Contents

Chapter 1

Solutions for exercises to chapter 1

Problem 1.1 - Closed form solution to polynomial regression

We use a slightly better notation to write this problem. Let X be the matrix of the form

$$X = \begin{bmatrix} x_1^0 & x_1^1 & \cdots & x_1^M \\ x_2^0 & x_2^1 & \cdots & x_2^M \\ \vdots & \vdots & \ddots & \vdots \\ x_N^0 & x_N^1 & \cdots & x_N^M \end{bmatrix}, \quad t = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{bmatrix}$$

The the problem can be rewritten in the following form:

$$E(w) = \frac{1}{2}\left((Xw-t)^T(Xw-t)\right).$$

Now we differentiate w.r.t w, note that

$$\begin{aligned} E(w+h) &= \frac{1}{2}(X(w+h)-t)^T(X(w+h)-t) \\ &= \frac{1}{2}\left((Xw-t)^T + (Xh)^T\right)(Xw-t+Xh) \\ &= \frac{1}{2}\left[(Xw-t)^T(Xw-t) + (Xw-t)^T Xh + (Xh)^T(Xw-t) + (Xh)^T(Xh)\right] \\ &= E(w) + \left\langle (Xw-t)^T, Xh \right\rangle + \frac{1}{2}\langle Xh, Xh\rangle \\ &= E(w) + \left\langle X^T(Xw-t), h\right\rangle + \frac{1}{2}\langle Xh, Xh\rangle. \end{aligned}$$

Note that $\left\langle X^T(Xw-t), h\right\rangle \in \mathrm{Hom}(\mathbb{R}^{M+1}, \mathbb{R})$ and

$$\frac{1}{2}\langle Xh, Xh\rangle \le \frac{1}{2}\|Xh\|\|Xh\| \le \frac{C}{2}\|X\|_\infty^2\|h\| \xrightarrow{\|h\|\to 0} 0,$$

it follows that $\nabla E(w) = X^T(Xw-t)$. Set it to zero and we get

$$X^T(Xw-t) = 0 \iff X^T Xw = X^\top t.$$

So X^TX is the A proposed in the problem.

$$\left[X^TX\right]_{ij} = \sum_{n=1}^{N}\left(x_n^i x_n^j\right) = \sum_{n=1}^{N} x_n^{i+j}, \text{ and } \left[X^Tt\right]_i = \sum_{n=1}^{N} x_n^i t_n,$$

as desired.

Problem 1.2 - Closed form solution to regularized polynomial regression

We use the same notation as in the previous problem and still rewrite the loss function in matrix form as follows:

$$\widetilde{E}(w) = \frac{1}{2}\langle Xw - t, Xw - t\rangle + \frac{\lambda}{2}\langle w, w\rangle .$$

Still we differentiate the expression. Note that if we let $\varphi(w) = \frac{\lambda}{2}\langle w, w\rangle$,we have that

$$\begin{aligned}\varphi(w+h) &= \frac{\lambda}{2}(w+h)^T(w+h)\\ &= \frac{\lambda}{2}\left(w^Tw + w^Th + h^Tx + \|h\|\right)\\ &= \varphi(w) + \langle \lambda w, h\rangle + \underbrace{\frac{\lambda}{2}\|h\|^2}_{=o(\|h\|)}.\end{aligned}$$

Therefore, $\nabla\varphi(w) = \lambda w$, and as a result

$$\nabla\widetilde{E}(w) = \nabla E(w) + \nabla\varphi(w) = X^T(Xw - t) + \lambda w.$$

Setting it to zero:

$$X^T(Xw - t) + \lambda w = 0 \iff (X^TX + \lambda I)w = X^Tt.$$

Hence, $(X^TX + \lambda I)$ and X^Tt are the corresponding matrices.

Problem 1.3 - Bayes formula warm up

According to the Bayes formula, we get that

$$\begin{aligned}P(\text{apple}) &= P(\text{apple}|\text{r})P(\text{r}) + P(\text{apple}|\text{g})P(\text{g}) + P(\text{apple}|\text{b})P(\text{b})\\ &= \frac{3}{10}\cdot\frac{2}{10} + \frac{1}{2}\frac{2}{10} + \frac{3}{10}\frac{6}{10} = \frac{17}{50}.\end{aligned}$$

And again, we can use formula to get

$$\begin{aligned}P(\text{g}|\text{orange}) &= \frac{P(\text{orange}|\text{g})P(\text{g})}{P(\text{orange}|\text{g})P(\text{g}) + P(\text{orange}|\text{b})P(\text{b}) + P(\text{orange}|\text{r})P(\text{r})}\\ &= \frac{\frac{3}{10}\frac{6}{10}}{\frac{3}{10}\frac{6}{10} + \frac{2}{10}\frac{1}{2} + \frac{2}{10}\frac{4}{10}}\\ &= \frac{1}{2}.\end{aligned}$$

Problem 1.4 - Nonlinear transform of likelihood function doesn't preserve its extrema

We first observe that if x_* maximizes the likelihood function $p_x(x)$, then $p'_x(x_*) = 0$. By chain rule, we have that

$$\begin{aligned}\frac{dp_x(g(y))\,|g'(y)|}{dy} &= \frac{dp_x(g(y))}{dy}\,|g'(y)| + p_x(g(y))\frac{d\,|g'(y)|}{dy} \\ &= \frac{dp_x(g(y))}{dg(y)}\frac{dg(y)}{dy}\,|g'(y)| + p_x(g(y))\frac{d\,|g'(y)|}{dy}. \end{aligned} \tag{1}$$

Hence, if $x_* = g(y_*)$, the

$$\frac{dp_x(g(y_*))}{dg(y_*)} = \frac{dp_x(x_*)}{dx_*} = 0.$$

However, there is no guarantee that the second term of the RHS of Eq. 1 is zero. For example, if $p_x(x) = 2x$ for $0 \le x \le 1$ and $x = \sin(y)$, where $0 \le y \le \pi/2$. Then according to the transformation formula, we have that

$$p_y(y) = p_x(g(y))g'(y) = 2\sin(y)\cos(y) = \sin(2y) \text{ for } 0 \le y \le \frac{\pi}{2}.$$

Clearly, $p_y(y)$ reaches its peak at $y = \pi/4$ but $\sin(\pi/4) \neq x_* = 1$. Thus, we have found a counterexample.

On the other hand, if $g(y)$ is an affine map, then $g'(y)$ is a constant map and as a result

$$\frac{d\,|g'(y)|}{dy} = 0$$

Problem 1.5 - Characterization of variance

It suffices to show that $\mathrm{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$ since any a measurable function of a random variable is again a random variable and in this case f although is not mentioned, it is safe to assume in this context that f is measurable. So note

$$\begin{aligned}\mathrm{Var}[X] &= \mathbb{E}[X - \mathbb{E}[X]]^2 \\ &= \mathbb{E}[X^2] - 2\mathbb{E}[X]\mathbb{E}[X] + (\mathbb{E}[X])^2 \\ &= \mathbb{E}[X^2] - \mathbb{E}[X]^2\end{aligned}$$

as desired.

Problem 1.6 - Covariance of two independent r.v. is zero

Since $X \perp Y$, then it follows that $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$. Then we have

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= \mathbb{E}[XY - X\mathbb{E}[Y] - \mathbb{E}[X]Y + \mathbb{E}[X]\mathbb{E}[Y]] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y] \\ &= \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y]\end{aligned}$$

$$= 0.$$

Problem 1.7 - Gaussian integral via polar coordinate

First, we write

$$\begin{aligned} I^2 &= \left(\int_{\mathbb{R}} \exp\left\{-\frac{1}{2\sigma^2}x^2\right\}dx\right)\left(\int_{\mathbb{R}} \exp\left\{-\frac{1}{2\sigma^2}y^2\right\}dy\right) \\ &= \int\int_{\mathbb{R}\times\mathbb{R}} \exp\left\{-\frac{1}{2\sigma^2}\left(x^2+y^2\right)\right\}dxdx. \end{aligned}$$

Now using polar coordinate - let $x = r\cos\theta$ and $y = r\sin\theta$. Then we get the Jacobian matrix as

$$\frac{\partial(x,y)}{\partial(r,\theta)} = \begin{bmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{bmatrix} \implies \left|\frac{\partial(x,y)}{\partial(r,\theta)}\right| = r(\cos\theta^2 + \sin\theta^2) = r.$$

Hence, as a result

$$\begin{aligned} I^2 &= \int_0^{2\pi}\int_0^{\infty} \exp\left\{-\frac{r^2}{2\sigma^2}\right\} r dr d\theta \\ &=_1 \int_0^{2\pi}\int_0^{\infty} \exp(-u)\sigma^2 du d\theta \\ &= \int_0^{2\pi} \sigma^2 d\theta \int_0^{\infty} \exp(-u)du \\ &= 2\pi\sigma^2\left[-\exp(-u)\right]_0^{\infty} = 2\pi\sigma^2. \end{aligned}$$

Problem 1.8 - Second moment of gaussian integral via Feymann's trick

The differentiation under the integral needs a bit more theoretical justification. We won't reproduce the related theorems here. But they could be found in e.g. Theorem 3.2, Theorem 3.3 in Chapter XIII of [Lan97] or in [Con00]. With this in mind, we get

$$\begin{aligned} \frac{d}{d\sigma^2}\int_{\mathbb{R}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}dx &= \int_{\mathbb{R}} \frac{d}{d\sigma^2}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}dx \\ &= \int_{\mathbb{R}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}(x-\mu)^2\left(-\frac{1}{2}\right)(\sigma^{-2})^2 dx \end{aligned}$$

On the the other hand, we have

$$\frac{d}{d\sigma^2}(2\pi\sigma^2)^{1/2} = -\frac{1}{2}(2\pi)(\sigma^2)^{-1/2}.$$

So combined together, we get

$$\int_{\mathbb{R}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}(x-\mu)^2\left(-\frac{1}{2}\right)(\sigma^{-2})^2 dx = \left(-\frac{1}{2}\right)(2\pi)^{1/2}(\sigma^2)^{-1/2}.$$

One step of reduction, we get

$$\begin{aligned}\mathbb{E}[(x-\mathbb{E}[x])^2] &= \mathrm{Var}[x] \\ &= \frac{1}{(2\pi\sigma^2)^{1/2}}\int_{\mathbb{R}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}(x-\mu)^2 dx \\ &= \sigma^2.\end{aligned}$$

And as a result,

$$\mathbb{E}[x^2] = \mathrm{Var}[x] + (\mathbb{E}[x])^2 = \sigma^2 + \mu^2.$$

Problem 1.9 - Gaussian density peaks at mean

It suffices to show the result holds in the multidimensional case since 1-dim is just a special case. Recall that the density of the Gaussian distribution in D dimension is

$$N(x|u,\Sigma) = \frac{1}{(2\pi)^{D/2}\left|\Sigma\right|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}.$$

Differentiate w.r.t. x and we get:

$$\nabla_x N(x|u,\Sigma) = \frac{1}{(2\pi)^{D/2}\left|\Sigma\right|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\} \nabla_x\left(\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right).$$

Now note that $\varphi(x) = (x-\mu)^T\Sigma^{-1}(x-\mu)$ for $x \in \mathbb{R}^d$, then note for any $h \in \mathbb{R}^D$

$$\begin{aligned}\varphi(x+h) &= (x-u+h)^T\Sigma^{-1}(x-\mu+h) \\ &= (x-\mu)^T\Sigma^{-1}(x-\mu+h) + h^T\Sigma^{-1}(x-\mu+h) \\ &= (x-\mu)^T\Sigma^{-1}(x-\mu) + (x-\mu)^T\Sigma^{-1}h + h^T\Sigma^{-1}(x-\mu) + h^T\Sigma^{-1}h \\ &= (x-\mu)^T\Sigma^{-1}(x-\mu) + \left\langle 2\Sigma^{-1}(x-\mu), h\right\rangle + h^T\Sigma^{-1}h\end{aligned}$$

Note that and

$$h^T\Sigma^{-1}h = \left\langle h\Sigma^{-1/2}, h\Sigma^{-1/2}\right\rangle \le \left\|h\Sigma^{-1/2}\right\|^2 \le C\left\|h\right\|^2\left\|\Sigma\right\|_\infty^2 = o(\|h\|),$$

and that $\left\langle 2\Sigma^{-1}(x-\mu), h\right\rangle \in \mathrm{Hom}(\mathbb{R}^d, \mathbb{R})$. It follows that

$$\nabla_x\varphi(x) = 2\Sigma^{-1}(x-\mu),$$

whence

$$\nabla_x\varphi(x) = 0 \iff 2\Sigma^{-1}(x-\mu) = 0 \iff x = \mu.$$

Problem 1.10 - Linearity of expectation and variance

1. Note

$$\mathbb{E}\left[x+y\right] = \int_{\mathrm{supp}(x)}\int_{\mathrm{supp}(y)} (x+y) f_{(x,y)}(x,y)dxdy$$

$$
\begin{aligned}
&= \int_{\text{supp}(x)} \int_{\text{supp}(y)} (x+y) f_x(x) f_y(y) dxdy \\
&= \int_{\text{supp}(x)} \int_{\text{supp}(y)} x f_x(x) f_y(y) dxdy + \int_{\text{supp}(x)} \int_{\text{supp}(y)} y f_x(x) f_y(y) dxdy \\
&= \int_{\text{supp}(x)} x f_x(x) dx \int_{\text{supp}(y)} f_y(y) + \int_{\text{supp}(x)} f_x(x) dx \int_{\text{supp}(y)} y f_y(y) \\
&= \mathbb{E}[x] + \mathbb{E}[y].
\end{aligned}
$$

2. Note

$$
\begin{aligned}
\text{Var}[x+y] &= \mathbb{E}[x+y]^2 - (\mathbb{E}[x+y])^2 \\
&= \mathbb{E}[x^2] + \mathbb{E}[y^2] + \underbrace{2\mathbb{E}[xy]}_{\mathbb{E}[x]\mathbb{E}[y]} - (\mathbb{E}[x])^2 - (\mathbb{E}[y])^2 - 2\mathbb{E}[x]\mathbb{E}[y] \\
&= \mathbb{E}[x^2] - (\mathbb{E}[x])^2 + \mathbb{E}[y^2] - (\mathbb{E}[y])^2 \\
&= \text{Var}[x] + \text{Var}[y].
\end{aligned}
$$

Problem 1.11 - MLE of gaussian

Recall that the log-likelihood function for Gaussian distribution is

$$
\ln p(x|\mu, \sigma^2) = -\frac{1}{2\sigma^2} \sum_{i=1}^{N} (x_n - \mu)^2 - \frac{N}{2} \ln \sigma^2 - \frac{N}{2} \ln(2\pi).
$$

Now we differentiate it w.r.t. μ and setting it to zero:

$$
\frac{\partial \ln p(x|\mu, \sigma^2)}{\partial \mu} = -\frac{1}{2\sigma^2} \cdot 2 \cdot \sum_{i=1}^{N} (x_n - \mu) = 0 \iff \sum_{i=1}^{N} (x_n - \mu) = 0 \iff \mu_{ML} = \frac{1}{n} \sum_{i=1}^{N} x_n.
$$

Now we differentiate it w.r.t. σ^2 and setting it to zero:

$$
\frac{\partial \ln(p|\mu, \sigma^2)}{\partial \sigma^2} = \underbrace{\sum_{n=1}^{N} (x_n - \mu)^2 \left(-\frac{1}{2}\right) (-1)(\sigma^2)^{-2} - \frac{N}{2\sigma^2} = 0}_{(\star)}.
$$

To rearrange, we get

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

Plug in $\mu = \mu_{ML}$ we get $\sigma^2_{ML} = \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu)^2$ as desired.

Problem 1.12 - Inconsistency gaussian MLE

Problem 1.14 - Independent terms of 2-nd order term in polynomial

We rewrite the sum in matrix form: $\sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}x_ix_j = x^TWx$, where $[W]_{ij} = w_{ij}$. Define

$$W_S = \frac{1}{2}(W + W^T) \text{ and } W_A = \frac{1}{2}(W - W^T).$$

Clearly, W_S is symmetric and $W_A^T = \frac{1}{2}(W^T - W) = -W_A$ is anti-symmetric and $W_S + W_A = W$. Therefore,

$$x^TWx = x^T(W_S + W_A)x = x^TW_Sx + x^TW_Ax.$$

Notice that

$$x^TW_Ax = \frac{1}{2}(x^TW_Sx - x^TW^Tx) = \frac{1}{2}(x^TW_Ss - x^TWx) = 0,$$

where the last inequality follows from the fact that x^TW^Tx is a scalar and is equal to x^TWx. Since we have shown the sum, $\sum_{i,j} w_{ij}x_ix_j$, only depends on a symmetric matrix, W_S, whose independent items is of the cardinality of $\sum_{i=1}^{D} i = D(D+1)/2$ if we assume its of dimension $D \times D$, we have established our claim.

Problem 1.15 - Independent terms of M-th order term in polynomial

1. Since by writing the M-th order in the form of

$$\sum_{i_1=1}^{D}\sum_{i_2=1}^{D}\cdots\sum_{i_M=1}^{D} w_{i_1,i_2,\cdots,i_M}x_{i_1}x_{i_2}\cdots x_{i_M}$$

introduces duplicate terms, e.g. if $w_{1,3,2}x_1x_3x_2$ and $w_{2,3,1}x_2x_3x_1$ are the same and can be combined into $(w_{1,3,2} + w_{2,3,1})x_1x_2x_3$, we can introduce an ordering that prevents such duplication from happening. Rewrite the sum in the newly introduced ordering yields

$$\sum_{i_1=1}^{D}\sum_{i_2=1}^{i_1}\cdots\sum_{i_M=1}^{i_{M-1}} w_{i_1,i_2,\cdots,i_M}x_{i_1}x_{i_2}\cdots x_{i_M}.$$

Thus, we have ll

$$\begin{aligned} n(D,M) &= \sum_{i_1=1}^{D}\sum_{i_2=1}^{i_1}\cdots\sum_{i_M=1}^{i_{M-1}} w_{i_1,i_2,\cdots,i_M}x_{i_1}x_{i_2}\cdots x_{i_M} \\ &= \sum_{i_1=1}^{D}\left(\sum_{i_2=1}^{i_1}\cdots\sum_{i_M=1}^{i_{M-1}} w_{i_1,i_2,\cdots,i_M}x_{i_1}x_{i_2}\cdots x_{i_M}\right) \\ &= \sum_{i_1=1}^{D} n(i_1, M-1). \end{aligned}$$

2. To show the equality holds using induction, we note for the base case of $D = 1$,

$$\text{LHS} = \frac{(1+M-2)!}{0!(M-1)!} = \frac{(M-1)!}{(M-1)!} = 1.$$

And

$$\text{RHS} = \frac{(1+M-1)!}{(D-1)!M!} = \frac{M!}{M!} = 1.$$

Now suppose $D = k$ and the equality holds. Then

$$\begin{aligned}
\sum_{i=1}^{k+1} \frac{(i+M-2)!}{(i-1)!(M-1)!} &= \sum_{i=1}^{k} \frac{(i+M-2)!}{(i-1)!(M-1)!} + \frac{(k+1+M-2)!}{k!(M-1)!} \\
&= \frac{(k+M-1)!}{(k-1)!M!} + \frac{(k+M-1)!}{k!(M-1)!} \qquad (1) \\
&= \frac{(k+M-1)!(k+M)}{k!(M-1)!} \\
&= \frac{(k+M)!}{k!M!} \\
&= \frac{((k+1)+M-1)!}{(k+1-1)!M!},
\end{aligned}$$

where Eq. (1) follows from induction hypothesis.

3. We establish the identity by inducting on M. By Problem 1.14, it follows that

$$n(D,2) = \frac{1}{2}D(D+1) = \frac{(D+2-1)!}{(D-1)!2!} = \frac{(D+1)!}{(D-1)!2!},$$

which proves the base case. Now suppose the statement holds for $M = k$. Then for $M = k+1$, we have

$$n(D,k+1) = \sum_{i=1}^{D} n(i,k) = \sum_{i=1}^{D} \frac{(i+M-2)!}{(i-1)!(M-1)!} = \frac{(D+M-1)!}{(D-1)!M!}$$

using part-2.

Problem 1.16 - Independent terms of high order polynomial

1. The first equality just follows from that summing up all the independent terms:

$$N(D,M) = \sum_{i=0}^{M} n(D,i).$$

2. We prove this inequality by inducting on M. Now for the base case, $M = 0$, we note that

$$\text{LHS} = n(D,0) = \frac{(D+0-1)!}{(D-1)!0!} = 1 = \frac{(D+0)!}{D!0!} = \text{RHS}.$$

Now assume that the claim holds for $M = k$. Then for $M = k + 1$, we have

$$\begin{aligned} N(D, k+1) &= \sum_{i=0}^{k} n(D, i) + n(D, k+1) \\ &= \frac{(D+k)!}{D!k!} + \frac{(D+k+1-1)!}{(D-1)!(k+1)!} \\ &= \frac{(D+k)!(D+k+1)}{D!(k+1)!} \\ &= \frac{(D+k+1)!}{D!(k+1)!}, \end{aligned}$$

proving the inducting step.

3. Now we show that $N(D, M)$ grows in polynomial fashion like D^M. Assume $D \ll M$. First, we write

$$\begin{aligned} N(D, M) &= \frac{(D+M)!}{D!M!} \\ &\simeq \frac{(D+M)^{D+M} e^{-(D+M)}}{D!M^M e^{-M}} && \text{(by Stirling's approximation)} \\ &= \frac{1}{D!M^M} \left(1 + \frac{D}{M}\right)^{D+M} M^{D+M} \frac{e^{-(D+M)}}{e^{-M}} \\ &= \frac{e^{-D}}{D!} \left(1 + \frac{D}{M}\right)^{D+M} M^D. && (1) \end{aligned}$$

Now we take a more delicate look at the term $(1 + \frac{D}{M})^{D+M}$. Note that

$$\begin{aligned} \left(1 + \frac{D}{M}\right)^{D+M} &= \left(1 + \frac{D}{M}\right)^{M} \left(1 + \frac{D}{M}\right)^{D} \\ &= \left(\left(1 + \frac{1}{M/D}\right)^{M/D}\right)^{D} \left(1 + \frac{D}{M}\right)^{D} \\ &\leq e^D 2^D, \end{aligned}$$

where the inequality comes from the fact that $(1+1/x)^x$ is an increasing function and $D < M \Rightarrow D/M \leq 1$. Substitution back into Eq (1), we get

$$N(D, M) \leq \frac{e^{-D}}{D!} e^D 2^D M^D = \frac{2^D}{D!} M^D.$$

The case for $M \ll D$ follows by symmetry.

Problem 1.17 - Gamma density warmup

1. Note

$$\begin{aligned} \Gamma(x+1) &= \int_0^\infty u^x e^{-u} du \\ &= \left[-u^x e^{-u}\right]_{u=0}^{\infty} + \int_0^\infty x u^{x-1} e^{-u} du \end{aligned}$$

$$= x\Gamma(x).$$

2. We note that

$$\Gamma(1) = \int_0^\infty e^{-u} du = \left[e^{-u}\right]_0^\infty = 1.$$

And as a result, by recursion

$$\Gamma(x+1) = x\Gamma(x) = \cdots = x! \text{ for } x \in \mathbb{N}.$$

Problem 1.18 - Volume of unit sphere in n-space

To state the problem statement in a clearer manner, we solve this problem in several steps. In this problem, we let $d\mu$ denote the Lebesgue measure.

1. First we derive Eq (1.142) in the book. We first rewrite the LHS in the following way. Let $x \in \mathbb{R}^D$ be arbitrary, then

$$\begin{aligned}\int_{\mathbb{R}^d} e^{-\|x\|^2} dx &= \int_{\mathbb{R}} \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} e^{-(x_1^2+x_2^2+\cdots+x_n^2)} dx_1 dx_2 \cdots dx_n \\ &= \prod_{i=1}^{D} \int_{\mathbb{R}} e^{-x_i^2} dx_i.\end{aligned}$$

Next, we evaluate this integral. In order to make the computation easier, we choose to let the integrand be $e^{-\pi|x|^2}$ instead (it doesn't effect the final result, and one could always get the original integral by scaling). Note that using the same argument as above, we have

$$\int_{\mathbb{R}^D} e^{-\pi\|x\|^2} dx = \left(\int_{\mathbb{R}} e^{-\pi x^2} dx\right)^D.$$

Next, we have

$$\begin{aligned}\left(\int_{\mathbb{R}} e^{-\pi x^2} dx\right)^2 &= \left(\int_{\mathbb{R}} e^{-\pi x_1^2} dx_1\right)\left(\int_{\mathbb{R}} e^{-\pi x_2^2} dx_2\right) \\ &= \int_{\mathbb{R}\times\mathbb{R}} e^{-\pi(x_1^2+x_2^2)} d(x_1 \times x_2) && \text{(by Fubini's theorem)} \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}} e^{-\pi(x_1^2+x_2^2)} dx_1 dx_2 && \text{(by Fubini's theorem)} \\ &= \int_{[0,2\pi]} \int_{\mathbb{R}} e^{-\pi r^2} r dr d\theta && \text{(switch to polar coordinates)} \\ &= \int_{[0,2\pi]} d\theta \int_{\mathbb{R}} e^{-\pi r^2} r dr \\ &= 2\pi \left[-\frac{1}{2\pi} e^{-\pi r^2}\right]_0^\infty \\ &= 1.\end{aligned}$$

Since $\int_{\mathbb{R}} e^{\pi x^2} dx > 0$, it follows that $\int_{\mathbb{R}^D} e^{-\pi\|x\|^2} dx = 1$.

2. Consider the function $f : \mathbb{R}^D \to \mathbb{R}; x \mapsto e^{-\pi\|x\|^2}$. We just showed in part-1 that $f \in L^1(\mathbb{R}^D)$. Therefore, using generalized spherical coordinate (e.g. Theorem 6.3.4 in [Ste05]), we have that

$$\begin{aligned}
1 = \int_{\mathbb{R}^D} f(x)dx &= \int_{S^{D-1}} \left(\int_{\mathbb{R}^+} f(r\gamma) r^{D-1} dr \right) d\sigma(\gamma) \\
&= \int_{S^{D-1}} \left(\int_{\mathbb{R}^+} e^{-\pi\|r\gamma\|^2} r^{D-1} dr \right) d\sigma(\gamma) \\
&= \int_{S^{D-1}} \left(\int_{\mathbb{R}^+} e^{-\pi r^2} r^{D-1} dr \right) d\sigma(\gamma) \\
&= \int_{S^{D-1}} d\sigma(r) \int_{\mathbb{R}^+} e^{-\pi r^2} r^{D-1} dr \\
&= \sigma(S^{D-1}) \int_{\mathbb{R}^+} e^{-\pi r^2} r^{D-1} dr.
\end{aligned}$$

Now we evaluate the integral on the RHS:

$$\begin{aligned}
\int_{\mathbb{R}^+} e^{-\pi r^r} r^{D-1} dr &= \int_0^\infty e^{-u} \left(\frac{u}{\pi}\right)^{\frac{D-1}{2}} \frac{1}{2\pi(u/\pi)^{1/2}} du \\
&= \frac{1}{2\pi} \int_0^\infty e^{-u} \left(\frac{u}{\pi}\right)^{\frac{D}{2}-1} du \\
&= \frac{1}{2\pi} \pi^{1-\frac{D}{2}} \int_0^\infty e^{-u} u^{\frac{D}{2}-1} du \\
&= \frac{1}{2} \pi^{-\frac{D}{2}} \Gamma\left(\frac{D}{2}\right).
\end{aligned}$$

Therefore, substituting back we get

$$\sigma(S^{D-1}) = \frac{1}{\int_{\mathbb{R}^+} e^{-\pi r^2} r^{D-1} dr} = \frac{2\pi^{D/2}}{\Gamma(D/2)}.$$

This $\sigma(S^{D-1})$ is the S_D in the problem.

