A concise correction

EO

October 2022

1 Week 1

1.1 Exercise 2.1

1. For a fixed $\mathbf{X}$, $\lim_{\|w\|\to\infty} E(w) = \infty$, thus there always exists a minimum to this continuous function.

2. $X^TXw = 0 \Rightarrow w^TX^TXw = 0 \Rightarrow \|Xw\| = 0$; at the same time, $\|Xw\| = 0 \Rightarrow Xw = 0 \Rightarrow XX^Tw = 0$. Thus, $\ker(X^TX) = ker(X)$.

3. Here, $\inf_w \|Xw - y\|^2 = \inf_{z\in\text{span}(X)} \|z - y\|^2$. Thus, by orthogonal projection theorem on a subspace, the minimum is reached for $z = \pi_{\text{span}(X)}(y)$, where $\pi_{\text{span}(X)}$ is the orthogonal projection on span(X). It implies that for any $w \in \text{span}(X)$, $w \perp \pi_{\text{span}(X)}(y) - y$, equivalent to $x_i \perp \pi_{\text{span}(X)}(y) - y$ and thus, $x_i^T\pi_{\text{span}(X)}(y) = x_i^Ty$. It implies that $X^T\pi_{\text{span}(X)}y = X^Ty$; but by definition, there is $\hat{w}$ such that $\pi_{\text{span}(X)}y = X\hat{w}$. This writes $X^TX\hat{w} = X^Ty$, and thus $\hat{w} = (X^TX)^{-1}X^Ty$

4. Note that $y = Xw + \sigma\epsilon$. $\hat{w}$ is a linear combination of random Gaussian variables and thus is a Gaussian. To characterize it, we note that, $\mathbb{E}[\hat{w}] = (X^TX)^{-1}(X^TXw) = w$. Next, $\mathbb{E}[\hat{w}] = \mathbb{E}[\sigma\big((X^TX)^{-1}X^T\big)\big((X^TX)^{-1}X^T\big)^T] = \sigma(X^TX)^{-1}$. The same applies for $\hat{\epsilon}$, and it's clear that $\mathbb{E}[\hat{\epsilon}] = Xw - Xw = 0$. Next, $\mathbb{E}[\hat{\epsilon}\hat{\epsilon}^T] = [I - X(X^TX)^{-1}X^T][I - X(X^TX)^{-1}X^T]^T = I - X(X^TX)^{-1}X^T$

5. $\hat{w}$ is an unbiased estimator of w and $\mathbb{E}[\|\hat{\epsilon}\|^2] = \sigma^2$ thus $\hat{\sigma} = \|\hat{\epsilon}\|^2$ is an unbiased estimator.

6. Let $\tilde{w}_A$ associated to A. It implies that for any $w \in \mathbb{R}$, $\mathbb{E}[\tilde{w}_A] = w = A\mathbb{E}[y] = AXw$, which means that $AX = I$. Next, $\mathbb{E}[(\tilde{w}_A - w)(\tilde{w}_A - w)^T] = \mathbb{E}[(Ay - AXw)(Ay - AXw)^T] = A\mathbb{E}[(y - Xw)(y - Xw)^T]A^T = \sigma^2AA^T$. Now $\mathbf{I} = \mathbf{I} - X(X^TX)^{-1}X^T + X(X^TX)^{-1}X^T$ which is a sum of two projections. Thus, $AA^T = A\big(\mathbf{I} - X(X^TX)^{-1}X^T + X(X^TX)^{-1}X^T\big)A^T = A\big(\mathbf{I} - X(X^TX)^{-1}X^T\big)A^T + (X^TX)^{-1}$ and we get the desired result, as the left term is a positive semi-definite matrix.

2 Week 2

2.1 Exercise 3.1

1. Let $u = (u_1, ..., u_n)$, then, $u^T X = \sum_{i=1}^n u_i G_i$. We admit that the sum of two independent real-valued Gaussian is a Gaussian, thus by induction, $u^T X$ is a real-valued Gaussian. Furthermore, $\mathbb{E}[u^T X] = \sum_{i=1}^n u_i \mathbb{E}[G_i] = 0$ and by independence, $\mathbb{E}[(u^T X)^2] = \sum_{i=1}^n u_i^2 \mathbb{E}[G_i^2] = \|u\|$, thus $X \sim \mathcal{N}(0, \sigma \mathbf{I})$

2. Write $\Sigma = PDP^*$ where $D = (d_1..., d_n)$ is a diagonal matrix, set $\sqrt{D} = (\sqrt{d_1}..., \sqrt{d_n})$ set $L = P\sqrt{D}P^*$, then $LL^T = \Sigma$. Now, let $Y = LX + m$ so that $\mathbb{E}[Y] = m$ and $\mathbb{E}[(Y - m)(Y - m)^T] = \Sigma$. Thus, it's enough to simulate n independent real-valued standard normal variables and to perform the affine transformation $x \to Lx + m$.

