimum.

Solution: Let

$$\boldsymbol{w}^T = [\boldsymbol{a}^T, 1]^T.$$

and

$$\boldsymbol{z}_n = (\boldsymbol{x}_n, y_n) \in \mathbb{R}^{l+1},\ n = 1, 2, \ldots, N.$$

Then we will search for the $\boldsymbol{w}$ given by

$$\begin{aligned} \underset{\boldsymbol{w}}{\text{minimize}} \qquad & \sum_{n=1}^{N} |\boldsymbol{w}^T \boldsymbol{z}_n|^2 \\ \text{s.t.} \qquad & \|\boldsymbol{w}\|^2 = 1. \end{aligned}$$

The constraint is used since we are only interested in the direction. If $\|\boldsymbol{w}\| \neq 1$, then the square distance of a point from the hyperplane would be equal to,

$$\frac{|\boldsymbol{w}^T \boldsymbol{z}_n|^2}{\|\boldsymbol{w}\|^2}.$$

Let

$$Z = \begin{bmatrix} \boldsymbol{x}_1^T & y_1 \\ \vdots & \vdots \\ \boldsymbol{x}_N^T & y_N \end{bmatrix} = \begin{bmatrix} \boldsymbol{z}_1^T \\ \vdots \\ \boldsymbol{z}_N^T \end{bmatrix}.$$

Then,

$$\sum_{n=1}^{N} |\boldsymbol{w}^T \boldsymbol{z}_n|^2 = \boldsymbol{w}^T Z^T Z \boldsymbol{w}.$$

Hence, the Lagrangian becomes

$$L(\boldsymbol{w}, \bar{\lambda}) = \boldsymbol{w}^T Z^T Z \boldsymbol{w} - \bar{\lambda}(\boldsymbol{w}^T \boldsymbol{w} - 1)$$

or

$$\begin{aligned} \frac{\partial L(\boldsymbol{w}, \bar{\lambda})}{\partial \boldsymbol{w}} &= 2(Z^T Z \boldsymbol{w} - \bar{\lambda} \boldsymbol{w}) = \boldsymbol{0} \\ &\Rightarrow Z^T Z \boldsymbol{w} = \bar{\lambda} \boldsymbol{w}. \end{aligned}$$

Thus, $\bar{\lambda}$ is an eigenvalue of ZZ^T. Let $\bar{\lambda}_1 \geq \bar{\lambda}_2 \geq \ldots \geq \bar{\lambda}_l > \bar{\lambda}_{l+1} > 0$. That is, we assume that ZZ^T is full rank and also the smallest eigenvalue is single. Then,

$$\boldsymbol{w} = \bar{\boldsymbol{v}}_{l+1}.$$

Note that in this case,

$$\boldsymbol{w}^T Z^T Z \boldsymbol{w} = \bar{\lambda}_{\min} \|\bar{\boldsymbol{v}}_{l+1}\|^2 = \bar{\lambda}_{\min},$$

since $\bar{\boldsymbol{v}}_{l+1}$ is the normalized eigenvector, in order to guarantee the constrain. For any other choice of $\bar{\lambda}$, the value of loss would be larger, hence not minimum. Thus, the hyperplane is defined by the eigenvector corresponding to the minimum single singular value of

$$[X|y].$$

Note that, in general, there is not a single solution.

Solutions To Problems of Chapter 7

7.1. Show that the Bayesian classifier is optimal, in the sense that it minimizes the probability of error.

Hint: Consider a classification task of M classes and start with the probability of correct label prediction, $P(C)$. Then the probability of error will be $P(e) = 1 - P(C)$.

Solution: Let $P(C)$ be the probability of correct classification. Then

$$P(C) \,=\, \sum_{i=1}^{M} P(\boldsymbol{x} \in R_i, \omega_i) \,=\, \sum_{i=1}^{M} P(\omega_i) P(\boldsymbol{x} \in R_i|\omega_i),$$

or

$$P(C) \,=\, \sum_{i=1}^{M} P(\omega_i) \int_{R_i} P(\boldsymbol{x}|\omega_i) d\boldsymbol{x} \,=\, \sum_{i=1}^{M} \int_{R_i} P(\omega_i) p(\boldsymbol{x}|\omega_i) d\boldsymbol{x},$$

or

$$P(C) \,=\, \sum_{i=1}^{M} \int_{R_i} P(\omega_i|\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x}.$$

For minimum classification error $P(e)$, $P(C)$ must be maximum $(P(C) + P(e) = 1)$. Thus $P(C)$ is maximized if the regions R_i are chosen so that in each region the corresponding integrals, which are all positive, have the maximum possible value. That is,

$$R_i \,:\, P(\omega_i|\boldsymbol{x}) p(\boldsymbol{x}) \,>\, P(\omega_j|\boldsymbol{x}) p(\boldsymbol{x}) \quad \forall\, i \neq j$$

or

$$R_i \,:\, P(\omega_i|\boldsymbol{x}) \,>\, P(\omega_j|\boldsymbol{x}) \quad \forall\, i \neq j$$

7.2. Show that if the data follow the Gaussian distribution in an M class task, with equal covariance matrices in all classes, the regions formed by the Bayesian classifier are convex.

Solution: Consider two points, $\boldsymbol{x}_1,\ \boldsymbol{x}_2$, lying in R_i. Then any point lying on the line segment, which connects these two points, can be written as

$$\boldsymbol{x} = a\boldsymbol{x}_1 + (1-a)\boldsymbol{x}_2,\ a \in [0,1].$$

In this case, the decision is taken according to a linear discriminant function, i.e,

$$\boldsymbol{\theta}_i^T \boldsymbol{x} + \theta_{0i} = a\left(\boldsymbol{\theta}_i^T \boldsymbol{x}_1 + \theta_{0i}\right) + (1-a)\left(\boldsymbol{\theta}_i^T \boldsymbol{x}_2 + \theta_{0i}\right).$$

However, since by the definition of the partition,

$$\boldsymbol{\theta}_i^T \boldsymbol{x}_1 + \theta_{0i} > \boldsymbol{\theta}_j^T \boldsymbol{x}_1 + \theta_{0j},\ j \neq i,$$

and

$$\boldsymbol{\theta}_i^T \boldsymbol{x}_2 + \theta_{0i} > \boldsymbol{\theta}_j^T \boldsymbol{x}_2 + \theta_{0j},\ j \neq i,$$

it turns out that $\boldsymbol{x} \in R_i$, which proves the claim.

7.3. Derive the form of the Bayesian classifier for the case of two equiprobable classes, when the data follow the Gaussian distribution of the same covariance matrix. Furthermore, derive the equation that describes the LS linear classifier. Compare and comment on the results.

Solution: For the scenario of this problem, we already know that the Bayesian classifier is equivalent to the minimum Mahalanobis distance classifier, i.e.,

$$(\boldsymbol{x} - \boldsymbol{\mu}_1)^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_1) > \ (<) (\boldsymbol{x} - \boldsymbol{\mu}_2)^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_2).$$

Expanding and rearranging the terms, we obtain that the decision surface is of the following form

$$\begin{aligned} g(\boldsymbol{x}) &= -\boldsymbol{x}^T \Sigma^{-1} \boldsymbol{x} + 2\boldsymbol{x}^T \Sigma^{-1} \boldsymbol{\mu}_1 - \boldsymbol{\mu}_1^T \Sigma^{-1} \boldsymbol{\mu}_1 \\ &\quad + \boldsymbol{x}^T \Sigma^{-1} \boldsymbol{x} - 2\boldsymbol{x}^T \Sigma^{-1} \boldsymbol{\mu}_2 + \boldsymbol{\mu}_2^T \Sigma^{-1} \boldsymbol{\mu}_2 = 0. \end{aligned}$$

Rearranging the terms one can easily obtain that

$$g(\boldsymbol{x}) = \boldsymbol{\theta}^T (\boldsymbol{x} - \boldsymbol{\theta}_0) = 0,$$

where

$$\boldsymbol{\theta} := \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \quad \boldsymbol{\theta}_0 \ = \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2).$$

Let us now turn our attention to the LS solution. By extending the involved training feature vectors by one, in order to accommodate the bias term, the LS classifier is given by the solution of the following set of linear equations:

$$\frac{1}{N}\left(\sum_{n=1}^{N} \begin{bmatrix} \boldsymbol{x}_n \\ 1 \end{bmatrix} [\boldsymbol{x}_n^T, 1]\right) \begin{bmatrix} \boldsymbol{\theta} \\ \theta_0 \end{bmatrix} = \frac{1}{N} \sum_{n=1}^{N} y_n \begin{bmatrix} \boldsymbol{x}_n \\ 1 \end{bmatrix},$$

where we have divided both sides by $1/N$. For large enough N and for equiprobable classes, we can assume that both classes are represented with equal number of points $N/2$ and by the law of large numbers we can equate sample means with expectations. Then the previous formula can be written as

$$\begin{bmatrix} R & \mathbb{E}[\mathbf{x}] \\ \mathbb{E}[\mathbf{x}^T] & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{\theta} \\ \theta_0 \end{bmatrix} = \begin{bmatrix} \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\ 0 \end{bmatrix} \tag{1}$$

where the zero comes from $\sum_n y_n$, in which half of the terms are $+1$ and the rest -1. In the previous formula we have that

$$R := \mathbb{E}[\mathbf{x}\mathbf{x}^T] = \Sigma_b + \boldsymbol{\mu}\boldsymbol{\mu}^T,$$
$$\boldsymbol{\mu} := \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) = \mathbb{E}[\mathbf{x}] = P(\omega_1)\,\mathbb{E}[\mathbf{x}|\omega_1] + P(\omega_2)\,\mathbb{E}[\mathbf{x}|\omega_2],$$

and Σ_b is the covariance matrix for the two classes together; that is, around the overall mean value $\boldsymbol{\mu}$. Substituting the previous definitions in (1), and solving the system of equations, it is trivial to see that

$$\theta_0 = -\boldsymbol{\theta}^T\boldsymbol{\mu},$$
$$\boldsymbol{\theta} = \frac{1}{2}\Sigma_b^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

Combining the previous finding we obtain the LS optimal linear classifier as

$$g(\boldsymbol{x}) = \boldsymbol{\theta}^T\boldsymbol{x} + \theta_0 = \boldsymbol{\theta}^T(\boldsymbol{x} - \boldsymbol{\theta}_0) = 0,$$

where

$$\boldsymbol{\theta} = \Sigma_b^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \quad \boldsymbol{\theta}_0 \equiv \boldsymbol{\mu} = \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2).$$

However,

$$\Sigma_b = P(\omega_1)\,\mathbb{E}\left[(\mathbf{x} - \boldsymbol{\mu}_1)(\mathbf{x} - \boldsymbol{\mu}_1)^T\right] + P(\omega_2)\,\mathbb{E}\left[(\mathbf{x} - \boldsymbol{\mu}_2)(\mathbf{x} - \boldsymbol{\mu}_2)^T\right] = \Sigma.$$

That is, under the adopted assumptions, the LS classifier tends asymptotically to the optimal Bayesian classifier.

7.4. Show that the ML estimate of the covariance matrix of a Gaussian distribution, based on N i.i.d, observations, $\boldsymbol{x}_n$, $n = 1, 2, \ldots, N$, is given by,

$$\hat{\Sigma}_{ML} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n - \hat{\boldsymbol{\mu}}_{ML})(\boldsymbol{x}_n - \hat{\boldsymbol{\mu}}_{ML})^T,$$

where

$$\hat{\boldsymbol{\mu}}_{ML} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n.$$

Solution: We shall focus on the simplest case where $\Sigma = \sigma^2 I$. The log-likelihood function is

$$L(\boldsymbol{\mu}, \sigma^2) \;=\; \sum_{n=1}^{N}\ln p(\boldsymbol{x}_n; \boldsymbol{\mu}, \sigma^2),$$
$$= -\frac{N}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu})^T(\boldsymbol{x}_n - \boldsymbol{\mu}).$$

where constants have been omitted. The unknown set of parameters is now $\boldsymbol{\theta}^T = [\boldsymbol{\mu}^T, \sigma^2]$, hence,

$$\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \begin{bmatrix} \frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\mu}} \\ \frac{\partial L(\boldsymbol{\theta})}{\partial \sigma^2} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2}\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu}) \\ -\frac{N}{2\sigma^2} + \frac{\sum_{n=1}^{N}(\boldsymbol{x}_n-\boldsymbol{\mu})^T(\boldsymbol{x}_n-\boldsymbol{\mu})}{2\sigma^4} \end{bmatrix} = 0.$$

Solving the above system w.r. to $\boldsymbol{\mu}$ and σ^2 results in,

$$\hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{k=1}^{N}\boldsymbol{x}_n,$$

and

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu})^T(\boldsymbol{x}_n - \boldsymbol{\mu}).$$

7.5. Prove that the covariance estimate

$$\hat{\Sigma} = \frac{1}{N-1}\sum_{k=1}^{N}(\boldsymbol{x}_k - \hat{\boldsymbol{\mu}})(\boldsymbol{x}_k - \hat{\boldsymbol{\mu}})^T$$

corresponds to an unbiased estimator, where,

$$\hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{k=1}^{N}\boldsymbol{x}_k.$$

Solution: We have that

$$\begin{aligned}
\mathbb{E}[\hat{\Sigma}] &= \frac{1}{N-1}\sum_{k=1}^{N}\mathbb{E}\left[\left((\mathbf{x}_k - \boldsymbol{\mu}) - (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})\right)\left((\mathbf{x}_k - \boldsymbol{\mu}) - (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})\right)^T\right] \\
&= \frac{1}{N-1}\sum_{k=1}^{N}\mathbb{E}\left[(\mathbf{x}_k - \boldsymbol{\mu})(\mathbf{x}_k - \boldsymbol{\mu})^T\right] + \frac{1}{N-1}\sum_{k=1}^{N}\mathbb{E}\left[(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})^T\right] - \\
&\quad \frac{1}{N-1}\sum_{k=1}^{N}\mathbb{E}\left[(\mathbf{x}_k - \boldsymbol{\mu})(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})^T\right] - \\
&\quad \frac{1}{N-1}\sum_{k=1}^{N}\mathbb{E}\left[(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})(\mathbf{x}_k - \boldsymbol{\mu})^T\right]. \qquad (2)
\end{aligned}$$

However,

$$\begin{aligned}
\mathbb{E}\left[(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})^T\right] &= \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \boldsymbol{\mu})\frac{1}{N}\sum_{j=1}^{N}(\mathbf{x}_j - \boldsymbol{\mu})^T\right] \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T\right] \\
&= \frac{1}{N^2}N\Sigma = \frac{1}{N}\Sigma, \qquad (3)
\end{aligned}$$

where independence among the samples has been assumed, i.e.,

$$\mathbb{E}\left[(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_j - \boldsymbol{\mu})^T\right] = \delta_{ij}\Sigma.$$

Following a similar path, we end up with

$$\mathbb{E}\left[(\mathbf{x}_k - \boldsymbol{\mu})(\hat{\boldsymbol{\upmu}} - \boldsymbol{\mu})^T\right] = \mathbb{E}\left[(\hat{\boldsymbol{\upmu}} - \boldsymbol{\mu})(\mathbf{x}_k - \boldsymbol{\mu})^T\right] = \frac{1}{N}\Sigma. \tag{4}$$

Also,

$$\mathbb{E}\left[(\mathbf{x}_k - \boldsymbol{\mu})(\mathbf{x}_k - \boldsymbol{\mu})^T\right] = \Sigma. \tag{5}$$

Combining Eqs (2)-(5), we get

$$E[\hat{\Sigma}] = \frac{N}{N-1}\Sigma + \frac{1}{N-1}\Sigma - \frac{1}{N-1}\Sigma - \frac{1}{N-1}\Sigma = \Sigma.$$

Hence the estimator is an unbiased one.

7.6. Show that the derivative of the logistic link function is given by

$$\frac{d\sigma(t)}{dt} = \sigma(t)\big(1 - \sigma(t)\big).$$

Solution: We have that

$$\sigma(t) = \frac{1}{1 + \exp(-t)}.$$

Thus

$$\frac{d\sigma(t)}{dt} = \frac{\exp(-t)}{\big(1 + \exp(-t)\big)^2} = \sigma(t)\frac{\exp(-t)}{(1 + \exp(-t))} = \sigma(t)\left(1 - \sigma(t)\right).$$

7.7. Derive the gradient of the negative log-likelihood function associated with the two-class logistic regression.

Solution: The log-likelihood is given by

$$L(\boldsymbol{\theta}) = -\sum_{n=1}^{N}\big(y_n \ln s_n + (1 - y_n)\ln(1 - s_n)\big),$$

where

$$s_n = \sigma(\boldsymbol{\theta}^T \boldsymbol{x}_n).$$

Let

$$a_n := y_n \ln s_n + (1 - y_n)\ln(1 - s_n).$$

Hence,

$$\nabla_{\boldsymbol{\theta}} a_n = y_n \frac{1}{s_n}\left(s_n(1 - s_n)\right)\nabla_{\boldsymbol{\theta}} t_n - (1 - y_n)\frac{1}{1 - s_n}\left(s_n(1 - s_n)\right)\nabla_{\boldsymbol{\theta}} t_n,$$

where $t_n := \boldsymbol{\theta}^T \boldsymbol{x}_n$, or

$$\nabla_{\boldsymbol{\theta}} a_n = y_n(1-s_n)\boldsymbol{x}_n - (1-y_n)s_n\boldsymbol{x}_n = (y_n - s_n)\boldsymbol{x}_n.$$

Hence,

$$\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = \sum_{n=1}^{N}(s_n - y_n)\boldsymbol{x}_n = X^T(\boldsymbol{s} - \boldsymbol{y}).$$

7.8. Derive the Hessian matrix of the negative log-likelihood function associated with the two-class logistic regression.

Solution: From the previous problem, we have already derived the gradient, i.e.,

$$\nabla_{\boldsymbol{\theta}} a_n = (y_n - s_n)\boldsymbol{x}_n.$$

Hence, the i, j element of the respective Hessian matrix is given by,

$$\left[\nabla^2_{\boldsymbol{\theta}} a_n\right]_{ij} = \frac{\partial}{\partial \theta_i}\Big((s_n - y_n)x_n(j)\Big) = \left(s_n(1-s_n)\right)x_n(i)x_n(j),$$

which can equivalently be written as,

$$\left[\nabla^2_{\boldsymbol{\theta}} a_n\right]_{ij} = \left[\boldsymbol{x}_n r_n \boldsymbol{x}_n^T\right]_{ij},$$

where

$$r_n = s_n(1-s_n).$$

Hence,

$$\nabla^2_{\boldsymbol{\theta}} a_n = \boldsymbol{x}_n r_n \boldsymbol{x}_n^T,$$

which leads to

$$\nabla^2_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = \sum_{n=1}^{N} \boldsymbol{x}_n r_n \boldsymbol{x}_n^T = X^T R_n X,$$

where R_n is the diagonal matrix comprising, $r_n, \ n = 1, 2, \ldots, N$.

7.9. Show that the Hessian matrix of the negative log-likelihood function of the two-class logistic regression is a positive definite matrix.

Solution. We have seen that

$$\nabla^2_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = X^T R_n X.$$

Thus for any vector $\boldsymbol{x} \in \mathbb{R}^l$, the following is true,

$$\boldsymbol{x}^T X^T R_n X \boldsymbol{x} = \boldsymbol{y}^T R_n \boldsymbol{y} > 0,$$

due to the positive definite nature of R_n, where $\boldsymbol{y} := X\boldsymbol{x}$. This is the definition of a positive definite matrix.

7.10. Show that if

$$\phi_m = \frac{\exp(t_m)}{\sum_{j=1}^{M} \exp(t_j)},$$

the derivative with respect to $t_j,\ j = 1, 2, \ldots, M$ is given by,

$$\frac{\partial \phi_m}{\partial t_j} = \phi_m(\delta_{mj} - \phi_j).$$

Solution: a) Let $j = m$. Then we have

$$\frac{\partial \phi_m}{\partial t_m} = \frac{\exp(t_m) \sum_{j=1}^{M} \exp(t_j) - \exp(t_m)\exp(t_m)}{\left(\sum_{j=1}^{M} \exp(t_j)\right)^2},$$

or

$$\frac{\partial \phi_m}{\partial t_m} = \phi_m(1 - \phi_m).$$

Now, let $t_j \neq t_m$. Then, in this case,

$$\frac{\partial \phi_m}{\partial t_j} = \frac{-\exp(t_j)\exp(t_m)}{\left(\sum_{j=1}^{M} \exp(t_j)\right)^2} = -\phi_m \phi_j.$$

Thus,

$$\frac{\partial \phi_m}{\partial t_j} = \phi_m(\delta_{mj} - \phi_j),\ j = 1, 2, \ldots, M.$$

7.11. Derive the gradient of the negative log-likelihood for the multiclass logistic regression case.

Solution: Let

$$a_n := \sum_{m=1}^{M} y_{nm} \ln \phi_{nm}.$$

Then we have that,

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}_j} a_n &= \sum_{m=1}^{M} y_{nm} \frac{1}{\phi_{nm}} \frac{\partial \phi_{nm}}{\partial t_j} \frac{\partial t_j}{\partial \boldsymbol{\theta}_j} \\
&= \sum_{m=1}^{M} y_{nm} \frac{1}{\phi_{nm}} \phi_{nm}(\delta_{mj} - \phi_{nj}) \boldsymbol{x}_n \\
&= (y_{nj} - \phi_{nj}) \boldsymbol{x}_n,
\end{aligned}$$

since

$$\sum_{m=1}^{M} y_{nm} = 1.$$

Hence,

$$\nabla_{\boldsymbol{\theta}_j} L(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M) = \sum_{n=1}^{N} (\phi_{nj} - y_{nj}) \boldsymbol{x}_n.$$

7.12. Derive the j, k block element of the Hessian matrix of the negative log-likelihood function for the multiclass logistic regression.

Solution: From the previous problem, we have already obtained that,

$$\nabla_{\boldsymbol{\theta}_j} L(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M) = \sum_{n=1}^{N} (\phi_{nj} - y_{nj}) \boldsymbol{x}_n.$$

Hence,

$$\begin{aligned} \nabla_{\boldsymbol{\theta}_k} \nabla_{\boldsymbol{\theta}_j} L(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M) &= \sum_{n=1}^{N} \boldsymbol{x}_n \nabla^T_{\boldsymbol{\theta}_k} (\phi_{nj} - y_{nj}) \\ &= \sum_{n=1}^{N} \phi_{nj} (\delta_{kj} - \phi_{nk}) \boldsymbol{x}_n \boldsymbol{x}_n^T, \end{aligned}$$

where the result of Problem 7.10 has been used.

7.13. Consider the so-called Rayleigh ratio,

$$R = \frac{\boldsymbol{\theta}^T A \boldsymbol{\theta}}{||\boldsymbol{\theta}||^2},$$

where A is a symmetric matrix. Show that R is maximized, with respect to $\boldsymbol{\theta}$, if $\boldsymbol{\theta}$ is the eigenvector corresponding the maximum eigenvalue of A.

Solution: The maximization task is equivalent to

$$\begin{aligned} \text{maximize w.r. } \boldsymbol{\theta} \quad &: \quad \boldsymbol{\theta}^T A \boldsymbol{\theta}, \\ \text{s.t} \quad & \quad ||\boldsymbol{\theta}||^2 = 1, \end{aligned}$$

since we are only interested in the direction and the ratio does not depend on the norm of $\boldsymbol{\theta}$. Using Lagrange multipliers, the Lagrangian becomes,

$$L(\boldsymbol{\theta}, \lambda) = \boldsymbol{\theta}^T A \boldsymbol{\theta} - \lambda(||\boldsymbol{\theta}||^2 - 1),$$

and taking the gradient with respect to $\boldsymbol{\theta}$ and equating to zero, we obtain,

$$A\boldsymbol{\theta} = \lambda \boldsymbol{\theta}.$$

In other words, $\boldsymbol{\theta}$ is an eigenvector of A. In order to maximize the ratio, we choose the one which corresponds to the maximum eigenvalue, leading to

$$R = \boldsymbol{\theta}^T A \boldsymbol{\theta} = \lambda ||\boldsymbol{\theta}||^2 = \lambda.$$

7.14. Consider the generalized Rayleigh quotient,

$$R_g = \frac{\boldsymbol{\theta}^T B \boldsymbol{\theta}}{\boldsymbol{\theta}^T A \boldsymbol{\theta}}.$$

where A and B are a symmetric matrices. Show that R_g is maximized with respect to $\boldsymbol{\theta}$, if $\boldsymbol{\theta}$ is the eigenvector which corresponds to the maximum eigenvalue of $A^{-1}B$, assuming that the inversion is possible.

Solution: For the case of our problem (generalized Rayleigh quotient), let

$$\boldsymbol{y} := A^{1/2}\boldsymbol{\theta} \quad \Rightarrow \quad \boldsymbol{\theta} = A^{-1/2}\boldsymbol{y}.$$

Then the problem becomes equivalent with maximizing

$$\max_{\boldsymbol{y}} \frac{\boldsymbol{y}^T A^{-1/2} B A^{-1/2} \boldsymbol{y}}{\boldsymbol{y}^T \boldsymbol{y}},$$

where the symmetry of A has been taken into account. From the previous problem, the above is maximized if $\boldsymbol{y}$ is the eigenvector corresponding to the largest eigenvalue λ, i.e.,

$$A^{-1/2} B A^{-1/2} \boldsymbol{y} = \lambda \boldsymbol{y},$$

and finally, by replacing $\boldsymbol{y}$ by $\boldsymbol{\theta}$, if $\boldsymbol{\theta}$ is chosen to satisfy

$$B\boldsymbol{\theta} = \lambda A \boldsymbol{\theta}.$$

Equivalently we can solve the following eigenvalue task

$$A^{-1}B\boldsymbol{\theta} = \lambda \boldsymbol{\theta}.$$

The corresponding maximum value is

$$\frac{\boldsymbol{\theta}^T B \boldsymbol{\theta}}{\boldsymbol{\theta}^T A \boldsymbol{\theta}} = \lambda \frac{\boldsymbol{\theta}^T A \boldsymbol{\theta}}{\boldsymbol{\theta}^T A \boldsymbol{\theta}} = \lambda$$

which justifies the choice of the maximum eigenvalue.

7.15. Show that the between-classes scatter matrix Σ_b for an M class problem is of rank $M-1$.

Solution: We will show it for $M=2$ and $M=3$ and the result is easily generalized. We use $P(\omega_i) = P_i$. For $M=2$, we have $P_1 + P_2 = 1$. Thus,

$$\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0 = \boldsymbol{\mu}_1 - P_1\boldsymbol{\mu}_1 - P_2\boldsymbol{\mu}_2 = P_2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

Similarly we can show that

$$\boldsymbol{\mu}_2 - \boldsymbol{\mu}_0 = -P_1(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

Thus

$$\begin{aligned}\Sigma_b &= P_1(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)^T \\ &\quad +P_2(\boldsymbol{\mu}_2-\boldsymbol{\mu}_0)(\boldsymbol{\mu}_2-\boldsymbol{\mu}_0)^T \\ &= P_1P_2^2(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T \\ &\quad +P_2P_1^2(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T \\ &= P_1P_2(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T. \qquad (6)\end{aligned}$$

Thus Σ_b is a rank one matrix.

Also, for $M=3$, the vectors entering into the summation are,

$$\boldsymbol{\mu}_1-\boldsymbol{\mu}_0\,,\ \boldsymbol{\mu}_2-\boldsymbol{\mu}_0\,,\ \boldsymbol{\mu}_3-\boldsymbol{\mu}_0,$$

or

$$\begin{aligned}\boldsymbol{\mu}_1-\boldsymbol{\mu}_0 &= (1-P_1)\boldsymbol{\mu}_1-P_2\boldsymbol{\mu}_2-P_3\boldsymbol{\mu}_3 = P_2(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)+P_3(\boldsymbol{\mu}_1-\boldsymbol{\mu}_3),\\ \boldsymbol{\mu}_2-\boldsymbol{\mu}_0 &= (1-P_2)\boldsymbol{\mu}_2-P_1\boldsymbol{\mu}_1-P_3\boldsymbol{\mu}_3 = P_1(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)+P_3(\boldsymbol{\mu}_2-\boldsymbol{\mu}_3),\\ \boldsymbol{\mu}_3-\boldsymbol{\mu}_0 &= (1-P_3)\boldsymbol{\mu}_3-P_1\boldsymbol{\mu}_1-P_2\boldsymbol{\mu}_2 = P_1(\boldsymbol{\mu}_3-\boldsymbol{\mu}_1)+P_2(\boldsymbol{\mu}_3-\boldsymbol{\mu}_2),\end{aligned}$$

From the above, it is obvious that

$$P_1(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)+P_2(\boldsymbol{\mu}_2-\boldsymbol{\mu}_0) = -P_3(\boldsymbol{\mu}_3-\boldsymbol{\mu}_0).$$

Thus, again two of them produce the third. This is easily generalized to M, that is,

$$P_1(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)+\ldots+P_{M-1}(\boldsymbol{\mu}_{M-1}-\boldsymbol{\mu}_0) = -P_M(\boldsymbol{\mu}_M-\boldsymbol{\mu}_0)$$

Hence $\sum_{i=1}^M P_i(\boldsymbol{\mu}_i-\boldsymbol{\mu}_0)(\boldsymbol{\mu}_i-\boldsymbol{\mu}_0)^T$ is of rank $M-1$

7.16. Derive the arithmetic rule for combination, by minimizing the average KL divergence.

Solution: Let the output from each classifier be $P_j(\omega_i|\boldsymbol{x}) := P_j(\omega_i)$ and let $P(\omega_i|\boldsymbol{x}) := P(\omega_i)$ be the output of the combiner to be estimated.

Minimizing the average Kullback-Leibler divergence and employing the obvious constraint

$$\sum_{i=1}^{M} P(\omega_i) = 1, \qquad (7)$$

the corresponding Lagrangian becomes

$$\frac{1}{L}\sum_{j=1}^{L}\sum_{i=1}^{M} P_j(\omega_i)\ln P_j(\omega_i) - \frac{1}{L}\sum_{j=1}^{L}\sum_{i=1}^{M} P_j(\omega_i)\ln P(\omega_i) - \lambda\left(\sum_{i=1}^{M} P(\omega_i)-1\right). \qquad (8)$$

Taking the derivative with respect to $P(\omega_i)$ and equating to zero, it turns out that

$$P(\omega_i) = -\frac{1}{\lambda L}\sum_{j=1}^{L} P_j(\omega_i). \tag{9}$$

Substituting into the constraint equation, we obtain

$$\lambda = -\frac{1}{L}\sum_{i=1}^{M}\sum_{j=1}^{L} P_j(\omega_i), \tag{10}$$

and finally,

$$P(\omega_i) = \frac{1}{L}\sum_{j=1}^{L} P_j(\omega_i). \tag{11}$$

7.17. Derive the product rule via the minimization of the Kullback-Leibler divergence, as pointed out in the text.

Solution: We will estimate $P(\omega_i)$ by minimizing the Kullback-Leibler average divergence, i.e,

$$\frac{1}{L}\sum_{j=1}^{L}\sum_{i=1}^{M} P(\omega_i)\ln\frac{P(\omega_i)}{P_j(\omega_i)}, \tag{12}$$

subject to the constraint

$$\sum_{i=1}^{M} P(\omega_i) = 1. \tag{13}$$

Employing Lagrange multipliers, the above problem becomes equivalent with minimizing the Lagrangian

$$\frac{1}{L}\sum_{j=1}^{L}\sum_{i=1}^{M} P(\omega_i)\ln\frac{P(\omega_i)}{P_j(\omega_i)} - \lambda\left(\sum_{i=1}^{M} P(\omega_i) - 1\right). \tag{14}$$

Taking the derivative with respect to $P(\omega_i)$ we get

$$\frac{1}{L}\sum_{j=1}^{L}\ln P(\omega_i) + 1 - \frac{1}{L}\sum_{j=1}^{L}\ln P_j(\omega_i) - \lambda = 0, \tag{15}$$

and solving the above we have

$$\begin{aligned} P(\omega_i) &= \exp\{\lambda-1\}\exp\{\frac{1}{L}\sum_{j=1}^{L}\ln P_j(\omega_i)\} \\ &= \exp\{\lambda-1\}\exp\left(\ln\prod_j\left(P_j(\omega_i)\right)^{\frac{1}{L}}\right) \\ &= \exp\{\lambda-1\}\prod_j\left(P_j(\omega_i)\right)^{\frac{1}{L}}. \end{aligned} \tag{16}$$

Substituting the above in the constraint equation it turns out that,

$$\exp\{\lambda - 1\} = \frac{1}{\sum_{i=1}^{M} \prod_{j=1}^{L} \left(P_j(\omega_i)\right)^{\frac{1}{L}}} \tag{17}$$

Hence,

$$P(\omega_i) = \frac{\prod_{j=1}^{L} (P_j(\omega_i))^{\frac{1}{L}}}{\sum_{i=1}^{M} \prod_{j=1}^{L} (P_j(\omega_i))^{\frac{1}{L}}}. \tag{18}$$

7.18. Show that the error rate on the training set of the final classifier, obtained by boosting, tends to zero exponentially fast.

Solution The error rate on the training data set is given by

$$P_e^N = \frac{1}{N} \sum_{n=1}^{N} I(1 - y_n f(\boldsymbol{x}_n)) \leq \frac{1}{N} \sum_{n=1}^{N} \exp(-y_n F(\boldsymbol{x}_n)),$$

which by the definition of the combined classifier is written as

$$\begin{aligned} \frac{1}{N} \sum_{n=1}^{N} \exp(-y_n F(\boldsymbol{x}_n)) &= \frac{1}{N} \sum_{n=1}^{N} \exp\left(-y_n \sum_{k=1}^{K} a_k \phi(\boldsymbol{x}_n; \boldsymbol{\theta}_k)\right) \\ &= \sum_{n=1}^{N} \frac{1}{N} \prod_{k=1}^{K} \exp(-y_n a_k \phi(\boldsymbol{x}_n; \boldsymbol{\theta}_k)). \end{aligned} \tag{19}$$

However,

$$\begin{aligned} w_n^{(K+1)} &= \frac{w_n^{(K)} \exp(-y_n a_K \phi(\boldsymbol{x}_n; \boldsymbol{\theta}_K))}{Z_K} \\ &= \frac{w_n^{(K-1)} \exp(-y_n a_{K-1} \phi(\boldsymbol{x}_n; \boldsymbol{\theta}_{K-1})) \exp(-y_n a_K \phi(\boldsymbol{x}_n; \boldsymbol{\theta}_K))}{Z_{K-1} Z_K} \\ &= \frac{\prod_{k=1}^{K} \exp(-y_n a_k \phi(\boldsymbol{x}_n; \boldsymbol{\theta}_k))}{N \prod_{k=1}^{K} Z_k}. \end{aligned} \tag{20}$$

However, we know that the weights sum up to one. Thus

$$\sum_{n=1}^{N} \frac{\prod_{k=1}^{K} \exp(-y_n a_k \phi(\boldsymbol{x}_n; \boldsymbol{\theta}_k))}{N \prod_{k=1}^{K} Z_k} = 1. \tag{21}$$

Combining (19) and (21) results in,

$$P_e^N \leq \prod_{k=1}^{K} Z_k. \tag{22}$$

By the respective definition we have,

$$\begin{aligned} Z_k &= \sum_{n=1}^{N} w_n^{(k)} \exp(-y_n a_k \phi(\boldsymbol{x}_n; \boldsymbol{\theta}_k)) \\ &= \sum_{y_n \phi(\boldsymbol{x}_n; \boldsymbol{\theta}_k) < 0} w_n^{(k)} \exp(a_k) + \sum_{y_n \phi(\boldsymbol{x}_n; \boldsymbol{\theta}_k) > 0} w_n^{(k)} \exp(-a_k). \end{aligned}$$

Using the respective formulas, (7.93) and (7.94) from the proof in the book, the above becomes

$$Z_k = P_k \exp(a_k) + (1 - P_k) \exp(-a_k). \qquad (23)$$

Also we know that

$$a_k = \frac{1}{2} \ln \frac{1 - P_k}{P_k}. \qquad (24)$$

Combining (23) and (24), we obtain

$$Z_k = 2\sqrt{P_k(1 - P_k)}. \qquad (25)$$

Hence (22) can now be written as

$$P_e^N \leq \prod_{k=1}^{K} \left\{ 2\sqrt{P_k(1 - P_k)} \right\} = \prod_{k=1}^{K} \sqrt{1 - 4\gamma_k^2} \leq \exp(-2 \sum_{k=1}^{K} \gamma_k^2),$$

where by definition $\gamma_k \equiv 1/2 - P_k$. Since the base classifiers do better than a random guessing, we can write that $\gamma_k \geq \gamma > 0$. Thus the error is bounded by $\exp(-2K\gamma^2)$ and drops exponentially fast with K.

Solutions To Problems of Chapter 8

1. Prove the Cauchy - Schwartz's inequality in a general Hilbert space.

   Solution: We have to show that $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{H}$,
$$|\langle \boldsymbol{x},\ \boldsymbol{y}\rangle| \leq \|\boldsymbol{x}\|\|\boldsymbol{y}\|,$$
   and that equality holds only if $\boldsymbol{y} = a\boldsymbol{x}$, $a \in \mathbb{C}$.

   Solution: The inequality holds true for $\boldsymbol{x}$ and/or $\boldsymbol{y} = \boldsymbol{0}$. Let now $\boldsymbol{y},\ \boldsymbol{x} \neq \boldsymbol{0}$. We have
$$\begin{aligned} 0 &\leq \|\boldsymbol{x} - \lambda\boldsymbol{y}\|^2 = \langle \boldsymbol{x} - \lambda\boldsymbol{y},\ \boldsymbol{x} - \lambda\boldsymbol{y}\rangle \\ &= \|\boldsymbol{x}\|^2 + |\lambda|^2\|\boldsymbol{y}\|^2 - \lambda^*\langle \boldsymbol{x},\ \boldsymbol{y}\rangle - \lambda\langle \boldsymbol{y},\ \boldsymbol{x}\rangle. \end{aligned}$$
   Since the last inequality is valid for any $\lambda \in \mathbb{C}$, let
$$\lambda = \frac{\langle \boldsymbol{x},\ \boldsymbol{y}\rangle}{\|\boldsymbol{y}\|^2}.$$
   Thus
$$0 \leq \|\boldsymbol{x}\|^2 - \frac{|\langle \boldsymbol{x},\ \boldsymbol{y}\rangle|^2}{\|\boldsymbol{y}\|^2},$$
   from which a) the inequality results and b) the fact that the equality holds iff $\boldsymbol{x} = a\boldsymbol{y}$.

   Indeed, if $\boldsymbol{x} = a\boldsymbol{y}$, then equality is trivially shown. Let us now assume that equality holds true. Then
$$\langle \boldsymbol{x}, \boldsymbol{x}\rangle\langle \boldsymbol{y}, \boldsymbol{y}\rangle = \langle \boldsymbol{x}, \boldsymbol{y}\rangle^*\langle \boldsymbol{x}, \boldsymbol{y}\rangle$$
   and from the properties of the inner product in a Hilbert space we have
$$\langle \boldsymbol{x}, ||\boldsymbol{y}||^2\boldsymbol{x}\rangle = \langle \boldsymbol{x}, \langle \boldsymbol{x}, \boldsymbol{y}\rangle\boldsymbol{y}\rangle,$$
   from which it is readily seen that
$$\boldsymbol{x} = \frac{\langle \boldsymbol{x}, \boldsymbol{y}\rangle}{||\boldsymbol{y}||^2}\boldsymbol{y},$$
   which proves the claim.

2. Show a) that the set of points in a Hilbert space $\mathbb{H}$,
$$C = \{\boldsymbol{x} :\ \|\boldsymbol{x}\| \leq 1\}$$
   is a convex set, and b) the set of points
$$C = \{\boldsymbol{x} :\ \|\boldsymbol{x}\| = 1\}$$

is a nonconvex one.

Solution: From the definition of a Hilbert space (see Appendix of this chapter) the norm is the induced by the inner product norm, i.e.,

$$\|\boldsymbol{x}\| = (\langle \boldsymbol{x}, \boldsymbol{x} \rangle)^{\frac{1}{2}} .$$

a) Let us now consider two points, $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{H}$, such that

$$\|\boldsymbol{x}_1\| \leq 1, \quad \|\boldsymbol{x}_2\| \leq 1,$$

and let

$$\boldsymbol{x} = \lambda \boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2, \quad \lambda \in [0,1].$$

Then, by the triangle inequality property of a norm

$$\|\boldsymbol{x}\| = \|\lambda \boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2\| \leq \|\lambda \boldsymbol{x}_1\| + \|(1-\lambda)\boldsymbol{x}_2\|,$$

and since $\lambda \in [0,1]$

$$\|\boldsymbol{x}\| \leq \lambda \|\boldsymbol{x}_1\| + (1-\lambda)\|\boldsymbol{x}_2\| \leq (\lambda + 1 - \lambda)1 \leq 1.$$

b) Let two points such that,

$$\|\boldsymbol{x}_1\| = 1, \quad \|\boldsymbol{x}_2\| = 1,$$

and

$$\boldsymbol{x} = \lambda \boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2.$$

Then we have that

$$\begin{aligned} \|\boldsymbol{x}\|^2 &= \langle \lambda \boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2,\ \lambda \boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2 \rangle \\ &= \lambda^2 \|\boldsymbol{x}_1\|^2 + (1-\lambda)^2 \|\boldsymbol{x}_2\|^2 + 2\lambda(1-\lambda)\langle \boldsymbol{x}_1,\ \boldsymbol{x}_2 \rangle \\ &= \lambda^2 + (1-\lambda)^2 + 2\lambda(1-\lambda)\langle \boldsymbol{x}_1,\ \boldsymbol{x}_2 \rangle. \end{aligned} \tag{1}$$

From the Schwartz inequality (Problem 1), we have that

$$|\langle \boldsymbol{x}_1,\ \boldsymbol{x}_2 \rangle| \leq \|\boldsymbol{x}_1\| \|\boldsymbol{x}_2\|, \tag{2}$$

or

$$-1 \leq \langle \boldsymbol{x}_1,\ \boldsymbol{x}_2 \rangle \leq 1. \tag{3}$$

From (1) and (3) is readily seen that

$$\|\boldsymbol{x}\|^2 \leq 1.$$

As a matter of fact, the only way for $\|\boldsymbol{x}\|^2 = 1$, is that

$$\langle \boldsymbol{x}_1,\ \boldsymbol{x}_2 \rangle = 1 = \|\boldsymbol{x}_1\| \|\boldsymbol{x}_2\|.$$

However, this is not possible. Equality in (2) is attained if

$$\boldsymbol{x}_1 = a \boldsymbol{x}_2,$$

and since $\|\boldsymbol{x}_1\| = \|\boldsymbol{x}_2\| = 1$, this can only happen in the trivial case of $\boldsymbol{x}_1 = \boldsymbol{x}_2$.

3. Show the first order convexity condition.

   Solution: Assume that f is convex. Then

   $$f\big(\lambda \boldsymbol{y} + (1-\lambda)\boldsymbol{x}\big) \leq \lambda f(\boldsymbol{y}) + (1-\lambda) f(\boldsymbol{x})$$

   or

   $$f\big(\boldsymbol{x} + \lambda(\boldsymbol{y}-\boldsymbol{x})\big) - f(\boldsymbol{x}) \leq \lambda\big(f(\boldsymbol{y}) - f(\boldsymbol{x})\big).$$

   Taking $\lambda \longrightarrow 0$, we can employ the Taylor expansion and get

   $$f\big(\boldsymbol{x}+\lambda(\boldsymbol{y}-\boldsymbol{x})\big)-f(\boldsymbol{x}) \approx f(\boldsymbol{x})+\lambda\nabla^T f(\boldsymbol{x})(\boldsymbol{y}-\boldsymbol{x})-f(\boldsymbol{x}) \leq \lambda\big(f(\boldsymbol{y})-f(\boldsymbol{x})\big),$$

   from which, in the limit we obtain

   $$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla^T f(\boldsymbol{x})(\boldsymbol{y}-\boldsymbol{x}). \tag{4}$$

   b) Assume that

   $$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla^T f(\boldsymbol{x})(\boldsymbol{y}-\boldsymbol{x}),$$

   is valid $\forall \boldsymbol{x}, \boldsymbol{y} \in X$, where X is the domain of definition of f. Then, we have

   $$f(\boldsymbol{y}_1) \geq f(\boldsymbol{x}) + \nabla^T f(\boldsymbol{x})(\boldsymbol{y}_1-\boldsymbol{x}), \tag{5}$$

   and

   $$f(\boldsymbol{y}_2) \geq f(\boldsymbol{x}) + \nabla^T f(\boldsymbol{x})(\boldsymbol{y}_2-\boldsymbol{x}). \tag{6}$$

   Combining the previous two inequalities together, we obtain

   $$\begin{aligned}\lambda f(\boldsymbol{y}_1) + (1-\lambda) f(\boldsymbol{y}_2) \quad \geq \quad & \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{x}) + \\ & \lambda \nabla^T f(\boldsymbol{x})(\boldsymbol{y}_1-\boldsymbol{x}) + \\ & (1-\lambda)\nabla^T f(\boldsymbol{x})(\boldsymbol{y}_2-\boldsymbol{x}), \end{aligned} \tag{7}$$

   for $\lambda \in (0,1)$. Since this is true for any $\boldsymbol{x}$, it will also be true for

   $$\boldsymbol{x} = \lambda \boldsymbol{y}_1 + (1-\lambda)\boldsymbol{y}_2,$$

   which results in

   $$f\big(\lambda \boldsymbol{y}_1 + (1-\lambda)\boldsymbol{y}_2\big) \leq \lambda f(\boldsymbol{y}_1) + (1-\lambda) f(\boldsymbol{y}_2), \tag{8}$$

   which proves the claim.

4. Show that a function f is convex, iff the one-dimensional function,

   $$g(t) := f(\boldsymbol{x} + t\boldsymbol{y}),$$

   is convex, $\forall \boldsymbol{x},\ \boldsymbol{y}$ in the domain of definition of f.

   Solution: Observe that,

   $$\begin{aligned} g(\lambda t_1 + (1-\lambda)t_2) \quad &= \quad f(\boldsymbol{x} + \lambda t_1 \boldsymbol{y} + (1-\lambda)t_2 \boldsymbol{y}) \\ &= \quad f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{x} + \lambda t_1 \boldsymbol{y} + (1-\lambda) t_2 \boldsymbol{y}) \\ &= \quad f\big(\lambda(\boldsymbol{x} + t_1\boldsymbol{y}) + (1-\lambda)(\boldsymbol{x} + t_2 \boldsymbol{y})\big). \end{aligned}$$

Also note that

$$g(t_1) = f(\boldsymbol{x} + t_1\boldsymbol{y}), \quad g(t_2) = f(\boldsymbol{x} + t_2\boldsymbol{y}),$$

and taking the definition of convexity, the claim is now straightforward to be shown.

5. Show the second order convexity condition.
Hint: Show the claim first for the one-dimensional case and then use the result of the previous problem, for the generalization.

Solution: We start with the one-dimensional case. Let a function $f(x)$ be convex. Then we know from the first order convexity condition that

$$f'(x)(y - x) \leq f(y) - f(x) \leq f'(y)(y - x),$$

and dividing both sides by the positive quantity $(y - x)^2$, we get

$$\frac{f'(y) - f'(x)}{y - x} \geq 0,$$

and taking the limit $y \longrightarrow x$ we obtain

$$f^{''}(x) \geq 0. \tag{9}$$

Assume now that the second derivative is non-negative everywhere. Then select $y > x$ and we get

$$\begin{aligned} 0 &\leq \int_x^y f^{''}(z)(y - z)dz \\ &= f'(z)(y - z) \mid_{z=x}^{z=y} + \int_x^y f'(z)dz \\ &= -f'(x)(y - x) + f(y) - f(x). \end{aligned} \tag{10}$$

The above is true for $y > x$. Note that we can also show that

$$f(x) \geq f'(y)(x - y) + f(y),$$

by using the identity

$$0 \leq \int_x^y f^{''}(z)(z - x)dz.$$

Thus we proved f is convex. For the more general case, consider

$$g(t) = f(\boldsymbol{x} + t\boldsymbol{y}),$$

form which we get

$$g^{''}(t) = \boldsymbol{y}^T \nabla^2 f(\boldsymbol{x} + t\boldsymbol{y})\boldsymbol{y}.$$

Since this is true for any $\boldsymbol{x}$, $\boldsymbol{y}$ and t and using the previously obtained results, the claim is readily shown.

6. Show that a function

$$f : \mathbb{R}^l \longmapsto \mathbb{R}$$

is convex iff its epigraph is convex.

Solution: a) Assume f to be convex. We have to show that its epigraph is a convex set. Let two points, $\boldsymbol{x}_1,\ \boldsymbol{x}_2$. From the convexity of f we have

$$f(\lambda\boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2) \leq \lambda f(\boldsymbol{x}_1) + (1-\lambda)f(\boldsymbol{x}_2). \tag{11}$$

Consider two points in the epigraph $\boldsymbol{y}_1 = (\boldsymbol{x}_1, r_1)$ and $\boldsymbol{y}_2 = (\boldsymbol{x}_2, r_2)$. Then we have

$$\lambda\boldsymbol{y}_1 + (1-\lambda)\boldsymbol{y}_2 := \boldsymbol{y} = (\boldsymbol{x}, r)$$

with

$$\boldsymbol{x} = \lambda\boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2, \tag{12}$$

$$r = \lambda r_1 + (1-\lambda)r_2, \tag{13}$$

and since $\boldsymbol{y}_1, \boldsymbol{y}_2 \in \text{epi}(f)$

$$f(\boldsymbol{x}_1) \leq r_1,\ f(\boldsymbol{x}_2) \leq r_2. \tag{14}$$

Combining (11) and (14) we get

$$f(\boldsymbol{x}) \leq \lambda f(\boldsymbol{x}_1) + (1-\lambda)f(\boldsymbol{x}_2) \leq \lambda r_1 + (1-\lambda)r_2 = r,$$

hence $\boldsymbol{y} = (\boldsymbol{x}, r) \in \text{epi}(f)$ and the epigraph is convex.

b) Assume the epigraph to be convex. Then

$$\boldsymbol{y} = \lambda\boldsymbol{y}_1 + (1-\lambda)\boldsymbol{y}_2 \in \text{epi}(f),$$

hence

$$r = \lambda r_1 + (1-\lambda)r_2 \geq f(\lambda\boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2) \tag{15}$$

for any

$$r_1 \geq f(\boldsymbol{x}_1),\ r_2 \geq f(\boldsymbol{x}_2).$$

Thus (15) is also valid for $r_1 = f(\boldsymbol{x}_1),\ r_2 = f(\boldsymbol{x}_2)$ and therefore

$$f(\lambda\boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2) \leq \lambda f(\boldsymbol{x}_1) + (1-\lambda)f(\boldsymbol{x}_2).$$

7. Show that if a function is convex, then its lower level set is convex for any ξ.

Solution: Let the function f be convex and two points, $\boldsymbol{x},\ \boldsymbol{y}$, which lie in the $\text{lev}_{\leq\xi}(f)$. Then,

$$f(\boldsymbol{x}) \leq \xi, \quad f(\boldsymbol{y}) \leq \xi.$$

Hence, by the definition of convexity,

$$f(\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \leq \lambda f(\boldsymbol{x}) + (1-\lambda)f(\boldsymbol{y}) \leq \lambda\xi + (1-\lambda)\xi \leq \xi,$$

which proves the claim, that $\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y} \in \text{lev}_{\leq\xi}(f)$.

8. Show that in a Hilbert space, $\mathbb{H}$, the parallelogram rule,

$$\|\boldsymbol{x}+\boldsymbol{y}\|^2 + \|\boldsymbol{x}-\boldsymbol{y}\|^2 = 2\left(\|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2\right), \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{H}.$$

holds true.

Solution: The proof is straightforward from the respective definitions and the properties of the inner product relation.

$$\begin{aligned} \|\boldsymbol{x}+\boldsymbol{y}\|^2 &= \|\boldsymbol{x}\|^2 + 2\langle \boldsymbol{x},\ \boldsymbol{y}\rangle + \|\boldsymbol{y}\|^2, \\ \|\boldsymbol{x}-\boldsymbol{y}\|^2 &= \|\boldsymbol{x}\|^2 - 2\langle \boldsymbol{x},\ \boldsymbol{y}\rangle + \|\boldsymbol{y}\|^2, \end{aligned}$$

from which the parallelogram rule is obtained.

9. Show that if $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{H}$, where $\mathbb{H}$ is a Hilbert space, then the induced by the inner product norm satisfies the triangle inequality, as required by any norm, i.e.,

$$\|\boldsymbol{x}+\boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$$

Solution: By the respective definitions we have

$$\|\boldsymbol{x}+\boldsymbol{y}\|^2 := \langle \boldsymbol{x}+\boldsymbol{y},\ \boldsymbol{x}+\boldsymbol{y}\rangle = \|\boldsymbol{x}\|^2 + \langle \boldsymbol{x},\ \boldsymbol{y}\rangle + \langle \boldsymbol{y},\ \boldsymbol{x}\rangle + \|\boldsymbol{y}\|^2,$$

or

$$\begin{aligned} \|\boldsymbol{x}+\boldsymbol{y}\|^2 &= \|\boldsymbol{x}\|^2 + 2\text{Real}(\langle \boldsymbol{x},\ \boldsymbol{y}\rangle) + \|\boldsymbol{y}\|^2 \\ &\leq \|\boldsymbol{x}\|^2 + 2|\langle \boldsymbol{x},\ \boldsymbol{y}\rangle| + \|\boldsymbol{y}\|^2 \\ &\leq \|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2 + 2\|\boldsymbol{x}\|\|\boldsymbol{y}\| = (\|\boldsymbol{x}\| + \|\boldsymbol{y}\|)^2. \end{aligned}$$

where the Cauchy-Schwartz inequality has been used.

10. Show that if a point $\boldsymbol{x}_*$ is a local minimizer of a convex function, it is necessarily a global one. Moreover, it is the unique minimizer if the function is strictly convex.

Solution: Let $\boldsymbol{x}_*$ be a local minimizer. Then, there exists $\epsilon > 0$ and $\boldsymbol{y}_* \notin B[0, \epsilon]$ such that

$$f(\boldsymbol{x}_*) \leq f(\boldsymbol{x}_* + \Delta),\ \forall \Delta \in B[0, \epsilon],$$

and

$$f(\boldsymbol{y}_*) < f(\boldsymbol{x}_*).$$

Let

$$\lambda := \frac{\epsilon}{2||\boldsymbol{y}_* - \boldsymbol{x}_*||}.$$

Then,

$$\lambda(\boldsymbol{y}_* - \boldsymbol{x}_*) \in B[0, \epsilon].$$

Hence,

$$f\big(\boldsymbol{x}_* + \lambda(\boldsymbol{y}_* - \boldsymbol{x}_*)\big) \leq (1-\lambda)f(\boldsymbol{x}_*) + \lambda f(\boldsymbol{y}_*) < f(\boldsymbol{x}_*),$$

which is not possible, since $\boldsymbol{x}_*$ is a local minimizer.

Assume now that f is strictly convex and that there exist two minimizers, $\boldsymbol{x}_* \neq \boldsymbol{y}_*$. Then, by the definition of strict convexity, we have that

$$f(\frac{1}{2}\boldsymbol{x}_* + \frac{1}{2}\boldsymbol{y}_*) < \frac{1}{2}f(\boldsymbol{x}_*) + \frac{1}{2}f(\boldsymbol{y}_*) = f(\boldsymbol{x}_*),$$

which is not possible, since $\boldsymbol{x}_*$ is a global minimizer.

11. Let C be a closed convex set in a Hilbert space, $\mathbb{H}$. Then show that $\forall \boldsymbol{x} \in \mathbb{H}$, there exists a point, denoted as $P_C(\boldsymbol{x}) \in C$, such that

$$\|\boldsymbol{x} - P_C(\boldsymbol{x})\| = \min_{\boldsymbol{y}\in C} \|\boldsymbol{x} - \boldsymbol{y}\|.$$

Solution: Let $\boldsymbol{x} \notin C$, otherwise it is trivial. Let ρ be the largest lower bound of $\|\boldsymbol{x} - \boldsymbol{y}\|$, $\boldsymbol{y} \in C$, i.e.,

$$\rho := \inf_{\boldsymbol{y}\in C} \|\boldsymbol{x} - \boldsymbol{y}\| > 0.$$

Consider the sequence,

$$\rho_n = \rho + \frac{1}{n}.$$

By the definition of the infimum, for each n, there will be at least one element $\boldsymbol{x}_n \in C$, such that

$$\|\boldsymbol{x} - \boldsymbol{x}_n\| < \rho_n,$$

which then defines a sequence, $\{\boldsymbol{x}_n\}$, of points for which we have that

$$\rho \leq \|\boldsymbol{x} - \boldsymbol{x}_n\| < \rho_n,$$

or

$$\rho \leq \lim_{n\to\infty} \|\boldsymbol{x} - \boldsymbol{x}_n\| < \lim_{n\to\infty} \rho_n = \rho,$$

which necessarily leads to

$$\lim_{n\to\infty} \|\boldsymbol{x} - \boldsymbol{x}_n\| = \rho \tag{16}$$

From the parallelogram law, we can write

$$\|(\boldsymbol{x}-\boldsymbol{x}_m)+(\boldsymbol{x}-\boldsymbol{x}_n)\|^2+\|(\boldsymbol{x}-\boldsymbol{x}_m)-(\boldsymbol{x}-\boldsymbol{x}_n)\|^2 = 2\big(\|\boldsymbol{x}-\boldsymbol{x}_m\|^2+\|\boldsymbol{x}-\boldsymbol{x}_n\|^2\big),$$

or

$$\|\boldsymbol{x}_n - \boldsymbol{x}_m\|^2 = 2(\|\boldsymbol{x} - \boldsymbol{x}_m\|^2 + \|\boldsymbol{x} - \boldsymbol{x}_n\|^2) - 4\|\boldsymbol{x} - \frac{1}{2}(\boldsymbol{x}_n + \boldsymbol{x}_m)\|^2.$$

However, since C is convex, the point $\frac{1}{2}(\boldsymbol{x}_n + \boldsymbol{x}_m) \in C$ and we deduce that

$$\|\boldsymbol{x}_n - \boldsymbol{x}_m\|^2 \leq 2\left(\|\boldsymbol{x} - \boldsymbol{x}_m\|^2 + \|\boldsymbol{x} - \boldsymbol{x}_n\|^2\right) - 4\|\rho\|^2.$$

Taking the limit for $n, m \to \infty$ on both sides we get

$$\lim_{n,m\to\infty} \|\boldsymbol{x}_m - \boldsymbol{x}_n\|^2 \leq 0 \Rightarrow$$
$$\lim_{n,m\to\infty} \|\boldsymbol{x}_m - \boldsymbol{x}_n\| = 0.$$

That is, $\boldsymbol{x}_n$ is a Cauchy sequence and since $\mathbb{H}$ is Hilbert the sequence converges to a point $\boldsymbol{x}_*$. Moreover it converges in C, since C is closed, i.e., $\boldsymbol{x} \in C$. Hence we have

$$\begin{aligned} \|\boldsymbol{x} - \boldsymbol{x}_*\| &= \|\boldsymbol{x} - \boldsymbol{x}_n + \boldsymbol{x}_n - \boldsymbol{x}_*\| \\ &\leq \|\boldsymbol{x} - \boldsymbol{x}_n\| + \|\boldsymbol{x}_n - \boldsymbol{x}_*\|. \end{aligned}$$

Taking the limit and using (16) we obtain

$$\|\boldsymbol{x} - \boldsymbol{x}_*\| \leq \rho.$$

However, since $\boldsymbol{x}_* \in C$,

$$\|\boldsymbol{x} - \boldsymbol{x}_*\| \geq \rho,$$

which means that

$$\|\boldsymbol{x} - \boldsymbol{x}_*\| = \rho,$$

that is the infimum is attained, which proves the claim. Uniqueness has been established in the text.

12. Show that the projection of a point $\boldsymbol{x} \in \mathbb{H}$ onto a non-empty closed convex set, $C \subset \mathbb{H}$, lies on the boundary of C.

Solution: Assume that $P_C(\boldsymbol{x})$ is an interior point of C. By the definition of interior points, $\exists \delta > 0$ such that

$$S_\delta := \{\boldsymbol{y} : \ \|\boldsymbol{y} - P_C(\boldsymbol{x})\| < \delta\} \subset C.$$

Let

$$\boldsymbol{z} := P_C(\boldsymbol{x}) + \frac{\delta}{2} \cdot \frac{\boldsymbol{x} - P_C(\boldsymbol{x})}{\|\boldsymbol{x} - P_C(\boldsymbol{x})\|},$$

where by assumption $\|\boldsymbol{x} - P_C(\boldsymbol{x})\| > 0$, since $\boldsymbol{x} \notin C$. Obviously $\boldsymbol{z} \in S_\delta$. Hence,

$$\|\boldsymbol{x} - \boldsymbol{z}\| = \left\| \left(|\boldsymbol{x} - P_C(\boldsymbol{x})\right) \left(1 - \frac{\delta}{2\|\boldsymbol{x} - P_C(\boldsymbol{x})\|}\right) \right\|.$$

However, δ can be chosen arbitrarily small, thus choose

$$\delta < \|\boldsymbol{x} - P_C(\boldsymbol{x})\|.$$

However, this is not possible, since in this case

$$\|\boldsymbol{x}-\boldsymbol{z}\| < \|\boldsymbol{x}-P_C(\boldsymbol{x})\|,$$

which violates the definition of projection. Thus, $P_C(\boldsymbol{x})$ lies onto the boundary of C.

13. Derive the formula for the projection onto a hyperplane in a (real) Hilbert space, $\mathbb{H}$.

Solution: Let us first show that a hyperplane, H, is a closed convex set. Convexity is shown trivially. To show closeness, let $\boldsymbol{y}_n \in H \longrightarrow \boldsymbol{y}_*$. We will show that $\boldsymbol{y}_* \in H$. Let

$$\begin{aligned} 0 \leq \|\langle\boldsymbol{\theta},\boldsymbol{y}_*\rangle+\theta_0\|^2 &= \|\langle\boldsymbol{\theta},\boldsymbol{y}_*\rangle+\langle\boldsymbol{\theta},\boldsymbol{y}_n\rangle-\langle\boldsymbol{\theta},\boldsymbol{y}_n\rangle+\theta_0\|^2 \\ &= \|\langle\boldsymbol{\theta},\boldsymbol{y}_*-\boldsymbol{y}_n\rangle\|^2. \end{aligned}$$

Hence,

$$0 \leq \|\langle\boldsymbol{\theta},\boldsymbol{y}_*\rangle+\theta_0\|^2 = \lim_{n\to\infty}\|\langle\boldsymbol{\theta},\boldsymbol{y}_*-\boldsymbol{y}_n\rangle\|^2 = 0,$$

or

$$\langle\boldsymbol{\theta},\boldsymbol{y}_*\rangle+\theta_0 = 0,$$

which proves the claim.

Let now $\boldsymbol{z} \in H$ be the projection of $\boldsymbol{x} \in \mathbb{H}$, i.e., $\boldsymbol{z} : P_C(\boldsymbol{x})$. Then by the definition

$$\boldsymbol{z} := \arg\min_{\langle\boldsymbol{\theta},\boldsymbol{z}\rangle+\theta_0=0}\langle\boldsymbol{x}-\boldsymbol{z},\boldsymbol{x}-\boldsymbol{z}\rangle.$$

Using Lagrange multipliers, we obtain the Lagrangian

$$L(\boldsymbol{z},\lambda) = \langle\boldsymbol{x}-\boldsymbol{z},\boldsymbol{x}-\boldsymbol{z}\rangle-\lambda\big(\langle\boldsymbol{\theta},\boldsymbol{z}\rangle+\theta_0\big).$$

For those not familiar with infinite dimensional spaces, it suffices to say that similar rules of differentiation apply, although the respective definitions are different (more general).

After differentiation of the Lagrangian, we obtain

$$2\boldsymbol{z}-2\boldsymbol{x}-\lambda\boldsymbol{\theta}=0$$

or

$$\boldsymbol{z}=\frac{1}{2}(2\boldsymbol{x}+\lambda\boldsymbol{\theta}).$$

Plugging into the constraint, we obtain

$$\lambda = -2\frac{\langle\boldsymbol{\theta},\boldsymbol{x}\rangle+\theta_0}{\|\boldsymbol{\theta}\|^2},$$

which then results in the solution.

