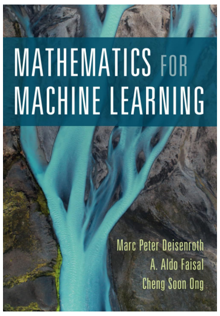
Mathematics for Machine Learning (Deisenroth, Faisal, Ong)




Chapter 6: Probability and Distributions

Full solutions to exercises 6.1 - 6.13

## CONTENTS (clickable links)

6.1 a) $p(x_i) = \sum_{j=1}^{3} p(x_i|y_j)p(y_j) = \sum_{j=1}^{3} p(x_i, y_j)$

i.e. sum the columns.

$p(x_1) = 0.16$

$p(x_2) = 0.17$

$p(x_3) = 0.11$

$p(x_4) = 0.22$

$p(x_5) = 0.34$

$p(y_j) = \sum_{i=1}^{5} p(y_j|x_i)p(x_i) = \sum_{i=1}^{5} p(y_j, x_i)$

i.e sum the rows.

$p(y_1) = 0.26$

$p(y_2) = 0.47$

$p(y_3) = 0.27$

b) $p(x|y_1) = \frac{p(x, y_1)}{p(y_1)} = \frac{p(x, y_1)}{0.26}$

| | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ |
|---|---|---|---|---|---|
| $p(x_i\|y_1)$ | $1/26$ | $2/26$ | $3/26$ | $10/26$ | $10/26$ |

$p(y|x_3) = \frac{p(y, x_3)}{p(x_3)} = \frac{p(y, x_3)}{0.11}$

| | $y_1$ | $y_2$ | $y_3$ |
|---|---|---|---|
| $p(y_j\|x_3)$ | $3/11$ | $5/11$ | $3/11$ |

6.2 a) $p(x) = 0.4\mathcal{N}(10, 1) + 0.6\mathcal{N}(0, 8.4)$

$p(y) = 0.4\mathcal{N}(2, 1) + 0.6\mathcal{N}(0, 1.7)$

b) mean : $\mathbb{E}(x) = 0.4(10) + 0.6(0) = 4$

$\mathbb{E}(y) = 0.4(2) + 0.6(0) = 0.8$

(mode & median must be found numerically.
See MMLCh6 HelperCode. ipynb.)

mode : For $p(x)$ – there are two; 0 and 9.998.

For $p(y)$ – the mode is 1.33

median : For $p(x)$ – the median is 2.803

For $p(y)$ – the median is 0.8686

c) mean : $\mathbb{E}[(x,y)] = 0.4(10, 2) + 0.6(0, 0) = (4, 0.8)$

modes : (calculated numerically)

$(9.99854, 2.00038)$, $(0, 0)$. (5 decimal places).

6.3) Let $y = (x_1, \dots, x_N) \in \mathbb{R}^N$ be series of $N$ Bernoulli trials.

Beta distribution is conjugate for Bernoulli likelihood.

$\mu \sim \text{Beta}(\alpha, \beta)$

$$p(\mu | \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \mu^{\alpha-1} \mu^{\beta-1}.$$

Then $p(\mu | y) \propto p(y|\mu) \cdot p(\mu | \alpha, \beta)$.

$$p(y|\mu) = p((x_1, \dots, x_N)|\mu) = \mu^{x_1}(1-\mu)^{1-x_1} \cdots \mu^{x_N}(1-\mu)^{1-x_N}$$

$$= \mu^{s}(1-\mu)^{1-s}, \quad \text{where } s = \sum_{i=1}^{N} x_i.$$

Note: This is <u>NOT</u> binomial, since the order of the $x_i$ matters

Then $p(y|\mu) \cdot p(\mu|\alpha,\beta)$

$$\propto \mu^{s}(1-\mu)^{1-s} \cdot \mu^{\alpha-1}\mu^{\beta-1}$$

$$= \mu^{\alpha+s-1}(1-\mu)^{1+\beta-s-1}$$

(Ignore $\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$ since we have to renormalise later anyway.)