2.2 Exercise 3.2

1. X_3 and (X_1, X_2) are independent.

2. $(X_1, X_2) \sim \mathcal{N}(0, \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix})$

3. By linear operations on Gaussian variables, it remains a Gaussian vector.

4. $\mathbb{E}(X_2(X_2 + aX_1)) = 2 + a$. If $a = -2$, we get independence. Thus, $\mathbb{E}[X_2 - 2X_1|X_2] = \mathbb{E}[X_2 - 2X_1] = 0$ and $\mathbb{E}[X_1|X_2] = \frac{X_2}{2}$.

2.3 Exercise 3.3

Again, the symbol $\propto$ means "up to a constant not crucial for integration".

1. By direct calculations.

2. Here:

$$-\frac{\beta}{2}(y_i - \sum_{j=1}^{d} w_j \phi_j(x_i))^2 = -\frac{\beta}{2}(y_i - \phi(x_i)^T w)^2$$

. We thus write, $y = [y_1...y_n]^T$ and $\Phi_x = [\phi(x_1)...\phi(x_n)]^T$

3. Here:

$$\begin{aligned}
p(w|\mathcal{D}) &= \frac{p(w)p_w(\mathcal{D})}{p(\mathcal{D}} && (1)\\
&\propto p(w)p_w(\mathcal{D}) && (2)\\
&= \exp(-\frac{\beta}{2}\|y - \Phi_x w\|^2)\exp(-\frac{1}{2}(w - m_0)^T S_0^{-1}(w - m_0)) && (3)\\
&\propto \exp(-\frac{1}{2}w^T(\Phi_x^T \Phi_x + S_0^{-1})w - w^T(\beta\Phi_x^T y + S_0^{-1}m_0)) && (4)\\
&\propto p_{\mathcal{N}(\mu,\Sigma)}(w) && (5)
\end{aligned}$$

where, from the first question, $\Sigma = (\Phi_x^T \Phi_x + S_0^{-1})^{-1}$ and $\mu = \Sigma\beta\Phi_x^T y + S_0^{-1}m_0$. Now the final argument is as follow: since,

$$\int_w p(w|\mathcal{D}) = 1\,,$$

then we can identity $p(w|\mathcal{D})$ to $p_{\mathcal{N}(\mu,\Sigma)}(w)$.

4. - Maximum a posteriori Estimation (MAP): we remind that its definition given by:

$$\hat{w}^{\mathbf{MAP},\mathcal{D}} \triangleq \arg\max_w p(w|\mathcal{D})$$

.

Taking the gradient w.r.t. w, we note that $\nabla_w p(w|\mathcal{D}) = 0$ is equivalent to:

$$w^{\mathbf{MAP},\mathcal{D}} = (\Phi_x^T \Phi_x + S_0^{-1})^{-1} \big(\beta \Phi_x^T y + S_0^{-1} m_0\big)$$

In the context of the question,

$$w^{\mathbf{MAP},\mathcal{D}} = \beta(\Phi_x^T \Phi_x + \alpha \mathbf{I})^{-1} \Phi_x^T y$$

- Posterior mean: we remind that by definition,

$$w^{\mathbf{PM},\mathcal{D}} \triangleq \int_w w p(w|\mathcal{D})\, dw$$

In the context of the question, given the distribution corresponds to a Gaussian,

$$w^{\mathbf{PM},\mathcal{D}} = \beta \Phi_x^T y$$

5. Our goal is to compute $p(\tilde{y}_{new}|\tilde{x}_{new}, \mathcal{D})$. One (informal) way to understand this quantity is that (y, x) are new variables (obtained from a similar process as $\mathcal{D}$), which will not affect the current estimate of w which is obtained from $\mathcal{D}$. !! Be careful, the notation $p(a, b, c, d, ...)$ for the density is overloaded and can have very different meaning depending on the context... !! Consequently, via Bayes rule:

$$\begin{aligned}
p(\tilde{y}_{new}|\tilde{x}_{new}, \mathcal{D}) &\triangleq \frac{p(\tilde{y}_{new})}{p(\tilde{x}_{new}, \mathcal{D})} && (6)\\
&= \int_w \frac{p(\tilde{y}_{new}|w, \tilde{x}_{new}) p(w, \tilde{x}_{new})}{p(\tilde{x}_{new}, \mathcal{D})} && (7)\\
&= \int_w \frac{p(\tilde{y}_{new}|w, \tilde{x}_{new}) p(w) p(\tilde{x}_{new})}{p(\tilde{x}_{new}) p(\mathcal{D})}, \text{ by the mentioned independence} && (8)\\
&= \int_w p(\tilde{y}_{new}|w, \tilde{x}_{new}) p(w|\mathcal{D})\, dw && (9)
\end{aligned}$$