14. Derive the formula for the projection onto a closed ball, $B[\mathbf{0}, \rho]$.

Solution: The closed ball is defined as

$$B[\mathbf{0}, \rho] = \{\boldsymbol{y} : \|\boldsymbol{y}\| \leq \rho\}.$$

We have already seen (Problem 2) that it is convex. Let us show the closeness. Let

$$\boldsymbol{y}_n \in B[\mathbf{0}, \rho] \to \boldsymbol{y}^*.$$

We have to show that $\boldsymbol{y}^* \in B[\mathbf{0}, \rho]$.

$$\begin{aligned} \|\boldsymbol{y}^*\|^2 &= \|\boldsymbol{y}^* - \boldsymbol{y}_n + \boldsymbol{y}_n\|^2 \leq \|\boldsymbol{y}_n\|^2 + \|\boldsymbol{y}^* - \boldsymbol{y}_n\|^2 \\ &\leq \rho^2 + \|\boldsymbol{y}^* - \boldsymbol{y}_n\|^2, \end{aligned}$$

or

$$\|\boldsymbol{y}^*\|^2 \leq \rho^2 + \lim_{n\to\infty} \|\boldsymbol{y}^* - \boldsymbol{y}_n\|^2,$$

or

$$\|\boldsymbol{y}^*\|^2 \leq \rho^2,$$

which proves the claim.

To derive the projection, we follow similar steps as in Problem 13, by replacing the constraint by

$$\|\boldsymbol{z}\|^2 = \rho^2,$$

since the projection is on the boundary. Taking the gradient of the Lagrangian we get

$$2(\boldsymbol{z} - \boldsymbol{x}) + 2\lambda\boldsymbol{z} = 0,$$

or

$$\boldsymbol{z} = \frac{1}{1+\lambda}\boldsymbol{x}.$$

Plugging into the constraint, we get

$$|1+\lambda| = \frac{1}{\rho}\|\boldsymbol{x}\|.$$

When $1 + \lambda = \frac{1}{\rho}\|\boldsymbol{x}\|$, we get

$$\boldsymbol{z} = \frac{\rho}{\|\boldsymbol{x}\|}\boldsymbol{x},$$

and when $1 + \lambda = -\frac{1}{\rho}\|\boldsymbol{x}\|$

$$\boldsymbol{z} = -\frac{\rho}{\|\boldsymbol{x}\|}\boldsymbol{x}.$$

From the two possible vectors, we have to keep the one that has the smaller distance from $\boldsymbol{x}$. However,

$$\left\| \boldsymbol{x} - \frac{\rho}{\|\boldsymbol{x}\|}\boldsymbol{x} \right\| < \left\| \boldsymbol{x} + \frac{\rho}{\|\boldsymbol{x}\|}\boldsymbol{x} \right\|,$$

since $\frac{\rho}{\|\boldsymbol{x}\|} < 1$. Thus,

$$1 + \lambda = \frac{1}{\rho}\|\boldsymbol{x}\|,$$

and the projection is equal to

$$P_{B[0,\rho]}(\boldsymbol{x}) = \begin{cases} \frac{\rho}{\|\boldsymbol{x}\|}\boldsymbol{x}, & \|\boldsymbol{x}\| > \rho \\ \boldsymbol{x} & \text{otherwise.} \end{cases}$$

15. Find an example of a point whose projection on the ℓ_1 ball is not unique.

Solution: The ℓ_1 ball of radius $\rho = 1$ in $\mathbb{R}^l$ is defined as

$$S_1[\mathbf{0}, \rho] = \{\boldsymbol{y} : \sum_{i=1} |y_i| \leq \rho\}.$$

Let the point

$$\boldsymbol{x} = [1, 1]^T \in \mathbb{R}^2,$$

which obviously does not lie inside the ℓ_1 ball of radius $\rho = 1$, since $\|\boldsymbol{x}\|_1 = 2 > 1$. For any point $\boldsymbol{y} \in S_1[\mathbf{0}, 1]$ we have

$$\|\boldsymbol{x} - \boldsymbol{y}\|_1 = |1 - y_1| + |1 - y_2| \geq 1 - |y_1| + 1 - |y_2|$$

or

$$\|\boldsymbol{x} - \boldsymbol{y}\|_1 \geq 2 - \|\boldsymbol{y}\|_1 \geq 1.$$

That is, the ℓ_1 norm of $\boldsymbol{x}$ from any point in the set $S_1[\mathbf{0}, 1]$ is bounded below by 1. Consider the two points

$$\boldsymbol{y}_1 = [1, 0]^T \text{ and } \boldsymbol{y}_1 = [0, 1]^T.$$

For both of them the lower bound is achieved, i.e.,

$$\|\boldsymbol{x} - \boldsymbol{y}_1\| = \|\boldsymbol{x} - \boldsymbol{y}_2\| = 1.$$

Moreover, one can easily check out that all points on the line segment

$$\boldsymbol{y} : y_1 + y_2 = 1$$

can be projection points of $\boldsymbol{x}$.

16. Show that if $C \subset \mathbb{H}$, is a closed convex set in a Hilbert space, then $\forall \boldsymbol{x} \in \mathbb{H}$ and $\forall \boldsymbol{y} \in C$, the projection $P_C(\boldsymbol{x})$ satisfies the following properties:

- $\text{Real}\{\langle \boldsymbol{x} - P_C(\boldsymbol{x}),\ \boldsymbol{y} - P_C(\boldsymbol{x})\rangle\} \leq 0.$
- $\|P_C(\boldsymbol{x}) - P_C(\boldsymbol{y})\|^2 \leq \text{Real}\{\langle \boldsymbol{x} - \boldsymbol{y},\ P_C(\boldsymbol{x}) - P_C(\boldsymbol{y})\rangle\}.$

Solution: We know that $P_C(\boldsymbol{x}) \in C$. Hence, due to the convexity of C, $\forall \lambda \in [0, 1]$

$$\lambda \boldsymbol{y} + (1 - \lambda) P_C(\boldsymbol{x}) \in C.$$

Hence

$$\begin{aligned} \|\boldsymbol{x} - P_C(\boldsymbol{x})\|^2 &\leq \|\boldsymbol{x} - (\lambda \boldsymbol{y} + (1 - \lambda) P_C(\boldsymbol{x}))\|^2 \\ &\leq \|(\boldsymbol{x} - P_C(\boldsymbol{x})) - \lambda(\boldsymbol{y} - P_C(\boldsymbol{x}))\|^2, \end{aligned}$$

which by the definition of the norm, gives

$$\begin{aligned} \|\boldsymbol{x} - P_C(\boldsymbol{x})\|^2 \leq\ & \|\boldsymbol{x} - P_C(\boldsymbol{x})\|^2 + \lambda^2 \|\boldsymbol{y} - P_C(\boldsymbol{x})\|^2 - \\ & 2\lambda \text{Real}\{\langle \boldsymbol{x} - P_C(\boldsymbol{x}),\ \boldsymbol{y} - P_C(\boldsymbol{x})\rangle\}. \end{aligned}$$

or

$$\text{Real}\{\langle \boldsymbol{x} - P_C(\boldsymbol{x}),\ \boldsymbol{y} - P_C(\boldsymbol{x})\rangle\} \leq \frac{\lambda}{2} \|\boldsymbol{y} - P_C(\boldsymbol{x})\|^2.$$

Taking the limit $\lambda \to 0$, we prove the first property.

To prove the second property, since $P_C(\boldsymbol{y}) \in C$, we apply the previous property with $P_C(\boldsymbol{y})$ in place of $\boldsymbol{y}$, i.e.,

$$\text{Real}\{\langle \boldsymbol{x} - P_C(\boldsymbol{x}),\ P_C(\boldsymbol{y}) - P_C(\boldsymbol{x})\rangle\} \leq 0.$$

Similarly

$$\text{Real}\{\langle \boldsymbol{y} - P_C(\boldsymbol{y}),\ P_C(\boldsymbol{x}) - P_C(\boldsymbol{y})\rangle\} \leq 0.$$

After adding the above inequalities together and rearranging the terms we obtain the second property.

17. Prove that if S is a closed subspace $S \subset \mathbb{H}$ in a Hilbert space $\mathbb{H}$, then $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{H}$,

$$\langle \boldsymbol{x}, P_S(\boldsymbol{y})\rangle = \langle P_S(\boldsymbol{x}), \boldsymbol{y}\rangle = \langle P_S(\boldsymbol{x}), P_S(\boldsymbol{y})\rangle.$$

and

$$P_S(a\boldsymbol{x} + b\boldsymbol{y}) = aP_S(\boldsymbol{x}) + bP_S(\boldsymbol{y}).$$

Hint: Use the result of Problem 18.

Solution: We have that

$$\begin{aligned} \langle \boldsymbol{x}, P_S(\boldsymbol{y})\rangle &= \langle P_S(\boldsymbol{x}) + (\boldsymbol{x} - P_S(\boldsymbol{x})), P_S(\boldsymbol{y})\rangle \\ &= \langle P_S(\boldsymbol{x}), P_S(\boldsymbol{y})\rangle, \end{aligned}$$

since

$$\boldsymbol{x} - P_S(\boldsymbol{x}) \perp P_S(\boldsymbol{y})$$

from Problem 18. Similarly, we can show that

$$\langle P_S(\boldsymbol{x}), \boldsymbol{y}\rangle = \langle P_S(\boldsymbol{x}), P_S(\boldsymbol{y})\rangle.$$

Hence

$$\langle \boldsymbol{x}, P_S(\boldsymbol{y})\rangle = \langle P_S(\boldsymbol{x}), \boldsymbol{y}\rangle.$$

For the linearity, we have

$$\begin{array}{rcl} \boldsymbol{x} & = & P_S(\boldsymbol{x}) + (\boldsymbol{x} - P_S(\boldsymbol{x})), \\ \boldsymbol{y} & = & P_S(\boldsymbol{y}) + (\boldsymbol{y} - P_S(\boldsymbol{y})), \end{array}$$

where $P_S(\boldsymbol{x}),\ P_S(\boldsymbol{y}) \in S$ and $\boldsymbol{x} - P_S(\boldsymbol{x}) \in S^\perp$ and $\boldsymbol{y} - P_S(\boldsymbol{y}) \in S^\perp$. Hence

$$a\boldsymbol{x} + b\boldsymbol{y} = (aP_S(\boldsymbol{x}) + bP_S(\boldsymbol{y})) + (a(\boldsymbol{x} - P_S(\boldsymbol{x})) + b(\boldsymbol{y} - P_S(\boldsymbol{y}))),$$

and since the term in the second parenthesis on the right hand side lies in $S^\perp$ we readily obtain that

$$P_S(a\boldsymbol{x} + b\boldsymbol{y}) = aP_S(\boldsymbol{x}) + bP_S(\boldsymbol{y}).$$

18. Let S be a closed convex subspace in a Hilbert space $\mathbb{H}$, $S \subset \mathbb{H}$. Let $S^\perp$ be the set of all elements $\boldsymbol{x} \in \mathbb{H}$ which are orthogonal to S. Then show that, a) $S^\perp$ is also a closed, subspace, b) $S \cap S^\perp = \{\boldsymbol{0}\}$, c) $\mathbb{H} = S \oplus S^\perp$; that is, $\forall \boldsymbol{x} \in \mathbb{H}$, $\exists \boldsymbol{x}_1 \in S$ and $\boldsymbol{x}_2 \in S^\perp :\ \boldsymbol{x} = \boldsymbol{x}_1 + \boldsymbol{x}_2$, where $\boldsymbol{x}_1,\ \boldsymbol{x}_2$ are *unique*.

   Solution:
   a) We will first prove that $S^\perp$ is a subspace. Indeed if $\boldsymbol{x}_1 \in S^\perp$ and $\boldsymbol{x}_2 \in S^\perp$ then

$$\langle \boldsymbol{x}_1, \boldsymbol{y}\rangle = \langle \boldsymbol{x}_2, \boldsymbol{y}\rangle = 0,\ \forall \boldsymbol{y} \in S.$$

   or

$$\langle a\boldsymbol{x}_1 + b\boldsymbol{x}_2, \boldsymbol{y}\rangle = 0 \Rightarrow a\boldsymbol{x}_1 + b\boldsymbol{x}_2 \in S^\perp.$$

   Also, $\boldsymbol{0} \in S^\perp$ since $\langle \boldsymbol{x}, \boldsymbol{0}\rangle = 0$. Hence $S^\perp$ is a subspace.

   We will prove that $S^\perp$ is also closed. Let $\{\boldsymbol{x}_n\} \in S^\perp$ and

$$\lim_{n\to\infty} \boldsymbol{x}_n = \boldsymbol{x}_*.$$

   We will show that $\boldsymbol{x}_* \in S^\perp$. By the definition

$$\langle \boldsymbol{x}_n, \boldsymbol{y}\rangle = 0,\ \forall \boldsymbol{y} \in S.$$

   Moreover,

$$\begin{array}{rcl} |\langle \boldsymbol{x}_*, \boldsymbol{y}\rangle| & = & |\langle \boldsymbol{x}_n, \boldsymbol{y}\rangle - \langle \boldsymbol{x}_*, \boldsymbol{y}\rangle| \\ & = & |\langle \boldsymbol{x}_n - \boldsymbol{x}_*, \boldsymbol{y}\rangle| \leq \|\boldsymbol{x}_n - \boldsymbol{x}_*\|\|\boldsymbol{y}\| \to 0, \end{array}$$

where the Cauchy-Schwartz inequality has been used. The last inequality leads to

$$\langle \boldsymbol{x}_*, \boldsymbol{y} \rangle = 0 \Rightarrow \boldsymbol{x}_* \in S^{\perp}.$$

b) Let $\boldsymbol{x} \in S \cap S^{\perp}$. By definition, since it belongs to both subspaces,

$$\langle \boldsymbol{x}, \boldsymbol{x} \rangle = 0 \Rightarrow \boldsymbol{x} = \boldsymbol{0},$$

c) Let $\boldsymbol{x} \in \mathbb{H}$. We have that

$$\boldsymbol{x} = P_S(\boldsymbol{x}) + (\boldsymbol{x} - P_S(\boldsymbol{x})).$$

We will first show that $\boldsymbol{x} - P_S(\boldsymbol{x}) \in S^{\perp}$. Then we will show that this decomposition is unique. We already know that

$$\text{Real}\{\langle \boldsymbol{x} - P_S(\boldsymbol{x}), \boldsymbol{y} - P_S(\boldsymbol{x}) \rangle\} \leq 0, \ \forall \boldsymbol{y} \in S.$$

Also, since S is a subspace, $a\boldsymbol{y} \in S$, $\forall a \in \mathbb{R}$, hence

$$\text{Real}\{\langle \boldsymbol{x} - P_S(\boldsymbol{x}), a\boldsymbol{y} - P_S(\boldsymbol{x}) \rangle\} \leq 0,$$

or

$$a\text{Real}\{\langle \boldsymbol{x} - P_S(\boldsymbol{x}), \boldsymbol{y} \rangle\} \leq \text{Real}\{\langle \boldsymbol{x} - P_S(\boldsymbol{x}), P_S(\boldsymbol{y}) \rangle\},$$

which can only be true if

$$\text{Real}\{\langle \boldsymbol{x} - P_S(\boldsymbol{x}), \boldsymbol{y} \rangle\} = 0.$$

We apply the same for $j\boldsymbol{y}$. Then we have that

$$\langle \boldsymbol{x} - P_S(\boldsymbol{x}), j\boldsymbol{y} \rangle = -j \langle \boldsymbol{x} - P_S(\boldsymbol{x}), \boldsymbol{y} \rangle.$$

Recall that if $c \in \mathbb{C}$

$$\text{Imag}\{c\} = \text{Real}\{-jc\},$$

Hence,

$$\text{Real}\{-j\langle \boldsymbol{x} - P_S(\boldsymbol{x}), \boldsymbol{y} \rangle\} = 0 = \text{Imag}\{\langle \boldsymbol{x} - P_S(\boldsymbol{x}), \boldsymbol{y} \rangle\}.$$

Thus,

$$\langle \boldsymbol{x} - P_S(\boldsymbol{x}), \boldsymbol{y} \rangle\} = 0, \ \forall \boldsymbol{y} \in S,$$

and

$$\boldsymbol{x} - P_S(\boldsymbol{x}) \in S^{\perp}.$$

Thus

$$\boldsymbol{x} = \boldsymbol{x}_1 + \boldsymbol{x}_2, \ \boldsymbol{x}_1 = P_S(\boldsymbol{x}) \in S, \ \boldsymbol{x}_2 = \boldsymbol{x} - P_S(\boldsymbol{x}) \in S^{\perp}.$$

Let us now assume that there is another decomposition,

$$\boldsymbol{x} = \boldsymbol{x}_3 + \boldsymbol{x}_4, \ \boldsymbol{x}_3 \in S, \ \boldsymbol{x}_4 \in S^{\perp}.$$

Then

$$\boldsymbol{x}_1 + \boldsymbol{x}_2 = \boldsymbol{x}_3 + \boldsymbol{x}_4$$

or

$$S \to \boldsymbol{x}_1 - \boldsymbol{x}_3 = \boldsymbol{x}_4 - \boldsymbol{x}_2 \in S^{\perp},$$

which necessarily implies that they are equal to the single point comprising $S \cap S^{\perp}$, i.e.,

$$\boldsymbol{x}_1 - \boldsymbol{x}_3 = 0 = \boldsymbol{x}_4 - \boldsymbol{x}_2,$$

hence the decomposition is unique and we have proved the claim.

Let us elaborate a bit more. we will show that,

$$P_{S^{\perp}}(\boldsymbol{x}) = \boldsymbol{x} - P_S(\boldsymbol{x}).$$

Indeed, $P_{S^{\perp}}(\boldsymbol{x})$ is unique. Also,

$$\boldsymbol{x} = P_S(\boldsymbol{x}) + (\boldsymbol{x} - P_S(\boldsymbol{x})).$$

or

$$\begin{aligned} P_{S^{\perp}}(\boldsymbol{x}) &= \mathbf{0} + P_{S^{\perp}}(\boldsymbol{x} - P_S(\boldsymbol{x})) \\ &= \boldsymbol{x} - P_S(\boldsymbol{x}), \end{aligned}$$

since $\boldsymbol{x} - P_S(\boldsymbol{x}) \in S^{\perp}$. Note that we used the fact that if $\boldsymbol{y} \in S$ then

$$P_{S^{\perp}}(\boldsymbol{y}) = \mathbf{0}.$$

Indeed,

$$\|\boldsymbol{y} - \mathbf{0}\|^2 = \|\boldsymbol{y}\|^2 < \|\boldsymbol{y} - \boldsymbol{a}\|^2, \ \forall \boldsymbol{a} \neq \mathbf{0} \in S^{\perp}$$

since

$$\begin{aligned} \|\boldsymbol{y} - \boldsymbol{a}\|^2 &= \|\boldsymbol{y}\|^2 + \|\boldsymbol{a}\|^2 - 2\text{Real}\langle \boldsymbol{y}, \boldsymbol{a}\rangle \\ &= \|\boldsymbol{y}\|^2 + \|\boldsymbol{a}\|^2 \end{aligned}$$

19. Show that the relaxed projection operator is a non-expansive mapping.

Solution: By the respective definitions we have,

$$\begin{aligned} \|T_C(\boldsymbol{x}) - T_C(\boldsymbol{y})\| &= \|\boldsymbol{x} + \mu(P_C(\boldsymbol{x}) - \boldsymbol{x}) - \boldsymbol{y} - \mu(P_C(\boldsymbol{y}) - \boldsymbol{y})\| \\ &= \|(1-\mu)(\boldsymbol{x} - \boldsymbol{y}) + \mu(P_C(\boldsymbol{x}) - P_C(\boldsymbol{y}))\|. \end{aligned}$$

a) $\mu \in (0, 1]$. Recalling the triangle property of a norm (Appendix of Chapter 8), we get

$$\begin{aligned} \|T_C(\boldsymbol{x}) - T_C(\boldsymbol{y})\| &\leq |1-\mu|\|\boldsymbol{x} - \boldsymbol{y}\| + \mu\|P_C(\boldsymbol{x}) - P_C(\boldsymbol{y})\| \\ &\leq |1-\mu|\|\boldsymbol{x} - \boldsymbol{y}\| + \mu\|\boldsymbol{x} - \boldsymbol{y}\| \\ &\leq \|\boldsymbol{x} - \boldsymbol{y}\|. \end{aligned}$$

b) $\mu \in (1, 2)$. In this case

$$\begin{aligned}
\|T_C(\boldsymbol{x}) - T_C(\boldsymbol{y})\|^2 &= (1-\mu)^2\|\boldsymbol{x}-\boldsymbol{y}\|^2 + \mu^2\|P_C(\boldsymbol{x}) - P_C(\boldsymbol{y})\|^2 \\
&+ 2\mu(1-\mu)\text{Real}\{\langle \boldsymbol{x}-\boldsymbol{y}, P_C(\boldsymbol{x}) - P_C(\boldsymbol{y})\rangle\} \\
&\leq (1-\mu)^2\|\boldsymbol{x}-\boldsymbol{y}\|^2 + \mu^2\|P_C(\boldsymbol{x}) - P_C(\boldsymbol{y})\|^2 \\
&+ 2\mu(1-\mu)\|P_C(\boldsymbol{x}) - P_C(\boldsymbol{y})\|^2 \\
&= (1-\mu)^2\|\boldsymbol{x}-\boldsymbol{y}\|^2 + \mu(2-\mu)\|P_C(\boldsymbol{x}) - P_C(\boldsymbol{y})\|^2 \\
&\leq (1-\mu)^2\|\boldsymbol{x}-\boldsymbol{y}\|^2 + \mu(2-\mu)\|\boldsymbol{x}-\boldsymbol{y}\|^2 \\
&\leq \|\boldsymbol{x}-\boldsymbol{y}\|^2.
\end{aligned}$$

To derive the bounds we used that $1-\mu < 0$ and $2-\mu > 0$, for $\mu \in (1, 2)$.

20. Show that the relaxed projection operator is a strongly attractive mapping.

Solution: By the respective definition we have,

$$\begin{aligned}
\|T_C(\boldsymbol{x}) - \boldsymbol{y}\|^2 &= \|\boldsymbol{x} + \mu(P_C(\boldsymbol{x}) - \boldsymbol{x}) - \boldsymbol{y}\|^2 \\
&= \|\boldsymbol{x}-\boldsymbol{y}\|^2 + \mu^2\|P_C(\boldsymbol{x}) - \boldsymbol{x}\|^2 + 2\mu\text{Real}\{\langle \boldsymbol{x}-\boldsymbol{y}, P_C(\boldsymbol{x}) - \boldsymbol{x}\rangle\} \\
&= \|\boldsymbol{x}-\boldsymbol{y}\|^2 + \mu^2\|P_C(\boldsymbol{x}) - \boldsymbol{x}\|^2 \\
&+ 2\mu\text{Real}\{\langle \boldsymbol{x} - P_C(\boldsymbol{x}) + P_C(\boldsymbol{x}) - \boldsymbol{y}, P_C(\boldsymbol{x}) - \boldsymbol{x}\rangle\} \\
&= \|\boldsymbol{x}-\boldsymbol{y}\|^2 + \mu^2\|P_C(\boldsymbol{x}) - \boldsymbol{x}\|^2 - 2\mu\|P_C(\boldsymbol{x}) - \boldsymbol{x}\|^2 \\
&+ 2\mu\text{Real}\{\langle P_C(\boldsymbol{x}) - \boldsymbol{y}, P_C(\boldsymbol{x}) - \boldsymbol{x}\rangle\} \\
&= \|\boldsymbol{x}-\boldsymbol{y}\|^2 + \mu^2\|P_C(\boldsymbol{x}) - \boldsymbol{x}\|^2 - 2\mu\|P_C(\boldsymbol{x}) - \boldsymbol{x}\|^2 \\
&+ 2\mu\text{Real}\{\langle \boldsymbol{x} - P_C(\boldsymbol{x}), \boldsymbol{y} - P_C(\boldsymbol{y})\rangle\} \\
&\leq \|\boldsymbol{x}-\boldsymbol{y}\|^2 - \mu(2-\mu)\|P_C(\boldsymbol{x}) - \boldsymbol{x}\|^2.
\end{aligned}$$

where (8.15) has been used. Thus

$$\begin{aligned}
\|T_C(\boldsymbol{x}) - \boldsymbol{y}\|^2 &\leq \|\boldsymbol{x}-\boldsymbol{y}\|^2 - \frac{\mu(2-\mu)}{\mu^2}\|T_C(\boldsymbol{x}) - \boldsymbol{x}\|^2 \\
&= \|\boldsymbol{x}-\boldsymbol{y}\|^2 - \frac{(2-\mu)}{\mu}\|T_C(\boldsymbol{x}) - \boldsymbol{x}\|^2,
\end{aligned}$$

where we used that

$$P_C(\boldsymbol{x}) - \boldsymbol{x} = \frac{1}{\mu}(T_C(\boldsymbol{x}) - \boldsymbol{x}).$$

21. Give an example of a sequence in a Hilbert space $\mathbb{H}$, which converges weakly but not strongly.

Solution: Define the sequence of points $\boldsymbol{x}_n \in l^2$,

$$\boldsymbol{x}_n = \{0, 0, \ldots, 1, 0, 0, \ldots\} := \{\delta_{ni}\}, \; i = 0, 1, 2, \ldots$$

That is, each point, $\boldsymbol{x}_n$ is itself a sequence, with zeros everywhere, except at time index, n, where it is 1. For every point (sequence) $\boldsymbol{y} \in l^2$, we have that

$$\|\boldsymbol{y}\|^2 := \sum_{n=1}^{\infty} |y_n|^2 = \sum_{n=1}^{\infty} |\langle \boldsymbol{x}_n, \boldsymbol{y}\rangle|^2 < \infty,$$

by the definition of l^2 space (Appendix of Chapter 8). The previous inequality implies that

$$\langle \boldsymbol{x}_n, \boldsymbol{y}\rangle \xrightarrow[n\to\infty]{} 0 = \langle \boldsymbol{0}, \boldsymbol{y}\rangle.$$

On the other hand,

$$\boldsymbol{x}_n \not\longrightarrow \boldsymbol{0},$$

since

$$\|\boldsymbol{x}_n - \boldsymbol{0}\| = \|\boldsymbol{x}_n\| = 1.$$

22. Prove that if $C_1 \dots C_K$ are closed convex sets in a Hilbert space $\mathbb{H}$, then the operator

$$T = T_{C_K} \cdots T_{C_1},$$

is a *regular* one; that is,

$$\|T^{n-1}(\boldsymbol{x}) - T^n(\boldsymbol{x})\| \longrightarrow 0, \ n \longrightarrow \infty,$$

where $T^n := TT \dots T$ is the application of T n successive times.

Solution:
Fact 1:

$$T = T_{C_K} T_{C_{K-1}} \cdots T_{C_1} := T_K \cdots T_1$$

is a non-expansive mapping.

Indeed, $\forall \boldsymbol{x}, \ \boldsymbol{y} \in \mathbb{H}$

$$\begin{aligned}
\|T(\boldsymbol{x}) - T(\boldsymbol{y})\| &= \|T_K(T_{K-1}\cdots T_1)(\boldsymbol{x}) - T_K(T_{K-1}\cdots T_1)(\boldsymbol{y})\| \\
&\leq \|T_{K-1}\cdots T_1(\boldsymbol{x}) - T_{K-1}\cdots T_1(\boldsymbol{y})\| \\
&\leq \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\
&\leq \|T_1(\boldsymbol{x}) - T_1(\boldsymbol{y})\| \leq \|\boldsymbol{x} - \boldsymbol{y}\|.
\end{aligned}$$

Fact 2:

$$\text{Fix}(T) = \bigcap_{k=1}^{K} C_k := C.$$

Indeed, if $\boldsymbol{x} \in C$ then

$$\begin{aligned}
T_K T_{K-1} \cdots T_1(\boldsymbol{x}) &= T_K T_{K-1} \cdots T_2(\boldsymbol{x}) = \dots \\
&= T_K(\boldsymbol{x}) = \boldsymbol{x}.
\end{aligned}$$

Moreover, let us assume that $\exists\ \boldsymbol{x} \notin C$:

$$T_K(\boldsymbol{x}) = \boldsymbol{x}.$$

Then $\forall \boldsymbol{y} \in C$ we have

$$\|\boldsymbol{x} - \boldsymbol{y}\| = \|T(\boldsymbol{x}) - T(\boldsymbol{y})\| \leq \|T_1(\boldsymbol{x}) - T_1(\boldsymbol{y})\|,$$

as shown before. Thus,

$$\|\boldsymbol{x} - \boldsymbol{y}\| = \|T_1(\boldsymbol{x}) - T_1(\boldsymbol{y})\| = \|T_1(\boldsymbol{x}) - \boldsymbol{y}\| \leq \|\boldsymbol{x} - \boldsymbol{y}\|$$

or

$$\|T_1(\boldsymbol{x}) - \boldsymbol{y}\| = \|\boldsymbol{x} - \boldsymbol{y}\|,\ \forall \boldsymbol{y} \in C.$$

which can only be true if $\boldsymbol{x} \in C$ and hence $T_1(\boldsymbol{x}) = \boldsymbol{x}$. Note that the previous two facts are valid for general Hilbert spaces.

Fact3: If C is a closed subspace, then $T_k,\ k = 1, \ldots, K$, and $T = T_K T_{K-1} \cdots T_1$ are linear operators. The proof is trivial from the respective linearity of the projection operators, $P_k,\ k = 1, \ldots, K$. This is also true for general Hilbert spaces.

Fact 4: The operator T is a *regular* one, i.e.,

$$\|T^n(\boldsymbol{x}) - T^{n-1}(\boldsymbol{x})\| \underset{n \longrightarrow \infty}{\longrightarrow} 0.$$

Recall from Problem 20 that $\forall \boldsymbol{x} \in \mathbb{H}$,

$$\|T_1(\boldsymbol{x}) - \boldsymbol{y}\|^2 \leq \|\boldsymbol{x} - \boldsymbol{y}\|^2 - \mu_1(2 - \mu_1)\|\boldsymbol{x} - P_1(\boldsymbol{x})\|^2$$

or

$$\|\boldsymbol{x} - P_1(\boldsymbol{x})\|^2 \leq \frac{1}{\mu_1(2 - \mu_1)} \left(\|\boldsymbol{x} - \boldsymbol{y}\|^2 - \|T_1(\boldsymbol{x}) - \boldsymbol{y}\|^2\right),$$

and by the definition

$$T_1(\boldsymbol{x}) = \boldsymbol{x} + \mu_1(P_1(\boldsymbol{x}) - \boldsymbol{x}),$$

we get

$$\|\boldsymbol{x} - T_1(\boldsymbol{x})\| = \mu_1^2\|\boldsymbol{x} - P_1(\boldsymbol{x})\|^2 \leq \frac{\mu_1}{2 - \mu_1} \left(\|\boldsymbol{x} - \boldsymbol{y}\|^2 - \|T_1(\boldsymbol{x}) - \boldsymbol{y}\|^2\right).$$

Let now

$$\begin{aligned}
\|\boldsymbol{x} - T_2T_1(\boldsymbol{x})\|^2 &= \|\boldsymbol{x} - T_1(\boldsymbol{x}) + T_1(\boldsymbol{x}) - T_2T_1(\boldsymbol{x})\|^2 \\
&\leq (\|\boldsymbol{x} - T_1(\boldsymbol{x})\| + \|T_2T_1(\boldsymbol{x}) - T_1(\boldsymbol{x})\|)^2 \\
&\leq 2\left(\|\boldsymbol{x} - T_1(\boldsymbol{x})\|^2 + \|T_2T_1(\boldsymbol{x}) - T_1(\boldsymbol{x})\|^2\right),
\end{aligned}$$

or

$$\begin{aligned}\|\boldsymbol{x} - T_2 T_1(\boldsymbol{x})\|^2 &\leq \frac{2\mu_1}{2-\mu_1}(\|\boldsymbol{x}-\boldsymbol{y}\|^2 - \|T_1(\boldsymbol{x}) - \boldsymbol{y}\|^2) \\ &+ \frac{2\mu_2}{2-\mu_2}(\|T_1(\boldsymbol{x})-\boldsymbol{y}\|^2 - \|T_2T_1(\boldsymbol{x}) - \boldsymbol{y}\|^2).\end{aligned}$$

Let

$$b_2 = \max\left\{\frac{2\mu_1}{2-\mu_1},\ \frac{2\mu_2}{2-\mu_2}\right\}.$$

Then obviously, we can write

$$\|\boldsymbol{x} - T_2T_1(\boldsymbol{x})\|^2 \leq 2b_2(\|\boldsymbol{x}-\boldsymbol{y}\|^2 - \|T_{12}(\boldsymbol{x}) - \boldsymbol{y}\|^2),$$

where

$$T_{12}(\boldsymbol{x}) = T_2\big(T_1(\boldsymbol{x})\big).$$

Following a similar rationale and by induction we can show that

$$\|\boldsymbol{x} - T(\boldsymbol{x})\|^2 \leq b_K 2^{K-1}(\|\boldsymbol{x}-\boldsymbol{y}\|^2 - \|T(\boldsymbol{x}) - \boldsymbol{y}\|^2), \tag{17}$$

where,

$$T = T_K T_{K-1}\cdots T_1,$$

and

$$b_K = \max_{1\leq k\leq K}\left\{\frac{\mu_k}{2-\mu_k}\right\}.$$

Now by induction,

$$\|T(\boldsymbol{x}) - T^2(\boldsymbol{x})\|^2 \leq b_K 2^{K-1}(\|T(\boldsymbol{x}) - \boldsymbol{y})\|^2 - \|T^2(\boldsymbol{x}) - \boldsymbol{y}\| \tag{18}$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$\|T^{n-1}(\boldsymbol{x}) - T^n(\boldsymbol{x})\|^2 \leq b_K 2^{K-1}(\|T^{n-1}(\boldsymbol{x}) - \boldsymbol{y})\|^2 - \|T^n(\boldsymbol{x}) - \boldsymbol{y}\| \tag{19}$$

Summing by parts (17)–(19), we obtain

$$\sum_{n=1}^{\infty}\|T^{n-1}(\boldsymbol{x}) - T^n(\boldsymbol{x})\|^2 \leq b_K 2^{K-1}\|\boldsymbol{x}-\boldsymbol{y}\|^2 < +\infty.$$

Hence,

$$\lim_{n\longrightarrow\infty}\|T^{n-1}(\boldsymbol{x}) - T^n(\boldsymbol{x})\| = 0.$$

Note that till now, everything is valid for general Hilbert spaces.

23. Show the fundamental POCS theorem for the case of closed subspaces in a Hilbert space, $\mathbb{H}$.

    Solution: Fact 1: The relaxed projection operator is self adjoint, i.e.,

    $$\langle \boldsymbol{x}, T_{C_i}(\boldsymbol{y})\rangle = \langle T_{C_i}(\boldsymbol{x}), \boldsymbol{y}\rangle,\ \forall \boldsymbol{x},\boldsymbol{y}\in\mathbb{H}.$$

This is a direct consequence of the self-adjoint property of the projection, when C_k is a closed subspace, i.e.,

$$\langle \boldsymbol{x}, P_{C_k}(\boldsymbol{y})\rangle = \langle P_{C_k}(\boldsymbol{x}), \boldsymbol{y}\rangle = \langle P_{C_k}(\boldsymbol{x}), P_{C_k}(\boldsymbol{y})\rangle.$$

Fact 2: For a closed subspace, C_k, the respected relaxed projection operator is linear, i.e.,

$$T_{C_k}(a\boldsymbol{x} + b\boldsymbol{y}) = aT_{C_k}(\boldsymbol{x}) + bT_{C_k}(\boldsymbol{y}),\ \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{H}.$$

This is also a direct consequence of the linear property of the projection operator onto subspaces. This property is easily checked out that it is readily transferred to $T := T_{C_K} \cdots T_{C_1}$.

Fact 3:

$$\begin{array}{rcl} \langle \boldsymbol{x}, T(\boldsymbol{y})\rangle & = & \langle \boldsymbol{x}, T_{C_K} \cdots T_{C_1}(\boldsymbol{y}) \\ & = & \langle T_{C_K}(\boldsymbol{x}), T_{C_{K-1}} \cdots T_{C_1}(\boldsymbol{y}) \\ & = & \ldots \\ & = & \langle T_{C_1} \cdots T_{C_K}(\boldsymbol{x}), \boldsymbol{y}\rangle \\ & = & \langle T^*(\boldsymbol{x}), \boldsymbol{y}\rangle, \end{array}$$

where $T^* = T_{C_1}T_{C_2} \cdots T_{C_K}$.

Fact 4: Let the operator $T = T_{C_K} \cdots T_{C_1}$, with $\text{Fix}(T) = C$, where C is a closed subspace. Then, the set

$$S := \{\boldsymbol{y} : \boldsymbol{y} = (I - T)(\boldsymbol{z}),\ \forall \boldsymbol{z} \in \mathbb{H}\}$$

is also a (closed) subspace and it is the orthogonal complement of C, i.e.,

$$S = C^{\perp}.$$

The proof that S is a subspace is trivial, from the linearity of T. Also, let $\boldsymbol{x} \in S^{\perp}$. Then, by the respective definition,

$$\begin{array}{rcl} 0 & = & \langle \boldsymbol{x}, (I - T)(\boldsymbol{z})\rangle = \langle \boldsymbol{x}, \boldsymbol{z}\rangle - \langle \boldsymbol{x}, T(\boldsymbol{z})\rangle \\ & = & \langle \boldsymbol{x}, \boldsymbol{z}\rangle - \langle T^*(\boldsymbol{x}), \boldsymbol{z}\rangle = \langle (I - T^*)(\boldsymbol{x}), \boldsymbol{z}\rangle,\ \forall \boldsymbol{z} \in \mathbb{H}. \end{array}$$

Hence,

$$(I - T^*)(\boldsymbol{x}) = \boldsymbol{x} - T^*(\boldsymbol{x}) = \boldsymbol{0},$$

or

$$T^*(\boldsymbol{x}) = \boldsymbol{x},$$

and since T^* and T have the same fixed point set (the proof trivial),

$$S^{\perp} \subseteq C.$$

Let now $\boldsymbol{x} \in C$. Then

$$\begin{aligned}\langle \boldsymbol{x}, (I-T)\boldsymbol{z}\rangle &= \langle \boldsymbol{x}, \boldsymbol{z}\rangle - \langle \boldsymbol{x}, T(\boldsymbol{z})\rangle \\ &= \langle \boldsymbol{x}, \boldsymbol{z}\rangle - \langle T^*(\boldsymbol{x}), \boldsymbol{z}, \rangle \\ &= \langle \boldsymbol{x} - T^*(\boldsymbol{x}), \boldsymbol{z}\rangle = 0,\end{aligned}$$

since

$$T^*(\boldsymbol{x}) = \boldsymbol{x},$$

which proves that

$$S^{\perp} = C.$$

Note that what we have said so far is a generalization of Problem 18. We are now ready to establish *strong convergence*.

The repeated application of T on any $\boldsymbol{x} \in \mathbb{H}$ leads to $T^n(\boldsymbol{x}) = (TTT)^n(\boldsymbol{x})$. We know that $\forall \boldsymbol{x} \in \mathbb{H}$ there is a unique decomposition into two orthogonal complement (closed) subspaces, i.e.,

$$\boldsymbol{x} = \boldsymbol{y} + \boldsymbol{z},\ \boldsymbol{y} \in C \text{ and } \boldsymbol{z} \in C^{\perp},\ \forall \boldsymbol{x} \in \mathbb{H}$$

and that $\boldsymbol{y} = P_C(\boldsymbol{x})$.

Hence, due to the linearity of T^n (C subspace in $\mathbb{H}$)

$$\begin{aligned}T^n(\boldsymbol{x}) &= T^n(\boldsymbol{y}) + T^n(\boldsymbol{z}) \\ &= \boldsymbol{y} + T^n(\boldsymbol{z}),\end{aligned}$$

since $C = \text{Fix}(T^n)$. However,

$$\begin{aligned}T^n(\boldsymbol{z}) &= T^n(I-T)\boldsymbol{w},\ \text{for some} \boldsymbol{w} \in \mathbb{H} \\ &= T^n(\boldsymbol{w}) - T^{n+1}(\boldsymbol{w})\end{aligned}$$

and we know that

$$\|T^n(\boldsymbol{z})\| = \|T^n(\boldsymbol{w}) - T^{n+1}(\boldsymbol{w})\| \longrightarrow 0.$$

Thus,

$$\|T^n(\boldsymbol{x}) - P_C(\boldsymbol{x})\| \longrightarrow 0.$$

which proves the claim.

24. Derive the subdifferential of the metric distance function $d_C(\boldsymbol{x})$, where C is a closed convex set $C \subseteq \mathbb{R}^l$ and $\boldsymbol{x} \in \mathbb{R}^l$.

Solution: By definition we have

$$\partial d_C(\boldsymbol{x}) = \{\boldsymbol{g} : \boldsymbol{g}^T(\boldsymbol{y} - \boldsymbol{x}) + d_C(\boldsymbol{x}) \leq d_C(\boldsymbol{y}),\ \forall \boldsymbol{y} \in \mathbb{R}^l\}.$$

Thus let $\boldsymbol{g}$ be a subgradient, then

$$\boldsymbol{g}^T(\boldsymbol{y} - \boldsymbol{x}) \leq d_C(\boldsymbol{y}) - d_C(\boldsymbol{x}) \leq \|\boldsymbol{y} - \boldsymbol{x}\|.$$

The above is easily shown by the respective definition

$$\begin{aligned} d_C(\boldsymbol{y}) = \min_{\boldsymbol{z}\in C} \|\boldsymbol{y}-\boldsymbol{z}\| &\leq \min_{\boldsymbol{z}\in C}(\|\boldsymbol{y}-\boldsymbol{x}\| + \|\boldsymbol{x}-\boldsymbol{z}\|) \\ &\leq \|\boldsymbol{y}-\boldsymbol{x}\| + \min_{\boldsymbol{z}\in C}\|\boldsymbol{x}-\boldsymbol{z}\|, \end{aligned}$$

or

$$d_C(\boldsymbol{y}) - d_C(\boldsymbol{x}) \leq \|\boldsymbol{y}-\boldsymbol{x}\|.$$

Hence,

$$\boldsymbol{g}^T(\boldsymbol{y}-\boldsymbol{x}) \leq \|\boldsymbol{y}-\boldsymbol{x}\|.$$

Since this is true $\forall \boldsymbol{y}$, let

$$\boldsymbol{y} : \boldsymbol{y}-\boldsymbol{x} = \boldsymbol{g} \Rightarrow \|\boldsymbol{g}\|^2 \leq \|\boldsymbol{g}\| \Rightarrow \|\boldsymbol{g}\| \leq 1,$$

or

$$\boldsymbol{g} \in B[\mathbf{0}, 1].$$

a) Let $\boldsymbol{x} \notin C$ and $\boldsymbol{g}$ any subgradient. For any $\boldsymbol{y} \in \mathbb{R}^l$,

$$\boldsymbol{g}^T(\boldsymbol{y} + P_C(\boldsymbol{x}) - \boldsymbol{x}) \leq d_C(\boldsymbol{y} + P_C(\boldsymbol{x})) - d_C(\boldsymbol{x}).$$

However,

$$\begin{aligned} d_C(\boldsymbol{y} + P_C(\boldsymbol{x})) &= \min_{\boldsymbol{z}\in C} \|\boldsymbol{y} + P_C(\boldsymbol{x}) - \boldsymbol{z}\| \\ &\leq \|\boldsymbol{y}\| + \min_{\boldsymbol{z}\in C}\|P_C(\boldsymbol{x}) - \boldsymbol{z}\|, \end{aligned}$$

and letting $\boldsymbol{z} = P_C(\boldsymbol{x})$

$$d_C(\boldsymbol{y} + P_C(\boldsymbol{x})) \leq \|\boldsymbol{y}\|.$$

Hence, we can write that

$$\boldsymbol{g}^T(\boldsymbol{y} + P_C(\boldsymbol{x}) - \boldsymbol{x}) \leq \|\boldsymbol{y}\| - d_C(\boldsymbol{x}),\ \forall \boldsymbol{y} \in \mathbb{R}^l.$$

Set $\boldsymbol{y} = \mathbf{0}$. Then,

$$-\boldsymbol{g}^T(\boldsymbol{x} - P_C(\boldsymbol{x})) \leq -\|\boldsymbol{x} - P_C(\boldsymbol{x})\|,$$

or

$$\boldsymbol{g}^T(\boldsymbol{x} - P_C(\boldsymbol{x})) \geq \|\boldsymbol{x} - P_C(\boldsymbol{x})\|.$$

However,

$$\|\boldsymbol{g}\| \leq 1,$$

and recalling the Cauchy-Schwartz inequality, we obtain

$$\|\boldsymbol{x} - P_C(\boldsymbol{x})\|\|\boldsymbol{g}\| \geq \|\boldsymbol{x} - P_C(\boldsymbol{x})\| \Rightarrow \|\boldsymbol{g}\| = 1,$$

and

$$\boldsymbol{g}^T(\boldsymbol{x} - P_C(\boldsymbol{x})) = \|\boldsymbol{x} - P_C(\boldsymbol{x})\|,$$

which implies (recall condition for equality in the Cauchy-Schwartz theorem)

$$\boldsymbol{g} = \frac{(\boldsymbol{x} - P_C(\boldsymbol{x}))}{\|(\boldsymbol{x} - P_C(\boldsymbol{x}))\|},$$

which proves the claim.
b) Let $\boldsymbol{x} \in C$. Then by definition, we have

$$\boldsymbol{g}^T(\boldsymbol{y} - \boldsymbol{x}) \leq d_C(\boldsymbol{y}) - d_C(\boldsymbol{x}),$$

and for any $\boldsymbol{y} \in C$

$$\boldsymbol{g}^T(\boldsymbol{y} - \boldsymbol{x}) \leq 0, \ \|\boldsymbol{g}\| \leq 1. \tag{20}$$

If in addition $\boldsymbol{x}$ is an interior point, there will be $\varepsilon > 0 : \forall \boldsymbol{z} \in \mathbb{R}^l$

$$\boldsymbol{g}^T(\boldsymbol{x} - \varepsilon(\boldsymbol{z} - \boldsymbol{x}) - \boldsymbol{x}) \leq 0,$$

since $\boldsymbol{x} - \varepsilon(\boldsymbol{z} - \boldsymbol{x}) \in C$ and the condition (20) has been used. Thus,

$$\boldsymbol{g}^T(\boldsymbol{z} - \boldsymbol{x}) \leq 0, \ \forall \boldsymbol{z} \in \mathbb{R}^l.$$

Set $\boldsymbol{z} - \boldsymbol{x} = \boldsymbol{g}$, which leads to

$$\boldsymbol{g} = \boldsymbol{0}.$$

This completes the proof.

25. Derive the bound in (8.55).
Solution: Subtracting θ_* from both sides of the recursion, squaring and taking into account the definition of the subgradient, it is readily shown that,

$$\begin{aligned} ||\boldsymbol{\theta}^{(i)} - \boldsymbol{\theta}_*||^2 &\leq ||\boldsymbol{\theta}^{(i-1)} - \boldsymbol{\theta}_*||^2 - 2\mu_i\big(J(\boldsymbol{\theta}^{(i-1)}) - J(\boldsymbol{\theta}_*)\big) + \\ & \quad \mu_i^2||J'(\boldsymbol{\theta}^{(i-1)})||^2. \end{aligned}$$

Applying the previous recursively, we obtain

$$\begin{aligned} ||\boldsymbol{\theta}^{(i)} - \boldsymbol{\theta}_*||^2 &\leq ||\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}_*||^2 - 2\sum_{k=1}^{i}\mu_k\big(J(\boldsymbol{\theta}^{(k-1)}) - J(\boldsymbol{\theta}_*)\big) + \\ & \quad \sum_{k=1}^{i}\mu_k^2||J'(\boldsymbol{\theta}^{(k-1)})||^2. \end{aligned}$$

Taking into account the bound of the subgradient and the fact that the left hand side of the inequality is a non-negative number, we obtain

$$2\sum_{k=1}^{i}\mu_k\big(J(\boldsymbol{\theta}^{(k-1)}) - J(\boldsymbol{\theta}_*)\big) \leq ||\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}_*||^2 + \sum_{k=1}^{i}\mu_k^2 G^2. \tag{21}$$

However, by the respective definition we get

$$J(\boldsymbol{\theta}^{(k-1)}) - J(\boldsymbol{\theta}_*) \geq J_*^{(i)} - J(\boldsymbol{\theta}_*),\ k = 1, \ldots, i.$$

Employing the previous bound in (21), the claim is readily obtained, i.e.,

$$J_*^{(i)} - J(\boldsymbol{\theta}) \leq \frac{||\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}_*||^2}{2\sum_{k=1}^{i} \mu_k} + \frac{\sum_{k=1}^{i} \mu_k^2}{2\sum_{k=1}^{i} \mu_k} G^2.$$

26. Show that if a function is γ-Lipschitz, then any of its subgradients is bounded.

Solution: By the definition of the subgradient we have that $\forall u, v$ we have that

$$f(u) - f(v) \geq < f^{'}(v)(u-v) > .$$

Since this is true for all, u, v, I can always select a u so that $u - v$ to be parallel to $f^{'}(v)$, Then

$$< f^{'}(v)(u-v) >= | < f^{'}(v)(u-v) > | = ||f^{'}(v)||||u-v||.$$

Then, if we plug in the Lipschitz condition, we show that $||f^{'}(v)||$ is bounded.

27. Show the convergence of the generic projected subgradient algorithm in (8.61).

Solution: Let us break the iteration into two steps,

$$\boldsymbol{z}^{(i)} = \boldsymbol{\theta}^{(i-1)} - \mu_i J'(\boldsymbol{\theta}^{(i-1)}), \tag{22}$$

$$\boldsymbol{\theta}^{(i)} = P_C(\boldsymbol{z}^{(i)}). \tag{23}$$

Then, following the same arguments as the ones adopted in Problem 25, we get

$$\begin{aligned} ||\boldsymbol{z}^{(i)} - \boldsymbol{\theta}_*||^2 \leq\ & ||\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}_*||^2 - 2\sum_{k=1}^{i} \mu_k \big(J(\boldsymbol{\theta}^{(k-1)}) - J(\boldsymbol{\theta}_*)\big) + \\ & \sum_{k=1}^{i} \mu_k^2 ||J'(\boldsymbol{\theta}^{(k-1)})||^2. \end{aligned} \tag{24}$$

However, from the non-expansive property of the projection operator, and taking into account that $\boldsymbol{\theta}_* \in C$, since it is a solution,

$$||\boldsymbol{\theta}^{(i)} - \boldsymbol{\theta}_*||^2 = ||P_C(\boldsymbol{z}^{(i)}) - P_C(\boldsymbol{\theta}_*)||^2 \leq ||\boldsymbol{z}^{(i)} - \boldsymbol{\theta}_*||^2. \tag{25}$$

Combining the last two formulas the proof proceeds as in Problem 25.

28. Derive equation (8.100).

Solution: By the definition

$$J_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{k=1}^{n} \mathcal{L}(y_k, \boldsymbol{x}_k, \boldsymbol{\theta}).$$

Hence,

$$\begin{aligned} J_n(\boldsymbol{\theta}) &= \frac{1}{n}\left(\sum_{k=1}^{n-1} \mathcal{L}(y_k, \boldsymbol{x}_k, \boldsymbol{\theta}) + \mathcal{L}(y_n, \boldsymbol{x}_n, \boldsymbol{\theta})\right) \\ &= \frac{n-1}{n} J_{n-1}(\boldsymbol{\theta}) + \frac{1}{n}\mathcal{L}(y_n, \boldsymbol{x}_n, \boldsymbol{\theta}), \end{aligned}$$

or

$$\nabla J_n(\boldsymbol{\theta}) = \frac{n-1}{n}\nabla J_{n-1}(\boldsymbol{\theta}) + \frac{1}{n}\nabla\mathcal{L}(y_n, \boldsymbol{x}_n, \boldsymbol{\theta}).$$

Hence

$$\nabla J_n(\boldsymbol{\theta}_*(n-1)) = 0 + \frac{1}{n}\nabla\mathcal{L}(y_n, \boldsymbol{x}_n, \boldsymbol{\theta}_*(n-1)).$$

Expanding the left hand side to a first order Taylor approximation we get,

$$\nabla J_n(\boldsymbol{\theta}_*(n)) = \mathbf{0} = \nabla J_n(\boldsymbol{\theta}_*(n-1)) + \nabla^2 J_n(\boldsymbol{\theta}_*(n-1))\left(\boldsymbol{\theta}_*(n) - \boldsymbol{\theta}_*(n-1)\right),$$

or

$$\nabla J_n(\boldsymbol{\theta}_*(n-1)) = \nabla^2 J_n(\boldsymbol{\theta}_*(n-1))\left(\boldsymbol{\theta}_*(n) - \boldsymbol{\theta}_*(n-1)\right),$$

which finally proves the claim.

29. Consider the online version of PDMb in (8.64), i.e.,

$$\boldsymbol{\theta}_n = \begin{cases} P_C\Big(\boldsymbol{\theta}_{n-1} - \mu_n \frac{J(\boldsymbol{\theta}_{n-1})}{||J'(\boldsymbol{\theta}^{n-1})||^2} J'(\boldsymbol{\theta}^{n-1})\Big), & \text{If } J'(\boldsymbol{\theta}^{n-1}) \neq \mathbf{0}, \\ P_C(\boldsymbol{\theta}^{n-1}), & \text{If } J'(\boldsymbol{\theta}^{n-1}) = \mathbf{0}, \end{cases} \tag{26}$$

where we have assumed that $J_* = 0$. If this is not the case, a shift can accommodate for the difference. Thus we assume that we know the minimum. For example, this is the case for a number tasks, such as the hinge loss function, assuming linearly separable classes, or the linear ϵ-insensitive loss function, for bounded noise. Assume that

$$\mathcal{L}_n(\boldsymbol{\theta}) = \sum_{k=n-q+1}^{n} \frac{\omega_k d_{C_k}(\boldsymbol{\theta}_{n-1})}{\sum_{k=n-q+1}^{n}\omega_k d_{C_k}(\boldsymbol{\theta}_{n-1})} d_{C_k}(\boldsymbol{\theta})$$

Then derive that APSM algorithm of (8.39).

Solution: Let the loss function be

$$\mathcal{L}_n(\boldsymbol{\theta}) = \sum_{k=n-q+1}^{n} \frac{\omega_k d_{C_k}(\boldsymbol{\theta}_{n-1})}{\sum_{k=n-q+1}^{n}\omega_k d_{C_k}(\boldsymbol{\theta}_{n-1})} d_{C_k}(\boldsymbol{\theta}) = \sum_{k=n-q+1}^{n} \beta_k d_{C_k}(\boldsymbol{\theta}),$$

where

$$\sum_{k=n-q+1}^{n} \beta_k = 1.$$

Then for the recursion (26), we need to compute the subgradient of $\mathcal{L}_n(\boldsymbol{\theta})$, which by Example 8.5 (and for $\boldsymbol{\theta} \notin C_k$) becomes

$$\acute{\mathcal{L}}_n(\boldsymbol{\theta}_{n-1}) = \sum_{k=n-q+1}^{n} \beta_k d'_{C_k}(\boldsymbol{\theta})\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{n-1}} = \sum_{k=n-q+1}^{n} \beta_k \frac{\boldsymbol{\theta}_{n-1} - P_{C_k}(\boldsymbol{\theta}_{n-1})}{d_{C_k}(\boldsymbol{\theta}_{n-1})},$$

or

$$\begin{aligned} \acute{\mathcal{L}}_n(\boldsymbol{\theta}_{n-1}) &= \frac{1}{L} \sum_{k=n-q+1}^{n} \omega_k(\boldsymbol{\theta}_{n-1} - P_{C_k}(\boldsymbol{\theta}_{n-1})), \text{ with} \\ L &= \sum_{k=n-q+1}^{n} \omega_k d_{C_k}(\boldsymbol{\theta}_{n-1}). \end{aligned}$$

Hence, (26) now becomes (using μ'_n instead, for reasons to become apparent soon),

$$\begin{aligned} \boldsymbol{\theta}_n &= \boldsymbol{\theta}_{n-1} - \mu'_n \frac{\frac{1}{L}\sum_{k=n-q+1}^{n} \omega_k d^2_{C_k}(\boldsymbol{\theta}_{n-1})}{\frac{1}{L}^2 \| \sum_{k=n-q+1}^{n} \omega_k(\boldsymbol{\theta}_{n-1} - P_{C_k}(\boldsymbol{\theta}_{n-1})\|^2} \\ &\quad \frac{1}{L} \sum_{k=n-q+1}^{n} \omega_k(\boldsymbol{\theta}_{n-1} - P_{C_k}(\boldsymbol{\theta}_{n-1})) \\ &= \boldsymbol{\theta}_{n-1} + \mu'_n M \left(\sum_{k=n-q+1}^{n} \omega_k(P_{C_k}(\boldsymbol{\theta}_{n-1}) - \boldsymbol{\theta}_{n-1})) \right) \\ &= \boldsymbol{\theta}_{n-1} + \mu'_n M \left(\sum_{k=n-q+1}^{n} \omega_k P_{C_k}(\boldsymbol{\theta}_{n-1}) - \boldsymbol{\theta}_{n-1} \right) \end{aligned}$$

and

$$M := \frac{\sum_{k=n-q+1}^{n} \omega_k d^2_{C_k}(\boldsymbol{\theta}_{n-1})}{\| \sum_{k=n-q+1}^{n} \omega_k P_{C_k}(\boldsymbol{\theta}_{n-1}) - \boldsymbol{\theta}_{n-1}\|^2}.$$

Setting

$$\mu_n = \mu'_n M \in (0, 2M)$$

the APSM algorithm results.

30. Derive the regret bound for the subgradient algorithm in (8.83).

Solution: From the text, we have that

$$\begin{aligned}\mathcal{L}_n(\boldsymbol{\theta}_{n-1}) - \mathcal{L}_n(\boldsymbol{h}) &\leq \boldsymbol{g}_n^T(\boldsymbol{\theta}_{n-1} - \boldsymbol{h}) \\ &\leq \frac{1}{2\mu_n}\Big(||\boldsymbol{\theta}_{n-1} - \boldsymbol{h}||^2 - ||\boldsymbol{\theta}_n - \boldsymbol{h}||^2\Big) + \\ &\quad \frac{\mu_n}{2}G^2. \qquad (27)\end{aligned}$$

Summing up both sides, results in

$$\begin{aligned}\sum_{n=1}^{N}\mathcal{L}_n(\boldsymbol{\theta}_{n-1}) - \sum_{n=1}^{N}\mathcal{L}_n(\boldsymbol{h}) &\leq \sum_{n=1}^{N}\frac{1}{2\mu_n}\Big(||\boldsymbol{\theta}_{n-1} - \boldsymbol{h}||^2 - ||\boldsymbol{\theta}_n - \boldsymbol{h}||^2\Big) + \\ &\quad \frac{G^2}{2}\sum_{n=1}^{N}\mu_n. \qquad (28)\end{aligned}$$

Carrying out the summations on the left hand side we get

$$\begin{aligned}A &:= \frac{1}{2\mu_1}||\boldsymbol{\theta}_0 - \boldsymbol{h}||^2 - \frac{1}{2\mu_1}||\boldsymbol{\theta}_1 - \boldsymbol{h}||^2 + \\ &\quad \frac{1}{2\mu_2}||\boldsymbol{\theta}_1 - \boldsymbol{h}||^2 - \frac{1}{2\mu_2}||\boldsymbol{\theta}_2 - \boldsymbol{h}||^2 + \\ &\quad \cdots\cdots\cdots \\ &\quad \frac{1}{2\mu_N}||\boldsymbol{\theta}_{N-1} - \boldsymbol{h}||^2 - \frac{1}{2\mu_N}||\boldsymbol{\theta}_N - \boldsymbol{h}||^2 + \\ &\quad \frac{G^2}{2}\sum_{n=1}^{N}\mu_n,\end{aligned}$$

or

$$\begin{aligned}A &\leq \frac{1}{2\mu_1}||\boldsymbol{\theta}_0 - \boldsymbol{h}||^2 - \frac{1}{2\mu_1}||\boldsymbol{\theta}_1 - \boldsymbol{h}||^2 + \\ &\quad \frac{1}{2\mu_2}||\boldsymbol{\theta}_1 - \boldsymbol{h}||^2 - \frac{1}{2\mu_2}||\boldsymbol{\theta}_2 - \boldsymbol{h}||^2 + \\ &\quad \cdots\cdots\cdots \\ &\quad \frac{1}{2\mu_N}||\boldsymbol{\theta}_{N-1} - \boldsymbol{h}||^2 + \frac{G^2}{2}\sum_{n=1}^{N}\mu_n,\end{aligned}$$

Taking into account the bound $||\boldsymbol{\theta}_n - \boldsymbol{h}||^2 \leq F^2$, and selecting the step-size to be a decreasing sequence, we readily get

$$A \leq F^2\Big(\frac{1}{2\mu_1} + \frac{1}{2}\sum_{n=2}^{N}(\frac{1}{\mu_n} - \frac{1}{\mu_{n-1}})\Big) + \frac{G^2}{2}\sum_{n=1}^{N}\mu_n, \qquad (29)$$

which then easily leads to

$$A \leq \frac{1}{2\mu_N}F^2 + \frac{G^2}{2}\sum_{n=1}^{N}\mu_n, . \qquad (30)$$

Combining the above with (28), the claim is proved.

31. Show that a function $f(\boldsymbol{x})$ is σ-strongly convex if and only if the function $f(\boldsymbol{x}) - \frac{\sigma}{2}||\boldsymbol{x}||^2$ is convex.

Solution:
a) Assume that

$$f(\boldsymbol{x}) - \frac{\sigma}{2}||\boldsymbol{x}||^2,$$

is convex. Then, by the definition of the subgradient at $\boldsymbol{x}$, we have

$$f(\boldsymbol{y}) - \frac{\sigma}{2}||\boldsymbol{y}||^2 - f(\boldsymbol{x}) + \frac{\sigma}{2}||\boldsymbol{x}||^2 \geq \boldsymbol{g}^T(\boldsymbol{y}-\boldsymbol{x}) - \sigma\boldsymbol{x}^T(\boldsymbol{y}-\boldsymbol{x}), \tag{31}$$

which readily implies that

$$f(\boldsymbol{y}) - f(\boldsymbol{x}) \geq \boldsymbol{g}^T(\boldsymbol{y}-\boldsymbol{x}) + \frac{\sigma}{2}||\boldsymbol{y}-\boldsymbol{x}||^2, \tag{32}$$

from which the strong convexity of $f(\boldsymbol{x})$ is deduced.

b) Assume that $f(\boldsymbol{x})$ is strongly convex. Then by its definition we have,

$$\begin{aligned} f(\boldsymbol{y}) - f(\boldsymbol{x}) \quad \geq \quad & \boldsymbol{g}^T(\boldsymbol{y}-\boldsymbol{x}) + \frac{\sigma}{2}||\boldsymbol{y}||^2 + \frac{\sigma}{2}||\boldsymbol{x}||^2 - \sigma\boldsymbol{x}^T\boldsymbol{y} \\ & +\sigma||\boldsymbol{x}|||^2 - \sigma||\boldsymbol{x}||^2, \end{aligned}$$

or

$$f(\boldsymbol{y}) - f(\boldsymbol{x}) \geq \boldsymbol{g}^T(\boldsymbol{y}-\boldsymbol{x}) + \frac{\sigma}{2}||\boldsymbol{y}||^2 - \frac{\sigma}{2}||\boldsymbol{x}||^2 - \sigma\boldsymbol{x}^T\boldsymbol{y} + \sigma||\boldsymbol{x}||^2, \tag{33}$$

from which we obtain,

$$f(\boldsymbol{y}) - \frac{\sigma}{2}||\boldsymbol{y}||^2 - f(\boldsymbol{x}) + \frac{\sigma}{2}||\boldsymbol{x}||^2 \geq \boldsymbol{g}^T(\boldsymbol{y}-\boldsymbol{x}) - \sigma\boldsymbol{x}^T(\boldsymbol{y}-\boldsymbol{x}), \tag{34}$$

which proves the claim that $f(\boldsymbol{x}) - \frac{\sigma}{2}||\boldsymbol{x}||^2$ is convex.