Normalising gives
$$p(y|\mu) = \frac{\Gamma(1+\alpha+\beta)}{\Gamma(\alpha+s)\Gamma(1+\beta-s)} \mu^{\alpha+s-1}(1-\mu)^{1+\beta-s-1}.$$

where $y = (x_1, \dots, x_N)$, and $s = \sum_{i=1}^{N} x_i$.

6.4) We have

| | | Fruit | |
|---|---|---|---|
| | | Mango | Apple |
| Bag | 1 | $0.6(\frac{2}{3})$ | $0.6(\frac{1}{3})$ |
| | 2 | $0.4(\frac{1}{2})$ | $0.4(\frac{1}{2})$ |

=

| | | Fruit | |
|---|---|---|---|
| | | Mango | Apple |
| Bag | 1 | 0.4 | 0.2 |
| | 2 | 0.2 | 0.2 |

$$p(\text{bag } 2 \mid \text{mango}) = \frac{p(\text{mango} \mid \text{bag } 2) \cdot p(\text{bag } 2)}{p(\text{mango})}$$

$$= \frac{\frac{0.2}{(0.2+0.2)} \cdot 0.4}{(0.4+0.2)} = \frac{0.2}{0.6} = \frac{1}{3}.$$

6.5 a) $p(x_0, \ldots, x_T)$ is Gaussian distributed.

Let $X_0$ be r.v. with $p(X_0 = x_0) = \mathcal{N}(x_0 \mid \mu_0, \Sigma_0)$

Let $X_{t+1} = AX_t + W$ where $A$ fixed, $p(W=w) = \mathcal{N}(w \mid 0, Q)$.

- $X_0$ is Gaussian distributed
- $X_{t+1} = AX_t + W = A^{t+1}X_0 + A^t W + A^{t-1} W + \ldots + AW + W$

So each $X_{t+1}$ is Gaussian distributed as it is the sum of <u>independent</u> r.vs.

(It's essential that the r.vs $W$ are i.i.d.)

Fact: If $\{V_i\}_1^k$ are Gaussian distributed r.v.s, then the r.v $(V_1, V_2, \dots, V_k)$ is (multivariate) Gaussian distributed $\Leftrightarrow$ $\sum a_i V_i$ is Gaussian distributed for all $a_i \neq 0$.

- $aX_0 + bX_1 + \dots$ is an affine combination of $X_0$ and the i.i.d r.v.s $W$, which is Gaussian.

- $(X_0, X_1, \dots, X_T)$ is Gaussian distributed.

b). Assuming $p(x_t | y_1, \dots, y_t) = N(\mu_t, \Sigma_t)$

1) $p(x_{t+1} | y_1, \dots, y_t) = p(Ax_t + w | y_1, \dots, y_t)$.

We have $AX_t \sim N(A\mu_t, A\Sigma_t A^T)$ (Eq (6.88))

$W \sim N(0, Q)$

As $AX_t$ & $W$ are independent, we have by Eq (6.78):

$$X_{t+1} = AX_t + W \sim N(A\mu_t, A\Sigma_t A^T + Q).$$

2) $p(x_{t+1}, y_{t+1} | y_1, \dots, y_t) = \underbrace{p(y_{t+1} | x_{t+1}, y_1, \dots, y_t)} \cdot \underbrace{p(x_{t+1} | y_1, \dots, y_t)}_{\text{see previous part}}$.

key observation: $y_{t+1} = Cx_{t+1} + v$ and $x_{t+1}$ is given. So the only random part is $v$.

$V$ independent of $X_i, Y_i$.

Thus: $p_{Y_{t+1}}(y_{t+1} | x_{t+1}, y_1, \dots, y_t) = p_V(v | x_{t+1}, y_1, \dots, y_t) \overset{\downarrow}{=} p_V(v)$.

so $p(x_{t+1}, y_{t+1} | y_1, \dots, y_t) = p_v(v)\, p_{x_{t+1}}(x_{t+1} | y_1, \dots, y_t)$

where $p_*$ denotes probability from distribution of r.v $*$.

& where $y_{t+1} = Cx_{t+1} + v$.