We identify $p_w(y|x)$ with $p(y|w, x)$ and we omit the "new" to mention the variables. At this stage, we remark that:

$$p_w(y|x) \triangleq \frac{p_w(y, x)}{p_w(x)} \propto e^{-\frac{\beta}{2}(y - w^T \phi(x))^2}$$

At the same time, from question 3,

$$p(w|\mathcal{D}) \propto p_{\mathcal{N}(\beta\Phi_x^T y, (\Phi_x^T\Phi_x + \alpha\mathbf{I}))}$$

We thus recognize the law of $Z = Y + W^T\phi(x)$ where $Y \sim \mathcal{N}(0, \beta^{-1})$ and $W \sim \mathcal{N}((\Phi_x^T\Phi_x + \alpha\mathbf{I})^{-1}\beta\Phi_x^T y, (\Phi_x^T\Phi_x + \alpha\mathbf{I}))$ are two independent variables. By linearity of Gaussian variables, this is as well a Gaussian and:

$$\mathbb{E}[Z] = (\Phi_x^T\Phi_x + \alpha\mathbf{I})^{-1}\beta\Phi_x^T y^T\phi(x)$$

and

$$\mathbb{E}[(Z - \mathbb{E}Z)^2] = \frac{1}{\beta^2} + \phi(x)^T(\Phi_x^T\Phi_x + \alpha\mathbf{I})^{-1}\phi(x)$$

Thanks to this, we could for instance compute the posterior predictive mean (see slide 34/42, which leads to the Bayes estimator), given by:

$$\hat{y}^{*,\mathcal{D}}(x) \triangleq \int_y yp(y|x, \mathcal{D})\, dy (\triangleq \mathbb{E}[Y|X = x, \mathcal{D}])$$

which recovers the result. Note that the slide 34/42 has a typo, as the dependency in $\mathcal{D}$ should be explicit in Eq 69/70.

6. This question is straightforward given the previous one. Again, we recognize that $p(y|x)$ is the density of:

$$Z = Y + W^T\phi(x)\,,$$

where $Y \sim \mathcal{N}(0, \beta^{-1})$ and $W \sim \mathcal{N}(0, \alpha^{-1}\mathbf{I})$. Its density is thus clearly, following the same argument: $\mathcal{N}(0, \beta^{-1} + \alpha^{-1}\phi(x)^T\phi(x))$ (check it!)

Note how significantly different are the result of question 5 and 6. In particular, note that the estimator $\hat{y}^{*,\mathcal{D}}$ of question 5 will be specific to the data $\mathcal{D}$.

3 Week 3:

3.1 Exercise 4.1

Reminder:

$$\int_{\mathbb{R}_+} y^n e^{-y} dy = n!$$

Method 1. By Bayes rule, $p_{(X,Y)}(n, y) = p_Y(y)p_{X|Y}(n|y) = e^{-y} \times \frac{y^n}{n!}e^{-y}$. Next, X is a marginal of (X, Y), so:

$$p_X(n) = \int_{\mathbb{R}^+} p_{(X,Y)}(n, y)\, dy = \int_{\mathbb{R}} \frac{y^n}{n!}e^{-2y}\, dy = \frac{1}{2^{n+1}}$$

Next, $p_{Y|X}(n|y) = \frac{p_{(X,Y)(n,y)}}{p_X(n)} = \frac{y^n}{n!}e^{-2y}2^{n+1}$ and:

$$\mathbb{E}[Y|X = n] = \int_{\mathbb{R}_+} y \times \frac{y^n}{n!}e^{-2y}2^{n+1}\, dy = \frac{n+1}{2}$$

Thus, $\mathbb{E}[Y|X] = \frac{X+1}{2}$.

Method 2. (longer but can be safer)

Reminder: p is the density of X if for any continuous bounded function,

$$\mathbb{E}[f(X)] = \int_{\mathcal{X}} f(x)p(x)d\mu(x)$$

and $Z = \mathbb{E}[X|Y]$ if and only if $Z = \psi(Y)$ for some ψ, and for any f bounded continuous:

$$\mathbb{E}[f(Y)X] = \mathbb{E}[f(Y)Z]$$

.