32. Show that if the loss function is σ-strongly convex, then if $\mu_n = \frac{1}{\sigma n}$, the regret bound for the subgradient algorithm becomes

$$\frac{1}{N}\sum_{n=1}^{N}\mathcal{L}_n(\boldsymbol{\theta}_{n-1}) \leq \frac{1}{N}\sum_{n=1}^{N}\mathcal{L}_n(\boldsymbol{\theta}_*) + \frac{G^2(1+\ln N)}{2\sigma N}. \tag{35}$$

Solution: Taking into account the strong convexity we have that,

$$\mathcal{L}_n(\boldsymbol{\theta}_{n-1}) - \mathcal{L}_n(\boldsymbol{\theta}_*) \leq \boldsymbol{g}_n^T(\boldsymbol{\theta}_{n-1} - \boldsymbol{\theta}_*) - \frac{\sigma}{2}||\boldsymbol{\theta}_{n-1} - \boldsymbol{\theta}_*||^2, \tag{36}$$

and following similar arguments as for Problem 30, we get

$$\begin{aligned} \mathcal{L}_n(\boldsymbol{\theta}_{n-1}) - \mathcal{L}_n(\boldsymbol{\theta}_*) \quad \leq \quad & \frac{1}{2\mu_n}\Big(||\boldsymbol{\theta}_{n-1} - \boldsymbol{\theta}_*||^2 - ||\boldsymbol{\theta}_n - \boldsymbol{\theta}_*||^2\Big) - \\ & \frac{\sigma}{2}||\boldsymbol{\theta}_{n-1} - \boldsymbol{\theta}_*||^2 + \frac{\mu_n}{2}G^2. \end{aligned} \tag{37}$$

Using $\mu_n = \frac{1}{\sigma n}$, results in

$$2\Big(\mathcal{L}_n(\boldsymbol{\theta}_{n-1}) - \mathcal{L}_n(\boldsymbol{\theta}_*)\Big) \leq \sigma n\Big(||\boldsymbol{\theta}_{n-1} - \boldsymbol{\theta}_*||^2 - ||\boldsymbol{\theta}_n - \boldsymbol{\theta}_*||^2\Big) - \sigma||\boldsymbol{\theta}_{n-1} - \boldsymbol{\theta}_*||^2 + \frac{1}{\sigma n}G^2. \tag{38}$$

Summing up both sides we obtain

$$\begin{aligned}
&2\sum_{n=1}^{N}\Big(\mathcal{L}_n(\boldsymbol{\theta}_{n-1}) - \mathcal{L}_n(\boldsymbol{\theta}_*)\Big) \leq \\
&\sigma\Big(||\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*||^2 - ||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_*||^2\Big) - \sigma||\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*||^2 + \\
&2\sigma\Big(||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_*||^2 - ||\boldsymbol{\theta}_2 - \boldsymbol{\theta}_*||^2\Big) - \sigma||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_*||^2 + \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
&N\sigma\Big(||\boldsymbol{\theta}_{N-1} - \boldsymbol{\theta}_*||^2 - ||\boldsymbol{\theta}_N - \boldsymbol{\theta}_*||^2\Big) - \sigma||\boldsymbol{\theta}_{N-1} - \boldsymbol{\theta}_*||^2 + \\
&G^2\sum_{n=1}^{N}\frac{1}{\sigma n},
\end{aligned}$$

or

$$\begin{aligned}
&2\sum_{n=1}^{N}\Big(\mathcal{L}_n(\boldsymbol{\theta}_{n-1}) - \mathcal{L}_n(\boldsymbol{\theta}_*)||^2\Big) \leq \\
&\sigma\Big(||\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*||^2 - ||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_*\Big) - \sigma||\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*||^2 + \\
&2\sigma\Big(||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_*||^2 - ||\boldsymbol{\theta}_2 - \boldsymbol{\theta}_*\Big) - \sigma||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_*||^2 + \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
&N\sigma\Big(||\boldsymbol{\theta}_{N-1} - \boldsymbol{\theta}_*||^2\Big) - \sigma||\boldsymbol{\theta}_{N-1} - \boldsymbol{\theta}_*||^2 + \\
&G^2\sum_{n=1}^{N}\frac{1}{\sigma n} \\
&\leq G^2\sum_{n=1}^{N}\frac{1}{\sigma n}.
\end{aligned}$$

Using now the bound

$$\sum_{n=1}^{N}\frac{1}{n} \leq 1 + \int_1^N \frac{1}{t}dt = (1 + \ln N),$$

the claim is proved.

33. Consider a batch algorithm that computes the minimum of the empirical loss function, $\boldsymbol{\theta}_*(N)$, having a quadratic convergence rate, i.e.,

$$\ln\ln \frac{1}{||\boldsymbol{\theta}^{(i)} - \boldsymbol{\theta}_*(N)||^2} \sim i.$$

Show that an online algorithm, running for n time instants so that to spend the same computational processing resources as the batch one, achieves for large values of N better performance than the batch algorithm, i.e., [12],

$$||\boldsymbol{\theta}_n - \boldsymbol{\theta}_*||^2 \sim \frac{1}{N\ln\ln N} << \frac{1}{N} \sim ||\boldsymbol{\theta}_*(N) - \boldsymbol{\theta}_*||^2.$$

Hint: Use the fact that

$$||\boldsymbol{\theta}_n - \boldsymbol{\theta}_*||^2 \sim \frac{1}{n}, \quad \text{and} \quad ||\boldsymbol{\theta}_*(N) - \boldsymbol{\theta}_*||^2 \sim \frac{1}{N}.$$

Solution: Let K be the number of operations per iteration for the online algorithm. This amounts to a total of Kn operations. The batch algorithm, in order to make sense, should perform $\mathcal{O}(\ln\ln N)$ operations, so that to get close to $||\boldsymbol{\theta}^{(i)} - \boldsymbol{\theta}_*(N)||^2 \sim 1/N$. Assuming that at each iteration it performs, approximately, $K_1 N$ operations, this amounts to a total of $K_1 N \ln\ln N$ operations. To keep the same load for both algorithms, it should be,

$$Kn = K_1 N \ln\ln N.$$

This leads to the following approximate accuracies,

$$||\boldsymbol{\theta}_n - \boldsymbol{\theta}_*||^2 \sim \frac{1}{N\ln\ln N} << \frac{1}{N} \sim ||\boldsymbol{\theta}_*(N) - \boldsymbol{\theta}_*||^2,$$

which proves the claim. Note that in practice, the values of K and K_1 play an important role as well.

34. Show property (8.111) for the proximal operator.

Solution: Assume first that $\boldsymbol{p} = \mathrm{Prox}_{\lambda f}(\boldsymbol{x})$. By definition,

$$f(\boldsymbol{p}) + \frac{1}{2\lambda}\left\|\boldsymbol{x} - \boldsymbol{p}\right\|^2 \leq f(\boldsymbol{v}) + \frac{1}{2\lambda}\left\|\boldsymbol{x} - \boldsymbol{v}\right\|^2, \quad \forall \boldsymbol{v} \in \mathbb{R}^l.$$

Since the previous inequality holds true for any $\boldsymbol{v} \in \mathbb{R}^l$, it also holds true for $\alpha\boldsymbol{v} + (1-\alpha)\boldsymbol{p}$, where $\boldsymbol{v}$ is any vector in $\mathbb{R}^l$, and α any real number within $(0,1)$. Hence,

$$\begin{aligned}
\lambda f(\boldsymbol{p}) &\leq \lambda f(\alpha \boldsymbol{v} + (1-\alpha)\boldsymbol{p}) + \frac{1}{2}\left\|\boldsymbol{x} - \alpha \boldsymbol{v} - (1-\alpha)\boldsymbol{p}\right\|^2 - \frac{1}{2}\left\|\boldsymbol{x} - \boldsymbol{p}\right\|^2 \\
&\leq \lambda \alpha f(\boldsymbol{v}) + \lambda(1-\alpha) f(\boldsymbol{p}) + \frac{1}{2}\left\|\boldsymbol{x} - \boldsymbol{p}\right\|^2 \\
&\quad + \frac{1}{2}\alpha^2 \left\|\boldsymbol{v} - \boldsymbol{p}\right\|^2 - \alpha \left\langle \boldsymbol{x} - \boldsymbol{p}, \boldsymbol{v} - \boldsymbol{p} \right\rangle - \frac{1}{2}\left\|\boldsymbol{x} - \boldsymbol{p}\right\|^2 \\
&= \lambda \alpha f(\boldsymbol{v}) + \lambda(1-\alpha) f(\boldsymbol{p}) + \frac{1}{2}\alpha^2 \left\|\boldsymbol{v} - \boldsymbol{p}\right\|^2 - \alpha \left\langle \boldsymbol{x} - \boldsymbol{p}, \boldsymbol{v} - \boldsymbol{p} \right\rangle .
\end{aligned}$$

After re-arranging terms in the previous relation,

$$\lambda f(\boldsymbol{p}) \leq \lambda f(\boldsymbol{v}) + \frac{1}{2}\alpha \left\|\boldsymbol{v} - \boldsymbol{p}\right\|^2 - \left\langle \boldsymbol{x} - \boldsymbol{p}, \boldsymbol{v} - \boldsymbol{p} \right\rangle , \quad \forall \alpha \in (0,1).$$

Application of $\lim_{\alpha \to 0}$ on both sides of the previous inequality results in the desired $\langle \boldsymbol{v} - \boldsymbol{p}, \boldsymbol{x} - \boldsymbol{p} \rangle \leq \lambda\big(f(\boldsymbol{v}) - f(\boldsymbol{p})\big)$.

Conversely, assume that $\langle \boldsymbol{v} - \boldsymbol{p}, \boldsymbol{x} - \boldsymbol{p} \rangle / \lambda \leq f(\boldsymbol{v}) - f(\boldsymbol{p})$. Then,

$$\begin{aligned}
f(\boldsymbol{p}) + \frac{1}{2\lambda}\left\|\boldsymbol{x} - \boldsymbol{p}\right\|^2 &\leq f(\boldsymbol{v}) + \frac{1}{2\lambda}\left\|\boldsymbol{x} - \boldsymbol{p}\right\|^2 - \frac{1}{\lambda}\left\langle \boldsymbol{v} - \boldsymbol{p}, \boldsymbol{x} - \boldsymbol{p} \right\rangle \\
&= f(\boldsymbol{v}) + \frac{1}{2\lambda}\left\|(\boldsymbol{x} - \boldsymbol{v}) + (\boldsymbol{v} - \boldsymbol{p})\right\|^2 - \frac{1}{\lambda}\left\langle \boldsymbol{v} - \boldsymbol{p}, \boldsymbol{x} - \boldsymbol{p} \right\rangle \\
&= f(\boldsymbol{v}) + \frac{1}{2\lambda}\left\|\boldsymbol{x} - \boldsymbol{v}\right\|^2 + \frac{1}{2\lambda}\left\|\boldsymbol{v} - \boldsymbol{p}\right\|^2 \\
&\quad + \frac{1}{\lambda}\left\langle \boldsymbol{v} - \boldsymbol{p}, \boldsymbol{x} - \boldsymbol{v} \right\rangle - \frac{1}{\lambda}\left\langle \boldsymbol{v} - \boldsymbol{p}, \boldsymbol{x} - \boldsymbol{p} \right\rangle \\
&= f(\boldsymbol{v}) + \frac{1}{2\lambda}\left\|\boldsymbol{x} - \boldsymbol{v}\right\|^2 + \frac{1}{2\lambda}\left\|\boldsymbol{v} - \boldsymbol{p}\right\|^2 - \frac{1}{\lambda}\left\|\boldsymbol{v} - \boldsymbol{p}\right\|^2 \\
&= f(\boldsymbol{v}) + \frac{1}{2\lambda}\left\|\boldsymbol{x} - \boldsymbol{v}\right\|^2 - \frac{1}{2\lambda}\left\|\boldsymbol{v} - \boldsymbol{p}\right\|^2 \\
&\leq f(\boldsymbol{v}) + \frac{1}{2\lambda}\left\|\boldsymbol{x} - \boldsymbol{v}\right\|^2 , \quad \forall \boldsymbol{v} \in \mathbb{R}^l.
\end{aligned}$$

The previous inequality clearly suggests that $\boldsymbol{p} = \operatorname{Prox}_{\lambda f}(\boldsymbol{x})$.

35. Show property (8.112) for the proximal operator.

Solution: For compact notations, define $\boldsymbol{p}_j := \operatorname{Prox}_{\lambda f}(\boldsymbol{x}_j)$, $j = 1, 2$. Then,

$$\begin{aligned}
\langle \boldsymbol{p}_2 - \boldsymbol{p}_1, \boldsymbol{x}_1 - \boldsymbol{p}_1 \rangle &\leq \lambda\big(f(\boldsymbol{p}_2) - f(\boldsymbol{p}_1)\big), \\
\langle \boldsymbol{p}_1 - \boldsymbol{p}_2, \boldsymbol{x}_2 - \boldsymbol{p}_2 \rangle &\leq \lambda\big(f(\boldsymbol{p}_1) - f(\boldsymbol{p}_2)\big).
\end{aligned}$$

Adding the previous inequalities results into

$$\langle \boldsymbol{p}_1 - \boldsymbol{p}_2, (\boldsymbol{p}_1 - \boldsymbol{p}_2) - (\boldsymbol{x}_1 - \boldsymbol{x}_2) \rangle \leq 0,$$

which in turn leads to the desired $\left\|\boldsymbol{p}_1 - \boldsymbol{p}_2\right\|^2 \leq \langle \boldsymbol{p}_1 - \boldsymbol{p}_2, \boldsymbol{x}_1 - \boldsymbol{x}_2 \rangle$.

36. Prove that the recursion in (8.118) converges to a minimizer of f.

Solution: Define the mapping $R := 2\operatorname{Prox}_{\lambda f} - I$. Then, (8.118) takes the following form:

$$\begin{aligned}\boldsymbol{x}_{k+1} &= \boldsymbol{x}_k + \frac{\mu_k}{2}\big(R(\boldsymbol{x}_k) - \boldsymbol{x}_k\big)\\ &= \Big(1 - \frac{\mu_k}{2}\Big)\boldsymbol{x}_k + \frac{\mu_k}{2}R(\boldsymbol{x}_k).\end{aligned}$$

Notice that R is non-expansive: $\forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^l$,

$$\begin{aligned}\left\|R(\boldsymbol{x}_1) - R(\boldsymbol{x}_2)\right\|^2 &= \left\|2\big(\operatorname{Prox}_{\lambda f}(\boldsymbol{x}_1) - \operatorname{Prox}_{\lambda f}(\boldsymbol{x}_2)\big) - (\boldsymbol{x}_1 - \boldsymbol{x}_2)\right\|^2\\ &= 4\left\|\operatorname{Prox}_{\lambda f}(\boldsymbol{x}_1) - \operatorname{Prox}_{\lambda f}(\boldsymbol{x}_2)\right\|^2 + \left\|\boldsymbol{x}_1 - \boldsymbol{x}_2\right\|^2\\ &\quad - 4\left\langle \operatorname{Prox}_{\lambda f}(\boldsymbol{x}_1) - \operatorname{Prox}_{\lambda f}(\boldsymbol{x}_2), \boldsymbol{x}_1 - \boldsymbol{x}_2\right\rangle\\ &\leq 4\left\|\operatorname{Prox}_{\lambda f}(\boldsymbol{x}_1) - \operatorname{Prox}_{\lambda f}(\boldsymbol{x}_2)\right\|^2 + \left\|\boldsymbol{x}_1 - \boldsymbol{x}_2\right\|^2\\ &\quad - 4\left\|\operatorname{Prox}_{\lambda f}(\boldsymbol{x}_1) - \operatorname{Prox}_{\lambda f}(\boldsymbol{x}_2)\right\|^2\\ &= \left\|\boldsymbol{x}_1 - \boldsymbol{x}_2\right\|^2.\end{aligned}$$

In turn, let $\boldsymbol{z}$ be a fixed point, then

$$\begin{aligned}\left\|\boldsymbol{x}_{k+1} - \boldsymbol{z}\right\|^2 &= \left\|\Big(1 - \frac{\mu_k}{2}\Big)(\boldsymbol{x}_k - \boldsymbol{z}) + \frac{\mu_k}{2}\big(R(\boldsymbol{x}_k) - \boldsymbol{z}\big)\right\|^2\\ &= \Big(1 - \frac{\mu_k}{2}\Big)\left\|\boldsymbol{x}_k - \boldsymbol{z}\right\|^2 + \frac{\mu_k}{2}\left\|R(\boldsymbol{x}_k) - \boldsymbol{z}\right\|^2\\ &\quad - \frac{\mu_k}{2}\Big(1 - \frac{\mu_k}{2}\Big)\left\|R(\boldsymbol{x}_k) - \boldsymbol{x}_k\right\|^2\\ &= \Big(1 - \frac{\mu_k}{2}\Big)\left\|\boldsymbol{x}_k - \boldsymbol{z}\right\|^2 + \frac{\mu_k}{2}\left\|2\big(\operatorname{Prox}_{\lambda f}(\boldsymbol{x}_k) - \boldsymbol{z}\big) - (\boldsymbol{x}_k - \boldsymbol{z})\right\|^2\\ &\quad - \mu_k(2 - \mu_k)\left\|\operatorname{Prox}_{\lambda f}(\boldsymbol{x}_k) - \boldsymbol{x}_k\right\|^2\\ &= \Big(1 - \frac{\mu_k}{2}\Big)\left\|\boldsymbol{x}_k - \boldsymbol{z}\right\|^2 + \frac{\mu_k}{2}\left\|\boldsymbol{x}_k - \boldsymbol{z}\right\|^2\\ &\quad + 2\mu_k\left\|\operatorname{Prox}_{\lambda f}(\boldsymbol{x}_k) - \boldsymbol{z}\right\|^2 - 2\mu_k\left\langle \operatorname{Prox}_{\lambda f}(\boldsymbol{x}_k) - \boldsymbol{z}, \boldsymbol{x}_k - \boldsymbol{z}\right\rangle\\ &\quad - \mu_k(2 - \mu_k)\left\|\operatorname{Prox}_{\lambda f}(\boldsymbol{x}_k) - \boldsymbol{x}_k\right\|^2\\ &\leq \left\|\boldsymbol{x}_k - \boldsymbol{z}\right\|^2 + 2\mu_k\left\|\operatorname{Prox}_{\lambda f}(\boldsymbol{x}_k) - \boldsymbol{z}\right\|^2 - 2\mu_k\left\|\operatorname{Prox}_{\lambda f}(\boldsymbol{x}_k) - \boldsymbol{z}\right\|^2\\ &\quad - \mu_k(2 - \mu_k)\left\|\operatorname{Prox}_{\lambda f}(\boldsymbol{x}_k) - \boldsymbol{x}_k\right\|^2\\ &= \left\|\boldsymbol{x}_k - \boldsymbol{z}\right\|^2 - \mu_k(2 - \mu_k)\left\|\operatorname{Prox}_{\lambda f}(\boldsymbol{x}_k) - \boldsymbol{x}_k\right\|^2.\end{aligned}$$

Hence, $\forall k$,

$$\mu_k(2 - \mu_k)\left\|\operatorname{Prox}_{\lambda f}(\boldsymbol{x}_k) - \boldsymbol{x}_k\right\|^2 \leq \left\|\boldsymbol{x}_k - \boldsymbol{z}\right\|^2 - \left\|\boldsymbol{x}_{k+1} - \boldsymbol{z}\right\|^2.$$

Given any non-negative integer k_0, the previous telescoping inequality is utilized for all $k \in \{0, \dots, k_0\}$ to produce

$$\begin{aligned}\sum_{k=0}^{k_0} \mu_k(2-\mu_k) \left\|\operatorname{Prox}_{\lambda f}(\boldsymbol{x}_k) - \boldsymbol{x}_k\right\|^2 &\leq \left\|\boldsymbol{x}_0 - \boldsymbol{z}\right\|^2 - \left\|\boldsymbol{x}_{k_0+1} - \boldsymbol{z}\right\|^2 \\ &\leq \left\|\boldsymbol{x}_0 - \boldsymbol{z}\right\|^2 .\end{aligned}$$

Since the previous relation holds for any k_0, applying $\lim_{k_0 \to \infty}$ on both sides of the inequality results into

$$\sum_{k=0}^{+\infty} \mu_k(2-\mu_k) \left\|\operatorname{Prox}_{\lambda f}(\boldsymbol{x}_k) - \boldsymbol{x}_k\right\|^2 < +\infty. \tag{39}$$

Moreover, notice that

$$\begin{aligned}&\left\|\operatorname{Prox}_{\lambda f}(\boldsymbol{x}_{k+1}) - \boldsymbol{x}_{k+1}\right\| = \frac{1}{2}\left\|R(\boldsymbol{x}_{k+1}) - \boldsymbol{x}_{k+1}\right\| \\ &\quad = \frac{1}{2}\left\|R(\boldsymbol{x}_{k+1}) - R(\boldsymbol{x}_k) + \Big(1 - \frac{\mu_k}{2}\Big)\big(R(\boldsymbol{x}_k) - \boldsymbol{x}_k\big)\right\| \\ &\quad \leq \frac{1}{2}\left\|R(\boldsymbol{x}_{k+1}) - R(\boldsymbol{x}_k)\right\| + \frac{1}{2}\Big(1 - \frac{\mu_k}{2}\Big)\left\|R(\boldsymbol{x}_k) - \boldsymbol{x}_k\right\| \\ &\quad \leq \frac{1}{2}\left\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\right\| + \Big(1 - \frac{\mu_k}{2}\Big)\left\|\operatorname{Prox}_{\lambda f}(\boldsymbol{x}_k) - \boldsymbol{x}_k\right\| \\ &\quad = \frac{\mu_k}{2}\left\|\operatorname{Prox}_{\lambda f}(\boldsymbol{x}_k) - \boldsymbol{x}_k\right\| + \Big(1 - \frac{\mu_k}{2}\Big)\left\|\operatorname{Prox}_{\lambda f}(\boldsymbol{x}_k) - \boldsymbol{x}_k\right\| \\ &\quad = \left\|\operatorname{Prox}_{\lambda f}(\boldsymbol{x}_k) - \boldsymbol{x}_k\right\| .\end{aligned}$$

Since $(\left\|\operatorname{Prox}_{\lambda f}(\boldsymbol{x}_k) - \boldsymbol{x}_k\right\|)_{k\in\mathbb{N}}$ is monotonically non-increasing, and bounded from below, it converges. Necessarily, $\lim_{k\to\infty} \left\|\operatorname{Prox}_{\lambda f}(\boldsymbol{x}_k) - \boldsymbol{x}_k\right\|^2 = 0$. Otherwise, there exists an $\epsilon > 0$ and a subsequence $(k_m)_{m\in\mathbb{N}}$ such that

$$\left\|\operatorname{Prox}_{\lambda f}(\boldsymbol{x}_{k_m}) - \boldsymbol{x}_{k_m}\right\|^2 \geq \epsilon, \quad \forall m \in \mathbb{N}.$$

This together with the fact that $\lim_{m\to\infty} \sum_{i=0}^{k_m} \mu_i(2-\mu_i) = +\infty$, and (39) imply that

$$\begin{aligned}+\infty &> \sum_{k=0}^{+\infty} \mu_k(2-\mu_k) \left\|\operatorname{Prox}_{\lambda f}(\boldsymbol{x}_k) - \boldsymbol{x}_k\right\|^2 \\ &\geq \sum_{m=0}^{+\infty} \mu_{k_m}(2-\mu_{k_m}) \left\|\operatorname{Prox}_{\lambda f}(\boldsymbol{x}_{k_m}) - \boldsymbol{x}_{k_m}\right\|^2 \geq \epsilon \sum_{m=0}^{+\infty} \mu_{k_m}(2-\mu_{k_m}) = +\infty,\end{aligned}$$

which is clearly absurd.

Let $\boldsymbol{x}_*$ be an arbitrary cluster point. Notice that

$$\begin{aligned}
&\|\boldsymbol{x}_* - \mathrm{Prox}_{\lambda f}(\boldsymbol{x}_*)\|^2 \\
&\quad = \|\boldsymbol{x}_* - \boldsymbol{x}_{k_m} + \boldsymbol{x}_{k_m} - \mathrm{Prox}_{\lambda f}(\boldsymbol{x}_*)\|^2 \\
&\quad = \|\boldsymbol{x}_* - \boldsymbol{x}_{k_m}\|^2 + \|\boldsymbol{x}_{k_m} - \mathrm{Prox}_{\lambda f}(\boldsymbol{x}_{k_m}) + \mathrm{Prox}_{\lambda f}(\boldsymbol{x}_{k_m}) - \mathrm{Prox}_{\lambda f}(\boldsymbol{x}_*)\|^2 \\
&\qquad + 2\left\langle \boldsymbol{x}_* - \boldsymbol{x}_{k_m}, \boldsymbol{x}_{k_m} - \mathrm{Prox}_{\lambda f}(\boldsymbol{x}_*)\right\rangle \\
&\quad = \|\boldsymbol{x}_* - \boldsymbol{x}_{k_m}\|^2 + \|\boldsymbol{x}_{k_m} - \mathrm{Prox}_{\lambda f}(\boldsymbol{x}_{k_m})\|^2 + \|\mathrm{Prox}_{\lambda f}(\boldsymbol{x}_{k_m}) - \mathrm{Prox}_{\lambda f}(\boldsymbol{x}_*)\|^2 \\
&\qquad + 2\left\langle \boldsymbol{x}_{k_m} - \mathrm{Prox}_{\lambda f}(\boldsymbol{x}_{k_m}), \mathrm{Prox}_{\lambda f}(\boldsymbol{x}_{k_m}) - \mathrm{Prox}_{\lambda f}(\boldsymbol{x}_*)\right\rangle \\
&\qquad + 2\left\langle \boldsymbol{x}_* - \boldsymbol{x}_{k_m}, \boldsymbol{x}_{k_m} - \boldsymbol{x}_* + \boldsymbol{x}_* - \mathrm{Prox}_{\lambda f}(\boldsymbol{x}_*)\right\rangle \\
&\quad \leq \|\boldsymbol{x}_* - \boldsymbol{x}_{k_m}\|^2 + \|\boldsymbol{x}_{k_m} - \mathrm{Prox}_{\lambda f}(\boldsymbol{x}_{k_m})\|^2 + \|\boldsymbol{x}_{k_m} - \boldsymbol{x}_*\|^2 \\
&\qquad + 2\left\langle \boldsymbol{x}_{k_m} - \mathrm{Prox}_{\lambda f}(\boldsymbol{x}_{k_m}), \mathrm{Prox}_{\lambda f}(\boldsymbol{x}_{k_m}) - \mathrm{Prox}_{\lambda f}(\boldsymbol{x}_*)\right\rangle \\
&\qquad - 2\|\boldsymbol{x}_* - \boldsymbol{x}_{k_m}\|^2 + 2\left\langle \boldsymbol{x}_* - \boldsymbol{x}_{k_m}, \boldsymbol{x}_* - \mathrm{Prox}_{\lambda f}(\boldsymbol{x}_*)\right\rangle \\
&\quad \leq \|\boldsymbol{x}_{k_m} - \mathrm{Prox}_{\lambda f}(\boldsymbol{x}_{k_m})\|^2 \\
&\qquad + 2\|\boldsymbol{x}_{k_m} - \mathrm{Prox}_{\lambda f}(\boldsymbol{x}_{k_m})\|\,\|\mathrm{Prox}_{\lambda f}(\boldsymbol{x}_{k_m}) - \mathrm{Prox}_{\lambda f}(\boldsymbol{x}_*)\| \\
&\qquad + 2\|\boldsymbol{x}_* - \boldsymbol{x}_{k_m}\|\,\|\boldsymbol{x}_* - \mathrm{Prox}_{\lambda f}(\boldsymbol{x}_*)\| \\
&\quad \leq \|\boldsymbol{x}_{k_m} - \mathrm{Prox}_{\lambda f}(\boldsymbol{x}_{k_m})\|^2 + 2\|\boldsymbol{x}_{k_m} - \mathrm{Prox}_{\lambda f}(\boldsymbol{x}_{k_m})\|\,\|\boldsymbol{x}_{k_m} - \boldsymbol{x}_*\| \\
&\qquad + 2\|\boldsymbol{x}_* - \boldsymbol{x}_{k_m}\|\,\|\boldsymbol{x}_* - \mathrm{Prox}_{\lambda f}(\boldsymbol{x}_*)\|\,.
\end{aligned}$$

Applying $\lim_{m\to\infty}$ on both sides of the previous inequality results into $\boldsymbol{x}_* = \mathrm{Prox}_{\lambda f}(\boldsymbol{x}_*) \Leftrightarrow \boldsymbol{x}_* \in \mathrm{Fix}(\mathrm{Prox}_{\lambda f})$. Since $\boldsymbol{x}_*$ was chosen arbitrarily within the set of all cluster points of $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$, then it can be readily seen that all cluster points belong to $\mathrm{Fix}(\mathrm{Prox}_{\lambda f})$.

We have already seen that the sequence $(\|\boldsymbol{x}_n - \boldsymbol{x}\|^2)_{n\in\mathbb{N}}$ converges for any $\boldsymbol{x} \in \mathrm{Fix}(\mathrm{Prox}_{\lambda f})$. Moreover, any cluster point of $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ belongs to $\mathrm{Fix}(\mathrm{Prox}_{\lambda f})$. Let us show now that $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ possesses only one cluster point. To this end, assume two cluster points $\boldsymbol{x}, \boldsymbol{y}$ of $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$. This means that there exist subsequences $(\boldsymbol{x}_{k_m})_{m\in\mathbb{N}}$ and $(\boldsymbol{x}_{l_m})_{m\in\mathbb{N}}$ which converge to $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. Moreover, notice that

$$\langle \boldsymbol{x}_k, \boldsymbol{x} - \boldsymbol{y}\rangle = \frac{1}{2}\Big(\|\boldsymbol{x}_k - \boldsymbol{y}\|^2 - \|\boldsymbol{x}_k - \boldsymbol{x}\|^2 + \|\boldsymbol{x}\|^2 - \|\boldsymbol{y}\|^2\Big).$$

Since both $(\|\boldsymbol{x}_k - \boldsymbol{x}\|^2)_{k\in\mathbb{N}}$ and $(\|\boldsymbol{x}_k - \boldsymbol{y}\|^2)_{k\in\mathbb{N}}$ converge, so does also the sequence $(\langle \boldsymbol{x}_k, \boldsymbol{x} - \boldsymbol{y}\rangle)_{k\in\mathbb{N}}$. Hence,

$$\begin{aligned}
\langle \boldsymbol{x}, \boldsymbol{x} - \boldsymbol{y}\rangle &= \Big\langle \lim_{m\to\infty} \boldsymbol{x}_{k_m}, \boldsymbol{x} - \boldsymbol{y}\Big\rangle = \lim_{m\to\infty} \langle \boldsymbol{x}_{k_m}, \boldsymbol{x} - \boldsymbol{y}\rangle \\
&= \lim_{k\to\infty} \langle \boldsymbol{x}_k, \boldsymbol{x} - \boldsymbol{y}\rangle = \lim_{m\to\infty} \langle \boldsymbol{x}_{l_m}, \boldsymbol{x} - \boldsymbol{y}\rangle \\
&= \Big\langle \lim_{m\to\infty} \boldsymbol{x}_{l_m}, \boldsymbol{x} - \boldsymbol{y}\Big\rangle = \langle \boldsymbol{y}, \boldsymbol{x} - \boldsymbol{y}\rangle\,,
\end{aligned}$$

and in turn, $\|\boldsymbol{x} - \boldsymbol{y}\|^2 = 0 \Rightarrow \boldsymbol{x} = \boldsymbol{y}$. To conclude, $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ converges to a point in $\mathrm{Fix}(\mathrm{Prox}_{\lambda f}) = \arg\min_{\boldsymbol{v}\in\mathbb{R}^l} f(\boldsymbol{v})$.

37. Derive (8.122) from (8.121).

Solution: Use the matrix inversion lemma

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

with $B = C = I$, $D = A^{-1}$ and $A = \epsilon I$, which gives

$$\begin{aligned}(A + \epsilon I)^{-1} &= \frac{1}{\epsilon}I - \frac{1}{\epsilon^2}\big(\frac{1}{\epsilon}I + A^{-1}\big)^{-1} \\ &= \frac{1}{\epsilon}I - \frac{1}{\epsilon^2}\big(\frac{1}{\epsilon}A + I\big)^{-1}A \\ &= \frac{1}{\epsilon}I - \frac{1}{\epsilon}\big(A + \epsilon I\big)^{-1}A,\end{aligned}$$

which finally leads to the result.

Solutions To Problems of Chapter 9

9.1. Show that if $x_i,\ y_i,\ i = 1, 2, \ldots, l$, are real numbers, then prove the Cauchy-Schwarz inequality:

$$\left(\sum_{i=1}^{l} x_i y_i\right)^2 \leq \left(\sum_{i=1}^{l} x_i^2\right)\left(\sum_{i=1}^{l} y_i^2\right).$$

Solution: Consider the identity

$$\sum_{i=1}^{l} (z x_i + y_i)^2 \geq 0,\ \forall z \in \mathbb{R}.$$

Expanding the previous we get

$$z^2 \sum_{i=1}^{l} x_i^2 + 2z \sum_{i=1}^{l} x_i y_i + \sum_{i=1}^{l} y_i^2 \geq 0,\ \forall z.$$

Since the previous quadratic expansion is nonnegative for every z, we know from our early college mathematics that its discriminant must be non-positive, i.e.,

$$\left(\sum_{i=1}^{l} x_i y_i\right)^2 - \left(\sum_{i=1}^{l} x_i^2\right)\left(\sum_{i=1}^{l} y_i^2\right) \leq 0,$$

which proves the required inequality.

9.2. Prove that the l_2 (Euclidean) norm is a true norm, i.e., it satisfies the four conditions that define a norm.
Hint: To prove the triangle inequality, use the Cauchy-Schwarz inequality.

Solution: To prove the three first conditions, i.e.,

- $||\boldsymbol{x}||_2 \geq 0$
- $||\boldsymbol{x}||_2 = 0 \Leftrightarrow \boldsymbol{x} = \boldsymbol{0}$
- $||\alpha \boldsymbol{x}||_2 = |\alpha| ||\boldsymbol{x}||_2$

is straightforward. For the triangle inequality, recall the Cauchy-Schwartz inequality

$$\left(\sum_{i=1}^{l} x_i y_i\right)^2 \leq \left(\sum_{i=1}^{l} x_i^2\right)\left(\sum_{i=1}^{l} y_i^2\right),$$

or

$$\sum_{i=1}^{l} x_i y_i \leq \sqrt{\sum_{i=1}^{l} x_i^2}\sqrt{\sum_{i=1}^{l} y_i^2}.$$

Multiplying both terms by two and adding the same nonnegative factors in both sides we get

$$\sum_{i=1}^{l} x_i^2 + \sum_{i=1}^{l} y_i^2 + 2\sum_{i=1}^{l} x_i y_i \leq \sum_{i=1}^{l} x_i^2 + \sum_{i=1}^{l} y_i^2 + 2\sqrt{\sum_{i=1}^{l} x_i^2}\sqrt{\sum_{i-1}^{l} y_i^2},$$

or

$$\sum_{i=1}^{l} (x_i + y_i)^2 \leq \left(\sqrt{\sum_{i=1}^{l} x_i^2} + \sqrt{\sum_{i-1}^{l} y_i^2}\right)^2,$$

and taking the square root of both sides we prove the inequality.

9.3. Prove that any function that is a norm is also a convex function.

Solution: Let a function $f : \mathbb{R}^l \longmapsto [0, \infty)$, which is a norm. Then by the definition of a norm we have

$$f(a\boldsymbol{x} + b\boldsymbol{y}) \leq |a|f(\boldsymbol{x}) + |b|f(\boldsymbol{y}),$$

and using λ and $1 - \lambda$ in place of a and b, we prove the claim.

9.4. Show Young's inequality for nonnegative real numbers a and b,

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q},$$

for $\infty > p > 1$ and $\infty > q > 1$ such that

$$\frac{1}{p} + \frac{1}{q} = 1.$$

Solution: First, it is obvious that the inequality holds if a and/or b are zero. Then, we turn our attention to the case that both of them are positive. We know that the $\ln(\cdot)$ function is a concave one. Then for $t = \frac{1}{p}$ and $1 - t = \frac{1}{q}$ we have that,

$$\ln\left(ta^p + (1-t)b^q\right) \geq t\ln(a^p) + (1-t)\ln(b^q) = \ln a + \ln b = \ln(ab).$$

where the claim is proved by removing the ln and keeping in mind that ln is a monotonically increasing function.

9.5. Prove Holder's inequality for l_p norms,

$$||\boldsymbol{x}^T\boldsymbol{y}||_1 = \sum_{i=1}^{l} |x_i y_i| \leq ||\boldsymbol{x}||_p ||\boldsymbol{y}||_q = \left(\sum_{i=1}^{l} |x_i|^p\right)^{\frac{1}{p}} \left(\sum_{i=1}^{q} |y_i|^q\right)^{\frac{1}{q}},$$

for $p \geq 1$ and $q \geq 1$ such that

$$\frac{1}{p} + \frac{1}{q} = 1.$$

Hint: Use Young's inequality.

Solution: First note that if $p = 1$ and $q = \infty$ then the inequality is obvious, since in this case $||\boldsymbol{y}||_\infty = \max\{|y_1|, \ldots, |y_l|\}$. Thus we concentrate in the case where are $\infty > p > 1$ and $\infty > q > 1$. We use Young's inequality with $\frac{|x_i|}{||\boldsymbol{x}||_p}$, and $\frac{|y_i|}{||\boldsymbol{y}||_q}$ in place of a and b and take the sum. This results in

$$\frac{1}{||\boldsymbol{x}||_p ||\boldsymbol{y}||_q} \sum_{i=1}^{l} |x_i y_i| \leq \frac{1}{p} \sum_{i=1}^{l} \frac{|x_i|^p}{||\boldsymbol{x}||_p^p} + \frac{1}{q} \sum_{i=1}^{l} \frac{|y_i|^q}{||\boldsymbol{y}||_q^q} = \frac{1}{p} + \frac{1}{q} = 1,$$

which proves the claim.

9.6. Prove Minkowski's inequality,

$$\left(\sum_{i=1}^{l} (|x_i| + |y_i|)^p\right)^{\frac{1}{p}} \leq \left(\sum_{i=1}^{l} |x_i|^p\right)^{\frac{1}{p}} + \left(\sum_{i=1}^{l} |y_i|^p\right)^{\frac{1}{p}},$$

for $p \geq 1$.
Hint: Use Holder's inequality together with the identity:

$$(|a| + |b|)^p = (|a| + |b|)^{p-1}|a| + (|a| + |b|)^{p-1}|b|.$$

Solution: The result is obvious for $p = 1$.We will prove it for $p > 1$. In the given identity, put in place of a x_i and in the place of b y_i and add to get

$$\sum_{i=1}^{l} (|x_i| + y_i|)^p = \sum_{i=1}^{l} (|x_i| + |y_i|)^{p-1} |x_i| + \sum_{i=1}^{l} (|x_i| + |y_i|)^{p-1} |y_i|.$$

Using Holder's inequality we have

$$\sum_{i=1}^{l} (|x_i| + |y_i|)^{p-1} |x_i| \leq \left(\sum_{i=1}^{l} (|x_i| + |y_i|)^{q(p-1)}\right)^{\frac{1}{q}} \left(\sum_{i=1}^{l} |x_i|^p\right)^{\frac{1}{p}},$$

and similarly

$$\sum_{i=1}^{l} (|x_i| + |y_i|)^{p-1} |y_i| \leq \left(\sum_{i=1}^{l} (|x_i| + |y_i|)^{q(p-1)}\right)^{\frac{1}{q}} \left(\sum_{i=1}^{l} |y_i|^p\right)^{\frac{1}{p}}.$$

If we add both inequalities together and take into account that $q(p-1) = p$, we get

$$\sum_{i=1}^{l}(|x_i|+|y_i|)^p \leq \left(\sum_{i=1}^{l}(|x_i| + |y_i|)^p\right)^{\frac{1}{q}} \left(\left(\sum_{i=1}^{l} |x_i|^p\right)^{\frac{1}{p}} + \left(\sum_{i=1}^{l} |y_i|^p\right)^{\frac{1}{p}}\right).$$

Dividing by $\left(\sum_{i=1}^{l}(|x_i| + |y_i|)^p\right)^{\frac{1}{q}}$, Minkowski's inequality results.

9.7. Prove that for $p \geq 1$ the l_p norm is a true norm.

Solution: The first three conditions for a function to be a norm are obvious. The triangle inequality is a direct consequence of Minkowski's inequality, if we recall that

$$|a + b| \leq |a| + |b|.$$

9.8. Use a counterexample to show that the l_p norm for $0 < p < 1$ is not a true norm and it violates the triangle condition.

Solution: Consider the two vectors in the l-dimensional space

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

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

$$||\boldsymbol{x} + \boldsymbol{y}||_p = 2^{\frac{1}{p}} \leq ||\boldsymbol{x}||_p + ||\boldsymbol{y}||_p = 1 + 1 = 2,$$

which is not possible for $p < 1$.

9.9. Show that the null space of a full rank $N \times l$ matrix X is a subspace of dimensionality $l - N$, for $N < l$.

Solution: By the definition of a null space, i.e.,

$$\mathcal{N} = \{\boldsymbol{z} : X\boldsymbol{z} = \boldsymbol{0}\},$$

it is straightforward to see that the three conditions that define a linear space are met, i.e.,

- $\boldsymbol{0} \in \mathcal{N}$
- If $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ belong to the null subspace so does their sum

- If $\boldsymbol{x}$ belongs to the space so does $c\boldsymbol{x}$, $c \in \mathbb{R}$

To show that its dimensionality is $l - N$, recall from linear algebra (and it is readily established by the respective definitions) that the dimensionality of the range space is N, which is equal to the rank of the matrix, since it has been assumed to be full rank. The null space, being orthogonal to the range space, will necessarily have dimension equal to $l - N$

9.10. Show, using Lagrange multipliers, that the l_2 minimizer in (9.18) accepts the closed form solution

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

Solution: The ℓ_2 minimizer task is given by

$$\begin{aligned} \text{minimize} \quad & ||\boldsymbol{\theta}||_2^2 \\ \text{s.t.} \quad & \boldsymbol{x}_n^T \boldsymbol{\theta} = y_n, \;\; n = 1, 2, \ldots, N, \end{aligned} \tag{1}$$

where by assumption $N < l$. The equivalent Lagrangian is given by

$$L(\boldsymbol{\theta}) = \boldsymbol{\theta}^T \boldsymbol{\theta} + \sum_{n=1}^{N} \lambda_n (y_n - \boldsymbol{x}_n^T \boldsymbol{\theta}).$$

Taking the gradient w.r. to $\boldsymbol{\theta}$ and equating to zero we obtain

$$\boldsymbol{\theta} = X^T \boldsymbol{\lambda},$$

where $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \ldots, \lambda_N]^T$, and X being the input matrix

$$X = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_N^T \end{bmatrix}.$$

Plugging the solution into the set of constraints, which can be written as

$$\boldsymbol{y} = X\boldsymbol{\theta},$$

we obtain the equation

$$X\boldsymbol{\theta} = \boldsymbol{y} = XX^T \boldsymbol{\lambda} \Rightarrow \boldsymbol{\lambda} = \left(XX^T\right)^{-1} \boldsymbol{y},$$

which finally results in the solution

$$\boldsymbol{\theta} = X^T \left(XX^T\right)^{-1} \boldsymbol{y}.$$

9.11. Show that the necessary and sufficient condition for a $\boldsymbol{\theta}$ to be a minimizer of

$$\begin{aligned} \text{minimize :} \quad & ||\boldsymbol{\theta}||_1 \\ \text{s.t.} \quad & X\boldsymbol{\theta} = \boldsymbol{y}, \end{aligned}$$

is the following

$$\left| \sum_{i:\theta_i \neq 0} \text{sign}(\theta_i) z_i \right| \leq \sum_{i:\theta_i = 0} |z_i|, \ \forall \boldsymbol{z} \in \mathcal{N},$$

where $\mathcal{N}$ is the null space of X. Moreover, if the minimizer is unique the previous inequality becomes a strict one.

Solution: (a) Sufficiency

Let $\boldsymbol{\theta}$ be a minimizer, then for any $\boldsymbol{z} \in \mathcal{N}$ and $\forall t \in \mathbb{R}$ we have

$$\sum_{i=1}^{l} |\theta_i + t z_i| \geq \sum_{i}^{l} |\theta_i|,$$

by the definition of the ℓ_1 norm. Fix $\boldsymbol{z}$. Then, for sufficiently small value of t, θ_i and $\theta_i + t z_i$ will have the same sign, whenever $\theta_i \neq 0$. Hence, the previous inequality becomes

$$\begin{aligned} \sum_{i:\theta_i = 0} |t z_i| + \sum_{i:\theta_i \neq 0} \text{sign}(\theta_i)(\theta_i + t z_i) \quad & \geq \quad \sum_{i:\theta_i \neq 0} \text{sign}(\theta_i)\theta_i, \quad \text{or} \\ t \sum_{i:\theta_i \neq 0} \text{sign}(\theta_i) z_i + \sum_{i:\theta_i = 0} |t z_i| \quad & \geq \quad 0, \text{or} \\ -t \sum_{i:\theta_i \neq 0} \text{sign}(\theta_i) z_i \quad & \leq \quad |t| \sum_{i:\theta_i = 0} |z_i|. \end{aligned}$$

Since the previous is valid $\forall t$, sufficiently small, we can choose the sign of t so that to obtain

$$\left| \sum_{i:\theta_i \neq 0} \text{sign}(\theta_i) z_i \right| \leq \sum_{i:\theta_i = 0} |z_i|.$$

Obviously, if $\boldsymbol{\theta}$ is a strict minimizer the previous is true with the strict inequality.

(b) Sufficiency
Let

$$\left| \sum_{i:\theta_i \neq 0} \text{sign}(\theta_i) z_i \right| \leq \sum_{i:\theta_i = 0} |z_i| \qquad (2)$$

be true. Then the following are valid,

$$\begin{aligned}\sum_{i=1}^{l}|\theta_i| &= \sum_{i:\theta_i\neq 0}\operatorname{sign}(\theta_i)\theta_i \\ &= \sum_{i:\theta_i\neq 0}\operatorname{sign}(\theta_i)(\theta_i+z_i) - \sum_{i:\theta_i\neq 0}\operatorname{sign}(\theta_i)z_i. \qquad (3)\end{aligned}$$

However, (2) implies that

$$-\sum_{i:\theta_i=0}|z_i| \leq \sum_{i:\theta_i\neq 0}\operatorname{sign}(\theta_i)z_i \leq \sum_{i:\theta_i=0}|z_i|. \qquad (4)$$

Combining (4), and (3) we get

$$\begin{aligned}\sum_{i=1}^{l}|\theta_i| &\leq \sum_{i:\theta_i\neq 0}\operatorname{sign}(\theta_i)(\theta_i+z_i) + \sum_{i:\theta_i=0}|z_i| \\ &\leq \sum_{i:\theta_i\neq 0}|\theta_i+z_i| + \sum_{i:\theta_i=0}|z_i| = \sum_{i}^{l}|\theta_i+z_i|.\end{aligned}$$

Moreover, if the strict inequality holds in (2), then the minimizer is unique.

9.12. Prove that if the ℓ_1 norm minimizer is unique, then the number of its components, which are identically zero, must be at least as large as the dimensionality of the null space of the corresponding input matrix.

Solution: Assume that

$$\xi := \operatorname{card}\{i : \hat{\theta}_i = 0\} < \dim(\operatorname{null}(X)) := m.$$

There exists a set of vectors $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m\}$ which constitutes a basis for null(X). Hence, for any vector in the solutions set, $\boldsymbol{\theta} \in \Theta$, there exist nonzero vectors $\boldsymbol{z} \in \operatorname{null}(X)$ and $\boldsymbol{a} \in \mathbb{R}^m$, such that

$$\boldsymbol{\theta} = \boldsymbol{z} + \hat{\boldsymbol{\theta}} = \sum_{j=1}^{m} a_j\boldsymbol{u}_j + \hat{\boldsymbol{\theta}}.$$

As a result, $\forall i$ such that $\hat{\theta}_i = 0$, we have that $\theta_i = z_i = \sum_{j=1}^{m} a_j u_{ji}$, which generates the following system of equations:

$$\begin{bmatrix}\theta_{i_1}\\ \vdots \\ \theta_{i_\xi}\end{bmatrix} = \begin{bmatrix}u_{1i_1} & \cdots & u_{mi_1}\\ \vdots & \ddots & \vdots \\ u_{1i_\xi} & \cdots & u_{mi_\xi}\end{bmatrix}\begin{bmatrix}a_1\\ \vdots \\ a_m\end{bmatrix}.$$

Since $\xi < m$, there exists a nonzero $\boldsymbol{a}$ such that $\theta_{i_1} = \cdots = \theta_{i_\xi} = 0$. This implies also that the nonzero $\boldsymbol{z}$, defined as $\boldsymbol{z} = \sum_{j=1}^{m} a_j \boldsymbol{u}_j$, satisfies $z_{i_1} = \cdots = z_{i_\xi} = 0$. If we plug this last result into Lemma **??** then we end up with the following absurd relation:

$$0 \leq \left| \sum_{i:\ \hat{\theta}_i \neq 0} \operatorname{sgn}(\hat{\theta}_i) z_i \right| < \sum_{i:\ \hat{\theta}_i = 0} |z_i| = 0.$$

9.13. Show that the l_1 norm is a convex function (as all norms), yet it is not strictly convex. In contrast, the squared Euclidean norm is a strictly convex function.

Solution: From the definition of convexity we have that

$$f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \leq \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y}),$$

and for strict convexity the strict inequality is valid. For the l_1 norm we have

$$\sum_i (|\lambda x_i + (1-\lambda) y_i|) \leq \lambda \sum_i |x_i| + (1-\lambda) \sum_i |y_i|,$$

but there is no reason for strict inequality. Take for example the case where all x_i and y_i to be nonnegative.

However, things are different with the square Euclidean norm. Indeed, for this case we have.

$$\sum_i (\lambda x_i + (1-\lambda) y_i)^2 < \lambda \sum_i x_i^2 + (1-\lambda) \sum_i y_i^2.$$

By expanding the term on the left hand side we obtain that

$$\lambda^2 \sum_i x_i^2 + (1-\lambda)^2 \sum_i y_i^2 + 2\lambda(1-\lambda) x_i y_i < \lambda \sum_i x_i^2 + (1-\lambda) \sum_i y_i^2,$$

or

$$2\lambda(1-\lambda) x_i y_i < \lambda(1-\lambda) \left(\sum_i x_i^2 + \sum_i y_i^2 \right),$$

and finally

$$\sum_i (x_i - y_i)^2 > 0, \text{for } \boldsymbol{x} \neq \boldsymbol{y}.$$

Following a similar path, and exploiting the Cauchy-Schwartz inequality (which becomes equality for co-linear vectors) show that the ℓ_2 norm function is not strictly convex.

9.14. Construct in the 5-dimensional space a matrix that has a) rank equal to five and spark equal to four, b) rank equal to five and spark equal to three and c) rank and spark equal to four.

Solution: What we do here can obviously be applied in any $\mathbb{R}^l$ space. Consider a basis in the space, e.g.,

$$\boldsymbol{e}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \boldsymbol{e}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \boldsymbol{e}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \boldsymbol{e}_4 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \boldsymbol{e}_5 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}.$$

a) Combine the previous by taking three of them at a time. Obviously we can form $\begin{pmatrix} 5 \\ 3 \end{pmatrix}$ more vectors. Take, for example, the following combinations

$$\boldsymbol{e}_{123} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \boldsymbol{e}_{234} = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \\ 0 \end{bmatrix}, \boldsymbol{e}_{345} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \boldsymbol{e}_{134} = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \\ 0 \end{bmatrix}, \boldsymbol{e}_{135} = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix}.$$

The array that has as columns ALL the previous vectors, from $\boldsymbol{e}_1$ to $\boldsymbol{e}_{135}$, has rank equal to five and spark equal to four. Every combination of three of them are linearly independent. For the case b), combine the vectors of the basis in pairs, then follow the same procedure and it is straightforward to see that the resulting matrix has rank five and spark three. For the case c), take four of the five basis vector together with the ones that result from the combinations formed in a) and the corresponding matrix will be or rank 4 and spark 4.

9.15. Let X be a full row rank $N \times l$ matrix, with $l > N$. Derive the Welch bound for the mutual coherence $\mu(X)$,

$$\mu(X) \geq \sqrt{\frac{l-N}{N(l-1)}}. \tag{5}$$

The bound for $N = l$ is obviously zero, since the matrix can be orthogonal.

Solution: Without harming generality, let us assume that the columns

$\boldsymbol{x}_i^c, i = 1, 2, \ldots, l$, are normalized to unit norm. Form the Gram matrix

$$G = X^T X = \begin{bmatrix} \boldsymbol{x}_1^{cT}\boldsymbol{x}_1^c & \ldots & \boldsymbol{x}_1^{cT}\boldsymbol{x}_l^c \\ \boldsymbol{x}_2^{cT}\boldsymbol{x}_1^c & \ldots & \boldsymbol{x}_2^{cT}\boldsymbol{x}_l^c \\ \vdots & \ddots & \vdots \\ \boldsymbol{x}_l^{cT}\boldsymbol{x}_1 & \ldots & \boldsymbol{x}_l^{cT}\boldsymbol{x}_l^c \end{bmatrix}. \tag{6}$$

We know that the trace of G is equal to the sum of its eigenvalues. Moreover, since G is of rank N and semidefinite, only N of its eigenvalues are positive and the rest are zero. Let the nonzero eigenvalues be $\lambda_1, \ldots, \lambda_N$. Form the two vectors $\mathbf{1} = [1, 1, \ldots, 1]^T$ and $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_N]^T$. Apply the Cauchy-Schwartz inequality on these two vectors to obtain

$$(\text{trace}(G))^2 = \left(\sum_{i=1}^{N} \lambda_i\right)^2 \leq N \sum_{i=1}^{N} \lambda_i^2, \tag{7}$$

and taking into account that the trace of G is equal to l (due to the normalization of the columns of X) we obtain

$$\sum_{i=1}^{N} \lambda_i^2 \geq \frac{l^2}{N}. \tag{8}$$

However from linear algebra, we know that the Frobenius norm satisfies the following: ([Golu 83]. This is true for any matrix and it is easily shown by the definition of the Frobenius norm and the fact that the sum of the square singular values is equal to the trace of the respective Gram)

$$||G||_F^2 \equiv \sum_{i=1}^{l}\sum_{j=1}^{l} |\boldsymbol{x}_i^{cT}\boldsymbol{x}_j^c|^2 = \sum_{i=1}^{N} \sigma_i^2, \tag{9}$$

where $\sigma_i, i = 1, 2, \ldots, N$, are the singular values of G. In addition, we know that the square singular values of a matrix are the eigenvalues of its Gram, [Golu 83]. The Gram of G is equal to $G^T G = GG$, since G is symmetric. Hence the eigenvalues of $G^T G$ are equal to the square of the eigenvalues of G. Thus $\sigma_i^2 = \lambda_i^2, i = 1, 2, \ldots, l$. Therefore, we get that

$$\sum_{i=1}^{l}\sum_{j=1}^{l} |\boldsymbol{x}_i^{cT}\boldsymbol{x}_j^c|^2 = \sum_{i=1}^{N} \lambda_i^2. \tag{10}$$

Combining (8) and (10) results to

$$\sum_{i=1}^{l}\sum_{j=1}^{l} |\boldsymbol{x}_i^{cT}\boldsymbol{x}_j^c|^2 \geq \frac{l^2}{N}. \tag{11}$$

Taking into account that all the elements in the diagonal of G are equal to one, we get

$$l+\sum_{i=1}^{l}\sum_{j=1,j\neq i}^{l}|\boldsymbol{x}_i^{cT}\boldsymbol{x}_j^c|^2\geq\frac{l^2}{N},\tag{12}$$

or

$$\sum_{i=1}^{l}\sum_{j=1,j\neq i}^{l}|\boldsymbol{x}_i^{cT}\boldsymbol{x}_j^c|^2\geq\frac{l(l-N)}{N}.\tag{13}$$

Also, by the definition of the mutual coherence, it becomes obvious that (the maximum is greater than the average value)

$$(\mu(X))^2:=(\max_{i\neq j}(|\boldsymbol{x}_i^{cT}\boldsymbol{x}_j^c|^2)\geq\frac{1}{l^2-l}\sum_{j=1,j\neq i}^{l}|\boldsymbol{x}_i^{cT}\boldsymbol{x}_j^c|^2,\tag{14}$$

and combining it with (13), we finally obtain that

$$\mu(X)\geq\sqrt{\frac{l-N}{N(l-1)}}.\tag{15}$$

9.16. Let X be a $N\times l$ matrix. Then prove that its spark is bounded as

$$\text{spark}(X)\geq 1+\frac{1}{\mu(X)},$$

where $\mu(X)$ is the mutual coherence of the matrix.
Hint: Consider the Gram matrix X^TX and the following theorem, concerning positive definite matrices: An $m\times m$ matrix A is positive definite if

$$|A(i,i)|>\sum_{j=1,j\neq i}^{m}|A(i,j)|,\ \forall i=1,2,\dots,m,$$

see, e.g., [Horn 85].

Solution: Let the matrix X be

$$X=[\boldsymbol{x}_1^c,\boldsymbol{x}_2^c,\dots,\boldsymbol{x}_l^c]\,.$$

Assume that the columns are normalized to unity. Otherwise, we have to normalize them, which does not affect the spark and the mutual coherence. Then the corresponding Gram matrix is

$$G=X^TX=\begin{bmatrix}\boldsymbol{x}_1^{cT}\boldsymbol{x}_1^c & \dots & \boldsymbol{x}_1^{cT}\boldsymbol{x}_l^c\\ \boldsymbol{x}_2^{cT}\boldsymbol{x}_1^c & \dots & \boldsymbol{x}_2^{cT}\boldsymbol{x}_l^c\\ \vdots & \ddots & \vdots\\ \boldsymbol{x}_l^{cT}\boldsymbol{x}_1 & \dots & \boldsymbol{x}_l^{cT}\boldsymbol{x}_l^c\end{bmatrix}.\tag{16}$$

Observe that all diagonal elements are equal to one. Moreover, all the off-diagonal elements have absolute value less than or equal to the mutual coherence, by its definition. Let $1 \leq p \leq N$ be the maximum number for which the following holds:

$$(p-1)\mu(X) < 1. \tag{17}$$

This means that any submatrix G of $X^T X$, that results by combining any p columns of X, will be positive definite, according to the matrix theorem, stated in the hint, since the diagonal elements are 1 and the $p-1$ off-diagonals are bounded by $\mu(X)$ in absolute value. Moreover, (17) implies that this will also be true for any $m \leq p$. Thus, if this is valid, then any $m \leq p$ columns are linearly independent, hence

$$\text{spark}(X) \geq p+1, \tag{18}$$

However, since p is the maximum value for which (17) is valid, then

$$1 \leq p\mu(X). \tag{19}$$

Combining (19) with (17) results in

$$\text{spark}(X) \geq 1 + \frac{1}{\mu(X)}.$$

Note that if for $p = N$ (17) is valid, then spark attains its maximum value of $N+1$. If $p = 1$, this necessarily means that the mutual coherence is 1, hence the spark is bounded by two (only one column can be guaranteed to be linearly independent)

9.17. Show that if the underdetermined system of equations $\boldsymbol{y} = X\boldsymbol{\theta}$ accepts a solution such that

$$||\boldsymbol{\theta}||_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(X)}\right)$$

then the l_1 minimizer is equivalent to the l_0 one. Assume that the columns of X are normalized.

Solution: Recall that in this case, the l_0 minimizer is unique, so any other solution is bound to have larger l_0 norm. We only have to check that if this condition is valid, then any other solution will also have larger l_1 norm.

Let us assume that there exists a solution with smaller ℓ_1 norm. Then, this means that there will be a vector $\boldsymbol{z}$ in the null space of X such that

$$||\boldsymbol{\theta} + \boldsymbol{z}||_1 - ||\boldsymbol{\theta}||_1 \leq 0, \quad X\boldsymbol{z} = \boldsymbol{0},$$

where $\boldsymbol{\theta}$ is the solution with l_0 norm that satisfies the given condition. Let us assume, without loss of generality, that the $k = ||\boldsymbol{\theta}||_0$ nonzero elements are the first k elements of the vector. Then we obtain

$$\sum_{j=1}^{k} (|\theta_j + z_j| - |\theta_j|) + \sum_{j>k} |z_j| \leq 0.$$

Taking into consideration that $|a+b| - |b| \geq -|a|$, the previous inequality, if it is true, should also satisfy the following

$$-\sum_{j=1}^{k} |z_j| + \sum_{j>k} |z_j| \leq 0, \tag{20}$$

or adding and subtracting the term $\sum_{j=1}^{k} |z_j|$, the previous becomes equivalent to

$$||\boldsymbol{z}||_1 - 2\sum_{j=1}^{k} |z_j| \leq 0, \quad X\boldsymbol{z} = \boldsymbol{0}.$$

The second of the two requirements (i.e., that $\boldsymbol{z}$ lies in the null space) implies that

$$X^T X\boldsymbol{z} = \boldsymbol{0} \Rightarrow -\boldsymbol{z} = \left(X^T X - I\right)\boldsymbol{z}.$$

Hence the following is true

$$-z_i = \sum_{j=1}^{l} \boldsymbol{x}_i^{cT} \boldsymbol{x}_j^c z_j - z_i = \sum_{j=1,\ j\neq i}^{l} \boldsymbol{x}_i^{cT} \boldsymbol{x}_j^c z_j,$$

where we have used the definition of $X^T X$ in terms of its columns and the fact that X is normalized, hence the diagonal elements $\boldsymbol{x}_i^{cT} \boldsymbol{x}_i^c = 1$. Taking the absolute values of the previously obtained relationship we have

$$\begin{aligned} |z_i| \;&=\; |\sum_{j=1,\ j\neq i}^{l} \boldsymbol{x}_i^{cT} \boldsymbol{x}_j^c z_j| \leq \sum_{j=1,\ j\neq i}^{l} |\boldsymbol{x}_i^{cT} \boldsymbol{x}_j^j||z_j| \leq \mu(X) \sum_{j=1,\ j\neq i}^{l} |z_j| = \\ &\mu(X)||\boldsymbol{z}||_1 - \mu(X))|z_i|,\ i = 1, 2, \ldots, l \end{aligned}$$

or

$$|z_i| \leq \frac{\mu(X)}{1+\mu(X)} ||\boldsymbol{z}||_1 \Rightarrow -2\sum_{i=1}^{k} |z_i| \geq -2k\frac{\mu(X)}{1+\mu(X)} ||\boldsymbol{z}||_1.$$

The last one combined with (20) results in

$$||\boldsymbol{z}||_1 \left(1 - 2k\frac{\mu(X)}{1+\mu(X)}\right) \leq ||\boldsymbol{z}||_1 - 2\sum_{i=1}^{k} |z_i| \leq 0.$$

From the last one we obtain that

$$1 - 2k\frac{\mu(X)}{1+\mu(X)} \leq 0 \Rightarrow 1 \leq 2k\frac{\mu(X)}{1+\mu(X)},$$

and finally that

$$k \geq \frac{1}{2}\left(1 + \frac{1}{\mu(X)}\right),$$

which is obviously a contradiction to the condition which we have initially assumed to hold true, i.e., $k < \frac{1}{2}\left(1 + \frac{1}{\mu(X)}\right)$. Hence this proves the claim.

9.18. Prove that if the RIP of order k is valid for a matrix X and $\delta_k < 1$, then any $m < k$ columns of X are necessarily linearly independent.

Solution: Recall from the definition of RIP that it is valid for any k sparse vector, that is, for any vector that has up to k nonzero elements. Assume, now, that there exist $m \leq k$ columns that are linearly dependent. Without loss of generality assume these to be $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$. Then, there will be a sparse vector $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_m, 0, 0, \ldots, 0]^T$, $\theta_i \neq 0$, $i = 1, 2, \ldots, m$, such that

$$\sum_{i=1}^{m} \theta_i \boldsymbol{x}_i = X\boldsymbol{\theta} = 0 \Rightarrow ||X\boldsymbol{\theta}||_2^2 = 0,$$

which is a contradiction, since

$$||X\boldsymbol{\theta}||_2^2 \geq (1 - \delta_k)||\boldsymbol{\theta}||_2^2 > 0.$$

9.19. Show that if X satisfies the RIP of order k and some isometry constant δ_k so does the product $X\Psi$ if Ψ is an orthonormal matrix.

Solution: This is a direct consequence of the properties of the matrix orthonormality.

Bibliography

[Horn 85] Horn R.A., Johnson C.R. *Matrix Analysis*, Cambridge University Press, New York 1985.

[Golu 83] Golub G.H., Van Loan C.F. *Matrix Computations*, The John Hopkins University Press, Baltimore, 1983.