3. Now we calculate the volume of the ball. Let B_1 denote the unit ball in $\mathbb{R}^D$. Note that again by generalized spherical coordinate,

$$\begin{aligned}
V_D &= \int_{\mathbb{R}^D} \mathbb{1}_{B_1}(x) d\mu \\
&= \int_{S^{D-1}} \int_{\mathbb{R}^+} \mathbb{1}_{B_1}(r\gamma) r^{D-1} d\sigma(\gamma) \\
&= \int_{S^{D-1}} \left(\int_{[0,1]} r^{D-1} dr \right) d\sigma(\gamma) \\
&= \left(\int_{S^{D-1}} d\sigma(\gamma) \right) \left(\int_{[0,1]} r^{D-1} dr \right) \\
&= \sigma(S^{D-1}) \left[\frac{1}{D} r^D \right]_0^1 \\
&= \frac{\pi^{D/2}}{\Gamma(D/2)(D/2)}
\end{aligned}$$

$$= \frac{\pi^{D/2}}{\Gamma(D/2+1)}.$$

as desired.

4. When $D = 2$, we get

$$S_D = \frac{2\pi^{2/2}}{\Gamma(1)} = 2\pi \text{ and } V_D = \frac{S_D}{D} = \pi.$$

When $D = 2$, we get

$$S_D = \frac{2\pi^{3/2}}{\Gamma(3/2)} = \frac{2\pi^{3/2}}{\pi^{1/2}/2} = 4\pi \text{ and } V_D = \frac{4}{3}\pi.$$

Remark 1.1. This problem could have been solved heuristically. But it loses rigor. What was showed was a rigorous mathematical way to treat this problem.

Problem 1.19 - High dimensional cubes concentrate on corners

1. Using the result of the previous problem, and the fact that $m_d(rB) = r^d m(B)$, where m_d is the Lebesgue measure in d-dimensional Euclidean space (e.g. Exercise 1.6 in [Ste05]), we have that

$$\begin{aligned}
\frac{V_{\text{sphere}}}{V_{\text{cube}}} &= \frac{\pi^{D/2}a^D}{\Gamma(D/2+1)2^D a^D} = \frac{\pi^{D/2}}{\Gamma(D/2+1)2^D} \\
&\simeq \frac{\pi^{D/2}}{(2\pi)^{1/2}e^{-D/2}(D/2)^{D/2+1/2}2^D} && \text{(by Stirling formula)} \\
&= C\frac{\pi^{D/2}e^{D/2}}{(D/2)^{D/2}}\frac{1}{D^{1/2}}2^{-D} && (C \text{ is some constant}) \\
&= C\left(\frac{2\pi e}{D}\right)^{D/2}\frac{1}{D^{1/2}2^D} \xrightarrow{D\to\infty} 0.
\end{aligned}$$

2. On the other hand, we have

$$\begin{aligned}
\text{dist(center to corner)} &= \sqrt{Da^2} = a\sqrt{D} \\
\text{dist(center to top)} &= a.
\end{aligned}$$

And thus the ratio is $\sqrt{D}$.

Problem 1.20 - High dimensional gaussian concentrate on a thin strip

First, note that the density given in the problem is that of a Gaussian in D dimensional Euclidean space with $\Sigma = \text{diag}(\sigma^2)$.

1. To show that the density is of the form exhibited in (1.148), we note that again by generalized spherical coordinate we have

$$\int_{\mathbb{R}^D} p(x)dx = \int_{S^{D-1}}\int_{\mathbb{R}^+} p(\gamma r)drd\sigma(\gamma)$$

$$
\begin{aligned}
&= \int_{S^{D-1}} \int_{\mathbb{R}^+} \frac{1}{(2\pi\sigma^2)^{D/2}} \exp\left\{-\frac{\|\gamma r\|^2}{2\sigma^2}\right\} r^{D-1} dr d\sigma(\gamma) \\
&= \int_{S^{D-1}} \int_{\mathbb{R}^+} \frac{1}{(2\pi\sigma^2)^{D/2}} \exp\left\{-\frac{\|\gamma\|^2 r^2}{2\sigma^2}\right\} r^{D-1} dr d\sigma(\gamma) \\
&= \int_{S^{D-1}} d\sigma(\gamma) \int_{\mathbb{R}^+} \frac{1}{(2\pi\sigma^2)^{D/2}} \exp\left\{-\frac{r^2}{2\sigma^2}\right\} r^{D-1} dr \\
&= \sigma(S^{D-1}) \int_{\mathbb{R}^+} \frac{1}{(2\pi\sigma^2)^{D/2}} \exp\left\{-\frac{r^2}{2\sigma^2}\right\} r^{D-1} dr.
\end{aligned}
$$

This is the formula in (1.148) if we relabel $\sigma(S^{D-1}) = S_D$.

2. First, we note

$$
\begin{aligned}
\frac{d}{dr}p(r) &= C \cdot \frac{d}{dr}\left[r^{D-1}\exp\left\{-\frac{r^2}{2\sigma^2}\right\}\right] \\
&= C \cdot \left[(D-1)r^{D-2}\exp\left\{-\frac{r^2}{2\sigma^2}\right\} + r^{D-1}\exp\left\{-\frac{r^2}{2\sigma^2}\right\}\left(-\frac{1}{2\sigma^2}\right)2r\right] \\
&= C\left[(D-1)r^{D-2} - \frac{r^D}{\sigma^2}\right]\exp\left\{-\frac{r^2}{2\sigma^2}\right\}.
\end{aligned}
$$

To find the stationary point, we set it to zero:

$$
\begin{aligned}
\frac{d}{dr}p(r) = 0 &\iff C\left[(D-1)r^{D-2} - \frac{r^D}{2\sigma^2}\right]\exp\left\{-\frac{r^2}{2\sigma^2}\right\} = 0 \\
&\iff (D-1)r^{D-2} - \frac{r^D}{\sigma^2} = 0 \\
&\iff \hat{r} = \sqrt{(D-1)\sigma^2} \simeq \sqrt{D}\sigma,
\end{aligned}
$$

where the approximation follows since $\sqrt{D+1} = \sqrt{D}$ for large D.

3. To show (1.149), first we note

$$
\begin{aligned}
p(\hat{r}+\varepsilon) &= \frac{S_D(\hat{r}+\varepsilon)^{D-1}}{(2\pi\sigma^2)^{D/2}}\exp\left\{-\frac{(\hat{r}+\varepsilon)^2}{2\sigma^2}\right\} \\
&= \frac{S_D}{(2\pi\sigma^2)^{D/2}}\exp\left\{-\frac{(\hat{r}+\varepsilon)^2}{2\sigma^2} + (D-1)\log(\hat{r}+\varepsilon)\right\} \\
&= \frac{S_D}{(2\pi\sigma^2)^{D/2}}\exp\left\{-\frac{(\hat{r}+\varepsilon)^2}{2\sigma^2} + (D-1)\left[\log\left(1+\frac{\varepsilon}{\hat{r}}\right) + \log\hat{r}\right]\right\} \\
&= \frac{S_D\hat{r}^{D-1}}{(2\pi\sigma^2)^{D/2}}\exp\left\{-\frac{\hat{r}^2}{2\sigma^2} - \frac{\hat{r}\varepsilon}{\sigma^2} - \frac{\varepsilon^2}{2\sigma^2} + (D-1)\left(\frac{\varepsilon}{\hat{r}} - \frac{\varepsilon^2}{2\hat{\gamma}^2} + o\left(\frac{\varepsilon^2}{\hat{\gamma}^2}\right)\right)\right\} \\
&= \underbrace{\frac{S_D\hat{r}^{D-1}}{(2\pi\sigma^2)^{D/2}}\exp\left\{-\frac{\hat{r}^2}{2\sigma^2}\right\}}_{=p(r)}\underbrace{\exp\left\{-\frac{\hat{r}\varepsilon}{\sigma^2} - \frac{\varepsilon^2}{2\sigma^2} + (D-1)\left(\frac{\varepsilon}{\hat{r}} - \frac{\varepsilon^2}{2\hat{\gamma}^2} + o\left(\frac{\varepsilon^2}{\hat{\gamma}^2}\right)\right)\right\}}_{:=\mathcal{E}(\varepsilon,\sigma,\hat{r})}. \qquad (1)
\end{aligned}
$$

Now, we just need to massage last term in the RHS of (1): since $\hat{r} = \sqrt{D-1}\sigma$, we get

$$\begin{aligned}\mathcal{E}(\varepsilon, \sigma, \hat{r}) &= \exp\left\{-\frac{\sqrt{D}-1\varepsilon}{\sigma} - \frac{\varepsilon^2}{2\sigma^2} + \frac{\sqrt{D-1}\varepsilon}{\sigma} - \frac{\varepsilon^2}{2\sigma^2} + o\left(\frac{\varepsilon^2}{\hat{\gamma}^2}\right)\right\} \\ &= \exp\left\{-\frac{\varepsilon^2}{\sigma^2}\right\} \exp\left\{o\left(\frac{\varepsilon^2}{\hat{\gamma}^2}\right)\right\}.\end{aligned}$$

Since by assumption $\varepsilon \ll \hat{r}$, it follows that $\mathcal{E}(\varepsilon, \sigma, \hat{r}) \simeq \exp\{-\varepsilon^2/\sigma^2\}$. Substituting back we get

$$p(\hat{r} + \varepsilon) = p(r) \exp\left\{-\frac{\varepsilon^2}{\sigma^2}\right\}$$

as desired.

4. Note that we have

$$p(x = 0) = \frac{1}{(2\pi\sigma^2)^{D/2}},$$

and

$$\begin{aligned}p(x \in \Gamma | \Gamma = \{\gamma \in \mathbb{R}^d | \, \|\gamma\| = \sqrt{D-1}\sigma\}) &= \frac{1}{(2\pi\sigma^2)^{D/2}} \exp\left\{-\frac{(D-1)\sigma^2}{2\sigma^2}\right\} \\ &= \frac{1}{(2\pi\sigma^2)^{D/2}} \exp\left\{-\frac{D-1}{2}\right\},\end{aligned}$$

whence

$$\begin{aligned}\frac{p(x \in \Gamma | \Gamma = \{\gamma \in \mathbb{R}^d | \, \|\gamma\| = \sqrt{D-1}\sigma\})}{p(x=0)} &= \exp\left\{-\frac{D-1}{2}\right\} \\ &\simeq \exp\left\{-\frac{D}{2}\right\} \text{ when } D \text{ is large}\end{aligned}$$

Problem 1.21 - Upper bound of bayesian classification error

1. Since $x \mapsto \sqrt{x}$ is monotonically increasing and $a \leq b$, it follows that $0 \leq a^{1/2} \leq b^{1/2}$, which then implies $a \leq a^{1/2} b^{1/2}$ after multiplying both sides with $a^{1/2}$.

2. To show the desired inequality, we note (for notation, we let $\mathcal{X}$ be the ambient input space),

$$\begin{aligned}\mathbb{P}(\text{mistake}) &= \int_{\mathcal{R}_1} \mathbb{P}(x, \mathcal{C}_2) dx + \int_{\mathcal{R}_2} \mathbb{P}(x, \mathcal{C}_1) dx \\ &\leq \int_{\mathcal{R}_1} \mathbb{P}(x, \mathcal{C}_1)\mathbb{P}(x, \mathcal{C}_2) dx + \int_{\mathcal{R}_2} \mathbb{P}(x, \mathcal{C}_1)\mathbb{P}(x, \mathcal{C}_2) dx \qquad \text{(by part-1)} \\ &= \int_{\mathcal{R}_1 \cup \mathcal{R}_2} \mathbb{P}(x, \mathcal{C}_1)\mathbb{P}(x, \mathcal{C}_2) dx \\ &= \int_{\mathcal{X}} \mathbb{P}(x, \mathcal{C}_1)\mathbb{P}(x, \mathcal{C}_2) dx,\end{aligned}$$

where the last inequality follows since we are working in a two-class setting and the fact that decision regions partition the input space.

Problem 1.22 - Uniform loss maximizes posterior probability

For concise notation, we write the loss matrix as $L = \mathbb{1}\mathbb{1}^T - I$, where here $\mathbb{1}$ stands for vector of 1's and $\vec{\mathbb{P}}(\mathcal{C}|x)$ as a vector of $\mathbb{P}(\mathcal{C}_k, x)$'s. Then we can rewrite Eq. (1.81) in the book as

$$\begin{aligned}\min_j \sum_k L_{kj}\mathbb{P}(\mathcal{C}_k|x) &= \min_j \vec{\mathbb{P}}(\mathcal{C}|x)^T(\mathbb{1}\mathbb{1}^T - I)e_j \\ &= \min_j \vec{\mathbb{P}}(\mathcal{C}|x)^T\mathbb{1} - \mathbb{P}(\mathcal{C}_j|x) \\ &= \min_j 1 - \mathbb{P}(\mathcal{C}_j|x) \\ &= \max_j \mathbb{P}(\mathcal{C}_j|x).\end{aligned}$$

where the second equality follows from the fact the conditional distribution sums to 1.

We can interpret this loss in the following way: this loss assigns unit weight to each misclassified labels and zero weight to correctly classified labels and therefore minimizing the expectation represents minimizing the misclassification rate.

Problem 1.23 - Characterization for minimizing general expected loss

Note

$$\sum_k L_{kj}\mathbb{P}(\mathcal{C}_k|x) = \frac{1}{p(x)}\sum_k L_{kj}\mathbb{P}(x|\mathcal{C}_k)\mathbb{P}(\mathcal{C}_k).$$

Suppose $m = \min(\sum_k L_{kj}\mathbb{P}(x|\mathcal{C}_k)$, if we increase $\mathbb{P}(\mathcal{C}_k)$, we would have to decrease L_{kj} to keep the minimum. Hence, there is a direct trade-off between $\mathbb{P}(\mathcal{C}_k)$ and L_{kj}.

Problem 1.24 - Duality between decision and rejection criterion

1. According to Eq. (1.81) in the book, the decision of labels is found by computing $\arg\min_j \sum_k L_{kj}\mathbb{P}(\mathcal{C}_k|x)$. Since rejection option is also used, let $\hat{j}$ be the minimum, then the decision criterion can be modeled as a function $\varphi : \mathbb{N} \to \mathbb{N} \cup \{\emptyset\}$ by

$$j \mapsto \begin{cases} \arg\min_j \sum_k L_{kj}\mathbb{P}(\mathcal{C}_k|x) & \text{if } \min_j \sum_k L_{kj}\mathbb{P}(\mathcal{C}_k|x) \\ \emptyset & \text{otherwise} \end{cases}.$$

   Note the j defined in φ by default refers to the minimizer of $\sum_k L_{kj}\mathbb{P}(\mathcal{C}_k|x)$, and the mapping to empty set means rejection.

2. When $L = \mathbb{1}\mathbb{1}^T - I$, then we have by previous part that

$$\begin{aligned}\varphi(\hat{j}) = j &\iff \min_j \sum_k L_{kj}\mathbb{P}(\mathcal{C}_k|x) \le \lambda \\ &\iff \min_j 1 - \mathbb{P}(\mathcal{C}_j|x) \le \lambda \qquad \text{(by Problem 1.22)} \\ &\iff \max \mathbb{P}(\mathcal{C}_k|x) \ge 1 - \lambda.\end{aligned}$$

Note that the last stipulation is equivalent to $\theta = 1 - \lambda$ in the reject option definition. Hence, the two criteria coincide when $\theta = 1 - \lambda$.

Problem 1.25 - Generalized squared loss function

We follow the same procedure as in the 1 dimensional case. Note

$$\begin{aligned}\frac{\delta \mathbb{E}[L]}{\delta L} &= \frac{\delta}{\delta L}\left[\int\int \|y(x) - t\|^2 \, p(t,x) dx dt\right] \\ &= \int 2(y(x) - t) p(t,x) dt.\end{aligned}$$

Setting it to zero yields:

$$\begin{aligned}y(x)\int p(t,x)dt = \int tp(t,x)dt \iff y(x) &= \frac{\int tp(t,x)dt}{\int p(t,x)dt} \\ \iff y(x) &= \frac{\int tp(t,x)dt}{p(x)} = \int tp(t|x)dt\end{aligned}$$

as desired.

Problem 1.26 - Decomposition of expected squared loss

We use the similar argument as in deriving Eq. (1.90) in here. We write

$$\begin{aligned}\|y(x) - t\|^2 &= \|y(x) - \mathbb{E}[t|x] + \mathbb{E}[t|x] - t\|^2 \\ &= \|y(x) - \mathbb{E}[t|x]\|^2 - 2(y(x) - \mathbb{E}[y|x])^T(\mathbb{E}[t|x] - t) + \|\mathbb{E}[t|x] - t\|^2 .\end{aligned}$$

Also note that we can rewrite $\mathbb{E}[t|x] - t = \mathbb{E}[t|x] - \mathbb{E}[\mathbb{E}[t|x]]$ and that $\mathbb{E}[y(x) - \mathbb{E}[y|x]] = \mathbb{E}[y] - \mathbb{E}[y] = 0$. Hence

$$\begin{aligned}\mathbb{E}\left[\|y(x) - t\|^2\right] &= \int \|y(x) - \mathbb{E}[t|x]\|^2 p(x)dx + \int \|\mathbb{E}[t|x - t\|^2 p(x)dx \\ &= \int \|y(x) - \mathbb{E}[t|x]\|^2 p(x)dx + \int \mathrm{Var}[t|x]p(x)dx.\end{aligned}$$

Hence, we see that $\mathbb{E}[\|y(x) - t\|^2]$ is minimized when $y(x) = \mathbb{E}[t|x]$, which is analogues to Eq. (1.90).

Problem 1.27 - Maximizer of L_1, L_{0^+} expected loss

According to Eq. (1.91), an application of Fubini's theorem we can rewrite the expected Minkowski loss in the following form:

$$\mathbb{E}[L] = \int \underbrace{\int |y(x) - t|^q \, p(x,t)dt}_{:=G(x,y(x))} \, dx$$

Here we need assume $G(x, y(x))$ converges uniformly so that we can differentiate under the improper (possibly) integral. As usual, we compute the first variation:

$$\begin{aligned}\frac{\delta \mathbb{E}[L]}{\delta y(x)} &= \frac{\partial G(x, y(x))}{\partial y(x)} = \int q\,|y(x)-t|^{q-1}\,\frac{(y(x)-t)}{|y(x)-t|}p(x,t)dt \\ &= p(x)\int q\,|y(x)-t|^{q-1}\,\mathrm{sgn}(y(x)-t)p(t|x)dt \\ &= p(x)\left(\int_{\{t\le y(x)\}} q\,|y(x)|^{q-1}\,p(t|x)dt - \int_{t>y(x)} q\,|y(x)-t|^{q-1}\,p(t|x)dt\right) \end{aligned} \tag{1}$$

To find the stationary point when $q = 1$, we set Eq.(1) to zero:

$$\begin{aligned} & p(x)\left(\int_{\{t\le y(x)\}} q\,|y(x)|^{q-1}\,p(t|x)dt - \int_{\{t>y(x)\}} q\,|y(x)-t|^{q-1}\,p(t|x)dt\right) = 0 \\ \implies {}_1 & \int_{\{t\le y(x)\}} p(t|x)dt = \int_{\{t>y(x)\}} p(t|x), \end{aligned} \tag{2}$$

where $\implies_1$ follows since we only need to care about $x \in \mathrm{supp}(p(x))$. Hence, the $y(x)$ that maximizes the expected loss function with $q = 1$ satisfies Eq.(2), which is the definition of the median.

Now we consider the case when $p \to 0$. Instead of taking the functional derivative, we will use a more delicate and analytical approach. First, we write

$$\mathbb{E}[L] = \lim_{q\to 0}\int |y(x)-t|^q\,d(F_x \times F_t).$$

Observe that $\lim_{q\to 0}|y(x)-t|^q = \mathbb{1}_{\{y(x)\neq t\}}$, which is in $L_2(\Omega)$ (we use Ω to denote the probability space). An application of DCT yields

$$\begin{aligned}\mathbb{E}[L] &= \int \mathbb{1}_{\{y(x)\neq t\}} d(F_x \times F_t) \\ &= \int d(F_x \times F_t) - \int \mathbb{1}_{\{y(x)=t\}} d(F_x \times F_t) \\ &= 1 - \underbrace{\int\int \mathbb{1}_{\{y(x)=t\}} p(x,t)dtdx}_{:=\mathcal{I}_1(y(x),x,t)}. \qquad \text{(by change of variable theorem)}\end{aligned}$$

In order to minimize $\mathbb{E}[L]$, it suffices to find $\arg\max_{y(x)} \mathcal{I}_1(y(x), x, t)$. First, we rewrite

$$\mathcal{I}_1(y(x),x,t) = \int p(x)\underbrace{\int \mathbb{1}_{\{t=y(x)\}} p(t|x)dt}_{:=\mathcal{I}_2(y(x),x,t)}dx.$$

Since $p(x) \ge 0$ for any $x \in \mathbb{R}^n$ and $\mathcal{I}_2(y(x), x, t) \ge 0$, it follows that $\arg\max_{y(x)} \mathcal{I}_1(y(x), x, t) = \arg\max_{y(x)} \mathcal{I}_2(y(x), x, t)$. Note that $\mathcal{I}_2(y(x), x, t) = 0$ since it is an integral w.r.t to a singleton point whose Lebesgue measure is zero. However, we can circumvent this in the following manner: note that

$$\mathcal{I}_2(y(x),x,t) = \int \lim_{n\to\infty} \mathbb{1}_{\{t\in(y(x)-\frac{1}{2n},y(x)+\frac{1}{2n})\}} p(t|x)dt$$

$$
\begin{aligned}
&= \lim_{n\to\infty} \int \mathbb{1}_{\{t\in(y(x)-\frac{1}{2n},y(x)+\frac{1}{2n})\}} p(t|x)dt \\
&\leq \lim_{n\to\infty} \frac{1}{n} \sup_{t\in(y(x)-\frac{1}{2n},y(x)+\frac{1}{2n})} p(t|x) \qquad (3)
\end{aligned}
$$

If we define $F_n(y(x)) = \frac{1}{n}\sup_{t\in(y(x)-1/(2n),y(x)+1/(2n))} p(t|x)$, then it follows from Eq.(3) that $F_n(y(x)) \to 0$ as $n \to \infty$ for any $y(x) \in \mathbb{R}$. However, we would like to find a $\tilde{y}(x)$ s.t. $F_n(\tilde{y}(x)) \leq F_n(y(x))$ for any other choice of $y(x)$. We claim that $\tilde{y}(x) = \arg\max_{t\in\mathbb{R}} p(t|x)$. Indeed, if so, we have

$$
\begin{aligned}
F_n(\tilde{y}(x)) &= \frac{1}{n} \sup_{t\in(\arg\max_t p(t|x)-\frac{1}{2n},\arg\max_t p(t|x)+\frac{1}{2n})} p(t|x) = \frac{1}{n}\sup_{t\in\mathbb{R}^n} p(t|x) \\
&\geq \frac{1}{n} \sup_{t\in(y(x)-1/(2n),y(x)+1/(2n))} p(t|x) = F(y(x)).
\end{aligned}
$$

So to translate into heuristic terms, $y(x) = \arg\max_t p(t|x)$ minimizes the loss function in "each step" of the process of "approaching the limit of $q \to 0$".

Problem 1.28 - Derivation of information content

Assuming the randoms variables to be discrete does simplifies the argument but it also losses rigor. To achieve maximum amount of rigor possible, we use a measure theoretic language. For this reason, we will use a slightly different formulation, but the idea remains the same.

Instead of using x, y to denote random variable, we use X, Y. Note that for any $A \in \mathcal{B}(X)$, $B \in \mathcal{B}(Y)$, where $\mathcal{B}$ denote the Borel sets, if X and Y are independent,

$$
\begin{aligned}
h(X \in A, Y \in B) &= h\bigg(\int_{A\times B} d(F_X \times F_Y)\bigg) = h\bigg(\int_A dF_X \cdot \int_B dF_Y\bigg). \qquad \text{(by independence)} \\
&= h(X \in A) + h(Y \in B) = h\bigg(\int_A dF_X\bigg) + h\bigg(\int_B dF_Y\bigg).
\end{aligned}
$$

If we let $x = \int_A dF_X$ and $y = \int_B dF_Y$, then this problem reduces to the following form: find a representation of h such that $h(xy) = h(x) + h(y)$ for any $x, y \in [0, 1]$. This is variant of the Cauchy Functional Equation problem.

Recall that the Cauchy functional equation in its standard form is as follows: find a function f that satisfies $f(x + y) = f(x) + f(y)$. The obvious solution to f is the linear one: $x \mapsto cx$ for $x \in \mathbb{R}^n$. However, without additional assumptions, one can obtain other complicated solutions as well. But generally these solutions serves as pedagogical example. A classical result is that if we assume f to be either continuous or monotone, then $f(x) = cx$ for arbitrary $c \in \mathbb{R}$ is the only solution. (cf. [Kuc09]).

With this in mind, to solve for h, we define $g(x) = h(e^x)$. Then we see that

$$
g(x + y) = h(e^x e^y) = h(e^x) + h(e^y) = g(x) + g(y).
$$

Then if we require g to be continuous, then $g(x)$ is uniquely represented as cx for any $c \in \mathbb{R}$. Then note that for any $x \in \mathbb{R}^+$,

$$
h(x) = h(e^{\ln x}) = g(\ln x) = c \ln x, \text{ for any } c \in \mathbb{R}.
$$

Therefore, we have $h(x) \propto \ln(x)$ as desired.

Problem 1.29 - Upper bound for entropy of discrete variables

We directly apply Jensen's inequality:

$$H(x) = -\sum_{i=1}^{M} p(x_i)\ln(x_i) = \sum_{i=1}^{M} p(x_i)\ln\frac{1}{p(x_i)} \leq \ln\left(\sum_{i=1}^{M} p(x_i)\frac{1}{p(x_i)}\right) = \ln M,$$

where the $\leq$ follows since $\ln(x)$ is concave.