Let $f : \mathbb{N} \times \mathbb{R} \to \mathbb{R}$ bounded such that $x \to f(n,x)$ is continuous, then

$$\mathbb{E}[f(X,Y)] = \int_{\mathbb{R}} \sum_n f(n,y)e^{-y}\frac{y^n}{n!}e^{-y}\,dy \tag{10}$$

$$= \int_{\mathbb{R}} \sum_n f(n,y)\frac{y^n}{n!}e^{-2y}\,dy \tag{11}$$

Consquently, the density of (X,Y) is given by $p_{(X,Y)}(n,y) = \frac{y^n}{n!}e^{-2y}$. Then,

$$\mathbb{E}[f(X)] = \int_{\mathbb{R}} \sum_n f(n)\frac{y^n}{n!}e^{-2y}\,dy \tag{12}$$

$$= \sum_n f(n) \int_{\mathbb{R}} \frac{y^n}{n!}e^{-2y}\,dy \tag{13}$$

$$= \sum_n f(n)\frac{1}{2^{n+1}} \tag{14}$$

The density of X is given by $p_X(n) = \frac{1}{2^{n+1}}$, which is the density of a geometric law with parameter $\frac{1}{2}$. Next, let f bounded continuous, then:

$$\mathbb{E}[Yf(X)] = \int_{\mathbb{R}} \sum_n yf(n)\frac{y^n}{n!}e^{-2y}\,dy \tag{15}$$

$$= \sum_n f(n) \int_{\mathbb{R}} \frac{y^{n+1}}{n!}e^{-2y}\,dy \tag{16}$$

$$= \sum_n f(n)\frac{n+1}{2^{n+2}} \tag{17}$$

$$= \sum_n f(n)\frac{n+1}{2} \times \frac{1}{2^{n+1}} \tag{18}$$

$$\tag{19}$$

$$= \mathbb{E}[f(X)\frac{X+1}{2}] \tag{20}$$

Thus, $\mathbb{E}[Y|X] = \frac{X+1}{2}$

3.2 Exercise 4.2

(i) $\mathcal{P}(\lambda + \mu)$

Method 1: (Getting (iii) without (ii) - faster here if one doesn't recognize a standard density)

With $k = n + m$

$$\mathbb{E}[Xf(S)] = \mathbb{E}[Xf(X+Y)] \tag{21}$$
$$= \sum_{n,m} n \frac{\lambda^n}{n!} e^{-\lambda} \frac{\mu^m}{m!} e^{-\mu} f(n+m) \tag{22}$$
$$= \sum_k f(k) e^{-\mu-\lambda} \sum_{n=1}^{k} \frac{1}{(k-n)!(n-1)!} \lambda^n \mu^{k-n} \tag{23}$$
$$= \sum_k f(k) e^{-\mu-\lambda} \sum_{n'=0}^{k-1} \frac{1}{(k-1-n')!n'!} \lambda^{n'+1} \mu^{k-1-n'} \text{ with } n' = n-1 \tag{24}$$
$$= \sum_k f(k) e^{-\mu-\lambda} \frac{1}{(k-1)!} \lambda (\lambda+\mu)^{k-1} \tag{25}$$
$$= \sum_k f(k) \frac{\lambda k}{\lambda+\mu} \times \frac{(\lambda+\mu)^k}{k!} e^{-\mu-\lambda} \tag{26}$$
$$= \mathbb{E}[f(S) S \frac{\lambda}{\lambda+\mu}] \tag{27}$$

Thus, $\mathbb{E}[X|S] = \frac{\lambda}{\lambda+\mu} S$

Method 2: (ii) (given some computations are done above, I skip them) for $n \leq k$ $p_{(X|S)}(n,k) = \frac{p_{(X,S)}(n,k)}{p_S(k)} = \frac{\mathbb{P}(X=n)\mathbb{P}(Y=k-n)}{\mathbb{P}(S=k)} = \binom{k}{n} (\frac{\lambda}{\lambda+\mu})^n (\frac{\mu}{\lambda+\mu})^{k-n}$ The conditional density is a Binomial of parameter $k, \frac{\lambda}{\lambda+\mu}$.

(iii) Next, either one remember the mean of a binomial, either:

$$\mathbb{E}[X|S=k] = \sum_{n=0}^{k} n \binom{k}{n} (\frac{\lambda}{\lambda+\mu})^n (\frac{\mu}{\lambda+\mu})^{k-n} \tag{28}$$
$$= ... \tag{29}$$
$$= \frac{\lambda k}{\lambda+\mu} \tag{30}$$

Thus, $\mathbb{E}[X|S] = \frac{\lambda S}{\lambda+\mu}$.