Solutions To Problems of Chapter 10

10.1. Show that the step, in a greedy algorithm, that selects the column of the sensing matrix, so that to maximize the correlation between the column and the currently available error vector $\boldsymbol{e}^{(i-1)}$, is equivalent with selecting the column that reduces the l_2 norm of the error vector. Hint: All the parameters obtained in previous steps are fixed, and the optimization is with respect to the new column as well as the corresponding weighting coefficient in the estimate of the parameter vector.

Solution: Let $\boldsymbol{e}^{(k-1)} = \boldsymbol{y} - X\boldsymbol{\theta}^{(k-1)}$ be the current estimate of the residual vector. We know that

$$\boldsymbol{e}^{(k-1)} \perp \boldsymbol{x}_{i_1}^c, \boldsymbol{x}_{i_2}^c, \ldots, \boldsymbol{x}_{i_{k-1}}^c.$$

Let $\boldsymbol{x}_{i_k}^c$ be the new column to be selected and θ_{i_k} the associated coefficient in $\boldsymbol{\theta}^{(k)}$. Then, if we fix all the previously selected columns and their associated coefficients in $\boldsymbol{\theta}^{(k)}$, the square ℓ_2 norm of the error becomes

$$||\boldsymbol{y} - X\boldsymbol{\theta}^{(k-1)} - \boldsymbol{x}_{i_k}^c \theta_{i_k}||_2^2 = ||\boldsymbol{e}^{(k-1)} - \boldsymbol{x}_{i_k}^c \theta_{i_k}||_2^2.$$

Minimizing the previous norm w.r.t. to θ_{i_k}, we easily obtain that

$$\hat{\theta}_{i_k} = \frac{\boldsymbol{x}_{i_k}^{c\ T} \boldsymbol{e}^{(k-1)}}{\boldsymbol{x}_{i_k}^{c\ T} \boldsymbol{x}_{i_k}^c}.$$

Plugging this optimal value into the error norm, it is easily shown that the error norm, corresponding to the optimal value, is equal to

$$||\boldsymbol{e}^{(k-1)} - \boldsymbol{x}_{i_k}^c \hat{\theta}_{i_k}||_2^2 = ||\boldsymbol{e}^{(k-1)}||_2^2 - \hat{\theta}_{i_k}^2 \boldsymbol{x}_{i_k}^{c\ T} \boldsymbol{x}_{i_k}.$$

Hence, the error norm is minimized if we select the column to satisfy

$$i_k = \arg\max_i \frac{|\boldsymbol{x}_i^{cT} \boldsymbol{e}^{(k-1)}|}{||\boldsymbol{x}_i^c||_2}.$$

10.2. Prove the proposition stating that if there is a sparse solution to the linear system $\boldsymbol{y} = X\boldsymbol{\theta}$ such that

$$k_0 = ||\boldsymbol{\theta}||_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(X)}\right),$$

where $\mu(X)$ is the mutual coherence of X, then the column selection procedure in a greedy algorithm will always select a column among

the active columns X, that correspond to the support of $\boldsymbol{\theta}$; that is, the columns that take part in the representation of $\boldsymbol{y}$ in terms of the columns of X.
Hint: Assume that

$$\boldsymbol{y} = \sum_{i=1}^{k_0} \theta_i \boldsymbol{x}_i^c.$$

Solution: Let $m \leq k_0$ be the index for which

$$|\theta_m|||\boldsymbol{x}_m^c|| \geq |\theta_i|||\boldsymbol{x}_i^c||, i \neq m,\ i \leq k_0,$$

where the vector norms are assumed to be the ℓ_2 norms. Then if we show that

$$\frac{|\boldsymbol{x}_m^{c\ T}\boldsymbol{y}|}{||\boldsymbol{x}_m^c||} > \frac{|\boldsymbol{x}_j^{cT}\boldsymbol{y}|}{||\boldsymbol{x}_j^c||}, \forall j > k_0, \tag{1}$$

then there is no way for any column with index outside the support to be selected in the first iteration; there is at least one column, with index within the support, whose correlation with $\boldsymbol{y}$ is greater than that of any column whose index is outside the support.

Combining (1) with the expansion equation leads to

$$\left|\sum_{i=1}^{k_0} \theta_i \frac{\boldsymbol{x}_m^{c\ T}\boldsymbol{x}_i^c}{||\boldsymbol{x}_m^c||}\right| > \left|\sum_{i=1}^{k_0} \theta_i \frac{\boldsymbol{x}_j^{cT}\boldsymbol{x}_i^c}{||\boldsymbol{x}_j^c||}\right|,\ j > k_0. \tag{2}$$

Using known identities of the absolute values, we have

$$\begin{aligned}
\left|\sum_{i=1}^{k_0} \theta_i \frac{\boldsymbol{x}_m^{c\ T}\boldsymbol{x}_i^c}{||\boldsymbol{x}_m^c||}\right| &\geq |\theta_m|||\boldsymbol{x}_m^c|| - \sum_{i=1, i\neq m}^{k_0} |\theta_i| \left|\frac{\boldsymbol{x}_m^{c\ T}\boldsymbol{x}_i^c}{||\boldsymbol{x}_m^c||}\right| \\
&\geq |\theta_m|||\boldsymbol{x}_m^c|| - \sum_{i=1, i\neq m}^{k_0} |\theta_m|||\boldsymbol{x}_m^c||\mu(X) \\
&\geq |\theta_m|||\boldsymbol{x}_m^c|| \left(1 - \mu(X)(k_0 - 1)\right).
\end{aligned}$$

On the other hand, we have that

$$\begin{aligned}
\left|\sum_{i=1}^{k_0} \theta_i \frac{\boldsymbol{x}_j^{cT}\boldsymbol{x}_i^c}{||\boldsymbol{x}_j^c||}\right| &\leq \sum_{i=1}^{k_0} |\theta_i| \left|\frac{\boldsymbol{x}_j^{cT}\boldsymbol{x}_i^c}{||\boldsymbol{x}_j^c||}\right| \\
&\leq \sum_{i=1}^{k_0} |\theta_i|||\boldsymbol{x}_i^c||\mu(X) \\
&\leq |\theta_m|||\boldsymbol{x}_m^c||\mu(X)k_0.
\end{aligned}$$

Hence, in order (1) to hold true it suffices that

$$1 - \mu(X)k_0 + \mu(X) \geq \mu(X)k_0 \Rightarrow k_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(X)}\right).$$

For the second iteration step, we have that

$$\boldsymbol{e}^{(1)} = \boldsymbol{y} - \theta_{i_1}\boldsymbol{x}_{i_1}^c = \sum_{i=1}^{k_0} \tilde{\theta}_i \boldsymbol{x}_i^c,$$

and we can repeat the previous arguments in a similar way.

10.3. Give an explanation to justify why in step 4 of the CoSaMP algorithm the value of t is taken to be equal to $2k$

Solution: The estimates, $\boldsymbol{\theta}^{(i)}$, obtained at each iteration are always k-sparse vectors, due to the hard thresholding operation in step 7. Hence

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

and for near orthogonal matrices, $X^TX \approx I$, the difference between two k-sparse vectors, the true one and the currently available estimate, will be $2k$-sparse, approximately.

10.4. Show that if

$$J(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) = \frac{1}{2}||\boldsymbol{y} - X\boldsymbol{\theta}||_2^2 + \lambda||\boldsymbol{\theta}||_1 + \frac{1}{2}d(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}),$$

where

$$d(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) := c||\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}||_2^2 - ||X\boldsymbol{\theta} - X\tilde{\boldsymbol{\theta}}||_2^2,$$

then minimization results to

$$\hat{\boldsymbol{\theta}} = S_{\lambda/c}\left(\frac{1}{c}X^T(\boldsymbol{y} - X\tilde{\boldsymbol{\theta}}) + \tilde{\boldsymbol{\theta}}\right).$$

Solution: Expanding the surrogate function, we obtain

$$\begin{aligned} J(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) \;=\; & \frac{1}{2}\boldsymbol{y}^T\boldsymbol{y} + \frac{1}{2}\boldsymbol{\theta}^TX^TX\boldsymbol{\theta} - \boldsymbol{y}^TX\boldsymbol{\theta} + \lambda||\boldsymbol{\theta}||_1 + \\ & \frac{1}{2}c(\boldsymbol{\theta}^T\boldsymbol{\theta} - 2\tilde{\boldsymbol{\theta}}^T\boldsymbol{\theta} + \tilde{\boldsymbol{\theta}}^T\tilde{\boldsymbol{\theta}}) - \\ & \frac{1}{2}\boldsymbol{\theta}^TX^TX\boldsymbol{\theta} + \frac{1}{2}\tilde{\boldsymbol{\theta}}^TX^TX\boldsymbol{\theta} + \frac{1}{2}\boldsymbol{\theta}^TX^TX\tilde{\boldsymbol{\theta}} - \frac{1}{2}\tilde{\boldsymbol{\theta}}^TX^TX\tilde{\boldsymbol{\theta}}. \end{aligned}$$

Note that the term that involves $\boldsymbol{\theta}^TX^TX\boldsymbol{\theta}$ is canceled out. When dealing with the LASSO in the theory, in order to come up with an analytic solution, we had to assume orthogonality on X. With the surrogate function, this term is canceled out by the extra term that we have added. Taking the subgradient of the previous function we have

$$c\boldsymbol{\theta} - c\tilde{\boldsymbol{\theta}} - X^T\boldsymbol{y} + X^TX\tilde{\boldsymbol{\theta}} + \lambda\partial||\boldsymbol{\theta}||_1 = \boldsymbol{0},$$

and following the same procedure as we did in the theory for the LASSO, for the orthogonal case, we end up with the required solution.

10.5. Prove the basic recursion of the parallel coordinate descent algorithm. Hint: Assume that at the ith iteration, it is the turn of the jth component to be updated, so that the following is minimized

$$J(\theta_j) = \frac{1}{2}||\boldsymbol{y} - X\boldsymbol{\theta}^{(i-1)} + \theta_j^{(i-1)}\boldsymbol{x_j} - \theta_j \boldsymbol{x}_j||_2^2 + \lambda|\theta_j|.$$

Solution: Basically in the cost, one removes the contribution of the jth component, as this was estimated by the previous step and attempts the find a new update of the corresponding coefficient that minimizes the function, by keeping the rest of the elements unchanged. Expanding the cost function we get

$$J(\theta_j) = \frac{1}{2}(\boldsymbol{y} - \boldsymbol{a})^T(\boldsymbol{y} - \boldsymbol{a}) - \theta_j(\boldsymbol{y} - \boldsymbol{a})^T\boldsymbol{x}_j + \frac{1}{2}\theta_j^2 \boldsymbol{x}_j^T\boldsymbol{x}_j + \lambda|\theta_j|,$$

where $\boldsymbol{a} = X\boldsymbol{\theta}^{(i-1)} - \theta_j^{(i-1)}\boldsymbol{x}_j$. Taking the subgradient with respect to θ_j and equating to zero, we obtain

$$-(\boldsymbol{y} - \boldsymbol{a})^T\boldsymbol{x}_j + \theta_j\boldsymbol{x}_j^T\boldsymbol{x}_j + \lambda\partial|\theta_j| = \boldsymbol{0},$$

and following the familiar procedure, as the one used to solve LASSO for the case of an orthogonal matrix X, we obtain

$$\theta_j^{(i)} = S_{\lambda/||\boldsymbol{x}_j||_2^2}\left(\frac{(\boldsymbol{y} - X\boldsymbol{\theta}^{(i-1)})^T\boldsymbol{x}_j + \boldsymbol{x}_j^T\boldsymbol{x}_j\theta_j^{(i-1)}}{\boldsymbol{x}_j^T\boldsymbol{x}_j}\right).$$

This completes the proof.

10.6. Derive the iterative scheme to minimize the weighted ℓ_1 ball, using a majorization-minimization procedure to minimize $\sum_{i=1}^{l}\ln(|\theta_i| + \epsilon)$, subject to the measurements set, $\boldsymbol{y} = X\boldsymbol{\theta}$.
Hint: Use the linearization of the logarithmic function to bound it from above, since it is a concave function and its graph is located below its tangent.

Solution: Our task is

$$\begin{array}{rl} \text{minimize w.r. } \boldsymbol{\theta} & \displaystyle\sum_{j=1}^{l}\ln(|\theta_j| + \epsilon), \\ \text{s.t.} & X\boldsymbol{\theta} = \boldsymbol{y}. \end{array}$$

It is known from optimization theory, and it is easily checked out, that the above task is equivalent to

$$\begin{array}{rl} \text{minimize w.r. } (\boldsymbol{\theta},\ \boldsymbol{u}) & \displaystyle\sum_{j=1}^{l}\ln(u_j + \epsilon). \\ \text{s.t.} & X\boldsymbol{\theta} = \boldsymbol{y} \\ & |\theta_j| \leq u_j,\ j = 1, 2, \ldots, l. \end{array}$$

Instead, at each iteration step, we will try to minimize the linearization of the function, around its previous estimate. We have seen that this is also an alternative way to derive the gradient descent, albeit there, we use regularized local optimization. Hence the task now becomes

$$\begin{aligned} \boldsymbol{\theta}^{(i)} &= \arg\min_{\boldsymbol{\theta},\boldsymbol{u}} \left(\sum_{j=1}^{l} \ln(u_j^{(i-1)} + \epsilon) + (\boldsymbol{u} - \boldsymbol{u}^{(i-1)})^T \nabla \sum_{j=1}^{l} (\ln(u_j + \epsilon)|_{u_j = u_j^{(i-1)}} \right) \\ \text{s.t.} \quad & X\boldsymbol{\theta} = \boldsymbol{y} \\ & |\theta_j| \leq u_j, \end{aligned}$$

which is equivalent with

$$\begin{aligned} \boldsymbol{\theta}^{(i)} &= \arg\min_{\boldsymbol{\theta},\boldsymbol{u}} \left(\sum_{j=1}^{l} \frac{u_j}{u_j^{(i-1)} + \epsilon} \right) \\ s.t. \quad & X\boldsymbol{\theta} = \boldsymbol{y} \\ & |\theta_j| \leq u_j, \end{aligned}$$

which is finally equivalent to

$$\begin{aligned} \boldsymbol{\theta}^{(i)} &= \arg\min_{\boldsymbol{\theta}} \sum_{j=1}^{l} \frac{|\theta_j|}{|\theta_j^{(i-1)}| + \epsilon} \\ s.t. \quad & X\boldsymbol{\theta} = \boldsymbol{y}. \end{aligned}$$

10.7. Show that the weighted ℓ_1 ball, used in SpAPSM, is upper bounded by the ℓ_0 norm of the target vector.

Solution: Assume that $\boldsymbol{\theta}(n)$ converges to the true vector. Moreover assume that $\acute{\epsilon}_n \geq \acute{\epsilon} > 0$. Then

$$\sum_{j=1}^{l} w_j(n)|\theta_j(n)| \leq \sum_{j=1}^{l} \frac{|\theta_j(n)|}{|\theta_j(n)| + \acute{\epsilon}}.$$

Hence the following is true

$$\begin{aligned} \limsup_{n\to\infty} \sum_{j=1}^{l} w_j(n)|\theta_j(n)| &\leq \limsup_{n\to\infty} \sum_{j=1}^{l} \frac{|\theta_j(n)|}{|\theta_j(n)| + \acute{\epsilon}} = \lim_{n\to\infty} \sum_{j=1}^{l} \frac{|\theta_j(n)|}{|\theta_j(n)| + \acute{\epsilon}} \\ &= \sum_{j\in\text{Support}(\boldsymbol{\theta}_*)} \frac{|\theta_{*,j}|}{|\theta_{*,j}| + \acute{\epsilon}} + \sum_{j\notin\text{Support}(\boldsymbol{\theta}_*)} \frac{|\theta_{*,j}|}{|\theta_{*,j}| + \acute{\epsilon}} \\ &< \sum_{j\in\text{Support}(\boldsymbol{\theta}_*)} \frac{|\theta_{*,j}|}{|\theta_{*,j}|} = ||\boldsymbol{\theta}_*||_0. \end{aligned}$$

10.8. Show that the canonical dual frame minimizes the total ℓ_2 norm of the dual frame, i.e.,

$$\sum_{i\in\mathcal{I}} \left\|\tilde{\boldsymbol{\psi}}_i\right\|_2^2.$$

Hint: Use the result of Problem 9.10.

Solution: We know from the theory that the matrices corresponding to the frame and its dual satisfy the following condition

$$\tilde{\Psi}\Psi^H = \Psi\tilde{\Psi}^H = I.$$

Let us derive from the above the columns of $\tilde{\Psi}^H$ one by one. We will work for the first and the rest can be obtained in a similar way. Let $\tilde{\boldsymbol{\psi}}_1^*$ denote the first column vector of $\tilde{\Psi}$. Then we have

$$\Psi\tilde{\boldsymbol{\psi}}_1^* = [1, 0, \dots, 0]^T.$$

Then we know that the minimum norm solution for the previous system is given by

$$\tilde{\boldsymbol{\psi}}_1^* = \Psi^H(\Psi\Psi^H)^{-1}[1, 0, \dots, 0]^T.$$

Repeating the above for all columns we get

$$\tilde{\Psi}^H = \Psi^H(\Psi\Psi^H)^{-1} \Longrightarrow \tilde{\Psi} = (\Psi\Psi^H)^{-1}\Psi$$

10.9. Show that Parseval's tight frames are self dual.

Solution: Let the dimensionality of the frame's matrix Ψ be $N \times p$. From the definition of the Parseval tight frame (PTFP) we have that

$$\forall \boldsymbol{s},\ \sum_{i=1}^{p} |\langle \boldsymbol{\psi}_i, \boldsymbol{s}\rangle|^2 = \sum_{i=1}^{p} \boldsymbol{s}^H\boldsymbol{\psi}_i\boldsymbol{\psi}_i^H\boldsymbol{s} = \boldsymbol{s}^H\boldsymbol{s}.$$

Hence

$$\sum_{i=1}^{p} \boldsymbol{s}^H\boldsymbol{\psi}_i\boldsymbol{\psi}_i^H\boldsymbol{s} = \boldsymbol{s}^H\sum_{i=1}^{p}\left(\boldsymbol{\psi}_i\boldsymbol{\psi}_i^H\right)\boldsymbol{s} = \boldsymbol{s}^H\Psi\Psi^H\boldsymbol{s} = \boldsymbol{s}^H\boldsymbol{s}.$$

Hence $\Psi\Psi^H = I$, which obviously proves the claim.

10.10. Prove that the bounds A, B of a frame coincide with the maximum and minimum eigenvalues of the matrix product $\Psi\Psi^H$.

Solution: From the definition of a frame we have

$$A\left\|\boldsymbol{s}\right\|_2^2 \leq \sum_{i=1}^{p} |\langle \boldsymbol{\psi}_i, \boldsymbol{s}\rangle|^2 = \boldsymbol{s}^H\Psi\Psi^H\boldsymbol{s} \leq B\left\|\boldsymbol{s}\right\|_2^2.$$

However, we know from linear algebra that the minimum and maximum values for the Rayleigh quotient

$$\frac{\boldsymbol{s}^H P \boldsymbol{s}}{\boldsymbol{s}^H \boldsymbol{s}},$$

for any Hermitian matrix P, are the maximum and minimum eigenvalues of P. This proves the claim.

Bibliography

[Horn 85] Horn R.A., Johnson C.R. *Matrix Analysis*, Cambridge University Press, New York 1985.

[Golu 83] Golub G.H., Van Loan C.F. *Matrix Computations*, The John Hopkins University Press, Baltimore, 1983.

Solutions To Problems of Chapter 11

11.1. Derive the formula for the number of groupings $\mathcal{O}(N, l)$ in Cover's theorem.
Hint: Show first the following recursion

$$\mathcal{O}(N+1, l) = \mathcal{O}(N, l) + \mathcal{O}(N, l-1).$$

To this end, start with N points and add an extra one. Show that the extra number of linear dichotomies is solely due to those, for the N data point case, which could be drawn via the new point.

Solution: Let us assume that we start with N points in the l-dimensional space, and we add an extra point P. Then, the $\mathcal{O}(N, l)$ old dichotomies fall into either of the two categories.

- For those dichotomies where the dichotomizing plane can not pass through P, the new point P can *only* belong to the old dichotomies.
- For those dichotomies for which the plane could pass through P, then to each one of the old dichotomies correspond two new dichotomies. This is because by shifting infinitesimally the plane, P can change class ownership.

Thus, the total number of dichotomies with the $N+1$ points will be

$$\mathcal{O}(N+1, l) \;=\; \mathcal{O}(N, l) + \mathcal{O}(N, l-1). \qquad (1)$$

The second term is the number of dichotomies of N points in the l-dimensional space that are constrained to pass through a specific point P. This is equivalent with drawing dichotomies of N points in the $(l-1)$-dimensional space.
In the sequel, we will show by induction that

$$\mathcal{O}(N, l) \;=\; \begin{pmatrix} k \\ 0 \end{pmatrix} \mathcal{O}(N-k, l) + \begin{pmatrix} k \\ 1 \end{pmatrix} \mathcal{O}(N-k, l-1) + \dots$$

$$+ \begin{pmatrix} k \\ k \end{pmatrix} \mathcal{O}(N-k, l-k) \;\; \texttt{for} \;\; k = 1, 2, \dots, N-1. \qquad (2)$$

The above is true for $k = 1$. Indeed

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix} = 1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Then, assume it is true for some k. We will show that it is also true for $k + 1$. We have that

$$\mathcal{O}(N-k, l) \;=\; \mathcal{O}(N-k-1, l) + \mathcal{O}(N-k-1, l-1).$$

Thus

$$\mathcal{O}(N,l) \;=\; \begin{pmatrix} k \\ 0 \end{pmatrix} \{\mathcal{O}(N-k-1,l) + \mathcal{O}(N-k-1,l-1)\}$$
$$+\begin{pmatrix} k \\ 1 \end{pmatrix} \{\mathcal{O}(N-k-1,l-1) + O(N-k-1,l-2)\} +$$
$$\cdots + \begin{pmatrix} k \\ k \end{pmatrix} \{\mathcal{O}(N-k-1,l-k) + \mathcal{O}(N-k-1,l-k-1)\}.$$

However,

$$\begin{pmatrix} k \\ r \end{pmatrix} + \begin{pmatrix} k \\ r+1 \end{pmatrix} \;=\; \begin{pmatrix} k+1 \\ r+1 \end{pmatrix}.$$

Indeed,

$$\begin{pmatrix} k \\ r \end{pmatrix} \;=\; \frac{k!}{r!(k-r)!}, = \frac{k!}{r!(k-r-1)!(k-r)},$$
$$\begin{pmatrix} k \\ r+1 \end{pmatrix} \;=\; \frac{k!}{r!(r+1)(k-r-1)!},$$

$$\begin{pmatrix} k \\ r \end{pmatrix} + \begin{pmatrix} k \\ r+1 \end{pmatrix} \;=\; \frac{(r+1)k! + (k-r)k!}{r!(r+1)(k-r-1)!(k-r)} \;=$$
$$\frac{(k+1)k!}{(r+1)!(k+1-r-1)!} \;=\; \begin{pmatrix} k+1 \\ r+1 \end{pmatrix}.$$

Also

$$\begin{pmatrix} k \\ 0 \end{pmatrix} \;=\; \begin{pmatrix} k+1 \\ 0 \end{pmatrix} \texttt{and} \begin{pmatrix} k \\ k \end{pmatrix} \;=\; \begin{pmatrix} k+1 \\ k+1 \end{pmatrix}.$$

Hence,

$$\mathcal{O}(N,l) = \begin{pmatrix} k+1 \\ 0 \end{pmatrix} \mathcal{O}(N-k-1,l) + \begin{pmatrix} k+1 \\ 1 \end{pmatrix} \mathcal{O}(N-k-1,l-1) +$$
$$\ldots + \begin{pmatrix} k+1 \\ k+1 \end{pmatrix} \mathcal{O}(N-k-1,l-k-1).$$

Thus, (2) is true. By convention $O(N,l) \;=\; 0$ for $l \;<\; 0$. Then, setting in (2) $k = N-2$ and taking into account that

$$O(1,l) \;=\; 2, \qquad l \;\geq\; 0,$$

we obtain

$$\mathcal{O}(N,l) \;=\; 2\sum_{i=0}^{l} \begin{pmatrix} N-1 \\ i \end{pmatrix}.$$

Note that the convention $\mathcal{O}(1,0) = 2$ is in line with the recursion in (1). Indeed, for $l = 1$ and $N = 1$ we have

$$\mathcal{O}(2,1) \;=\; \mathcal{O}(1,1) \;+\; \mathcal{O}(1,0).$$

However, $\mathcal{O}(2,1) = 4$, since 2 points in a line can be separated by a point in four possible groupings. Also $\mathcal{O}(1,1) = 2$, since a point can belong to either of two classes. Then $\mathcal{O}(1,0) = 2$ is in agreement with the recursion.

11.2. Show that if $N = 2(l+1)$, the number of linear dichotomies in Cover's theorem is equal to 2^{2l+1}.
Hint: Use the identity

$$\sum_{i=1}^{j} \binom{j}{i} = 2^j,$$

and recall that

$$\binom{2n+1}{n-i+1} = \binom{2n+1}{n+i}.$$

Solution: From the theory, we have that for $N = 2l+2$,

$$\sum_{i=0}^{2l+1} \binom{2l+1}{i} = 2^{2l+1}.$$

Then,

$$\mathcal{O}(2l+2, l) = 2\sum_{i=0}^{l} \binom{2l+1}{i} = 2\left[\frac{1}{2}\left[\sum_{i=0}^{l} \binom{2l+1}{i} + \sum_{i=l+1}^{2l+1} \binom{2l+1}{i}\right]\right]$$

$$= 2\Big[\frac{1}{2}\,2^{2l+1}\Big] = 2^{N-1}.$$

11.3. Show that the reproducing kernel is a positive definite one.

Solution: Consider $N > 0$, the real numbers $a_1, \ldots, a_N$, and the elements $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \mathcal{X}$. Then

$$\begin{aligned}
\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m \kappa(\boldsymbol{x}_n, \boldsymbol{x}_m) &= \sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m \langle \kappa(\cdot, \boldsymbol{x}_n), \kappa(\cdot, \boldsymbol{x}_m)\rangle \\
&= \sum_{n=1}^{N} a_n \left\langle \kappa(\cdot, \boldsymbol{x}_n), \sum_{m=1}^{N} a_m \kappa(\cdot, \boldsymbol{x}_m)\right\rangle \\
&= \left\langle \sum_{n=1}^{N} a_n \kappa(\cdot, \boldsymbol{x}_n), \sum_{m=1}^{N} a_m \kappa(\cdot, \boldsymbol{x}_m)\right\rangle \\
&= \left\| \sum_{n=1}^{N} a_n \kappa(\cdot, \boldsymbol{x}_n)\right\|^2 \geq 0.
\end{aligned}$$

11.4. Show that if $\kappa(\cdot,\cdot)$ is the reproducing kernel in a RKHS, $\mathbb{H}$, then

$$\mathbb{H} = \overline{\text{span}\{\kappa(\cdot, \boldsymbol{x}), x \in \mathcal{X}\}}.$$

Solution: We will show that the only function in $\mathbb{H}$ which is orthogonal to $A = \overline{\text{span}\{\kappa(\cdot, \boldsymbol{x}), x \in \mathcal{X}\}}$ is the zero function. Let $f \in \mathbb{H}$ be a function that is orthogonal to A. Then

$$f(\boldsymbol{x}) = \langle f, \kappa(\cdot, \boldsymbol{x})\rangle = 0, \ \forall \boldsymbol{x} \in X,$$

which holds true for $f = 0$. Thus

$$A^{\perp} = \overline{A^{\perp}} = \{0\}.$$

Let now $f \in \mathbb{H}$ such that $f \notin \bar{A}$. By its definition, $\bar{A}$ is closed and also is convex, being a subspace. Note also that

$$\bar{A} \subseteq \mathcal{H}.$$

This is because, $\mathbb{H}$ by its definition, is complete and hence it contains all the limit points of sequences in $\mathbb{H}$. Also, by its definition, any $\kappa(\cdot, \boldsymbol{x}) \in \mathbb{H}$, $\boldsymbol{x} \in X$. Thus according to what we have said in Chapter 8, there is a $g \in \bar{A}$ which is the projection of f on $\bar{A}$. Also, since $\bar{A}$ is a closed subspace, we know from (8.20)

$$\mathbb{H} = \bar{A} + \bar{A}^{\perp}.$$

Hence

$$f - g \ \in \bar{A}^{\perp},$$

which means that $f - g$ is orthogonal to any element in $\bar{A}$, hence in A. This leads to

$$f - g = 0.$$

In other words, if there exists such an f, it coincides with its projection on $\bar{A}$, hence

$$f \in \bar{A}.$$

This contradicts our assumption that $f \notin \bar{A}$.

11.5. Show the Cauchy-Schwarz inequality for kernels, that is,

$$\|\kappa(\boldsymbol{x}, \boldsymbol{y})\|^2 \leq \kappa(\boldsymbol{x}, \boldsymbol{x})\kappa(\boldsymbol{y}, \boldsymbol{y}).$$

Solution: Let

$$\phi(\boldsymbol{x}) := \kappa(\cdot, \boldsymbol{x}),$$

and

$$\phi(\boldsymbol{y}) := \kappa(\cdot, \boldsymbol{y}).$$

Then,

$$\|\langle\phi(\boldsymbol{x}), \phi(\boldsymbol{y})\rangle\| \leq \|\phi(\boldsymbol{x})\| \cdot \|\phi(\boldsymbol{y})\|,$$

or

$$\kappa(\boldsymbol{x}, \boldsymbol{y}) \leq \sqrt{\kappa(\boldsymbol{x}, \boldsymbol{x})}\sqrt{\kappa(\boldsymbol{y}, \boldsymbol{y})}.$$

11.6. Show that if

$$\kappa_i(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \longmapsto \mathbb{R},\ i = 1, 2$$

are kernels then:

- $\kappa(\boldsymbol{x}, \boldsymbol{y}) = \kappa_1(\boldsymbol{x}, \boldsymbol{y}) + \kappa_2(\boldsymbol{x}, \boldsymbol{y})$ is also a kernel.
- $a\kappa(\boldsymbol{x}, \boldsymbol{y}),\ a > 0$ is also a kernel.
- $\kappa(\boldsymbol{x}, \boldsymbol{y}) = \kappa_1(\boldsymbol{x}, \boldsymbol{y})\kappa_2(\boldsymbol{x}, \boldsymbol{y})$ is also a kernel.

.

Solution: Recall that it suffices to show that the respective kernel matrices of the new constructed functions are positive semidefinite.

- For the addition, the i, j element of $\mathcal{K}$ is given by

  $$[K]_{i,j} = \kappa_1(\boldsymbol{x}_i, \boldsymbol{x}_j) + \kappa_2(\boldsymbol{x}_i, \boldsymbol{x}_j) = [\mathcal{K}_1]_{i,j} + [\mathcal{K}_2]_{i,j}$$

  or

  $$\mathcal{K} = \mathcal{K}_1 + \mathcal{K}_2.$$

  Since $\mathcal{K}_1$ and $\mathcal{K}_2$ are positive semidefinite, so is their sum.
- For the case $\kappa(\boldsymbol{x}, \boldsymbol{y}) = a\kappa(\boldsymbol{x}, \boldsymbol{y}),\ a > 0$ it is trivial that $\mathcal{K}$ is also positive semidefinite for $a > 0$.
- The proof for the product is slightly more "clever". Note that the kernel matrix consists of respective products, e.g.,

  $$\mathcal{K} = \begin{bmatrix} \kappa_1(\boldsymbol{x}_1, \boldsymbol{y}_1)\kappa_2(\boldsymbol{x}_1, \boldsymbol{y}_1) & \kappa_1(\boldsymbol{x}_1, \boldsymbol{y}_2)\kappa_2(\boldsymbol{x}_1, \boldsymbol{y}_2) \\ \kappa_1(\boldsymbol{x}_2, \boldsymbol{y}_1)\kappa_2(\boldsymbol{x}_2, \boldsymbol{y}_1) & \kappa_1(\boldsymbol{x}_2, \boldsymbol{y}_2)\kappa_2(\boldsymbol{x}_2, \boldsymbol{y}_2) \end{bmatrix}$$

  for $N = 2$. Observe that $\mathcal{K}$ results as a principal submatrix of Kronecker product $\mathcal{K}_1 \otimes \mathcal{K}_2$, which is the $N^2 \times N^2$ matrix, that results if each element of $\mathcal{K}_1$ is replaced by $\mathcal{K}_2$ multiplied by the respective element. $\mathcal{K}$ results from $\mathcal{K}_1 \otimes \mathcal{K}_2$ by keeping a set of columns and the same set of rows. Verify it for the case $N = 2$. Hence $\forall \boldsymbol{a} \in \mathbb{R}^N,\ \exists \boldsymbol{b} \in \mathbb{R}^{N^2}$:

  $$\boldsymbol{a}^T K \boldsymbol{a} = \boldsymbol{b}^T (\mathcal{K}_1 \otimes \mathcal{K}_2)\boldsymbol{b} \geq 0.$$

  The latter inequality is true since we know from linear algebra that if $\mathcal{K}_1,\ \mathcal{K}_2$ are positive semidefinite, so their Kronecker product is.

11.7. Derive Equation (11.25).

Solution: The starting point is

$$J(\boldsymbol{\theta}) = \sum_{n=1}^{N} \left(y_n - \sum_{m=1}^{N} \theta_m \kappa(\boldsymbol{x}_n, \boldsymbol{x}_m) \right)^2 + \langle f, f \rangle.$$

Substitute in the regularizer the form of f to get

$$\begin{aligned}
\langle f, f\rangle &= \left\langle \sum_{n=1}^{N} \theta_n \kappa(\cdot, \boldsymbol{x}_n), \sum_{m=1}^{N} \theta_m \kappa(\cdot, \boldsymbol{x}_m) \right\rangle \\
&= \sum_{n=1}^{N} \theta_n \left\langle \kappa(\cdot, \boldsymbol{x}_n), \sum_{m=1}^{N} \theta_m \kappa(\cdot, \boldsymbol{x}_m) \right\rangle \\
&= \sum_{n=1}^{N} \theta_n \sum_{m=1}^{N} \theta_m \left\langle \kappa(\cdot, \boldsymbol{x}_n), \kappa(\cdot, \boldsymbol{x}_m) \right\rangle \\
&= \sum_{n=1}^{N} \theta_n \sum_{m=1}^{N} \theta_m \kappa(\boldsymbol{x}_n, \boldsymbol{x}_m) \\
&= \boldsymbol{\theta}^T \mathcal{K} \boldsymbol{\theta} = \boldsymbol{\theta}^T \mathcal{K}^T \boldsymbol{\theta}.
\end{aligned}$$

The first term can be written as

$$\sum_{n=1}^{N} \left(y_n - \boldsymbol{\theta}^T \boldsymbol{k}(\boldsymbol{x}_n) \right)^2 = \|\boldsymbol{y} - \mathcal{K}\boldsymbol{\theta}\|^2,$$

where

$$\boldsymbol{k}(\cdot) := [\kappa(\cdot, \boldsymbol{x}_1) \cdots \kappa(\cdot, \boldsymbol{x}_N)]^T,$$

with $\mathcal{K}$ being the kernel matrix. This proves the claim.

11.8. Show that the solution for the parameters, $\hat{\boldsymbol{\theta}}$, for the kernel ridge regression, if a bias term, b, is present, is given by

$$\begin{bmatrix} \mathcal{K} + CI & \mathbf{1} \\ \mathbf{1}^T \mathcal{K} & N \end{bmatrix} \begin{bmatrix} \boldsymbol{\theta} \\ b \end{bmatrix} = \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{y}^T \mathbf{1} \end{bmatrix},$$

where $\mathbf{1}$ is the vector with all its elements being equal to one. Invertibility of the kernel matrix has been assumed.

Solution: In this case, the unknown coefficients are estimated by minimizing $J(\boldsymbol{\theta}, b)$, where

$$J(\boldsymbol{\theta}, b) = \sum_{n=1}^{N} \left(y_n - \sum_{m=1}^{N} \theta_m \kappa(\boldsymbol{x}_n, \boldsymbol{x}_m) - b \right) + C\langle f, f\rangle. \tag{3}$$

Equation (3) can be recast in a more compact form as follows:

$$J(\boldsymbol{\theta}, b) = (\boldsymbol{y} - \mathcal{K}\boldsymbol{\theta} - b\mathbf{1})^T (\boldsymbol{y} - \mathcal{K}\boldsymbol{\theta} - b\mathbf{1}) + C\boldsymbol{\theta}^T \mathcal{K} \boldsymbol{\theta}. \tag{4}$$

As $J(\boldsymbol{\theta}, b)$ is a strictly convex function, the unique minimum can be obtained by solving

$$\frac{\partial J}{\partial \boldsymbol{\theta}} = 0, \quad \frac{\partial J}{\partial b} = 0.$$

Hence, we obtain the system of equations

$$\begin{aligned} (\mathcal{K}^2 + C\mathcal{K})\boldsymbol{\theta} + b\mathcal{K}\mathbf{1} &= \mathcal{K}\boldsymbol{y}, \\ \mathbf{1}^T\mathcal{K}\boldsymbol{\theta} + Nb &= \boldsymbol{y}^T\mathbf{1}. \end{aligned}$$

Assuming that $\mathcal{K}$ is invertible, we obtain the result.

11.9. Derive Equation (11.56).

Solution: The dual representation of the Lagrangian in (11.55) can be written as,

$$L(\boldsymbol{\lambda}) = \boldsymbol{y}^T\boldsymbol{\lambda} - \frac{1}{4C}\boldsymbol{\lambda}^T\mathcal{K}\boldsymbol{\lambda} - \frac{1}{4}\boldsymbol{\lambda}^T\boldsymbol{\lambda},$$

where $\mathcal{K}$ is the kernel matrix and

$$\boldsymbol{\lambda} = [\lambda_1, \dots, \lambda_N]^T.$$

Taking the gradient with respect to $\boldsymbol{\lambda}$ and equating to zero we obtain

$$\boldsymbol{\lambda} = 2C\left(\mathcal{K} + CI\right)^{-1}\boldsymbol{y}.$$

11.10. Derive the dual cost function associated with the linear ϵ-insensitive loss function.

Solution: From the text, the Lagrangian is given by

$$\begin{aligned} L(\boldsymbol{\theta}, \theta_0, \tilde{\boldsymbol{\xi}}, \boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\mu}) &= \frac{1}{2}\|\boldsymbol{\theta}\|^2 + C\left(\sum_{n=1}^N \xi_n + \sum_{n=1}^N \tilde{\xi}_n\right) \\ &+ \sum_{n=1}^N \tilde{\lambda}_n(y_n - \boldsymbol{\theta}^T\boldsymbol{x}_n - \theta_0 - \epsilon - \tilde{\xi}_n) \\ &+ \sum_{n=1}^N \lambda_n(\boldsymbol{\theta}^T\boldsymbol{x}_n + \theta_0 - y_n - \epsilon - \xi_n) \\ &- \sum_{n=1}^N \tilde{\mu}_n\tilde{\xi}_n - \sum_{n=1}^N \mu_n\xi_n. \end{aligned}$$

Let us plug in place of $\boldsymbol{\theta}$ the corresponding KKT condition, i.e.,

$$\boldsymbol{\theta} = \sum_{n=1}^N (\tilde{\lambda}_n - \lambda_n)\boldsymbol{x}_n.$$

Then we obtain

$$\begin{aligned} L &= \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}(\tilde{\lambda}_n-\lambda_n)(\tilde{\lambda}_m-\lambda_m)\boldsymbol{x}_n^T\boldsymbol{x}_m + C\left(\sum_{n=1}^{N}\xi_n+\sum_{n=1}^{N}\tilde{\xi}_n\right) \\ & -\sum_{n=1}^{N}\tilde{\lambda}_n\sum_{m=1}(\tilde{\lambda}_n-\lambda_n)\boldsymbol{x}_m^T\boldsymbol{x}_n + \sum_{n=1}^{N}\lambda_n\sum_{m=1}(\tilde{\lambda}_m-\lambda_m)\boldsymbol{x}_m^T\boldsymbol{x}_n \\ & +\sum_{n=1}^{N}\tilde{\lambda}_n y_n - \sum_{n=1}^{N}\tilde{\lambda}_n\theta_0 - \epsilon\sum_{n=1}^{N}\tilde{\lambda}_n - \sum_{n=1}^{N}\tilde{\lambda}_n\tilde{\xi}_n \\ & +\sum_{n=1}^{N}\lambda_n\theta_0 - \sum_{n=1}^{N}\lambda_n y_n - \epsilon\sum_{n=1}^{N}\lambda_n - \sum_{n=1}^{N}\lambda_n\xi_n \\ & -\sum_{n=1}^{N}\tilde{\mu}_n\tilde{\xi}_n - \sum_{n=1}^{N}\mu_n\xi_n. \end{aligned}$$

Taking into account, from the KKT conditions given in the text, that

$$C-\tilde{\lambda}_n-\tilde{\mu}_n = C-\lambda_n-\mu_n = 0,\ n=1,2,\ldots,N,$$

and that

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

we obtain

$$\begin{aligned} L &= \sum_{n=1}^{N}(\tilde{\lambda}_n-\lambda_n)y_n - \epsilon\sum_{n=1}^{N}(\tilde{\lambda}_n+\lambda_n) \\ &+ \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}(\tilde{\lambda}_n-\lambda_n)(\tilde{\lambda}_m-\lambda_m)\boldsymbol{x}_n^T\boldsymbol{x}_m. \end{aligned}$$

11.11. Derive the dual cost function for the separable class SVM formulation.

Solution: The Lagrangian is given by

$$L(\boldsymbol{\theta},\theta_0,\boldsymbol{\lambda}) = \frac{1}{2}\|\boldsymbol{\theta}\|^2 - \sum_{n=1}^{N}\lambda_n\big(y_n(\boldsymbol{\theta}^T\boldsymbol{x}_n+\theta_0)-1\big).$$

From the KKT condition, given in the text, we have

$$\boldsymbol{\theta} = \sum_{n=1}^{N}\lambda_n y_n\boldsymbol{x}_n.$$

Plugging it into the Lagrangian we obtain

$$\begin{aligned} L(\boldsymbol{\lambda}) &= \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\lambda_n\lambda_m y_n y_m \boldsymbol{x}_n^T\boldsymbol{x}_m - \sum_{n=1}^{N}\lambda_n y_n\left(\sum_{m=1}^{N}\lambda_m y_m \boldsymbol{x}_m^T\right)\boldsymbol{x}_n \\ &- \sum_{n=1}^{N}\lambda_n y_n\theta_0 + \sum_{n=1}^{N}\lambda_n. \end{aligned}$$

Taking into account from the text that $\sum_{n=1}^{N}\lambda_n y_n = 0$, we finally get

$$L(\boldsymbol{\lambda}) = \sum_{n=1}^{N}\lambda_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\lambda_n\lambda_m y_n y_m \boldsymbol{x}_n^T\boldsymbol{x}_m.$$

11.12. Derive the kernel approximation in Eq. (11.91)

Solution: The formula comes as a direct manipulation by plugging in the definition Euler's formula.

11.13. Derive the subgradient for the Huber loss function.

Solution: The Huber loss function is given by

$$L(y,z) = \begin{cases} \epsilon|y-z| - \frac{\epsilon^2}{2}, & \text{if} \quad |y-z| > \epsilon, \\ \frac{1}{2}|y-z|^2, & \text{if} \quad |y-z| \le \epsilon, \end{cases}$$

where $z = f(\boldsymbol{x})$. Hence, for $|y-z| > \epsilon$

$$\frac{\partial}{\partial z}L(y,z) = -\epsilon\,\mathrm{sgn}(y-z).$$

For $|y-z| < \epsilon$

$$\frac{\partial}{\partial z}L(y,z) = -(y-z).$$

At the discontinuities either of the two can be a subgradient, and we will choose the latter. Thus,

$$\frac{\partial L(y,z)}{\partial z} = \begin{cases} -\epsilon\,\mathrm{sgn}(y-z), & |y-z| > \epsilon \\ -(y-z), & |y-z| \le \epsilon. \end{cases}$$

Solutions To Problems of Chapter 12

12.1. Show that if

$$p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_z, \Sigma_z),$$

and

$$p(\boldsymbol{t}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{t}|A\boldsymbol{z}, \Sigma_{t|z}),$$

then

$$\mathbb{E}[\mathbf{z}|\boldsymbol{t}] = (\Sigma_z^{-1} + A^T \Sigma_{t|z}^{-1} A)^{-1}(A^T \Sigma_{t|z}^{-1}\boldsymbol{t} + \Sigma_z^{-1}\boldsymbol{\mu}_z)$$

Solution: We have shown in the Appendix of the chapter that,

$$\mathbb{E}[\mathbf{z}|\boldsymbol{t}] := \boldsymbol{\mu}_{z|t} = \boldsymbol{\mu}_z + (\Sigma_z^{-1} + A^T \Sigma_{t|z}^{-1} A)^{-1} A^T \Sigma_{t|z}^{-1}(\boldsymbol{t} - A\boldsymbol{\mu}_z) \tag{1}$$

or

$$\begin{aligned}\boldsymbol{\mu}_{z|t} =& (\Sigma_z^{-1} + A^T \Sigma_{t|z}^{-1} A)^{-1} A^T \Sigma_{t|z}^{-1}\boldsymbol{t} + \\ & \big(I - (\Sigma_z^{-1} + A^T \Sigma_{t|z}^{-1} A)^{-1} A^T \Sigma_{t|z}^{-1} A\big)\boldsymbol{\mu}_z.\end{aligned} \tag{2}$$

Recall the matrix inversion identity

$$(P^{-1} + B^T R^{-1} B)^{-1} B^T R^{-1} = P B^T (R + BPB^T)^{-1}. \tag{3}$$

Then the following is true

$$(\Sigma_z^{-1} + A^T \Sigma_{t|z}^{-1} A)^{-1} A^T \Sigma_{t|z}^{-1} = \Sigma_z A^T (\Sigma_{t|z} + A\Sigma_z A^T)^{-1}. \tag{4}$$

Hence

$$\begin{aligned}& I - (\Sigma_z^{-1} + A^T \Sigma_{t|z}^{-1} A)^{-1} A^T \Sigma_{t|z}^{-1} A = \\ & I - \Sigma_z A^T (\Sigma_{t|z} + A\Sigma_z A^T)^{-1} A = \\ & (\Sigma_z - \Sigma_z A^T (\Sigma_{t|z} + A\Sigma_z A^T)^{-1} A \Sigma_z)\Sigma_z^{-1}\end{aligned}$$

Recall now the other matrix inversion identity

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

Then we obtain

$$\begin{aligned}& (\Sigma_z - \Sigma_z A^T (\Sigma_{t|z} + A\Sigma_z A^T)^{-1} A \Sigma_z)\Sigma_z^{-1} = \\ & \quad (\Sigma_z^{-1} + A^T \Sigma_{t|z}^{-1} A)^{-1} \Sigma_z^{-1}\end{aligned} \tag{5}$$

Combining (2) and (5) proves the claim.

12.2. Let $\mathbf{x} \in \mathbb{R}^l$ be a random vector following the normal $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma)$. Consider $\boldsymbol{x}_n$, $n = 1, 2, \ldots, N$, to be i.i.d. observations. If the prior for $\boldsymbol{\mu}$ follows $\mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu}_0, \Sigma_0)$, show that the posterior $p(\boldsymbol{\mu}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ is normal $\mathcal{N}(\boldsymbol{\mu}|\tilde{\boldsymbol{\mu}}, \tilde{\Sigma})$ with

$$\tilde{\Sigma}^{-1} = \Sigma_0^{-1} + N\Sigma^{-1},$$

and

$$\tilde{\mu} = \tilde{\Sigma}(\Sigma_0^{-1}\boldsymbol{\mu}_0 + N\Sigma^{-1}\bar{\boldsymbol{x}})$$

where $\bar{\boldsymbol{x}} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n$.

Solution: We have that

$$p(\boldsymbol{\mu}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \propto p(\boldsymbol{\mu}|\boldsymbol{\mu}_0, Q_0^{-1}) \prod_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{\mu}; Q^{-1}), \tag{6}$$

where $Q_0 := \Sigma_0^{-1}$ and $Q := \Sigma^{-1}$. In turn, (6) is written as

$$p(\boldsymbol{\mu}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \propto \frac{|Q|^{N/2}}{(2\pi)^{\frac{Nl}{2}}} \exp\left(-\frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu})^T Q(\boldsymbol{x}_n - \boldsymbol{\mu})\right) \times \\ \frac{|Q_0|^{1/2}}{(2\pi)^{\frac{l}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^T Q_0(\boldsymbol{\mu} - \boldsymbol{\mu}_0)\right).$$

Keeping only the terms that depend on $\boldsymbol{\mu}$, we get

$$p(\boldsymbol{\mu}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \propto \exp\left(-\frac{1}{2}\boldsymbol{\mu}^T Q_0 \boldsymbol{\mu} + \boldsymbol{\mu}^T Q_0 \boldsymbol{\mu}_0\right) \times \\ \exp\left(-\frac{1}{2}\boldsymbol{\mu}^T (NQ)\boldsymbol{\mu} + \boldsymbol{\mu}^T (NQ)\bar{\boldsymbol{x}}\right),$$

or

$$p(\boldsymbol{\mu}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \propto \exp\Big(-\frac{1}{2}\boldsymbol{\mu}^T (Q_0 + NQ)\boldsymbol{\mu} \\ + \boldsymbol{\mu}^T (Q_0\boldsymbol{\mu}_0 + NQ\bar{\boldsymbol{x}})\Big).$$

Note that the exponent in the above is of a quadratic form, i.e., $\boldsymbol{\mu}^T \tilde{Q}\boldsymbol{\mu} + \boldsymbol{\mu}^T \tilde{\boldsymbol{p}}$, which proves the claim.

In words, the posterior mean is a weighted average (by the respective precision matrices) of the prior mean and the sample mean. Moreover the posterior precision matrix is the sum of the prior precision and the conditional precision matrix (weighted by N). Observe that as $N \longrightarrow \infty$ then

$$\tilde{\Sigma} \longrightarrow \frac{1}{N}\Sigma, \quad \tilde{\boldsymbol{\mu}} \longrightarrow \bar{\boldsymbol{x}}.$$

12.3. If $\mathcal{X}$ is the set of observed variables and $\mathcal{X}^l$ the set of the corresponding latent ones, show that

$$\frac{\partial \ln p(\mathcal{X};\boldsymbol{\xi})}{\partial \boldsymbol{\xi}} = \mathbb{E}\left[\frac{\partial \ln p(\mathcal{X},\mathcal{X}^l;\boldsymbol{\xi})}{\partial \boldsymbol{\xi}}\right],$$

where $\mathbb{E}[\cdot]$ is with respect to $p(\mathcal{X}^l|\mathcal{X};\boldsymbol{\xi})$ and $\boldsymbol{\xi}$ is an unknown vector parameter. Note that if one fixes the value of $\boldsymbol{\xi}$ in $p(\mathcal{X}^l|\mathcal{X};\boldsymbol{\xi})$, then one has obtained the M-step of the EM algorithm.

Solution: we have that

$$\begin{aligned}
\frac{\partial \ln p(\mathcal{X};\boldsymbol{\xi})}{\partial \boldsymbol{\xi}} &= \frac{\partial \ln \int_{-\infty}^{+\infty} p(\mathcal{X},\mathcal{X}^l;\boldsymbol{\xi})d\mathcal{X}^l}{\partial \boldsymbol{\xi}} \\
&= \frac{1}{\int_{-\infty}^{+\infty} p(\mathcal{X},\mathcal{X}^l;\boldsymbol{\xi})d\mathcal{X}^l}\frac{\partial \int_{-\infty}^{+\infty} p(\mathcal{X},\mathcal{X}^l;\boldsymbol{\xi})d\mathcal{X}^l}{\partial \boldsymbol{\xi}} \\
&= \int_{-\infty}^{+\infty} \frac{1}{p(\mathcal{X};\boldsymbol{\xi})}\frac{\partial p(\mathcal{X},\mathcal{X}^l;\boldsymbol{\xi})}{\partial \boldsymbol{\xi}} d\mathcal{X}^l \\
&= \int_{-\infty}^{+\infty} \frac{p(\mathcal{X},\mathcal{X}^l;\boldsymbol{\xi})}{p(\mathcal{X};\boldsymbol{\xi})}\frac{\partial \ln p(\mathcal{X},\mathcal{X}^l;\boldsymbol{\xi})}{\partial \boldsymbol{\xi}} d\mathcal{X}^l \\
&= \int_{-\infty}^{+\infty} p(\mathcal{X}^l|\mathcal{X};\boldsymbol{\xi})\frac{\partial \ln p(\mathcal{X},\mathcal{X}^l;\boldsymbol{\xi})}{\partial \boldsymbol{\xi}} d\mathcal{X}^l \\
&= \mathbb{E}\left[\frac{\partial \ln p(\mathcal{X},\mathcal{X}^l;\boldsymbol{\xi})}{\partial \boldsymbol{\xi}}\right]. \qquad (7)
\end{aligned}$$

12.4. Show equation (12.42).

Solution: By the definition of Eq. (12.40), in case the hyperparameters vector is considered to be random, we have,

$$\mathcal{Q}(\boldsymbol{\xi},\boldsymbol{\xi}^{(j)}) = \mathbb{E}\left[\ln p\left(\mathcal{X},\mathcal{X}^l,\boldsymbol{\xi}\right)\right]$$

which by the product rule of probabilities is written as

$$\mathcal{Q}(\boldsymbol{\xi},\boldsymbol{\xi}^{(j)}) = \mathbb{E}\left[\ln\left(p\left(\mathcal{X},\mathcal{X}^l|\boldsymbol{\xi}\right)p(\boldsymbol{\xi})\right)\right]$$

or

$$\mathcal{Q}(\boldsymbol{\xi},\boldsymbol{\xi}^{(j)}) = \mathbb{E}\left[\ln p\left(\mathcal{X},\mathcal{X}^l|\boldsymbol{\xi}\right)\right] + \ln p(\boldsymbol{\xi})$$

which gives the (12.42).

12.5. Let $\boldsymbol{y} \in \mathbb{R}^N$, $\boldsymbol{\theta} \in \mathbb{R}^l$ and Φ a matrix of appropriate dimensions. Derive the expected value of $\|\boldsymbol{y} - \Phi\boldsymbol{\theta}\|^2$ with respect to $\boldsymbol{\theta}$, given $\mathbb{E}[\boldsymbol{\theta}]$ and the corresponding covariance matrix Σ_θ.

Solution: Let $\boldsymbol{\phi} = \boldsymbol{y} - \Phi\boldsymbol{\theta}$. By definition we have

$$\begin{aligned} \Sigma_\phi &= \mathbb{E}[(\boldsymbol{\phi} - \mathbb{E}[\boldsymbol{\phi}])(\boldsymbol{\phi} - \mathbb{E}[\boldsymbol{up\phi}])^T] \\ &= \mathbb{E}[\boldsymbol{\phi}\boldsymbol{\phi}^T] - \mathbb{E}[\boldsymbol{\phi}]\,\mathbb{E}[\boldsymbol{\phi}^T], \end{aligned} \tag{8}$$

where

$$\boldsymbol{\mu}_\phi := \mathbb{E}[\boldsymbol{\phi}] = \boldsymbol{y} - \Phi\boldsymbol{\mu}_{\theta}. \tag{9}$$

Elaborating on (8) we get

$$\begin{aligned} \Sigma_\phi &= \mathbb{E}[(\boldsymbol{y} - \Phi\boldsymbol{\theta})(\boldsymbol{y} - \Phi\boldsymbol{\theta})^T] - (\boldsymbol{y} - \Phi\boldsymbol{\mu}_\theta)(\boldsymbol{y} - \Phi\boldsymbol{\mu}_\theta)^T \\ &= \boldsymbol{y}\boldsymbol{y}^T - \Phi\boldsymbol{\mu}_\theta\boldsymbol{y}^T - \boldsymbol{y}\boldsymbol{\mu}_\theta^T\Phi^T + \Phi\,\mathbb{E}[\boldsymbol{\theta}\boldsymbol{\theta}^T]\Phi^T \\ &\quad - \boldsymbol{y}\boldsymbol{y}^T + \Phi\boldsymbol{\mu}_\theta\boldsymbol{y}^T + \boldsymbol{y}\boldsymbol{\mu}_\theta^T\Phi^T - \Phi\boldsymbol{\mu}_\theta\boldsymbol{\mu}_\theta^T\Phi^T \\ &= \Phi\,\mathbb{E}[(\boldsymbol{\theta} - \boldsymbol{\mu}_\theta)(\boldsymbol{\theta} - \boldsymbol{\mu}_\theta)^T]\Phi^T \\ &= \Phi\Sigma_\theta\Phi^T. \end{aligned} \tag{10}$$

Hence,

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

12.6. Derive recursions (12.60)-(12.62).

Solution: Recall from Eq. (12.59) of the text that

$$\begin{aligned} \mathcal{Q}(\boldsymbol{\Xi}, \boldsymbol{P}; \boldsymbol{\Xi}^{(j)}, \boldsymbol{P}^{(j)}) &= \sum_{n=1}^{N} \mathbb{E}\left[\ln\Big(p(\boldsymbol{x}_n|k_n; \boldsymbol{\xi}_{k_n})P_{k_n}\Big)\right] \\ &:= \sum_{n=1}^{N}\sum_{k=1}^{K} P(k|\boldsymbol{x}_n; \boldsymbol{\Xi}^{(j)}, \boldsymbol{P}^{(j)})\Big(\ln P_k - \frac{l}{2}\ln\sigma_k^2 \\ &\quad -\frac{1}{2\sigma_k^2}\|\boldsymbol{x}_n - \boldsymbol{\mu}_k\|^2\Big) + C. \end{aligned} \tag{11}$$

- Iteration for the mean value: The above can be rewritten as

  $$\mathcal{Q}(\boldsymbol{\Xi}, \boldsymbol{P}; \boldsymbol{\Xi}^{(j)}, \boldsymbol{P}^{(j)} = -\frac{1}{2\sigma_k^2}\sum_{n=1}^{N}\sum_{k=1}^{K} P(k|\boldsymbol{x}_n; \boldsymbol{\Xi}^{(j)}, \boldsymbol{P}^{(j)})||\boldsymbol{x}_n - \boldsymbol{\mu}_k||^2 + \text{constant}$$

  where constant included all terms which are independent of $\boldsymbol{\mu}_k$. Taking the derivative with respect to a specific k, we obtain

  $$\sum_{n=1}^{N}\gamma_{kn}\boldsymbol{\mu}_k = \sum_{n=1}^{N}\gamma_{kn}\boldsymbol{x}_n,$$

which leads to the recursion.

- Recursion for the variance: Eq. (12.59) is now written as

$$\mathcal{Q}(\boldsymbol{\Xi}, \boldsymbol{P}; \boldsymbol{\Xi}^{(j)}, \boldsymbol{P}^{(j)} = -\sum_{n=1}^{N}\sum_{k=1}^{K} P(k|\boldsymbol{x}_n; \boldsymbol{\Xi}^{(j)}, \boldsymbol{P}^{(j)}) \left(l \ln \sigma_k^2 + \frac{1}{2\sigma_k^2}||\boldsymbol{x}_n - \boldsymbol{\mu}_k||^2 \right) + \text{constant}.$$

Taking the derivative with respect to σ_k^2, for a specific values k, obtain

$$\sum_{n=1}^{N} \gamma_{kn} \frac{l}{\sigma_k^2} = \sum_{n=1}^{N} \gamma_{kn} \frac{1}{\sigma_k^4} ||\boldsymbol{x}_n - \boldsymbol{\mu}_k||^2,$$

which gives the recursion.

- Recursion for the probabilities: Eq. (12.59) is now written as

$$\mathcal{Q}(\boldsymbol{\Xi}, \boldsymbol{P}; \boldsymbol{\Xi}^{(j)}, \boldsymbol{P}^{(j)} = \sum_{n=1}^{N}\sum_{k=1}^{K} P(k|\boldsymbol{x}_n; \boldsymbol{\Xi}^{(j)}, \boldsymbol{P}^{(j)}) \ln P_k + \text{constant}.$$

Also since we refer to probabilities, they have to sum up to one,

$$\sum_{k-1}^{K} P_k = 1.$$

Thus, we have an constrained optimization task. Using Lagrange multipliers, the following Lagrangian results

$$L(\lambda, P_k) = \sum_{n=1}^{N}\sum_{k=1}^{K} P(k|\boldsymbol{x}_n; \boldsymbol{\Xi}^{(j)}, \boldsymbol{P}^{(j)}) \ln P_k + \lambda \left(\sum_{k=1}^{K} P_k - 1 \right).$$

Taking the derivative of the Lagrangian with respect to P_k, for a specific value of k, we get

$$P_k = \frac{1}{\lambda} \sum_{n=1}^{N} \gamma_{kn}.$$

Substituting in the constraint equation, it turns out that $\lambda = N$ and finally the recursion results.

12.7. Show that the Kullback-Leibler divergence KL($p \parallel q$) is a nonnegative quantity.
Hint: Recall that $\ln(\cdot)$ is a concave function and use Jensen's inequality, that is,

$$f\left(\int g(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} \right) \leq \int f(g(\boldsymbol{x})) p(\boldsymbol{x}) d\boldsymbol{x},$$

where $p(\boldsymbol{x})$ is a pdf and f is a convex function

Solution: By definition of the KL$(q \parallel p)$ divergence and the fact that $-\ln(\cdot)$ is a convex function, we have

$$\begin{aligned}\mathrm{KL}(q \parallel p) &= -\int p(\boldsymbol{x}) \ln \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})} d\boldsymbol{x} \\ &\geq -\ln \int \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})} p(\boldsymbol{x}) d\boldsymbol{x} = -\ln \int q(\boldsymbol{x}) d\boldsymbol{x} = 0.\end{aligned}$$

12.8. Prove that the binomial and beta distributions are conjugate pairs with respect to the mean value.

Solution: The binomial distribution

$$P(x|\mu) = \left(\begin{array}{c} N \\ x \end{array} \right) \mu^x (1-\mu)^{N-x},$$

is a special case of the multinomial one, when only two events are possible. On the other hand, the beta distribution

$$p(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1} (1-\mu)^{b-1}$$

is a special case of the Dirichlet distribution when only two events are possible. Hence, since the multinomial and Dirichlet are conjugate pairs, the binomial and beta are conjugate pairs, too.

12.9. Show that the normalizing constant C in the Dirichlet pdf

$$\mathrm{Dir}(\boldsymbol{x}|\boldsymbol{a}) = C \prod_{k=1}^{K} x_k^{a_k - 1}, \ \sum_{k=1}^{K} x_k = 1,$$

is given by

$$C = \frac{\Gamma(a_1 + a_2 + \ldots + a_K)}{\Gamma(a_1)\Gamma(a_2)\ldots\Gamma(a_K)}$$

Hint: Use the property $\Gamma(a+1) = a\Gamma(a)$.
a) Use induction. Since the proposition is true for $k = 2$ (beta distribution), assume that it is true for $k = K-1$, and prove that it will be true for $k = K$.
b) Note that due to the constraint $\sum_{k=1}^{K} x_k = 1$, only $K-1$ of the variables are independent. So, basically the Dirichlet pdf implies that

$$p(x_1, x_2, \ldots, x_{K-1}) = C \prod_{k=1}^{K-1} x_k^{a_k - 1} \left(1 - \sum_{k=1}^{K-1} x_k\right)^{a_K - 1}.$$

Solution: We will integrate x_{K-1} out. Then we get

$$p(x_1, x_2, \ldots, x_{K-2}) = \\ C \prod_{k=1}^{K-2} x_k^{a_k - 1} \int_0^{1-\sum_{k=1}^{K-2} x_k} x_{K-1}^{a_{K-1}-1} \left(1 - \sum_{k=1}^{K-1} x_k\right)^{a_K - 1} dx_{K-1}. \tag{12}$$

The upper limit in the integration takes into account that x_k are probabilities (sum to one); hence, given the values $x_1, \ldots, x_{K-2}$, the limits in which x_{K-1} can lie range from zero to $1 - \sum_{k=1}^{K-2} x_k$. Set

$$x_{K-1} = t \left(1 - \sum_{k=1}^{K-2} x_k\right). \tag{13}$$

Hence,

$$dx_{K-1} = \left(1 - \sum_{k=1}^{K-2} x_k\right) dt, \tag{14}$$

and

$$1 - \sum_{k=1}^{K-1} x_k = x_{K-1} \frac{1-t}{t} = (1-t)\left(1 - \sum_{k=1}^{K-2} x_k\right). \tag{15}$$

From (12)-(15), we obtain

$$p(x_1, x_2, \ldots, x_{K-2}) = C \prod_{k=1}^{K-2} x_k^{a_k - 1} \left(1 - \sum_{k=1}^{K-2} x_k\right)^{(a_{K-1} + a_K - 1)} \\ \int_0^1 t^{a_{K-1}-1} (1-t)^{a_k - 1} dt. \tag{16}$$

However, recalling the result from beta pdf in Problem 12.8 the integral is equal to $\frac{\Gamma(a_{K-1})\Gamma(a_K)}{\Gamma(a_{K-1}+a_K)}$. Hence, (16) can be rewritten as

$$p(x_1, x_2, \ldots, x_{K-2}) = C \prod_{k=1}^{K-2} x_k^{a_k - 1} \left(1 - \sum_{k=1}^{K-2} x_k\right)^{(a_{K-1} + a_K - 1)} \\ \frac{\Gamma(a_{K-1})\Gamma(a_K)}{\Gamma(a_{K-1} + a_K)}. \tag{17}$$

Looking at (17) it is readily seen that $p(x_1, x_2, \ldots, x_{K-2})$ is a Dirichlet pdf, hence by assumption

$$C \frac{\Gamma(a_{K-1})\Gamma(a_K)}{\Gamma(a_{K-1} + a_K)} = \frac{\Gamma(a_1 + a_2 + \ldots + a_{K-2} + a_{K-1} + a_K)}{\Gamma(\alpha_1) \ldots \Gamma(\alpha_{K-2})\Gamma(a_{K-1} + a_K)},$$

or

$$C = \frac{\Gamma(a_1 + a_2 + \ldots + a_K)}{\Gamma(a_1) \ldots \Gamma(a_K)}.$$

12.10. Show that $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma)$ for known Σ is of an exponential form and that its conjugate prior is also Gaussian.