Problem 1.30 - KL-divergence for Gaussian

We use the original definition of KL-divergence:

$$KL(p||q) = \underbrace{-\int p(x)\ln q(x)dx}_{(1)} - \underbrace{\left(-\int p(x)\ln p(x)dx\right)}_{(2)}.$$

We compute it term by term, first note that

$$\begin{aligned}
(1) =& -\int \varphi(x|\mu,\sigma^2)\ln\left[\varphi(x|m,s^2)\right]dx \\
=& -\int \varphi(x|\mu,\sigma^2)\left\{\ln\frac{1}{(2\pi s^2)^{1/2}} - \frac{(x-m)^2}{2s^2}\right\} \\
=& \int \varphi(x|\mu,\sigma^2)\left[\frac{1}{2}\ln(2\pi s^2) + \frac{(x-m)^2}{2s^2}\right]dx \\
=& \frac{1}{2}\int \varphi(x|\mu,\sigma^2)\ln(2\pi s^2)dx + \frac{1}{2s^2}\left[\int \varphi(x|\mu,\sigma^2)x^2dx - 2m\int \varphi(x|\mu,\sigma^2)xdx + \int \varphi(x|\mu,\sigma^2)m^2dx\right] \\
=& \frac{1}{2}\ln(2\pi s^2) + \frac{1}{2s^2}\left[\sigma^2+\mu^2-2m\mu+m^2\right].
\end{aligned}$$

And similarly

$$\begin{aligned}
(2) =& -\int \varphi(x|\mu,\sigma^2)\ln\left[\varphi(x|\mu,\sigma^2)\right]dx \\
=& \frac{1}{2}\int \varphi(x|\mu,\sigma^2)\ln(2\pi\sigma^2)dx + \frac{1}{2\sigma^2}\left[\int \varphi(x|\mu,\sigma^2)x^2dx - 2\mu\int \varphi(x|\mu,\sigma^2)xdx + \mu^2\int \varphi(x|\mu,\sigma^2)dx\right] \\
=& \frac{1}{2}\ln(2\pi\sigma^2) + \frac{1}{2\sigma^2}\left[\sigma^2+\mu^2-2\mu^2+\mu^2\right] \\
=& \frac{1}{2}\ln(2\pi\sigma^2) + \frac{1}{2}.
\end{aligned}$$

Hence, it follows that

$$KL(p||q) = \frac{1}{2}\ln(2\pi s^2) + \frac{1}{2s^2}\left[\sigma^2+\mu^2-2m\mu+m^2\right] - \frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2}$$

$$= \frac{1}{2s^2}\left[(m-\mu)^2 + (\sigma^2 - s^2) + s^2 \log \frac{s^2}{\sigma^2}\right]$$

Problem 1.31 - Differential entropy and independence

In this problem, we extend the definition to of KL-divergence to a more general setting as follows:

Definition 1.1. If P and Q are probability measures over a set Ω, if P is absolutely continuous w.r.t. Q, then the KL divergence is defined as

$$KL(P||Q) = \int_\Omega \log \frac{dP}{dQ} dP,$$

where dP/dQ is the Radon-Nikodym derivative, whose existence is guaranteed by the fact that P is absolutely continuous w.r.t Q.

Lemma 1.1. *$KL(P||Q) \geq 0$ for any pair of probability measures P and Q such that $P \ll Q$, the equality if P and Q are equal.*

Proof. This is a directly application of Jensen's inequality. Note that

$$KL(P||Q) = -\int_\Omega \log \frac{dQ}{dP} dP \geq -\log\left(\int_\Omega \frac{dQ}{dP} dP\right) = -\log \int_\Omega dQ = 0.$$

Recall that Jensen's inequality attains the equality if and only if when the function is affine or its argument is constant. In this case, $\log(t)$ is not constant, and thus $KL(P||Q) = 0$ iff $dP/dQ = C$ for some constant $C \in \mathbb{R}$. We claim that $C = 1$, since otherwise we would have

$$\int_\Omega dP = \int_\Omega \frac{dP}{dQ} dQ = C \int_\Omega dQ = C \neq 1,$$

which is a contradiction since P is a probability measure. Then we claim that P and Q are equal. For any set A in the (predefined) sigma algebra, we have

$$P(A) = \int_A dP = \int_A \frac{dP}{dQ} dQ = \int_A 1 dQ = Q(A).$$

Hence, $P = Q$. □

Now we come back to the problem. We instead use X and Y two denote the random variables and $f_X, f_Y, f_{X,Y}$, to denote their (marginal) densities. Suppose X and Y are independent. Then it follows that

$$\begin{aligned}
H(X,Y) &= \int\int f_{X,Y}(x,y) \log f_{X,Y}(x,y) dxdy \\
&= \int\int f_X(x) f_Y(y) (\log f_X(x) + \log f_Y(y)) dxdy \\
&= \int\int f_X(x) f_Y(y) \log f_X(x) dxdy + \int\int f_X(x) f_Y(y) \log f_Y(y) dxdy \\
&= \int f_X(x) \log f_X(x) dx + \int f_Y(y) \log f_Y(y) dy \\
&= H(X) + H(Y).
\end{aligned}$$

Now on the other hand, suppose $H(X,Y) = H(X)+H(Y)$. Then since $H(X,Y) = H(Y|X)+H(X)$ according to Eq. (1.112), it follows that $H(Y|X) = H(Y)$. Note that

$$\begin{aligned} H(Y|X) - H(Y) &= -\int\int f(x,y)\log f(y|x)dxdy + \int\int f(x,y)\log f(y)dxdy \\ &= \int\int f(x,y)\log\frac{f(y)}{f(y|x)}dxdy \\ &= KL(f_Y(y)||f_{Y|X}(y|x)). \end{aligned}$$

Then according to Lem. 1.1, $f_Y(y) = f_{Y|X}(y|x)$ almost surely, and as a result $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ which implies that X and Y are independent.

Problem 1.32 - Entropy under linear transformation

Since this problem uses transformation theorem, we first recall this classical result:

Theorem 1.1 ([Bil12, Thm. 17.2])**.** *Let T be a continuously differentiable map of the open set U onto V. Suppose that T is injective an that $J(x) \neq 0$ for all x. If f is non-negative, then*

$$\int_U f(Tx)\,|J(x)|\,dx = \int_{V=TU} f(y)dy.$$

Remark 1.2. We can use this theorem to the get change of variable formula for random variables in $\mathbb{R}^d$ in the following way. Suppose X is a random variable in $\mathbb{R}^d$ with density f_X and $g(\cdot)$ is a C^1 diffeomorphism in $\mathbb{R}^d$, whose inverse is denoted as T and $J_T(x) \neq 0$, then it follows that

$$\mathbb{P}[g(X) \in A] = \mathbb{P}[X \in g^{-1}(A)] = \mathbb{P}[X \in TA] = \int_{TA} f_X(y)dy.$$

Now apply Thm. 1.1, and we get

$$\int_{TA} f_X(y)dy = \int_A f_X(Tx)|J_T(x)|dx = \int_A f_X(g^{-1}(x))|J_{g^{-1}}(x)|dx.$$

Hence, from

$$\begin{aligned} \mathbb{P}(g(X) \in A) &= \int \mathbb{1}_A dF_{g(X)} = \int \mathbb{1}_A \frac{dF_{g(X)}}{dx}dx \qquad (dx \text{ refers to Lebesgue measure}) \\ &= \int_A f_X(g^{-1}(x))\left|J_{g^{-1}}(x)\right|dx \end{aligned}$$

and the fact that Radon-Nikodym derivative is unique it follows that $g(X)$ has density of the form $f_X(g^{-1}(x))\left|J_{g^{-1}}(x)\right|$.

Now we return to the problem. We instead use $f_Y(y)$ and $f_X(x)$ to denote the density function for X and Y. First, by previous remark, we see that $f_Y(y) = f_X(A^{-1}y)\,|J_{A^{-1}}| = f_X(A^{-1}y)\left|\det(A)^{-1}\right|$. So,

$$\begin{aligned} H(Y) &= -\int \ln f_Y(y)dF_Y = \int \ln f_X(A^{-1}y)\left|\det(A)^{-1}\right|dF_Y \\ &= -\int \ln\left[f_X(A^{-1}y)\left|\det(A)^{-1}\right|\right] f_X(A^{-1}y)\left|\det(A)^{-1}\right|dy \end{aligned}$$

$$
\begin{aligned}
&= -\int \ln\left[f_X(A^{-1}Ax)\left|\det(A)^{-1}\right|\right] f_X(A^{-1}Ax)\left|\det(A)^{-1}\right| |\det(A)|\, dx \qquad (1)\\
&= -\int \ln\left[f_X(x)\left|\det(A)^{-1}\right|\right] f_X(x)dx\\
&= -\int f_X(x)\ln f_X(x)dx + \int (\ln\det A) f_X(x)dx\\
&= H(X) + \ln(\det A)
\end{aligned}
$$

as desired. Note that the justification for Eq. (1) is as follows: we abbreviate

$$\varphi(x) = f_X(x)\left|\det(A)^{-1}\right| f_X(x)\left|\det(A)^{-1}\right|,$$

then again by an application of Thm. 1.1

$$(1) = -\int \varphi \circ L d\mu = -\det(L^{-1})\int \varphi \circ L \circ L^{-1} d\mu = -\det(L^{-1})\int \varphi d\mu.$$

where μ is the Lebesgue measure. Here since L is represented by A^{-1}, L^{-1} is thus represented by A.

Problem 1.33 - Zero conditional entropy implies singleton concentration

Instead of x, y, we use X, Y to denote random variables. First, we reformulate $H(Y|X)$ as follows:

$$
\begin{aligned}
H(Y|X) &= -\sum_i \sum_j \mathbb{P}(X = x_i, Y = y_j)\log \mathbb{P}(Y_j = y_j | X = x_i)\\
&= -\sum_i \sum_j \mathbb{P}(Y = y_j | X = x_i)\mathbb{P}(X = x_i)\log \mathbb{P}(Y_j = y_j | X = x_i)\\
&= \sum_i \mathbb{P}(X = x_i)\sum_j f(x_{ij}),
\end{aligned}
$$

where $f : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}$ is defined as $x \mapsto -x\log x$. We now observe that f is strictly positive for $x \in (0, 1)$ and zero for $x =$ 1or 0. The latter is straightforward by direct substitution. To see the former, note

$$f(x) = x\log\frac{1}{x} > x\log 1 = 0 \text{ for } x \in (0, 1).$$

Without loss of generality, we assume $\mathbb{P}(X = x_i) > 0$ since otherwise we get remove these zeros terms without affect the sum. Note that

$$
\begin{aligned}
H(Y|X) = 0 &\implies \sum_i x\mathbb{P}(X = x_i)\sum_j f(x_{ij}) = 0\\
&\implies \sum_j f(x_{ij}) = 0 \quad \text{for any given } i. \qquad (\text{since } \mathbb{P}(X = x_i) > 0 \text{ for any } i)
\end{aligned}
$$

Since $f(x) = 0$ iff $x_{ij} = 0$ or 1, it follows that for any given i, $\mathbb{P}(Y = y_j|X = x_i) = 0$ or 1 for any j. Clearly, there must be only j such that $\mathbb{P}(Y = y_j|X = x_i) = 1$ and $\mathbb{P}(Y = y_j|X = x_i) = 0$ for all other j's since otherwise $\sum_j \mathbb{P}(Y = y_j|X = x_i) \neq 0$, causing a contradiction.

Problem 1.34 - Gaussian distribution maximizes entropy under constraints

To facilitate the notation, we define

$$F(p(x) = -\int_{\mathbb{R}} p(x)\ln p(x)dx + \lambda_1\left(\int_{\mathbb{R}} p(x)dx - 1\right) + \lambda_2\left(\int_{\mathbb{R}} xp(x)dx - \mu\right) + \lambda_3\left(\int_{\mathbb{R}} (x-\mu)^2 p(x)dx - \sigma^2\right).$$

First, we rearrange to get

$$F(p(x)) = \int_{\mathbb{R}} \underbrace{-p(x)\ln p(x)dx + \lambda_1 p(x) + \lambda_2 xp(x) + \lambda_3(x-\mu)^2 p(x)}_{:=G(p(x),x)} dx - (\lambda_1 + \lambda_2\mu + \lambda_3\sigma^2).$$

To get the stationary point, we take the functional derivative:

$$\frac{\delta F(p(x))}{\delta p(x)} = \frac{\partial G(p(x),x)}{\partial p(x)} = -\ln p(x) - 1 + \lambda_1 + \lambda_2 x + \lambda_3(x-\mu)^2.$$

Setting it to zero yields,

$$\ln(p(x)) = \lambda_1 + \lambda_2 x + \lambda_3(x-\mu)^2 - 1 \implies p(x) = \exp\{\lambda_1 + \lambda_2 x + \lambda_3(x-\mu)^2 - 1\}.$$

Now we need to eliminate the λ's by substituting back to the constraints

1. $\int p(x)dx = 1$
2. $\int xp(x)dx = \mu$
3. $\int (x-\mu)^2 p(x)dx = \sigma^2$.

This is system of integral equations. To solve it using first principles would require a lot more work (plus I don't know if Gaussian density is the unique solution). But since we are only required to show that Gaussian density is indeed one solution, we are relieved from the burden of proving uniqueness. And we can just directly compare the coefficients. Note that

$$\int_{\mathbb{R}} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx = 1.$$

Hence, if we let

$$\exp\left\{\lambda_2 x + \lambda_3(x-\mu)^2\right\} = \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \implies \lambda_3 = -\frac{1}{2\sigma^2}, \lambda_2 = 0 \text{ is a solution}$$

and

$$\exp\{\lambda_1 - 1\} = \frac{1}{(2\pi\sigma^2)^{1/2}} \implies \lambda_1 = 1 - \frac{1}{2}\ln 2\pi\sigma^2 \text{ is a solution.}$$

Hence, we have shown that we can find admissible $\lambda_1, \lambda_2, \lambda_3$ such that $p(x)$ satisfies the constraint, and the resulting distribution with this set of λ's is Gaussian. Therefore, Gaussian distribution is a minimizer.

Remark 1.3. One can potentially ask is Gaussian a unique minimizer for this optimization problem? I don't know on the top of my head. This is equivalent to showing that the solution to the integral constraints with $p(x) = \exp\{\lambda_1 + \lambda_2 x + \lambda_3(x-\mu)^2 - 1\}$, has unique solution. I would guess some deep theorems are needed to prove this result, assuming it is true.

Problem 1.35 - Entropy of Gaussian

We let $\varphi(x|\mu,\sigma^2)$ denote the density of Gaussian distribution. Let X be a Gaussian random variable, then

$$\begin{aligned}
H(X) &= -\int \varphi(x|\mu,\sigma^2)\ln\left[\varphi(x|\mu,\sigma^2)\right]dx \\
&= \frac{1}{2}\int \varphi(x|\mu,\sigma^2)\ln(2\pi\sigma^2)dx + \frac{1}{2\sigma^2}\left[\int \varphi(x|\mu,\sigma^2)x^2dx - 2\mu\int \varphi(x|\mu,\sigma^2)xdx + \mu^2\int \varphi(x|\mu,\sigma^2)dx\right] \\
&= \frac{1}{2}\ln(2\pi\sigma^2) + \frac{1}{2\sigma^2}\left[\sigma^2+\mu^2-2\mu^2+\mu^2\right] \\
&= \frac{1}{2}\ln(2\pi\sigma^2) + \frac{1}{2} \\
&= \frac{1}{2}(1+\ln(2\pi\sigma^2))
\end{aligned}$$

as desired.

Problem 1.36 - Second order characterization of convexity

We prove a slightly more generalized version. First, we recall the definition of the convexity.

Definition 1.2. A function $f:\mathbb{R}^n\to\mathbb{R}$ is convex if its domain $\mathcal{D}_f$ is a convex set and for any $x,y\in\mathcal{D}_f$ and $\lambda\in[0,1]$,

$$f(\lambda x+(1-\lambda)y)\le\lambda f(x)+(1-\lambda)f(y). \tag{1}$$

The result of this problem is an direct consequence of the following proposition.

Proposition 1.1. *Let $f:\mathbb{R}^n\to\mathbb{R}$ be a function. Then the following statements are equivalent.*

1. *f is convex.*
2. *$f(y)\ge f(x)+\nabla f(x)^T(y-x)$ assuming f is differentiable.*
3. *The Hessian matrix $H_f(x)$ is positive semidefinite, assuming f is twice differentiable and $\mathcal{D}_f$ is open.*

Proof. (1) $\Rightarrow$ (2). Suppose f is convex. Then by definition for any $y,x\in\mathcal{D}_f$,

$$\begin{aligned}
f(\lambda y+(1-\lambda)x) &= f(x+\lambda(y-x)) \\
&\le \lambda f(y)+(1-\lambda)f(x) \\
&= f(x)+\lambda(f(y)-f(x)).
\end{aligned}$$

Rearranging the expression yields

$$\frac{f(x+\lambda(y-x))-f(x)}{\lambda}\le f(y)-f(x).$$

Now we take the limit:

$$\lim_{\lambda\to0}\frac{f(x+\lambda(y-x))-f(x)}{\lambda}=\nabla f(x)^T(y-x).$$

This equality can be derived from the following argument: note the Taylor expansion of f at $x+h$ is

$$f(x+th)=f(x)+t\left\langle\nabla f(x),h\right\rangle+o(\|th\|).$$

Then by rearranging we get

$$\frac{f(x+th)-f(x)}{t} = \langle \nabla f(x), h\rangle + \frac{o(\|th\|)}{t\|h\|}\|h\| \xrightarrow{t\to 0} \langle \nabla f(x), h\rangle = \nabla f(x)^T h.$$

Hence, we have

$$\nabla f(x)^T(y-x) \le f(y)-f(x) \iff f(y) \ge f(x) + f(x)^T(y-x)$$

as desired.

(2) $\Rightarrow$ (1). Now assume $f(y) \ge f(x) + \nabla f(x)^T(y-x)$ for any $x, y \in \mathcal{D}_f$. Fix $x, y \in \mathcal{D}_f$. Then note that since $\mathcal{D}_f$ is convex, $\lambda x + (1-\lambda)y \in \mathcal{D}_f$. We first apply it to the pair $(\lambda x + (1-\lambda)y, y)$:

$$\begin{aligned} f(y) &\ge f(\lambda x + (1-\lambda)y) + \nabla f(\lambda x + (1-\lambda)y)^T(y - \lambda x - (1-\lambda)y) \\ &= f(\lambda x + (1-\lambda)y) + \nabla f(\lambda x + (1-\lambda)y)^T \lambda(y-x). \end{aligned} \tag{2}$$

Similarly, we apply it to the pair $(\lambda x + (1-\lambda)y, x)$:

$$f(x) \ge f(\lambda x + (1-\lambda)y) + \nabla f(\lambda x + (1-\lambda)y)(1-\lambda)(x-y). \tag{3}$$

Now, we note that for $\lambda \in (0,1)$,

$$\begin{aligned} (1-\lambda)\times \text{Eq.}(2) + \lambda \times \text{Eq.}(3) =& (1-\lambda f(y)) + \lambda f(x) \\ \ge& (1-\lambda+\lambda) f(\lambda x + (1-\lambda)y) \\ =& f(\lambda x + (1-\lambda)y), \end{aligned}$$

which is the definition of convexity in defined in Eq. (1).

(2) $\Rightarrow$ (3). Pick arbitrary $x, h \in \mathcal{D}_f$. Since $\mathcal{D}_f$ is open, we can find a sufficiently small λ such that $x+\lambda h \in \mathcal{D}_f$. We first write out the second order Taylor expansion of f at $x+\lambda h$,

$$f(x+\lambda h) = f(x) + \lambda \langle \nabla f(x), h\rangle + \frac{\lambda^2}{2} H_f(x)(h,h) + o(\|\lambda h\|^2). \tag{4}$$

Since f is convex, it follows that $f(x+\lambda h) \ge f(x) + \lambda \langle \nabla f(x), h\rangle$. Substituting back to Eq.(4) yields

$$\begin{aligned} \lambda^2 H_f(x)(h,h) + o(\|\lambda h\|^2) \ge 0 &\implies H_f(x)(h,h) + \frac{o(\|\lambda h\|^2)}{\|\lambda h\|^2}\|h\|^2 \ge 0 \qquad \text{(any } \lambda \in (0,1)\text{)} \\ &\implies \lim_{\lambda \to 0^+} \left[H_f(x)(h,h) + \frac{o(\|\lambda h\|^2)}{\|\lambda h\|^2}\|h\|^2 \right] \ge 0 \\ &\implies H_f(x)(h,h) \ge 0. \end{aligned}$$

Since h is arbitrary, it follows that $H_f(x)$ is positive semidefinite.

(3) $\Rightarrow$ (2). Suppose H_f is positive semidefinite. Then for any $x, y \in \mathcal{D}_f$, since $\mathcal{D}_f$ is convex, $\lambda x + (1-\lambda)y \in \mathcal{D}_f$ for any $\lambda \in (0,1)$. Then by a second order Taylor formula, we can write

$$f(y) = f(x) + \langle \nabla f(x), y-x\rangle + \frac{1}{2} H_f(z)(y-x, y-x)$$

for some z in the segment $[x,y] := \{\text{all points of form } \lambda x + (1-\lambda)y \text{ for } \lambda \in (0,1)\}$. Since H_f is positive

semidefinite, it follows that $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$. □

Problem 1.37 - Decomposition of joint entropy

We instead use X, Y to denote the random variable and f_X, f_Y denote marginal distribution and $f_{X,Y}$ joint distribution. Note that

$$\begin{aligned}
H(X,Y) &= \int\int f_{X,Y}(x,y) \log f_{X,Y}(x,y) dxdy \\
&= \int\int f_{X,Y}(x,y) \log f_{Y|X}(y|x) f_X(x) dxdy \\
&= \int\int f_{X,Y}(x,y) \log f_{Y|X}(y|x) dxdy + \int\int f_{X,Y}(x,y) \log f_X(x) dxdy \\
&= H(Y|X) + \int \log f_X(x) \left(\int f_{X,Y}(x,y) dy \right) dx \\
&= H(Y|X) + \int f_X(x) \log f_X(x) dx \\
&= H(Y|X) + H(X),
\end{aligned}$$

as desired.

Problem 1.38 - Proof of discrete Jensen's inequality

We would like to show $f(\sum_{i=1}^M \lambda_i x_i) \leq \sum_{i=1}^M \lambda_i f(x_i)$ for any set of point $\{x_i\}_{i=1}^M$ under the assumption that $\lambda_i \geq 0$ and $\sum \lambda_i = 1$ and f is convex. We show this by inducting on M. For the base case, note that $M = 2$ holds trivially, since by definition of convexity,

$$f(\lambda_1 x_1 + \lambda_2 x_2) = f(\lambda_1 x_1 + (1 - \lambda_1) x_2) \leq \lambda_1 f(x_1) + (1 - \lambda_2) f(x_2) = \lambda_1 f(x_1) + \lambda_2 f(x_2).$$

Now suppose the claim holds for $M = k$. Then for $M = k + 1$, we have

$$\begin{aligned}
f\left(\sum_{i=1}^{k+1} \lambda_i x_i \right) &= f\left(\sum_{i=1}^{k} \lambda_i x_i + \lambda_{k+1} x_{k+1} \right) \\
&= f\left((1 - \lambda_{k+1}) \left(\sum_{i=1}^{k} \frac{\lambda_i}{1 - \lambda_{k+1}} x_i \right) + \lambda_{k+1} x_{k+1} \right) \\
&\leq (1 - \lambda_{k+1}) f\left(\sum_{i=1}^{k} \frac{\lambda_i}{1 - \lambda_{k+1}} x_i \right) + \lambda_{k+1} f(x_{k+1}) \qquad (1)
\end{aligned}$$

where the last inequality follows by treating $\sum_{i=1}^k \lambda_i x_i / (1 - \lambda_{k+1})$ as a singleton point and applying the base case. Now note

$$\sum_{i=1}^{k} \frac{\lambda_i}{1 - \lambda_{k+1}} = \frac{1 - \lambda_{k+1}}{1 - \lambda_{k+1}} = 1.$$

It follows from induction hypothesis that

$$f\left(\sum_{i=1}^{k}\frac{\lambda_i}{1-\lambda_{k+1}}x_i\right)\leq\sum_{i=1}^{k}\frac{\lambda_i}{1-\lambda_{k+1}}f(x_i).$$

Now substituting it back to Eq.(1) and we get

$$\text{Eq.(1)}\leq(1-\lambda_{k+1})\sum_{i=1}^{k}\frac{\lambda_i}{1-\lambda_{k+1}}f(x_i)+\lambda_{k+1}f(x_{k+1})=\sum_{i=1}^{k+1}\lambda_i f(x_i)$$

as desired.

Problem 1.39 - Calculation of entropy and mutual information

1. To find $H(X)$, note

$$H(X)=-\sum_{x\in\{0,1\}}f_X(x)\log f_X(x)=-\frac{2}{3}\log\frac{2}{3}-\frac{1}{3}\log\frac{1}{3}.$$

2. To find $H(Y)$, note

$$H(Y)=-\sum_{y\in\{0,1\}}f_Y(y)\log f_Y(y)=-\frac{2}{3}\log\frac{2}{3}-\frac{1}{3}\log\frac{1}{3}.$$

3. To find $H(X|Y)$, we need to find $f_{X|Y}(x|y)$. Note that

$$f_{X|Y}(x|y)=\frac{f_{X,Y}(x,y)}{f_Y(y)}=\begin{cases}1 & \text{if } x=0,y=0\\ 0 & \text{if } x=1,y=0\\ \frac{1}{2} & \text{if } x=1,y=1 \text{ or } x=0,y=1.\end{cases}$$

   Hence, it follows that

$$H(X|Y)-=\sum_{(x,y)\in\{0,1)\times\{0,1\}}f_{X,Y}(x,y)\log f_{X|Y}(x|y)=-\frac{2}{3}\log\frac{1}{2}.$$

4. Similarly, to find $H(Y|X)$, note that since

$$f_{Y|X}(y|x)=\frac{f_{X,Y}(x,y)}{f_Y(y)}=\begin{cases}0 & \text{if } x=1,y=0\\ 1 & \text{if } x=1,y=1\\ \frac{1}{2} & \text{if } x=0,y=0 \text{ or } x=0,y=1\end{cases},$$

   it follows that

$$H(Y|X)=-\sum_{(x,y)\in\{0,1\}\times\{0,1\}}f_{X,Y}(x,y)\log f_{Y|X}(y|x)=-\frac{2}{3}\log\frac{1}{2}.$$

5. To find $H(X,Y)$, note

$$H(X,Y) = -\sum_{(x,y)\in\{0,1\}\times\{0,1\}} f_{X,Y}(x,y)\log f_{X,Y}(x,y) = \log 3.$$

6. Finally, to find $I(X,Y)$, we note that

$$I(X,Y) = H(X) - H(X|Y) = \frac{2}{3}\log\frac{3}{4} + \frac{1}{3}\log 3.$$

Problem 1.40 - Proof of AM-GM using Jensen's inequality

Since the function $x \mapsto \log x$ is concave, it follows that for any set of points $\{x_i\}_{i=1}^N, N \in \mathbb{N}$ we have

$$\ln\left(\sum_{i=1}^N \frac{1}{N}x_i\right) \geq \sum_{i=1}^N \frac{1}{N}\log(x_i) = \log\left(\prod_{i=1}^N \sqrt[N]{x_i}\right).$$

Next, since $x \mapsto \exp(x)$ preserves monotonicity, it follows that $\sum_{i=1}^N \frac{1}{N}x_i \geq \prod_{i=1}^N \sqrt[N]{x_i}$as desired.

Problem 1.41 - Characterization of mutual information

To show that desired equality, note that

$$\begin{aligned}
I(X,Y) &= -\int\int f_{X,Y}(x,y)\log\frac{f_X(x)f_Y(y)}{f_{X,Y}(x,y)}dxdy \\
&= -\int\int f_{X,Y}(x,y)\log f_X(x)dxdy - \left(-\int\int f_{X,Y}(x,y)\log\frac{f_{X,Y}(x,y)}{f_Y(y)}dxdy\right) \\
&= -\int \log f_X(x)\left(\int f_{X,Y}(x,y)dy\right)dx - \left(-\int\int f_{X,Y}(x,y)\log f_{X|Y}(x|y)dxdy\right) \\
&= \left(-\int f_X(x)\log f_X(x)dx\right) - \left(-\int\int f_{X,Y}(x,y)\log f_{X|Y}(x|y)dxdy\right) \\
&= H(X) - H(X|Y).
\end{aligned}$$

That $I(X,Y) = H(Y) - H(Y|X)$ follows by the same argument but swapping X and Y.