(iv) Clear, as the variance of a Poisson law $\mathcal{P}(\lambda)$ is λ.

3.3 Exercise 4.3

(i) $p_X(x) = 1_{[-1,0]}(1+x) + 1_{[0,1]}(1-x)$

(ii) $p_{X|D}(x|d) = 1_{[d,1]}\frac{1}{1-d}(x)$
(iii) $\mathbb{E}[X|D] = \frac{1+D}{2}$

3.4 Exercise 4.4

(i) $p_S(x) = \lambda^2 x e^{-\lambda x}$ (Erlang distribution)
(ii) $p(x|s) = \frac{1}{s}1_{x\le s}$ (uniform law)
(iii) $\mathbb{E}[X|S] = \frac{S}{2}$

4 Week 4

4.1 Exercise 5.1

1. We write, using a property of conditional expectations at each line:

$$\begin{aligned}
\mathrm{Risk}(\mathcal{C}) &= \mathbb{E}[1_{Y\neq\mathcal{C}(X)}] &(31)\\
&= \mathbb{E}[1_{Y=0}1_{\mathcal{C}(X)=1} + 1_{Y=1}1_{\mathcal{C}(X)=0}] &(32)\\
&= \mathbb{E}[\mathbb{E}[1_{Y=0}1_{\mathcal{C}(X)=1} + 1_{Y=1}1_{\mathcal{C}(X)=0}|X]] &(33)\\
&= \mathbb{E}[1_{\mathcal{C}(X)=1}\mathbb{E}[1_{Y=0}|X] + 1_{\mathcal{C}(X)=0}\mathbb{E}[1_{Y=1}|X]] &(34)\\
&= \mathbb{E}[1_{\mathcal{C}(X)=1}\mathbb{E}[(1-Y)|X] + 1_{\mathcal{C}(X)=0}\mathbb{E}[Y|X]] &(35)\\
&= \mathbb{E}[\mathcal{C}(X)(1-\eta(X)) + (1-1_{\mathcal{C}(X)=1})\eta(X)] &(36)\\
&= \mathbb{E}[\mathcal{C}(X)(1-2\eta(X)) + \eta(X)] &(37)
\end{aligned}$$

Now, we note that for any $x \in \mathbb{R}^d$, $\mathcal{C}(x)(1-2\eta(x)) \ge \mathcal{C}^*(x)(1-2\eta(x))$. Thus,

$$\mathrm{Risk}(\mathcal{C}) \ge \mathrm{Risk}(\mathcal{C}^*)$$

2. In this case $\mathcal{C}(x)(1-2\eta(x)) - \mathcal{C}^*(x)(1-2\eta(x)) \ge 0$ and has expectation 0. Thus $\mathcal{C}(X)(1-2\eta(X)) - \mathcal{C}^*(X)(1-2\eta(X)) = 0$ almost surely. Thus, almost surely again:

$$\mathcal{C}(X)(1-2\eta(X)) - \mathcal{C}^*(X)(1-2\eta(X)) = 0$$

Thus multiplying by $1_{\eta(X)\neq\frac{1}{2}}$

$$1_{\eta(X)\neq\frac{1}{2}}\mathcal{C}(X)(1-2\eta(X)) = 1_{\eta(X)\neq\frac{1}{2}}\mathcal{C}^*(X)(1-2\eta(X))$$

Thus(as the right term becomes non 0)

$$\mathcal{C}(X)1_{\eta(X)\neq\frac{1}{2}} = \mathcal{C}^*(X)1_{\eta(X)\neq\frac{1}{2}}$$

4.2 Exercise 5.2

If (x_i), (x'_i) are linearly separable, then there is $w, b, \epsilon > 0$ such that $x_i^T w + b \ge \epsilon$ for $i \le n$ and $x_i'^T w + b \le -\epsilon$ for $i \le n'$. Thus, if $\mathrm{conv}(x_i) \cap \mathrm{conv}(x_i) \neq \emptyset$, then there exists some λ_i, λ'_i such that $x = \sum_i \lambda_i x_i = \sum_i \lambda'_i x'_i$, where $\sum_i \lambda_i =$

$\sum_i \lambda'_i = 1$ and $\lambda_i \geq 0, \lambda'_i \geq 0$. In this case, $x^T w + b = \sum_i \lambda_i (x_i^T w + b) \geq \sum_i \lambda_i \epsilon = \epsilon$. At the same time, $x^T w + b \leq -\epsilon$ which is absurd.