Solution:

$$\begin{aligned} p(\boldsymbol{x}|\boldsymbol{\mu}) &= \frac{1}{(2\pi)^{l/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \\ &= \frac{1}{(2\pi)^{l/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}\boldsymbol{x}^T\Sigma^{-1}\boldsymbol{x}\right) \exp\left(-\frac{1}{2}\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu}\right) \times \\ &\quad \exp(\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{x}). \end{aligned}$$

Hence

$$\begin{aligned} \boldsymbol{\phi}(\boldsymbol{\mu}) &= \boldsymbol{\mu} \\ \boldsymbol{u}(\boldsymbol{x}) &= \Sigma^{-1}\boldsymbol{x} \\ f(\boldsymbol{x}) &= \exp\left(-\frac{1}{2}\boldsymbol{x}^T\Sigma^{-1}\boldsymbol{x}\right) \frac{1}{(2\pi)^{l/2}|\Sigma|^{1/2}}, \end{aligned}$$

and from the respective definition and some trivial algebraic manipulation

$$g(\boldsymbol{\mu}) = \exp\left(-\frac{1}{2}\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu}\right).$$

For the conjugate prior, we have

$$\begin{aligned} p(\boldsymbol{\mu}|\lambda, \boldsymbol{v}) &= \exp\left(-\frac{\lambda}{2}\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu}\right) h(\lambda, \boldsymbol{v}) \exp(\boldsymbol{\mu}^T\boldsymbol{v}) \\ &= h(\lambda, \boldsymbol{v}) \exp\left(-\frac{1}{2}\boldsymbol{\mu}^T\left(\frac{1}{\lambda}\Sigma\right)^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^T\boldsymbol{v}\right). \end{aligned}$$

This is of an exponential form and quadratic with respect to $\boldsymbol{\mu}$, hence Gaussian.

12.11. Show that the conjugate prior of the multivariate Gaussian with respect to the precision matrix, Q, is a Wishart distribution.

Solution: We have that,

$$\begin{aligned} p(\boldsymbol{x}|Q) &= \frac{|Q|^{1/2}}{(2\pi)^{l/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T Q(\boldsymbol{x}-\boldsymbol{\mu})\right) \\ &= \frac{|Q|^{1/2}}{(2\pi)^{l/2}} \exp\left(-\frac{1}{2}\text{Trace}\{Q(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T\}\right), \end{aligned} \tag{18}$$

where we used the identity from matrix theory trace$\{AB\}$ = trace$\{BA\}$. We will now bring (18) into the standard form of the exponential family,

$$p(\boldsymbol{x}|Q) = \frac{|Q|^{1/2}}{(2\pi)^{l/2}} \exp\left(-\frac{1}{2}[\boldsymbol{q}_1^T, \boldsymbol{q}_2^T, \dots, \boldsymbol{q}_l^T]\boldsymbol{u}(\boldsymbol{x})\right),$$

where $\boldsymbol{q}_i^T,\ i = 1, 2, \ldots, l$, are the row vectors of Q, i.e.,

$$Q := \begin{bmatrix} \boldsymbol{q}_1^T \\ \boldsymbol{q}_2^T \\ \vdots \\ \boldsymbol{q}_l^T \end{bmatrix}$$

and $\boldsymbol{u}(\boldsymbol{x})$ comprises the columns of $(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^T$, stacked in sequence one below the other. Hence $p(\boldsymbol{x}|Q)$ is of an exponential form. Its conjugate prior will be given by

$$p(Q; \lambda, \boldsymbol{v}) = h(\lambda, \boldsymbol{v})|Q|^{\frac{\lambda}{2}} \exp\left(\frac{1}{2}[\boldsymbol{q}_1^T, \boldsymbol{q}_2^T, \ldots, \boldsymbol{q}_l^T]\boldsymbol{v}\right). \tag{19}$$

Eq. (19) can be brought in a slightly different form, which will facilitate the computation of $h(\lambda, \boldsymbol{v})$. Consider $\boldsymbol{v}$ as comprising the column vectors of a matrix W^{-1}, and set $\lambda = \nu - l - 1$ then (19) is written as

$$p(Q; \nu, W) = h(\nu, W)|Q|^{\frac{\nu-l-1}{2}} \exp\left(-\frac{1}{2}\text{trace}\{W^{-1}Q\}\right),$$

which is a Wishart distribution and the normalizing constant then necessarily becomes

$$h(\nu, W) = |W|^{-\frac{\nu}{2}} \left(2^{\frac{\nu l}{2}} \pi^{\frac{l(l-1)}{4}} \prod_{j=1}^{l} \Gamma\left(\frac{\nu + 1 - j}{2}\right)\right)^{-1}.$$

12.12. Show that the conjugate prior of the univariate Gaussian $\mathcal{N}(x|\mu, \sigma^2)$ with respect to the mean and the precision $\beta = \frac{1}{\sigma^2}$, is the Gaussian-gamma product

$$p(\mu, \beta; \lambda, \boldsymbol{v}) = \mathcal{N}\left(\mu\middle|\frac{v_2}{\lambda}, (\lambda\beta)^{-1}\right) \text{Gamma}\left(\beta\middle|\frac{\lambda+1}{2}, \frac{v_1}{2} - \frac{v_2^2}{2\lambda}\right)$$

where $\boldsymbol{v} := [v_1, v_2]^T$.

Solution: We have shown in the respective section in the book that

$$p(\mu, \beta; \lambda, \boldsymbol{v}) = h(\lambda, \boldsymbol{v})\beta^{\frac{\lambda}{2}} \exp\left(\frac{-\lambda\beta\mu^2}{2}\right) \exp\left([-\frac{\beta}{2}, \beta\mu]\begin{bmatrix} v_1 \\ v_2 \end{bmatrix}\right),$$

which leads to

$$\begin{aligned} p(\mu, \beta; \lambda, \boldsymbol{v}) &= h(\lambda, \boldsymbol{v}) \exp\left(-\frac{\lambda\beta}{2}\mu^2 + v_2\beta\mu - \frac{v_2^2\beta}{2\lambda} + \frac{v_2^2\beta}{2\lambda}\right) \times \\ &\quad \beta^{\frac{\lambda}{2}} \exp(-\frac{v_1}{2}\beta) \\ &= h(\lambda, \boldsymbol{v}) \exp\left(-\frac{\lambda\beta}{2}\left(\mu - \frac{v_2}{\lambda}\right)^2\right) \beta^{\frac{\lambda}{2}} \exp\left(-\beta\left(\frac{v_1}{2} - \frac{v_2^2}{2\lambda}\right)\right) \\ &= h(\lambda, \boldsymbol{v})\beta^{\frac{1}{2}} \exp\left(-\frac{\lambda\beta}{2}\left(\mu - \frac{v_2}{\lambda}\right)^2\right) \beta^{\frac{\lambda-1}{2}} \exp\left(-\beta\left(\frac{v_1}{2} - \frac{v_2^2}{2\lambda}\right)\right). \end{aligned}$$

The last formula is of the form $p(\mu, \beta) = p(\mu|\beta)p(\beta)$, which readily suggests that $h(\lambda, \boldsymbol{v})$ should be such that to lead to

$$p(\mu, \beta; \lambda, \boldsymbol{v}) = \mathcal{N}\left(\mu \middle| \frac{v_2}{\lambda}, \frac{\beta^{-1}}{\lambda}\right) \text{Gamma}\left(\beta \middle| \frac{\lambda+1}{2}, \frac{v_1}{2} - \frac{v_2^2}{2\lambda}\right).$$

12.13. Show that the multivariate Gaussian $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, Q^{-1})$ has as a conjugate prior, with respect to the mean and the precision matrix, Q, the Gaussian-Wishart product.

Solution: Let

$$\begin{aligned} p(\boldsymbol{x}|\boldsymbol{\mu}, Q) &= \frac{|Q|^{1/2}}{(2\pi)^{l/2}} \exp\left(-\frac{1}{2}\boldsymbol{x}^T Q \boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}^T Q \boldsymbol{\mu} + \boldsymbol{\mu}^T Q \boldsymbol{x}\right) \\ &= \frac{|Q|^{1/2}}{(2\pi)^{l/2}} \exp\left(-\frac{1}{2}\boldsymbol{\mu}^T Q \boldsymbol{\mu}\right) \exp\left(-\frac{1}{2}\boldsymbol{x}^T Q \boldsymbol{x} + \boldsymbol{\mu}^T Q \boldsymbol{x}\right). \end{aligned}$$

Let also

$$p(\boldsymbol{\mu}, Q; \boldsymbol{\mu}_0, \lambda, W, \nu) = \mathcal{N}\left(\boldsymbol{\mu}|\boldsymbol{\mu}_0, (\lambda Q)^{-1}\right) \mathcal{W}\left(Q|W, \nu\right).$$

Then, the posterior is of the form

$$\begin{aligned} &p(\boldsymbol{\mu}, Q|\boldsymbol{x}) \propto \frac{|Q|^{1/2}}{(2\pi)^{l/2}} \exp\left(-\frac{1}{2}\boldsymbol{\mu}^T Q \boldsymbol{\mu}\right) \exp\left(-\frac{1}{2}\boldsymbol{x}^T Q \boldsymbol{x} + \boldsymbol{\mu}^T Q \boldsymbol{x}\right) \times \\ &\frac{|\lambda Q|^{1/2}}{(2\pi)^{l/2}} \exp\left(-\frac{1}{2}\boldsymbol{\mu}^T (\lambda Q) \boldsymbol{\mu} - \frac{1}{2}\boldsymbol{\mu}_0^T (\lambda Q) \boldsymbol{\mu}_0 + \boldsymbol{\mu}^T (\lambda Q) \boldsymbol{\mu}_0\right) \times \\ &h|Q|^{\frac{\nu-\lambda-1}{2}} \exp\left(-\frac{1}{2}\text{trace}\{W^{-1}Q\}\right), \end{aligned}$$

which after some trivial manipulations becomes

$$\begin{aligned} p(\boldsymbol{\mu}, Q|\boldsymbol{x}) \propto \exp&\left(-\frac{1}{2}\boldsymbol{\mu}^T (\lambda+1) Q \boldsymbol{\mu} + \boldsymbol{\mu}^T (Q\boldsymbol{x} + \lambda Q \boldsymbol{\mu}_0)\right) \times \\ &\exp\Big(-\frac{1}{2}\text{trace}\{W^{-1}Q\} - \frac{1}{2}\text{trace}\{\boldsymbol{x}\boldsymbol{x}^T Q\} - \\ &\frac{1}{2}\text{trace}\{\lambda \boldsymbol{\mu}_0 \boldsymbol{\mu}_0^T Q\}\Big) |Q|^{\frac{\nu-l}{2}}, \end{aligned}$$

which is rewritten as

$$\begin{aligned} p(\boldsymbol{\mu}, Q|\boldsymbol{x}) \propto \exp&\left(-\frac{1}{2}(\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}_0)^T (\tilde{\lambda} Q)(\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}_0)\right) \times \\ &|Q|^{\frac{\tilde{\nu}-l-1}{2}} \exp\left(-\frac{1}{2}\text{trace}\{\tilde{W}^{-1}Q\}\right), \end{aligned}$$

where

$$\begin{aligned}\tilde{\nu} &= \nu + 1\\ \tilde{\lambda} &= \lambda + 1\\ \tilde{\boldsymbol{\mu}}_0 &= (\tilde{\lambda}Q)^{-1}(Q\boldsymbol{x} + \lambda Q\boldsymbol{\mu}_0)\\ \tilde{W}^{-1} &= W^{-1} + \boldsymbol{x}\boldsymbol{x}^T + \lambda\boldsymbol{\mu}_0\boldsymbol{\mu}_0^T.\end{aligned}$$

Since $p(\boldsymbol{\mu}, Q|\boldsymbol{x}) = p(\boldsymbol{\mu}|Q, \boldsymbol{x})p(Q|\boldsymbol{x})$ must integrate to one, then the missing constants must be such that

$$p(\boldsymbol{\mu}, Q|\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{\mu}|\tilde{\boldsymbol{\mu}}_0, (\tilde{\lambda}Q)^{-1}\right)\mathcal{W}\left(Q|\tilde{W}, \tilde{\nu}\right).$$

12.14. Show that the distribution

$$P(x|\mu) = \mu^x(1-\mu)^{1-x},\ x \in \{0, 1\},$$

is of an exponential form and derive its conjugate prior with respect to μ.

Solution: We have

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

from which we obtain

$$g(\mu) = 1-\mu,\ \phi(\mu) = \ln\frac{\mu}{1-\mu},\ u(x) = x.$$

Hence, the conjugate prior will be given by

$$\begin{aligned}p(\mu|\lambda, v) &= h(1-\mu)^{\lambda}\exp\left(v\ln\frac{\mu}{(1-\mu)}\right)\\ &= h(1-\mu)^{\lambda}\frac{\mu^v}{(1-\mu)^v} = h(1-\mu)^{\lambda-v}\mu^v,\end{aligned}$$

which is of a beta distribution form

$$\text{Beta}(\mu|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1},$$

with $a = v+1,\ b = \lambda - v + 1$. Hence

$$h = \frac{\Gamma(\lambda+2)}{\Gamma(v+1)\Gamma(\lambda-v+1)}.$$

12.15. Show that estimating an unknown pdf by maximizing the respective entropy, subject to a set of empirical expectations, results in a pdf that belongs to the exponential family.

Solution: The problem is cast as

$$\begin{aligned}\text{maximize w.r.t } p(x) \qquad & \int_{\mathcal{A}_x} p(x)\ln p(x)dx, \\ \text{s.t.} \qquad & \mathbb{E}[\mathrm{u}_i(\mathrm{x})] = \int_{\mathcal{A}_x} p(x)u_i(x)dx, \ i \in \mathcal{I}.\end{aligned}$$

Using Lagrange multipliers, the Lagrangian becomes

$$\begin{aligned} L(p(x),\boldsymbol{\theta}) &= \int_{\mathcal{A}_x} p(x)\ln p(x)dx \\ &\quad -\sum_{i\in\mathcal{I}}\theta_i\left(\int_{\mathcal{A}_x} p(x)u_i(x)dx - \hat{\mu}_i\right) - \theta_0\left(\int_{\mathcal{A}_x} p(x)dx - 1\right).\end{aligned}$$

Taking the variational derivative with respect to $p(x)$ and equating to zero we get

$$\begin{aligned}\frac{\partial L}{\partial p(x)} &= \int_{\mathcal{A}_x} \ln p(x)dx + \int_{\mathcal{A}_x} dx \\ &\quad -\sum_{i\in\mathcal{I}}\theta_i\int_{\mathcal{A}_x} u_i(x)dx - \theta_0\int_{\mathcal{A}_x} dx = \\ &\quad \int_{\mathcal{A}_x}\left(\ln p(x) - \sum_{i\in\mathcal{I}}\theta_i u_i(x) - (\theta_0 - 1)\right)dx = 0,\end{aligned}$$

which results in

$$\ln p(x) = \sum_{i\in\mathcal{I}}\theta_i u_i(x) + (\theta_0 - 1)$$

and finally

$$p(x) = C\exp\left(\sum_{i\in\mathcal{I}}\theta_i u_i(x)\right).$$

Solutions To Problems of Chapter 13

13.1. Show Eq. (13.5).

Solution: The functional $\mathcal{F}(q)$ is defined as

$$\mathcal{F}(q) = \int q(\mathcal{X}^l, \boldsymbol{\theta}) \ln \frac{p(\mathcal{X}, \mathcal{X}^l, \boldsymbol{\theta})}{q(\mathcal{X}^l, \boldsymbol{\theta})} d\mathcal{X}^l d\boldsymbol{\theta}. \tag{1}$$

Plugging in the mean field approximation, $q(\mathcal{X}^l, \boldsymbol{\theta}) = q_{\mathcal{X}^l}(\mathcal{X}^l) q_\theta(\boldsymbol{\theta})$, the previous equation becomes

$$\mathcal{F}(q) = \int q_{\mathcal{X}^l}(\mathcal{X}^l) q_\theta(\boldsymbol{\theta}) \ln \frac{p(\mathcal{X}, \mathcal{X}^l, \boldsymbol{\theta})}{q_{\mathcal{X}^l}(\mathcal{X}^l) q_\theta(\boldsymbol{\theta})} d\mathcal{X}^l d\boldsymbol{\theta}, \tag{2}$$

or

$$\begin{aligned}\mathcal{F}(q) \quad = \quad & \int q_{\mathcal{X}^l}(\mathcal{X}^l) \left(\int q_\theta(\boldsymbol{\theta}) \ln p(\mathcal{X}, \mathcal{X}^l, \boldsymbol{\theta}) d\boldsymbol{\theta} \right) d\mathcal{X}^l \\ & - \int q_{\mathcal{X}^l}(\mathcal{X}^l) \ln q_{\mathcal{X}^l}(\mathcal{X}^l) \left(\int q_\theta(\boldsymbol{\theta}) d\boldsymbol{\theta} \right) d\mathcal{X}^l \\ & - \int q_\theta(\boldsymbol{\theta}) \ln q_\theta(\boldsymbol{\theta}) \left(\int q_{\mathcal{X}^l}(\mathcal{X}^l) d\mathcal{X}^l \right) d\boldsymbol{\theta}, \end{aligned} \tag{3}$$

which then leads trivially to the claim.

13.2. Show equation (13.38).

Solution: From Eq. (13.37) in the text we have

$$\begin{aligned}\ln q_\alpha^{(j+1)}(\boldsymbol{\alpha}) = {} & \mathbb{E}_{q_\theta^{(j+1)}} \left[\frac{1}{2} \sum_{k=0}^{K-1} \ln \alpha_k - \frac{1}{2} \sum_{k=0}^{K-1} \alpha_k \theta_k^2 \right] && (4) \\ & - \sum_{k=0}^{K-1} \ln \Gamma(a) + a \sum_{k=0}^{K-1} \ln b + (a-1) \sum_{k=0}^{K-1} \ln \alpha_k && (5) \\ & - b \sum_{k=0}^{K-1} \alpha_k + \text{constant} && (6) \\ = {} & \left(a - 1 + \frac{1}{2} \right) \sum_{k=0}^{K-1} \ln \alpha_k - \sum_{k=0}^{K-1} \left(b + \frac{1}{2} \mathbb{E}_{q_\theta^{(j+1)}}[\theta_k^2] \right) \alpha_k && (7) \\ & - K \ln \Gamma(a) + aK \ln b + \text{constant}, && (8)\end{aligned}$$

or

$$q_\alpha^{(j+1)}(\boldsymbol{\alpha}) \propto \prod_{k=0}^{K-1} \alpha_k^{\left(a-1+\frac{1}{2}\right)} \exp\left(- \left(b + \frac{1}{2} \mathbb{E}_{q_\theta^{(j+1)}}[\theta_k^2] \right) \alpha_k \right),$$

or

$$q_\alpha^{(j+1)}(\boldsymbol{\alpha}) = \prod_{k=0}^{K-1} \text{Gamma}\left(\alpha_k; a+\frac{1}{2}, b+\frac{1}{2}E_{q_\theta^{(j+1)}}[\theta_k^2]\right),$$

which proves the claim.

13.3. Show equations (13.43)-(13.45).

Solution: From the text we have

$$\begin{aligned}\ln q_\beta^{(j+1)}(\beta) &= \mathbb{E}_{q_\theta^{(j+1)} q_\alpha^{(j+1)}}[\ln p(\boldsymbol{y}|\boldsymbol{\theta},\beta) + \ln p(\beta)] + \text{constant}\\ &= \mathbb{E}_{q_\theta^{(j+1)}}\left[\frac{N}{2}\ln\beta - \frac{1}{2}\beta\|\boldsymbol{y}-\Phi\boldsymbol{\theta}\|^2\right] - \ln\Gamma(c)\\ &+ c\ln b + (c-1)\ln\beta - d\beta\\ &= \left(c-1+\frac{N}{2}\right)\ln\beta - \left(d+\frac{1}{2}\,\mathbb{E}_{q_\theta^{(j+1)}}\,\|\boldsymbol{y}-\Phi\boldsymbol{\theta}\|^2\right)\beta\\ &+ \text{constant},\end{aligned}$$

which then results to

$$q_\beta^{(j+1)}(\beta) \propto \beta^{\tilde{c}}\exp(-\tilde{d}\beta),$$

with

$$\begin{aligned}\tilde{c} &= c+\frac{N}{2}\\ \tilde{d} &= d+\frac{1}{2}\,\mathbb{E}_{q_\theta^{(j+1)}}[\|\boldsymbol{y}-\Phi\boldsymbol{\theta}\|^2].\end{aligned}$$

Thus,

$$q_\beta^{(j+1)}(\beta) = \frac{1}{\Gamma(\tilde{c})}\tilde{d}^{\tilde{c}}\beta^{(\tilde{c}-1)}\exp(-\tilde{d}\beta).$$

The proportionality constant has to be equal to

$$\frac{\tilde{d}^{\tilde{c}}}{\Gamma(\tilde{c})}$$

in order $q_\beta^{(j+1)}(\beta)$ to integrate to one.

13.4. Show that if

$$p(x) \propto \frac{1}{x},$$

then the random variable $\mathrm{z} = \ln \mathrm{x}$ follows a uniform distribution.

Solution: We know that

$$p_\mathrm{z}(z) \propto p_\mathrm{x}(x(z))\frac{dx}{dz} = \exp(-z)\exp(z) = 1, \qquad (9)$$

where we used the fact that $x = \exp(z)$.

13.5. Derive the lower bound after convergence of the variational Bayesian EM for the linear regression task which is modeled as in Section 13.3.

Solution: The lower bound after convergence to $\tilde{q}_\theta(\boldsymbol{\theta})$, $\tilde{q}_\alpha(\boldsymbol{\alpha})$, $\tilde{q}_\beta(\boldsymbol{\beta})$, which are defined by $\tilde{\boldsymbol{\mu}}_\theta$, $\tilde{\Sigma}_\theta$ for the Gaussian $\tilde{q}_\theta$, by $(\tilde{a}, \tilde{b}_i)$, $i = 1, 2, \dots, l$ for the gamma $\tilde{q}_\alpha(\boldsymbol{\alpha})$ and $\tilde{c}$, $\tilde{d}$ for the gamma $\tilde{q}_\beta(\beta)$, will be

$$\begin{aligned}\mathcal{F}(\tilde{q}_\theta,\tilde{q}_\alpha,\tilde{q}_\beta) &= \mathbb{E}_\theta\left[\mathbb{E}_\alpha\left[\mathbb{E}_\beta\left[\ln p(\boldsymbol{y},\theta,\boldsymbol{\alpha},\beta)\right]\right]\right] \\ &\quad - \mathbb{E}_{\boldsymbol{\theta}}\left[\ln q_\theta(\theta)\right] - \mathbb{E}_{\boldsymbol{\alpha}}\left[\ln q_\alpha(\boldsymbol{\alpha})\right] - \mathbb{E}_\beta\left[\ln q_\beta(\beta)\right].\end{aligned}$$

We have that

$$\ln p(\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{\alpha},\beta) = \ln p(\boldsymbol{y}|\boldsymbol{\theta},\beta) + \ln p(\boldsymbol{\theta}|\boldsymbol{\alpha}) + \ln p(\boldsymbol{\alpha}) + \ln p(\beta),$$

where

(a) $\ln p(\boldsymbol{y}|\boldsymbol{\theta},\beta) = \ln \beta^{\frac{N}{2}} - \frac{N}{2}\ln(2\pi) - \frac{\beta}{2}\|\boldsymbol{y} - \Phi\boldsymbol{\theta}\|^2$.

(b) $\ln p(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{1}{2}\sum_{k=0}^{K-1}\ln\alpha_k - \frac{K}{2}\ln(2\pi) - \frac{1}{2}\sum_{k=0}^{K-1}\alpha_k\theta_k^2$.

(c) $\ln p(\boldsymbol{\alpha}) = -K\ln\Gamma(a) + Ka\ln b + (a-1)\sum_{k=0}^{K-1}\ln\alpha_k - b\sum_{k=0}^{K-1}\alpha_k$.

(d) $\ln p(\beta) = -\ln\Gamma(c) + c\ln d + (c-1)\ln\beta - d\beta$.

Using identities from the Appendix of the chapter and the independence among $\tilde{q}_\theta$, $\tilde{q}_\alpha$, $\tilde{q}_\beta$, we get:

(a)

$$\begin{aligned}A_1 := \mathbb{E}_\theta\,\mathbb{E}_\beta\left[\ln p(\boldsymbol{y}|\theta,\beta)\right] &= \frac{N}{2}\,\mathbb{E}_\beta[\ln\beta] - \frac{N}{2}\ln(2\pi) - \frac{1}{2}\,\mathbb{E}_\beta[\beta]\,\mathbb{E}_\theta[\|\boldsymbol{y} - \Phi\theta\|^2] \\ &= \frac{N}{2}[\psi(\tilde{c}) - \ln\tilde{d}] - \frac{N}{2}\ln(2\pi) - \frac{1}{2}\frac{\tilde{c}}{\tilde{d}}\Big\{\Big\|\boldsymbol{y} - \Phi\tilde{\boldsymbol{\mu}}_\theta\|^2 + \mathrm{trace}\{\Phi\tilde{\Sigma}_\theta\Phi^T\}\Big\},\end{aligned}$$

where the results of Problem 12.13 have been used.

(b)

$$\begin{aligned}A_2 := \mathbb{E}_\theta\,\mathbb{E}_\alpha\left[\ln p(\theta|\boldsymbol{\alpha})\right] &= \frac{1}{2}\sum_{k=0}^{K-1}\mathbb{E}_\alpha[\ln\alpha_k] - \frac{K}{2}\ln(2\pi) - \frac{1}{2}\sum_{k=0}^{K-1}\mathbb{E}_\alpha[\alpha_k]\,\mathbb{E}_\theta[\theta_k^2] \\ &= \frac{1}{2}\sum_{k=0}^{K-1}[\psi(\tilde{a}) - \ln\tilde{b}_k] - \frac{K}{2}\ln(2\pi) - \frac{1}{2}\sum_{k=0}^{K-1}\frac{\tilde{a}}{\tilde{b}_k}\left(\tilde{\Sigma}_\theta + \tilde{\boldsymbol{\mu}}_\theta\tilde{\boldsymbol{\mu}}_\theta^T\right)_{kk}.\end{aligned}$$

(c)

$$\begin{aligned}A_3 := \mathbb{E}_\alpha[\ln p(\boldsymbol{\alpha})] &= -K\ln\Gamma(a) - Ka\ln b - (a-1)\sum_{k=0}^{K-1}\mathbb{E}_\alpha[\ln\alpha_k] - b\sum_{k=0}^{K-1}\mathbb{E}_\alpha[\alpha_k] \\ &= K\ln\Gamma(a) - Ka\ln b - (a-1)\sum_{k=0}^{K-1}[\psi(\tilde{a}) - \ln\tilde{b}_k] \\ &\quad - b\sum_{k=0}^{K-1}\frac{\tilde{a}}{\tilde{b}_k}.\end{aligned}$$

(d)

$$\begin{aligned}A_4 &:= \mathbb{E}_\beta[\ln p(\beta)] = -\ln\Gamma(c) + c\ln d + (c-1)\,\mathbb{E}_\beta[\ln\beta] - d\,\mathbb{E}_\beta[\beta]\\ &= -\ln\Gamma(c) + c\ln d + (c-1)(\psi(\tilde{c}) - \ln\tilde{d}) - d\frac{\tilde{c}}{\tilde{d}}.\end{aligned}$$

In the sequel the respective entropies have to be computed

(a)

$$\begin{aligned}B_1 &:= \mathbb{E}_\theta[\ln\tilde{q}_\theta(\boldsymbol{\theta})] = \mathbb{E}_\theta\Big[-\frac{1}{2}\ln|\tilde{\Sigma}_\theta| - \frac{K}{2}\ln(2\pi) -\\ &\quad \frac{1}{2}(\boldsymbol{\theta}-\tilde{\boldsymbol{\mu}}_\theta)^T\tilde{\Sigma}_\theta^{-1}(\boldsymbol{\theta}-\tilde{\boldsymbol{\mu}}_\theta)\Big]\\ &= -\frac{1}{2}\ln|\tilde{\Sigma}_\theta| - \frac{K}{2}\ln(2\pi) - \frac{1}{2}\,\mathbb{E}_\theta[\boldsymbol{\theta}^T\tilde{\Sigma}_\theta^{-1}\boldsymbol{\theta}] - \frac{1}{2}\tilde{\boldsymbol{\mu}}_\theta^T\tilde{\Sigma}_\theta^{-1}\tilde{\boldsymbol{\mu}}_\theta\\ &\quad + \tilde{\boldsymbol{\mu}}_\theta^T\tilde{\Sigma}_\theta^{-1}\,\mathbb{E}[\boldsymbol{\theta}]\\ &= -\frac{1}{2}\ln|\tilde{\Sigma}_\theta| - \frac{K}{2}\ln(2\pi) + \frac{1}{2}\tilde{\boldsymbol{\mu}}_\theta^T\tilde{\Sigma}_\theta^{-1}\tilde{\boldsymbol{\mu}}_\theta\\ &\quad - \frac{1}{2}\,\mathbb{E}_\theta\Big[\mathrm{trace}\{\tilde{\Sigma}_\theta^{-1}\boldsymbol{\theta}\boldsymbol{\theta}^T\}\Big]\\ &= -\frac{1}{2}\ln|\tilde{\Sigma}_\theta| - \frac{K}{2}\ln(2\pi) + \frac{1}{2}\tilde{\boldsymbol{\mu}}_\theta^T\tilde{\Sigma}_\theta^{-1}\tilde{\boldsymbol{\mu}}_\theta\\ &\quad - \frac{1}{2}\mathrm{trace}\{\tilde{\Sigma}_\theta^{-1}(\tilde{\Sigma}_\theta + \tilde{\boldsymbol{\mu}}_\theta\tilde{\boldsymbol{\mu}}_\theta^T)\}\\ &= -\frac{1}{2}\ln|\tilde{\Sigma}_\theta| - \frac{K}{2}\ln(2\pi) + \frac{1}{2}\tilde{\boldsymbol{\mu}}_\theta^T\tilde{\Sigma}_\theta^{-1}\tilde{\boldsymbol{\mu}}_\theta\\ &\quad - \frac{1}{2}\mathrm{Trace}\{I + \tilde{\Sigma}_\theta^{-1}\tilde{\boldsymbol{\mu}}_\theta\tilde{\boldsymbol{\mu}}_\theta^T\},\end{aligned}$$

where Eq. 12.48 from the text has been used.

(b)

$$\begin{aligned}B_2 &:= \mathbb{E}_\beta[\ln\tilde{q}_\beta(\beta)] = \mathbb{E}_\beta[-\ln\Gamma(\tilde{c}) + \tilde{c}\ln\tilde{d} - \tilde{d}\beta + (\tilde{c}-1)\ln\beta]\\ &= -\ln\Gamma(\tilde{c}) + \tilde{c}\ln\tilde{d} - \tilde{d}\frac{\tilde{c}}{\tilde{d}} + (\tilde{c}-1)(\psi(\tilde{c}) - \ln\tilde{d}).\end{aligned}$$

(c)

$$\begin{aligned}B_3 &:= \mathbb{E}_\alpha[\ln\tilde{q}_\alpha(\boldsymbol{\alpha})] = \mathbb{E}_\alpha[-K\ln\Gamma(\tilde{a}) + \tilde{a}\sum_{k=0}^{K-1}\ln\tilde{b}_k + (\tilde{a}-1)\sum_{k=0}^{K-1}\ln\alpha_k - \sum_{k=0}^{K-1}\tilde{b}_k\alpha_k]\\ &= -K\ln\Gamma(\tilde{a}) + \tilde{a}\sum_{k=0}^{K-1}\ln\tilde{b}_k + (\tilde{a}-1)\sum_{k=0}^{K-1}(\psi(\tilde{a}) - \ln\tilde{b}_k) - \sum_{k=0}^{K-1}\tilde{b}_k\frac{\tilde{a}}{\tilde{b}_k}\\ &= -K\ln\Gamma(\tilde{a}) + \tilde{a}\sum_{k=0}^{K-1}\ln\tilde{b}_k + (\tilde{a}-1)\sum_{k=0}^{K-1}(\psi(\tilde{a}) - \ln\tilde{b}_k) - K\tilde{a},\end{aligned}$$

which concludes the derivation.

13.6. Consider the Gaussian mixture model

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} P_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, Q_k^{-1}),$$

with priors

$$p(\boldsymbol{\mu}_k) = \mathcal{N}(\boldsymbol{\mu}_k|0, \beta^{-1}I), \tag{10}$$

and

$$p(Q_k) = \mathcal{W}(Q_k|\nu_0, W_0).$$

Given the set of observations $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, $\boldsymbol{x} \in \mathbb{R}^l$, derive the respective variational Bayesian EM algorithm, using the mean field approximation for the involved posterior pdfs. Consider P_k, $k = 1, 2, \ldots, K$, as deterministic parameters and optimize the respective lower bound of the evidence with respect to the P_k's.

Solution: Consider

$$q(\mathcal{Z}, \boldsymbol{\mu}_{1:K}, Q_{1:K}) = q(\mathcal{Z})q(\boldsymbol{\mu}_{1:K})q(Q_{1:K}),$$

where the notation has been introduced in Section 13.4. From the theory we have:
Step 1a:

$$\begin{aligned}\ln q_z^{(j+1)}(\mathcal{Z}) &= \mathbb{E}_{q_\mu^{(j)} q_Q^{(j)}}[\ln p(\mathcal{X}, \mathcal{Z}, \boldsymbol{\mu}_{1:K}, \mathrm{Q}_{1:K})] + \text{constant} \\ &= \mathbb{E}_{q_\mu^{(j)} q_Q^{(j)}}[\ln p(\mathcal{X}|\mathcal{Z}, \boldsymbol{\mu}_{1:K}, \mathrm{Q}_{1:K})] + \\ &\ln P(\mathcal{Z}|\boldsymbol{P}^{(j)})] + \text{constants}.\end{aligned}$$

Recall that

$$p(\mathcal{X}|\mathcal{Z}, \boldsymbol{\mu}_{1:K}, Q_{1:K}) = \prod_{n=1}^{N}\prod_{k=1}^{K} \left(\mathcal{N}\left(\boldsymbol{x}_n|\boldsymbol{\mu}_k, Q_k^{-1}\right)\right)^{z_{n_k}},$$

and

$$P(\mathcal{Z}|\boldsymbol{P}) = \prod_{n=1}^{N}\prod_{k=1}^{K} P_k^{z_{n_k}}.$$

Hence, we have that

$$\begin{aligned}\ln q_z^{(j+1)}(\mathcal{Z}) &= \sum_{n=1}^{N}\sum_{k=1}^{K} z_{n_k}\,\mathbb{E}_{q_\mu^{(j)} q_Q^{(j)}}\left[\ln P_k^{(j)} + \frac{1}{2}\ln|\mathrm{Q}_k| - \frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T\mathrm{Q}_k(\boldsymbol{x}_n - \boldsymbol{\mu}_k)\right]\\ &+ \text{constants}\\ &= \sum_{n=1}^{N}\sum_{k=1}^{K} z_{n_k}\Big(\ln P_k^{(j)} + \frac{1}{2}\,\mathbb{E}_{q_Q^{(j)}}[\ln|\mathrm{Q}_k|]\\ &\quad - \frac{1}{2}\text{trace}\Big\{\,\mathbb{E}_{q_Q^{(j)}}[\mathrm{Q}_k](\boldsymbol{x}_n\boldsymbol{x}_n^T - \mathbb{E}_{q_\mu^{(j)}}[\boldsymbol{\mu}_k]\boldsymbol{x}_n^T -\\ &\boldsymbol{x}_n\,\mathbb{E}_{q_\mu^{(j)}}[\boldsymbol{\mu}_k^T] + \mathbb{E}_{q_\mu^{(j)}}[\boldsymbol{\mu}_k\boldsymbol{\mu}_k^T])\Big\}\Big) + \text{constants},\end{aligned}$$

or if we set

$$\begin{aligned}\pi_{n_k} &= P_k^{(j)}\exp\Big(\frac{1}{2}\,\mathbb{E}_{q_Q^{(j)}}[\ln|\mathrm{Q}_k|]\\ &- \frac{1}{2}\text{trace}\Big\{\,\mathbb{E}_{q_Q^{(j)}}[\mathrm{Q}_k](\boldsymbol{x}_n\boldsymbol{x}_n^T - \mathbb{E}_{q_\mu^{(j)}}[\boldsymbol{\mu}_k]\boldsymbol{x}_n^T - \boldsymbol{x}_n\,\mathbb{E}_{q_\mu^{(j)}}[\boldsymbol{\mu}_k^T] + \mathbb{E}_{q_\mu^{(j)}}[\boldsymbol{\mu}_k\boldsymbol{\mu}_k^T])\Big\}\Big),\end{aligned}$$

then we have that

$$q_z^{(j+1)}(\mathcal{Z}) = \prod_{n=1}^{N}\prod_{k=1}^{K}\rho_{n_k}^{z_{n_k}},$$

where we have taken care for the normalization of the respective probabilities, hence

$$\rho_{n_k} = \frac{\pi_{n_k}}{\sum_{k=1}^{K}\pi_{n_k}}.$$

Also note that $\mathbb{E}_{q_z^{(j+1)}}[\mathrm{z}_{n_k}] = \rho_{n_k}$, by the binary nature of z_{n_k}.

Step 1b:

$$\begin{aligned}\ln q_\mu^{(j+1)}(\boldsymbol{\mu}_{1:K}) &= \mathbb{E}_{q_z^{(j+1)} q_Q^{(j)}}[\ln p(\mathcal{X}|\mathcal{Z},\boldsymbol{\mu}_{1:K},\mathcal{Q}_{1:K}) + \ln p(\boldsymbol{\mu}_{1:K})] + \text{constants}\\ &= \sum_{k=1}^{K}\Big(\sum_{n=1}^{N}\mathbb{E}_{q_z^{(j+1)}}[\mathrm{z}_{n_k}]\Big(\frac{1}{2}\,\mathbb{E}_{q_Q^{(j)}}[\ln|\mathrm{Q}_k|] -\\ &\frac{1}{2}\,\mathbb{E}_{q_Q^{(j)}}[\boldsymbol{x}_n^T\mathrm{Q}_k\boldsymbol{x}_n] + \boldsymbol{\mu}_k^T\,\mathbb{E}_{q_Q^{(j)}}[\mathrm{Q}_k]\boldsymbol{x}_n - \frac{1}{2}\boldsymbol{\mu}_k^T\,\mathbb{E}_{q_Q^{(j)}}[\mathrm{Q}_k]\boldsymbol{\mu}_k\Big)\\ &- \frac{1}{2}\beta\boldsymbol{\mu}_k^T\boldsymbol{\mu}_k\Big) + \text{constants},\end{aligned}$$

or

$$\begin{aligned}\ln q_\mu^{(j+1)}(\boldsymbol{\mu}_{1:K}) = \sum_{k=1}^{K}\Big(-\frac{1}{2}\boldsymbol{\mu}_k^T\left(\beta I + \mathbb{E}_{q_Q^{(j)}}[\mathrm{Q}_k]\sum_{n=1}^{N}\rho_{n_k}\right)\boldsymbol{\mu}_k\\ + \boldsymbol{\mu}_k^T\,\mathbb{E}_{q_Q^{(j)}}[\mathrm{Q}_k]\sum_{n=1}^{N}\rho_{n_k}\boldsymbol{x}_n\Big) + \text{constants},\end{aligned}$$

and since the exponent is of quadratic form we get,

$$q_{\mu}^{(j+1)}(\boldsymbol{\mu}_{1:K}) = \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{\mu}_k|\tilde{\boldsymbol{\mu}}_k, \tilde{Q}_k),$$

where

$$\tilde{Q}_k = \beta I + \mathbb{E}_{q_Q^{(j)}}[\mathrm{Q}_k] \sum_{k=1}^{K} \rho_{n_k}$$

and

$$\tilde{\boldsymbol{\mu}}_k = \tilde{Q}_k^{-1}\, \mathbb{E}_{q_Q^{(j)}}[\mathrm{Q}_k] \sum_{k=1}^{K} \rho_{n_k} \boldsymbol{x}_n.$$

Moreover, following (12.48) we have that

$$\mathbb{E}_{q_{\mu}^{(j+1)}}[\boldsymbol{\mu}_k \boldsymbol{\mu}_k^T] = \tilde{\varSigma}_k + \tilde{\boldsymbol{\mu}}_k \tilde{\boldsymbol{\mu}}_k^T.$$

Hence, we have now obtained the recursions for all the statistics that are required by the previous step.

Step 1c:

$$\begin{aligned}
\ln q_Q^{(j+1)}(Q_{1:K}) = {} & \mathbb{E}_{q_z^{(j+1)} q_{\mu}^{(j+1)}} \Big[\ln p(\mathcal{X}|\mathcal{Z}, \boldsymbol{\mu}_{1:K}, Q_{1:K}) + \\
& \ln p(Q_{1:K}) \Big] + \text{constants} \\
= {} & \mathbb{E}_{q_z^{(j+1)} q_{\mu}^{(j+1)}} \Big[\sum_{k=1}^{K} \Big(\sum_{n=1}^{N} \mathrm{z}_{n_k} \Big(\frac{1}{2} \ln|Q_k| - \\
& \frac{1}{2} \text{trace}\Big\{ (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T Q_k \Big\} \Big) + \\
& \frac{\nu_0 - l - 1}{2} \ln|Q_k| - \frac{1}{2} \text{trace}\Big\{ W_0^{-1} Q_k \Big\} \Big) \Big] + \text{constants}
\end{aligned}$$

or

$$\begin{aligned}
\ln q_Q^{(j+1)}(Q_{1:K}) = {} & \sum_{k=1}^{K} \Big(\frac{1}{2} \ln|Q_k| \sum_{n=1}^{N} \rho_{n_k} - \\
& \frac{1}{2} \text{trace}\Big\{ \sum_{n=1}^{N} \rho_{n_k} \Big(\boldsymbol{x}_n \boldsymbol{x}_n^T - \tilde{\boldsymbol{\mu}}_k \boldsymbol{x}_n^T - \boldsymbol{x}_n \tilde{\boldsymbol{\mu}}_k^T \\
& + \mathbb{E}_{q_{\mu}^{(j+1)}}[\boldsymbol{\mu}_k \boldsymbol{\mu}_k^T] \Big) Q_k \Big\} + \\
& \frac{\nu_0 - l - 1}{2} \ln|Q_k| - \frac{1}{2} \text{trace}\{ W_0^{-1} Q_k \} \Big) + \text{constants},
\end{aligned}$$

which results in

$$q_Q^{(j+1)}(Q_{1:K}) = \prod_{k=1}^{K} \mathcal{W}(Q_k|\tilde{\nu}_k, \tilde{W}_k),$$

where

$$\begin{aligned}\tilde{\nu}_k &= \nu_0 + \sum_{n=1}^{N} \rho_{n_k}\\ \tilde{W}_k^{-1} &= \tilde{W}_0^{-1} + \sum_{n=1}^{N} \rho_{n_k} (\boldsymbol{x}_n \boldsymbol{x}_n^T - \tilde{\boldsymbol{\mu}}_k \boldsymbol{x}_n^T - \boldsymbol{x}_n \tilde{\boldsymbol{\mu}}_k^T +\\ &\quad \mathbb{E}_{q_\mu^{(j+1)}}[\boldsymbol{\mu}_k \boldsymbol{\mu}_k^T]).\end{aligned}$$

Since $q_Q^{(j+1)}(Q_{1:K})$ is a product of Wisharts, then t is known from statistics that the respective mean values are given by,

$$\begin{aligned}\mathbb{E}_{q_Q^{(j+1)}}[\mathrm{Q}_k] &= \tilde{\nu}_k \tilde{W}_k\\ \mathbb{E}_{q_Q^{(j+1)}}[\ln |\mathrm{Q}_k|] &= \sum_{i=1}^{l} \psi\left(\frac{\tilde{\nu}_k + 1 - i}{2}\right) + l \ln 2 + \ln |\tilde{W}_k|.\end{aligned}$$

where $\psi(\cdot)$ is the digamma function defined in the text.
Step 2: The lower bound is given by

$$\begin{aligned}\mathcal{F}(q; \boldsymbol{P}) &= \mathbb{E}_{q_z^{(j+1)} q_\mu^{(j+1)} q_Q^{(j+1)}}[\ln p(\mathcal{X}, \mathcal{Z}, \boldsymbol{\mu}_{1:K}, \mathrm{Q}_{1:K})]\\ &\quad - \mathbb{E}_{q_z^{(j+1)}}[\ln q_z^{(j+1)}(\mathcal{Z})] - \mathbb{E}_{q_\mu^{(j+1)}}[\ln q_\mu^{(j+1)}(\boldsymbol{\mu}_{1:K})]\\ &\quad - \mathbb{E}_{q_Q^{(j+1)}}[\ln q_Q^{(j+1)}(\mathrm{Q}_{1:K})]\\ &= \mathbb{E}_{q_z^{(j+1)}}\left[\sum_{n=1}^{N} \sum_{k=1}^{N} \mathrm{z}_{n_k} \ln P_k\right] + \text{constants},\end{aligned}$$

where constants are independent on $\boldsymbol{P}$. Hence,

$$\mathcal{F}(q; \boldsymbol{P}) = \sum_{n=1}^{N} \sum_{k=1}^{N} \rho_{n_k} \ln P_k + \text{constants}. \tag{11}$$

$\boldsymbol{P}$ now results by maximizing $\mathcal{F}(q; \boldsymbol{P})$ with respect to $\boldsymbol{P}$ subject to the constraint

$$\sum_{k=1}^{K} P_k = 1. \tag{12}$$

Thus,

$$\frac{\partial}{\partial \boldsymbol{P}} \left[\sum_{k=1}^{K} \ln P_k (\sum_{n=1}^{N} \rho_{n_k}) - \lambda \sum_{k=1}^{K} P_k\right] = 0,$$

or

$$P_k = \frac{1}{\lambda} \sum_{n=1}^{N} \rho_{n_k},\ \ k = 1, 2, \ldots, K. \tag{13}$$

Plugging (13) into (12)

$$\lambda = \sum_{n=1}^{N}\sum_{k=1}^{K} \rho_{n_k} = N.$$

or

$$P_k = \frac{1}{N}\sum_{n=1}^{N} \rho_{n_k},\ k = 1, 2, \ldots, K.$$

13.7. Consider the Gaussian Mixture model of Problem 13.6, with the following priors imposed on $\boldsymbol{\mu}$, Q, and $\mathbf{P}$:

$$\begin{aligned} p(\boldsymbol{\mu}, Q) &= p(\boldsymbol{\mu}|Q)p(Q) \\ &= \prod_{k=1}^{K} \mathcal{N}\left(\boldsymbol{\mu}_k|\mathbf{0}, (\lambda Q_k)^{-1}\right) \mathcal{W}(Q_k|\nu_0, W_o), \end{aligned}$$

that is, a Gaussian-Wishart product and

$$p(\boldsymbol{P}) = \mathrm{Dir}(\boldsymbol{P}|a) \propto \prod_{k=1}^{K} P_k^{a-1},$$

i.e., a Dirichlet prior. That is, $\mathbf{P}$ is treated as a random vector. Derive the E algorithmic steps of the variational Bayesian approximation adopting the mean field approximation for the involved posterior pdfs. We have adopted the notation $\boldsymbol{\mu}$ in place of $\boldsymbol{\mu}_{1:K}$ and Q in place of $Q_{1:K}$, for notational simplicity.

Solution: If $\mathcal{Z}$ is the set of latent variables associated with the mixture indices, we have

$$q(\mathcal{Z}, \boldsymbol{P}, \boldsymbol{\mu}, Q) = q(\mathcal{Z})q(\boldsymbol{P})q(\boldsymbol{\mu}, Q).$$

Step 1a: We have that

$$\begin{aligned} \ln q_z^{(j+1)}(\mathcal{Z}) &= \mathbb{E}_{q_P^{(j)} q_{\mu,Q}^{(j)}}[\ln p(\mathcal{X}, \mathcal{Z}, \mathbf{P}, \boldsymbol{\mu}, \mathrm{Q})] \\ &= \mathbb{E}_{q_P^{(j)} q_{\mu,Q}^{(j)}}[\ln p(\mathcal{X}|\mathcal{Z}, \boldsymbol{\mu}, \mathrm{Q}) + \ln p(\mathcal{Z}|\mathbf{P})] + \text{constants}, \end{aligned}$$

or

$$\begin{aligned} \ln q_z^{(j+1)}(\mathcal{Z}) &= \mathbb{E}_{q_P^{(j)}}[\ln p(\mathcal{Z}|\boldsymbol{P})] + \mathbb{E}_{q_{\mu,Q}^{(j)}}[\ln p(\mathcal{X}|\mathcal{Z}, \boldsymbol{\mu}, \mathrm{Q})] \\ &= \mathbb{E}_{q_P^{(j)}}[\sum_{n=1}^{N}\sum_{k=1}^{K} z_{n_k} \ln \mathrm{P}_k] + \sum_{n=1}^{N}\sum_{k=1}^{K} z_{n_k} \Big(\mathbb{E}_{q_{\mu,Q}^{(j)}}\left[\frac{1}{2}\ln|\mathrm{Q}_k|\right] \\ &\quad - \mathbb{E}_{q_{\mu,Q}^{(j)}}\left[\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T \mathrm{Q}_k (\boldsymbol{x}_n - \boldsymbol{\mu}_k)\right] \Big), \end{aligned}$$

or if we define

$$\ln \pi_{n_k} = \mathbb{E}_{q_P^{(j)}}[\ln \mathrm{P}_k] + \frac{1}{2}\,\mathbb{E}_{q^{(j)}_{\mu,Q}}\,[\ln|\mathrm{Q}_k|] - \frac{1}{2}\,\mathbb{E}_{q^{(j)}_{\mu,Q}}[(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T \mathrm{Q}_k (\boldsymbol{x}_n - \boldsymbol{\mu}_k)], \tag{14}$$

then

$$q_z^{(j+1)}(\mathcal{Z}) = \prod_{n=1}^{N}\prod_{k=1}^{N} \rho_{n_k}^{z_{n_k}}, \tag{15}$$

with

$$\rho_{n_k} = \frac{\pi_{n_k}}{\sum_{k=1}^{K} \pi_{n_k}}, \tag{16}$$

with $v_{q_z^{(j+1)}}[\mathrm{z}_{n_k}] = \rho_{n_k}$, by definition and the binary nature of z_{n_k}.

Step 1b:

$$\begin{aligned}
\ln q_P^{(j+1)}(\boldsymbol{P}) &= \mathbb{E}_{q_z^{(j+1)} q^{(j)}_{\mu,Q}}\Big[\ln p(\mathcal{X}|\mathcal{Z}, \boldsymbol{\mu}, \mathrm{Q}) + \ln p(\mathcal{Z}|\boldsymbol{P}) + \ln(\boldsymbol{P})\Big] + \text{constants}\\
&= \mathbb{E}_{q_z^{(j+1)}}\left[\sum_{n=1}^{N}\sum_{k=1}^{K} \mathrm{z}_{n_k} \ln P_k\right] + (a-1)\sum_{k=1}^{K} \ln P_k + \text{constants}\\
&= \sum_{k=1}^{K}\left(\sum_{n=1}^{N} \rho_{n_k} + a - 1\right) \ln P_k + \text{constants},
\end{aligned}$$

from which we obtain

$$q_P^{(j+1)}(\boldsymbol{P}) \propto \prod_{k=1}^{K} P_k^{a_k - 1},$$

with

$$a_k = a + \sum_{n=1}^{N} \rho_{n_k},$$

that is, it remains Dirichlet.

Step 1c:

$$\begin{aligned}
&\ln q_{\mu,Q}^{(j+1)}(\boldsymbol{\mu}, Q) = \mathbb{E}_{q_z^{(j+1)} q_P^{(j+1)}}\Big[\ln p(\mathcal{X}|\mathcal{Z}, \boldsymbol{\mu}, Q, P) + \ln p(\mathcal{Z}|P) + \ln p(P) + \ln p(\boldsymbol{\mu}, Q)\Big]\\
&\quad = \sum_{n=1}^{N}\sum_{k=1}^{K} \mathbb{E}_{q^{(j+1)}}[\mathrm{z}_{n_k}]\left(\frac{1}{2}\ln|Q_k| - \frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T Q_k (\boldsymbol{x}_n - \boldsymbol{\mu}_k)\right)\\
&\quad + \frac{1}{2}\sum_{k=1}^{K} \ln(|\lambda Q_k|) - \frac{1}{2}\sum_{k=1}^{K} \boldsymbol{\mu}_k^T (\lambda Q_k) \boldsymbol{\mu}_k + \frac{\nu_0 - l - 1}{2}\sum_{k=1}^{K} \ln|Q_k|\\
&\quad - \frac{1}{2}\sum_{k=1}^{K} \mathrm{trace}\{W_0^{-1} Q_k\} + \text{constants},
\end{aligned}$$

or

$$\begin{aligned}\ln q_{\mu,Q}^{(j+1)} = \sum_{k=1}^{K} \Big(& \underbrace{\frac{1}{2}\ln|Q_k|\sum_{k=1}^{N}\rho_{n_k}}_{A} - \underbrace{\frac{1}{2}\mathrm{trace}\{Q_k\sum_{n=1}^{N}\rho_{n_k}\boldsymbol{x}_n\boldsymbol{x}_n^T\}}_{C}\Big) \\ & \underbrace{-\frac{1}{2}\boldsymbol{\mu}_k^T]\left(Q_k\sum_{n=1}^{N}\rho_{n_k}\right)\boldsymbol{\mu}_k}_{B} + \underbrace{\boldsymbol{\mu}_k^T Q_k\sum_{n=1}^{N}\rho_{n_k}\boldsymbol{x}_n}_{B} \\ & + \underbrace{\frac{1}{2}\ln|\lambda Q_k|}_{A}\underbrace{-\frac{1}{2}\boldsymbol{\mu}_k^T(\lambda Q_k)\boldsymbol{\mu}_k}_{B} + \underbrace{\frac{\nu_0 - l - 1}{2}\ln|Q_k|}_{A} \\ & \underbrace{-\frac{1}{2}\mathrm{trace}\{W_0^{-1}Q_k\}}_{C} + \text{constants.}\Big)\end{aligned}$$

Combining all the terms A, B, C respectively together we obtain

$$\begin{aligned}A &: \frac{\nu_0 + 1 + \sum_{n=1}^{N}\rho_{n_k} - l - 1}{2}\ln|Q_k| \\ B &: -\frac{1}{2}\boldsymbol{\mu}_k^T(\lambda + \sum_{n=1}^{N}\rho_{n_k})Q_k\boldsymbol{\mu}_k + \boldsymbol{\mu}_k^T Q_k\sum_{n=1}^{N}\rho_{n_k}\boldsymbol{x}_n,\end{aligned}$$

which is quadratic with respect to $\boldsymbol{\mu}_k$ and it is rewritten, according to the appendix of the chapter, as

$$B: \ -\frac{1}{2}(\boldsymbol{\mu}_k - \hat{\boldsymbol{\mu}}_k)^T(\tilde{\lambda}Q_k)(\boldsymbol{\mu}_k - \hat{\boldsymbol{\mu}}_k) + \frac{1}{2}\hat{\boldsymbol{\mu}}_k^T(\tilde{\lambda}Q_k)\hat{\boldsymbol{\mu}}_k,$$

where

$$\tilde{\lambda} = \lambda + \sum_{n=1}^{N}\rho_{n_k}$$

$$\hat{\boldsymbol{\mu}}_k = \frac{1}{\tilde{\lambda}}\sum_{n=1}^{N}\rho_{n_k}\boldsymbol{x}_n$$

$$C: \frac{1}{2}\mathrm{trace}\{(W_0^{-1} + \sum_{n=1}^{N}\rho_{n_k}\boldsymbol{x}_n\boldsymbol{x}_n^T)Q_k\}.$$

Combining A, B and C, we finally obtain

$$\begin{aligned}\ln q_{\mu,Q}^{(j+1)}(\boldsymbol{\mu}, Q) = \sum_{k=1}^{K}\Big(& -\frac{1}{2}(\boldsymbol{\mu}_k - \hat{\boldsymbol{\mu}}_k)^T(\tilde{\lambda}Q_k)(\boldsymbol{\mu}_k - \hat{\boldsymbol{\mu}}_k) + \\ & \frac{\tilde{\nu}_k - l - 1}{2}\ln|Q_k| - \frac{1}{2}\mathrm{trace}\{\tilde{W}_k^{-1}Q_k\}\Big),\end{aligned}$$

where

$$\tilde{\nu}_k = \nu_0 + \sum_{n=1}^{N} \rho_{n_k} + 1,$$

and

$$\tilde{W}_k^{-1} = \tilde{W}_0^{-1} + \tilde{\lambda}\hat{\boldsymbol{\mu}}_k\hat{\boldsymbol{\mu}}_k^T - \sum_{n=1}^{N} \rho_{n_k}\boldsymbol{x}_n\boldsymbol{x}_n^T.$$

Hence,

$$q_{\mu,Q}^{(j+1)} = \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{\mu}_k|\hat{\boldsymbol{\mu}}_k, (\tilde{\lambda}Q_k)^{-1})\mathcal{W}(Q_k|\tilde{\nu}_k, \tilde{W}_k).$$

Thus, it is a Gaussian-Wishart product. Note that for the algorithm to be complete the expected values of $\ln|Q_k|$ and $(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T Q_k(\boldsymbol{x}_n - \boldsymbol{\mu}_k)$, w.r. to $q_{\mu,Q}^{(j+1)}$ have to be computed. The former one is obtained from the Wishart properties, as in Problem 13.6

13.8. If $\boldsymbol{\mu}$ and Q are distributed according to a Gaussian-Wishart product

$$p(\boldsymbol{\mu}, Q) = \mathcal{N}(\boldsymbol{\mu}|\hat{\boldsymbol{\mu}}, (\lambda Q)^{-1})\mathcal{W}(Q|\nu, W),$$

then compute the expectation

$$\mathbb{E}[\boldsymbol{\mu}^T \mathrm{Q}\boldsymbol{\mu}].$$

Solution: We have that

$$\begin{aligned}\mathbb{E}[\boldsymbol{\mu}^T \mathrm{Q}\boldsymbol{\mu}] = \mathbb{E}[\mathrm{trace}\{\mathrm{Q}\boldsymbol{\mu}\boldsymbol{\mu}^T\}] = \\ \mathbb{E}_Q\,\mathbb{E}_{\mu|Q}[\mathrm{trace}\{\mathrm{Q}\boldsymbol{\mu}\boldsymbol{\mu}^T\}] = \\ \mathrm{trace}\{\mathbb{E}_Q[\mathrm{Q}\,\mathbb{E}_{\mu|Q}[\boldsymbol{\mu}\boldsymbol{\mu}^T]]\},\end{aligned}$$

which from Eq. (12.74) in the text becomes

$$\begin{aligned}\mathbb{E}[\boldsymbol{\mu}^T \mathrm{Q}\boldsymbol{\mu}] &= \mathrm{trace}\{\mathbb{E}_Q[\mathrm{Q}((\lambda \mathrm{Q})^{-1} + \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^T)]\} \\ &= \mathrm{trace}\left\{\frac{1}{\lambda}I + \mathbb{E}_Q[\mathrm{Q}\hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^T]\right\} \\ &= \frac{l}{\lambda} + \mathrm{trace}\{\nu W\hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^T\} \\ &= \frac{l}{\lambda} + \nu\hat{\boldsymbol{\mu}}^T W\hat{\boldsymbol{\mu}},\end{aligned}$$

where l is the respective dimensionality.

13.9. Derive the Hessian matrix w.r. to $\boldsymbol{\theta}$ of the cost function

$$\begin{aligned}J(\boldsymbol{\theta}) = \sum_{n=1}^{N}[y_n \ln\sigma(\boldsymbol{\phi}^T(x_n)\boldsymbol{\theta}) + (1-y_n)\ln(1-\sigma(\boldsymbol{\phi}^T(x_n)\boldsymbol{\theta}))] \\ - \frac{1}{2}\boldsymbol{\theta}^T A\boldsymbol{\theta},\end{aligned}$$

where

$$\sigma(z) = \frac{1}{1 + \exp(-z)}.$$

Solution: Define

$$t = \sigma(z).$$

Then we have

$$\begin{aligned} \frac{\partial t}{\partial z} &= \frac{\partial}{\partial z} \frac{1}{1 + \exp(-z)} = \\ &= \frac{\exp(-z)}{(1 + \exp(-z))^2} = \\ &= \sigma(z)(1 - \sigma(z)) = t(1 - t). \end{aligned} \tag{17}$$

Now let

$$J_n(t) = y_n \ln t + (1 - y_n) \ln(1 - t_n). \tag{18}$$

We have

$$\frac{\partial J_n(t)}{\partial t} = y_n \frac{1}{t} - \frac{(1 - y_n)}{1 - t} = \frac{y_n - t}{t(1 - t)}. \tag{19}$$

Combining (17) and (19) and letting

$$z_n = \boldsymbol{\phi}^T(x_n)\boldsymbol{\theta},$$

we obtain

$$\begin{aligned} \frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \sum_{n=1}^{N} \frac{\partial J_n(t_n)}{\partial t_n} \frac{\partial t_n}{\partial z_n} \frac{\partial z_n}{\partial \boldsymbol{\theta}} - A\boldsymbol{\theta} \\ &= \sum_{n=1}^{N} \frac{y_n - t_n}{t_n(1 - t_n)} t_n(1 - t_n)\boldsymbol{\phi}(x_n) \\ &= \sum_{n=1}^{N} (y_n - t_n)\boldsymbol{\phi}(x_n) - A\boldsymbol{\theta} \\ &= \sum_{n=1}^{N} (y_n - \sigma(\boldsymbol{\phi}^T(x_n)\boldsymbol{\theta}))\boldsymbol{\phi}(x_n) - A\boldsymbol{\theta}. \end{aligned}$$

Hence, for the Hessian we have

$$\begin{aligned} \left[\frac{\partial^2 J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]_{ij} &= -[A]_{ij} - \sum_{n=1}^{N} \frac{\partial t_n}{\partial z_n} \frac{\partial z_n}{\partial \theta_j} \phi_i(\boldsymbol{x}_n) \\ &= -[A]_{ij} - \sum_{n=1}^{N} \phi_j(\boldsymbol{x}_n)(t_n(1 - t_n))\phi_i(\boldsymbol{x}_n) \\ &= -[A]_{ij} - \sum_{n=1}^{N} \phi_j(\boldsymbol{x}_n)\sigma(\boldsymbol{\phi}^T(\boldsymbol{x}_n)\boldsymbol{\theta})(1 - \sigma(\boldsymbol{\phi}^T(\boldsymbol{x}_n)\boldsymbol{\theta})\phi_i(\boldsymbol{x}_n), \end{aligned}$$

or

$$\frac{\partial^2 J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -A - \Phi^T T \Phi,$$

where $T = \text{diag}\{t_1, t_2, \ldots, t_N\}$.