Chapter 2

Solutions for exercises to chapter 2

Problem 2.1 - Bernoulli distribution's expectation, variance, normalization, entropy

In the discussion below, X is a random variable following Bernoulli distribution.

1. To check normalization, we note
$$\sum_{x=0}^{1} f_X(x|\mu) = \mu + (1-\mu) = 1.$$

2. To find the expectation, note that
$$\mathbb{E}[X] = \sum_{x=0}^{1} x f_X(x|\mu) = 1 \cdot \mu + 0 \cdot (1-\mu) = \mu.$$

3. To find the variance, we note that
$$\mathrm{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}X)^2 = \mu - \mu^2 = \mu(1-\mu).$$

4. To find the entropy, we note that
$$H(X) = -\sum_{x=0}^{1} f_X(x|\mu) f \log f_X(x|\mu) = -\mu \log \mu - (1-\mu) \log 1 - \mu.$$

Problem 2.2 - Symmetric Bernoulli distribution's expectation, variance, normalization, entropy

In the discussion below, X is a random variable following the distribution stipulated by Eq.(2.261).

1. To show it's normalized, we note
$$\sum_{x \in \{-1,1\}} f_X(x|\mu) = \left(\frac{1-\mu}{2}\right)^{2/2} \left(\frac{1+\mu}{2}\right)^{0} + \left(\frac{1-\mu}{2}\right)^{0} \left(\frac{1+\mu}{2}\right)^{1} = 1.$$

2. To find its expectation, we note

$$\mathbb{E}[X] = \sum_{x\in\{-1,1\}} x f_X(x|\mu) = \left(\frac{1+\mu}{2}\right) - \left(\frac{1-\mu}{2}\right) = \mu.$$

3. To find its variance, we note

$$\mathrm{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \left(\frac{1+\mu}{2}\right) + \left(\frac{1-\mu}{2}\right) - \mu^2 = 1 - \mu^2.$$

4. To find its entropy, we note

$$H(X) = -\sum_{x\in\{-1,1\}} f_X(x|\mu)\log f_X(x|\mu) = -\left(\frac{1-\mu}{2}\right)\log\frac{1-\mu}{2} - \left(\frac{1+\mu}{2}\right)\log\frac{1+\mu}{2}.$$

Problem 2.3 - Binomial distribution is normalized

1. First, we show Eq.(2.262) holds: note that

$$\begin{aligned}\binom{N}{m} + \binom{N}{m-1} &= \frac{N!}{m!(N-m)!} + \frac{N!}{(m-1)!(N-m+1)!} \\ &= \frac{N!(N-m+1)}{m!(N-m+1)!} + \frac{mN!}{m!(N-m+1)!} \\ &= \frac{(N+1)!}{m!((N+1)-m)!} \\ &= \binom{N+1}{m}.\end{aligned}$$

2. To prove the binomial theorem, we induce on N. For the base case $N = 1$ and 0, it is trivially true:

$$\begin{aligned}(1+x)^1 &= \binom{1}{0}x^0 + \binom{1}{1}x^1 = 1 + x, \\ (1+x)^0 &= \binom{0}{0}x^0 = 1.\end{aligned}$$

Now suppose the claim holds for $N = k$. Then for $N = k+1$ we have

$$\begin{aligned}(1+x)^{k+1} = (1+x)(1+x)^k &= (1+x)\sum_{m=0}^{N}\binom{N}{m}x^m \\ &= \sum_{m=0}^{M}\binom{N}{m}x^m + \sum_{m=0}^{N}\binom{N}{m}x^{m+1} \\ &= \binom{N}{0}x^0 + \sum_{m=1}^{M}\binom{N}{m}x^m + \sum_{m=1}^{N}\binom{N}{m-1}x^m + \binom{N+1}{N+1}x^{N+1} \\ &= \binom{N+1}{0}x^0 + \sum_{m=1}^{M}\left(\binom{N}{m} + \binom{N}{m-1}\right)x^m + \binom{N+1}{N+1}x^{N+1}\end{aligned}$$

$$= \binom{N+1}{0}x^0 + \sum_{m=1}^{M}\binom{N+1}{m}x^m + \binom{N+1}{N+1}x^{N+1}$$
$$= \sum_{m=0}^{N+1}\binom{N}{m}x^m.$$

3. Now to show that the binomial distribution is normalized, we note that

$$\begin{aligned}\sum_{m=0}^{N}\binom{N}{m}\mu^m(1-\mu)^{N-m} &= (1-\mu)^N\sum_{m=0}^{N}\binom{N}{m}\mu^m(1-\mu)^{-m}\\ &= (1-\mu)^N\sum_{m=0}^{N}\binom{N}{m}\left(\frac{\mu}{1-\mu}\right)^m\\ &= (1-\mu)^N\left(1+\frac{\mu}{1-\mu}\right)^N. \qquad \text{(by binomial theorem)}\\ &= \left[(1-\mu)\left(1+\frac{\mu}{1-\mu}\right)\right]^N\end{aligned}$$

Since

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

it follows that $\sum_{m=0}^{N}\binom{N}{m}\mu^m(1-\mu)^{N-m} = 1$, and thus the result follows.

Problem 2.4 - Binomial distribution's expectation and variance

1. Following the hint, we differentiate Eq.(2.264) w.r.t μ once:

$$\begin{aligned}\frac{\partial}{\partial\mu}\left\{\sum_{n=0}^{N}\binom{N}{n}\mu^n(1-\mu)^{N-n}\right\} =& n\cdot\sum_{n=1}^{N-1}\binom{N}{n}\mu^{n-1}(1-\mu)^{N-n} - (N-n)\cdot\sum_{n=1}^{N-1}\binom{N}{n}\mu^n(1-\mu)^{N-n-1}\\ &- N(1-\mu)^{N-1} + N\mu^{N-1}\\ =& n\cdot\sum_{n=1}^{N}\binom{N}{n}\mu^{n-1}(1-\mu)^{N-n} - (N-n)\cdot\sum_{n=0}^{N-1}\binom{N}{n}\mu^n(1-\mu)^{N-n-1}\\ =& n\cdot\sum_{n=0}^{N}\binom{N}{n}\mu^{n-1}(1-\mu)^{N-n} - (N-n)\cdot\sum_{n=0}^{N}\binom{N}{n}\mu^n(1-\mu)^{N-n-1}\\ =& \frac{n}{\mu}\cdot\sum_{n=0}^{N}\binom{N}{n}\mu^n(1-\mu)^{N-n} - \frac{N-n}{1-\mu}\cdot\sum_{n=0}^{N}\binom{N}{n}\mu^n(1-\mu)^{N-n}\\ =& \sum_{n=0}^{N}\binom{N}{n}\mu^n(1-\mu)^{N-n}\left(\frac{n}{\mu}-\frac{N-n}{1-\mu}\right).\end{aligned}$$

Since $\sum_{n=0}^{N}\binom{N}{n}\mu^n(1-\mu)^{N-n}=1$, it follows that

$$\begin{aligned}\sum_{n=0}^{N}\binom{N}{n}\mu^n(1-\mu)^{N-n}\left[\frac{n}{\mu}-\frac{N-n}{1-\mu}\right]=0 &\iff \sum_{n=0}^{N}\binom{N}{n}\mu^n(1-\mu)^{N-n}\left[\frac{n}{\mu}-\frac{N-n}{1-\mu}\right][\mu(1-\mu)]=0\\ &\iff \sum_{n=0}^{N}\binom{N}{n}\mu^n(1-\mu)^{N-n}[n(1-\mu)-(N-n)\mu]=0\\ &\iff \sum_{n=0}^{N}\binom{N}{n}\mu^n(1-\mu)^{N-n}(n-N\mu)=0. \qquad (1)\end{aligned}$$

Now we rearrange Eq.(1):

$$\sum_{n=0}^{N} n\cdot\binom{N}{n}\mu^n(1-\mu)^{N-n}=N\mu\left(\sum_{n=0}^{N}\binom{N}{n}\mu^n(1-\mu)^{N-n}\right)=N\mu.$$

The result follows by observing that

$$\mathbb{E}[X]=\sum_{n=0}^{N} n\cdot\binom{N}{n}\mu^n(1-\mu)^{N-n}$$

2. To facilitate notation, we let $\varphi(\mu)=\sum_{n=1}^{N}\binom{N}{n}\mu^n(1-\mu)^{N-n}\left[\frac{n}{\mu}-\frac{N-n}{1-\mu}\right]$. Then following the hint, we differentiate twice Eq.(2.264) w.r.t. μ and get

$$\begin{aligned}\frac{\partial^2}{\partial\mu^2}\left\{\sum_{n=0}^{N}\binom{N}{n}\mu^n(1-\mu)^{N-n}\right\}&=\frac{\partial\varphi(\mu)}{\partial\mu}\\ &=\sum_{n=0}^{N}\underbrace{\frac{\partial}{\partial\mu}\left\{\binom{N}{n}\mu^n(1-\mu)^{N-n}\left(\frac{n}{\mu}-\frac{N-n}{1-\mu}\right)\right\}}_{:=H(\mu).}.\end{aligned}$$

Hence, it suffices to evaluate $H(\mu)$

$$\begin{aligned}H(\mu)&=\frac{\partial}{\partial\mu}\left\{\binom{N}{n}\mu^n(1-\mu)^{N-n}\right\}\left(\frac{n}{\mu}-\frac{N-n}{1-\mu}\right)+\binom{N}{n}\mu^n(1-\mu)^{N-n}\frac{\partial}{\partial\mu}\left\{\frac{n}{\mu}-\frac{N-n}{1-\mu}\right\}\\ &=\binom{N}{n}\mu^n(1-\mu)^{N-n}\left(\frac{n}{\mu}-\frac{N-n}{1-\mu}\right)^2+\binom{N}{n}\mu^n(1-\mu)^{N-n}\left[-\frac{N-n}{(1-\mu)^2}-\frac{n}{\mu^2}\right]\\ &=\binom{N}{n}\mu^n(1-\mu)^{N-n}\left[\left(\frac{n}{\mu}-\frac{N-n}{1-\mu}\right)^2-\frac{N-n}{(1-\mu)^2}-\frac{n}{\mu^2}\right].\end{aligned}$$

Hence, it follows that

$$\frac{\partial^2}{\partial\mu^2}\left\{\sum_{n=0}^{N}\binom{N}{n}\mu^n(1-\mu)^{N-n}\right\}=\underbrace{\sum_{n=0}^{N}\binom{N}{n}\mu^n(1-\mu)^{N-n}\left[\left(\frac{n}{\mu}-\frac{N-n}{1-\mu}\right)^2-\frac{N-n}{(1-\mu)^2}-\frac{n}{\mu^2}\right]=0}_{(2)}.$$

Now, we arrange Eq.(2) and get

$$\begin{aligned}
&\sum_{n=0}^{N}\binom{N}{n}\mu^n(1-\mu)^{N-n}\left[\left(\frac{n}{\mu}-\frac{N-n}{1-\mu}\right)^2-\frac{N-n}{(1-\mu)^2}-\frac{n}{\mu^2}\right]=0\\
\iff &\sum_{n=0}^{N}\binom{N}{n}\mu^n(1-\mu)^{N-n}\left[\left(\frac{n}{\mu}-\frac{N-n}{1-\mu}\right)^2-\frac{N-n}{(1-\mu)^2}-\frac{n}{\mu^2}\right](\mu^2(1-\mu)^2)=0\\
\iff &\sum_{n=0}^{N}\binom{N}{n}\mu^n(1-\mu)^{N-n}\left[(n(1-\mu)-(N-n)\mu)^2-(N-n)\mu^2-n(1-\mu)^2\right]=0\\
\iff &\sum_{n=0}^{N}\binom{N}{n}\mu^n(1-\mu)^{N-n}\left[(n-N\mu)^2-(N-n)\mu^2-n(1-\mu)^2\right]=0\\
\iff &\sum_{n=0}^{N}\binom{N}{n}\mu^n(1-\mu)^{N-n}(n-N\mu)^2=\sum_{n=0}^{N}\binom{N}{n}\mu^n(1-\mu)^{N-n}\left[(N-n)\mu^2+n(1-\mu)^2\right]\\
\iff &\sum_{n=0}^{N}\binom{N}{n}\mu^n(1-\mu)^{N-n}(n-N\mu)^2=\sum_{n=0}^{N}\binom{N}{n}\mu^n(1-\mu)^{N-n}(N\mu^2+n-2n\mu)\\
\iff &\sum_{n=0}^{N}\binom{N}{n}\mu^n(1-\mu)^{N-n}(n-N\mu)^2=N\mu-N\mu^2=N\mu(1-\mu).
\end{aligned}$$

The conclusion can be drawn by observing that

$$\mathrm{Var}[X]=\mathbb{E}[(X-\mathbb{E}[X])^2]=\sum_{n=0}^{N}\sum_{n=0}^{N}\binom{N}{n}\mu^n(1-\mu)^{N-n}(n-N\mu)^2.$$

Problem 2.5 - Beta distribution is normalized

First, we note that

$$\begin{aligned}
\Gamma(a)\Gamma(b)&=\int_0^\infty e^{-x}x^{a-1}dx\int_0^\infty e^{-y}y^{b-1}dy\\
&=\int_0^\infty\int_0^\infty e^{-(x+y)}x^{a-1}y^{b-1}dydx. \qquad (1)
\end{aligned}$$

Now, we make a change of variable

$$x+y=t\implies\begin{cases}y=t-x\\ y\geq 0\Leftrightarrow t-x\geq 0\Leftrightarrow t\geq x\\ x\geq 0\\ dt=dy\end{cases}.$$

Therefore, it follows that

$$\text{Eq.(1)}=\int_0^\infty\int_x^\infty e^{-t}x^{a-1}(t-x)^{b-1}dtdx$$

$$
\begin{aligned}
&= \int_0^\infty \int_0^t e^{-t} x^{a-1} (t-x)^{b-1} dx dt && \text{(by Fubini's theorem)} \\
&= \int_0^\infty \int_0^1 e^{-t} (t\mu)^{a-1} (t-t\mu)^{b-1} t d\mu dt && (2) \\
&= \int_0^\infty e^{-t} t^{a-1} t^{b-1} t dt \int_0^1 \mu^{a-1} (1-\mu)^{b-1} d\mu \\
&= \Gamma(a+b) \int_0^1 \mu^{a-1} (1-\mu)^{b-1} d\mu,
\end{aligned}
$$

where Eq.(2) follows from a change of variables

$$
x = t\mu \implies \begin{cases} 0 \le x \le t \Leftrightarrow 0 \le t\mu \le t \Leftrightarrow 0 \le \mu \le t \\ dx = t d\mu \end{cases}.
$$

Hence, it follows that

$$
\int_0^1 \mu^{a-1} (1-\mu)^{b-1} d\mu = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)},
$$

and as a result the Beta density integrates to 1.

Problem 2.6 - Beta distribution's expectation, variance, mode

In the discussion below, let X be a random variable that follows Beta distribution with parameter $a, b \in \mathbb{R}^+$.

1. To find the expectation, note that

$$
\begin{aligned}
\mathbb{E}[X] &= \int_0^1 \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1} (1-x)^{b-1} x dx \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 x^{(a+1)-1} (1-x)^{b-1} dx \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} && \text{(by Problem 2.5)} \\
&= \frac{\Gamma(a+b) a \Gamma(a) \Gamma(b)}{\Gamma(a)\Gamma(b)\Gamma(a+b)\Gamma(a+b)} && \text{(since } \Gamma(x+1) = x\Gamma(x).\text{)} \\
&= \frac{a}{a+b}.
\end{aligned}
$$

2. To find the variance, we first note

$$
\begin{aligned}
\mathbb{E}[X^2] &= \int \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1} (1-x)^{b-1} x^2 dx \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 x^{a+2-1} (1-x)^{b-1} dx \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+2)\Gamma(b)}{\Gamma(a+b+2)} \\
&= \frac{a(a+1)}{(a+b+1)(a+b)}.
\end{aligned}
$$

Then it follows that

$$\begin{aligned}\text{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 &= \frac{a(a+1)}{(a+b+1)(a+b)} - \left(\frac{a}{a+b}\right)^2 \\ &= \frac{a(a+1)(a+b) - a^2(a+b+1)}{(a+b+1)(a+b)^2} \\ &= \frac{ab}{(a+b+1)(a+b)^2}.\end{aligned}$$

3. Since the mode of a continuous probability distribution is defined as its density function's critical point, it suffices for us to differentiate $f_X(x)$ and find the critical points. Note that

$$\begin{aligned}\frac{\partial f_X(x)}{\partial x} &= \frac{\partial}{\partial x}\left[\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}\right] \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\left[(a-1)x^{a-2}(1-x)^{b-1} + (b-1)x^{a-1}(1-x)^{b-2}\right].\end{aligned}$$

Setting it to zero yields

$$\begin{aligned}&\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\left[(a-1)x^{a-2}(1-x)^{b-1} + (b-1)x^{a-1}(1-x)^{b-2}\right] = 0 \\ \iff &\left[(a-1)x^{a-2}(1-x)^{b-1} + (b-1)x^{a-1}(1-x)^{b-2}\right] = 0 \\ \iff &(a-1)(1-x) = (b-1)x \\ \iff &x = \frac{a-1}{a+b-2}.\end{aligned}$$

Problem 2.7 - Comparison between posterior mean and MLE for Bernoulli model

The book didn't go through the details of deriving some of the calculations. Although these calculations are simple, they are worth doing by hand at least once. Hence, we show them here. For notation, we let $\mathcal{X}$ denote the sample data, $(x_1, \dots x_N)$.

First, we find the posterior mean for the Bernoulli model. By assumption, the parameter of interest, μ, follows beta distribution, i.e.

$$f(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}.$$

And the likelihood function after sampling the data is given by

$$f(\mathcal{X}|\mu) = \mu^{\sum_{i=1}^N x_i}(1-\mu)^{\sum_{i=1}^N (1-x_i)} = \mu^n(1-\mu)^m.$$

Therefore, we have the posterior as

$$\begin{aligned}f(\mu||\mathcal{X}) &\propto f(\mu|a,b)\cdot f(x_1,\dots,x_N|\mu) \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}\mu^n(1-\mu)^m \\ &\propto \mu^{a+n-1}(1-\mu)^{b+m-1}.\end{aligned}$$

Since $f(\mu|\mathcal{X})$ should integrates to 1 in order to be a valid probability density function, in view of Problem 2.5 we see that

$$f(\mu|\mathcal{X}) = \frac{\Gamma(a+b+n+m)}{\Gamma(a)\Gamma(b)}\mu^{a+n-1}(1-\mu)^{b+m-1} \sim \text{Beta}(a+n, b+m).$$

Hence, it follows that

$$\mathbb{E}_{\mu|\mathcal{X}}[\mu] = \frac{a+n}{a+b+n+m}$$

as desired.

Next, we find μ_{MLE}. First, we write out the likelihood equation,

$$f(\mathcal{X}|\mu) = \prod_{i=1}^{N} \mu^{x_i}(1-\mu)^{1-x_i} = \mu^{\sum_{i=1}^N x_i}(1-\mu)^{\sum_{i=1}^N (1-x_i)},$$

from which we can get the log-likelihood equation as

$$\ell(\mu) = \log f(\mathcal{X}|\mu) = \left(\sum_{i=1}^{N} x_i\right)\log\mu + \left(\sum_{i=1}^{N}(1-x_i)\right)\log(1-\mu).$$

Now we differentiate and set to zero

$$\begin{aligned}
&\frac{\partial \ell(\mu)}{\partial \mu} = \left(\sum_{i=1}^{N} x_i\right)\frac{1}{\mu} - \left(\sum_{i=1}^{N}(1-x_i)\right)\frac{1}{1-\mu} = 0 \\
\iff &\left(\sum_{i=1}^{N} x_i\right)(1-\mu) - \left(\sum_{i=1}^{N}(1-x_i)\right)\mu = 0 \\
\iff &\frac{1}{\mu} = \frac{\sum_{i=1}^N (1-x_i)}{\sum_{i=1}^N x_i} + 1 = \frac{\sum_{i=1}^N 1 - \sum_{i=1}^N x_i + \sum_{i=1}^N x_i}{\sum_{i=1}^N x_i} \\
\iff &\mu_{MLE} = \frac{n}{n+m}.
\end{aligned}$$

Now it suffices to show that

$$\frac{a+n}{a+b+n+m} \in \text{Seg}\left(\frac{a}{a+b}, \frac{n}{n+m}\right),$$

where Seg means the line segment whose endpoints are $a/(a+b)$ and $n/(n+m)$. To show this, it suffices to show that the solution, denoted as λ_*, to the equation

$$\lambda\left(\frac{a}{a+b}\right) + (1-\lambda)\frac{n}{n+m} = \frac{a+n}{a+b+n+m}$$

lies in $(0,1)$. Solving the equation yields

$$\lambda_* = \frac{a+b}{a+b+m+n}.$$

Then the claim is true since $a, b, n, m > 0$ by assumption.

Problem 2.9 - Dirichlet distribution is normalized

In the discussion below, we let $f_D(\mu)$ denote the density function for a Dirichlet distribution whose parameter μ is in K dimensional Euclidean space. We will use a slightly different approach from the one derived from the hint from the book.

We need to show that

$$\int_{\mathbb{S}_K} f_D(\mu)d\mu = \int_{\mathbb{S}_K} \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^K \Gamma(\alpha_i)} \prod_{i=1}^{K-1} \mu_i^{\alpha_i - 1} \left(1 - \sum_{i=1}^{K-1} \mu_i\right)^{\alpha_K - 1} d\mu = 1,$$

where $\mathbb{S}_k := \{x \in \mathbb{R}^k : \sum_{i=1}^k x_k = 1, x_i \geq 0, i = 0, ..., k\}$ is the k-simplex in Euclidean space. Following the idea in Problem 2.5, it suffices for us to show that

$$I_\mu(k) := \int_{\mathbb{S}_k} \prod_{i=1}^{k-1} \mu_i^{\alpha_i} \left(1 - \sum_{i=1}^{k-1} \mu_i\right)^{\alpha_k - 1} d\mu = \frac{\prod_{i=1}^k \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^k \alpha_i)},$$

for any $\mathbb{N} \ni k \geq 2$. We prove this using inducting on k. For the base case $k = 2$, note

$$\begin{aligned}
I_\mu(2) &= \int_{\{\mu \in \mathbb{R}^2 : \mu_1 + \mu_2 = 1, \mu_1 \geq 0, \mu \geq 0\}} \mu_1^{\alpha_1 - 1}(1 - \mu_1)^{\alpha_2 - 1} d\mu \\
&= \int_{\{\mu \in \mathbb{R}^2 : \mu_1 \times \mu_2 \in [0.1] \times [0,1]\}} \mu_1^{\alpha_1 - 1}(1 - \mu_1)^{\alpha_2 - 1} d\mu \qquad (1) \\
&= \int_0^1 d\mu_2 \int_0^1 \mu_1^{\alpha_1 - 1}(1 - \mu_1)^{\alpha_2 - 1} d\mu_1 \qquad \text{(by Fubini's theorem)} \\
&= \frac{\Gamma(a_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1 + \alpha_2)}.
\end{aligned}$$

where Eq.(1) follows from the observation that for any $\mathbb{N} \ni k \geq 2$

$$\begin{aligned}
\mathbb{S}_k &= \left\{ x \in \mathbb{R}^k : \sum_{i=1}^{k-1} x_i = 1 - x_k, x_k \in [0,1], x_i \geq 0, i = 1, ..., k \right\} \\
&= \left\{ x \in \mathbb{R}^k : \sum_{i=1}^{k-1} x_i \leq 1, x_k \in [0,1], x_i \geq 0, i = 1, ..., k-1 \right\},
\end{aligned}$$

where the equality can be verified by an element trace. Now assume the claim is true for $k = n$. Before going into the inductive step, we carefully formulate the inductive hypothesis: note that

$$\begin{aligned}
I_\mu(n) &= \int_{\mathbb{S}_n} \prod_{i=1}^{n-1} \mu_i^{\alpha_i - 1} \left(1 - \sum_{i=1}^{n-1} \mu_i\right)^{\alpha_n - 1} d\mu \\
&= \int_{\left\{\mu \in \mathbb{R}^n : \sum_{i=1}^{n-1} \mu_i \leq 1, \mu_n \in [0,1], \mu_{1 \leq i \leq n-1} \geq 0\right\}} \prod_{i=1}^{n-1} \mu_i^{\alpha_i - 1} \left(1 - \sum_{i=1}^{n-1} \mu_i\right)^{\alpha_n - 1} d\mu \\
&= \int_{\{\mu \in \mathbb{R}^n : \sum_{i=1}^{n-1} \mu_i \leq 1, \mu_{1 \leq i \leq n-1} \in [0,1]\}} \int_0^1 \prod_{i=1}^{n-1} \mu_i^{\alpha_i - 1} \left(1 - \sum_{i=1}^{n-1} \mu_i\right)^{\alpha_n} d\mu_n d(\times_{i=1}^{n-1} \mu_i)
\end{aligned}$$

$$
\begin{aligned}
&= \int_{\{\mu\in\mathbb{R}^n:\sum_{i=1}^{n-1}\mu_i\le 1,\mu_1\le i\le n-1\in[0,1]\}} \prod_{i=1}^{n-1}\mu_i^{\alpha_i-1}\left(1-\sum_{i=1}^{n-1}\mu_i\right)^{\alpha_n} d(\times_{i=1}^{n-1}\mu_i)\int_0^1 d\mu_n \\
&= \int_{\{\mu\in\mathbb{R}^n:\sum_{i=1}^{n-1}\mu_i\le 1,\mu_1\le i\le n-1\in[0,1]\}} \prod_{i=1}^{n-1}\mu_i^{\alpha_i-1}\left(1-\sum_{i=1}^{n-1}\mu_i\right)^{\alpha_n} d(\times_{i=1}^{n-1}\mu_i) \\
&= \int_0^1 \mu_1^{\alpha_1-1}\int_0^{1-\mu_i}\mu_2^{\alpha_2-1}\cdots\int_0^{1-\sum_{i=1}^{n-2}\mu_i}\mu_{n-1}^{\alpha_{n-1}-1}\left(1-\sum_{i=1}^{n-1}\mu_i\right)^{\alpha_n-1} d\mu_{n-1}d\mu_{n-2}\cdots d\mu_1 \qquad (2) \\
&= \frac{\prod_{i=1}^n\Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^n\alpha_i\right)}
\end{aligned}
$$

for any $\{\alpha_1,\ldots,\alpha_n\}$ s.t. $\sum_{i=1}^n\alpha_i = 1$. Also note that Eq.(2) follows from repeated application of Fubini's theorem in the following way:

Eq.(2)

$$
\begin{aligned}
&= \int_{\{\mu\in\mathbb{R}^n:\sum_{i=2}^{n-1}\mu_i\le 1-\mu_1,\mu_1\in[0,1],\mu_2\le i\le n-1\ge 0\}} \prod_{i=1}^{n-1}\mu_i^{\alpha_i-1}\left(1-\sum_{i=1}^{n-1}\mu_i\right)^{\alpha_n} d(\times_{i=1}^{n-1}\mu_i) \\
&= \int_0^1 \mu_1^{\alpha_1-1}\int_{\{(\mu_2,\ldots,\mu_n)\in\mathbb{R}^{n-1}:\sum_{i=2}^{n-1}\mu_i\le 1-\mu_1,\mu_2\le n-1\ge 0\}} \prod_{i=2}^{n-1}\mu_i^{\alpha_i-1}\left(1-\sum_{i=1}^{n-1}\mu_i\right)^{\alpha_n} d(\times_{i=1}^{n-1}\mu_i) \\
&= \int_0^1 \mu_1^{\alpha_1-1}\int_{\left\{\substack{(\mu_2,\ldots,\mu_n)\in\mathbb{R}^{n-1}:\sum_{i=3}^{n-1}\mu_i\le 1-\mu_1-\mu_2 \\ \mu_3\le i\le n-1\ge 0,\mu_2\in[0,1-\mu_1]}\right\}} \prod_{i=2}^{n-1}\mu_i^{\alpha_i-1}\left(1-\sum_{i=1}^{n-1}\mu_i\right)^{\alpha_n} d(\times_{i=2}^{n-1}\mu_i)d\mu_1 \\
&= \int_0^1 \mu_1^{\alpha_1-1}\int_0^{1-\mu_1}\mu_2^{\alpha_2-1}\int_{\left\{\substack{(\mu_3,\ldots,\mu_n)\in\mathbb{R}^{n-2}:\sum_{i=3}^{n-1}\mu_i\le 1-\mu_1-\mu_2, \\ \mu_3\le i\le n-1\ge 0}\right\}} \prod_{i=1}^{n-1}\mu_i^{\alpha_i-1}\left(1-\sum_{i=1}^{n-1}\mu_i\right)^{\alpha_n} d(\times_{i=3}^{n-1}\mu_i)d\mu_2 d\mu_1 \\
&\cdots \\
&= \int_0^1 \mu_1^{\alpha_1-1}\int_0^{1-\mu_i}\mu_2^{\alpha_2-2}\cdots\int_0^{1-\sum_{i=1}^{n-2}\mu_i}\mu_{n-1}^{\alpha_{n-1}-1}\left(1-\sum_{i=1}^{n-1}\mu_i\right)^{\alpha_n-1} d\mu_{n-1}d\mu_{n-2}\cdots d\mu_1 \qquad (2)
\end{aligned}
$$

We also prove a lemma to facilitate the inductive step.