Reciprocally (difficult), write $A = \text{conv}(x_i), B = \text{conv}(x'_i)$, then by compactness and continuity of the norm, there is $(a, b) \in A \times B$ such that $0 < \|a - b\| = \inf_{a' \in A, b' \in B} \|a' - b'\|$. Let $\varphi(\lambda) = \|\big(\lambda a + (1-\lambda)a'\big) - \big(\lambda b + (1-\lambda)b'\big)\|^2 \geq \|a - b\|^2$. We see that $\varphi(\lambda) = \lambda^2 \|a - b\|^2 + (1-\lambda)^2 \|a' - b'\|^2 + 2\lambda(1-\lambda)(a-b)^T(a' - b')$. It writes, after simplification by $(1-\lambda)$: $(1-\lambda)\|a' - b'\|^2 + 2\lambda(a-b)^T(a'-b') \geq (1+\lambda)\|a - b\|^2$ which implies for $\lambda \to 1$ that $2(a-b)^T(a'-b') \geq 2\|a-b\|^2$ for any $(a', b') \in A \times B$. With $b' = b$, $\langle a - b, a' \rangle \geq \langle a - b, a \rangle$ and $a' = a$, $\langle a - b, b' \rangle \leq \langle a - b, b \rangle$. Now, $\langle a - b, b \rangle - \langle a - b, b \rangle = \|a - b\|^2 > 0$: we have the linear separability.

5 Week 5

5.1 Exercise 6.1

Maximizing the likelihood is equivalent to maximizing the log-likelihood by monotony of the logarithm. The log-likelihood is given here, for $\mu \in \mathbb{R}^d \times \mathcal{S}_n^{++}$ by:

$$\mathcal{L}_{(\mu,\Sigma)}(X_1, ..., X_n) = \log(\prod_{i=1}^n \frac{1}{(2\pi)^{d/2} \det(\Sigma)^{1/2}} e^{-\frac{1}{2}(X_i - \mu)^T \Sigma^{,-1}(X_i - \mu))}) \tag{38}$$

$$= -\frac{nd}{2}\log(2\pi) - \frac{n}{2}\log\det(\Sigma^{)} - \frac{1}{2}\sum_{i=1}^n (X_i - \mu)^T \Sigma^{-1}(X_i - \mu) \tag{39}$$

Reminders: $\frac{\partial}{\partial x} x^T A x = Ax + A^T x$ and $\frac{\partial}{\partial x} b^T x = b^T$ (can be deduced from the formula on the trace) ; $(\mu, \Sigma) \to \mathcal{L}_{(\mu,\Sigma)}(X_1, ..., X_n)$ is concave, so that we can focus on critic points (ie, for which the gradient cancels).

1. An intermediary computation gives:

$$\nabla_\mu \mathcal{L}_{(\mu,\Sigma)}(X_1, ..., X_n) = \nabla_\mu \big(2 \times \frac{1}{2}\sum_{i=1}^n \mu^{,T} \Sigma^{-1} X_i - \frac{n}{2}\mu^{,T}\Sigma^{-1}\mu^T\big) \tag{40}$$

$$= \Sigma^{-1} \sum_{i=1}^n X_i - n\Sigma^{,-1}\mu \tag{41}$$

This vanishes iff:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^n X_i$$

2. For simplicity, we introduce $\tilde{\Sigma} = \Sigma^{-1}$, so that the log-likelyhood writes, using

the reminders,

$$\nabla_{\tilde{\Sigma}}\mathcal{L}_{(\mu,\Sigma)}(X_1,...,X_n) = \nabla_\Sigma(\frac{n}{2}\log\det(\tilde{\Sigma}^) - \frac{1}{2}\sum_{i=1}^{n}(X_i-\mu^)^T\tilde{\Sigma}^(X_i-\mu^)) \quad (42)$$

$$= \frac{n}{2}\tilde{\Sigma}^{,T,-1} - \frac{1}{2}\nabla_\Sigma \mathrm{Tr}(\sum_{i=1}^{n}(X_i-\mu^)^T\tilde{\Sigma}(X_i-\mu^)) \quad (43)$$

$$= \frac{n}{2}\tilde{\Sigma}^{T,-1} - \frac{1}{2}\nabla_\Sigma \mathrm{Tr}(\sum_{i=1}^{n}(X_i-\mu^)(X_i-\mu^)^T\tilde{\Sigma}) \quad (44)$$

$$= \frac{n}{2}\tilde{\Sigma}^{T,-1} - \frac{1}{2}\sum_{i=1}^{n}(X_i-\mu^)(X_i-\mu^)^T \quad (45)$$

We look for the points such that $\nabla_{(\mu,\Sigma)}\mathcal{L}_{(\hat{\mu},\hat{\Sigma})}(X_1,...,X_n) = 0$, leading to:

$$\hat{\Sigma} = \sum_{i=1}^{n}(X_i-\hat{\mu})(X_i-\hat{\mu})^T$$

Note: Here, if X_i sampled data are Gaussian and if $d \leq n$, then almost surely $\sum_{i=1}^{n}(X_i-\mu)(X_i-\mu)^T$ is positive definite. Otherwise, this must be an assumption of the exercise.