13.10. Show that the marginal of a Gaussian pdf with a gamma prior on the variance, after integrating out the variance, is the student's-t pdf, given by

$$\text{st}(x|\mu,\lambda,\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})}\left(\frac{\lambda}{\pi\nu}\right)^{1/2} \frac{1}{\left(1+\frac{\lambda(x-\mu)^2}{\nu}\right)^{\frac{\nu+1}{2}}}. \tag{20}$$

Solution: From the text in the chapter and for the one dimensional (parameter) case, we have that

$$\begin{aligned} p(\theta;a,b) &= \int p(\theta|\alpha)p(\alpha)d\alpha \\ &= \int \mathcal{N}(\theta|0,\alpha^{-1})\text{Gamma}(\alpha|a,b)d\alpha \\ &= \int \left(\frac{1}{2\pi}\right)^{\frac{1}{2}} \alpha^{\frac{1}{2}} \exp\left(-\frac{1}{2}\alpha(x-\mu)^2\right)\frac{1}{\Gamma(a)}b^a\alpha^{a-1}\exp\left(-b\alpha\right)d\alpha, \end{aligned} \tag{21}$$

or

$$\begin{aligned} p(\theta;a,b) &= \frac{1}{\Gamma(a)}\left(\frac{1}{2\pi}\right)^{\frac{1}{2}} b^a \int \alpha^{a-\frac{1}{2}}\exp\left(-(b+\frac{1}{2}(x-\mu)^2)\alpha\right)d\alpha \\ &= \frac{1}{\Gamma(a)}\left(\frac{1}{2\pi}\right)^{\frac{1}{2}} b^a \left(b+\frac{1}{2}(x-\mu)^2\right)^{-a-\frac{1}{2}}. \\ &\int \left(b+\frac{1}{2}(x-\mu)^2\right)^{a+\frac{1}{2}} \alpha^{a-\frac{1}{2}}\exp\left(-(b+\frac{1}{2}(x-\mu)^2)\alpha\right)d\alpha. \end{aligned} \tag{22}$$

Note that the quantity under the integral is a gamma distribution with parameters $\alpha + \frac{1}{2}$ and $b + \frac{1}{2}(x-\mu)^2$ and its integration results in the corresponding normalization constant, hence

$$p(\theta;a,b) = \frac{1}{\Gamma(a)}\left(\frac{1}{2\pi}\right)^{\frac{1}{2}} b^a \left(b+\frac{1}{2}(x-\mu)^2\right)^{-a-\frac{1}{2}}\Gamma(a+\frac{1}{2}), \tag{23}$$

and the standard form of the student's-t pdf results from the previous in a trivial way by introducing the change of variables, $\nu = 2a$ and $a = \lambda b$.

13.11. Derive the pair of recursions (13.62)-(13.63).

Solution: Our starting point is Eq. (13.60).

$$\begin{aligned} L(\boldsymbol{\alpha}, \beta) &= -\frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln|\beta^{-1}I + \Phi A^{-1}\Phi^T| \\ & -\frac{1}{2}\boldsymbol{y}^T\left(\beta^{-1}I + \Phi A^{-1}\Phi^T\right)^{-1}\boldsymbol{y}. \end{aligned} \tag{24}$$

However, using the identity concerning determinants from the matrix theory (Appendix A), we have

$$\begin{aligned} C &:= \ln|\beta^{-1}I + \Phi A^{-1}\Phi^T| = \ln\frac{|A|^{-1}}{\beta}|I||A + \beta\Phi^T\Phi| \\ &= -\ln|\mathit{\Sigma}| - \ln|A| - N\ln\beta, \end{aligned} \tag{25}$$

where,

$$\mathit{\Sigma} := \left(A + \beta\Phi^T\Phi\right)^{-1}. \tag{26}$$

Also, from the matrix inversion lemma we have,

$$D := \left(\beta^{-1} + \Phi A^{-1}\Phi^T\right)^{-1} = \beta I - \beta\Phi\left(A + \beta\Phi^T\Phi\right)^{-1}\Phi^T\beta. \tag{27}$$

Thus,

$$\begin{aligned} E &:= \boldsymbol{y}\left(\beta^{-1}I + \Phi A^{-1}\Phi^T\right)^{-1}\boldsymbol{y} = \beta\boldsymbol{y}^T\boldsymbol{y} - \\ & \beta\boldsymbol{y}^T\Phi\mathit{\Sigma}\Phi^T\beta\boldsymbol{y} \\ &= \beta\boldsymbol{y}^T\left(\boldsymbol{y} - \Phi\boldsymbol{\mu}\right), \end{aligned}$$

where

$$\boldsymbol{\mu} = \beta\mathit{\Sigma}\Phi^T\boldsymbol{y}. \tag{28}$$

From the previous definitions and after some algebra, one can show that

$$E = \beta\boldsymbol{y}^T\left(\boldsymbol{y} - \Phi\boldsymbol{\mu}\right) = \beta||\boldsymbol{y} - \Phi\boldsymbol{\mu}||^2 + \boldsymbol{\mu}^T A\boldsymbol{\mu}. \tag{29}$$

Indeed

$$\begin{aligned} \left(\boldsymbol{y} - \Phi\boldsymbol{\mu}\right)^T\left(\boldsymbol{y} - \Phi\boldsymbol{\mu}\right) &= \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\Phi\boldsymbol{\mu} + \\ & \boldsymbol{\mu}^T\Phi^T\Phi\boldsymbol{\mu} - \boldsymbol{\mu}^T\Phi^T\boldsymbol{y}, \end{aligned}$$

or

$$\begin{aligned} \beta\boldsymbol{y}^T\left(\boldsymbol{y} - \Phi\boldsymbol{\mu}\right) &= \beta||\boldsymbol{y} - \Phi\boldsymbol{\mu}||^2 + \beta\boldsymbol{\mu}^T\Phi^T\boldsymbol{y} - \\ & \beta\boldsymbol{\mu}^T\Phi^T\Phi^T\boldsymbol{\mu}, \end{aligned}$$

or taking into account that

$$\beta\Phi^T\Phi = \mathit{\Sigma}^{-1} - A,$$

we get

$$\begin{aligned}\beta \boldsymbol{y}^T\left(\boldsymbol{y}-\Phi\boldsymbol{\mu}\right) &= \beta||\boldsymbol{y}-\Phi\boldsymbol{\mu}||^2+\beta\boldsymbol{\mu}^T\Phi^T\boldsymbol{y}- \\ &\quad \boldsymbol{\mu}^T\left(\Sigma^{-1}-A\right)\boldsymbol{\mu} \\ &= \beta||\boldsymbol{y}-\Phi\boldsymbol{\mu}||^2+\beta\boldsymbol{\mu}^T\Phi^T\boldsymbol{y}- \\ &\quad \boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu}+\boldsymbol{\mu}^TA\boldsymbol{\mu}.\end{aligned}$$

Taking into account that

$$\Sigma^{-1}\boldsymbol{\mu}=\beta\Sigma^{-1}\Sigma\Phi^T\boldsymbol{y},$$

proves the claim.

Combining, the formulae for C, D, E, we obtain that

$$\begin{aligned}L(\boldsymbol{\alpha},\beta) &= -\frac{N}{2}\ln(2\pi)-\frac{1}{2}\ln|\Sigma^{-1}|+\frac{1}{2}\ln|A|+\frac{1}{2}N\ln\beta- \\ &\quad \frac{1}{2}\beta||\boldsymbol{y}-\Phi\boldsymbol{\mu}||^2-\frac{1}{2}\boldsymbol{\mu}^TA\boldsymbol{\mu}. \qquad (30)\end{aligned}$$

Using standard formulae for the differentiation for matrices and determinants, we obtain

$$\begin{aligned}\frac{\partial\ln|\Sigma^{-1}|}{\partial\alpha_k} &= \frac{1}{|\Sigma^{-1}|}\frac{\partial|\Sigma^{-1}|}{\partial\alpha_k} \\ &= \frac{1}{|\Sigma^{-1}|}|\Sigma^{-1}|\text{trace}\left\{\Sigma\frac{\partial\Sigma^{-1}}{\partial\alpha_k}\right\} \qquad (31)\end{aligned}$$

However,

$$\Sigma^{-1}=A+\beta\Phi^T\Phi,$$

hence

$$\frac{\partial\Sigma^{-1}}{\partial\alpha_k}=\text{diag}\{0,\ldots,1,0,\ldots,0\}, \qquad (32)$$

with an 1 at the kth position. Thus

$$\frac{\partial\ln|\Sigma^{-1}|}{\partial\alpha_k}=\Sigma_{kk}. \qquad (33)$$

Also

$$\frac{\partial\ln|A|}{\partial\alpha_k}=\frac{1}{|A|}\frac{\partial}{\partial\alpha_k}\prod_j\alpha_j=\frac{1}{\alpha_k}. \qquad (34)$$

Also,

$$\frac{\partial}{\partial\alpha_k}\left(\boldsymbol{\mu}^TA\boldsymbol{\mu}\right)=\mu_k^2. \qquad (35)$$

Thus we have

$$\frac{L(\boldsymbol{\alpha},\beta)}{\partial\alpha_k}=-\frac{1}{2}\Sigma_{kk}+\frac{1}{2}\frac{1}{\alpha_k}-\frac{1}{2}\mu_k^2. \qquad (36)$$

Equating to zero and setting

$$\gamma_k = 1 - \alpha_k \Sigma_{kk},$$

we obtain

$$\alpha_k = \frac{1}{\mu_k^2 + \Sigma_{kk}} = \frac{\gamma_k}{\mu_k^2}.$$

We now turn into the derivation with respect to β. We have that

$$\begin{aligned}
\frac{\partial \ln |\Sigma^{-1}|}{\partial \beta} &= \frac{1}{|\Sigma^{-1}|}|\Sigma^{-1}|\text{trace}\{\Sigma \frac{\partial \Sigma^{-1}}{\partial \beta}\} \\
&= \text{trace}\{\Sigma \Phi^T \Phi\} = \text{trace}\{\frac{1}{\beta}\Sigma \left(\beta \Phi^T \Phi + A - A\right)\} \\
&= \text{trace}\{\frac{1}{\beta}\Sigma \left(\Sigma^{-1} - A\right)\} \\
&= \frac{1}{\beta}\text{trace}\{(I - \Sigma A)\} = \frac{1}{\beta}\sum_k \gamma_k. \qquad (37)
\end{aligned}$$

Thus finally we have

$$\frac{\partial L(\boldsymbol{\alpha}, \beta)}{\partial \beta} = -\frac{1}{2}||\boldsymbol{y} - \Phi \boldsymbol{\mu}||^2 + \frac{1}{2}\frac{N}{\beta} - \frac{1}{2}\sum_k \gamma_k, \qquad (38)$$

which after setting it to zero leads to,

$$\beta = \frac{N - \sum_k \gamma_k}{||\boldsymbol{y} - \Phi \boldsymbol{\mu}||^2}, \qquad (39)$$

which concludes the proof.

13.12. Consider a two class classification task and assume that the feature vectors in each one of the two classes, $\omega_1,\ \omega_2$, are distributed according to the Gaussian pdf. Both classes share the same covariance matrix Σ, and the mean values are $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, respectively. Prove that, given an observed feature vector, $\boldsymbol{x} \in \mathbb{R}^l$, the posterior probabilities for deciding in favor of one of the classes is given by the logistic function, i.e.,

$$P(\omega_2|\boldsymbol{x}) = \frac{1}{1 + \exp\left(-\boldsymbol{\theta}^T \boldsymbol{x} + \theta_0\right)},$$

where

$$\boldsymbol{\theta} := \Sigma^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1),$$

and

$$\theta_0 = \frac{1}{2}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T \Sigma^{-1}(\boldsymbol{\mu}_2 + \boldsymbol{\mu}_1) + \ln \frac{P(\omega_1)}{P(\omega_2)}$$

Solution: We have that

$$p(\boldsymbol{x}|\omega_i) = \frac{1}{(2\pi)^{l/2}|\Sigma^{-1}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)^T \Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_i)\right),\ i = 1, 2.$$

Hence,

$$P(\omega_1|\boldsymbol{x})p(\boldsymbol{x}) = p(\boldsymbol{x}|\omega_1)P(\omega_1),$$

and

$$P(\omega_2|\boldsymbol{x})p(\boldsymbol{x}) = p(\boldsymbol{x}|\omega_2)P(\omega_2),$$

where $P(\omega_1)$, $P(\omega_2)$ are the respective prior probabilities. Dividing the previous equations by part, substituting the Gaussian pdf in place of $p(\boldsymbol{x})$, it is a matter of simple algebra to see that,

$$\frac{P(\omega_1|\boldsymbol{x})}{P(\omega_2|\boldsymbol{x})} = \exp\left(-\boldsymbol{\theta}^T\boldsymbol{x} + \frac{1}{2}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T \Sigma^{-1}(\boldsymbol{\mu}_2 + \boldsymbol{\mu}_1)\right) \exp\left(\ln \frac{P(\omega_1)}{P(\omega_2)}\right),$$

where

$$\boldsymbol{\theta} = \Sigma^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1).$$

Taking into account that

$$P(\omega_2|\boldsymbol{x}) = 1 - P(\omega_1|\boldsymbol{x}),$$

we prove the claim.

13.13. Derive equation (13.74).

Solution: Our starting point is the cost

$$J := \sum_{n=1}^{N} \Big(y_n \ln \sigma(\boldsymbol{\theta}^T \boldsymbol{\phi}_n) + (1 - y_n) \ln \big(1 - \sigma(\boldsymbol{\theta}^T \boldsymbol{\phi}_n)\big) \Big) - \frac{1}{2} \boldsymbol{\theta}^T A \boldsymbol{\theta}, \quad (40)$$

where $\boldsymbol{\phi}_n := \phi(\boldsymbol{x}_n)$. Taking the gradient with respect to $\boldsymbol{\theta}$, we get

$$\begin{aligned} \nabla_{\boldsymbol{\theta}} J \quad = \quad & \sum_{n=1}^{N} \Big(y_n \frac{1}{\sigma(\boldsymbol{\theta}^T \boldsymbol{\phi}_n)} \boldsymbol{\phi}_n \sigma^{'}(\boldsymbol{\theta}^T \boldsymbol{\phi}) + \\ & (1 - y_n) \frac{1}{1 - \sigma(\boldsymbol{\theta}^T \boldsymbol{\phi}_n)} \boldsymbol{\phi}_n \big(- \sigma^{'}(\boldsymbol{\theta}^T \boldsymbol{\phi}_n)\big) \Big) - \\ & A\boldsymbol{\theta} \qquad (41) \end{aligned}$$

Taking into account that (see also next problem)

$$\sigma^{'}(z) = \sigma(z)\big(1 - \sigma(z)\big),$$

and after some trivial algebra we end up with

$$\begin{aligned} \nabla_{\boldsymbol{\theta}} J \quad &= \quad \sum_{n=1}^{N} \big(y_n - \sigma(\boldsymbol{\theta}^T \boldsymbol{\phi}_n)\big) \boldsymbol{\phi}_n - A\boldsymbol{\theta} \\ &= \quad \Phi^T(\boldsymbol{y} - \boldsymbol{s}) - A\boldsymbol{\theta}, \qquad (42) \end{aligned}$$

which is set equal to zero and proves the claim.

13.14. Show Equation (13.75).

Solution: By the respective definition we have that

$$\sigma(t) = \frac{1}{1+\exp(-t)},$$

hence

$$\frac{\partial \sigma(t)}{\partial t} = \sigma(t)(1-\sigma(t)).$$

Also, we have

$$\frac{\partial}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}^T \boldsymbol{\phi}(\boldsymbol{x}_n)) = \boldsymbol{\phi}(\boldsymbol{x}_n).$$

Let now,

$$s_n := \sigma(\boldsymbol{\theta}^T \boldsymbol{\phi}(\boldsymbol{x}_n)).$$

Then, by combining the previous formulae we get

$$\frac{\partial s_n}{\partial \boldsymbol{\theta}} = s_n(1-s_n)\boldsymbol{\phi}(\boldsymbol{x}_n).$$

Define now

$$a_n := y_n \ln s_n + (1-y_n)\ln(1-s_n).$$

Then after some simple algebra and taking into account well known differentiation rules concerning the logarithm, we readily obtain that

$$\frac{\partial a_n}{\partial \boldsymbol{\theta}} = y_n \boldsymbol{\phi}(\boldsymbol{x}_n) - s_n \boldsymbol{\phi}(\boldsymbol{x}_n).$$

For the needs of the proof, we need to consider the second derivative. The respective (i,j) element is readily obtained by

$$\begin{aligned}\left[\frac{\partial^2 a_n}{\partial \boldsymbol{\theta}^2}\right]_{ij} &= -s_n(1-s_n)\phi_i(\boldsymbol{x}_n)\phi_j(\boldsymbol{x}_n) = -\phi_i(\boldsymbol{x}_n)t_n\phi_j(\boldsymbol{x}_n) \\ &= \left[\boldsymbol{\phi}(\boldsymbol{x}_n)t_n\boldsymbol{\phi}(\boldsymbol{x}_n)^T\right]_{ij}.\end{aligned}$$

where

$$t_n := s_n(1-s_n).$$

Recalling the definition,

$$\Phi = \begin{bmatrix} \boldsymbol{\phi}^T(\boldsymbol{x}_1) \\ \vdots \\ \boldsymbol{\phi}^T(\boldsymbol{x}_N) \end{bmatrix}$$

as well as that

$$\begin{aligned}\ln\left(P(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\alpha})\right) &= \sum_{n=1}^{N}\Big[y_n \ln\sigma(\boldsymbol{\theta}^T\boldsymbol{\phi}(\boldsymbol{x}_n)) + (1-y_n)\ln\big(1-\sigma(\boldsymbol{\theta}^T\boldsymbol{\phi}(\boldsymbol{x}_n))\big)\Big] - \\ & \frac{1}{2}\boldsymbol{\theta}^T A\boldsymbol{\theta} + \text{constant}, \qquad (43)\end{aligned}$$

it is now straightforward to write

$$\Sigma^{-1} = A + \Phi^T T \Phi, \qquad (44)$$

where we have used that

$$\frac{\partial^2(\boldsymbol{\theta}^T A\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^2} = A.$$

13.15. Derive the recursion (13.77).

Solution: Taking the logarithm of $P(\boldsymbol{y}|\boldsymbol{\alpha})$ in (13.75) and keeping only the terms which depend on $\boldsymbol{\alpha}$, we have,

$$\ln P(\boldsymbol{y}|\boldsymbol{\alpha}) = \ln p(\hat{\boldsymbol{\theta}}_{\text{MAP}}|\boldsymbol{\alpha}) - \frac{1}{2}\ln|\Sigma^{-1}| + \text{constants}.$$

Taking the derivative of the first term on the right hand side is trivial, since the pdf is Gaussian and we have done it many times already. The emphasis will be given on computing the determinant term. Using the formula form the linear algebra,

$$\frac{\partial \ln|\Sigma^{-1}|}{\partial\alpha_k} = \frac{1}{|\Sigma^{-1}|}|\Sigma^{-1}|\text{trace}\{\Sigma\frac{\partial\Sigma^{-1}}{\partial\alpha_k}\}.$$

However, from the text we know that

$$\Sigma^{-1} = A + \Phi^T T\Phi,$$

and since A is diagonal (see, also, Problem 13.11), we get that

$$-\frac{1}{2}\frac{\partial\ln|\Sigma^{-1}|}{\partial\alpha_k} = -\frac{1}{2}\Sigma_{kk},$$

which is the last term in (13.77).

13.16. Show that if f is a convex function $f:\mathbb{R}^l \to \mathbb{R}$, then it is equal to the conjugate of its conjugate, i.e., $(f^*)^* = f$.

Solution: Recall from the theory that

$$f^*(\boldsymbol{\xi}) = \boldsymbol{\xi}^T\boldsymbol{x}_* - f(\boldsymbol{x}_*), \qquad (45)$$

where

$$\boldsymbol{x}_* : \boldsymbol{x}_*(\boldsymbol{\xi}) : \nabla f(\boldsymbol{x}_*) = \boldsymbol{\xi}, \tag{46}$$

where we have suppressed the dependence of $\boldsymbol{x}_*$ on $\boldsymbol{\xi}$. Hence, combining (45) and (46) we get

$$\boldsymbol{x}^T\boldsymbol{\xi} - f^*(\boldsymbol{\xi}) = f(\boldsymbol{x}_*) + (\boldsymbol{x} - \boldsymbol{x}_*)^T\nabla f(\boldsymbol{x}_*). \tag{47}$$

However, we know that if f is convex then

$$\boldsymbol{x}^T\boldsymbol{\xi} - f^*(\boldsymbol{\xi}) = f(\boldsymbol{x}_*) + (\boldsymbol{x} - \boldsymbol{x}_*)^T\nabla f(\boldsymbol{x}_*) \leq f(\boldsymbol{x}),\ \forall \boldsymbol{x},\ \boldsymbol{x}_* \in \mathbb{R}^l,$$

and equality, that is the maximum is achieved for $\boldsymbol{x} = \boldsymbol{x}_*$.

Hence, due to the dependence of $\boldsymbol{x}_*$ on $\boldsymbol{\xi}$ (Eq. (46)), we can write

$$f(\boldsymbol{x}) : \max_{\boldsymbol{\xi}}(\boldsymbol{x}^T\boldsymbol{\xi} - f^*(\boldsymbol{\xi})).$$

13.17. Prove that

$$f(x) = \ln\frac{\lambda}{2} - \lambda\sqrt{x},\ x \geq 0,$$

is a convex function.

Solution: It is known from the theory of convex functions that if $\frac{d^2 f(x)}{dx^2} \geq 0$, then $f(x)$ is convex ([Boyd 04]). Hence,

$$f'(x) = -\frac{\lambda}{2}x^{-\frac{1}{2}},\ f''(x) = \frac{\lambda}{4}x^{-\frac{3}{4}} \geq 0,\ x \geq 0.$$

13.18. Derive variational bounds for the logistic regression function

$$\sigma(x) = \frac{1}{1 + e^{-x}},$$

one of them in terms of a Gaussian function. For the latter case, use the transformation, $t = \sqrt{x}$.

Solution: This is a typical case where the function itself is neither convex or concave. We will first transform it, by taking the logarithm, which leads to a concave one.

a) $f(x) = \ln\sigma(x) = -\ln(1 + e^{-x})$. This is a concave function since

$$f'(x) = \frac{e^{-x}}{1 + e^{-x}},\ f''(x) = \frac{-e^{-x}}{(1 + e^{-x})^2} < 0.$$

Now,

$$f^*(\xi) = \min_x(\xi x + \ln(1 + e^{-x})),$$

and taking the derivative we obtain

$$\begin{aligned} x_* : \xi - \frac{e^{-x}}{1+e^{-x}} = 0 \Rightarrow \\ e^{-x_*} = \frac{\xi}{1-\xi} \Rightarrow x_* = \ln\frac{1-\xi}{\xi},\ 0<\xi<1. \end{aligned}$$

Hence

$$\begin{aligned} f^*(\xi) &= \xi \ln\frac{1-\xi}{\xi} + \ln\left(1+\frac{\xi}{1-\xi}\right) \\ &= -(1-\xi)\ln(1-\xi) - \xi\ln\xi, \end{aligned}$$

which is the binary entropy function. Hence.

$$\ln\sigma(x) \equiv f(x) \le \xi x - f^*(\xi),$$

or

$$\sigma(x) \le \exp(\xi x - f^*(\xi)),\ 0<\xi<1,$$

and this bounds the sigmoid regression by exponential functions.

b) Following [Jaak 97] , we write,

$$\begin{aligned} \ln\sigma(x) &= -\ln\big(1+\exp(-x)\big) = -\ln\Big(\exp(-x/2)\big(\exp(x/2)+\exp(-x/2)\big)\Big) \\ &= \frac{x}{2} - \ln\big(\exp(x/2)+\exp(-x/2)\big). \end{aligned} \tag{48}$$

Let us now define

$$f(x) := -\ln\big(\exp(\sqrt{x}/2)+\exp(-\sqrt{x}/2)\big),\ x\ge 0. \tag{49}$$

We will first show that $f(x)$ is a convex function. To this end we will prove that the second derivative is nonnegative, in the respective domain of definition. To this end, set

$$t=\sqrt{x},\ \text{or}\ \frac{dt}{dx}=\frac{1}{2\sqrt{x}}.$$

Hence,

$$\frac{df}{dx} = \frac{1}{2\sqrt{x}}\frac{dg}{dt} = -\frac{1}{4\sqrt{x}}\tanh(\frac{\sqrt{x}}{2}),$$

where

$$g(t) = \ln\big(\exp(t/2)+\exp(-t/2)\big).$$

and

$$\tanh(t) = \frac{\exp(t)-\exp(-t)}{\exp(t)+\exp(-t)}.$$

Following similar arguments as before, the second derivative turns out to be

$$\frac{d^2 f}{dx^2} = \frac{1}{8x}\Big(\frac{\tanh(\frac{\sqrt{x}}{2})}{\sqrt{x}} - \frac{1}{2}\big(1 - \tanh^2(\frac{\sqrt{x}}{2})\big)\Big),$$

where we have used the chain differentiation rule as well as the property,

$$\frac{d\tanh(y)}{dy} = 1 - \tanh^2(y) = \frac{1}{\cosh^2(y)}.$$

Let now

$$y = \sqrt{x}/2.$$

It suffices to show that

$$\frac{\tanh(y)}{2y} \geq \frac{1}{2\cosh^2(y)},$$

or recalling the known properties form the analysis,

$$\tanh(y) = \frac{\sinh(y)}{\cosh(y)},$$

and

$$2\sinh(y)\cosh(y) = \sinh(2y),$$

we have to show that

$$\frac{\sinh(2y)}{2y} \geq 1.$$

However, this is always true for $y \geq 0$, from the known from the analysis expansion

$$\sinh(y) = y + \frac{y^3}{3!} + \frac{y^5}{5!} + \dots,$$

which proves the claim about convexity.

From the convexity property we can now write that

$$f^*(\xi) = \max_x \big(\xi x - f(x)\big),\ \xi < 0.$$

Note that the constraint is necessary in order to guarantee that the conjugate function remains finite. Otherwise, as x tends to infinity, the conjugate function also tends to infinity. The conjugate function can be obtained by the differentiation and finding the optimal value x_*. However, there is no need for it, since the exact form is of no interest for us.

Thus we can now write that

$$f^*(\xi) \geq \xi x - f(x),$$

or

$$f(x) \geq \xi x - f^*(\xi).$$

Hence,

$$\ln \sigma(x) = x/2 + f(x^2) \geq x/2 + \xi x^2 - f^*(\xi),$$

or

$$\sigma(x) \geq \exp(-f^*(\xi)) \exp\big(\frac{x}{2} + \xi x^2\big),$$

which provides a bound in terms of a Gaussian function.

13.19. Prove equation (13.100).

Solution: We know that

$$\begin{aligned} Q(\boldsymbol{\xi}, \beta; \boldsymbol{\xi}^{(j)},\beta^{(j)}) = {} & \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) - \frac{\beta}{2}\mathbb{E}[\|\boldsymbol{y} - \Phi\theta\|^2] \\ & + \sum_{k=0}^{K-1}\ln\phi(\xi_k) - \frac{K}{2}\ln(2\pi) - \frac{1}{2}\ln|\Xi| - \frac{1}{2}\sum_{k=0}^{K-1}\frac{\mathbb{E}[\theta_k^2]}{\xi_k}. \end{aligned}$$

Taking the derivative with respect to ξ_k, and recalling that

$$\begin{aligned} |\Xi| &= \prod_{k=0}^{K-1}\xi_k, \\ \phi(\xi_k) &= \frac{\lambda}{2}\sqrt{2\pi\xi_k}\exp\left(-\frac{\lambda^2}{2}\xi_k\right), \end{aligned}$$

we have

- $$\begin{aligned} \frac{\partial}{\xi_k}\sum_{k=0}^{K-1}\ln\phi(\xi_k) &= \frac{1}{\phi(\xi_k)}\phi'(\xi_k) \\ &= \frac{\exp\left(\frac{\lambda^2}{2}\xi_k\right)}{\frac{\lambda}{2}\sqrt{2\pi\xi_k}}\Big(\frac{\lambda}{4}\sqrt{2\pi}\xi_k^{-1/2}\exp\left(-\frac{\lambda^2}{2}\xi_k\right) \\ &\quad + \frac{\lambda}{2}\sqrt{2\pi\xi_k}\left(-\frac{\lambda^2}{2}\right)\exp\left(-\frac{\lambda^2}{2}\xi_k\right)\Big) \\ &= \frac{1}{2}\xi_k^{-1} - \frac{\lambda^2}{2}. \end{aligned}$$

- $$-\frac{1}{2}\frac{\partial}{\partial\xi_k}\ln|\Xi| = -\frac{1}{2}\xi_k^{-1}.$$

- $$\frac{\partial}{\partial\xi_k}\left[-\frac{1}{2}\sum_{k=0}^{K-1}\frac{\mathbb{E}[\theta_k^2]}{\xi_k}\right] = \frac{\mathbb{E}[\theta_k^2]}{\xi_k^2}.$$

Combining the previous steps and equating to zero we obtain

$$\xi_k = \sqrt{\frac{\mathbb{E}[\theta_k^2]}{\lambda^2}}.$$

13.20. Derive the mean and variance of $\mathrm{G}(T_k)$ for a DP process.

Solution: Recall the mean and variance values of a Dirichlet distribution from Chapter 2. Also, for the case of DPs, the parameters of the associated Dirichlet distribution are $a_k = \alpha G_0(T_k)$. Then we get,

$$\mathbb{E}[\mathrm{G}(T_k)] = \frac{\alpha G_0(T_k)}{\alpha \sum_{k=1}^{K} G_0(T_k)} = G_0(T_k).$$

For the variance we get,

$$\begin{aligned} \mathrm{var}[\mathrm{G}(T_k)] &= \frac{\alpha G_0(T_k)\left(\alpha \sum_{k=1}^{K} G_0(T_k) - \alpha G_0(T_k)\right)}{\alpha^2 \left(\sum_{k=1}^{K} G_0(T_k)\right)\left(\alpha \sum_{k=1}^{K} G_0(T_k) + 1\right)} \\ &= \frac{G_0(T_k)\left(1 - G_0(T_k)\right)}{1+\alpha} \end{aligned}$$

13.21. Show that the posterior DP, after having obtained n observations from the set Θ, is given by,

$$G|\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n \sim \mathrm{DP}\left(\alpha + n,\ \frac{1}{\alpha+n}\left(\alpha G_0 + \sum_{i=1}^{n} \delta_{\boldsymbol{\theta}_i}(\boldsymbol{\theta})\right)\right)$$

Solution: By definition we have that

$$\alpha^{'} G_0^{'}(T_k) = \alpha G_0(T_k) + n_k.$$

Moreover, the above is true for *all* finite (measurable) partitions, e.g., for disjoint T_ks. Adding over k and taking into account that probabilities add to one, we get,

$$\alpha^{'} \sum_{k=1}^{K} G_0^{'}(T_k) = \alpha^{'} = \left(\alpha \sum_{k=1}^{K} G_0(T_k) + \sum_{k=1}^{K} n_k\right) = \alpha + n.$$

Hence we can now write that

$$(\alpha+n) G_0^{'}(T_k) = \alpha G_0(T_k) + n_k = \alpha G_0(T_k) + \sum_{i=1:\boldsymbol{\theta}_i \in T_k}^{n} \delta_{\boldsymbol{\theta}_i}(\boldsymbol{\theta}).$$

The above is true for any partition, hence we can write that.

$$G_0^{'} = \frac{1}{\alpha+n}\left(\alpha G_0 + \sum_{i=1}^{n} \delta_{\boldsymbol{\theta}_i}(\boldsymbol{\theta})\right).$$

13.22. The stick breaking construction of a DP is built around the following rule: $P_1 = \beta_1 \sim \text{Beta}(\beta|1,\alpha)$ and,

$$\begin{aligned} \beta_i &\sim \text{Beta}(\beta|1,\alpha), \\ P_i &= \beta_i \prod_{j=1}^{i-1}(1-\beta_j),\ i \geq 2. \end{aligned}$$

Show that if the number of steps is finite, i.e., we assume that $P_i = 0,\ i > T$, for some T, then $\beta_T = 1$.

Solution: If we stop at a step T, then we should have,

$$\sum_{i=1}^{T} P_i = 1$$

or

$$\begin{aligned} P_T &= 1 - P_1 - \sum_{i=2}^{T-1} P_i = 1 - \beta_1 - \sum_{i=2}^{T-1} \beta_i \prod_{j=1}^{i-1}(1-\beta_j) \\ &= \prod_{j=1}^{T-1}(1-\beta_j),\ T \geq 2. \end{aligned}$$

The last equality can be shown recursively. First, it holds true for $T = 2$, because $P_2 = 1 - P_1 = 1 - \beta_1$. In the sequel, assume that it is true for some $T > 2$, i.e.,

$$1 - \beta_1 - \sum_{i=2}^{T-1} \beta_i \prod_{j=1}^{i-1}(1-\beta_j) = \prod_{j=1}^{T-1}(1-\beta_j).$$

We will show that it is true for $T+1$. Indeed,

$$\begin{aligned} 1 - \beta_1 - \sum_{i=2}^{T} \beta_i \prod_{j=1}^{i-1}(1-\beta_j) &= 1 - \beta_1 - \sum_{i=2}^{T-1} \beta_i \prod_{j=1}^{i-1}(1-\beta_j) - \\ &\quad \beta_T \prod_{j=1}^{T-1}(1-\beta_j) \\ &= \prod_{j=1}^{T-1}(1-\beta_j) - \beta_T \prod_{j=1}^{T-1}(1-\beta_j) \\ &= \prod_{j=1}^{T}(1-\beta_j), \end{aligned}$$

and the claim has been proved.

Hence,

$$P_T = \beta_T \prod_{j=1}^{T-1}(1-\beta_j) = \prod_{j=1}^{T-1}(1-\beta_j) \Longrightarrow \beta_T = 1.$$

13.23. Show that in CRP, the cluster assignments are exchangeable and do not depend on the sequence that customers arrive, up to a permutation of the labels of the tables.

Solution: Let us assume that n customers have arrived and K_n tables (clusters) have been formed. Associate with each customer a label, y_i, where $y_i = k$ means that the ith customer sits at table $k \in \{1, 2, \ldots, K_n\}$. According to the chain rule of probabilities, we get

$$P(y_1, y_2, \ldots, y_n) = P(y_1)P(y_2|y_1).....P(y_n|y_{n-1}, y_{n-2}, \ldots, 1). \qquad (50)$$

In the above product on the right hand side, collect all the terms where the corresponding label, y_i, takes the value k; that is, we consider together all customers that sit at table k. Note that the corresponding number of terms is equal to the number of customers sitting at the kth table, denoted as n_k. Each of these terms represents the respective probability, which has been assigned according to the rule in Eq. (13.119). Thus, we can write that this product is equal to

$$\frac{\alpha}{n(k,1)-1+\alpha} \times \frac{1}{n(k,2)-1+\alpha} \times \frac{2}{n(k,3)-1+\alpha} \times \ldots \times \frac{n_k - 1}{n(k,n_k)-1+\alpha}.$$

Indeed, one of the terms will be related to the first customer sitting at table k. This corresponds to the first term in the above product, where $n(k, 1)$ denotes the order in which the aforementioned customer arrived. For example, if he/she were the first customer, $n(k, 1) = 1$, if he/she were, say, the 9th customer in the queue entering the restaurant, then $n(k, 1) = 9$, etc. The rest of the terms are similarly defined. Now, the term of the product in Eq. (50) that corresponds to the second customer sitting at the kth table is equal to the second term above, and so on up to the n_kth term. Thus, we can now write the product in Eq. (50) as,

$$P(y_1, y_2, \ldots, y_n) = \frac{\alpha^{K_n} \prod_{i=1}^{K_n}(n_i - 1)!}{\prod_{i=1}^{n}(i - 1 + \alpha)}. \qquad (51)$$

The numerator is straightforward, since we have K_n groups of terms in Eq. (50). For the denominator, recall that each customer arrives *only once*. Hence, each one of the terms $n(k, j)$, $j = 1, 2, \ldots, n_k$, and $k = 1, 2, \ldots, K_n$, has a unique and distinct, from all the others, value in the set $\{1, 2, \ldots, n\}$. Thus, each one of the terms in the denominators in the last product is equal to $i - 1 + \alpha$ for a specific value of i in the previous set of integers up to n.

Observe that the product in Eq. (51) depends *only* on the number of tables/clusters, K_n, the number of customers, n, the number of customers sitting at the tables, and *not* on the sequence in which customers have arrived and sat at the tables.

13.24. Show that in an IBP, the probabilities for $P(Z)$ and the equivalence classes, $P([Z])$, are given by the formulae,

$$P(Z) = \prod_{k=1}^{K} \frac{\alpha}{K} \frac{\Gamma\left(m_k + \frac{\alpha}{K}\right) \Gamma\left(N - m_k + 1\right)}{\Gamma\left(N + 1 + \frac{\alpha}{K}\right)}.$$

and

$$P([Z]) = \frac{K!}{\prod_{h=0}^{2^N-1} K_h!} \prod_{k=1}^{K} \frac{\alpha}{K} \frac{\Gamma\left(m_k + \frac{\alpha}{K}\right) \Gamma\left(N - m_k + 1\right)}{\Gamma\left(N + 1 + \frac{\alpha}{K}\right)},$$

respectively. Note that $K_h,\ h = 1, 2, \ldots, 2^N - 1$, is the number of times the row vector associated with the hth nonzero binary number appears in Z.

Solution: From the text and the definition of $P(Z)$, taking into account that the probabilities are beta distributed, we get

$$\begin{aligned} P(Z) &= \prod_{k=1}^{K} \int_0^1 P_k^{m_k} (1 - P_k)^{N - m_k} \frac{P_k^{\frac{\alpha}{K} - 1}}{B\left(\frac{\alpha}{K}, 1\right)} dP_k \\ &= \prod_{k=1}^{K} \frac{1}{\mathrm{B}\left(\frac{\alpha}{K}, 1\right)} \int_0^1 P_k^{m_k + \frac{\alpha}{K} - 1} \left(1 - P_k\right)^{N - m_k} dP_k. \end{aligned}$$

The above integrand is the normalizing constant of a Beta $\left(P_k | m_k + \frac{\alpha}{K}, N - m_k + 1\right)$ and we know that this is equal to

$$\mathrm{B}\left(m_k + \frac{\alpha}{K}, N - m_k + 1\right) = \frac{\Gamma\left(m_k + \frac{\alpha}{K}\right) \Gamma\left(N - m_k + 1\right)}{\Gamma\left(N + 1 + \frac{\alpha}{K}\right)}.$$

Thus,

$$\begin{aligned} P(Z) &= \prod_{k=1}^{K} \frac{\Gamma\left(\frac{\alpha}{K} + 1\right)}{\Gamma\left(\frac{\alpha}{K}\right) \Gamma(1)} \frac{\Gamma\left(m_k + \frac{\alpha}{K}\right) \Gamma\left(N - m_k + 1\right)}{\Gamma\left(N + 1 + \frac{\alpha}{K}\right)} \\ &= \prod_{k=1}^{K} \frac{\alpha}{K} \frac{\Gamma\left(m_k + \frac{\alpha}{K}\right) \Gamma\left(N - m_k + 1\right)}{\Gamma\left(N + 1 + \frac{\alpha}{K}\right)}. \end{aligned}$$

For $P\left([Z]\right)$, we have that,

$$P\left([Z]\right) = \sum_{Z \in [Z]} P(Z).$$

Let K be the number of total rows in Z. Let K_0 be the number of times the 0th row vector is present, K_1 is the respective number for the row vector associated with the binary 1, and so on. The total number of permutations of K objects, grouped in K_0, K_1,…,K_{2^N-1} groups is known from combinatorics to be equal to

$$\left(\begin{array}{c} K \\ K_0, K_1, \ldots, K_{2^N-1} \end{array} \right) = \frac{K!}{\prod_{h=0}^{2^N-1} K_h!}. \tag{52}$$

Hence,

$$P\left([Z]\right) = \frac{K!}{\prod_{h=0}^{2^N-1} K_h!} \prod_{k=1}^{K} \frac{\alpha}{K} \frac{\Gamma\left(m_k + \frac{\alpha}{K}\right) \Gamma\left(N - m_k + 1\right)}{\Gamma\left(N + 1 + \frac{\alpha}{K}\right)}.$$

The previously used combinatorics formula in (52) can be shown by the following arguments. Out of the K rows, K_0 are the zero ones. Then the total number of permutations of these zero rows are,

$$\left(\begin{array}{c} K \\ K_0 \end{array} \right) = \frac{K!}{K_0!(K-K_0)!}.$$

Now, for each one of the above permuted matrices, we make all possible permutations for the row vectors associated with K_1. This number amounts to,

$$\left(\begin{array}{c} K - K_0 \\ K_1 \end{array} \right) = \frac{(K-K_0)!}{K_1!(K-K_0-K_1)!}.$$

Thus, the total number of permutations involving K_0 and K_1 amounts to,

$$\frac{K!}{K_0!K_1!(K-K_0-K_1)!}.$$

Taking the above rationale recursively, the total number of permutations will be given by involving successive products as before, which it proves formula in Eq. (52).

13.25. Show that the discarded pieces, π_k, in the stick-breaking construction of an IBP are equal to the sequence of probabilities produced in a DP stick-breaking construction.

Solution: The sequence of the discarded segments is equal to

$$\begin{aligned} \pi_1 &= 1 - \beta_1, \\ \pi_2 &= \beta_1 - \beta_1\beta_2 = (1-\beta_2)\beta_1, \\ \ldots &= \ldots \\ \pi_k &= \prod_{j=1}^{k-1} \beta_j - \prod_{j=1}^{k} \beta_j = (1-\beta_k) \prod_{j=1}^{k-1} \beta_j \end{aligned}$$

Set $\beta^{'} := 1 - \beta$. Then, the previous recursions can be written as

$$\begin{aligned}
\pi_1 &= \beta^{'}_1, \\
\pi_2 &= \beta^{'}_2(1-\beta^{'}_1), \\
\ldots &= \ldots \\
\pi_k &= = \beta^{'}_k \prod_{j=1}^{k-1}(1-\beta^{'}_j).
\end{aligned}$$

The above sequence is the same with that in the DP construction. It suffices to show that $\beta^{'}$ follow a $\mathrm{Beta}(1, \alpha)$ distribution. Indeed,

$$\beta \sim \mathrm{Beta}(\alpha, 1) = \frac{\beta^{\alpha-1}}{\mathrm{B}(\alpha, 1)}.$$

Hence,

$$\beta^{'} \sim \frac{(1-\beta^{'})^{\alpha-1}}{\mathrm{B}(\alpha, 1)} = \mathrm{Beta}(1, \alpha),$$

which proves the claim.

Solutions To Problems of Chapter 14

14.1. Show that if $F_{\mathrm{x}}(x)$ is the cumulative distribution function of a random variable x, then the random variable $\mathrm{u} = F_{\mathrm{x}}(\mathrm{x})$ follows the uniform distribution in $[0, 1]$.

Solution: Let

$$\mathrm{u} = F_{\mathrm{x}}(\mathrm{x}), \ 0 \leq \mathrm{u} \leq 1.$$

Then, we have that

$$\begin{aligned} F_{\mathrm{u}}(u) &:= \mathrm{Prob}\{\mathrm{u} \leq u\} = \mathrm{Prob}\{F_{\mathrm{x}}(\mathrm{x}) \leq u\} = \mathrm{Prob}\{\mathrm{x} \leq F_{\mathrm{x}}^{-1}(u)\} \\ &= \mathrm{Prob}\{\mathrm{x} \leq x\} = F_{\mathrm{x}}(x) = u. \end{aligned}$$

Hence,

$$F_{\mathrm{u}}(u) = u$$

which implies that u is uniform in $[0, 1]$.

14.2. Show that if u follows the uniform distribution and

$$\mathrm{x} = F_{\mathrm{x}}^{-1}(\mathrm{u}) := g(\mathrm{u}), \tag{1}$$

then indeed x is distributed according to $F_{\mathrm{x}}(x) = \int_{-\infty}^{x} p(x)dx$.

Solution: Due to the monotonicity of $F_{\mathrm{x}}(x)$, for each x there exists a unique u. Let $\tilde{p}(x)$ be the pdf of the generated x by (1). We have that

$$\mathrm{x} = F_{\mathrm{x}}^{-1}(\mathrm{u}) \ \text{ or}$$

$$\mathrm{u} = F_{\mathrm{x}}(\mathrm{x}).$$

Hence,

$$\begin{aligned} 1 = p(u) &= \frac{\tilde{p}(x)}{|F_{\mathrm{x}}'(x)|} = \frac{\tilde{p}(x)}{p(x)} \Rightarrow \\ & \tilde{p}(x) = p(x), \end{aligned}$$

where we have used the rule of transforming random variables and that

$$\int_0^1 c du = 1 \Rightarrow c = 1.$$

14.3. Consider the random variables r and ϕ with exponential and uniform distributions

$$p_{\mathrm{r}}(r) = \frac{1}{2} \exp\left(-\frac{r}{2}\right), \ r \geq 0$$

and

$$p_{\phi}(\phi) = \begin{cases} \frac{1}{2\pi} & 0 \leq \phi \leq 2\pi \\ 0 & \text{otherwise,} \end{cases}$$

respectively. Show that the transformation

$$\begin{aligned}\mathrm{x} &= \sqrt{\mathrm{r}}\cos\phi = g_{\mathrm{x}}(\mathrm{r},\phi),\\ \mathrm{y} &= \sqrt{\mathrm{r}}\sin\phi = g_{\mathrm{y}}(\mathrm{r},\phi),\end{aligned}$$

renders both x and y to follow the normalized Gaussian $\mathcal{N}(0,1)$.

Solution: We have that

$$\begin{aligned}\mathrm{r} &= \mathrm{x}^2+\mathrm{y}^2 := g_{\mathrm{r}}(\mathrm{x},\mathrm{y}),\\ \phi &= \arctan(\frac{\mathrm{y}}{\mathrm{x}}) := g_{\phi}(\mathrm{x},\mathrm{y}).\end{aligned}$$

The respective Jacobian is

$$J(\mathrm{x},\mathrm{y};\mathrm{r},\phi) = \begin{bmatrix}\frac{\partial g_{\mathrm{x}}}{\partial r} & \frac{\partial g_{\mathrm{x}}}{\partial \phi}\\ \frac{\partial g_{\mathrm{y}}}{\partial r} & \frac{\partial g_{\mathrm{y}}}{\partial \phi}\end{bmatrix} = \begin{bmatrix}\frac{1}{2}r^{-\frac{1}{2}}\cos\phi & -r^{\frac{1}{2}}\sin\phi\\ \frac{1}{2}r^{-\frac{1}{2}}\sin\phi & r^{\frac{1}{2}}\cos\phi\end{bmatrix}$$

and

$$|J(\mathrm{x},\mathrm{y};\mathrm{r},\phi)| = \frac{1}{2}.$$

Note that

$$J(\mathrm{r},\phi;\mathrm{x},\mathrm{y}) = \begin{bmatrix}2x & \frac{-y}{x^2}\frac{1}{1+\left(\frac{y}{x}\right)^2}\\ 2y & \frac{1}{x}\frac{1}{1+\left(\frac{y}{x}\right)^2}\end{bmatrix}$$

and $|J(\mathrm{r},\phi;\mathrm{x},\mathrm{y})| = 2$. Hence,

$$\begin{aligned}p_{\mathrm{x},\mathrm{y}}(x,y) &= p_{\mathrm{r}\phi}\left(g_{\mathrm{r}}(x,y), g_{\phi}(x,y)\right)\frac{1}{1/2}\\ &= \frac{1}{2\pi}\frac{1}{2}\exp\left(-\frac{x^2+y^2}{2}\right)2\\ &= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{y^2}{2}\right).\end{aligned}$$

14.4. Show that if

$$p_{\mathrm{x}}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|0,I),$$

then $\mathbf{y}$ given by the transformation

$$\mathbf{y} = L\mathbf{x} + \boldsymbol{\mu}$$

is distributed according to

$$p_{\mathrm{y}}(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{\mu},\Sigma),$$

where $\Sigma = LL^T$.

Solution: We have that,

$$\boldsymbol{y} = L\boldsymbol{x} + \boldsymbol{\mu},$$

is a linear transformation and it is readily checked out that the Jacobian matrix

$$J(\mathbf{y};\mathbf{x}) = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_l} \\ \vdots & \vdots & \vdots \\ \frac{\partial y_l}{\partial x_1} & \cdots & \frac{\partial y_l}{\partial x_l} \end{bmatrix} = L.$$

Also,

$$\mathbf{x} = L^{-1}(\mathbf{y} - \boldsymbol{\mu}),$$

where the inverse is assumed to exist (Σ is invertible). Then

$$p_{\mathrm{y}}(\boldsymbol{y}) = \frac{p_{\mathrm{x}}\left(L^{-1}(\boldsymbol{y} - \boldsymbol{\mu})\right)}{\left|\det\bigl(J(\mathbf{y};\mathbf{x})\right|},$$

or

$$\begin{aligned} p_{\mathrm{y}}(\boldsymbol{y}) &= \frac{1}{|\det L|(2\pi)^{\frac{l}{2}}} \exp\left(-\frac{1}{2}\left(L^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right)^T \left(L^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right)\right) \\ &= \frac{1}{|\det L|(2\pi)^{\frac{l}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T L^{-T} L^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right) \\ &= \frac{1}{(2\pi)^{\frac{l}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T L^{-T} L^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right), \end{aligned}$$

where $|\Sigma|$ is the determinant of Σ and we used that

$$\begin{aligned} |\Sigma| &= |L||L^T| = |L|^2 \Rightarrow \\ |L| &= |\Sigma|^{\frac{1}{2}}. \end{aligned}$$

Recall that $|\Sigma| > 0$ since Σ is positive definite.

14.5. Consider two Gaussians

$$p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|0, \sigma_p^2 I),\ \sigma_p^2 = 0.1$$

and

$$q(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|0, \sigma_q^2 I),\ \sigma_q^2 = 0.11$$

$\boldsymbol{x} \in R^l$. In order to use $q(\boldsymbol{x})$ for drawing samples from $p(\boldsymbol{x})$, via the rejection sampling method, a constant c has to be computed so that

$$cq(\boldsymbol{x}) \geq p(\boldsymbol{x}).$$

Show that

$$c \geq \left(\frac{\sigma_q}{\sigma_p}\right)^l,$$

and compute the probability of accepting samples.

Solution: The maximum values for the two distributions occur at $x = 0$ and they are

$$p(0) = \frac{1}{(2\pi)^{\frac{l}{2}}\sigma_p^l} = \frac{1}{(2\pi)^{\frac{l}{2}}(0.10)^{\frac{l}{2}}},$$

and

$$q(0) = \frac{1}{(2\pi)^{\frac{l}{2}}\sigma_q^l} = \frac{1}{(2\pi)^{\frac{l}{2}}(0.11)^{\frac{l}{2}}}.$$

Note that $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ are very close, since $\sigma_p^2 \simeq \sigma_q^2$. If

$$c \geq \left(\frac{\sigma_q^2}{\sigma_p^2}\right)^{\frac{l}{2}} = (1.1)^{\frac{l}{2}},$$

then $cq(\boldsymbol{x}) \geq p(\boldsymbol{x})$, due to the broader shape of $q(\boldsymbol{x})$ (draw the graphs to see it). Let us choose c to have the smallest possible value,

$$c = (1.1)^{\frac{l}{2}}.$$

For large l, this can be a very large value indeed. The corresponding probability for acceptance is equal to

$$\text{Prob}\{\text{acceptance}\} = \frac{1}{c}\int p(\boldsymbol{x})d\boldsymbol{x} = \frac{1}{c}.$$

which for large values of l, can be very small.

14.6. Show that using importance sampling leads to an unbiased estimator for the normalizing constant of the desired distribution,

$$p(\boldsymbol{x}) = \frac{1}{Z}\phi(\boldsymbol{x}).$$

However, the estimator $\mathbb{E}[f(\mathbf{x})]$, of a function $f(\cdot)$ is a biased one.

Solution: We know that

$$\hat{Z} = \frac{1}{N}\sum_{i=1}^{N} w(\boldsymbol{x}_i)$$

or

$$\mathbb{E}[\hat{\mathrm{Z}}] = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[w(\mathbf{x}_i)].$$

However,

$$\begin{aligned}\mathbb{E}[w(\mathbf{x})] &= \int \left(\frac{\phi(\boldsymbol{x})}{q(\boldsymbol{x})}\right) q(\boldsymbol{x})d\boldsymbol{x} \\ &= \int \phi(\boldsymbol{x})d\boldsymbol{x} = Z.\end{aligned}$$

Hence $\mathbb{E}[\hat{\mathrm{Z}}] = \frac{1}{N} NZ = Z$. Also, since $\hat{\mathrm{Z}}$ is the sum of N unbiased variables divided by N, we know (Chapter 3) that its variance will be $\frac{\sigma_w^2}{N}$, where σ_w^2 is the variance of $w(x)$, provided σ_w^2 is finite. Thus, $\hat{Z}$ converges to Z as $N \longrightarrow \infty$.

On the other hand, we have

$$\widehat{\mathbb{E}[f(\mathbf{x})]} = \frac{\sum_{i=1}^N f(\boldsymbol{x}_i) w(\boldsymbol{x}_i)}{\sum_{i=1}^N w(\boldsymbol{x}_i)}.$$

For the denominator, we know that $\mathbb{E}[\sum_{i=1}^N w(\mathbf{x}_i)] = NZ$. For the numerator, we get

$$\begin{aligned} \mathbb{E}\left[\sum_{i=1}^N f(\mathbf{x}_i) w(\mathbf{x}_i)\right] &= \sum_{i=1}^N \left(\int f(\boldsymbol{x}) \frac{\phi(\boldsymbol{x})}{q(\boldsymbol{x})} q(\boldsymbol{x}) d\boldsymbol{x} \right) \\ &= NZ \cdot \mathbb{E}[f(\mathbf{x})] \end{aligned}$$

However, for finite number of N, this does not mean that although the mean values of the numerator and denominator are $NZ \cdot \mathbb{E}[f(\mathbf{x})]$ and NZ, respectively, their ratio is necessarily an unbiased estimator of $\mathbb{E}[f(\mathbf{x})]$. Only for $N \to \infty$, we obtain the unbiased estimator, when we divide by NZ, since $\hat{Z}$ converges to Z.

14.7. Let $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\mathbf{0}, \sigma_1^2 I)$. Choose the proposal distribution for importance sampling as

$$q(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\mathbf{0}, \sigma_2^2 I).$$

The weights are computed as

$$w(\boldsymbol{x}) = \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}.$$

If $w(\mathbf{0})$ is the weight at $\boldsymbol{x} = \mathbf{0}$, then the ratio $\frac{w(\boldsymbol{x})}{w(\mathbf{0})}$ is given by

$$\frac{w(\boldsymbol{x})}{w(\mathbf{0})} = \exp \frac{1}{2} \left(\frac{\sigma_1^2 - \sigma_2^2}{\sigma_1^2 \sigma_2^2} \sum_{i=1}^l x_i^2 \right).$$

Observe that even for very good match between $q(\boldsymbol{x})$ and $p(\boldsymbol{x})$ ($\sigma_1^2 \simeq \sigma_2^2$), for large values of l, the values of the weights can change significantly, due to the exponential dependence.

Solution: We have that

$$p(\boldsymbol{x}) = \frac{1}{(2\pi\sigma_1^2)^{l/2}} \prod_{i=1}^l \exp\left(-\frac{x_i^2}{2\sigma_1^2}\right),$$

and

$$q(\boldsymbol{x}) = \frac{1}{(2\pi\sigma_2^2)^{l/2}} \prod_{i=1}^{l} \exp\left(-\frac{x_i^2}{2\sigma_2^2}\right).$$

Hence,

$$\begin{aligned} \frac{w(\boldsymbol{x})}{w(\boldsymbol{0})} &= \left(\frac{\sigma_1}{\sigma_2}\right)^l \left(\frac{\sigma_2}{\sigma_1}\right)^l \prod_{i=1}^{l} \exp\left(\left(\frac{1}{2\sigma_2^2} - \frac{1}{2\sigma_1^2}\right) x_i^2\right) \\ &= \prod_{i=1}^{l} \exp\frac{1}{2}\left(\frac{\sigma_1^2-\sigma_2^2}{\sigma_1^2\sigma_2^2} x_i^2\right) \\ &= \exp\frac{1}{2}\left(\frac{\sigma_1^2-\sigma_2^2}{\sigma_1^2\sigma_2^2} \sum_{i=1}^{l} x_i^2\right). \end{aligned}$$

14.8. Show a stochastic matrix P has always the value $\lambda = 1$ as its eigenvalue.

Solution: By definition we have

$$P\boldsymbol{a} = \lambda\boldsymbol{a},$$

and λ is computed so that the matrix $P-\lambda I$ to be singular, i.e., $|P-\lambda I| = 0$. However, since the entries in each column of P add to one we have that adding all the rows of $P - I$ together leads to zeros. Hence the rows are linearly dependent, $P - I$ is singular and $\lambda = 1$ is an eigenvalue.

14.9. Show that if the eigenvalue of a transition matrix is not equal to one, its magnitude can not be larger than one, i.e., $|\lambda| \leq 1$.

Solution: Let λ be an eigenvalue such that $|\lambda| > 1$, with $\boldsymbol{a}$ its corresponding eigenvector,

$$P\boldsymbol{a} = \lambda\boldsymbol{a}.$$

Then we know that

$$P^n\boldsymbol{a} = \lambda^n\boldsymbol{a},$$

and P^n diverges, as $n \to \infty$. However, this is not possible, since P^n is a stochastic matrix such that

$$\boldsymbol{p}_n = P^n\boldsymbol{p}_0.$$

and all its elements are between 0 and 1.

14.10. Prove that if P is a stochastic matrix and $\lambda \neq 1$, then the elements of the corresponding eigenvector add to zero.

Solution: Let

$$P\boldsymbol{a} = \lambda\boldsymbol{a}, \; \lambda \neq 1,$$

then

$$P_{11}a_1 + \ldots + P_{1K}a_K = \lambda a_1$$

$$P_{K1}a_1 + \ldots + P_{KK}a_K = \lambda a_K$$

Adding together we have that

$$(a_1 + \ldots + a_K) = \lambda(a_1 + \ldots + a_K).$$

Since the entries in each column of P add to one. Hence, either $\lambda = 1$ or $\sum_{k=1}^{K} a_k = 0$.

14.11. Prove the square root dependence of the distance travelled by a random walk, with infinite many integer states, on the time, n.

Solution: We have that

$$\begin{aligned}
\mathbb{E}[\mathrm{x}_n^2] &= \sum_x x^2 p_n(x) = \sum_x x^2 \left[q p_{n-1}(x) + \rho p_{n-1}(x-1) + \rho p_{n-1}(x+1)\right] \\
&= q\sum_x x^2 p_{n-1}(x) + \rho\sum_x (x+1)^2 p_{n-1}(x) + \rho\sum_x (x-1)^2 p_{n-1}(x) \\
&= q\,\mathbb{E}[\mathrm{x}_{n-1}^2] + \rho\,\mathbb{E}[(\mathrm{x}_{n-1}+1)^2] + \rho\,\mathbb{E}[(\mathrm{x}_{n-1}-1)^2] \\
&= q\,\mathbb{E}[\mathrm{x}_{n-1}^2] + 2\rho\,\mathbb{E}[\mathrm{x}_{n-1}^2] + 2\rho \\
&= \mathbb{E}[\mathrm{x}_{n-1}^2] + 2\rho \quad \text{or} \\
\mathbb{E}[\mathrm{x}_n^2] &= 2\rho n + \mathbb{E}[x_0^2] = 2\rho n.
\end{aligned}$$

14.12. Prove, using the detailed balance condition, that the invariant distribution associated with the Markov chain implied by the Metropolis-Hastings algorithm is the desired distribution, $p(\boldsymbol{x})$.

Solution: By the definition of the respective kernel density we have

$$\begin{aligned}
p(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{y}) &= \big(q(\boldsymbol{x}|\boldsymbol{y})\alpha(\boldsymbol{x}|\boldsymbol{y}) + \delta(\boldsymbol{x}-\boldsymbol{y})r(\boldsymbol{x})\big)p(\boldsymbol{y}) \\
&= q(\boldsymbol{x}|\boldsymbol{y})\min\left\{1, \frac{q(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})}{q(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{y})}\right\}p(\boldsymbol{y}) + \delta(\boldsymbol{x}-\boldsymbol{y})r(\boldsymbol{x})p(\boldsymbol{y}) \\
&= \min\left\{q(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{y}), q(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})\right\} + \delta(\boldsymbol{x}-\boldsymbol{y})r(\boldsymbol{x})p(\boldsymbol{y}) \\
&= q(\boldsymbol{y}|\boldsymbol{x})\min\left\{1, \frac{q(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{y})}{q(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})}\right\}p(\boldsymbol{x}) + \delta(\boldsymbol{y}-\boldsymbol{x})r(\boldsymbol{y})p(\boldsymbol{x}) \\
&= \big(q(\boldsymbol{y}|\boldsymbol{x})\alpha(\boldsymbol{y}|\boldsymbol{x}) + \delta(\boldsymbol{y}-\boldsymbol{x})r(\boldsymbol{y})\big)p(\boldsymbol{x}) \\
&= \kappa(\boldsymbol{y}|\boldsymbol{x}),
\end{aligned}$$

which proves the claim.

14.13. Show that in Gibbs sampling, the desired joint distribution is invariant with respect to each one of the base transition pdfs.

Solution: By the definition of the base transition pdfs we have

$$B_d(\boldsymbol{x}|\boldsymbol{y}) = p(x(d)|\{y(i)\} : i \neq d) \prod_{i \neq d} \delta(y(i) - x(i)).$$

Hence

$$\begin{aligned}
\int B_d(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{y})d\boldsymbol{y} &= \int p(x(d)|\{y(i)\} : i \neq d) \prod_{i \neq d} \delta(y(i) - x(i))p(\boldsymbol{y})d\boldsymbol{y} \\
&= \int p(x(d)|\{y(i)\} : i \neq d)p(y(d)|\{y(i)\} : i \neq d) \times \\
&\quad p(\{y(i)\} : i \neq d) \prod_{i \neq d} \delta(y(i) - x(i))d\boldsymbol{y} \\
&= \int p(x(d)|\{x(i)\} : i \neq d)p(y(d)|\{x(i)\} : i \neq d) \times \\
&\quad p(\{x(i)\} : i \neq d)d(y(d)) \\
&= p(x(d)|\{x(i)\} : i \neq d)p(\{x(i)\} : i \neq d) \times \\
&\quad \int p(y(d)|\{x(i)\} : i \neq d)d(y(d)) \\
&= p(\boldsymbol{x}) \times 1 = p(\boldsymbol{x}),
\end{aligned}$$

which proves the claim.

14.14. Show that the acceptance rate for the Gibbs sampling is equal to one.

Solution: We know that the acceptance ratio for the Metropolis-Hastings algorithm is given by

$$\alpha(\boldsymbol{x}|\boldsymbol{x}_n) = \min\left\{1, \frac{q(\boldsymbol{x}_{n-1}|\boldsymbol{x})p(\boldsymbol{x})}{q(\boldsymbol{x}|\boldsymbol{x}_{n-1})p(\boldsymbol{x}_{n-1})}\right\}.$$

For the Gibbs sampling, the second term in the min operator becomes

$$\begin{aligned}
& \frac{p(x_{n-1}(d)|\{x(i)\} : i \neq d) \prod_{i \neq d} \delta(x_{n-1}(i) - x(i))}{p(x(d)|\{x_{n-1}(i)\} : i \neq d) \prod_{i \neq d} \delta(x(i) - x_{n-1}(i))} \\
& \frac{p(x(d)|\{x(i)\} : i \neq d)p(\{x(i)\} : i \neq d)}{p(x_{n-1}(d)|\{x_{n-1}(i)\} : i \neq d)p(\{x_{n-1}(i)\} : i \neq d)} \\
= & \frac{p(x_{n-1}(d)|\{x_{n-1}(i)\} : i \neq d)p(x(d)|\{x(i)\} : i \neq d)p(\{x_{n-1}(i)\} : i \neq d)}{p(x(d)|\{x(i)\} : i \neq d)p(x_{n-1}(d)|\{x_{n-1}(i)\} : i \neq d)p(\{x_{n-1}(i)\} : i \neq d)} \\
= & \ 1,
\end{aligned}$$

which proves the claim.