Lemma 2.1. *For any $a\in\mathbb{R}-\{0\}$ and $m,n>0$, the following integral identity holds:*

$$\int_0^1 x^{m-1}(1-x)^{n-1}dx = \frac{1}{a^{m+n-1}}\int_0^a y^{m-1}(a-y)^{n-1}dy,$$

and as a result

$$\int_0^a y^{m-1}(a-y)^{n-1} = a^{m+n-1}\frac{\Gamma(m)\Gamma(n)}{\Gamma(m+n)}$$

Proof. By change of variable $x = y/a$,

$$
\begin{aligned}
\int_0^1 x^{m-1}(1-x)^{n-1}dx &= \frac{1}{a}\int_0^a\left(\frac{y}{a}\right)^{m-1}\left(1-\frac{y}{a}\right)^{n-1}dy \\
&= \frac{1}{a^{m+n-1}}\int_0^a a^{m+n-2}\left(\frac{y}{a}\right)^{m-1}\left(1-\frac{y}{a}\right)^{n-1}dy \\
&= \frac{1}{a^{m+n-1}}\int_0^a a^{m-1}\left(\frac{y}{a}\right)^{m-1}a^{n-1}\left(\frac{a-y}{a}\right)^{n-1}dy
\end{aligned}
$$

$$= \frac{1}{a^{m+n-1}} \int_0^a y^{m-1}(a-y)^{n-1} dy.$$

That $\int_0^a y^{m-1}(a-y)^{n-1} = \frac{1}{a^{m+n-1}} \frac{\Gamma(m)\Gamma(n)}{\Gamma(m+n)}$ then directly follows from Problem 2.5. □

Then for $k = n+1$, again by repeated application of Fubini's theorem we have

$$\begin{aligned} I_\mu(n+1) &= \int_{\mathbb{S}_{n+1}} \prod_{i=1}^n \mu_i^{\alpha_i - 1} \left(1 - \sum_{i=1}^n \mu_i\right)^{\alpha_{n+1}-1} d\mu \\ &= \int_0^1 \mu_1^{\alpha_1-1} \int_0^{1-\mu_1} \mu_2^{\alpha_2-1} \cdots \int_0^{1-\sum_{i=1}^{n-1}\mu_i} \mu_n^{\alpha_n-1} \left(1 - \sum_{i=1}^n \mu_i\right)^{\alpha_{n+1}-1} d\mu_n d\mu_{n-1} \cdots d\mu_1. \end{aligned} \tag{3}$$

Note that by Lem. 2.1

$$\begin{aligned} &\int_0^{1-\sum_{i=1}^{n-1}\mu_i} \mu_n^{\alpha_n-1} \left(1 - \sum_{i=1}^n \mu_i\right)^{\alpha_{n+1}-1} d\mu_n \\ = &\int_0^{1-\sum_{i=1}^{n-1}\mu_i} \mu_n^{\alpha_n-1} \left(1 - \sum_{i=1}^{n-1} \mu_i - \mu_n\right)^{\alpha_{n+1}-1} d\mu_n \\ = &\frac{\Gamma(\alpha_n)\Gamma(\alpha_{n+1})}{\Gamma(\alpha_n+\alpha_{n+1})} \left(1 - \sum_{i=1}^{n-1} \mu_i\right)^{\alpha_n+\alpha_{n+1}-1}. \end{aligned}$$

Therefore,

$$\begin{aligned} \text{Eq.(3)} &= \frac{\Gamma(\alpha_n)\Gamma(\alpha_{n+1})}{\Gamma(\alpha_n+\alpha_{n+1})} \int_0^1 \mu_1^{\alpha_1-1} \cdots \int_0^{1-\sum_{i=1}^{n-2}\mu_i} \mu_{n-1}^{\alpha_{n-1}-1} \left(1 - \sum_{i=1}^{n-1} \mu_i\right)^{\alpha_n+\alpha_{n+1}-1} d\mu_{n-1} \cdots d\mu_1 \\ &= \frac{\Gamma(\alpha_n)\Gamma(\alpha_{n+1})}{\Gamma(\alpha_n+\alpha_{n+1})} \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)\cdots\Gamma(\alpha_n+\alpha_{n+1})}{\Gamma(\alpha_1+\cdots+\alpha_{n+1})} \qquad \text{(by inductive hypothesis)} \\ &= \frac{\prod_{i=1}^n \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^n \alpha_i\right)} \end{aligned}$$

as desired.

Problem 2.10 - Dirichlet distribution's expectation, variance and covariance

In the discussion below, we let μ be a n-dimensional random vector s.t $\mu \sim \text{Dir}(\alpha)$ and $\mathbb{S}_n$ denote the standard simplex in $\mathbb{R}^n$.

1. To find the expectation, note that

$$\begin{aligned} \mathbb{E}[\mu_j] &= \int_{\mathbb{S}_n} \frac{\Gamma(\alpha_1)\cdots\Gamma(\alpha_n)}{\Gamma(\alpha_1+\cdots+\alpha_n)} \mu_j \prod_{i=1}^{n-1} \mu_i^{\alpha_i-1} \left(1 - \sum_{i=1}^{n-1} \mu_i\right)^{\alpha_n-1} d\mu \\ &= \begin{cases} \int_{\mathbb{S}_n} \frac{\Gamma(\alpha_1)\cdots\Gamma(\alpha_n)}{\Gamma(\alpha_1+\cdots+\alpha_n)} \prod_{i=1}^{n-1} \mu_i^{\alpha_i-1} \left(1 - \sum_{i=1}^{n-1} \mu_i\right)^{\alpha_n+1-1} d\mu & \text{if } j = n \\ \int_{\mathbb{S}_n} \frac{\Gamma(\alpha_1)\cdots\Gamma(\alpha_n)}{\Gamma(\alpha_1+\cdots+\alpha_n)} \prod_{i\geq 1, i\neq j}^{n-1} \mu_j^{\alpha_i-1} \mu_j^{\alpha_j+1-1} \left(1 - \sum_{i=1}^{n-1} \mu_i\right)^{\alpha_n-1} d\mu & \text{if } j \in \{1,...,n-1\} \end{cases} \end{aligned}$$

$$= \frac{\Gamma(\alpha_1 + \cdots + \alpha_n)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_n)} \frac{\prod_{i \geq 1, i \neq j}^{n-1} \Gamma(\alpha_i)\Gamma(\alpha_j + 1)}{\Gamma\left(\left(\sum_{i=1}^{n} \alpha_i\right) + 1\right)}$$
$$= \frac{\alpha_j}{\sum_{i=1}^{n} \alpha_i}.$$

2. To find the variance, note that using the same argument as in part(1),

$$\mathbb{E}[\mu_j^2] = \begin{cases} \int_{\mathbb{S}_n} \frac{\Gamma(\alpha_1) \cdots \Gamma(\alpha_n)}{\Gamma(\alpha_1 + \cdots + \alpha_n)} \prod_{i=1}^{n-1} \mu_i^{\alpha_i - 1} \left(1 - \sum_{i=1}^{n-1} \mu_i\right)^{\alpha_n + 2 - 1} d\mu & \text{if } j = n \\ \int_{\mathbb{S}_n} \frac{\Gamma(\alpha_1) \cdots \Gamma(\alpha_n)}{\Gamma(\alpha_1 + \cdots + \alpha_n)} \prod_{i \geq 1, i \neq j}^{n-1} \mu_j^{\alpha_i - 1} \mu_j^{\alpha_j + 2 - 1} \left(1 - \sum_{i=1}^{n-1} \mu_i\right)^{\alpha_n - 1} d\mu & \text{if } j \in \{1, ..., n-1\} \end{cases}$$
$$= \frac{\Gamma(\alpha_1 + \cdots + \alpha_n)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_n)} \frac{\prod_{i \geq 1, i \neq j}^{n-1} \Gamma(\alpha_i)\Gamma(\alpha_j + 2)}{\Gamma\left(\left(\sum_{i=1}^{n} \alpha_i\right) + 2\right)}$$
$$= \frac{\alpha_j(\alpha_j + 1)}{(\sum_{i=1}^{n} \alpha_i + 1)(\sum_{i=1}^{n} \alpha_i)}.$$

Hence,

$$\begin{aligned} \text{Var}[\mu_j] &= \mathbb{E}[\mu_j^2] - (\mathbb{E}[\mu_j])^2 \\ &= \frac{\alpha_j(\alpha_j + 1)}{(\sum_{i=1}^{n} \alpha_i + 1)(\sum_{i=1}^{n} \alpha_i)} - \frac{\alpha_j^2}{(\sum_{i=1}^{n} \alpha_i)^2} \\ &= \frac{(\sum_{i=1}^{n} \alpha_i)\alpha_j(\alpha_j + 1) - (\sum_{i=1}^{n} \alpha_i + 1)\alpha_j^2}{(\sum_{i=1}^{n} \alpha_i + 1)(\sum_{i=1}^{n} \alpha_i)^2} \\ &= \frac{\alpha_j(\sum_{i=1}^{n} \alpha_i - \alpha_j)}{(\sum_{i=1}^{n} \alpha_i + 1)(\sum_{i=1}^{n} \alpha_i)^2}. \end{aligned}$$

3. To find the covariance, note that

$$\mathbb{E}[\mu_i \mu_j] = \begin{cases} \int_{\mathbb{S}_n} \frac{\Gamma(\alpha_1) \cdots \Gamma(\alpha_n)}{\Gamma(\alpha_1 + \cdots + \alpha_n)} \prod_{k \geq 1, k \notin \{i,j\}}^{n-1} \mu_k^{\alpha_i - 1} \mu_{i \in \{i,j\}}^{\alpha_{i \in \{i,j\}} - 1} \left(1 - \sum_{i=1}^{n-1} \mu_i\right)^{\alpha_n + 1 - 1} d\mu & \text{if } i = n \text{ or } j = n \\ \int_{\mathbb{S}_n} \frac{\Gamma(\alpha_1) \cdots \Gamma(\alpha_n)}{\Gamma(\alpha_1 + \cdots + \alpha_n)} \prod_{k \geq 1, k \neq i,j}^{n-1} \mu_k^{\alpha_i - 1} \mu_i^{\alpha_i + 1 - 1} \mu_j^{\alpha_j + 1 - 1} \left(1 - \sum_{i=1}^{n-1} \mu_i\right)^{\alpha_n - 1} d\mu & \text{if } i \neq n \text{ and } j \neq n \end{cases}$$
$$= \frac{\Gamma(\alpha_1) \cdots \Gamma(\alpha_n)}{\Gamma(\alpha_1 + \cdots + \alpha_n)} \frac{\prod_{k \geq 1, k \neq i,j}^{n} \Gamma(\alpha_k)\Gamma(\alpha_i + 1)\Gamma(\alpha_j + 1)}{\Gamma(\sum_{i=k}^{n} \alpha_k + 2)}$$
$$= \frac{\alpha_i \alpha_j}{(\sum_{i=1}^{n} \alpha_i + 1)(\sum_{i=1}^{n} \alpha_i)}.$$

Therefore, it follows that

$$\begin{aligned} \text{Cov}[\mu_i, \mu_j] &= \mathbb{E}[\mu_i \mu_j] - \mathbb{E}[\mu_i]\mathbb{E}[\mu_j] \\ &= \frac{\alpha_i \alpha_j}{(\sum_{i=1}^{n} \alpha_i + 1)(\sum_{i=1}^{n} \alpha_i)} - \frac{\alpha_i \alpha_j}{(\sum_{i=1}^{n} \alpha_i)^2} \\ &= \frac{\alpha_i \alpha_j (\sum_{i=1}^{n} \alpha_i) - (\sum_{i=1}^{n} \alpha_i + 1)\alpha_i \alpha_j}{(\sum_{i=1}^{n} \alpha_i + 1)(\sum_{i=1}^{n} \alpha_i)^2} \\ &= -\frac{\alpha_i \alpha_j}{(\sum_{i=1}^{n} \alpha_i + 1)(\sum_{i=1}^{n} \alpha_i)^2}. \end{aligned}$$

Problem 2.11 - Expression for $\mathbb{E}[\log \mathrm{Dir}(\alpha)]$

In the discussion below, let X be a n-dimensional random vector such that $X \sim \mathrm{Dir}(\alpha)$.

Note that for $(\mu_1, \ldots, \mu_n)$ in the n-dimensional standard simplex, we have

$$\frac{\partial}{\partial \alpha_j}\left[\prod_{i=1}^{n} \mu_i^{\alpha_i - 1}\right] = \ln \mu_j \prod_{i=1}^{n} \mu_i^{\alpha_i - 1}.$$

Then it follows that

$$\begin{aligned}
\mathbb{E}[\ln \mu_j] &= \int_{\mathbb{S}_n} \frac{\Gamma(\sum_{i=1}^n \alpha_i)}{\Pi_{i=1}^n \Gamma(\alpha_i)} \ln \mu_j \prod_{i=1}^{n} \mu_i^{\alpha_i - 1} d\mu \\
&= \int_{\mathbb{S}_n} \frac{\Gamma(\sum_{i=1}^n \alpha_i)}{\Pi_{i=1}^n \Gamma(\alpha_i)} \frac{\partial}{\partial \alpha_j}\left[\prod_{i=1}^{n} \mu_i^{\alpha_i - 1}\right] d\mu \\
&= \frac{\Gamma(\sum_{i=1}^n \alpha_i)}{\Pi_{i=1}^n \Gamma(\alpha_i)} \frac{\partial}{\partial \alpha_j} \int \prod_{i=1}^{n} \mu_i^{\alpha_i - 1} d\mu \qquad \text{(by Leibniz rule)} \\
&= \frac{\Gamma(\sum_{i=1}^n \alpha_i)}{\Pi_{i=1}^n \Gamma(\alpha_i)} \frac{\partial}{\partial \alpha_j}\left[\frac{\Pi_{i=1}^n \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^n \alpha_i)}\right]. \qquad ((1))
\end{aligned}$$

Now we simplify Eq.(1),

$$\begin{aligned}
\text{Eq.}(1) &= \frac{\Gamma(\sum_{i=1}^n \alpha_i)}{\prod_{i=1}^n \Gamma(\alpha_i)}\left[\frac{\prod_{i\geq 1, i\neq j}^n \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^n \alpha_i)} \frac{\partial \Gamma(\alpha_j)}{\partial \alpha_j} - \frac{1}{\Gamma(\sum_{i=1}^n \alpha_i)^2} \frac{\partial \Gamma(\sum_{i=1}^n \alpha_i)}{\partial \alpha_j}\right] \\
&= \frac{\Gamma(\sum_{i=1}^n \alpha_i)}{\prod_{i=1}^n \Gamma(\alpha_i)}\left[\frac{\prod_{i\geq 1, i\neq j}^n \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^n \alpha_i)} \frac{\partial \Gamma(\alpha_j)}{\partial \alpha_j} - \frac{1}{\Gamma(\sum_{i=1}^n \alpha_i)^2} \frac{\partial \Gamma(\sum_{i=1}^n \alpha_i)}{\partial (\sum_{i=1}^n \alpha_i)} \underbrace{\frac{\partial (\sum_{i=1}^n \alpha_i)}{\partial \alpha_j}}_{=1}\right] \\
&= \frac{1}{\Gamma(\alpha_j)} \frac{\partial \Gamma(\alpha_j)}{\partial \alpha_j} - \frac{1}{\Gamma(\sum_{i=1}^n \alpha_i)} \frac{\partial \Gamma(\sum_{i=1}^n \alpha_i)}{\partial (\sum_{i=1}^n \alpha_i)} \\
&= \frac{\partial}{\partial \alpha_j} \ln \Gamma(\alpha_j) - \frac{\partial}{\partial (\sum_{i=1}^n \alpha_i)} \ln \Gamma\Big(\sum_{i=1}^{n} \alpha_i\Big).
\end{aligned}$$

Hence, $\mathbb{E}[\ln \mu_j] = \psi(\alpha_j) - \psi(\sum_{i=1}^n \alpha_i)$ as desired.

Problem 2.12 - Uniform distribution's normalization, expectation, variance

In the discussion below, the X be a random variable such that $X \sim$ Uniform(a, b) with $f_X(x) = \frac{1}{b-a}\mathbb{1}_{[a,b]}$.

1. To see the normalization, note that

$$\int \frac{1}{b-a}\mathbb{1}_{[a,b]} dx = \frac{1}{b-a}(b-a) = 1.$$

2. To find the expectation, note

$$\mathbb{E}[X] = \int_a^b x \frac{1}{b-a} dx = \left[\frac{x^2}{2}\right]_a^b (b-a) = \frac{b^2 - a^2}{2} \frac{1}{(b-a)} = \frac{a+b}{2}.$$

3. To find the variance, first we note that

$$\begin{aligned}\mathbb{E}[X^2] &= \int_a^b x^2 \frac{1}{b-a} dx = \left[\frac{x^3}{3}\right]_a^b (b-a) = \frac{b^3-a^3}{3}\frac{1}{(b-a)} \\ &= \frac{(b-a)(a^2+b^2+ab)}{3(b-a)} = \frac{a^2+b^2+ab}{3}.\end{aligned}$$

And thus

$$\begin{aligned}\mathrm{Var}[X] &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{a^2+b^2+ab}{3} - \left(\frac{a+b}{2}\right)^2 \\ &= \frac{4a^2+4b^2+4ab-3a^2-3b^2-6ab}{12} \\ &= \frac{a^2+b^2-2ab}{12} \\ &= \frac{(a-b)^2}{12}\end{aligned}$$

as desired.

Problem 2.14 - Multidimensional gaussian maximizes entropy

First, we write out the Lagrangian

$$\begin{aligned}\mathcal{L}(p(x)) &= -\int p(x)\ln p(x)dx + \left\langle \begin{bmatrix}\lambda_1\\ \lambda_2\\ \lambda_3\end{bmatrix}, \begin{bmatrix}\int p(x)dx - 1\\ \int p(x)xdx - \mu\\ \int p(x)(x-\mu)(x-\mu)^T dx - \Sigma\end{bmatrix}\right\rangle_{\text{prod}} \\ &= \int -p(x)\ln p(x)dx \\ &\quad + \left\langle \lambda_1, \int p(x)dx - 1\right\rangle + \left\langle \lambda_2, \int p(x)xdx - \mu\right\rangle + \left\langle \lambda_3, \int p(x)(x-\mu)(x-\mu)^T dx - \Sigma\right\rangle \\ &= \int -p(x)\ln p(x)dx \\ &\quad + \lambda_1\left(\int p(x)dx - 1\right) + \lambda_2^T\left(\int p(x)xdx - \mu\right) + \mathrm{tr}\left(\lambda_3^T\left(\int p(x)(x-\mu)(x-\mu)^T dx - \Sigma\right)\right) \\ &= \int -p(x)\ln p(x)dx \\ &\quad + \lambda_1\left(\int p(x)dx - 1\right) + \lambda_2^T\left(\int p(x)xdx - \mu\right) + \mathrm{tr}\left(\lambda_3\left(\int p(x)(x-\mu)(x-\mu)^T dx - \Sigma\right)\right) \quad (1) \\ &= \int -p(x)\ln p(x) + \lambda_1 p(x) + \lambda_2^T p(x)x + \mathrm{tr}((x-\mu)^T\lambda_3 p(x)(x-\mu))dx - (\lambda_1 + \lambda_2^T\mu + \lambda_3\Sigma) \\ &:= \int F(p(x)dx + C, \qquad \text{(by relabeling)}\end{aligned}$$

where Eq.(1) follow can be justified as follows. Note that

$$\int p(x)(x-\mu)(x-\mu)^T dx$$

$$
\begin{aligned}
&= \int \begin{bmatrix} p(x)(x_1-\mu)(x_1-\mu) & p(x)(x_1-\mu)(x_2-\mu) & \cdots & p(x)(x_1-\mu)(x_n-\mu) \\ p(x)(x_2-\mu)(x_1-\mu) & \ddots & \cdots & p(x)(x_2-\mu)(x_n-\mu) \\ \vdots & \vdots & \cdots & \vdots \\ p(x)(x_n-\mu)(x_1-\mu) & p(x)(x_n-\mu)(x_2-\mu) & \cdots & p(x)(x_n-\mu)(x_n-\mu) \end{bmatrix} dx \\
&= \begin{bmatrix} \int p(x)(x_1-\mu)(x_1-\mu)dx & \int p(x)(x_1-\mu)(x_2-\mu)dx & \cdots & \int p(x)(x_1-\mu)(x_n-\mu)dx \\ \int p(x)(x_2-\mu)(x_1-\mu)dx & \ddots & \cdots & \int p(x)(x_2-\mu)(x_n-\mu)dx \\ \vdots & \vdots & \cdots & \vdots \\ \int p(x)(x_n-\mu)(x_1-\mu)dx & \int p(x)(x_n-\mu)(x_2-\mu)dx & \cdots & \int p(x)(x_n-\mu)(x_n-\mu)dx \end{bmatrix},
\end{aligned}
$$

which is symmetric, whence by cyclic property of trace it follows that

$$
\begin{aligned}
\operatorname{tr}\left(\lambda_3^T\left(\int p(x)(x-\mu)(x-\mu)^T dx - \Sigma\right)\right) &= \operatorname{tr}\left(\lambda_3\left(\int p(x)(x-\mu)(x-\mu)^T dx - \Sigma\right)^T\right) \\
&= \operatorname{tr}\left(\lambda_3\left(\int p(x)(x-\mu)(x-\mu)^T dx - \Sigma\right)\right)
\end{aligned}
$$

To maximize, we take the functional derivative and set it to zero:

$$
\begin{aligned}
&\frac{\delta L(p(x)}{\delta p(x)} = \frac{\partial F(p(x))}{\partial p(x)} = -\ln p(x) - 1 + \lambda_1 + \lambda_2^T x + (x-\mu)^T \lambda_3 (x-\mu) = 0 \\
\implies & p(x) = \exp\{\lambda_1 - 1 + \lambda_2^T x + (x-\mu)^T \lambda_3 (x-\mu)\}.
\end{aligned}
$$

Now we substitute $p(x)$ into the constraints:

$$
\begin{aligned}
\int p(x)x dx &= \int \exp\{\lambda_1 - 1 + \lambda_2^T x + (x-\mu)^T \lambda_3 (x-\mu)\} x dx \\
&= \int \exp\left\{\left(x - \mu + \frac{1}{2}\lambda_3^{-1}\lambda_2\right)^T \lambda_3 \left(x - \mu + \frac{1}{2}\lambda_3^{-1}\lambda_2\right) - \frac{1}{4}\lambda_2\lambda_3^{-1}\lambda_2 + \lambda_2^T\mu + \lambda_1 - 1\right\} dx \\
&= \int \exp\left\{y^T\lambda_3 y - \frac{1}{4}\lambda_2\lambda_3^{-1}\lambda_2 + \lambda_2^T\mu + \lambda_1 - 1\right\}\left(y + \mu - \frac{1}{2}\lambda_3^{-1}\lambda_2\right) dy \qquad (2) \\
&= \int \exp\left\{y^T\lambda_3 y - \frac{1}{4}\lambda_2\lambda_3^{-1}\lambda_2 + \lambda_2^T\mu + \lambda_1 - 1\right\} y dy \qquad (3) \\
&\quad + \int \exp\left\{y^T\lambda_3 y - \frac{1}{4}\lambda_2\lambda_3^{-1}\lambda_2 + \lambda_2^T\mu + \lambda_1 - 1\right\}\left(\mu - \frac{1}{2}\lambda_3^{-1}\lambda_2\right) dy \qquad (4) \\
&= \mu,
\end{aligned}
$$

where Eq.(2) follows from change of variable $y = x - \mu + \frac{1}{2}\lambda_3^{-1}\lambda_2$. We take a closer look at Eq.(2). A couple of claims are in order.

Lemma 2.2. *The following identity holds:*

$$
\int_{\mathbb{R}^n} \exp\left\{y^T\lambda_3 y - \frac{1}{4}\lambda_2\lambda_3^{-1}\lambda_2 + \lambda_2^T\mu + \lambda_1 - 1\right\} y dy = 0,
$$

where $y \in \mathbb{R}^n$.