5.2 Exercise 6.2

1. Method a: Let $f : \mathbb{R}^d \times \{0,1\}$ bounded continuous, then:

$$\mathbb{E}[f(X,Y)] = \mathbb{E}[(1_{Y=0}+1_{Y=1})f(X,Y)] \quad (46)$$

$$= \mathbb{E}[\mathbb{E}[(1_{Y=0}+1_{Y=1})f(X,Y)|Y]] \quad (47)$$

$$= \mathbb{E}[(1_{Y=0}+1_{Y=1})\mathbb{E}[f(X,Y)|Y]] \quad (48)$$

$$= \mathbb{E}[1_{Y=0}\int_{\mathbb{R}^d} p_{\mathcal{N}(\mu_0,\Sigma_0)}(x)f(x,0)\,dx + 1_{Y=0}\int_{\mathbb{R}^d} p_{\mathcal{N}(\mu_0,\Sigma_0)}(x)f(x,1)\,dx] \quad (49)$$

$$= (1-\pi)\int_{\mathbb{R}^d} p_{\mathcal{N}(\mu_0,\Sigma_0)}(x)f(x,0)\,dx + \pi\int_{\mathbb{R}^d} p_{\mathcal{N}(\mu_1,\Sigma_1)}(x)f(x,1)\,dx \quad (50)$$

$$= \sum_{y\in\{0,1\}}\int_{\mathbb{R}^d} f(x,y)p(x,y)\,dx \quad (51)$$

where $p(x,0) = (1-\pi)p_{\mathcal{N}(\mu_0,\Sigma_0)}(x)$ and $p(x,1) = (1-\pi)p_{\mathcal{N}(\mu_1,\Sigma_1)}(x)$

Similarly, if we consider $\tilde{f}(x,y) = f(x)$, we obtain thus:

$$\mathbb{E}[f(X)] = \int_{\mathbb{R}^d} f(x)\big((1-\pi)p_{\mathcal{N}(\mu_0,\Sigma_0)}(x) + \pi p_{\mathcal{N}(\mu_1,\Sigma_1)}(x)\big)\,dx \quad (52)$$

Method b: By Bayes rule, $p(x,y) = p(y) \times p(x|y)$ and next $p(x) = (1-\pi)p(x,0) + \pi p(x,0)$

2. Method a. Let $f : \mathbb{R}^d \to \mathbb{R}$ bounded continuous, then:

$$
\begin{aligned}
\mathbb{E}[Yf(X)] &= \mathbb{E}[1_{Y=1} f(X)] && (53)\\
&= \mathbb{E}[\mathbb{E}[1_{Y=1} f(X)|Y]] && (54)\\
&= \mathbb{E}[1_{Y=1}\mathbb{E}[f(X)|Y]] && (55)\\
&= \mathbb{E}[1_{Y=1} \int_{\mathbb{R}^d} f(x) p_{\mathcal{N}(\mu_1,\Sigma_1)}(x)\, dx] && (56)\\
&= \pi \int_{\mathbb{R}^d} f(x) p_{\mathcal{N}(\mu_1,\Sigma_1)}(x)\, dx && (57)\\
&= \int_{\mathbb{R}^d} f(x) \frac{\pi p_{\mathcal{N}(\mu_1,\Sigma_1)}(x)}{(1-\pi)p_{\mathcal{N}(\mu_0,\Sigma_0)}(x) + \pi p_{\mathcal{N}(\mu_1,\Sigma_1)}(x)} \big((1-\pi)p_{\mathcal{N}(\mu_0,\Sigma_0)}(x) + \pi p_{\mathcal{N}(\mu_1,\Sigma_1)}(x)\big)\, dx && (58)
\end{aligned}
$$

Method b. Here, we can directly use the Bayes rule to get:

$$
\mathbb{E}[Y|X=x] = \mathbb{E}[1_{Y=1}|X=x] \triangleq p(y=1|X=x) = \frac{\pi p_{\mathcal{N}(\mu_1,\Sigma_1)}(x)}{(1-\pi)p_{\mathcal{N}(\mu_0,\Sigma_0)}(x) + \pi p_{\mathcal{N}(\mu_1,\Sigma_1)}(x)} \qquad (59)
$$