14.15. Derive the formulae for the conditional distributions of Section 14.11

Solution: The joint distribution of $n_0, \lambda_1, \lambda_2, x_{1:N}$ is given by

$$p(n_0, \lambda_1, \lambda_2, x_{1:N}) = \prod_{n=1}^{n_0} P(x_n|\lambda_1) \prod_{n=n_0+1}^{N} P(x_n|\lambda_2) p(\lambda_1) p(\lambda_2) \frac{1}{N}.$$

Taking the logarithm, we have

$$\begin{aligned} A := \ln p(n_0, \lambda_1, \lambda_2, x_{1:n}) &= \sum_{n=1}^{n_0} \ln P(x_n|\lambda_1) + \sum_{n=n_0+1}^{N} \ln P(x_n|\lambda_2) \\ &+ \ln p(\lambda_1) + \ln p(\lambda_2) - \ln(N) \end{aligned}$$

or

$$\begin{aligned} A = \quad & \sum_{n=1}^{n_0} (x_n \ln \lambda_1 - \lambda_1 - \ln(x_n!)) + \sum_{n=n_0+1}^{N} (x_n \ln \lambda_2 - \lambda_2 - \ln(x_n!)) + \\ & (a-1) \ln \lambda_1 - b\lambda_1 + a \ln b - \ln \Gamma(a) + \\ & (a-1) \ln \lambda_2 - b\lambda_2 + a \ln b - \ln \Gamma(a) - \ln N. \end{aligned}$$

To get the conditionals, it suffices to freeze the values of the rest of variables and consider them constants. For example,

$$\ln p(\lambda_1|n_0, \lambda_2, x_{1:n}) = \left(a - 1 + \sum_{n=1}^{n_0} x_n \right) \ln \lambda_1 - (n_0 + b)\lambda_1 + c_1,$$

where c_1 is a constant (normalizing). It is straightforward to see that the same result would occur by applying the Bayes theorem

$$p(\lambda_1|n_0, \lambda_2, x_{1:n}) = \frac{p(n_0, \lambda_1, \lambda_2, x_{1:n})}{p(n_0, \lambda_2, x_{1:n})}.$$

Hence

$$p(\lambda_1|n_0, \lambda_2, x_{1:n}) = \text{Gamma}(\lambda_1|a_1, b_1)$$

where

$$a_1 = a + \sum_{n=1}^{n_0} x_n, \quad b_1 = n_0 + b.$$

Similarly

$$\begin{aligned} p(\lambda_2|n_0, \lambda_1, x_{1:n}) &= \text{Gamma}(\lambda_2|a_2, b_2) \\ a_2 &= a + \sum_{n=n_0+1}^{N} x_n, \quad b_2 = N - n_0 + b. \end{aligned}$$

For n_0, we have

$$P(n_0|\lambda_1, \lambda_2, x_{1:n}) = \ln \lambda_1 \sum_{n=1}^{n_0} x_n - n_0\lambda_1 + \ln \lambda_2 \sum_{n_0+1}^{N} x_n - (N - n_0)\lambda_2 + c_3$$

for $n_0 = 1, 2, \ldots, N$. Note that this is a discrete distribution, providing the probability for each n_0, given the observations and λ_1, λ_2.

Solutions To Problems of Chapter 15

15.1. Show that in the product

$$\prod_{i=1}^{n}(1-x_i)$$

the number of cross product terms, $x_1x_2\cdots x_k,\ 1 \leq k \leq n$, for all possible combinations of $x_1, \cdots, x_n$ is equal to $2^n - n - 1$.

Solution: The claim is true for $n = 1$. Also, the total number of non-product terms is equal to n, excluding the term 1.

Assume that the claim is true for $n-1$. That is, the number of products in $\prod_{i=1}^{n-1}(1-x_i)$ is $2^{n-1}-(n-1)-1 = 2^{n-1}-n$. Then the number of products in

$$\prod_{i=1}^{n}(1-x_i) = \left(\prod_{i=1}^{n-1}(1-x_i)\right)(1-x_n)$$

will be $2(2^{n-1}-n)+n-1 = 2^n-n-1$, which proves the claim.
Take as examples:

- $(1-x_1)(1-x_2) = 1-x_1-x_2+x_1x_2:\ 2^2-2-1=1.$
- $(1-x_1)(1-x_2)(1-x_3) = (1-x_1-x_2+x_1x_2)(1-x_3) = 1-x_1-x_2+x_1x_2-x_3+x_1x_3+x_2x_3-x_1x_2x_3:\ 2^3-3-1=4.$

15.2. Prove that if a probability distribution p satisfies the Markov condition, as implied by a BN, then p is given as the product of the conditional distributions given the values of the parents.

Solution: We will prove it for the case of discrete probabilities, but similar is the proof for pdfs, as well. Let the involved random variables be $\mathrm{x}_1, \mathrm{x}_2, \cdots, \mathrm{x}_l$. We order them in the so called ancestral ordering. That is, every descendant variable comes after its parents. Let $\mathrm{x}_1, \mathrm{x}_2, \cdots, \mathrm{x}_l$ be the resulting ordering. We will show that

$$P(x_l, x_{l-1}, \cdots, x_1) = P(x_l|\mathrm{Pa}_l)P(x_{l-1}|\mathrm{Pa}_{l-1})\cdots P(x_1|\mathrm{Pa}_1).$$

For the proof, induction will be used.

(a) It is true for the first node for which $\mathrm{Pa}_1 = 0$

$$P(x_1|\mathrm{Pa}_1) = P(x_1)$$

(b) Assume that it is true up to the value $l-1$, i.e.,

$$P(x_{l-1}, \cdots, x_1) = \prod_{i=1}^{l-1} P(x_i|\mathrm{Pa}_i)$$

(c) Using Bayes theorem we get

$$\begin{aligned}
P(x_l, \cdots x_1) &= P(x_l|x_{l-1}, \cdots x_1)P(x_{l-1}, \cdots, x_1) \\
&= P(x_l|x_{l-1}, \cdots, x_1)\prod_{i=1}^{l-1} P(x_i|\mathrm{Pa}_i), \\
&= \prod_{i=1}^{l} P(x_l|\mathrm{Pa}_l)\prod_{i=1}^{l-1} P(x_i|\mathrm{Pa}_i),
\end{aligned}$$

which proves the claim.

15.3. Show that if a probability distribution factorizes according to a Bayesian network structure, then it satisfies the Markov condition.

Solution: The proof is the same with that in the second part of the next problem.

15.4. Consider a DAG and associate each node with a random variable. Define for each node the conditional probability of the respective variable given the values of its parents. Show that the product of the conditional probabilities yield a valid joint probability and that the Markov condition is satisfied.

Solution: To prove that the product, P, is a valid probability, we have to prove that $0 \leq P \leq 1$ and that

$$A := \sum_{x_1} \cdots \sum_{x_l} P(x_l|x_{l-1}, \cdots, x_1) = 1$$

Ordering the variables in ancestral order, we have

$$\begin{aligned}
A &:= \sum_{x_1} \cdots \sum_{x_l} P(x_1|\mathrm{Pa}_1)P(x_2|\mathrm{Pa}_2)\cdots P(x_l|\mathrm{Pa}_l) \\
&= \sum_{x_1} \cdots \sum_{x_{l-1}} P(x_1|\mathrm{Pa}_1)\cdots P(x_{l-1}|\mathrm{Pa}_{l-1})\sum_{x_l} P(x_l|\mathrm{Pa}_l) \\
&= \sum_{x_1} \cdots \sum_{x_{l-2}} P(x_1|\mathrm{Pa}_1)\cdots P(x_{l-2}|\mathrm{Pa}_{l-1})\sum_{x_{l-1}} P(x_{l-1}|\mathrm{Pa}_{l-1}) \\
&= 1
\end{aligned}$$

since the most right hand side summation always results in 1, due to the ancestral ordering. Also

$$P(x_l|\mathrm{Pa}_l)P(x_{l-1}|\mathrm{Pa}_{l-1})\cdots P(x_1|\mathrm{Pa}_1) \geq 0$$

since it is the product of non negative terns. To prove that the Markov condition holds, consider a node x_k. Order the variables so that all (and only) non-descendants precede x_k in the ordering. Thus

$$ND_k = \{\mathrm{x}_1, \mathrm{x}_2, \cdots, \mathrm{x}_{k-1}\}$$

and the set of descendants is given by

$$D_k = \{\mathrm{x}_{k+1}, \cdots, \mathrm{x}_l\}$$

Then we have

$$\begin{aligned} P(x_k|ND_k) &= \frac{P(x_k, ND_k)}{P(ND_k)} \\ &= \frac{\sum_{x_{k+1}} \cdots \sum_l \prod_{i=1}^l P(x_i|\mathrm{Pa}_i)}{\sum_{x_k} \cdots \sum_l \prod_{i=1}^l P(x_i|\mathrm{Pa}_i)} := \frac{A}{B} \end{aligned}$$

The numerator A becomes

$$\begin{aligned} A &= P(x_k|\mathrm{Pa}_k) \cdots P(x_1|\mathrm{Pa}_1) \sum_{x_{k+1}} \cdots \sum_{x_l} P(x_{k+1}|\mathrm{Pa}_{k+1}) \cdots P(x_l|\mathrm{Pa}_l) \\ &= P(x_k|\mathrm{Pa}_k) \cdots P(x_1|Pa_1) \end{aligned}$$

since all nodes involved in the summation form a subgraph which is also DAG and according to the first part of the problem the product of the conditional probabilities is the joint pdf of the involved variables. Following the same reasoning, we can show that

$$B = P(x_{k-1}|\mathrm{Pa}_{k-1}) \cdots P(x_1|Pa_1).$$

Hence,

$$P(x_k|ND_k) = P(x_k|\mathrm{Pa}_k),$$

which proves the claim.

15.5. Consider the graph in Figure 1. The r.v. x has two possible outcomes, with probabilities $P(x_1) = 0.3$ and $P(x_2) = 0.7$. Variable y has three possible outcomes with conditional probabilities,

$$\begin{aligned} &P(y_1|x_1) = 0.3,\ P(y_2|x_1) = 0.2,\ P(y_3|x_1) = 0.5, \\ &P(y_1|x_2) = 0.1,\ P(y_2|x_2) = 0.4,\ P(y_3|x_2) = 0.5. \end{aligned}$$

Finally, the conditional probabilities for z are

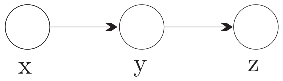

Figure 1: Graphical Model for Problem 15.5

$$\begin{aligned} &P(z_1|y_1) = 0.2,\ P(z_2|y_1) = 0.8, \\ &P(z_1|y_2) = 0.2,\ P(z_2|y_2) = 0.8, \\ &P(z_1|y_3) = 0.4,\ P(z_2|y_3) = 0.6. \end{aligned}$$

Show that this probability distribution, which factorizes over the graph, has x and z independent. However, x and z in the graph are not d–separated, since y is not instantiated.

Solution: It must be shown that $P(z|x) = P(z)$ for all combinations of the values z_1, z_2, x_1, x_2. We will sketch the derivation. Let

$$\begin{aligned}
P(z_1|x_1) &= \frac{P(z_1,x_1)}{P(x_1)} \\
&= \frac{P(z_1,y_1,x_1)+P(z_1,y_2,x_1)+P(z_1,y_3,x_1)}{P(x_1)} \\
&= P(y_1|x_1)P(z_1|y_1)+P(y_2|x_1)P(z_1|y_2) \\
&\quad +P(y_3|x_1)P(z_1|y_3) \\
&= 0.06+0.04+0.20 \\
&= 0.3.
\end{aligned}$$

Also,

$$P(z_1) = \sum_x \sum_y P(x,y,z_1) = \sum_x \sum_y P(z_1|y)P(y|x)P(x).$$

Carrying on the computations as before we obtain again 0.3. The same is true for all the other combinations.

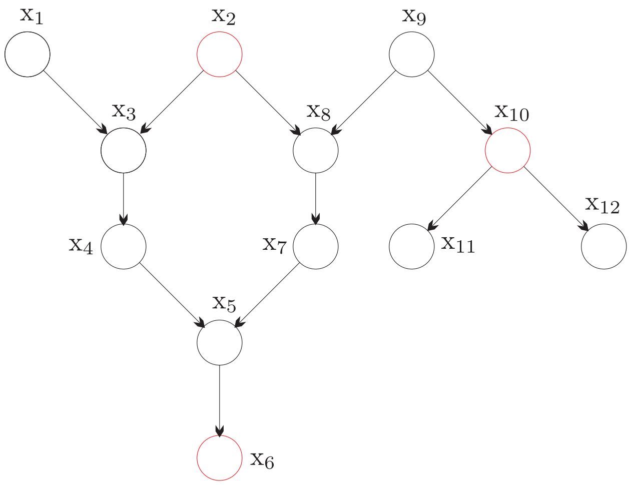

Figure 2: DAG for Problem 15.6. Nodes in red have been instantiated.

15.6. Consider the DAG in Figure 2. Detect the d–separations and d–connections in the graph.

Solution: Nodes x_{11} and x_{12} are d–separated from all the others, due to instantiation of x_{10} and the diverging connection. Then the rest of the uninstantiated nodes are d–connected. From x_1 evidence flows from x_3 and then to x_4 (unblocked serial connection the other path is blocked by x_2). Then it passes via x_5 to x_7 (converging node but the descendant of x_5 is instantiated). Then to x_8 and x_9 (serial unblocked connection). Recall from the theory that a chain of nodes is active if there are no instantiated nodes in it, and whenever a converging connection *along* the chain is present, $\mathrm{x}_{i-1} \rightarrow \mathrm{x}_i \leftarrow \mathrm{x}_{i+1}$, then a descendant node of x_i is instantiated.

15.7. Consider the DAG of Figure 3. Detect the blanket of node x_5 and verify that if all the nodes in the blanket are instantiated, then the node becomes *d*–separated from the rest of the nodes in the graph.

Solution: The blanket of x_5 comprises the nodes x_1,x_2, x_8,x_9 (children) and x_4,x_6 (share children with x_5). Check that any chain that connects x_5 with any node outside the blanket, once nodes in the blanket have been instantiated, is blocked.

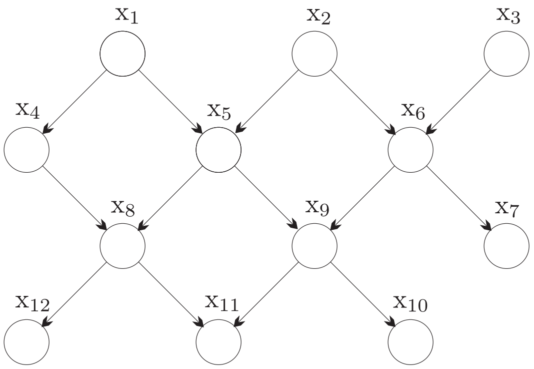

Figure 3: The graph structure for Problem 15.7

15.8. In a linear Gaussian Bayesian network model, derive the mean values and the respective covariance matrices for each one of the variables in a recursive manner

Solution: Since

$$P(x_i|\mathrm{Pa}_i) = \mathcal{N}\left(x_i|\sum_{k:\mathrm{x}_k\in\mathrm{Pa}_i} \theta_{ik}x_k + \theta_{i0}, \sigma_i^2\right),$$

we get that

$$\mathrm{E}[\mathrm{x}_i] = \sum_{k:\mathrm{x}_k\in\mathrm{pa}_i} \theta_{ik}\mathrm{E}[\mathrm{x}_k] + \theta_{i0}.$$

Thus, we can arrange our variables in ancestral order and compute the mean values recursively.

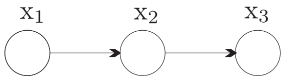

Figure 4: Network for Problem 15.9

For the computation of the covariance matrix, recall that

$$\mathrm{x}_i = \sum_{k:\mathrm{x}_k \in \mathrm{pa}_i} \theta_{ik}\mathrm{x}_k + \theta_{i0} + \sigma_i \mathrm{n}_i,$$

where n_i is a white (zero-mean) noise Gaussian source. Having adopted ancestral ordering and considering, say, x_i, we are certain that x_i is not involved in any one of the parent sets, $\mathrm{Pa}_1, \mathrm{Pa}_2, \cdots, \mathrm{Pa}_{i-1}$. We will derive the covariances inductively.

- We know σ_i^2
- Assume that we have computed all the covariances

$$\mathrm{cov}(\mathrm{x}_j, \mathrm{x}_r), \ \forall r, j : \ j < i \text{ and } r < i$$

- Based on the previous computations, we will obtain all the covariances

$$\mathrm{cov}(\mathrm{x}_i, \mathrm{x}_j), \ j \leq i.$$

For $j = 1, 2, \cdots, i-1$ we have,

$$\mathrm{E}\left[(\mathrm{x}_i - \mathrm{E}[\mathrm{x}_i])(\mathrm{x}_j - \mathrm{E}[\mathrm{x}_j])\right] =$$

$$\mathrm{E}\left[(\mathrm{x}_j - \mathrm{E}[\mathrm{x}_j])(\sum_{k:\mathrm{x}_k \in \mathrm{Pa}_i} \theta_{ik}\mathrm{x}_k + \theta_{i0} - \sum_{k:\mathrm{x}_k \in \mathrm{Pa}_i} \theta_{ik} E[\mathrm{x}_k] - \theta_{i0} + \sigma_i \mathrm{n}_i)\right] =$$

$$\mathrm{E}\left[(\mathrm{x}_j - \mathrm{E}[\mathrm{x}_j])(\sum_{k:\mathrm{x}_k \in \mathrm{Pa}_i} \theta_{ik}(\mathrm{x}_k - E[\mathrm{x}_k]) + \sigma_i \mathrm{n}_i)\right] =$$

$$\sum_{k:\mathrm{x}_k \in \mathrm{Pa}_i} \theta_{ik}\mathrm{E}\left[(\mathrm{x}_j - \mathrm{E}[\mathrm{x}_j])(\mathrm{x}_k - \mathrm{E}[\mathrm{x}_k])\right] + 0, \quad i \neq j,$$

and finally

$$\mathrm{cov}(\mathrm{x}_i, \mathrm{x}_j) = \sum_{k:\mathrm{x}_k \in \mathrm{Pa}_i} \theta_{ik}\mathrm{cov}(\mathrm{x}_j, \mathrm{x}_k) + \delta_{ij}\sigma_i^2, \quad j \leq i,$$

which results in a hierarchical computation of the covariances.

15.9. Assuming the variables associated with the nodes of the Bayesian structure of Figure 4 to be Gaussian, find the respective mean values and covariances.

Solution: Following the recursive manner of computation, as suggested in Problem 15.8, we have

$$\begin{aligned}
\mathrm{E}[\mathrm{x}_1] &= \theta_{10}, \\
\mathrm{E}[\mathrm{x}_2] &= \theta_{21}\mathrm{E}[\mathrm{x}_1] + \theta_{20} = \theta_{21}\theta_{10} + \theta_{20}, \\
\mathrm{E}[\mathrm{x}_3] &= \theta_{32}\mathrm{E}[\mathrm{x}_2] + \theta_{30}, \\
&= \theta_{32}\left(\theta_{21}\theta_{10} + \theta_{20}\right) + \theta_{30}, \\
&= \theta_{32}\theta_{21}\theta_{10} + \theta_{32}\theta_{20} + \theta_{30}.
\end{aligned}$$

For the covariances we have,

- $\mathrm{cov}(\mathrm{x}_1, \mathrm{x}_1) = \sigma_1^2$
- $\mathrm{cov}(\mathrm{x}_2, \mathrm{x}_1) = \theta_{21}\sigma_1^2$, $\mathrm{cov}(\mathrm{x}_2, \mathrm{x}_2) = \theta_{21}^2\sigma_1^2 + \sigma_2^2$
- $\mathrm{cov}(\mathrm{x}_3, \mathrm{x}_1) = \theta_{32}\mathrm{cov}(\mathrm{x}_2, \mathrm{x}_1) = \theta_{32}\theta_{21}\sigma_1^2$
- $\mathrm{cov}(\mathrm{x}_3, \mathrm{x}_2) = \theta_{32}\mathrm{cov}(\mathrm{x}_2, \mathrm{x}_2) = \theta_{32}\left(\theta_{21}^2\sigma_1^2 + \sigma_2^2\right)$
- $\mathrm{cov}(\mathrm{x}_3, \mathrm{x}_3) = \theta_{32}\mathrm{cov}(\mathrm{x}_3, \mathrm{x}_2) + \sigma_3^2 = \theta_{32}^2\left(\theta_{21}^2\sigma_1^2 + \sigma_2^2\right) + \sigma_3^2,$
- $\Sigma = \begin{bmatrix} \sigma_1^2 & \theta_{21}\sigma_1^2 & \theta_{32}\theta_{21}\sigma_1^2 \\ \theta_{21}\sigma_1^2 & \theta_{21}^2\sigma_1^2 + \sigma_2^2 & \theta_{32}\left(\theta_{21}^2\sigma_1^2 + \sigma_2^2\right) \\ \theta_{32}\theta_{21}\sigma_1^2 & \theta_{32}\left(\theta_{21}^2\sigma_1^2 + \sigma_2^2\right) & \theta_{32}^2\left(\theta_{21}^2\sigma_1^2 + \sigma_2^2\right) + \sigma_3^2 \end{bmatrix}.$

15.10. Prove that if p is a Gibbs distribution that factorizes over a MRF H, then H is an I-map for p.

Solution: Let X,Y,Z be any three disjoint sets such as Z separates X and Y. Then we will show that for any $\mathrm{x} \in X$ and $\mathrm{y} \in Y$, $\mathrm{x} \perp \mathrm{y}|Z$.

Let us start by assuming that $X \cup Y \cup Z = \mathcal{X}$, where $\mathcal{X}$ is the set of all points. Thus, any clique in H will be contained either in $X \cup Z$ or $Y \cup Z$, since there are no edges connecting X and Y directly, without passing via Z. Then, since p factorizes over H, it will be given as a product of potential functions over cliques such as

$$p\left(x_1, x_2, \ldots, x_l\right) = \frac{1}{Z} \prod_{c_i \in X \cup Z} \phi_{c_i}(\boldsymbol{x}_{c_i}) \prod_{c_j \in Z \cup Y} \phi_{c_j}(\boldsymbol{x}_{c_j}).$$

Thus, we can write that

$$p\left(x_1, x_2, \ldots, x_l\right) = \frac{1}{Z} f\left(X, Z\right) g\left(Y, Z\right),$$

where the first factor contains variables in X,Z and the second one variables in Y,Z. This implies that

$$p(x, y|z) = p(x|z)p(y|z),$$

i.e., independence among the variables in X and Y.

If $X \cup Y \cup Z \subset \mathcal{X}$, let $\Psi = \mathcal{X} - (X \cup Y \cup Z)$. Partition Ψ into two disjoint groups, such as Z separates $X \cup \Psi_1$ from $Y \cup \Psi_2$. Apply the previous result on $X \cup \Psi_1$ and $Y \cup \Psi_2$, and then prove that $\forall \mathrm{x} \in X \cup \Psi_1$ and $\forall \mathrm{y} \in Y \cup \Psi_2$, then

$$\mathrm{x} \perp \mathrm{y} | Z.$$

Then we readily have the claim.

15.11. Show that if H is the moral graph that results from moralization of a BN structure then

$$I(H) \subseteq I(G)$$

Solution: In an undirected graph, independencies are guaranteed by respective separations among nodes, given a set of other nodes. Take any node in the moral graph. The only separations that can be guaranteed of this node from any other node are those imposed by the nodes comprising the respective Markov blanket of the node in the original DAG. However, these independencies, is a subset of the set of d–separations of the original DAG. This proves the claim.

15.12. Consider a Bayesian network structure and a probability distribution p. Then show that if $I(G) \subseteq I(p)$, then p factorizes over G.

Solution: The proof consists of a combination of two theorems. We know that, given a BN structure the set of all d–separations does include the local independencies, on which the Bayesian structure is built upon. Since $I(G) \subseteq I(p)$, then these local independencies imply conditional independencies on p. Hence from Theorem 15.1, we have that p factorizes over G.

15.13. Show that in a undirected chain graphical model, the marginal probability $P(x_j)$ of a node, x_j, is given by

$$P(x_j) = \frac{1}{Z}\mu_f(x_j)\mu_b(x_j),$$

where $\mu_f(x_j)$ and $\mu_b(x_j)$ are the received by the node forward and backward messages.

Solution: We have that

$$P(x_j) = \frac{1}{Z}\sum_{x_1}\cdots\sum_{x_{j-1}}\sum_{x_{j+1}}\cdots\sum_{x_l}\prod_{i=1}^{l-1}\psi_{i,i+1}(x_i, x_{i+1}),$$

assuming l nodes, or

$$\begin{aligned} P(x_j) &= \frac{1}{Z}\sum_{x_1}\cdots\sum_{x_{j-1}}\prod_{i=1}^{j-1}\psi_{i,i+1}(x_i, x_{i+1})\sum_{x_{j+1}}\cdots\sum_{x_l}\prod_{i=j}^{l-1}\psi_{i,i+1}(x_i, x_{i+1}) \\ &= \frac{1}{Z}\mu_f(x_j)\mu_b(x_j), \end{aligned}$$

by the respective definitions.

15.14. Show that the joint distribution of two neighboring nodes in a undirected chain graphical model is given by

$$P(x_j, x_{j+1}) = \frac{1}{Z}\mu_f(x_j)\psi_{j,j+1}(x_j, x_{j+1})\mu_b(x_{j+1}).$$

Solution: Following the same steps as for the problem 15.13, we have

$$\begin{aligned}
P(x_j, x_{j+1}) &= \frac{1}{Z}\sum_{x_1}\cdots\sum_{x_{j-1}}\sum_{x_{j+2}}\cdots\sum_{x_l}\prod_{i=1}^{l-1}\psi_{i,i+1}(x_i, x_{i+1}) \\
&= \frac{1}{Z}\sum_{x_1}\cdots\sum_{x_{j-1}}\prod_{i=1}^{j-1}\psi_{i,i+1}(x_i, x_{i+1})\psi_{j,j+1}(x_j, x_{j+1}) \\
&\quad \sum_{x_{j+2}}\cdots\sum_{x_l}\prod_{i=j+1}^{l-1}\psi_{i,i+1}(x_i, x_{i+1}) \\
&= \frac{1}{Z}\mu_f(x_j)\psi_{j,j+1}(x_j, x_{j+1})\mu_b(x_{j+1}).
\end{aligned}$$

15.15. Using Figure 15.26, prove that if there is a second message passing, starting from x, towards the leaves, then any node will have the available information for the computation of the respective marginals.

Solution: Let us consider node x_4. The proof is similar for all the other nodes. In order to compute $P(x_4)$, we need $\mu_{f_d \to x_4}(x_4)$, $\mu_{f_e \to x_4}(x_4)$ and $\mu_{f_a \to x_4}(x_4)$, as Equation (15.55) suggests. Messages $\mu_{f_d \to x_4}(x_4)$ and $\mu_{f_e \to x_4}(x_4)$ have already been computed during the first pass. When message passing starts from x_1, then x_1 will propagate:

$$\mu_{x_1 \to f_a}(x_1) = \mu_{f_g \to x_1}(x_1)\mu_{f_h \to x_1}(x_1)$$

to factor node A. Although not of interest here, note that x_1 will also propagate

$$\mu_{x_1 \to f_g}(x_1) = \mu_{f_a \to x_1}(x_1)\mu_{f_h \to x_1}(x_1)$$

to factor node G and

$$\mu_{x_1 \to f_h}(x_1) = \mu_{f_g \to x_1}(x_1)\mu_{f_a \to x_1}(x_1)$$

to node H.

Thus, factor node A will propagate messages to $\mathrm{x}_2, \mathrm{x}_3$ and x_4 and so on. Focusing on x_4, the received message will be

$$\begin{aligned}
\mu_{f_a \to x_4}(x_4) &= \sum_{x_1}\sum_{x_2}\sum_{x_3} f_a(x_1, x_2, x_3, x_4)\mu_{x_2 \to f_a}(x_2)\mu_{x_3 \to f_a}(x_3) \\
&\quad \mu_{x_1 \to f_a}(x_1).
\end{aligned}$$

However, $\mu_{x_2 \to f_a}(x_2)$ and $\mu_{x_3 \to f_a}(x_3)$ have been computed during the first pass. Hence, all information for $P(x_4)$ is now available. The same is true for all nodes.

15.16. Consider the tree graph of Figure 15.26. Compute the marginal probability $P(x_1, x_2, x_3, x_4)$.

Solution: Let $V_{A'}$ be all the nodes of V_A excluding $\mathrm{x}_2, \mathrm{x}_3$ and x_4. Then following similar reasoning as for (15.46)

$$\begin{aligned} P(x_1, x_2, x_3, x_4) \quad &= \quad \frac{1}{Z} \sum_{\mathrm{x} \in V_{A'}} \sum_{\mathrm{x} \in V_F} \sum_{\mathrm{x} \in V_G} \Psi_A(x_1, x_2, x_3, x_4, \boldsymbol{x}_{A'}) \cdot \\ & \qquad \Psi_H(x_1, \boldsymbol{x}_H) \Psi_G(x_1, \boldsymbol{x}_G), \end{aligned}$$

or

$$\begin{aligned} P(x_1, x_2, x_3, x_4) \quad &= \quad \frac{1}{Z} \sum_{\boldsymbol{x}_{A'} \in V_{A'}} \Psi_A(x_1, x_2, x_3, x_4, \boldsymbol{x}_{A'}) \sum_{\boldsymbol{x}_H \in V_H} \Psi_H(x_1, \boldsymbol{x}_H) \\ & \qquad \sum_{\boldsymbol{x}_G \in V_G} \Psi_G(x_1, \boldsymbol{x}_G). \end{aligned}$$

Then, following exactly the same steps as the ones in order to prove Equation (15.54), we have that

$$\begin{aligned} \sum_{\boldsymbol{x}_H \in V_H} \Psi_F(x_1, \boldsymbol{x}_H) \quad &= \quad \mu_{f_h \to x_1}(x_1), \\ \sum_{\boldsymbol{x}_G \in V_G} \Psi_G(x_1, \boldsymbol{x}_G) \quad &= \quad \mu_{f_g \to x_1}(x_1). \end{aligned}$$

Furthermore, note that

$$\mu_{x_1 \to f_a}(x_1) = \mu_{f_h \to x_1}(x_1) \mu_{f_g \to x_1}(x_1).$$

Also, due to (15.48) (by neglecting the sum over x_2, x_3, x_4)

$$\begin{aligned} \sum_{\boldsymbol{x}_{A'} \in V_{A'}} \Psi_A(x_1, x_2, x_3, x_4, \boldsymbol{x}_{A'}) \quad &= \quad f_a(x_1, x_2, x_3, x_4) \mu_{x_2 \to f_a}(x_2) \\ & \qquad \mu_{x_3 \to f_a}(x_3) \mu_{x_4 \to f_a}(x_4). \end{aligned}$$

Thus, finally we obtain

$$\begin{aligned} P(x_1, x_2, x_3, x_4) \quad &= \quad \frac{1}{Z} f_a(x_1, x_2, x_3, x_4) \\ & \qquad \mu_{x_1 \to f_a}(x_1) \mu_{x_2 \to f_a}(x_2) \mu_{x_3 \to f_a}(x_3) \mu_{x_4 \to f_a}(x_4). \end{aligned}$$

15.17. Repeat the message procedure to find the optimal combination of variables for Example 15.4 using the logarithmic version and the max–sum algorithm

Solutions To Problems of Chapter 16

16.1. Prove that an undirected graph is triangulated if and only if its cliques can be organized into a join tree.

Solution: The proof follows [Jens 01].

a) Let the cliques be organized in a join tree. Let V be a leaf clique and let F be its unique neighbor. From the definition of a join tree and the running intersection property, if V has common nodes with any other clique, these nodes must also belong to F. Hence, there must be at least one node x $\in V$, such that x $\notin F$, otherwise $V \subset F$. Then we can eliminate this variable node, x, without adding any fill-ins, since x is a part of a single clique, namely V. We keep removing variable nodes such as x. Once we have completed it, we left with a graph, without V, which is also a join tree. Then the process continues and in this way we have generated a perfect elimination sequence. Hence the graph is triangulated.

b) Assume a triangulated graph. Then we will obtain a join tree by construction. We follow the algorithmic procedure of subsection 16.2.1. We must first prove that the generated structure by such a construction is a join tree. First, it is a tree, since every clique has at most one parent, and there cannot be multiple paths. To prove that it is a join tree, we have to show the running intersection property. Let V_i and V_j $(i < j)$ be two clique nodes sharing a set of variable nodes, X. We have proved that it is a tree, hence there is a unique path from V_i to V_j. The nodes in X cannot be eliminated from V_i (since they are shared with other cliques) and must be part of S_i and, by construction, part of V_i's parent, say, V_k. IF $j = k$, the proof is complete. If not, we continue the reasoning as before, with the min (k, j) clique.

16.2. For the graph of Figure 16.3a give all possible perfect elimination sequences and draw the resulting sequence of graphs.

Solution: Alternative perfect elimination sequences are:

$$x_5, x_6, x_3, x_1, x_2, x_4,$$

or

$$x_1, x_5, x_6, x_3, x_2, x_4,$$

or

$$x_6, x_1, x_3, x_5, x_2, x_4.$$

The sequence of the resulting graphs for the first one, is shown in Figure 1

16.3. Derive the formulas for the marginal probabilities of the variables in a) a clique node and b) in a separator node in a junction tree.

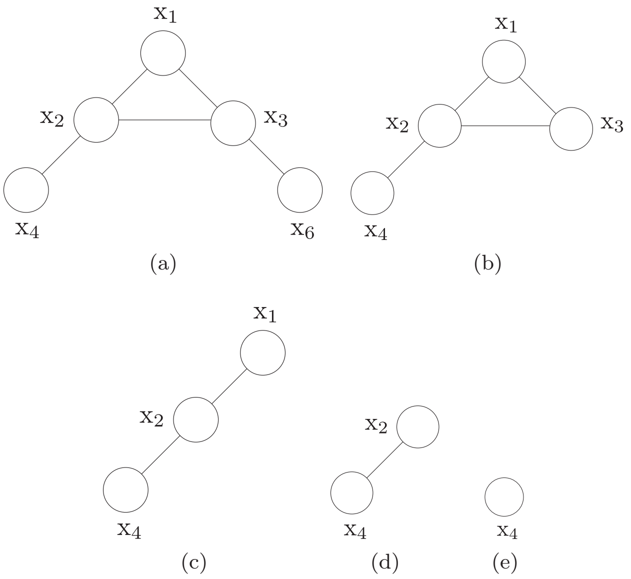

Figure 1: The resulting graphs associated with the elimination sequence: $x_5, x_6, x_3, x_1, x_2, x_4$, for Problem 16.2.

Solution: The proof is direct consequence of the definition of messages. Let $\psi_V(\boldsymbol{x}_V)$ be a clique of variable (nodes) which comprises $\boldsymbol{x}_V$. Some of them are not shared by any other clique, and some of them are. The shared ones will appear in the separators with whom V is connected to. We have

$$P(\boldsymbol{x}_V) = \frac{1}{Z} \sum_{\boldsymbol{x} \setminus \boldsymbol{x}_V} \prod_{c} \psi_c(\boldsymbol{x}_c),$$

where $\boldsymbol{x}$ is the set of all variables and c runs over all cliques. We can rewrite as

$$P(\boldsymbol{x}_V) = \frac{1}{Z} \psi_V(\boldsymbol{x}_V) \sum_{\boldsymbol{x} \setminus \boldsymbol{x}_V} \prod_{c \setminus V} \psi_c(\boldsymbol{x}_c).$$

However, the summation over the products is nothing but the product of the messages sent by the separators (which V is connected to) to V. Indeed, the message passed to a separator is the marginalization over all variables (except those in the separator) in the products of the factors along the path from a leaf to the separator. Similar arguments hold for the joint probability of the variables in a separator.

16.4. Prove that in a junction tree the joint pdf of the variables is given by (16.8).

Solution: We will show it first for a junction tree comprising two clique

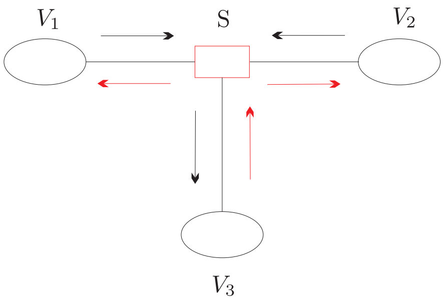

Figure 2: The graph structures for Problem 16.4. The arrows indicate two possible message flows (black and red) in different directions, which are required by the message passing algorithm.

nodes and one separator and then three clique nodes. Then the generalization is straightforward. The respective graphs are shown in Figure 2.
a) We have that

$$P(\boldsymbol{x}) = \frac{1}{Z}\psi_1(\boldsymbol{x}_{v_1})\psi_2(\boldsymbol{x}_{v_2}) \tag{1}$$

However, from (16.6) we get

$$P(\boldsymbol{x}_{v_1}) = \frac{1}{Z}\psi_1(\boldsymbol{x}_{v_1})\mu_{s\to v_1}(\boldsymbol{x}_s), \tag{2}$$

and

$$P(\boldsymbol{x}_{v_2}) = \frac{1}{Z}\psi_2(\boldsymbol{x}_{v_2})\mu_{s\to v_2}(\boldsymbol{x}_s). \tag{3}$$

Also, from (16.7) we get

$$P(\boldsymbol{x}_s) = \frac{1}{Z}\mu_{v_1\to s}(\boldsymbol{x}_s)\mu_{v_2\to s}(\boldsymbol{x}_s) \tag{4}$$

Moreover, take into account that

$$\mu_{v_1\to s}(\boldsymbol{x}_s) = \mu_{s\to v_2}(\boldsymbol{x}_s) \;\text{ and }\; \mu_{v_2\to s}(\boldsymbol{x}_s) = \mu_{s\to v_1}(\boldsymbol{x}_s). \tag{5}$$

Combining Equations (1)–(5) we obtain the result. The reason Z is eliminated, is that we have one separator and two clique nodes.

b) For the three clique nodes we have

$$P(\boldsymbol{x}) = \frac{1}{Z}\psi_1(\boldsymbol{x}_{v_1})\psi_2(\boldsymbol{x}_{v_2})\psi_3(\boldsymbol{x}_{v_3}), \tag{6}$$

and similarly

$$P(\boldsymbol{x}_{v_i}) = \frac{1}{Z}\psi_1(\boldsymbol{x}_{v_i})\mu_{s\to v_i}(\boldsymbol{x}_s), \quad i = 1, 2, 3. \tag{7}$$

Also,

$$P(\boldsymbol{x}_s) = \frac{1}{Z}\mu_{v_1\to s}(\boldsymbol{x}_s)\mu_{v_2\to s}(\boldsymbol{x}_s)\mu_{v_3\to s}(\boldsymbol{x}_s) \tag{8}$$

where

$$\mu_{s\to v_1}(\boldsymbol{x}_s) = \mu_{v_3\to s}(\boldsymbol{x}_s)\mu_{v_2\to s}(\boldsymbol{x}_s), \tag{9}$$

$$\mu_{s\to v_2}(\boldsymbol{x}_s) = \mu_{v_1\to s}(\boldsymbol{x}_s)\mu_{v_3\to s}(\boldsymbol{x}_s), \tag{10}$$

$$\mu_{s\to v_3}(\boldsymbol{x}_s) = \mu_{v_2\to s}(\boldsymbol{x}_s)\mu_{v_1\to s}(\boldsymbol{x}_s). \tag{11}$$

Combining (6)–(11), we obtain

$$\begin{aligned} P(\boldsymbol{x}) &= \frac{1}{Z}\Bigg(\frac{ZP(\boldsymbol{x}_{v_1})}{\mu_{v_3\to s}(\boldsymbol{x}_s)\mu_{v_2\to s}(\boldsymbol{x}_s)}\frac{ZP(\boldsymbol{x}_{v_2})}{\mu_{v_1\to s}(\boldsymbol{x}_s)\mu_{v_3\to s}(\boldsymbol{x}_s)} \\ &\quad \frac{ZP(\boldsymbol{x}_{v_3})}{\mu_{v_2\to s}(\boldsymbol{x}_s)\mu_{v_1\to s}(\boldsymbol{x}_s)}\Bigg) \\ &= \frac{P(\boldsymbol{x}_{v_1})P(\boldsymbol{x}_{v_2})P(\boldsymbol{x}_{v_3})}{(P_s(\boldsymbol{x}_s))^2}. \end{aligned}$$

16.5. Show that obtaining the marginal over a single variable is independent of which one from the cliques/separators nodes, which contain the variable, the marginalization is performed.
Hint: Prove it for the case of two neighboring clique nodes in the junction tree.

Solution: We will consider that the two neighboring nodes V_1, V_2 share one variable x, which also comprises their common separator, as shown in Figure 3. Assume for simplicity that V_1 is a leaf node, without harming generality. We have that

$$\begin{aligned} P(x) &= \sum_{\boldsymbol{x}_{V_1}} \psi_1(\boldsymbol{x}_{V_1}, x)\mu_{V_2\to S}(x) \\ &= \sum_{\boldsymbol{x}_{V_1}} \psi_1(\boldsymbol{x}_{V_1}, x)\sum_{\boldsymbol{x}_{V_2}} \psi_2(\boldsymbol{x}_{V_2}, x)\prod_s \mu_{s\to V_2}(\boldsymbol{x}_s), \end{aligned}$$

where $\boldsymbol{x}_s \subset (\boldsymbol{x}_{V_2}, x)$. Now let

$$\begin{aligned} P'(x) &= \sum_{\boldsymbol{x}_{V_2}} \psi_2(\boldsymbol{x}_{V_2}, x)\prod_s \mu_{s\to V_2}(\boldsymbol{x}_s)\mu_{S\to V_2}(x) \\ &= \sum_{\boldsymbol{x}_{V_2}} \psi_2(\boldsymbol{x}_{V_2}, x)\prod_s \mu_{s\to V_2}(\boldsymbol{x}_s)\left(\sum_{\boldsymbol{x}_{V_1}} \psi_1(\boldsymbol{x}_{V_1}, x)\right) \\ &= P(x). \end{aligned}$$

The proof is similar if one considers a separator node to compute the marginal.

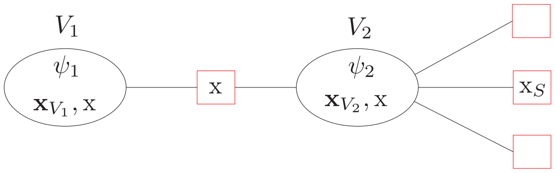
V_1
ψ_1
$\mathbf{x}_{V_1}, \mathrm{x}$
x
V_2
ψ_2
$\mathbf{x}_{V_2}, \mathrm{x}$
x_S

Figure 3: The set up for Problem 16.5

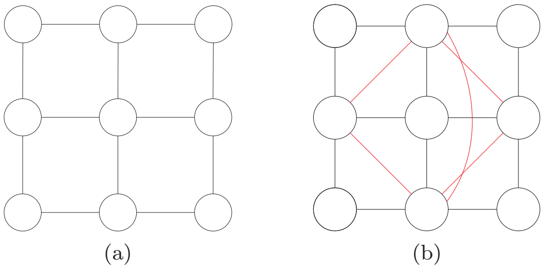

Figure 4: a) The graph for the Problem 16.6. b) A triangulated graph.

16.6. Consider the graph in Figure 4a. Obtain a triangulated version of it.

Solution: The triangulated version is shown in Figure 4b

16.7. Consider the Bayesian network structure given in Figure 5. Obtain a equivalent join tree.

Solution: First, we moralize the network, and the structure of Figure 6 results. Then we peel it off to form cliques and separators; as shown in Figure 7. Note that, at each time, nodes that belong to *exclusively* one clique can be removed. The resulting join tree is shown in Figure 8.

16.8. Consider the random variables A, B, C, D, E, F, G, H, I, J and assume that the joint distribution is given by the product of the following potential functions

$$p = \frac{1}{Z}\psi_1(A,B,C,D)\psi_2(B,E,D)\psi_3(E,D,F,I)\psi_4(C,D,G)\psi_5(C,H,G,J)$$

Construct an undirected graphical model on which the previous joint probability factorizes and in the sequence derive an equivalent junction tree.

Solution: The corresponding graph is shown in Figure 6. The nodes are eliminated in the following sequence: F, I, H, J, G, E, D, C, B, A with the following associated cliques and separators $V_2 = (D,E,F,I), S_2 = (D,E), V_4 = (C,H,G,J), S_4 = (C,G), V_5 = (C,G,D), S_5 = (C,D)$, $V_6 =$

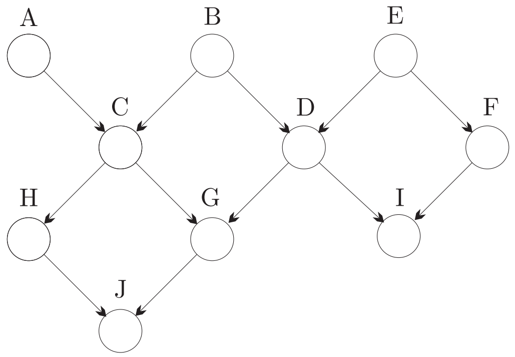

Figure 5: The Bayesian network structure for Problem 16.7

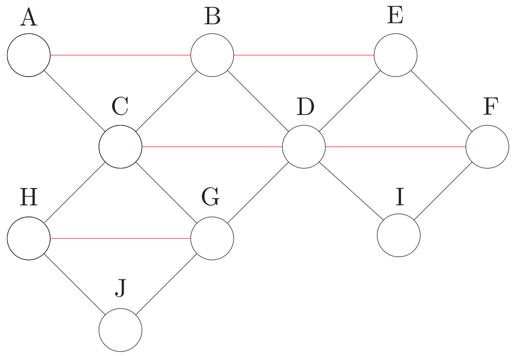

Figure 6: The moralized counterpart of the graph in 5

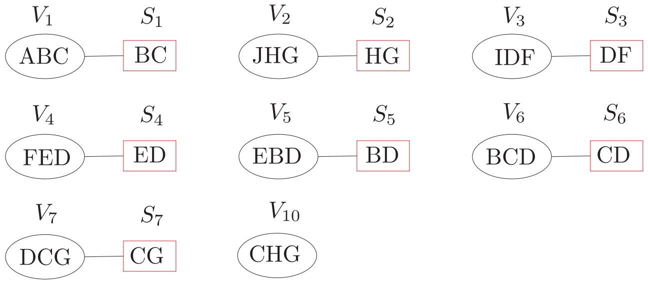

Figure 7: The set of cliques and separators after the sequence of nodes' elimination from Figure 6.

$(B, E, D), S_6 = (B, D), V_{10} = 9A, B, C, D)$. The resulting junction tree is shown in Figure 10

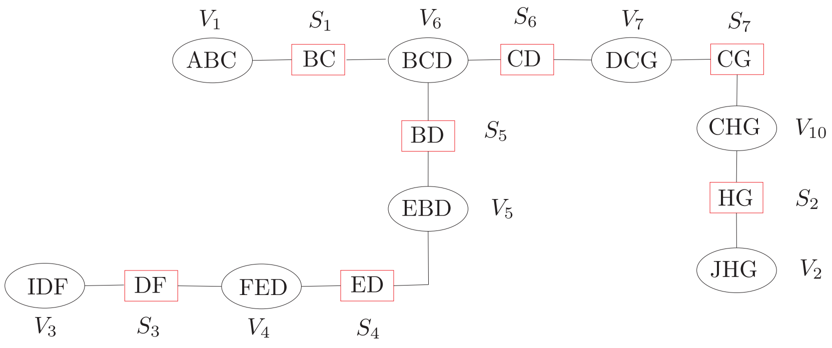

Figure 8: The Join tree for Problem 16.7.

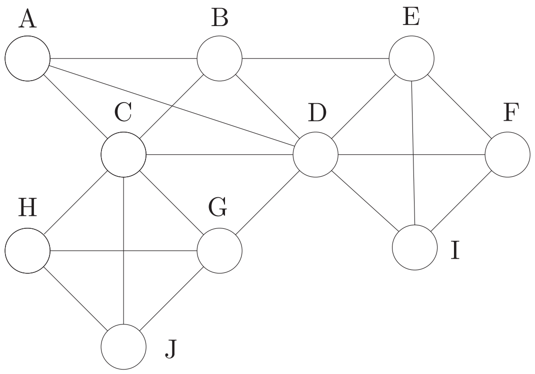

Figure 9: The undirected graph on which the joint Probability of Problem 16.8 factorizes.

16.9. Prove that the function

$$g(x) = 1 - \exp(-x),\ x > 0,$$

is log-concave.

Solution: Let

$$f(x) = \ln(1 - \exp(-x)).$$

Taking the derivative we have

$$f'(x) = \frac{\exp(-x)}{1 - \exp(-x)},$$

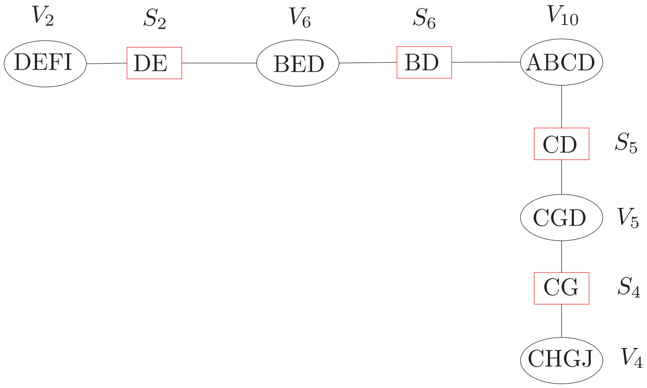

Figure 10: The resulting junction tree for Problem 16.8.

and the second derivative

$$\begin{aligned} f''(x) &= \frac{-\exp(-x)(1-\exp(-x)) - \exp(-x)\exp(-x)}{(1-\exp(-x))^2} \\ &= \frac{-\exp(-x)}{(1-\exp(-x))^2} < 0, \end{aligned}$$

hence the function is concave.

16.10. Derive the conjugate function of

$$f(x) = \ln(1-\exp(-x)).$$

Solution: By the definition and since $f(x)$ is concave we have

$$f^*(\xi) = \min_x \Big(\xi x - \ln(1-\exp(-x))\Big).$$

Taking the derivative with respect to x and equating to zero, we have

$$\begin{aligned} \xi - \frac{\exp(-x)}{1-\exp(-x)} &= 0 \Rightarrow \\ \exp(-x) &= \frac{\xi}{1+\xi} \quad \text{or} \\ x &= -\ln\xi + \ln(1+\xi). \end{aligned}$$

Also,

$$1-\exp(-x) = 1 - \frac{\xi}{1+\xi} = \frac{1}{1+\xi}.$$

Hence,

$$\begin{aligned} f^*(\xi) &= \xi(\ln(1+\xi) - \ln\xi) - \ln\frac{1}{1+\xi} \\ &= -\xi\ln\xi + (1+\xi)\ln(1+\xi). \end{aligned}$$

16.11. Show that minimizing the bound in Eq. (16.12) is a convex optimization task.

Solution: Instead of minimizing the bound itself we can minimize its logarithm, i.e.,

$$-f^*(\xi_i) + \sum_{j\in \mathrm{Pa}_i} \xi_i d_i \theta_{ij}.$$

However, $f^*(\xi)$ is concave, (Section 13.8) being the result of point-wise minimization of a set of affine functions. Hence $-f^*(\xi_i)$ is convex. Since the cost is a sum of convex functions, it is also convex.

16.12. Show that the function $1+\exp(-x),\ x\in\mathbb{R}$ is log-convex and derive the respective conjugate one.

Solution:

$$f(x) = \ln\big(1+\exp(-x)\big),$$

thus,

$$f'(x) = -\frac{\exp(-x)}{1+\exp(-x)},$$

and

$$\begin{aligned} f''(x) &= -\frac{-\exp(-x)(1+\exp(-x)) + \exp(-x)\exp(-x)}{(1+\exp(-x))^2} \\ &= \frac{\exp(-x)}{(1+\exp(-x))^2} > 0. \end{aligned}$$

The conjugate function is

$$f^*(\xi) = \max_x \Big(\xi x - \ln(1+\exp(-x)\Big),$$

or taking the derivative w.r. to x and equating to zero,

$$\xi - \frac{-\exp(-x)}{1+\exp(-x)} = 0,$$

or

$$x = -\ln\frac{-\xi}{1+\xi} \quad\text{and}\quad 1+\exp(-x) = \frac{1}{1+\xi}.$$

Hence,

$$\begin{aligned} f^*(\xi) &= \xi\ln\frac{1+\xi}{-\xi} - \ln\frac{1}{1+\xi} \\ &= (\xi+1)\ln(\xi+1) - \xi\ln(-\xi), \quad -1<\xi<0. \end{aligned}$$

16.13. Derive the KL divergence between $P(\mathcal{X}^l|\mathcal{X})$ and $Q(\mathcal{X}^l)$ for the mean field Boltzmann machine and obtain the respective variational l parameters.

Solution: The KL divergence is given by

$$\begin{aligned}
KL &= -\sum_{x_i\in\mathcal{X}^l} Q(\mathcal{X}^l|\mathcal{X};\boldsymbol{\mu}) \ln\left[\exp\left(-\sum_{i:x_i\in\mathcal{X}^l}\left(\sum_{\substack{j>i\\ j:x_j\in\mathcal{X}^l}} \theta_{ij}x_ix_j + \tilde{\theta}_{i0}x_i\right)\right)\right] \\
&+ \ln\tilde{Z} + \sum_{x_i\in\mathcal{X}^l} Q(\mathcal{X}^l|\mathcal{X};\boldsymbol{\mu}) \ln Q(\mathcal{X}^l|\mathcal{X};\boldsymbol{\mu}) \\
&= \sum_{x_i\in\mathcal{X}^l} Q(\mathcal{X}^l|\mathcal{X};\boldsymbol{\mu}) \left(\sum_{i:x_i\in\mathcal{X}^l}\left(\sum_{\substack{j>i\\ j:x_j\in\mathcal{X}^l}} \theta_{ij}x_ix_j + \tilde{\theta}_{i0}x_i\right)\right) \\
&+ \ln\tilde{Z} + \sum_{x_i\in\mathcal{X}^l} Q(\mathcal{X}^l|\mathcal{X};\boldsymbol{\mu}) \ln\left(\prod_{i:x_i\in\mathcal{X}^l} \mu_i^{x_i}(1-\mu_i)^{(1-x_i)}\right) \\
&= \sum_{i:x_i\in\mathcal{X}^l}\left(\sum_{\substack{j>i\\ j:x_j\in\mathcal{X}^l}}\left(\sum_{x_i\in\mathcal{X}^l} Q(\mathcal{X}^l|\mathcal{X};\boldsymbol{\mu})\theta_{ij}x_ix_j\right) + \sum_{x_i\in\mathcal{X}^l} Q(\mathcal{X}^l|\mathcal{X};\boldsymbol{\mu})\tilde{\theta}_{i0}x_i\right. \\
&+ \left.\sum_{x_i\in\mathcal{X}^l} Q(\mathcal{X}^l|\mathcal{X};\boldsymbol{\mu})\left(x_i\ln\mu_i + (1-x_i)\ln(1-x_i)\right)\right) + \ln\tilde{Z} \\
&= \sum_{i:x_i\in\mathcal{X}^l}\left(\sum_{\substack{j>i\\ j:x_j\in\mathcal{X}^l}} \theta_{ij}\mu_i\mu_j + \tilde{\theta}_{i0}\mu_i + \mu_i\ln\mu_i + (1-\mu_i)\ln(1-\mu_i)\right) \\
&+ \ln\tilde{Z},
\end{aligned}$$

where we assumed that under Q, x_i and x_j are independent. Also, due to the binomial assumption of each individual x_i, we have that $E[\mathrm{x}_i]=\mu_i$. Taking the derivative with respect to each μ_i and setting it equal to zero we get

$$\frac{\partial KL}{\partial\mu_i} = \sum_{j\neq i}\theta_{ij}\mu_j + \tilde{\theta}_{i0} + \ln\mu_i + 1 - \ln(1-\mu_i) - 1 = 0,$$

or

$$\mu_i = \frac{1}{1+\exp(-z)} = \sigma(z),$$

where

$$z = -\left(\sum_{j\neq i}\theta_{ij}\mu_j + \tilde{\theta}_{i0}\right).$$

16.14. Given a distribution in the exponential family,

$$p(\boldsymbol{x}) = \exp\left(\boldsymbol{\theta}^T \boldsymbol{u}(\boldsymbol{x}) - A(\boldsymbol{\theta})\right),$$

show that $A(\boldsymbol{\theta})$ generates the respective mean parameters which define the exponential family

$$\frac{\partial A(\boldsymbol{\theta})}{\partial \theta_i} = \mathbb{E}[u_i(\boldsymbol{x})] = \mu_i.$$

Also, show that $A(\theta)$ is a convex function.

Solution: By the definition of $A(\boldsymbol{\theta})$ we have

$$\frac{\partial A(\boldsymbol{\theta})}{\partial \theta_i} = \frac{\partial}{\partial \theta_i} \ln \int \exp\left(\boldsymbol{\theta}^T \boldsymbol{u}(\boldsymbol{x})\right) d\boldsymbol{x},$$

or

$$\begin{aligned}\frac{\partial A(\boldsymbol{\theta})}{\partial \theta_i} &= \frac{\int u_i(\boldsymbol{x}) \exp\left(\boldsymbol{\theta}^T \boldsymbol{u}(\boldsymbol{x})\right) d\boldsymbol{x}}{\int \exp\left(\boldsymbol{\theta}^T \boldsymbol{u}(\boldsymbol{x})\right) d\boldsymbol{x}} \\ &= \frac{\int u_i(\boldsymbol{x}) \exp\left(\boldsymbol{\theta}^T \boldsymbol{u}(\boldsymbol{x}) - A(\boldsymbol{\theta})\right) d\boldsymbol{x}}{\int \exp\left(\boldsymbol{\theta}^T \boldsymbol{u}(\boldsymbol{x}) - A(\boldsymbol{\theta})\right) d\boldsymbol{x}} \\ &= \int u_i(\boldsymbol{x}) p(\boldsymbol{x};\boldsymbol{\theta}) d\boldsymbol{x} = \mathbb{E}[u_i(\mathbf{x})].\end{aligned}$$

For the convexity we have

$$\begin{aligned}\frac{\partial^2 A}{\partial \theta_j \partial \theta_i} &= \frac{\partial \int u_i(\boldsymbol{x}) \exp\left(\boldsymbol{\theta}^T \boldsymbol{u}(\boldsymbol{x})\right) \exp\left(-A(\boldsymbol{\theta})\right) d\boldsymbol{x}}{\theta_j} \\ &= \int u_i(\boldsymbol{x}) u_j(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} - \mathbb{E}[u_i(\boldsymbol{x})]\mathbb{E}[u_j(\boldsymbol{x})] \\ &= \mathbb{E}\left[\left(u_i(\boldsymbol{x}) - \mathbb{E}[u_i(\boldsymbol{x})]\right)\left(u_j(\boldsymbol{x}) - \mathbb{E}[u_j(\boldsymbol{x})]\right)\right],\end{aligned}$$

and finally,

$$\frac{\partial^2 A}{\partial \boldsymbol{\theta}^2} = \Sigma := \mathbb{E}\left[\left(\boldsymbol{u}(\boldsymbol{x}) - \mathbb{E}[\boldsymbol{u}(\boldsymbol{x})]\right)\left(\boldsymbol{u}(\boldsymbol{x}) - \mathbb{E}[\boldsymbol{u}(\boldsymbol{x})]\right)^T\right], \tag{12}$$

which is a nonnegative matrix. Hence convexity has been shown.

16.15. Show that the conjugate function of $A(\boldsymbol{\theta})$, associated with an exponential distribution, such as that in Problem 16.14, is the corresponding negative entropy function. Moreover, if $\boldsymbol{\mu}$ and $\boldsymbol{\theta}(\boldsymbol{\mu})$ are doubly coupled, then

$$\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{u}(\boldsymbol{x})],$$

where $\mathbb{E}[\cdot]$ is with respect to $p(\boldsymbol{x};\boldsymbol{\theta}(\boldsymbol{\mu}))$.

Solution:

$$A^*(\boldsymbol{\mu}) = \max_{\boldsymbol{\theta}} \left(\boldsymbol{\theta}^T \boldsymbol{\mu} - A(\boldsymbol{\theta})\right),$$

or

$$\boldsymbol{\theta}(\boldsymbol{\mu}) : \boldsymbol{\mu} = \nabla A(\boldsymbol{\theta}) = \mathbb{E}[\boldsymbol{u}(\boldsymbol{x})],$$

where we used the result of Problem 16.14.

Thus, the negative entropy is given by

$$\begin{aligned} H &= \int p(\boldsymbol{x};\boldsymbol{\theta}(\boldsymbol{\mu})) \ln p(\boldsymbol{x};\boldsymbol{\theta}(\boldsymbol{\mu})) d\boldsymbol{x} \\ &= \boldsymbol{\theta}^T(\boldsymbol{\mu})\mathbb{E}[\boldsymbol{u}(\boldsymbol{x})] - A(\boldsymbol{\theta}) = A^*(\boldsymbol{\mu}). \end{aligned}$$

16.16. Derive a recursion for updating $\gamma(\boldsymbol{x}_n)$ in HMMs independent of $\beta(\boldsymbol{x}_n)$

Solution: We know from the text and the respective definition that

$$\begin{aligned} \gamma(\boldsymbol{x}_n) &= P(\boldsymbol{x}_n|Y) = \sum_{\boldsymbol{x}_{n+1}} P(\boldsymbol{x}_n, \boldsymbol{x}_{n+1}|Y) \\ &= \sum_{\boldsymbol{x}_{n+1}} P(\boldsymbol{x}_n|\boldsymbol{x}_{n+1}, Y_{[1:n]}, Y_{[n+1,N]})P(\boldsymbol{x}_{n+1}|Y). \end{aligned}$$

However, by the d–separation properties of the graph representing an HMM,

$$P(\boldsymbol{x}_n|\boldsymbol{x}_{n+1}, Y_{[1:n]}, Y_{[n+1,N]}) = P(\boldsymbol{x}_n|\boldsymbol{x}_{n+1}, Y_{[1:n]}).$$

Thus, we can write

$$\gamma(\boldsymbol{x}_n) = \sum_{\boldsymbol{x}_{n+1}} P(\boldsymbol{x}_n|\boldsymbol{x}_{n+1}, Y_{[1:n]})\gamma(\boldsymbol{x}_{n+1}).$$

Also,

$$\begin{aligned} P(\boldsymbol{x}_n|\boldsymbol{x}_{n+1}, Y_{[1:n]}) &\propto P(\boldsymbol{x}_n, \boldsymbol{x}_{n+1}|Y_{[1:n]}) \\ &= P(\boldsymbol{x}_{n+1}|\boldsymbol{x}_n)P(\boldsymbol{x}_n|Y_{[1:n]}), \end{aligned}$$

where again we have used that by the d–separation properties of the Bayesian network structure

$$P(\boldsymbol{x}_{n+1}|\boldsymbol{x}_n, Y_{[1:n]}) = P(\boldsymbol{x}_{n+1}|\boldsymbol{x}_n).$$

Thus, since

$$P(\boldsymbol{x}_n|Y_{[1:n]}) \propto \alpha(\boldsymbol{x}_n),$$

we can write

$$P(\boldsymbol{x}_n|\boldsymbol{x}_{n+1}, Y_{[1:n]}) \propto P(\boldsymbol{x}_{n+1}|\boldsymbol{x}_n)\alpha(\boldsymbol{x}_n),$$

and the proportionality constant is found by normalization. Note that

$$\gamma(\boldsymbol{x}_N) \propto \alpha(\boldsymbol{x}_N).$$

Thus, one computes $\alpha(\boldsymbol{x}_n)$ and then goes backwards to obtain $\gamma(\boldsymbol{x}_n)$, $n = N-1, N-2, \ldots, 1$.

16.17. Derive an efficient scheme for prediction in HMM models; that is, to obtain $p(\boldsymbol{y}_{N+1}|Y)$ where $Y = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_N\}$.