Proof. The key to proving it is that the integrand, denoted as $\varphi(y)$, is a "odd" function in multidimensional

space:

$$\begin{aligned}\varphi(-y) &= \exp\left\{(-y^T)\lambda_3(-y) - \frac{1}{4}\lambda_2\lambda_3^{-1}\lambda_2 + \lambda_2^T\mu + \lambda_1 - 1\right\}(-y)\\ &= -\exp\left\{y^T\lambda_3 y - \frac{1}{4}\lambda_2\lambda_3^{-1}\lambda_2 + \lambda_2^T\mu + \lambda_1 - 1\right\}y\\ &= -\varphi(y).\end{aligned}$$

Then note that

$$\mathbb{R}^n = \underbrace{\left(\bigcup_{(\#_1,\#_2,\cdots,\#_n)\in\prod_{i=1}^n\{+,-\}}\prod_{i=1}^n\mathbb{R}^{\#_i}\right)}_{:=P}\bigcup\underbrace{\left(\bigcup_{(\#_1,\#_2,\cdots,\#_n)\in\prod_{i=1}^n\{0,1\},\exists\#_i=0 \text{ for some } i}\prod_{i=1}^n\mathbb{R}^{\#_i}\right)}_{:=N},$$

where $\mathbb{R}^+ := \{x\in\mathbb{R}|x>0\}$, $\mathbb{R}^- := \{x\in\mathbb{R}|x<0\}$, $\mathbb{R}^0 := \{0\}$,and $\mathbb{R}^1 := \mathbb{R}$. Now we can rewrite the integral of interest as

$$\int_{P\cup N}\exp\left\{y^T\lambda_3 y - \frac{1}{4}\lambda_2\lambda_3^{-1}\lambda_2 + \lambda_2^T\mu + \lambda_1 - 1\right\}ydy = \int_N\varphi(y)dy + \int_P\varphi(y)dy.$$

Since $m(N)=0$, (c.f. [Ste05, Lemma 3.5]), $\int_N\varphi(y)dy = 0$. On the other hand, since P can be written as 2^n disjoint unions by definition (note that for $(\#_1,\ldots,\#_n)\neq(\tilde{\#}_1,\ldots,\tilde{\#}_2)$, we have $(\prod_{i=1}^n\mathbb{R}^{\#_i})\cap(\prod_{i=1}^n\mathbb{R}^{\tilde{\#}_i}) = \emptyset$), we have that

$$\int_P\varphi(y)dy = \int_{\sqcup_{i=1}^{2^n}P_i}\varphi(y)dy = \sum_{i=1}^{2^n}\int_{P_i}\varphi(y)dy.$$

Since for any $i\in\{1,...,2^n\}$, there exists some $j\neq i\in\{1,...,2^n\}$ such that $P_i = (-1)\cdot P_j$, we can rewrite the last term in the previous term as the sum over pairs

$$\begin{aligned}\sum_{i=1}^{2^n}\int_{P_i}\varphi(y)dy &= \sum_{(i,j)}\left(\int_{P_i}\varphi(y)dy + \int_{P_j}\varphi(y)dy\right) = \sum_{(i,j)}\left(\int_{P_i}-\varphi(-y)dy + \int_{P_j}\varphi(y)dy\right)\\ &= \sum_{(i,j)}\left(\int_{-P_i}\varphi(-(-x))dx + \int_{P_j}\varphi(y)dy\right) \qquad \text{(by Thm. 1.1)}\\ &= \sum_{(i,j)}\left(-\int_{P_j}\varphi(x)dx + \int_{P_j}\varphi(y)dy\right)\\ &= 0.\end{aligned}$$

Therefore, it follows that the desired integral is zero. □

Therefore, we have

$$\text{Eq.(2)} = \int\exp\left\{y^T\lambda_3 y - \frac{1}{4}\lambda_2\lambda_3^{-1}\lambda_2 + \lambda_2^T\mu + \lambda_2 - 1\right\}\left(\mu - \frac{1}{2}\lambda_3^{-1}\lambda_2\right)dy. \tag{5}$$

Now we break it down and evaluate Eq.(3) term by term, note that

$$1 = \int \exp\left\{\lambda_1 - 1 + \lambda_2^T x + (x-\mu)^T \lambda_3 (x-\mu)\right\} dx = \int \exp\left\{\lambda_1 - 1 + \lambda_2^T \mu + y^T \lambda_3 y - \frac{1}{4}\lambda_2 \lambda_3^{-1} \lambda_2\right\} dy,$$

by change of variable $y = x - \mu + \frac{1}{2}\lambda_3^{-1}\lambda_2$. Hence, substituting these results back, we get

$$\text{Eq.(2)} = \mu - \frac{1}{2}\lambda_3^{-1}\lambda_2 = \mu \iff \lambda_3^{-1}\lambda_2 = 0 \implies \lambda_2 = 0,$$

where the last implication can be justified as follows: suppose $\lambda_3 = 0$, then $p(x) = \exp\{\lambda_1 - 1 + \lambda_2^T x\}$ is a constant. And thus $\int_{\mathbb{R}^n} p(x)dx = \infty$ unless $p(x) = 0$, in which case the integral evaluates to 0, which does not satisfy Eq.(2.280). Hence, it follows that

$$p(x) = \exp\left\{\lambda_1 - 1 + (x-\mu)^T \lambda_3 (x-\mu)\right\}.$$

Now, we substitute back into the last constraint:

$$\int \exp\left\{\lambda_1 - 1 + (x-\mu)^T \lambda_3 (x-\mu)\right\} (x-\mu)(x-\mu)^T dx = \Sigma.$$

In order to find a solution, we recall that

$$\int \frac{1}{(2\pi)^{D/2}\left|\Sigma\right|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\} (x-\mu)(x-\mu)^T dx = \Sigma.$$

Hence, by comparison of the coefficients, we see that $\lambda_3 = -\frac{1}{2}\Sigma^{-1}$ and

$$\exp\{\lambda_1 - 1\} = \frac{1}{(2\pi)^{D/2}\left|\Sigma\right|^{1/2}} \implies \lambda_1 = \log\left(\frac{1}{(2\pi)^{D/2}\left|\Sigma\right|^{1/2}}\right) + 1$$

forms a set of admissible solution. With this set of λ's, we see that the $p(x)$ is the Gaussian density and thus it follows that multivariate Gaussian distribution is a minimizer of the calculus of variation program proposed in this problem.

Problem 2.15 - Entropy of multivariate gaussian

In the discussion below, let X be a random vector such that $X \sim \text{MVN}(\mu, \Sigma)$ with density $\varphi(x|\mu, \Sigma)$. First, we give a lemma to be used later.

Lemma 2.3. *Let $A \in \text{Mat}_{\mathbb{R}}(n, m)$ and $B(x) \in \text{Mat}_{\mathbb{R}}(m, n)$. Then the following identity holds:*

$$\text{tr}\left(\int AB(x)dx\right) = \int \text{tr}(AB(x))dx.$$

Proof. Just write out the equation and follow definitions:

$$\text{tr}\left(\int AB(x)dx\right) = \sum_{i=1}^{n}\left(\int AB(x)dx\right)_{ii} = \sum_{i=1}^{n}\left(\int \sum_{j=1}^{n} A_{ij}B_{ij}(x)dx\right)$$

$$= \int \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}B_{ij}(x)dx = \int \mathrm{tr}(AB(x))dx$$

as desired. □

Now, we go back to the proof. Note that

$$\begin{aligned}
H(X) &= -\int \varphi(x|\mu,\Sigma)\ln\varphi(x|\mu,\Sigma) \\
&= \int \varphi(x|\mu,\Sigma)\left(\log\frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} + \frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)d\mu \\
&= \log(2\pi)^{D/2} + \log|\Sigma|^{1/2} + \frac{1}{2}\int \varphi(x|\mu,\Sigma)(x-\mu)^T\Sigma^{-1}(x-\mu)d\mu \\
&= \log(2\pi)^{D/2} + \log|\Sigma|^{1/2} + \frac{1}{2}\int \varphi(x|\mu,\Sigma)\mathrm{tr}((x-\mu)^T\Sigma^{-1}(x-\mu))d\mu \\
&= \log(2\pi)^{D/2} + \log|\Sigma|^{1/2} + \frac{1}{2}\int \varphi(x|\mu,\Sigma)\mathrm{tr}(\Sigma^{-1}(x-\mu)(x-\mu)^T)d\mu \\
&= \log(2\pi)^{D/2} + \log|\Sigma|^{1/2} + \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\int \varphi(x|\mu,\Sigma)(x-\mu)(x-\mu)^T d\mu\right) \qquad \text{(by Lem. 2.3)} \\
&= \log(2\pi)^{D/2} + \log|\Sigma|^{1/2} + \frac{1}{2}\mathrm{tr}(\Sigma^{-1}\Sigma) \\
&= \log(2\pi)^{D/2} + \log|\Sigma|^{1/2} + \frac{1}{2}\mathrm{tr}(I_D) \\
&= \frac{D}{2}(\log 2\pi + 1) + \frac{1}{2}\log|\Sigma|\,.
\end{aligned}$$

Problem 2.16 - Entropy of sum of two gaussians

This problem can be solved in various ways. The method proposed by hint given in the problem is limited in the sense that it is hard to generalized to arbitrary transformations and requires a lot of computation, which is error prone. We shall take a different approach here.
First, we give a lemma.

Lemma 2.4. *Let X be a $\mathbb{R}^n$-valued random vector such that $X \sim \mathrm{MVN}(\mu,\Sigma)$. Then $Y = AX + b \sim \mathrm{MVN}(A\mu + b, A\Sigma A^*)$ for $A \in \mathrm{Mat}_{\mathbb{R}}(m,n)$ and $b \in \mathbb{R}^m$.*

To prove this lemma, we need to develop more theory and give more background knowledge about multivariate Gaussian distribution, which we present below.

Supplement knowledge

In the book, the notion of a Gaussian distribution was mainly introduced as a maximizer of a calculus of variation problem under some constraint. This is completely valid and useful. But the addition of some more auxiliary definitions and results will help us gain a more through understanding of this distribution.

In the discussion below, assume we have derived the univariate normal distribution using the book's view point. But now we use another route to push the result to the general setting. First, a few lemmas.

Lemma 2.5. *The characteristic function for the* $\mathrm{N}(\mu, \sigma^2)$ *is given by*

$$\varphi(t) = e^{i\theta a} e^{-\frac{1}{2}\sigma^2\theta^2}.$$

Proof. Following the definition, we have

$$\begin{aligned}
\varphi(t) = \mathbb{E}[\exp(ity)] &= \int_{\mathbb{R}} \exp(itx) \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx \\
&= \frac{1}{\sqrt{2\pi}\sigma} \int_{\mathbb{R}} \exp(it(y+\mu)) \exp\left(-\frac{y^2}{2\sigma^2}\right) dy \qquad \text{(by letting } y = x-\mu\text{)} \\
&= \frac{1}{\sqrt{2\pi}\sigma} e^{it\mu} \underbrace{\int_{\mathbb{R}} \exp(ity) \exp\left(-\frac{y^2}{2\sigma^2}\right) dy}_{:=\phi(t)}.
\end{aligned}$$

In order to find a more explicit form of $\widehat{\mu}(t)$, we evaluate $\phi(t)$. Now note that

$$\left|\frac{\partial}{\partial t}\left(\exp(ity)\exp\left(-\frac{y^2}{2\sigma^2}\right)\right)\right| = \left|iy\exp(ity)\exp\left(-\frac{y^2}{2\sigma^2}\right)\right| \le y\exp\left(-\frac{y^2}{2\sigma^2}\right) \in L^1(\mathbb{R}).$$

Then a corollary of DCT, we have

$$\begin{aligned}
\frac{\partial}{\partial t}\phi(t) &= \int \frac{\partial}{\partial t}\left\{\exp(ity)\exp\left(-\frac{y^2}{2\sigma^2}\right)\right\} dy = \int iy\exp(ity)\exp\left(-\frac{y^2}{2\sigma^2}\right) dy \\
&= \left[-(i\sigma^2)^2\exp(ity)\exp\left(-\frac{y^2}{2\sigma^2}\right)\right]_{-\infty}^{\infty} - \sigma^2 t\int_{-\infty}^{\infty}\exp(ity)\exp\left(-\frac{y^2}{2\sigma^2}\right) dy = -\sigma^2 t\phi(t).
\end{aligned}$$

Note that this is a first order differential equation. Moreover, observe that we also have following initial condition:

$$\phi(0) = \int_{\mathbb{R}} \exp\left(-\frac{y^2}{2\sigma^2}\right) dy = \sqrt{2\pi}\sigma \underbrace{\frac{1}{\sqrt{2\pi}\sigma}\int_{\mathbb{R}}\exp\left(-\frac{y^2}{2\sigma^2}\right) dy}_{=1 \text{ since it's gaussian density}} = \sqrt{2\pi}\sigma.$$

Using integrating factor, we find its general solution if $\phi(t) = ce^{-\frac{1}{2}\sigma^2 t^2}$. Substituting back into the initial condition, we get that $c = \sqrt{2\pi}\sigma$ and as a result $\phi(t) = ce^{-\frac{1}{2}\sigma^2 t^2}$. Hence, it follows that $\varphi(t) = e^{itx}e^{-\frac{1}{2}\sigma^2 t^2}$ as desired. □

Lemma 2.6. *The characteristic function for* $\mathrm{MVN}(\mu, \Sigma)$ *in n-dimensional Euclidean space is given by*

$$\varphi(t) = \exp(i\langle t, \mu\rangle)\exp\left(-\frac{1}{2}\langle \Sigma t, t\rangle\right)$$

Proof. First, we follow the definition of multivariate characteristic function to write

$$\begin{aligned}
\varphi(t) &= \int_{\mathbb{R}^n} \exp(i\langle t, x\rangle)\frac{1}{(2\pi)^{n/2}(\det\Sigma)^{1/2}}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu\right) \\
&= \int_{\mathbb{R}^n} \exp(i\langle t, y+\mu\rangle)\frac{1}{(2\pi)^{n/2}(\det\Sigma)^{1/2}}\exp\left(-\frac{1}{2}y^T\Sigma^{-1}y\right)|\det(J_\varphi(y))|\, dy \qquad (\text{let } x = \varphi(y) = y+\mu) \\
&= \frac{\exp(i\langle t, \mu\rangle)}{(2\pi)^{n/2}(\det\Sigma)^{1/2}} \underbrace{\int_{\mathbb{R}^n}\exp\left(i\langle t, y\rangle - \frac{1}{2}y^T\Sigma^{-1}y\right) dy}_{:=I_1(t)}. \qquad (1)
\end{aligned}$$

Note that since Σ^{-1} is symmetric (by Problem 2.22), it follows that Σ^{-1} has an eigen-decomposition in the form of $\Sigma^{-1} = V\Lambda V^*$ and $\Lambda = V^*\Sigma^{-1}V$, where Λ is a diagonal matrix containing eigen values which are real and $V \in \mathrm{O}(n)$ according to Problem 2.18. Now we make a change of variable as follows:

$$\varphi(x) : \mathbb{R}^n \to \mathbb{R}^n, x \mapsto Vx. \implies |\det J_\varphi(g)| = |\det V| = 1.$$

Then apply this change of variable to the $I_1(t)$ and we get

$$I_1(t) = \int_{\mathbb{R}^n} \exp\left(i\langle t, Vx\rangle - \frac{1}{2}x^T V^T \Sigma^{-1} V x\right) dx = \int_{\mathbb{R}^n} \exp\left(i\langle V^* t, x\rangle - \frac{1}{2}x^T \Lambda x\right) dx.$$

Now, we bring this result back to Eq.(1) and get

$$\begin{aligned}
\varphi(t) &= \frac{\exp(i\langle t, \mu\rangle)}{(2\pi)^{n/2}(\det\Sigma)^{1/2}} \int_{\mathbb{R}^n} \exp\left(i\langle s, x\rangle - \frac{1}{2}x^T\Lambda x\right) dx && \text{(where } s = V^*t\text{)} \\
&= \frac{\exp(i\langle t, \mu\rangle)}{(2\pi)^{n/2}(\det\Sigma)^{1/2}} \int_{\mathbb{R}^n} \exp\left(i\sum_{i=1}^n s_i x_i - \frac{1}{2}\sum_{i=1}^n \frac{x_i^2}{\lambda_i}\right) dx \\
&= \frac{\exp(i\langle t, u\rangle)}{(2\pi)^{n/2}(\det\Sigma)^{1/2}} \prod_{i=1}^n \int_{\mathbb{R}^n} \exp\left(is_i x_i - \frac{1}{2}\frac{x_i^2}{\lambda_i}\right) dx_i && \text{(by Fubini's theroem)} \\
&= \exp(i\langle t, u\rangle) \prod_{i=1}^n \int_{\mathbb{R}^n} \exp(its_i) \frac{1}{\sqrt{2\pi}\lambda_i^{1/2}} \exp\left(-\frac{1}{2}\frac{x_i^2}{\lambda_i}\right) dx_i \\
&= \exp(i\langle t, u\rangle) \prod_{i=1}^n \varphi_{X_i \sim \mathrm{N}(0,\lambda_i)}(s_i) = \exp(i\langle t, \mu\rangle) \exp\left(\sum_{i=1}^n -\frac{\lambda_i s_i^2}{2_i}\right) \\
&= \exp(i\langle t, \mu\rangle) \exp\left(-\frac{1}{2}t^*\Sigma t\right) = \exp(i\langle t, \mu\rangle) \exp\left(-\frac{1}{2}\langle \Sigma t, t\rangle\right),
\end{aligned}$$

as desired. □

Now since set of characteristic functions is an isomorphic to the set of probability distributions (cf.???), we can alternatively define Gaussian distribution using it's characteristic function. One advantage of this characterization is the following lemma.

Lemma 2.7. *Let X be an $\mathbb{R}^n$-valued random variable such that $X \sim \mathrm{MVN}(\mu, \Sigma)$. Then $X =_d \Sigma^{1/2}Z + \mu$, where $Z \sim \mathrm{MVN}(0, I)$.*

Proof. This is a standard result. For the sake of completeness, we provide a complete proof here. Recall a useful lemma:

Lemma 2.8. *Let X be an $\mathbb{R}^n$-valued random variable. Then the characteristic function for $AX + b$ where $A \in \mathrm{Mat}_{\mathbb{K}}(n, m)$ and $b \in \mathbb{R}^m$ can be characterized as $\varphi_{AX+b} = e^{i\langle t,b\rangle}\varphi_X(A^*t)$.*

Proof of Lem. 2.8. Following the definition, we have

$$\begin{aligned}
\varphi_{AX+b}(t) &= \int \exp\left(i\langle t, AX + b\rangle\right) d\Omega = \exp(i\langle t, b\rangle) \int \exp(i\langle t, Ax\rangle) d\Omega = \exp(i\langle t, b\rangle) \int \exp(i\langle A^*t, x\rangle) d\Omega \\
&= \exp(i\langle t, b\rangle \varphi_X(A^*t),
\end{aligned}$$

as desired. □

Now in view of Lem. 2.8 we have

$$\begin{aligned}\varphi_{\Sigma^{1/2}Z+\mu} &= \exp(i\langle t,b\rangle)\varphi_Z(\Sigma^{1/2}t) = \exp(i\langle t,b\rangle)\exp\left(-\frac{1}{2}\left\langle\Sigma^{1/2}t,\Sigma^{1/2}t\right\rangle\right)\\ &= \exp(i\langle t,b\rangle)\exp\left(-\frac{1}{2}\left\langle\Sigma^{1/2}\Sigma^{1/2}t,t\right\rangle\right) = \exp(i\langle t,b\rangle)\exp\left(-\frac{1}{2}\langle\Sigma t,t\rangle\right)\\ &= \varphi_X(t).\end{aligned}$$

Hence, it follows that $X =_d \Sigma^{1/2}Z+\mu$ as desired. □

Proof of Lem. 2.4. In view of Lem. 2.6, Lem. 2.8, we have

$$\begin{aligned}\varphi_{AX+b}(t) &= \exp(i\langle t,b\rangle)\varphi_X(A^*t)\\ &= \exp(i\langle t,b\rangle)\exp(i\langle A^*t,\mu\rangle)\exp\left(-\frac{1}{2}\langle\Sigma A^*t,A^*t\rangle\right)\\ &= \exp(i\langle t,A\mu,b\rangle)\exp\left(-\frac{1}{2}\langle A\Sigma A^*t,t\rangle\right)\\ &= \varphi_{Y\sim\mathrm{MVN}(A\mu+b,A\Sigma A^*)}(t).\end{aligned}$$

Hence, it follows that $AX+b =_d \mathrm{MVN}(A\mu+b,A\Sigma A^*)$. □

Now we go back to the problem itself. Instead of x_1,x_2, we use X_1,X_2 to denote the designated r.v, i.e., $X_1\sim\mathrm{N}(\mu_1,\tau_1^{-1})$ and $X_2\sim\mathrm{N}(\mu_2,\tau_2^{-1})$ and $X_1\perp X_2$. Then it follows that the random vector $\widetilde{X}=\begin{bmatrix}X_1\\X_2\end{bmatrix}\sim\mathrm{MVN}\left(\begin{bmatrix}\mu_1\\\mu_2\end{bmatrix},\begin{bmatrix}\tau_1^{-1}&0\\0&\tau_2^{-1}\end{bmatrix}\right)$. Note that since $X=\mathbf{1}^T\widetilde{X}$, an application of Lem. 2.4 we have that

$$X\sim\mathrm{N}\left(\mathbf{1}^T\begin{bmatrix}\mu_1\\\mu_2\end{bmatrix},\mathbf{1}^T\Sigma\mathbf{1}\right)=\mathrm{N}(\mu_1+\mu_2,\tau_1^{-1}+\tau_2^{-1}).$$

Then, by Problem 1.35, it follows that $H(X_1+X_2)=\frac{1}{2}(1+\ln(2\pi(\tau_1^{-1}+\tau_2^{-1})))$.

Alternative derivation of gaussian mean, covariance In the textbook, the mean and covariance matrix of MVN are derived using change of variables techniques when evaluating the integral. Since we have mentioned that we can characterize a distribution using its characteristic function. Naturally it comes the question of deriving moments of random variables from their characteristic function. Here, we provide a general solution to this problem and apply it to the Gaussian case.

Lemma 2.9. *Let X be a $\mathbb{R}^n$-valued random variable with $\mathbb{E}[\|X\|^N]<\infty$. Then*

$$\mathrm{D}^\alpha\varphi_X(t)=i^{|\alpha|}\int X^\alpha e^{i\langle t,X\rangle}d\Omega,$$

where $\alpha=(\alpha_1,\alpha_2,...,\alpha_n)$ denote the multi-index such that $|\alpha|:=\sum_{i=1}^n\alpha_i\le N$ and $X^\alpha:=X_1^{\alpha_1}X_2^{\alpha_2}\cdots X_n^{\alpha_n}$. As a result,

$$i^{|\alpha|}\mathbb{E}[X^\alpha]=\mathrm{D}^\alpha\varphi_X(0).$$

Proof. To prove this, we induct on N. Now for the base of $N = 1$, we note that $\mathrm{D}^\alpha \varphi_X(t)$ then is reduces to $\frac{\partial}{\partial t_i}\varphi_X(t)$ for some $i \in \{1, ..., n\}$. Note $\frac{\partial}{\partial t_i}\varphi_X(t) = \frac{\partial}{\partial t_i}\int e^{i\langle t, X\rangle} d\Omega$ and that

$$\left|\frac{\partial}{\partial t_i}\left(\exp(i\langle t, X\rangle)\right)\right| = \left|iX_i \exp\left(i\langle t, X\rangle\right)\right| \leq |X_i|.$$

We claim that $|X_i| \in L^1(\Omega)$. To see this, first we note that by Jensen's inequality $(\mathbb{E}[\|X\|])^N \leq \mathbb{E}[\|X\|^N] < \infty$. Take the N-th root, we see that $\|X\| \in L^1$. Clearly, we have $|X_i| \leq (\sum_{i=1}^n X_i^2)^{1/2}$. Hence, chaining these inequalities shows that $|X_i| \in L^1(\Omega)$. Then by a variant of DCT, we can move the differentiation inside and get

$$\frac{\partial}{\partial t_i}\varphi_X(t) = \int \frac{\partial}{\partial t_i}\exp(i\langle t, X\rangle)d\Omega = i\int X_i \exp(i\langle t, X\rangle)d\Omega.$$

Now assume that the claim holds for $N = n-1$. Then for $N = n$, we first note that for some α with $|\alpha| = n$, $\mathrm{D}^\alpha\varphi_X(t) = \frac{\partial}{\partial t_i}(\mathrm{D}^\beta\varphi_X(t))$ for some multi-index β such that $|\beta| = n-1$, and some $i \in \{1, ..., n\}$. Then, we note that first by inductive hypothesis, $\mathrm{D}\beta\varphi_X(t) = i^{|\beta|}\int X^\beta \exp^{i\langle t, X\rangle} d\Omega$ and second,

$$\left|\frac{\partial}{\partial t_i}X^\beta \exp(i\langle t, X\rangle)\right| = \left|iX_iX^\beta \exp(i\langle t, X\rangle)\right| \leq |X^\alpha| = \prod_{i=1}^n |X_i|^{\alpha_i}.$$

Now, we claim that $\prod_{i=1}^n |X_i|^{\alpha_i} \in L^1(\Omega)$. To see this we note that for since $\sum_{i=1}^n \alpha_i = N$, it follows that $\alpha_i \leq N$. As a result, $|X_i|^{\alpha_i} \leq \|X\|^{\alpha_i}$. Therefore, $\prod_{i=1}^n |X_i|^{\alpha_i} \leq \prod_{i=1}^n \|X\|^{\alpha_i} = \|X\|^{\sum_{i=1}^n \alpha_i} = \|X\|^N \in L^1(\Omega)$. Again, by the variant of DCT, we have

$$\begin{aligned}
\mathrm{D}^\alpha\varphi_X(t) &= \frac{\partial}{\partial t_i} i^{|\beta|}\int X^\beta \exp(i\langle t, X\rangle)d\Omega = i^{|\beta|}\int \frac{\partial}{\partial t_i}\left(X^\beta \exp(i\langle t, X\rangle)\right)d\Omega \\
&= i^{|\beta|}\int iX^\alpha \exp(i\langle t, X\rangle)d\Omega = i^{|\beta|+1}\int X^\alpha \exp(i\langle t, X\rangle)d\Omega \\
&= i^{|\alpha|}\int X^\alpha \exp(i\langle t, X\rangle)d\Omega,
\end{aligned}$$

as desired. □

Now we use this result to derive the mean and covariance of a MVN random variable, which we denote as $X \sim \mathrm{MVN}(\mu, \Sigma)$. Recall that by Lem. 2.6, $\varphi_X(t) = \exp(i\langle t, \mu\rangle - \frac{1}{2}\langle \Sigma t, t\rangle)$. Now we find the Frechet derivative w.r.t t: note that by the chain rule:

$$\mathrm{D}\varphi_X(t) = \mathrm{D}\left(\exp(t)\right) \circ \left(t \mapsto i\langle t, \mu\rangle - \frac{1}{2}\langle \Sigma t, t\rangle\right) \circ \mathrm{D}\Bigg(\underbrace{i\langle t, \mu\rangle - \frac{1}{2}\langle \Sigma t, t\rangle}_{:= H_1(t)}\Bigg).$$

Note that

$$\begin{aligned}
H_1(t+h) &= i\langle t+h, \mu\rangle - \frac{1}{2}\langle \Sigma(t+h), t+h\rangle \\
&= H_1(t) + i\langle h, \mu\rangle - \frac{1}{2}\langle \Sigma h, t\rangle - \frac{1}{2}\langle \Sigma t, h\rangle - \frac{1}{2}\langle \Sigma h, h\rangle \\
&= H_1(t) + i\langle h, \mu\rangle - \frac{1}{2}\langle \Sigma h, t\rangle - \frac{1}{2}\langle t, \Sigma^* h\rangle - \frac{1}{2}\langle \Sigma h, h\rangle \\
&= H_1(t) + i\langle h, \mu\rangle - \langle \Sigma h, t\rangle - \frac{1}{2}\langle \Sigma h, h\rangle
\end{aligned}$$

$$= H_1(t) + \langle i\mu - \Sigma t, h\rangle - \frac{1}{2}\langle \Sigma h, h\rangle .$$

Since $\langle \Sigma h, h\rangle \leq \|\Sigma\|_\infty \|h\|^2 \to 0$ at $\|h\| \to 0$, it follows that $\frac{1}{2}\langle \Sigma h, h\rangle = o(\|h\|)$. As $h \mapsto \langle i\mu - \Sigma t, h\rangle \in \mathrm{Hom}(\mathbb{R}^n, \mathbb{R})$, it follows that $\mathrm{D}H_1(t) = i\mu - \Sigma t$. Since $\mathrm{D}(\exp(t)) = \exp(t)$ by elementary calculus, it follows that

$$\mathrm{D}\varphi_X(t) = \exp(i\langle t, \mu\rangle - \frac{1}{2}\langle \Sigma t, t\rangle)(i\mu - \Sigma t).$$

Therefore, $\mathbb{E}[X] = \mathrm{D}\varphi_X(0) = i^{-1}i\mu = \mu$.