3/4. We thus have $\mathcal{C}^*(x) = 1$ if:

$$
\frac{\pi p_{\mathcal{N}(\mu_1,\Sigma_1)}(x)}{(1-\pi)p_{\mathcal{N}(\mu_0,\Sigma_0)}(x) + \pi p_{\mathcal{N}(\mu_1,\Sigma_1)}(x)} \geq \frac{1}{2} \qquad (60)
$$

This implies that:

$$
\pi p_{\mathcal{N}(\mu_1,\Sigma_1)}(x) \geq (1-\pi) p_{\mathcal{N}(\mu_0,\Sigma_0)}(x)
$$

This writes:

$$
\log \pi - \frac{1}{2}(x-\mu_1)^T \Sigma^{-1}(x-\mu_1) \geq \log(1-\pi) - \frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)
$$

This also writes:

$$
w^T x + b \geq 0
$$

with $b = \log \frac{\pi}{1-\pi} + \frac{1}{2}(\mu_0^T \Sigma^{-1}\mu_0 - \mu_1^T\Sigma^{-1}\mu_1)$ and $w = \Sigma^{-1}(\mu_1 - \mu_0)$.

5. Here,

$$
\begin{aligned}
\mathbb{P}(\mathcal{C}^*(X) = 1 | Y = 0) &= \int_{\mathbb{R}^d} 1_{w^T x + b \geq 0} p_{\mathcal{N}(\mu_0,\Sigma_0)}(x)\, dx && (61)\\
&= \mathbb{P}(w^T X + b \geq 0) && (62)
\end{aligned}
$$

where $X \sim \mathcal{N}(\mu_0, \Sigma_0)$. Now,

$$
\mathbb{E}[w^T X + b] = \mu_0 \Sigma^{-1}(\mu_1 - \mu_0) + \frac{1}{2}(\mu_0^T\Sigma^{-1}\mu_0 - \mu_1^T\Sigma^{-1}\mu_1) = -\frac{1}{2}(\mu_1-\mu_0)^T\Sigma^{-1}(\mu_1-\mu_0)
$$

At the same time,

$$\mathbb{E}[w^T(X-\mathbb{E}[X])(X-\mathbb{E}[X])^T w] = (\mu_1-\mu_0)\Sigma^{-1}\Sigma\Sigma^{-1}(\mu_1-\mu_0)$$

Following this, we note that:

$$\mathbb{P}(\mathcal{C}^*(X)=1|Y=0) = \mathbb{P}(\sqrt{d}Z - \frac{d}{2} \geq 0) = \mathbb{P}(Z \geq \frac{\sqrt{d}}{2}) \tag{63}$$

Now, by symmetry of the cumulative distribution of a Gaussian (which is $t \to \mathbb{P}(X \leq t)$, we get the result. Next, it's clear, by Bayes rule that and symmetry:

$$\mathbb{P}(\mathcal{C}^*(X) \neq Y) = \mathbb{P}(\mathcal{C}^*(X)=1|Y=0)\mathbb{P}(Y=0) + \mathbb{P}(\mathcal{C}^*(X)=0|Y=1)\mathbb{P}(Y=1) \tag{64}$$

$$= \Phi(-\frac{\sqrt{d}}{2}) \tag{65}$$

6. For Y, we write: $\mathcal{L}_\pi(Y_1,...,Y_n) = (1-\pi)^{n-\sum_i Y_i}\pi^{\sum_i Y_i}$, where $m = \sum_i Y_i$ This leads to minimizing: $(n-m)\log(1-\pi) + m\log\pi$ and taking the deriative leads to $\frac{n-m}{1-\hat{\pi}} = \frac{m}{\hat{\pi}}$ ie $\hat{\pi} = \frac{m}{n}$

We know that the samples Y_k for $k \in \{i, Y_i = 0\}$ or $k \in \{i, Y_i = 1\}$ follow the distribution $\mathcal{N}(\mu_0, \Sigma_0)$ and $\mathcal{N}(\mu_1, \Sigma_1)$ respectively. We can apply the result of exercise 6.1 to the likelihood of $p(x, y)$ to get:

$$\hat{\mu}_1 = \frac{1}{m}\sum_i Y_i X_i\,,$$

$$\hat{\mu}_0 = \frac{1}{n-m}\sum_i (1-Y_i)X_i\,,$$

$$\hat{\Sigma} = \frac{1}{n}\sum_i (X_i - \hat{\mu}_{Y_i})(X_i - \hat{\mu}_{Y_i})^T$$