Solution: We have that

$$\begin{aligned} p(\boldsymbol{y}_{N+1}|Y) &= \sum_{\boldsymbol{x}_{N+1}} p(\boldsymbol{y}_{N+1}, \boldsymbol{x}_{N+1}|Y) \\ &= \sum_{\boldsymbol{x}_{N+1}} p(\boldsymbol{y}_{N+1}|\boldsymbol{x}_{N+1}) P(\boldsymbol{x}_{N+1}|Y) \\ &= \sum_{\boldsymbol{x}_{N+1}} p(\boldsymbol{y}_{N+1}|\boldsymbol{x}_{N+1}) \sum_{\boldsymbol{x}_N} P(\boldsymbol{x}_{N+1}, \boldsymbol{x}_N|Y). \end{aligned}$$

Note that the second sum is over $\xi(\boldsymbol{x}_{N+1}, \boldsymbol{x}_N)$. However, to obtain $\xi(\cdot,\cdot)$ we need two passes. This is not necessary if prediction is our focus. Let us carry on the previous analysis to obtain,

$$\begin{aligned} p(\boldsymbol{y}_{N+1}|Y) &= \sum_{\boldsymbol{x}_{N+1}} p(\boldsymbol{y}_{N+1}|\boldsymbol{x}_{N+1}) \sum_{\boldsymbol{x}_N} P(\boldsymbol{x}_{N+1}|\boldsymbol{x}_N) P(\boldsymbol{x}_N|Y) \\ &= \sum_{\boldsymbol{x}_{N+1}} p(\boldsymbol{y}_{N+1}|\boldsymbol{x}_{N+1}) \sum_{\boldsymbol{x}_N} P(\boldsymbol{x}_{N+1}|\boldsymbol{x}_N) \frac{p(\boldsymbol{x}_N, Y)}{p(Y)} \\ &= \frac{1}{p(Y)} \sum_{\boldsymbol{x}_{N+1}} p(\boldsymbol{y}_{N+1}|\boldsymbol{x}_{N+1}) \sum_{\boldsymbol{x}_N} P(\boldsymbol{x}_{N+1}|\boldsymbol{x}_N) \alpha(\boldsymbol{x}_N), \end{aligned}$$

where

$$p(Y) = \sum_{\boldsymbol{x}_N} \alpha(\boldsymbol{x}_N).$$

Thus for prediction, the forward message passing is sufficient.

16.18. Prove the estimation formulas for the probabilities P_k, $k = 1, 2, \dots, K$, and P_{ij}, $i, j = 1, 2, \dots, K$, in the context of the forward-backward algorithm for training HMM.

Solution: From the text, we have

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

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

P_{ij}. The Lagrangian becomes

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

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

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

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

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

The reestimation rule for the probabilities $P_k^{(t+1)}$ is obtained in a similar way.

16.19. Consider the Gaussian Bayesian network of Section 15.3.5 defined by the local conditional pdfs

$$p(x_i|\mathrm{Pa}_i)=\mathcal{N}\left(x_i \mid \sum_{k:x_k\in\mathrm{Pa}_i}\theta_{ik}x_k+\theta_{i0},\sigma^2\right),\quad i=1,2,\ldots,l.$$

Assume a set of N observations, $x_i(n)$, $n=1,2,\ldots,N$, $i=1,2,\ldots,l$, and derive a maximum likelihood estimate of the parameters $\boldsymbol{\theta}$; assume the common variance σ^2 to be known.

Solution: The likelihood is given by

$$L(X;\boldsymbol{\theta})=\sum_{n=1}^{N}\sum_{i=1}^{l}\left(-\frac{1}{2\sigma^2}(x_i(n)-\boldsymbol{\theta}_i^T\boldsymbol{h}_i(n))^2\right),$$

where $\boldsymbol{h}_i(n)$ is the vector with values of $x_k : k\in\mathrm{Pa}_i$, extended by one dimension to include 1 as its last element, so that $\boldsymbol{\theta}_i$ includes θ_{i0} as well. That is, $\boldsymbol{\theta}_i=[\theta_{i1},\ldots,\theta_{iK},\theta_{i0}]^T$, where K is the number of parents of x_i and $\boldsymbol{h}_i=[h_{i1},h_{i2},\ldots,h_{iK},1]^T$. Since the log-likelihood is obviously decomposable, we will optimize each factor

$$l_i(X;\boldsymbol{\theta}_i)=\sum_{n=1}^{N}\left(-\frac{1}{2\sigma^2}(x_i(n)-\boldsymbol{\theta}_i^T\boldsymbol{h}_i(n))^2\right),\quad i=1,2,\ldots,l,$$

individually. Taking the gradient with respect to $\boldsymbol{\theta}_i$ and equating to zero, we obtain

$$\sum_{n=1}^{N}(\boldsymbol{h}_i(n)\boldsymbol{h}_i^T(n))\boldsymbol{\theta}_i=\sum_{n=1}^{N}x_i(n)\boldsymbol{h}_i(n),$$

or

$$\hat{\boldsymbol{\theta}}_i = \left(\sum_{n=1}^{N} \boldsymbol{h}_i(n) \boldsymbol{h}_i^T(n) \right)^{-1} \left(\sum_{n=1}^{N} x_i(n) \boldsymbol{h}_i(n) \right).$$

In other words, $\hat{\boldsymbol{\theta}}_i$ is nothing else but the LS solution if one tries to predict $x_i(n)$ via the values of its parents.

Solutions To Problems of Chapter 17

17.1. Let

$$\mu := \mathbb{E}[f(\mathbf{x})] = \int f(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}$$

and $q(\boldsymbol{x})$ be the proposal distribution. Show that if

$$w(\boldsymbol{x}) := \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})},$$

and

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} w(\boldsymbol{x}_i)f(\boldsymbol{x}_i),$$

then the variance

$$\sigma_f^2 = \mathbb{E}\left[(\hat{\mu} - \mathbb{E}\left[\hat{\mu}\right])^2\right] = \frac{1}{N}\left(\int \frac{f^2(\boldsymbol{x})p^2(\boldsymbol{x})}{q(\boldsymbol{x})}d\boldsymbol{x} - \mu^2\right).$$

Observe that if $f^2(\boldsymbol{x})p^2(\boldsymbol{x})$ goes to zero slower than $q(\boldsymbol{x})$, then for fixed N $\sigma_f^2 \longrightarrow \infty$.

Solution: Following the solution of Problem 14.6, using p in place of ϕ, we have that $\hat{\mu}$ is unbiased, hence

$$\mathbb{E}[\hat{\mu}] = \mu.$$

where $\mathbb{E}[\cdot]$ is with respect to q. Hence,

$$\mathbb{E}\left[(\hat{\mu} - \mu)^2\right] = \mathbb{E}[(\hat{\mu})^2] + \mu^2 - 2\mu\,\mathbb{E}[\hat{\mu}] = \mathbb{E}[(\hat{\mu})^2] - \mu^2.$$

However

$$\begin{aligned}
\mathbb{E}[(\hat{\mu})^2] &= \mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^{N} w(\mathbf{x}_i)f(\mathbf{x}_i)\right)^2\right] \\
&= \frac{1}{N^2}\left(\sum_{i=1}^{N}\mathbb{E}[w^2(\mathbf{x}_i)f^2(\mathbf{x}_i)] + \sum_{\substack{i=1\\ i\neq j}}^{N}\sum_{j=1}^{N}\mathbb{E}[w(\mathbf{x}_i)f(\mathbf{x}_i)w(\mathbf{x}_j)f(\mathbf{x}_j)]\right) \\
&= \frac{1}{N^2}\left(N\int \frac{p^2(\boldsymbol{x})f^2(\boldsymbol{x})}{q^2(\boldsymbol{x})}q(\boldsymbol{x})d\boldsymbol{x} + (N^2 - N)\left(\mathbb{E}[w(\mathbf{x}_i)f(\mathbf{x}_i)]\right)^2\right),
\end{aligned}$$

due to independence of $\mathbf{x}_i$ and $\mathbf{x}_j$. Thus,

$$\begin{aligned}
\mathbb{E}[(\hat{\mu})^2] &= \frac{1}{N^2}\left(N\int \frac{p^2(\boldsymbol{x})f^2(\boldsymbol{x})}{q(\boldsymbol{x})}d\boldsymbol{x} + (N^2 - N)\left(\int \frac{p(\boldsymbol{x})f(\boldsymbol{x})}{q(\boldsymbol{x})}q(\boldsymbol{x})d\boldsymbol{x}\right)^2\right) \\
&= \frac{1}{N}\left(\int \frac{p^2(\boldsymbol{x})f^2(\boldsymbol{x})}{q(\boldsymbol{x})}d\boldsymbol{x} + (N-1)\mu^2\right)
\end{aligned}$$

Hence

$$\sigma_f^2 = \frac{1}{N}\left(\int \frac{f^2(\boldsymbol{x})p^2(\boldsymbol{x})}{q(\boldsymbol{x})}d\boldsymbol{x} - \mu^2\right).$$

17.2. In Importance sampling, with weights defined as

$$w(\boldsymbol{x}) = \frac{\phi(\boldsymbol{x})}{q(\boldsymbol{x})},$$

where

$$p(\boldsymbol{x}) = \frac{1}{Z}\phi(\boldsymbol{x}),$$

we know from Problem 14.6 that the estimator

$$\hat{Z} = \frac{1}{N}\sum_{i=1}^{N} w(\boldsymbol{x}_i).$$

is an unbiased estimator of the normalizing constant, Z. Show that the respective variance is given by

$$\mathrm{var}[\hat{\mathrm{Z}}] = \frac{Z^2}{N}\left(\int \frac{p^2(\boldsymbol{x})}{q(\boldsymbol{x})}d\boldsymbol{x} - 1\right).$$

Solution: The proof proceeds in exactly the same line as Problem 17.1. This is easily checked out if in the previous problem one sets $f(\boldsymbol{x}) = 1$ and $Zw(\boldsymbol{x})$ instead of $w(\boldsymbol{x})$.

17.3. Show that using resampling in importance sampling, then as the number of particles tends to infinity the approximating, by the respective discrete random measure, distribution, $\bar{p}$, tends to the true (desired) one, p.
Hint: Consider the one-dimensional case.

Solution: Recall from the text that after resampling each resampled particle appears $N^{(i)}$ times out of N. Hence for large enough N, every particle has a probability of occurrence equal to $\frac{N^{(i)}}{N} \approx p^{(i)} = w^{(i)}$. Thus we can write,

$$\begin{aligned} P(x \leq z) &= \sum_{i:x_i \leq z} W^{(i)} = \sum_{i:x_i \leq z} \frac{\phi(x_i)/q(x_i)}{\sum_{i=1}^{N}\phi(x_i)/q(x_i)} \\ &= \frac{1}{\sum_{i=1}^{N}\frac{\phi(x_i)}{q(x_i)}} \sum_{i=1}^{N} \chi_{[0,z]}(x_i)\frac{\phi(x_i)}{q(x_i)}. \end{aligned}$$

where $\chi_{[\cdot,\cdot]}(\cdot)$ is the characteristic function, defined in the text. For large values of N, and assuming all the required regularity conditions are valid, the summation tends to

$$\begin{aligned} P(x \leq z) &= \frac{\int \chi_{[0,z]}(x) \frac{\phi(x)}{q(x)} q(x) dx}{\int \frac{\phi(x)}{q(x)} q(x) dx} \\ &= \int_{-\infty}^{z} p(x) dx, \end{aligned}$$

which is the cumulative desired distribution and the claim is proved.

17.4. Show that in sequential importance sampling, the proposal distribution that minimizes the variance of the weight at time n, conditioned on $\boldsymbol{x}_{1:n-1}$, is given by

$$q_n^{opt}(\boldsymbol{x}_n|\boldsymbol{x}_{1:n-1}) = p_n(\boldsymbol{x}_n|\boldsymbol{x}_{1:n-1})$$

Solution: Note from the text that we have

$$\begin{aligned} w(x_{1:n}) &= w(x_{1:n-1}) \frac{\phi_n(\boldsymbol{x}_{1:n})}{\phi_{n-1}(\boldsymbol{x}_{1:n-1}) q_n(\boldsymbol{x}_n|\boldsymbol{x}_{1:n-1})} \\ &:= a \frac{\phi_n(\boldsymbol{x}_{1:n})}{q_n(\boldsymbol{x}_n|\boldsymbol{x}_{1:n-1})}, \end{aligned}$$

since we assume the past (depending on $\boldsymbol{x}_{1:n-1}$) is known and fixed. Then,

$$\mathbb{E}[w(\mathbf{x}_{1:n})] = a \int \phi_n(\boldsymbol{x}_{1:n}) dx_n := a Z_n p_n(x_{1:n-1}),$$

from the respective definition

$$p_n(\boldsymbol{x}_{1:n}) = \frac{\phi_n(\boldsymbol{x}_{1:n})}{Z_n},$$

and where all means are with respect to the unknown $q_n(\boldsymbol{x}_n|\boldsymbol{x}_{0:n-1})$. Hence,

$$\text{var}[w(\mathbf{x}_{1:n})] = \mathbb{E}[w^2(\mathbf{x}_{1:n})] - a^2 Z_n^2 p_n^2(\boldsymbol{x}_{1:n-1}).$$

Also,

$$\begin{aligned} \mathbb{E}[w^2(\mathbf{x}_{1:n})] &= a^2 \int \frac{\phi_n^2(\boldsymbol{x}_{1:n})}{q_n^2(\boldsymbol{x}_n|\boldsymbol{x}_{1:n-1})} q_n(\boldsymbol{x}_n|\boldsymbol{x}_{1:n-1}) d\boldsymbol{x}_n \\ &= a^2 \int \frac{\phi_n^2(\boldsymbol{x}_{1:n})}{q_n(\boldsymbol{x}_n|\boldsymbol{x}_{1:n-1})} d\boldsymbol{x}_n. \end{aligned}$$

Using a) calculus of variations, b) the constraint

$$\int q_n(\boldsymbol{x}_n|\boldsymbol{x}_{1:n-1}) d\boldsymbol{x}_n = 1 \qquad (1)$$

and c) Lagrange multipliers, we obtain

$$\begin{aligned}
\frac{\phi_n^2(\boldsymbol{x}_{1:n})}{q_n^2(\boldsymbol{x}_n|\boldsymbol{x}_{1:n-1})} &= \lambda \Rightarrow \\
q_n(\boldsymbol{x}_n|\boldsymbol{x}_{1:n-1}) &= \lambda\phi_n(x_{1:n}) \Rightarrow \\
q_n(\boldsymbol{x}_n|\boldsymbol{x}_{1:n-1}) &= \frac{\phi_n(\boldsymbol{x}_{1:n})}{\int \phi_n(\boldsymbol{x}_{1:n})dx_n} = \frac{\phi_n(\boldsymbol{x}_{1:n})}{Z_n p_n(\boldsymbol{x}_{1:n-1})}.
\end{aligned}$$

Then,

$$\begin{aligned}
\mathbb{E}[w^2(\mathbf{x}_{1:n})] &= a^2 Z_n p_n(\boldsymbol{x}_{1:n-1}) \int \phi_n(\boldsymbol{x}_{1:n})d\boldsymbol{x}_n \\
&= a^2 Z_n^2 p_n^2(\boldsymbol{x}_{1:n-1}),
\end{aligned}$$

leading to zero variance. Note that one could put $q_n(\boldsymbol{x}_n|\boldsymbol{x}_{1:n-1}) = \frac{\phi_n(\boldsymbol{x}_{1:n})}{Z_n p_n(\boldsymbol{x}_{1:n-1})}$ straight away and since this makes the variance zero, it is the optimal value without having to use calculus of variation. Thus,

$$q_n(\boldsymbol{x}_n|\boldsymbol{x}_{1:n-1}) = \frac{\phi_n(\boldsymbol{x}_{1:n})}{Z_n p_n(\boldsymbol{x}_{1:n-1})} = \frac{p_n(\boldsymbol{x}_{1:n})}{p_n(\boldsymbol{x}_{1:n-1})} = p_n(\boldsymbol{x}_n|\boldsymbol{x}_{1:n-1}).$$

17.5. In a sequential importance sampling task, let

$$\begin{aligned}
p_n(\boldsymbol{x}_{1:n}) &= \prod_{k=1}^{n} \mathcal{N}(x_k|0,1) \\
\phi_n(\boldsymbol{x}_{1:n}) &= \prod_{k=1}^{n} \exp\left(-\frac{x_k^2}{2}\right)
\end{aligned}$$

and let the proposal distribution be

$$q_n(\boldsymbol{x}_{1:n}) = \prod_{k=1}^{n} \mathcal{N}(x_k|0,\sigma^2).$$

Let the estimator of $Z_n = (2\pi)^{\frac{n}{2}}$, be

$$\hat{Z}_n = \frac{1}{N}\sum_{i=1}^{N} w(\boldsymbol{x}_{1:n}^{(i)}).$$

Show that the variance of the estimator is given by:

$$\mathrm{var}[\hat{\mathrm{Z}}_n] = \frac{Z_n^2}{N}\left(\left(\frac{\sigma^4}{2\sigma^2-1}\right)^{\frac{n}{2}} - 1\right).$$

Observe that for $\sigma^2 > 1/2$, which is the range of values for which the above formula makes sense and guarantees a finite value for the variance, the variance exhibits an exponential increase with respect to n. To keep the variance small one has to make N very large; that is, to generate a very large number of particles ([15]).

Solution: From the Problem 17.2 we have that

$$\mathrm{var}[\hat{\mathrm{Z}}_n] = \frac{Z_n^2}{N}\left(\int \frac{p_n^2(\boldsymbol{x}_{1:n})}{q_n(\boldsymbol{x}_{1:n})} d\boldsymbol{x}_{1:n} - 1\right).$$

However,

$$p_n^2(\boldsymbol{x}_{1:n}) = \frac{1}{Z_n^2}\prod_{k=1}^{n} \exp\left(-x_k^2\right)$$

and

$$q_n(\boldsymbol{x}_{1:n}) = \frac{1}{(2\pi)^{\frac{n}{2}}(\sigma^2)^{\frac{n}{2}}}\prod_{k=1}^{n} \exp\left(-\frac{x_k^2}{2\sigma^2}\right).$$

Hence,

$$\begin{aligned}\int \frac{p_n^2(\boldsymbol{x}_{1:n})}{q_n(\boldsymbol{x}_{1:n})} d\boldsymbol{x}_{1:n} &= \frac{(2\pi)^{\frac{n}{2}}(\sigma^2)^{\frac{n}{2}}}{Z_n^2}\prod_{k=1}^{n}\int \exp\left(-\frac{2\sigma^2-1}{2\sigma^2}x_k^2\right) dx_k \\ &= \frac{(2\pi)^{n+1}}{Z_n^2}\left(\frac{\sigma^4}{2\sigma^2-1}\right)^{\frac{n}{2}} = \left(\frac{\sigma^4}{2\sigma^2-1}\right)^{\frac{n}{2}}.\end{aligned}$$

or

$$\mathrm{var}[\hat{\mathrm{Z}}_n] = \frac{Z_n^2}{N}\left(\left(\frac{\sigma^4}{2\sigma^2-1}\right)^{\frac{n}{2}} - 1\right).$$

17.6. Prove that the use of the optimal proposal distribution in particle filtering leads to

$$w_n(\boldsymbol{x}_{1:n}) = w_{n-1}(\boldsymbol{x}_{1:n-1})p(\boldsymbol{y}_n|\boldsymbol{x}_{n-1}).$$

Solution: Recall that the general update recursion is given by

$$w_n(\boldsymbol{x}_{1:n}) = w_{n-1}(\boldsymbol{x}_{1:n-1})\frac{p(\boldsymbol{y}_n|\boldsymbol{x}_n)p(\boldsymbol{x}_n|\boldsymbol{x}_{n-1})}{q_n(\boldsymbol{x}_n|\boldsymbol{x}_{1:n-1},\boldsymbol{y}_{1:n})}. \tag{2}$$

We know that the optimal proposal distribution is given by

$$\begin{aligned}q(\boldsymbol{x}_n|\boldsymbol{x}_{1:n-1},\boldsymbol{y}_{1:n}) &= p(\boldsymbol{x}_n|\boldsymbol{x}_{n-1},\boldsymbol{y}_n) \\ &= \frac{p(\boldsymbol{y}_n|\boldsymbol{x}_n,\boldsymbol{x}_{n-1})p(\boldsymbol{x}_n|\boldsymbol{x}_{n-1})}{p(\boldsymbol{y}_n|\boldsymbol{x}_{n-1})} \\ &= \frac{p(\boldsymbol{y}_n|\boldsymbol{x}_n)p(\boldsymbol{x}_n|\boldsymbol{x}_{n-1})}{p(\boldsymbol{y}_n|\boldsymbol{x}_{n-1})} \qquad (3)\end{aligned}$$

Combining (2) and (3), proves the claim.

Solutions To Problems of Chapter 18

18.1. Prove that the perceptron algorithm, in its pattern-by-pattern mode of operation, converges in a finite number of iteration steps. Assume that $\boldsymbol{\theta}^{(0)} = \mathbf{0}$.

Hint: Note that since classes are assumed to be linearly separable, there exists a normalized hyperplane, $\boldsymbol{\theta}_{\boldsymbol{*}}$, and a $\gamma > 0$, so that

$$\gamma \leq y_n \boldsymbol{\theta}_*^T \boldsymbol{x}_n, \; n = 1, 2, \ldots, N,$$

where, y_n is the respective label, being $+1$ for ω_1 and -1 for ω_2. By the term normalized hyperplane, we mean that,

$$\boldsymbol{\theta}_*^T = [\hat{\boldsymbol{\theta}}_*, \theta_{0*}]^T, \text{ with } ||\hat{\boldsymbol{\theta}}_*|| = 1.$$

In this case, $y_n \boldsymbol{\theta}_*^T \boldsymbol{x}_n$ is the distance of $\boldsymbol{x}_n$ from the hyperplane $\boldsymbol{\theta}_*$ ([164]).

Solution: Let us assume that we are currently at iteration i. This means that so far $i-1$ update corrections have been performed by the algorithm. The ith update will be performed *only* if the current sample, $\boldsymbol{x}_{(i)}$, is misclassified, i.e., only if

$$y_{(i)} \boldsymbol{x}_{(i)}^T \boldsymbol{\theta}^{(i-1)} \leq 0,$$

and the update will be

$$\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^{(i-1)} + \mu y_{(i)} \boldsymbol{x}_{(i)}. \tag{1}$$

Let us now define,

$$R := \max_n ||\boldsymbol{x}_n||, \tag{2}$$

and let $\boldsymbol{\theta}_*$ be a normalized solution. Then from (1) we obtain,

$$\begin{aligned} \boldsymbol{\theta}_*^T \boldsymbol{\theta}^{(i)} &= \boldsymbol{\theta}_*^T \boldsymbol{\theta}^{(i-1)} + \mu y_{(i)} \boldsymbol{\theta}_*^T \boldsymbol{x}_{(i)} \\ &\geq \boldsymbol{\theta}_*^T \boldsymbol{\theta}^{(i-1)} + \mu\gamma, \end{aligned}$$

which by induction and starting from a zero initial condition, implies that

$$\boldsymbol{\theta}_*^T \boldsymbol{\theta}^{(i)} \geq i\mu\gamma. \tag{3}$$

Also from (1), we have that

$$\begin{aligned} ||\boldsymbol{\theta}^{(i)}||^2 &= ||\boldsymbol{\theta}^{(i-1)}||^2 + 2\mu y_{(i)} \boldsymbol{x}_{(i)}^T \boldsymbol{\theta}^{(i-1)} + \\ & \mu^2 y_{(i)}^2 ||\boldsymbol{x}_{(i)}||^2 \\ &\leq ||\boldsymbol{\theta}^{(i-1)}||^2 + \mu^2 y_{(i)}^2 ||\boldsymbol{x}_{(i)}||^2 \\ &\leq ||\boldsymbol{\theta}^{(i-1)}||^2 + \mu^2 R^2, \end{aligned} \tag{4}$$

which by induction leads to

$$||\boldsymbol{\theta}^{(i)}||^2 \leq i\mu^2 R^2. \tag{5}$$

Thus combining (3) and (5) and using the Schwartz inequality, we obtain

$$||\boldsymbol{\theta}_*||\sqrt{i}\mu R \geq ||\boldsymbol{\theta}_*||||\boldsymbol{\theta}^{(i)}|| \geq \boldsymbol{\theta}_*^T \boldsymbol{\theta}^{(i)} \geq i\mu\gamma, \tag{6}$$

or

$$i \leq \left(\frac{R}{\gamma}\right)^2 ||\boldsymbol{\theta}_*||^2, \tag{7}$$

which proves that i will remain finite.

18.2. The derivative of the sigmoid functions has been computed in Problem 7.6. Compute the derivative of the hyperbolic tangent activation function and show that it is equal to,

$$f'(z) = ac\big(1 - f^2(z)\big).$$

Solution: We have that,

$$f(z) \ = \ a \tanh(cz) \ \Rightarrow \ f'(z) \ = \ (ac) \cosh^{-2} cz)$$

or

$$f'(z) \ = \ ac(1 - \tanh^2(cz))$$

or

$$f'(z) \ = \ ac(1 - f^2(z))$$

18.3. Show that the effect of the momentum term in the gradient descent backpropagation scheme is to effectively increase the learning convergence rate of the algorithm.
Hint: Assume that the gradient is approximately constant over I successive iterations.

Solution Let i denote the current iteration, i.e.,

$$\Delta\boldsymbol{\theta}_j^r(i) = a\Delta\boldsymbol{\theta}_j^r(i-1) - \mu\boldsymbol{g}, \tag{8}$$

since the gradient has been assumed to be constant. Repeating the iteration over I successive steps, we get

$$\Delta\boldsymbol{\theta}_j^r(I) = -\mu\sum_{i=0}^{I-1} a^i \boldsymbol{g} + a^I \Delta\boldsymbol{\theta}_j^t(0). \tag{9}$$

Since a is less than one, the rightmost term in the previous recursion tends to zero and the remaining term, for large values of I, becomes

$$\Delta\boldsymbol{\theta}_j^r(I) \approx -\mu(1 + a + a^2 + \ldots + a^{I-1})\boldsymbol{g} \approx -\frac{\mu}{1-a}\boldsymbol{g}. \tag{10}$$

18.4. Show that if a) the activation function is the hyperbolic tangent, b) the input variables are normalized to zero mean and unit variance, then in order to guarantee that all the outputs of the neurons are zero mean and unit variance, the weights must be drawn from a distribution of zero mean and standard deviation equal to

$$\sigma = m^{-1/2},$$

where m is the number of synaptic weights associated with the corresponding neuron.
Hint: For simplicity, consider the bias to be zero, and also that the inputs to each neuron are mutually uncorrelated.

Solution: Let us focus on the weights of a specific neuron, and denote by $y_i,\ i = 1, 2, \ldots, m$, the respective inputs. Then, the input to the activation functions will be given by

$$z = \sum_{i=1}^{m} \theta_i y_i. \tag{11}$$

Assume that the inputs to the neuron are zero mean and unit variance and are also uncorrelated, then

$$\sigma_i^2 = \mathbb{E}[\mathrm{y}_i^2] \tag{12}$$

and

$$\mathbb{E}[\mathrm{y}_i \mathrm{y}_j] = \delta_{ij}. \tag{13}$$

Recall that our goal is to produce a variable z of zero mean and of unit variance, since for the selected activation function, this guarantees values which are small but not very small, and correspond to the linear operation region of it. Let us select all weights to be of zero mean. Since the weights are generated independently of the inputs, then we have that the mean value of z is zero. Hence,

$$\begin{aligned} \sigma_z^2 &= \mathbb{E}[\mathrm{z}^2] = \mathbb{E}\left[\sum_i \sum_j \theta_i \theta_j \mathrm{y}_i \mathrm{y}_j\right] \\ &= \sum_i \sum_j \mathbb{E}[\theta_i \theta_j]\, \mathbb{E}[\mathrm{y}_i \mathrm{y}_j] \\ &= \sum_{i=1}^{m} \mathbb{E}[\theta_i^2] = m\sigma_\theta^2. \end{aligned} \tag{14}$$

Since we want the variance of z to be one, then

$$\sigma_\theta = m^{-1/2}.$$

18.5. Consider the sum of error squares cost function

$$J = \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{k_L}(\hat{y}_{nm} - y_{nm})^2. \tag{15}$$

Compute the elements of the Hessian matrix

$$\frac{\partial^2 J}{\partial\theta_{kj}^{r}\partial\theta_{k'j'}^{r'}}. \tag{16}$$

Near the optimum, show that the second order derivatives can be approximated by

$$\frac{\partial^2 J}{\partial\theta_{kj}^{r}\partial\theta_{k'j'}^{r'}} = \sum_{n=1}^{N}\sum_{m=1}^{k_L}\frac{\partial\hat{y}_{nm}}{\partial\theta_{kj}^{r}}\frac{\partial\hat{y}_{nm}}{\partial\theta_{k'j'}^{r'}}. \tag{17}$$

In other words, the second order derivatives can be approximated as products of the first order derivatives. The derivatives can be computed by following similar arguments as the gradient descent backpropagation scheme, [71].

Solution:

$$\begin{aligned}
J &= \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{k_L}(\hat{y}_{nm} - y_{nm})^2 \\
\frac{\partial J}{\partial\theta_{kj}^{r}} &= \sum_{n=1}^{N}\sum_{m=1}^{k_L}\frac{\partial\hat{y}_{nm}}{\partial\theta_{kj}^{r}}(\hat{y}_{nm} - y_{nm}) \\
\frac{\partial^2 J}{\partial\theta_{k'j'}^{r'}\partial\theta_{kj}^{r}} &= \sum_{n=1}^{N}\sum_{m=1}^{k_L}\frac{\partial^2\hat{y}_{nm}}{\partial\theta_{k'j'}^{r'}\partial\theta_{kj}^{r}}(\hat{y}_{nm} - y_{nm}) \\
&+ \sum_{n=1}^{N}\sum_{m=1}^{k_L}\frac{\partial\hat{y}_{nm}}{\partial\theta_{kj}^{r}}\frac{\partial\hat{y}_{nm}}{\partial\theta_{k'j'}^{r'}}
\end{aligned}$$

If we are near a good minimum, $\hat{y}_{nm} \simeq y_{nm}$ thus,

$$\frac{\partial^2 J}{\partial\theta_{k'j'}^{r'}\partial\theta_{kj}^{r}} \simeq \sum_{n=1}^{N}\sum_{m=1}^{k_L}\frac{\partial\hat{y}_{nm}}{\partial\theta_{kj}^{r}}\frac{\partial\hat{y}_{nm}}{\partial\theta_{k'j'}^{r'}}$$

18.6. It is common when computing the Hessian matrix, to assume that it is diagonal. Show, that under this assumption, the quantities

$$\frac{\partial^2 E}{\partial(\theta_{kj})^2},$$

where

$$E := \sum_{m=1}^{k_L}\left(f(z_m^L) - y_m\right)^2,$$

propagate backwards according to the following,

- $\frac{\partial^2 E}{\partial(\theta_{kj}^r)^2} \;=\; \frac{\partial^2 E}{\partial(z_j^r)^2}(y_k^{r-1})^2.$
- $\frac{\partial^2 E}{\partial(z_L^r)^2} = f''(z_j^L)e_j + (f'(z_j^L))^2.$
- $\frac{\partial^2 E}{\partial(z_k^{r-1})^2} \;=\; (f'(z_j^{r-1}))^2 \sum_k \frac{\partial^2 E}{\partial(z_k^r)^2}(\theta_{kj}^r)^2 + f''(z_j^{r-1}) \sum_{k=1}^{k_r} \theta_{kl}^r \delta_k^r.$

Solution:

$$\frac{\partial^2 E}{\partial(\theta_{kj}^r)^2} \;=\; \frac{\partial}{\partial\theta_{kj}^r}[\frac{\partial E}{\partial\theta_{kj}^r}] \;=\; \frac{\partial}{\partial\theta_{kj}^r}[\frac{\partial E}{\partial z_j^r}][\frac{\partial z_j^r}{\partial\theta_{kj}^r}] = \tag{18}$$

$$\frac{\partial}{\partial\theta_{kj}^r}\delta_j^r y_k^{r-1} \;=\; \frac{\partial\delta_j^r}{\partial\theta_{kj}^r} y_k^{r-1}.$$

However,

$$\frac{\partial\delta_j^r}{\partial\theta_{kj}^r} \;=\; \frac{\partial\delta_j^r}{\partial z_j^r}\frac{\partial z_j^r}{\partial\theta_{kj}^r} \;=\; y_k^{r-1}\frac{\partial\delta_j^r}{\partial z_j^r}. \tag{19}$$

Hence,

$$\frac{\partial^2 E}{\partial(\theta_{kj}^r)^2} \;=\; \frac{\partial^2 E}{\partial(z_j^r)^2}(y_k^{r-1})^2. \tag{20}$$

For the backpropagation of the involved second derivatives, we have

I: $r = L$:

$$\frac{\partial E}{\partial z_j^L} \;=\; f'(z_j^L)e_j$$

and

$$\begin{aligned} \frac{\partial^2 E}{\partial(z_L^r)^2} \;&=\; f''(z_j^L)e_j + (f'(z_j^L)\frac{\partial e_j}{\partial z_j^L} \\ &=\; f''(z_j^L)e_j + (f'(z_j^L))^2 \end{aligned}$$

II $r < L$:
We know from the text that,

$$\delta_j^{r-1} \;=\; \sum_{k=1}^{k_r}\delta_k^r\frac{\partial z_k^r}{\partial z_j^{r-1}} \;=\; \sum_{k=1}^{k_r}\delta_k^r\theta_{kj}^r f'(z_j^{r-1}) \tag{21}$$

and

$$\begin{aligned} \frac{\partial^2 E}{\partial(z_j^{r-1})^2} \;=\; \frac{\partial\delta_j^{r-1}}{\partial z_j^{r-1}} \;&=\; \sum_{k=1}^{k_l}\theta_{kj}^r\frac{\partial\delta_k^r}{\partial z_j^{r-1}} f'(z_j^{r-1}) \\ &\quad + \sum_{k=1}^{k_l}\theta_{kj}^r\delta_k^r\frac{\partial f'(z_j^{r-1})}{\partial z_j^{r-1}}, \end{aligned}$$

or

$$\frac{\partial^2 E}{\partial(z_j^{r-1})^2} = f''(v_j^{r-1})\sum_{k=1}^{k_l}\theta_{kj}^r\delta_k^r + \sum_{k=1}^{k_l}\theta_{kj}^r\frac{\partial\delta_k^r}{\partial z_j^{r-1}} f'(z_j^{r-1}) \tag{22}$$

also

$$\begin{aligned}\frac{\partial \delta_k^r}{\partial z_j^{r-1}} &= \sum_{k'} \frac{\partial \delta_k^r}{\partial z_{k'}^r}\frac{\partial z_{k'}^r}{\partial z_j^{r-1}} = \sum_{k'} \frac{\partial^2 E}{\partial z_k^r \partial z_{k'}^r}\theta_{k'j}^r f'(z_j^{r-1}) \\ &= \frac{\partial^2 E}{\partial (z_k^r)^2}\theta_{kj}^r f'(z_j^{r-1}).\end{aligned}$$

Thus, finally we get,

$$\frac{\partial^2 E}{\partial (z_k^{r-1})^2} = (f'(z_j^{r-1}))^2 \sum_k \frac{\partial^2 E}{\partial (z_k^r)^2}(\theta_{kj}^r)^2 + f''(z_j^{r-1})\sum_{k=1}^{k_r}\theta_{kl}^r \delta_k^r$$

18.7. Show that if the activation function is the logistic sigmoid and the loss function in (18.44) is used, then δ_{nj}^L in (18.21) becomes

$$\delta_{nj}^L = a(\hat{y}_{nj} - y_{nj}).$$

If the cross-entropy in Eq. (18.41) is used, then it is equal to

$$\delta_{nj}^L = a y_{nj}(\hat{y}_{nj} - 1).$$

Solution: The cost function is

$$J = \sum_{n=1}^N J_n,$$

and

$$J_n = -\sum_{k=1}^{k_L}\left(y_{nk}\ln \hat{y}_{nk} + (1-y_{nk})\ln(1-\hat{y}_{nk})\right).$$

Hence,

$$\frac{\partial J_n}{\partial z_{nk}} = \frac{\partial J_n}{\partial \hat{y}_{nk}}\frac{\partial \hat{y}_{nk}}{\partial z_{nk}}.$$

However,

$$\frac{\partial J_n}{\partial \hat{y}_{nk}} = -\frac{y_{nk}}{\hat{y}_{nk}} + \frac{1-y_{nk}}{1-\hat{y}_{nk}} = \frac{\hat{y}_{nk} - y_{nk}}{\hat{y}_{nk}(1-y_{nk})}.$$

Also,

$$\frac{\partial \hat{y}_{nk}}{\partial z_{nk}} = \sigma^{'}(z_{nk}) = a\sigma(z_{nk})(1-\sigma(z_{nk})) = a\hat{y}_{nk}(1-\hat{y}_{nk}).$$

Combining the above we finally obtain that

$$\delta_{nk}^L := \frac{\partial J_n}{\partial z_{nk}} = a(\hat{y}_{nk} - y_{nk}).$$

The proof for the cross-entropy is straightforward from the above.

18.8. Show that if the activation function is the logistic sigmoid and the relative entropy loss function is used, then δ_{nj}^L in (18.21) becomes,

$$\delta_{nj}^L = a(\hat{y}_{nj} - 1)y_{nj}.$$

Solution: From the theory, we have that

$$\delta_{nj}^L := \frac{\partial J_n}{\partial z_{nj}^L},$$

and

$$\hat{y}_{nj} = f(z_{nj}^L).$$

The cost function is

$$J = \sum_{n=1}^{N} J_n,$$

with

$$\begin{aligned} J_n &= -\sum_{k=1}^{k_L} y_{nk} \ln \frac{\hat{y}_{nk}}{y_{nk}} \\ &= -\sum_{k=1}^{k_L} y_{nk} \ln \frac{f(z_{nk}^L)}{y_{nk}} \end{aligned}$$

$$\begin{aligned} \delta_{nj}^L := \frac{\partial J_n}{\partial z_{nj}^L} &= -\frac{y_{nj}}{f(z_{nj}^L)} f'(z_{nj}^L) \\ &= -\frac{y_{nj}}{f(z_{nj}^L)} a f(z_{nj}^L)(1 - f(z_{nj}^L)) \end{aligned}$$

or

$$\delta_{nj}^L = a(\hat{y}_{nj} - 1)y_{nj}$$

18.9. Show that the cross-entropy loss function depends on the relative output errors.

Solution: We will prove it for one of the versions, and similar is the proof for the other. Taking into consideration the limiting value $0 \ln 0 = 0$, we can write the cross-entropy cost as,

$$J = -\sum_{n=1}^{N} \sum_{k=1}^{k_L} \Big(y_{nk} \ln \frac{\hat{y}_{nk}}{y_{nk}} + (1 - y_{nk}) \ln \frac{1 - \hat{y}_{nk}}{1 - y_{nk}} \Big).$$

Let,

$$\hat{y}_{nk} = y_{nk} + \epsilon_{nk}, \text{ and } (1 - \hat{y}_{nk}) = (1 - y_{nk}) + \epsilon_{nk}^{'}$$

where ϵ_{nk} and $\epsilon^{'}_{nk}$ are the respective errors. Then,

$$J \;=\; -\sum_{n=1}^{N}\sum_{k=1}^{k_L}\Big(y_{nk}\ln\frac{y_{nk}+\epsilon_{nk}}{y_{nk}} + (1-y_{nk})\ln\frac{1-y_{nk}+\epsilon^{'}_{nk}}{1-y_{nk}}\Big).$$

Thus J depends on the relative errors.

18.10. Show that if the activation function is the softmax and the loss function is the cross-entropy (or the loss in (18.44)), then δ^L_{nj} in (18.21) does not depend on the derivatives of the nonlinearity.

Solution: We will show it for Eq. (18.44) and for the cross-entropy it is just an obvious variant. We have that

$$J_n = -\sum_{k=1}^{k_L}\left(y_{nk}\ln\hat{y}_{nk} + (1-y_{nk})\ln(1-\hat{y}_{nk})\right),$$

where

$$\hat{y}_{nk} = \frac{\exp(z^L_{nk})}{\sum_{m=1}^{k_L}\exp(z^L_{nm})}.$$

Thus,

$$\frac{\partial J_n}{\partial z^L_{nj}} = \frac{\partial J_n}{\partial \hat{y}_{nj}}\frac{\partial \hat{y}_{nj}}{\partial z^L_{nj}}.$$

Now,

$$\frac{\partial \hat{y}_{nj}}{\partial z^L_{nj}} = \frac{\partial}{\partial z^L_{nj}}\left[\frac{\exp(z^L_{nj})}{\sum_{m=1}^{k_L}\exp(z^L_{nm})}\right],$$

or

$$\begin{aligned}\frac{\partial \hat{y}_{nj}}{\partial z^L_{nj}} \;&=\; \frac{\exp(z^L_{nj})\sum_{m=1}^{k_L}\exp(z^L_{nm}) - \exp(2z^L_{nj})}{\left(\sum_{m=1}^{k_L}\exp(z^L_{nm})\right)^2}\\ &=\; \frac{\exp(z^L_{nj})}{\sum_{m=1}^{k_L}\exp(z^L_{nm})} - \left(\frac{\exp(z^L_{nj)}}{\sum_{m=1}^{k_L}\exp(z^L_{nm})}\right)^2\\ &=\; \hat{y}_{nj}(1-\hat{y}_{nj}). \end{aligned} \qquad (23)$$

Also,

$$\frac{\partial J_n}{\partial \hat{y}_{nj}} = -\frac{y_{nj}}{\hat{y}_{nj}} + \frac{1-y_{nj}}{1-\hat{y}_{nj}} = \frac{\hat{y}_{nj}-y_{nj}}{\hat{y}_{nj}(1-\hat{y}_{nj})}. \qquad (24)$$

Combining Eqs. (23) and (24), we finally obtain that

$$\frac{\partial J_n}{\partial z^L_{nj}} = \hat{y}_{nj} - y_{nj}.$$

18.11. As in the previous problem, use the relative entropy as the cost function and the softmax activation function. Then show that,

$$\delta_{nj}^{L} = \hat{y}_{nj} - y_{nj}.$$

Solution:

$$\begin{aligned}
J_n &= -\sum_{k=1}^{k_L} y_{nk} \ln \frac{\hat{y}_{nk}}{y_{nk}}, \\
\hat{y}_{nk} &= \frac{\exp(z_{nk}^{L})}{\sum_m \exp(z_{nm}^{L})},
\end{aligned}$$

$$\begin{aligned}
\delta_{nj}^{L} \quad = \quad \frac{\partial J_n}{\partial z_{nj}^{L}} &= -\frac{\partial}{\partial z_{nj}^{L}} \Big(\sum_{k=1}^{k_L} y_{nk} z_{nk}^{L} - \\
& \sum_{k=1}^{k_L} y_{nk} \ln \big(\sum_m \exp(z_{nm}^{L}) \big) \Big), \\
&= - y_{nj} + \sum_{k=1}^{k_L} y_{nk} \frac{\exp(z_{nj}^{L})}{\sum_m \exp(z_{nm}^{L})}, \\
&= -y_{nj} + \hat{y}_{nj} \sum_{k=1}^{k_L} y_{nk} \;=\; \hat{y}_{nj} - y_{nj}
\end{aligned}$$

18.12. Derive the backpropagation through time algorithm for training recurrent neural networks.

Solutions: The starting set of equations are those defining an RNN, i.e.,

$$\begin{aligned}
\boldsymbol{h}_n &= f\left(U\boldsymbol{x}_n + W\boldsymbol{h}_{n-1} + \boldsymbol{b}\right), \\
\hat{\boldsymbol{y}}_n &= g\left(V\boldsymbol{h}_n + \boldsymbol{c}\right).
\end{aligned}$$

Let us start with the gradient of the cost J w.r. to the elements of matrix W. By definition we have,

$$\frac{\partial J}{\partial W} := \begin{bmatrix} \frac{\partial J}{\partial w_{11}} & \cdots & \frac{\partial J}{\partial w_{1k_h}} \\ \vdots & \ddots & \vdots \\ \frac{\partial J}{\partial w_{k_h 1}} & \cdots & \frac{\partial J}{\partial w_{k_h k_h}} \end{bmatrix},$$

where k_h is the dimensionality of the state vector.

For the computation of the above, we have the following chain rule,

$$\frac{\partial J}{\partial w_{ij}} = \sum_{n=1}^{N} \frac{\partial J}{\partial h_{ni}} \frac{\partial h_{ni}}{\partial w_{ij}}. \tag{25}$$

Note from the defining RNN equations, all the w_{ij}, $j = 1, 2, \ldots, k_h$, elements affect only the corresponding ith component of $\boldsymbol{h}_n$. Also,

$$\frac{\partial h_{ni}}{\partial w_{ij}} = f^{'}(z) h_{(n-1)j}, \quad z := \left(U\boldsymbol{x}_n + W\boldsymbol{h}_{n-1} + \boldsymbol{b}\right)_i .$$

Let us focus on the tanh nonlinearity. Then,

$$f^{'}(z) = 1 - (\tanh(z))^2 = 1 - h_{ni}^2,$$

or

$$\frac{\partial h_{ni}}{\partial w_{ij}} = \left(1 - h_{ni}^2\right) h_{(n-1)j}. \tag{26}$$

Combining Eqs. (25) and (26) in a vector form, we get

$$\frac{\partial J}{\partial W} = \sum_{n=1}^{N} \text{diag}\left(1 - h_{ni:1,k_h}^2\right) \frac{\partial J}{\partial \boldsymbol{h}_n} \boldsymbol{h}_{n-1}^T, \tag{27}$$

where diag $\left(1 - h_{ni:1,k_h}^2\right)$ is a diagonal matrix with its (i, i) element being equal to h_{ni}^2, $i = 1, 2, \ldots, k_h$. In the above formula, the gradient with respect to the state vectors are computed via the chain rule, in the backward pass, as it is explained in the text, i.e.,

$$\frac{\partial J}{\partial \boldsymbol{h}_n} = \left(\frac{\partial \boldsymbol{h}_{n+1}}{\partial \boldsymbol{h}_n}\right)^T \frac{\partial J}{\partial \boldsymbol{h}_{n+1}} + \left(\frac{\partial \hat{\boldsymbol{y}}_n}{\partial \boldsymbol{h}_n}\right)^T \frac{\partial J}{\partial \hat{\boldsymbol{y}}_n}. \tag{28}$$

For the above to be complete, we need to compute $\frac{\partial J}{\partial \boldsymbol{h}_N}$ (where the recursion starts), $\frac{\partial \hat{\boldsymbol{y}}_n}{\partial \boldsymbol{h}_n}$ and $\frac{\partial \boldsymbol{h}_{n+1}}{\partial \boldsymbol{h}_n}$. The gradient of the cost w.r. to the outputs is straightforward, and depends on the specific loss function, see, e.g., Problem 18.10.

Before we proceed further, let us introduce the variable, $\boldsymbol{z}_n = V\boldsymbol{h}_n + \boldsymbol{c}$, i.e., prior to the output nonlinearity. We will express derivatives w.r. to the latter, since it makes the formulae simpler and more general, independent of the specific form of the output nonlinearity.

(a) For $n = N$, we have,

$$\frac{\partial J}{\partial \boldsymbol{h}_N} = V^T \frac{\partial J}{\partial \boldsymbol{z}_N}.$$

where the computation of the latter gradient is straightforward and it depends on the loss function and the output nonlinearity, e.g., Problem 18.10.

(b) Also, following similar arguments as before,

$$\frac{\partial \boldsymbol{h}_{n+1}}{\partial \boldsymbol{h}_n} = \text{diag}(1 - h_{(n+1)i:1,k_h}^2) W.$$

(c) Taking into account that the output nonlinearity is the softmax and recalling the results from Problem 18.10, we get

$$\frac{\partial \hat{\boldsymbol{y}}_n}{\partial \boldsymbol{h}_n} = \text{diag}\big(\hat{y}_i(1-\hat{y}_i)\big)V.$$

Note that Eq. (28) can also be expressed in terms of gradients w.r. to $\boldsymbol{z}_n$. Then, instead of $\frac{\partial \hat{\boldsymbol{y}}_n}{\partial \boldsymbol{h}_n}$ we would have to compute $\frac{\partial \boldsymbol{z}_n}{\partial \boldsymbol{h}_n}$, which is straightforward.

The above have completed the calculations of the gradient of J w.r. to W.

- For the gradient w.r. to V, we get,

$$\frac{\partial J}{\partial V} = \sum_{n=1}^{N} \frac{\partial J}{\partial \boldsymbol{z}_n} \boldsymbol{h}_n^T$$

- For the gradient w.r. to U and following similar arguments as for W we obtain

$$\frac{\partial J}{\partial U} = \sum_{n=1}^{N} \text{diag}\left(1 - h_{ni:1,k_h}^2\right) \frac{\partial J}{\partial \boldsymbol{h}_n} \boldsymbol{x}_{n-1}^T,$$

- For the gradient w.r. to $\boldsymbol{c}$, we get

$$\frac{\partial J}{\partial \boldsymbol{c}} = \sum_{n=1}^{N} \left(\frac{\partial \boldsymbol{z}_n}{\partial \boldsymbol{c}}\right)^T \frac{\partial J}{\partial \boldsymbol{z}_n} = \sum_{n=1}^{N} \frac{\partial J}{\partial \boldsymbol{z}_n}$$

- The gradient w.r. to $\boldsymbol{b}$ is equal to,

$$\begin{aligned} \frac{\partial J}{\partial \boldsymbol{b}} &= \sum_{n=1}^{N} \left(\frac{\partial \boldsymbol{h}_n}{\partial \boldsymbol{b}}\right)^T \frac{\partial J}{\partial \boldsymbol{h}_n} \\ &= \sum_{n=1}^{N} \text{diag}\left(1 - h_{ni:1,k_h}^2\right) \frac{\partial J}{\partial \boldsymbol{h}_n} \end{aligned}$$

We have now derived the required gradients of the cost J w.r. to all parameters and matrices.

18.13. Derive the gradient of the log-likelihood in (18.74).

Solutions: To simplify slightly the notation, we will consider only one term of the sum. Thus we have,

$$\begin{aligned} \frac{\partial P(\boldsymbol{v};\Theta)}{\partial \theta_{ij}} &= \frac{\partial}{\partial \theta_{ij}} \ln\Big(\sum_{\boldsymbol{h}} \exp\left(-E\right)\Big) - \qquad (29) \\ & \frac{\partial}{\partial \theta_{ij}} \ln\Big(\sum_{\boldsymbol{v}} \sum_{\boldsymbol{h}} \exp\left(-E\right)\Big), \end{aligned}$$

or

$$\begin{aligned}\frac{\partial P(\boldsymbol{v};\Theta)}{\partial \theta_{ij}} &= -\frac{1}{\sum_{\boldsymbol{h}} \exp(-E)} \sum_{\boldsymbol{h}} \exp(-E) \frac{\partial E}{\partial \theta_{ij}} + \\ & \frac{1}{Z} \sum_{\boldsymbol{v}} \sum_{\boldsymbol{h}} \exp(-E) \frac{\partial E}{\partial \theta_{ij}}, \end{aligned} \tag{30}$$

or

$$\begin{aligned}\frac{\partial P(\boldsymbol{v};\Theta)}{\partial \theta_{ij}} &= \sum_{\boldsymbol{h}} \frac{\exp(-E)}{\sum_{\boldsymbol{h}} \exp(-E)} h_i v_j - \\ & \sum_{v} \sum_{h} P(\boldsymbol{v},\boldsymbol{h}) h_i v_j. \end{aligned} \tag{31}$$

Taking into account that

$$P(\boldsymbol{h}|\boldsymbol{v}) = \frac{\exp(-E)}{\sum_{\boldsymbol{h}} \exp(-E)},$$

the desired formula is derived.

18.14. Prove that for the case of RBMs, the conditional probabilities are given by the following factorized form,

$$P(\boldsymbol{h}|\boldsymbol{v}) = \prod_{i=1}^{I} \frac{\exp\left(\sum_{j=1}^{J} \theta_{ij} v_j + b_i\right) h_i}{\sum_{h_i'} \left[\exp\left(\sum_{j=1}^{J} \theta_{ij} v_j + b_i\right) h_i'\right]}.$$

Solution: Our starting point is

$$P(\boldsymbol{h}|\boldsymbol{v}) = \frac{\exp\left(\sum_{i=1}^{M}\sum_{j=1}^{K} \theta_{ij} h_i v_j + \sum_{i=1}^{M} b_i h_i\right)}{\sum_{\boldsymbol{h}} \exp\left(\sum_{i=1}^{M}\sum_{j=1}^{K} \theta_{ij} h_i v_j + \sum_{i=1}^{M} b_i h_i\right)}.$$

The numerator is written as,

$$A := \prod_{i=1}^{M} \exp\Big((\sum_{j=1}^{K} \theta_{ij} v_j + b_i) h_i \Big).$$

The denominator becomes

$$\begin{aligned} B: &= \sum_{h_1}\sum_{h_2}\cdots\sum_{h_M}\prod_{i=1}^{M} \exp\Big((\sum_{j=1}^{K} \theta_{ij} v_j + b_i) h_i \Big) \\ &= \prod_{i=1}^{M}\sum_{h_i} \exp\Big((\sum_{j=1}^{K} \theta_{ij} v_j + b_i) h_i \Big). \end{aligned}$$

Combining A and B we obtain the result.

18.15. Derive Eq. (18.83).

Solution: The integrand is equal to

$$p_r(\boldsymbol{x})\ln(D(\boldsymbol{x})) + p_g(\boldsymbol{x})\ln\big(1 - D(\boldsymbol{x})\big).$$

Taking the derivative w.r. to D and setting it equal to zero, the claim is readily obtained.

Solutions To Problems of Chapter 19

19.1. Show that the second principal component in PCA is given as the eigenvector corresponding to the second largest eigenvalue.

Solution As pointed out in the text, the following optimization task is in order,

$$\begin{aligned} \text{maximize} \quad & \boldsymbol{v}^T \hat{\Sigma} \boldsymbol{v}, & (1)\\ \text{s.t.} \quad & \boldsymbol{v}^T \boldsymbol{v} = 1, & (2)\\ \text{s.t.} \quad & \boldsymbol{v}^T \boldsymbol{v}_1 = 0. & (3) \end{aligned}$$

Forming the Lagrangian, taking the gradient and setting it equal to zero we readily obtain

$$\big(\hat{\Sigma} - aI\big)\boldsymbol{v} = b\boldsymbol{v}_1, \tag{4}$$

where we used a and b for the Lagrange multipliers, to avoid confusion with the eigenvalues. Left multiplying both sides with $\boldsymbol{v}_1^T$, using the second of the constraints in the last equation and recalling that $\boldsymbol{v}_1$ is an eigenvector of the sample covariance matrix, results in $b = 0$, or

$$\hat{\Sigma}\boldsymbol{v} = a\boldsymbol{v}.$$

Thus the maximizing vector is an eigenvector. Plugging it into the cost and taking into account the second of the constraints we have that

$$\boldsymbol{v}^T \hat{\Sigma} \boldsymbol{v} = a$$

which is maximized if a is equal to the second largest eigenvalue. Note that since the matrix is symmetric, eigenvectors corresponding to different eigenvalues are orthogonal.

For the general case, using induction, the proof follows similar arguments.

19.2. Show that the pair of directions, associated with CCA, which maximize the respective correlation coefficient, satisfy the following pair of relations,

$$\begin{aligned} \Sigma_{xy}\boldsymbol{u}_y &= \lambda \Sigma_{xx}\boldsymbol{u}_x,\\ \Sigma_{yx}\boldsymbol{u}_x &= \lambda \Sigma_{yy}\boldsymbol{u}_y. \end{aligned}$$

Solution: The corresponding Lagrangian is given by

$$L(\boldsymbol{u}_x, \boldsymbol{u}_y, \lambda_x, \lambda_y) = \boldsymbol{u}_x^T \Sigma_{xy} \boldsymbol{u}_y - \frac{\lambda_x}{2}\left(\boldsymbol{u}_x^T \Sigma_{xx} \boldsymbol{u}_x - 1\right) - \frac{\lambda_y}{2}\left(\boldsymbol{u}_y^T \Sigma_{yy} \boldsymbol{u}_y - 1\right),$$

and taking the gradients and equating to zero we get

$$\begin{aligned} \Sigma_{xy}\boldsymbol{u}_y &= \lambda_x \Sigma_{xx}\boldsymbol{u}_x,\\ \Sigma_{yx}\boldsymbol{u}_x &= \lambda_y \Sigma_{yy}\boldsymbol{u}_y, \end{aligned}$$

where we used the fact that $\Sigma_{xy} = \Sigma_{yx}^T$. Multiplying the first one by $\boldsymbol{u}_x^T$ and the second one by $\boldsymbol{u}_y^T$ and taking into account the constraints, it is readily seen that

$$\lambda_x = \lambda_y := \lambda,$$

which proves the claim.

19.3. Establish the arguments that verify the convergence of the k-SVD.

Solution: Let us first assume that we can perform the sparse coding stage perfectly; that is, we can retrieve the best approximation to $\boldsymbol{x}_n$, $n = 1, 2, \ldots, N$. that contains no more than T_0 nonzero entries. In this case, and assuming a fixed dictionary, each sparse coding step decreases the total representation error; this is the way that any greedy-like algorithm works.

During the second stage, since the Frobenius error norm is minimized, the error is guaranteed not to increase, and at the same time the sparsity constraint is not violated. Executing a series of such steps ensures a monotonic MSE reduction, and therefore the algorithm converges in a local minimum.

19.4. Prove that (19.83) and (19.89) are the same.

Proof: We have that

$$\hat{\boldsymbol{z}} = \frac{1}{\sigma^2}\Sigma_{z|x}A^T\boldsymbol{x} = \frac{\sigma^2}{\sigma^2}\left(\sigma^2 I + A^T A\right)^{-1} A^T\boldsymbol{x},$$

or

$$\hat{\boldsymbol{z}} = \frac{1}{\sigma^2}\left(I + A^T\frac{A}{\sigma^2}\right)^{-1} A^T\boldsymbol{x}.$$

Using the matrix inversion lemma

$$(I + AB)^{-1}A = A(I + BA)^{-1},$$

we obtain

$$\hat{\boldsymbol{z}} = \frac{1}{\sigma^2}A^T\left(I + \frac{A}{\sigma^2}A^T\right)^{-1}\boldsymbol{x} = A^T\Sigma_x^{-1}\boldsymbol{x},$$

where

$$\Sigma_x = \sigma^2 I + AA^T.$$

19.5. Show that the ML PPCA tends to PCA as $\sigma^2 \rightarrow 0$.

Solution: Recall that PCA relies on the transformation

$$\boldsymbol{y} = \tilde{A}^T\boldsymbol{x},$$

where $\tilde{A}$ comprises the principal eigenvectors of the covariance (for zero mean) matrix and that in this case

$$\Sigma_y = \tilde{A}^T \Sigma_x \tilde{A} = \Lambda.$$

If we impose on $\boldsymbol{y}$ to have identity covariance matrix as the latent variables in PPCA, we make the transformation

$$\boldsymbol{z} = \Lambda^{-1/2}\boldsymbol{y} = \Lambda^{-1/2}\tilde{A}^T\boldsymbol{x},$$

since this results in $\Sigma_z = I$. Recall now that the ML estimation in PPCA, for $\sigma^2 = 0$ results in

$$\Sigma_x = \mathbf{0} + AA^T = \tilde{A}\Lambda\tilde{A}^T.$$

Thus, taking into account that A is a factor of Σ_x and recalling from the theory that

$$\begin{aligned}\hat{\boldsymbol{z}} &= A^T\Sigma_x^{-1}\boldsymbol{x} = \Lambda^{1/2}\tilde{A}^T(\tilde{A}\Lambda^{-1}\tilde{A}^T)\boldsymbol{x} \\ &= \Lambda^{-1/2}\tilde{A}^T\boldsymbol{x}.\end{aligned}$$

hence $\hat{\boldsymbol{z}} = \boldsymbol{z}$. Note that we have used the orthogonality property $\tilde{A}^{-1} = \tilde{A}^T$.

19.6. Show Eqs. (19.91)-(19.92)

Solution: Our starting point are the relations given in the text,

$$\begin{aligned}\mathcal{Q}(A,\beta;A^{(j)},\beta^{(j)}) = -\sum_{n=1}^{N}\Big(&-\frac{l}{2}\ln\beta + \frac{1}{2}\|\boldsymbol{\mu}_{z|x}^{(j)}(n)\|^2 + \frac{1}{2}\text{trace}\{\Sigma_{z|x}^{(j)}\} \\ &+\frac{\beta}{2}\|\boldsymbol{x}_n - A\boldsymbol{\mu}_{z|x}^{(j)}(n)\|^2 + \frac{\beta}{2}\text{trace}\{A\Sigma_{z|x}^{(j)}A^T\}\Big) + C,\end{aligned} \tag{5}$$

where C is a constant and

$$\boldsymbol{\mu}_{z|x}^{(j)}(n) = \beta^{(j)}\Sigma_{z|x}^{(j)}A^{(j)T}\boldsymbol{x}_n,\ \Sigma_{z|x}^{(j)} = (I + \beta^{(j)}A^{(j)T}A^{(j)})^{-1}.$$

Eq. (5) is rewritten as

$$\begin{aligned}\mathcal{Q}(A,\beta;A^{(j)},\beta^{(j)}) = \frac{lN\ln\beta}{2} &- \frac{\beta}{2}\sum_{n=1}^{N}(\boldsymbol{x}_n - A\boldsymbol{\mu}(n))^T(\boldsymbol{x}_n - A\boldsymbol{\mu}(n)) \\ &- \frac{N\beta}{2}\text{trace}\{A\Sigma A^T\} + \text{constants},\end{aligned}$$

where for notational simplicity we have used $\boldsymbol{\mu}(n)$ instead of $\boldsymbol{\mu}_{z|x}^{(j)}(n)$ and Σ in place of $\Sigma_{z|x}^{(j)}$. Elaborating on the previous equation we get

$$\begin{aligned}\mathcal{Q}(A,\beta;A^{(j)},\beta^{(j)}) = & \frac{lN\ln\beta}{2} - \frac{\beta}{2}\sum_{n=1}^{N}\boldsymbol{x}_n^T\boldsymbol{x}_n + \beta\text{trace}\left\{A^T\sum_{n=1}^{N}\boldsymbol{x}_n\boldsymbol{\mu}^T(n)\right\} \\ & - \frac{\beta}{2}\text{trace}\left\{A\left(\sum_{n=1}^{N}\boldsymbol{\mu}(n)\boldsymbol{\mu}^T(n)\right)A^T\right\} - \frac{N\beta}{2}\text{trace}\{A\Sigma A^T\} \\ & + \text{constant}. \qquad (6)\end{aligned}$$

Recalling now the following identities

$$\begin{aligned}\frac{\partial}{\partial A}\text{trace}\{ABA^T\} &= A(B+B^T), \\ \frac{\partial}{\partial A}\text{trace}\{A^TB\} &= B, \\ \frac{\partial}{\partial A}\text{trace}\{AB\} &= B^T,\end{aligned}$$

and using them to compute the derivative with respect to A of the right hand side of (6) and equating to zero, we get

$$\beta\sum_{n=1}^{N}\boldsymbol{x}_n\boldsymbol{\mu}^T(n) - \beta A\sum_{n=1}^{N}\boldsymbol{\mu}(n)^T\boldsymbol{\mu}(n) - \beta NA\Sigma = O$$

which then results in Eq. (19.102).

Eq. (19.88) results in a straightforward way by taking the derivative w.r to β and equating to zero.

19.7. Show equation (19.102).

Solution: Our starting point is the eigenvalue/eigenvector equation

$$\Sigma\boldsymbol{u} = \lambda\boldsymbol{u}, \qquad (7)$$

where

$$\boldsymbol{u} = \sum_{n=1}^{N}a_n\boldsymbol{\phi}(\boldsymbol{x}_n) = \Phi^T\boldsymbol{a}, \qquad (8)$$

where

$$\Phi^T := [\boldsymbol{\phi}(\boldsymbol{x}_1),\ldots,\boldsymbol{\phi}(\boldsymbol{x}_N)].$$

Employing the last definition and (8), we can rewrite (7) as

$$\frac{1}{N}\Phi^T\Phi\Phi^T\boldsymbol{a} = \lambda\Phi^T\boldsymbol{a},$$

or finally

$$\Phi^T\big(\frac{1}{N}\Phi\Phi^T\boldsymbol{a} - \lambda\boldsymbol{a}\big) = \mathbf{0}.$$

If Φ is full rank, then the claim has been proved.

19.8. Show that the number of degrees of freedom of a rank r matrix is equal to $r(l_1 + l_2) - r^2$.

Solution: A rank r matrix can be expressed via SVD as in (19.119). The first column of U has l_1 parameters whereas the second one, since it is orthogonal to the first one, it is fully described with $l_1 - 1$ parameters. In the same manner, the third one is fully described with $l_1 - 2$ parameters because it is orthogonal the the first two, etc. The same is true for V and r extra parameters are needed for the description of the singular values. So in total, the number of degrees of freedom is $r(l_1+l_2)-2(1+, \ldots, l_1-1)+r$ which directly leads to the answer.