Now to find the covariance, we can either partial differentiate term by term or use the same notion of Frechet derivative. We adopt the second method since it is consistent with our previous method, and also yields the total derivative in its matrix form directly, which is more elegant than piecing together terms. Before we go into the calculation, we prepare ourselves with a tool to facilitate the calculation - generalized product rule.

Lemma 2.10. *Suppose the mapping $B : X_1 \times X_2 \to Y$ is bilinear and bounded, i.e.,*

$$\|B(x_1, x_2)\| \leq C\|x_1\|\|x_2\| \text{ for all } x_1 \in X_1, x_2 \in X_2$$

where C is fixed and B linear in each argument. Suppose further that the maps $f_i : X \to X_i$, $i = 1, 2$ are Frechet differentiable at x, and there exist and open set U such that $x \in U$ and $U \subseteq \mathcal{D}_{f_i}$. Then the function $H(x) = B(f_1(x), f_2(x))$ is differentiable at x, and

$$\mathrm{D}H(x)(h) = B(\mathrm{D}f_1(x)(h), f_2(x)) + B(f_1(x), \mathrm{D}f_2'(x)(h)).$$

Note: X_1, X_2, X, Y are all assumed to be Banach spaces.

Proof. We follow the definition and write out $H(x+h) - H(x)$ for later analysis. To facilitate notation, we let $f_i^x := f_i(x)$ for $i = 1, 2$.

$$\begin{aligned}
H(x+h) - H(x) &= B(f_1^{x+h}, f_2^{x+h}) - B(f_1^x, f_2^x) \\
&= B(f_1^{x+h}, f_2^{x+h}) - B(f_1^{x+h}, f_2^x) + B(f_1^{x+h}, f_2^x) - B(f_1^x, f_2^x) \\
&= B(f_1^{x+h}, f_2^{x+h} - f_2^x) + B(f_1^{x+h} - f_1^x, f_2^x) \\
&= B(f_1^x + \mathrm{D}f_1^x(h) + \|h\| r_1(h), \mathrm{D}f_2^x(h) + \|h\| r_2(h)) + B(\mathrm{D}f_1^x(h) + \|h\| r_1(h), f_2^x) \\
&= T_x(h) + R_x(h),
\end{aligned}$$

where

$$\begin{cases}
T_x(h) = & B(f_1^x, \mathrm{D}f_2^x(h)) + B(\mathrm{D}f_1^x(h), f_2^x) \\
R_x(h) = & B(f_1^x, \|h\| r_2(h)) + B(\mathrm{D}f_1^x(h), \mathrm{D}f_2^x(h)) + B(\mathrm{D}f_1^x(h), \|h\| r_2(h)) + B(\|h\|_1 r_1(h), \mathrm{D}f_2^x(h)) \\
& + B(\|h\| r_1(h), \|h\| r_2(h)) + B(\|h\| r_1(h), f_2^x).
\end{cases}$$

In order to show that $T_x(h) = \mathrm{D}H(x) \circ h$. We first need to show that $T(h) \in \mathrm{Hom}(X, Y)$. Indeed,

$$\begin{aligned}
T_x(\alpha h + \beta g) &= B(f_1^x, \mathrm{D}f_2(\alpha h + \beta g)) + B(\mathrm{D}f_1^x(\alpha h + \beta g), f_2^x) \\
&= \alpha B(f_1^x, \mathrm{D}f_2(h)) + \beta B(f_1^x, \mathrm{D}f_2(g)) + \alpha B(\mathrm{D}f_1^x(h), f_2^x) + \beta B(\mathrm{D}f_1^x(g), f_2^x) \\
&= \alpha T_x(h) + \beta T_x(g).
\end{aligned}$$

Next, we need to show that $R_x(h) = o(\|h\|)$. We analyze $R_x(h)$ term by term as follows

$$\begin{cases} B(f_1^x, \|h\| r_2(h)) \le C \|f_1^x\| \|h\| \|r_2(h)\| = o(\|h\|) \\ B(\mathrm{D}f_1^x(h), \mathrm{D}f_2^x(h)) \le C \|\mathrm{D}f_1^x\| \|\mathrm{D}f_2^x\| \|h\|^2 = o(\|h\|) \\ B(\mathrm{D}f_1^x(h), \|h\| r_2(h) \le C \|\mathrm{D}f_1^x\| \|h\|^2 \|r_2(h)\| = o(\|h\|) \\ B(\|h\|_1 r_1(h), \mathrm{D}f_2^x(h)) \le C \|\mathrm{D}f_2^x\| \|h\|^2 \|r_1(h)\| = o(\|h\|) \\ B(\|h\| r_1(h), \|h\| r_2(h)) \le C \|h\|^2 \|r_1(h)\| \|r_2(h)\| = o(\|h\|) \\ B(\|h\| r_1(h), f_2^x) \le C \|f_2^x\| \|h\| \|r_1(h)\| = o(\|h\|). \end{cases}.$$

Since $R_x(h)$ is the sum of these terms, it follows that $R_x(h) = o(\|h\|)$. Hence, it follows that $T_x(h) = \mathrm{D}H(x) \circ h$. □

Now, observe that $\mathrm{D}\varphi_X(t) = \exp(i\langle t, \mu\rangle - \frac{1}{2}\langle \Sigma t, t\rangle)(i\mu - \Sigma t) = B(\exp(i\langle t, \mu\rangle - \frac{1}{2}\langle \Sigma t, t\rangle), i\mu - \Sigma t)$, where $B \in \mathrm{Hom}(\mathbb{R}, \mathbb{R}^n; \mathbb{R}^n)$ is defined by $(x, y) \mapsto xy$. Hence, by Lem. 2.10, it follows that

$$\begin{aligned} \mathrm{D}^2\varphi_X(t)(h) &= B\left(\left\langle \exp\left(i\langle t, \mu\rangle - \frac{1}{2}\langle \Sigma t, t\rangle\right)(i\mu - \Sigma t), h\right\rangle, i\mu - \Sigma t\right) + B\left(\exp\left(i\langle t, \mu\rangle - \frac{1}{2}\langle \Sigma t, t\rangle\right), -\Sigma h\right) \\ &= \exp\left(i\langle t, \mu\rangle - \frac{1}{2}\langle \Sigma t, t\rangle\right)(i\mu - \Sigma t)(i\mu - \Sigma t)^T h - \exp\left(i\langle t, \mu\rangle - \frac{1}{2}\langle \Sigma t, t\rangle\right)\Sigma h \\ &= \exp\left(i\langle t, \mu\rangle - \frac{1}{2}\langle \Sigma t, t\rangle\right)((i\mu - \Sigma t)(i\mu - \Sigma t)^T h - \Sigma h). \end{aligned}$$

Therefore,

$$\mathrm{D}\varphi_X(0)h = (i^2\mu\mu^T - \Sigma)h = (-\mu\mu^T - \Sigma)h \implies \mathbb{E}[XX^T] = -\mathrm{D}^2\varphi_X(0) = \mu\mu^T + \Sigma.$$

As a result, we have according to definition the covariance matrix is $\mathbb{E}[XX^T] - \mu\mu^T = \Sigma$.

Problem 2.17 - Suffices to assume the parameter Σ in Gaussian to be symmetric

This is an direct application of Problem 1.14. Recall the MVN in n-dimensional space has density in the following form:

$$\frac{1}{(2\pi)^{n/2}(\det \Sigma)^{1/2}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right).$$

Since $(x-\mu)^T\Sigma(x-\mu)$ is a bilinear form. In problem 1.14, we showed that it suffices to assume $\Sigma^{-1} = \Sigma_S^{-1} = \frac{1}{2}(\Sigma^{-1} + (\Sigma^{-1})^T)$, which is symmetric since $(x - \mu)^T\Sigma^{-1}(x - \mu) = (x - \mu)^T\Sigma_S^{-1}(x - \mu)$ for all $x \in \mathbb{R}^n$.

Problem 2.18 - Eigen-decomposition for symmetric matrices

Before go into the proof, a lemma. This lemma is usually known as Gram-Schmidt orthogonalization. We prove it in the context of Hilbert space. Proof of this result at various level of generality can be found in any standard linear algebra textbook.

Lemma 2.11. *Let $\mathcal{H}$ be an Hilbert space. Suppose $\mathcal{U} = \{u_1, u_2, \ldots, u_n\}$ is a set of linearly independent vectors of V. Then there exists a set $\mathcal{V} = \{v_1, v_2, \ldots, v_n\}$ of elements of $\mathcal{H}$ such that*

1. $\|v_i\| = 1$ *for* $1 \leq i \leq n$.

2. $\langle v_i, v_j \rangle = 0$ *for* $1 \leq i, j, \leq n$ *with* $i \neq j$.

3. $\text{span}(\mathcal{V}) = \text{span}(\mathcal{U})$.

Proof. We prove this constructively. The construction goes as follows: first we let $v_1 = \frac{u_1}{\|u_1\|}$ and $v_2 = \frac{w_2}{\|w_2\|}$, where $w_2 = u_2 - \langle u_2, v_1 \rangle v_1$, and inductively $v_i = \frac{w_i}{\|w_i\|}$, where $w_i = u_i - \sum_{j=1}^{i-1} \langle u_i, v_j \rangle v_j$, assuming $v_1, ..., v_{i-1}$ have already been defined.

To prove the correctness of our construction, we induct on n. For the base case where $k = 1$. We see that (2) and (3) is trivially satisfied for $v_1 = u_1 / \|u_1\|$. Also, we have $\|v_1\| = \|v_1\| / \|v_1\| = 1$. Now suppose, the construction yields the desired set of vectors $\mathcal{V}_{k=n-1}$ that satisfies condition (1)-(3) for a given set of vectors $\mathcal{U}_{n-1}$. Then for $k = n$,we are given a set of elements in $\mathcal{H}$, $\mathcal{U}_n = \{u_1, \ldots, u_n\}$. By induction hypothesis, we can construct a set of vectors $\mathcal{V}_{n-1} = \{v_1, \ldots, v_{n-1}\}$ from the set $\mathcal{U}_n - \{u_n\}$ that satisfies the conditions (1)-(3) stipulated above with $\mathcal{V} = \mathcal{V}_{n-1}$ and $\mathcal{U} = \mathcal{U}_{n-1}$. Now let $v_n = w_n / \|w_n\|$, where $w_n = u_n - \sum_{i=1}^{n-1} \langle u_n, v_i \rangle v_i$. First, we show that v_n is well-defined. To prove this, we show that $w_n \notin \text{span}(\mathcal{V}_{n-1})$. Suppose otherwise, then $w_n = \sum_{i=1}^{n-1} \alpha_i v_n$ for some scalars $\{\alpha_i\}_{i=1}^{n-1}$. Then we have

$$w_n = \sum_{i=1}^{n-1} \alpha_i v_i = u_n - \sum_{i=1}^{n-1} \langle u_n, v_i \rangle v_i \implies u_n = \sum_{i=1}^{n-1} (\alpha_i + \langle u_n, v_i \rangle) v_i \in \text{span}(\mathcal{V}_{n-1}).$$

Since $\text{span}(\mathcal{V}_{n-1}) = \text{span}(\mathcal{U}_{n-1})$, it follows that $u_n \in \text{span}(\mathcal{U}_{n-1})$. This is a contradiction since then u_n are independent of $u_1, ..., u_{n-1}$ by assumption. We claim that $\mathcal{V}_{n-1} \cup \{v_n\}$ satisfies conditions (1)-(3) with $\mathcal{U} = \mathcal{U}_n$ and $\mathcal{V} = \mathcal{V}_{n-1} \cup \{v_n\}$. Note that by construction $\|v_n\| = 1$. And since for $j \in \{1, ..., n-1\}$

$$\begin{aligned}
\langle w_n, v_j \rangle &= \left\langle u_n - \sum_{i=1}^{n-1} \langle u_n, v_i \rangle v_i, v_j \right\rangle \\
&= \langle u_n, v_j \rangle - \sum_{i=1}^{n-1} (\langle u_n, v_i \rangle \langle v_i, v_j \rangle) \\
&= \langle u_n - v_j \rangle - \langle u_n - v_j \rangle = 0,
\end{aligned}$$

it follows that $\langle w_n / \|w_n\|, v_j \rangle = \langle v_n, v_j \rangle = 0$ for all $j = 1, ..., n$. What's left to prove is that $\text{span}(\mathcal{V}_{n-1} \cup \{v_n\}) = \text{span}(\mathcal{U}_n)$. Pick $\text{span}(\mathcal{V}_{n-1} \cup \{v_n\}) \ni x = \sum_{i=1}^{n} \alpha_i v_i$. If $\alpha_n = 0$, then $x \in \text{span}(\mathcal{V}_{n-1}) = \text{span}(\mathcal{U}_{n-1}) \subset \text{span}(\mathcal{U}_n)$. If $\alpha_n \neq 0$, then

$$\begin{aligned}
x &= \sum_{i=1}^{n-1} \alpha_i v_i + \frac{\alpha_n}{\left\| u_n - \sum_{i=1}^{n-1} \langle u_n, v_i \rangle v_i \right\|} \left(u_n - \sum_{i=1}^{n-1} \langle u_n, v_i \rangle v_i \right) \\
&= \sum_{i=1}^{n-1} (\alpha_i - \langle u_i, v_i \rangle) v_i + \frac{\alpha_n}{\left\| u_n - \sum_{i=1}^{n-1} \langle u_n, v_i \rangle v_i \right\|} u_n \qquad \text{(combining coefficients)} \\
&= \sum_{i=1}^{n-1} \beta_i u_i + \frac{\alpha_n}{\left\| u_n - \sum_{i=1}^{n-1} \langle u_n, v_i \rangle v_i \right\|} u_n.
\end{aligned}$$

Therefore, $x \in \text{span}(\mathcal{U}_n)$. On the other hand, suppose $x \in \text{span}(\mathcal{U}_n)$, i.e. $x = \sum_{i=1}^{n} \alpha_i u_i$. If $\alpha_n = 0$, then

$x \in \text{span}(\mathcal{U}_{n-1}) = \text{span}(\mathcal{V}_{n-1})$ by induction hypothesis. If $\alpha_n \neq 0$, then we have

$$\begin{aligned} x &= \sum_{i=1}^{n-1} \alpha_i u_i + u_n = \sum_{i=1}^{n-1} \beta_i v_i + u_n - \sum_{i=1}^{n-1} \langle u_n, v_i \rangle v_i + \sum_{i=1}^{n-1} \langle u_n, v_i \rangle v_i \\ &= \sum_{i=1}^{n-1} (\beta_i + \langle u_n, v_i \rangle) v_i + v_n, \end{aligned}$$

whence $x \in \text{span}(\mathcal{V}_{n-1} \cup \{v_n\})$. □

1. We first show that the eigenvalues are real. First, we note that by definition, the pair (λ_i, μ_i) is an eigenvector, eigenvalue pair iff $\Sigma\mu_i = \lambda_i \mu$. Now we fix one such pair (μ_i, λ_i). Multiplying μ_i^* on the left on both side of the equation yields $\mu^* \Sigma \mu_i = \lambda_i \mu_i^* \mu_i$. Now since Σ is symmetric, it follows that $(\mu^* \Sigma \mu)^* = \mu^* \Sigma \mu = \lambda_i^* \mu_i^* \mu_i$. Hence, it follows that

$$\lambda_i^* \mu_i^* \mu_i = \lambda_i \mu_i^* \mu_i \iff (\lambda_i^* - \lambda_i) \mu_i^* \mu_i = 0 \iff \lambda_i^* = \lambda_i,$$

since we assume $\mu_i \neq 0$. Therefore, λ_i is real. Since (μ_i, λ_i) is chosen to be arbitrary, it follows that all eigen values in this case are real.

2. Next, we show that for eigen-pair $(\lambda_i, \mu_i), (\lambda_j, \mu_j)$ with $\lambda_i \neq \lambda_j$, and $\lambda_i, \lambda_j \neq 0$, we have $\langle \mu_i, \mu_j \rangle = 0$. First, note the following identity:

$$\lambda_j \mu_j^* \mu_i = (\Sigma \mu_j)^* \mu_i = (\Sigma^* \mu_j)^* \mu_i = \mu_j^* \Sigma \mu_i = \mu_j^* \lambda_i \mu_i = \lambda_i \mu_j^* \mu_i,$$

which implies $\lambda_i \mu_j^* \mu_i = \lambda_j \mu_j^* \mu_i \Leftrightarrow \mu_j^* \mu_i (\lambda_i - \lambda_j) = 0$. Since $\lambda_i \neq \lambda_j$ by assumption, it follows that $\langle \mu_j, \mu_i \rangle = \mu_j^* \mu_i = 0$, as desired.

3. Now we show that for eigen-pair $(\lambda_i, \mu_i), (\lambda_j, \mu_j)$ with $\lambda_i = \lambda_j$ and $\lambda_i, \lambda_j \neq 0$, we can still have $\langle \mu_i, \mu_j \rangle = 0$. For better notation, suppose $\lambda_1 = \lambda_2 := \lambda \in \mathbb{R}$. First, we show that any linear combination of μ_i and μ_j is also an eigen vector with the same eigen value, i.e., $(\lambda, \alpha\mu_i + \beta\mu_2)$ is a valid eigen pair as well for any $\alpha, \beta \in \mathbb{R} - \{0\}$. Indeed, we have

$$\Sigma(\alpha\mu_i + \beta\mu_j) = \alpha\Sigma\mu_i + \beta\Sigma\mu_j = \lambda\alpha\mu_i + \lambda\beta\mu_j = \lambda(\alpha\mu_i + \beta\mu_j).$$

Therefore, in view of Lem. 2.11, we can orthonormalize μ_i and μ_j to $\widetilde{\mu}_i$ and $\widetilde{\mu}_j$ such that $(\lambda, \tilde{\mu}_i)$ and $(\lambda, \widetilde{\mu}_j)$ are eigen-pairs as well ($\widetilde{\mu}_1, \widetilde{\mu}_2$ are linear combination of μ_1 and μ_2).

4. Now we show that for eigen-pair $(\lambda_i, \mu_i), (\lambda_j, \mu_j)$ with at least one of λ_i and λ_j being equal to 0, we can still have $\langle \mu_i, \mu_j \rangle = 0$. Without loss of generality, suppose $\lambda_i = 0$. Then, this means that $\Sigma\mu_i = 0$. Now suppose $\lambda_j \neq 0$, then we have that $\Sigma\mu_j = \lambda_j \mu_j \implies \mu_j = \Sigma\mu_j/\lambda_j$. Then it follows that

$$\langle \mu_i, \mu_j \rangle = \frac{1}{\lambda_j} \mu_i^T \Sigma \mu_j = \frac{1}{\lambda_j} (\Sigma \mu_i)^T \mu_j = 0.$$

Suppose otherwise that $\lambda_j = 0$. Then $\mu_i, \mu_j \in \ker(\Sigma)$, which is a subspace. Therefore, we can orthonormalize μ_i, μ_j by applying Lem. 2.11.

Problem 2.19 - Characterization of Σ, Σ^{-1} in Gaussian distribution

Note that Eq.(2.45) $= \sum_{i=1}^{n} \lambda_i u_i u_i^* = U\Lambda U^*$, where $U = [u_i]$ are vertical stack of eigen vectors of Σ. On the other hand, note that $\Sigma U = U\Lambda$, which implies $\Lambda = U^*\Sigma U$. Therefore, substitute back we get $U\Lambda U^* = UU^*\Sigma UU^* = \Sigma$ as desired.

Problem 2.20 - Positive definite has positive eigenvalues

This is an important result. Hence, we will prove a stronger version by extending the matrix to the $\mathbb{C}$. For later use, we pack the result in the following lemma.

Lemma 2.12. *A matrix $M \in \mathrm{Mat}_{\mathbb{C}}(n, n)$ for some $n \in \mathbb{N}$ is symmetric positive definite*[1] *if and only if all of the eigen values of M_i are positive.*

Proof. $\Leftarrow$ Suppose all of the eigen values of M, denoted by $\{\lambda_i\}_{i=1}^n$ are positive. Note that in view of Problem 2.18, we have $M = V\Lambda V^*$, where V is the matrix containing eigen vectors and Λ is the diagonal matrix of eigen values. So we have for any $x \in \mathbb{C}^n$ that $x^*Mx = (x^*V)\Lambda(V^*x) = \sum_{i=1}^n s_i^2/\lambda_i$, where s_i is the i-th term in the vector V^*x. Sicne $\lambda_i > 0$ for all $i \in \{1, ..., n\}$, it follows that $x^*Mx > 0$ as desired.

$\Rightarrow$ On the other hand, suppose M is positive definite, i.e. $x^*Mx > 0$ for all $x \in \mathbb{C}^n$.[2] Let (λ, v) be an eigen-pair of M. We show that $\lambda > 0$ by case analysis. Suppose $\lambda \leq 0$, then we have $v^*Mv = v^*\lambda v = \lambda |v|^2 \leq 0$, which is a contradiction to the fact that M is positive definite. As a result, $\lambda > 0$. Since λ is chosen arbitrarily, we have shown that any eigen value of M is positive as desired. □

Problem 2.21 - Independent parameter for symmetric matrix

Clearly, once we have known the upper triangular part of the matrix plus the diagonal, we will have known the whole matrix. As a result, the number of independent parameters for symmetric matrix is $\sum_{i=1}^{D} i = D(D+1)/2$.

Problem 2.22 - Inverse of symmetric matrix is symmetric

First, we fix some notations. Let $M \in \mathrm{GL}_{\mathbb{R}}(n)$ be a symmetric matrix. It suffices to show that $(M^{-1})^T = M^{-1}$. Since $M^{-1}M = MM^{-1} = I$, it follows that

$$(MM^{-1})^T = ((M^{-1})^T M^T) = (M^{-1})^T M = I.$$

Hence, it follows that $(M^{-1})^T M = M^{-1}M$. Now multiplying both sides of the previous expression by M^{-1}, we get $(M^{-1})^T MM^{-1} = M^{-1}MM^{-t}$, which implies that $(M^{-1})^T = M^{-1}$.

[1] Although there are matrices that are postivie definite but not symmetric. It suffices for us to assume that M is symmetric in view of Problem 1.14.

[2] One detail we left out here is to show that x^*Mx is real for any $x \in \mathbb{C}^n$. To see this, we observe that $(x^*Mx)^* = x^*M^*x = x^*Mx$, whence is real since its complex conjugate is equal to itself.

Problem 2.23 - Volume of hyperellipsoid in n-dimensional space

The wording of this problem is a bit confusing. Here, we give a clarification: recall that the unit sphere in n-dimensional Euclidean space is given by

$$V_D := \int_{\mathbb{R}^n} \mathbb{1}(\|x\| \leq 1)dx = \int_{\mathbb{R}^n} \mathbb{1}(x^T x \leq 1)dx.$$

Also recall that the solution to this integral has been worked out in Problem 1.18. In the textbook, the hyperellipsoid with Mahalanobis distance Δ is defined to be the set of the form $\{x \in \mathbb{R}^n : (x-\mu)^T\Sigma^{-1}(x-\mu) \leq \Delta^2, \Sigma^{-1}$ is positive (semi)definite$\}$. So to show that the volumne of hyperellipsoid with Mahalanobis distance Δ is equal to $V_D \det|\Sigma|^{1/2}\Delta^D$ is equivalent to showing the following:

$$\int_{\mathbb{R}^n} \mathbb{1}((x-\mu)^T\Sigma^{-1}(x-\mu))dx = V_D \det|\Sigma|^{1/2}\Delta^D.$$

We show this using a series of change of variables(c.f. Thm. 1.1) : note that

$$\begin{aligned}
\int_{\mathbb{R}^n} \mathbb{1}((x-\mu)^T\Sigma^{-1}(x-\mu) \leq \Delta^2)dx &= \int_{\mathbb{R}^n} \mathbb{1}(y^T\Sigma^{-1}y \leq \Delta^2)dy && \text{(by letting } x = y+\mu\text{)} \\
&= \int_{\mathbb{R}^n} \mathbb{1}(y^T\Sigma^{-1/2}\Sigma^{-1/2}y \leq \Delta^2)dy && \text{(since } \Sigma^{-1} \text{ is p.d.)} \\
&= \int_{\mathbb{R}^n} \mathbb{1}(w^Tw \leq \Delta^2)\left|\det\Sigma^{1/2}\right|dw && \text{(by letting } y = \Sigma^{1/2}w\text{)} \\
&= \int_{\mathbb{R}^n} \mathbb{1}(z^Tz \leq 1)\left|\det\Sigma^{1/2}\right||\det\Delta I|\,dz && \text{(by letting } w = (\Delta I)z\text{)} \\
&= \left|\det\Sigma^{1/2}\right|\det|\Delta I|\,V_D \\
&= \det|\Sigma|^{1/2}\Delta^D V_D,
\end{aligned}$$

where the last equality follows since $\det(\Sigma^{1/2}\Sigma^{1/2}) = \det(\Sigma^{1/2})\det(\Sigma^{1/2}) = \det(\Sigma)$, which implies that $\det(\Sigma^{1/2}) = (\det\Sigma)^{1/2}$.

Problem 2.24 - Block matrix inversion formula

To facilitate notation, we use F to denote the matrix defined in the RHS of Eq.(2.76) and F the LHS. To show the identity claimed in Eq.(2.76) holds, we need to show that $FE = EF = I$(we assume the dimensions match in all the partitions and parts of the parition in F are necessarily invertible). Note that the following lemma shows that it suffices for us to show either $FE = I$ or $EF = I$.

Lemma 2.13. *Let $A \in \mathrm{Mat}_{\mathbb{C}}(n,m)$ and $B \in \mathrm{Mat}_{\mathbb{C}}(m,n)$. Then if $AB = I$, it follows that $BA = I$.*

Proof. By assumption we have $AB - I = 0$. We multiply on the left by B to get $BAB - B = (BA-I)B = 0$. Now we let $\{e_i\}_{i=1}^n$ be the standard basis of $\mathbb{R}^n$. We claim that $\{Be_i\}_{i=1}^n$ is also a standard basis. To see this, suppose $\sum_{i=1}^n \alpha_i Be_i = 0$. Multiplying both sides on the left by A and we get $\sum_{i=1}^n \alpha_i ABe_i = 0$, which reduces to $\sum_{i=1}^n \alpha_i e_i = 0$ since $AB = I$ by assumption. Since $\{e_i\}_{i=1}^n$ is the standard basis, it follows that $\alpha_i = 0$ for all $i \in \{1,...,n\}$. As a result $\{Be_i\}_{i=1}^n$ is a set of basis in $\mathbb{R}^n$ as desired. Now we come back to the $(BA-I)B = 0$. Note that from which we can see that $(BA-I)Be_i$ for all $i = 1,...,n$. Since $\{Be_i\}_{i=1}^n$ is a

set of basis, it follows that for any $v \in \mathbb{R}^n$, we have

$$(BA - I)v = (BA - I)\left(\sum_{i=1}^{n} \alpha_i Be_i\right) = \sum_{i=1}^{n} \alpha_i (BA - I)Be_i = 0,$$

whence $BA = I$ as desired. □

In view of previous lemma, we go ahead and show that $FE = I$. Writing the terms out explicitly yields:

$$\begin{aligned} FE &= \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} M & -MBD^{-1} \\ -D^{-1}CM & D^{-1} + D^{-1}CMBD^{-1} \end{bmatrix} \\ &= \begin{bmatrix} AM - B(D^{-1}CM) & -AMBD^{-1} + BD^{-1} + BD^{-1}CMBD^{-1} \\ CM - DD^{-1}CM & -CMBD^{-1} + I + CMBD^{-1} \end{bmatrix} \\ &:= \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix}. \end{aligned}$$

Now we evaluate the blocks P_{ij} term by term:

$$\begin{aligned} &P_{11} = A(A - BD^{-1}C)^{-1} - BD^{-1}C(A - BD^{-1}C)^{-1} = (A - BD^{-1}C)(A - BD^{-1}C)^{-1} = I; \\ &P_{12} = (-AM + I + BD^{-1}CM)BD^{-1} = (-\underbrace{(A - BD^{-1}C)M}_{=I} + I)BD^{-1} = 0; \\ &P_{21} = CM - CM = 0; \\ &P_{22} = I. \end{aligned}$$

Therefore, the results follows.

Problem 2.25 - Marginal and conditional expectation of multivariate gaussian

Although this textbook has derived the result for marginal and conditional distribution for multivariate gaussian distributions using completing squares. It is not done in the most rigiorous manner and left out many of the calculations. Hence, we rederive the results here. However, we will be using a slightly different approach in the derivation, which is more rigorous and slightly more general. Details of the derivation proposed by the book will also be discussed in the reading notes.

To begin with, we introduce a few auxiliary lemmas to help proving later results.

Lemma 2.14. *Suppose $A \in \mathrm{Mat}_{\mathbb{R}}(n, n)$ is positive definite(symmetric) and partitioned as $A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$, then A_{11} and A_{22} are positive definite as well.*

Proof. It suffices to show that A_{11} and A_{22} are p.d since p.d. matrices are invertible. Without loss of generality, we assume A_{11} is of dimension $n_1 \times n_1$ and A_{22} of $n_2 \times n_2$ such that $n_1 + n_2 = n$. Note that since A is p.d., $x^T Ax > 0$ for all $x \in \mathbb{R}^n$. Then for $x = \begin{bmatrix} x_1 \\ 0 \end{bmatrix}$ in block form, where $x_1 \in \mathbb{R}^{n_1}$ is arbitrarily chosen, it follows that $x^T Ax = x_{11}^T A_{11} x_{11} > 0$. Therefore, A_{11} is p.d. as well. On the other hand, if we let $x = \begin{bmatrix} 0 \\ x_2 \end{bmatrix}$

for some arbitrarily chosen $x_2 \in \mathbb{R}^{n_2}$, it also follows that $x^T Ax = x_2^T A_{22} x_2 > 0$. Since x_2 is arbitrarily, A_{22} is p.d. as well. □

Lemma 2.15. *Let $A \in \mathrm{Mat}_{\mathbb{C}}(n,n)$, $D \in \mathrm{Mat}_{\mathbb{C}}(m,m), B \in \mathrm{Mat}_{\mathbb{C}}(n,m)$ for arbitrary $n, m \in \mathbb{N}$. Then it follows that*

$$\det \begin{bmatrix} A & B \\ 0 & D \end{bmatrix} = \det(A) \det(D).$$

Proof. We layout our proof in several steps as below.

1. We first prove the basic case where $D = I$. We claim that $\det \begin{bmatrix} A & B \\ 0 & I \end{bmatrix} = \det A$. To prove this claim, we induct on I's size, denoted as m. For the base case where $m = 1$, we have by Laplace expansion on the last row that
$$\det \begin{bmatrix} A & B \\ 0 & 1 \end{bmatrix} = (-1)^{(n+1)+(n+1)} \det(A) = \det A.$$
Now suppose $m = k$ holds. Then we have that
$$\det \begin{bmatrix} A & B \\ 0 & I_{k+1} \end{bmatrix} = \det \begin{bmatrix} A & B_1 & B_2 \\ 0 & I_k & 0 \\ 0 & 0 & 1 \end{bmatrix} = (-1)^{(n+m+1)+(n+m+1)} \det \begin{bmatrix} A & B \\ 0 & I_k \end{bmatrix} = \det(A),$$
where $B_1 \in \mathrm{Mat}_{\mathbb{C}}(n,m)$ and that $B_2 \in \mathrm{Mat}_{\mathbb{C}}(n,1)$ and the last equality follows from inductive hypoethesis. We also note that using the same argument we can also show that
$$\det \begin{bmatrix} I & B \\ 0 & D \end{bmatrix} = \det(D).$$

2. Now we go back to the proof of the lemma. We first deal the case where A is inverstible, i.e. $\det A \neq 0$. Recall that $\det(MN) = \det(M)\det(N)$ for abitrary compatible matrices M and N. Therefore, it follows that
$$\det \begin{bmatrix} A & B \\ 0 & D \end{bmatrix} = \det \left(\begin{bmatrix} A & 0 \\ 0 & I_m \end{bmatrix} \begin{bmatrix} I_n & A^{-1}B \\ 0 & D \end{bmatrix} \right) = \det \begin{bmatrix} A & 0 \\ 0 & I_m \end{bmatrix} \det \begin{bmatrix} I_n & A^{-1}B \\ 0 & D \end{bmatrix} = \det(A)\det(D),$$
where the last equality follows from step-1.

3. It remains to deal with the case where A is not invertible. Note that if A is not invertible, then the columns of A are not linearly independent. From this we see that the first n columns of $\begin{bmatrix} A & B \\ 0 & D \end{bmatrix}$ are not linearly independent as well since we are stacking 0's under A and as a result the spanning set of first n columns is thus homeomorphic to that of A. Therefore, $\begin{bmatrix} A & B \\ 0 & D \end{bmatrix}$ is not invertible.

□

The following result is a classical theorem, whose proof can be easily found in any measure theoretic probability books. We state without proof.

Theorem 2.1 (Uniqueness of fourier transform)**.** *The Fourier transform of a probability measure on $\mathbb{R}^n$ characterizes μ, that is if two probabilty measures on $\mathbb{R}^n$ admit the same Fourier transform, they are equal.*

In later results, we will construct some independent variables from family of distributions decided by random variables that are not necessarily independent. The proof of this theorem is quite complicated. It's closely related to the Kolmogrov extension theorem.

Theorem 2.2 (Creation of new, independent random variables)**.** *Let $(X_\alpha)_{\alpha\in A}$ be a family of random varibales(not necessarily independent or finite), and let $(\mu_\beta)_{\beta\in B}$ be a collection (not necessarily finite) of probability measures on measurable spaces $(R_\beta)_{\beta\in B}$. Then after extending the sample spaces if necessary, one can find a family $(Y_\beta)_{\beta\in B}$ of independent random variables such that each Y_β has distribution μ_β, and the two families $(X_\alpha)_{\alpha\in A}$ and $(Y_\beta)_{\beta\in B}$ are independent of each other.*

One direct application of Thm. 2.1 is the following lemma.

Lemma 2.16. *Let X be an $\mathbb{R}^n$ valued random variable that has partition of the form $\left[X_1; X_2\right]$, where $X_1 \in \mathbb{R}^k$ and $X_2 \in \mathbb{R}^{n-k}$. Then $X_1 \perp X_2$ iff $\varphi_X(t) = \varphi_{X_1}(t_1)\varphi_{X_2}(t_2)$.*

Proof. $\Rightarrow$ Suppose $X_1 \perp X_2$. Then by a well known result (cf. [Res14, Exercise 4.15]), we have that

$$\varphi_X(t) = \mathbb{E}[e^{i\langle t,X\rangle}] = \mathbb{E}[e^{i\langle t_1,X_1\rangle}e^{i\langle t_2,X_2\rangle}] = \mathbb{E}[e^{i\langle t_1,X_1\rangle}]\mathbb{E}[e^{i\langle t_2,X_2\rangle}] = \varphi_{X_1}(t_1)\varphi_{X_2}(t_2).$$

$\Leftarrow$ We first construct $\widetilde{X}_1$ and $\widetilde{X}_2$ such that $X_1 =_d \widetilde{X}_1$ and $X_2 =_d \widetilde{X}_1$, as well as $\widetilde{X}_1 \perp \widetilde{X}_2$ (the existence of $\widetilde{X}_1$ and $\widetilde{X}_2$ is guaranteed by Thm. 2.2). Then we have

$$\begin{aligned}\varphi_{(X_1,X_2)}((t_1,t_2)) &= \varphi_{X_1}(t_1)\varphi_{X_2}(t_2)\\ &= \varphi_{\tilde{X}_1}(t_1)\varphi_{\tilde{X}_2}(t_2) && \text{(by definition of characteristic functions)}\\ &= \varphi_{(\tilde{X}_1,\tilde{X}_2)}((t_1,t_2)).\end{aligned}$$

Therefore, $\mathbb{P}_{(X_1,X_2)} = \mathbb{P}_{(\widetilde{X}_1,\widetilde{X}_2)}$. Hence, as a result, for any $A \in \mathcal{B}(\mathbb{R}^k), B \in \mathcal{B}(\mathbb{R}^{n-k})$, we have

$$\begin{aligned}\mathbb{P}_{(X_1,X_2)}(X_1 \in A, X_2 \in B) &= \mathbb{P}_{(\widetilde{X}_1,\widetilde{X}_2)}(\widetilde{X}_1 \in A, \widetilde{X}_2 \in B) = \mathbb{P}(\tilde{X}_1 \in A)\mathbb{P}(\tilde{X}_2 \in B)\\ &= \mathbb{P}(X_1 \in A)\mathbb{P}(X_2 \in B).\end{aligned}$$

Hence, X and Y are independent. □

Now, we state and prove some useful result of multivariate gaussian distribution for later use.

Lemma 2.17. *Let X be a $\mathbb{R}^n$ valued random variable such that $X \sim \mathrm{MVN}(\mu, \Sigma)$ and X is paritioned as (for some fixed k)*

$$X = \begin{bmatrix}X_1\\X_2\end{bmatrix}, \quad \mu = \begin{bmatrix}\mu_1\\\mu_2\end{bmatrix} \text{ and } \Sigma = \begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\\Sigma_{21} & \Sigma_{22}\end{bmatrix},$$

where $X_1, \mu_1 \in \mathbb{R}^k$ and $X_2, \mu_2 \in \mathbb{R}^{n-k}$, for $k = 1, ..., n-1$ ($\Sigma_{11}, \Sigma_{12}, \Sigma_{21}, \Sigma_{22}$ are of dimension $k\times k, k\times(n-k), (n-k)\times k, (n-k)\times(n-k)$). Then the follow holds:

1. *$X_1 \sim \mathrm{MVN}(\mu_1, \Sigma_{11})$ and $X_2 \sim \mathrm{MVN}(\mu_2, \Sigma_{22})$;*
2. *$X_1 \perp X_2$ iff $\Sigma_{12} = 0$;*

3. *the conditional distribution of* X_1 *given that* $X_2 = x_2$ *is* $\text{MVN}(\mu_{1\cdot 2}, \Sigma_{11\cdot 2})$, *where* $\mu_{1\cdot 2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)$, *and* $\Sigma_{11\cdot 2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

Proof. 1. To prove this, we take the characteristic function approach. We fist talk about the general case. If we don't make any distributional assumption about X and only assume that it has some density function which we denote as $f_X(x)$. Then we have that

$$\begin{aligned}
\varphi_{X_1}(t) &= \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \exp\left(i\left(\sum_{i=1}^{n_1} t_i x_i\right)\right) f(x_1, ..., x_{n_1}) dx_1 dx_2 \cdots dx_{n_1} \\
&= \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \exp\left(i\left(\sum_{i=1}^{n_1} t_i x_i\right)\right) \left(\int_{\mathbb{R}} \cdots \int_{\mathbb{R}} f(x_1, \ldots, x_n) dx_{n_1+1} \cdots dx_n\right) dx_1 \cdots dx_n \\
&= \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \exp\left(i\left(\sum_{i=1}^{n_1} t_i x_i + \sum_{i=n_{1+1}}^{n} 0 x_i\right)\right) f(x_1, \ldots, x_n) dx_1 \cdots dx_n \\
&= \varphi_X(t, 0),
\end{aligned}$$

where $t \in \mathbb{R}^{n_1}$. Hence, by applying this fact to the gaussian case along with Lem. 2.6 we see that

$$\begin{aligned}
\varphi_{X_1}(t) &= \varphi_X\left(\begin{bmatrix} t \\ 0 \end{bmatrix}\right) \\
&= \exp\left(\left(i\left\langle \begin{bmatrix} t \\ 0 \end{bmatrix}, \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \right\rangle\right) \exp\left(-\frac{1}{2} \begin{bmatrix} t \\ 0 \end{bmatrix}^T \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \begin{bmatrix} t \\ 0 \end{bmatrix}\right) \right. \\
&= \exp(i \langle t, x \rangle) \exp\left(-\frac{1}{2} t^T \Sigma_{11} t\right).
\end{aligned}$$

Hence, $X_1 \sim \text{MVN}(\mu_1, \Sigma_{11})$. That $X_2 \sim \text{MVN}(\mu_2, \Sigma_{22})$ follows from a similar argument.

2. To prove this, note that

$$\Sigma_{12} = 0 \iff t^T \Sigma t \iff t_1^T \Sigma_{11} t_1 + t_2^T \Sigma_{22} t_2,$$

for any $t \in \mathbb{R}^n$ and $t_1 \in \mathbb{R}^k, t_2 \in \mathbb{R}^{n-k}$. Then it follows that

$$\begin{aligned}
\varphi_X(t) &= \exp(i \langle t, x \rangle) \exp\left(-\frac{1}{2} t^T \Sigma_{11} t\right) \\
&= \exp(i \langle t_1, x_1 \rangle) \exp\left(-\frac{1}{2} t_1^T \Sigma_{11} t_1\right) \exp(i \langle t_2, x_2 \rangle) \exp\left(-\frac{1}{2} t_2^T \Sigma_{22} t_2\right) \\
&= \varphi_{X_1}(t_1) \varphi_{X_2}(t_2).
\end{aligned}$$

Then the result follows by an application of Lem. 2.16

3. First, we consider a linear transformation of X in the form as

$$CX = \begin{bmatrix} I_k & -B \\ 0 & I_{n-k} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} X_1 - BX_2 \\ X_2 \end{bmatrix} := \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} := Y$$

Then by Lem. 2.4, it follows that $Y \sim \text{MVN}(\mu_Y, \Sigma_Y)$, where

$$\mu_Y = \begin{bmatrix} I_k & -B \\ 0 & I_{n-k} \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} \mu_1 - B\mu_2 \\ \mu_2 \end{bmatrix};$$
$$\Sigma_Y = \begin{bmatrix} I_k & -B \\ 0 & I_{n-k} \end{bmatrix} \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \begin{bmatrix} I_k & -B^T \\ 0 & I_{n-k} \end{bmatrix} = \begin{bmatrix} \Sigma_{11} - B\Sigma_{12} + \Sigma_{12}B^T - B\Sigma_{22}B^T & \Sigma_{12} - B\Sigma_{22} \\ \Sigma_{12} - B^T\Sigma_{22} & \Sigma_{22} \end{bmatrix}.$$

In view of part-2, if we let $B = \Sigma_{12}\Sigma_{22}^{-1}$, we have that

$$Y \sim \text{MVN}\left(\begin{bmatrix} \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & 0 \\ 0 & \Sigma_{22} \end{bmatrix} \right) = \text{MVN}\left(\widetilde{\mu} = \begin{bmatrix} \mu_{1\cdot 2} \\ \mu_2 \end{bmatrix}, \widetilde{\Sigma} = \begin{bmatrix} \Sigma_{11\cdot 2} & 0 \\ 0 & \Sigma_{22} \end{bmatrix} \right)$$

Then in view of part 1 and 2, it follows that $Y_1 \sim \text{MVN}(\mu_{11\cdot 2}, \Sigma_{11\cdot 2})$ and $Y_2 \sim \text{MVN}(\mu_2, \Sigma_{22})$, and moreover $Y_1 \perp Y_2$. Therefore, Y has the density function in the following form

$$\begin{aligned} f_Y(y) &= f_{(Y_1,Y_2)}((y_1, y_2)) \\ &= \frac{1}{(2\pi)^{n/2}(\det \widetilde{\Sigma})^{1/2}} \exp\left(-\frac{1}{2}(y-\mu)^T\Sigma^{-1}(y-\mu) \right) \\ &= \frac{1}{(2\pi)^{k/2}(\det \Sigma_{11\cdot 2})} \exp\left(-\frac{1}{2}(y_1-\mu_{1\cdot 2})^T\Sigma_{11\cdot 2}^{-1}(y_1-\mu_{1\cdot 2}) \right) \qquad \text{(by Lem. 2.15)} \\ &\quad \times \frac{1}{(2\pi)^{(n-k)/2}\det(\Sigma_{22})} \exp\left(-\frac{1}{2}(y_2-\mu_2)^T\Sigma_{22}^{-1}(y_2-\mu_2) \right) \\ &= f_{Y_1}(y_1) f_{Y_2}(y_2). \end{aligned}$$

Now note that since

$$\begin{bmatrix} I_k & \Sigma_{12}\Sigma_{22} \\ 0 & I_{n-k} \end{bmatrix} Y = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}, \text{ and } \begin{bmatrix} I_k & \Sigma_{12}\Sigma_{22} \\ 0 & I_{n-k} \end{bmatrix}^{-1} = \begin{bmatrix} I_k & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I_{n-k} \end{bmatrix} := M,$$

it follows from Thm. 1.1

$$\begin{aligned} f_{(X_1,X_2)}(x_1, x_2) = f_{(Y_1,Y_2)}\left(\begin{bmatrix} I_k & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I_{n-k} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \right) &= f_{(Y_1,Y_2)}(x_1 - \Sigma_{12}\Sigma_{22}^{-1}x_2, x_2) \\ &= f_{Y_1}(x_1 - \Sigma_{12}\Sigma_{22}^{-1}x_2) f_{X_2}(x_2). \end{aligned}$$

On the other hand, since $f_{(X_1,X_2)}(x_1, x_2) = f_{X_1|X_2}(x_1|x_2) f_{X_2}(x_2)$, the follows that $f_{Y_1}(x_1 - \Sigma_{12}\Sigma_{22}^{-1}x_2) = f_{X_1|X_2}(x_1|x_2)$. Since

$$x_1 - \Sigma_{12}\Sigma_{22}^{-1}x_2 - \mu_1 + \Sigma_{12}\Sigma_{22}\mu_2 = x_1 - (\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) = \mu_{1\cdot 2},$$

we further expand the expression and get

$$\begin{aligned} f_{Y_1}(x_1 - \Sigma_{12}\Sigma_{22}^{-1}x_2) &= \frac{1}{(2\pi)^{k/2}\left|\det \Sigma_{11\cdot 2}\right|^{1/2}} \exp\left(-\frac{1}{2}(x_1-\mu_{1\cdot 2})^T\Sigma_{11\cdot 2}^{-1}(x_1-\mu_{1\cdot 2}) \right) \\ &= f_{\text{MVN}(\mu_{1\cdot 2},\Sigma_{11\cdot 2})}(x_1). \end{aligned}$$

Hence, it follows that $X_1|X_2 = x_2 \sim \text{MVN}(\mu_{1\cdot 2}, \Sigma_{11\cdot 2})$ as desired.

□

Now we go back to the solution to the problem. Since (X_a, X_b, X_c) and $((X_a, X_b), X_c)$ are homeomorphic, it follows from Lem. 2.17 that

$$\begin{bmatrix} X_a \\ X_b \end{bmatrix} \sim \text{MVN}\left(\mu_{a\cdot b}, \Sigma_{a\cdot b}\right), \text{where } \mu_{a\cdot b} = \begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix}, \Sigma_{a\cdot b} = \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix}.$$

Another application of Lem. 2.17-(3), yields that

$$f_{X_a|X_b}(x_a|x_b) = \frac{1}{(2\pi)^{(\dim X_a + \dim X_b)/2}(\det \Sigma_{aa\cdot 2})} \exp\left(-\frac{1}{2}(x - \mu_{a\cdot 2})^T \Sigma_{aa\cdot 2}^{-1}(x - \mu_{a\cdot 2})\right),$$

where $\mu_{a\cdot 2} = \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b - \mu_b)$, and $\Sigma_{aa\cdot 2} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}$.

Problem 2.26 - Woodbury matrix inversion formula

In view of Lem. 2.13, it suffices to show one of the left and right inverse. We show the left inverse here. Note

$$\begin{aligned}
& (A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1})(A + BCD) \\
= \; & I + A^{-1}B(C^{-1} + DA^{-1}B)^{-1}D - A^{-1}BCD - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1}BCD \\
= \; & I + A^{-1}BCD - A^{-1}B(C^{-1} + DA^{-1}B)(DA^{-1}BC + I)D \\
= \; & I + A^{-1}BCD - A^{-1}B(C^{-1} + DA^{-1}B)(DA^{-1}B + C^{-1})CD \\
= \; & I + A^{-1}BCD - A^{-1}BCD \\
= \; & I.
\end{aligned}$$

Problem 2.27 - Linearity of expectation and covariance (multivariate case)

1. For expectation, we note that

$$\begin{aligned}
\mathbb{E}[X + Y] &= \int\int (x + y) f_{(X,Y)}(x, y) dxdy \\
&= \int\int (x + y) f_X(x) f_Y(y) dxdy \\
&= \int_{\text{supp}(X)} x f(x) \left(\int_{\text{supp}(Y)} f(y)dy\right) dx + \int_{\text{supp}(Y)} y f(y) \left(\int_{\text{supp}(X)} f(x)dx\right) dy \\
&= \mathbb{E}[X] + \mathbb{E}[Y].
\end{aligned}$$

2. For covariance, note that

$$\begin{aligned}
\text{Cov}[X + Y] &= \mathbb{E}[(X + Y - \mathbb{E}[X + Y])(X + Y - \mathbb{E}[X + Y])^T] \\
&= \mathbb{E}[(X - \mathbb{E}[X] + Y - \mathbb{E}[Y])(X - \mathbb{E}[X] + Y - \mathbb{E}[Y])^T]
\end{aligned}$$

$$
\begin{aligned}
&= \mathrm{Cov}[X] + \mathrm{Cov}[Y] + \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])^T] + \mathbb{E}[(Y - \mathbb{E}[Y])(X - \mathbb{E}[X])^T] \\
&= \mathrm{Cov}[X] + \mathrm{Cov}[Y] + \mathbb{E}[(X - \mathbb{E}[X])]\mathbb{E}[(Y - \mathbb{E}[Y])^T] + \mathbb{E}[(Y - \mathbb{E}[Y]]\mathbb{E}[X - \mathbb{E}[X]]^T \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (\text{since } X \perp Y) \\
&= \mathrm{Cov}[X] + \mathrm{Cov}[Y].
\end{aligned}
$$

Problem 2.28 - Conditional distribution from joint gaussian

This problem can be solved using the hint provided by the book which adopts the technique of completing squares. However, here we will solve it using the theory we have developed in a few previous exercises. But first, we would like to rephrase this problem to make things clearer: given a joint normal variable Z such that

$$
Z = \begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathrm{MVN}\left(\mu_Z := \begin{bmatrix} \mu \\ A\mu + b \end{bmatrix}, \Sigma_Z := \begin{bmatrix} \Lambda^{-1} & \Lambda^{-1}A^T \\ A\Lambda^{-1} & L^{-1} + A\Lambda^{-1}A^T \end{bmatrix} \right),
$$

find the conditional distribution of $Y|X$, which we denote as $f_{Y|X}(y|x)$.

Without loss generality, assume that $X \in \mathbb{R}^n$ and $Y \in \mathbb{R}^m$. Since $\begin{bmatrix} 0 & I_m \\ I_n & 0 \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} Y \\ X \end{bmatrix}$, it follows from Lem. 2.4 that

$$
\begin{aligned}
\begin{bmatrix} Y \\ X \end{bmatrix} &\sim \mathrm{MVN}\left(\widehat{\mu} = \begin{bmatrix} 0 & I_m \\ I_n & 0 \end{bmatrix} \begin{bmatrix} \mu \\ A\mu + b \end{bmatrix}, \widetilde{\Sigma} = \begin{bmatrix} 0 & I_m \\ I_n & 0 \end{bmatrix} \begin{bmatrix} \Lambda^{-1} & \Lambda^{-1}A^T \\ A\Lambda^{-1} & L^{-1} + A\Lambda^{-1}A^T \end{bmatrix} \begin{bmatrix} 0 & I_m \\ I_n & 0 \end{bmatrix}^T \right) \\
&= \mathrm{MVN}\left(\widetilde{\mu} = \begin{bmatrix} A\mu + b \\ \mu \end{bmatrix}, \widetilde{\Sigma} = \begin{bmatrix} A\Lambda^{-1} & L^{-1} + A\Lambda^{-1}A^T \\ \Lambda^{-1} & \Lambda^{-1}A^T \end{bmatrix} \begin{bmatrix} 0 & I_n \\ I_m & 0 \end{bmatrix} \right) \\
&= \mathrm{MVN}\left(\widetilde{\mu} = \begin{bmatrix} A\mu + b \\ \mu \end{bmatrix}, \widetilde{\Sigma} = \begin{bmatrix} L^{-1} + A\Lambda^{-1}A^T & A\Lambda^{-1} \\ \Lambda^{-1}A^T & \Lambda^{-1} \end{bmatrix} \right).
\end{aligned}
$$

Then we can apply Lem. 2.17-(3) and get

$$
\begin{aligned}
Y|X = x \sim \mathrm{MVN}(&\mu_{Y|X=x} = A\mu_1 + b + (A\Lambda^{-1})\Lambda(x - \mu), \Sigma_{Y|X} = L^{-1} + A\Lambda^{-1}A^T - A\Lambda^{-1}\Lambda\Lambda^{-1}A^T) \\
= \mathrm{MVN}(&\mu_{Y|X=x} = Ax + b, \Sigma_{Y|X} = L^{-1}.
\end{aligned}
$$

as desired.

Problem 2.29 - Verify Eq.(2.105)

Recall from Problem 2.24, for an arbitrary block matrix of the form $\begin{bmatrix} A & B \\ C & D \end{bmatrix}$ d

Bibliography

B

[Bil12] Patrick Billingsley. *Probability and measure*. Wiley, Hoboken, N.J, 2012. 23

C

[Con00] Keith Conrad. Differentiating under the integral sign. 2000. 6

K

[Kuc09] Marek Kuczma. *An introduction to the theory of functional equations and inequalities : Cauchy's equation and Jensen's inequality*. Birkhauser, Basel Boston, 2009. 20

L

[Lan97] Serge Lang. *Undergraduate Analysis*. Springer-Verlag New York, 2 edition, 1997. 6

R

[Res14] Sidney I. Resnick. *A Probability Path*. Modern Birkhäuser classics. Birkhäuser/Springer, New York, 2014. OCLC: ocn869789842. 61

S

[Ste05] Elias Stein. *Real analysis : measure theory, integration, and Hilbert spaces*. Princeton University Press, Princeton, N.J. Oxford, 2005. 13, 14, 46