(Intuitively, this makes sense, as to get a particular value of $x_{t+1}$ AND $y_{t+1}$, you first need to get the value $x_{t+1}$. Then you need to get the value $v = y_{t+1} - Cx_{t+1}$.)

3) $y_{t+1} = Cx_{t+1} + v \iff x_{t+1} = (C^TC)^{-1}C^T(y_{t+1} - v)$

so $p(x_{t+1} | y_1, \dots, y_{t+1})$

$$= p\left(-(C^TC)^{-1}C^Tv + (C^TC)^{-1}C^Ty_{t+1} \,\middle|\, y_1, \dots, y_{t+1}\right)$$

$$\sim N\left((C^TC)^{-1}C^Ty_{t+1}, (C^TC)^{-1}C^TRC(C^TC)^{-T}\right).$$

(using that $p(Av+b) \sim N(A(0)+b, ARA^T) = N(b, ARA^T)$.).

6.6) $$V_x[x] := \mathbb{E}_x[(x - \mathbb{E}[x])^2]$$
$$= \mathbb{E}[x^2 - 2\mathbb{E}[x]x + \mathbb{E}[x]^2]$$
$$= \mathbb{E}[x^2] - 2\mathbb{E}[x]\mathbb{E}[x] + \mathbb{E}[\mathbb{E}[x]^2]$$
$$= \mathbb{E}[x^2] - 2\mathbb{E}[x]^2 + \mathbb{E}[x]^2$$
$$= \mathbb{E}[x^2] - \mathbb{E}[x]^2.$$

6.7) $\frac{1}{N^2} \sum_{i,j=1}^{N} (x_i - x_j)^2 = \frac{1}{N^2} \sum_{i,j} x_i^2 - 2x_i x_j + x_j^2$

$$= \frac{1}{N^2} \left( \sum_{i=1}^{N} N x_i^2 + \sum_{j=1}^{N} N x_j^2 - 2 \sum_{i,j=1}^{N} x_i x_j \right)$$

$$= \frac{1}{N^2} \left( 2N \sum_{i=1}^{N} x_i^2 - 2 \sum_{i=1}^{N} x_i \sum_{j=1}^{N} x_j \right)$$

$$= 2 \left( \frac{1}{N} \sum_{i=1}^{N} x_i^2 - \left( \frac{1}{N} \sum_{i=1}^{N} x_i \right)^2 \right).$$

6.8) If $x \sim \text{Bernoulli}(\mu)$ then $p(x|\mu) = \mu^x (1-\mu)^{1-x}$.

By (6.113d) we have:

$$p(x|\mu) = \exp\left( x \log \frac{\mu}{1-\mu} + \log(1-\mu) \right)$$

which is in the form of (6.107) with natural parameter $\theta = \log \frac{\mu}{1-\mu}$, and sufficient statistic $\phi(x) = x$.

6.9) If $m \sim \text{Binomial}(\mu, N)$ then $p(m|\mu, N) = \binom{N}{m} \mu^m (1-\mu)^{N-m}$.

Assume $N$ = number of trials is fixed. ← otherwise I don't know what to do.

Then $p(m|\mu, N) = \binom{N}{m} \exp\left( \log\left( \mu^m (1-\mu)^{N-m} \right) \right)$

$$= \binom{N}{m} \exp\left( m \log \mu + N \log(1-\mu) - m \log(1-\mu) \right)$$

$$= \binom{N}{m} \exp\left( m \log \frac{\mu}{1-\mu} + N \log(1-\mu) \right).$$

So: natural parameter is $\log \frac{\mu}{1-\mu}$.
sufficient statistic is $m$
log-partition function is $-N \log(1-\mu) = N \log\left(\frac{1}{1-\mu}\right)$.

If $x \sim \text{Beta}(\alpha, \beta)$ then $p(x|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$.

$$p(x|\alpha,\beta) = \exp\left(\log\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}\right)\right)$$

$$= \exp\left(\log\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right) + \alpha\log x - \log x + \beta\log(1-x) - \log(1-x)\right)$$

$$= \frac{1}{x(1-x)} \exp\left(\alpha\log x + \beta\log(1-x) - \log\left(\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}\right)\right)$$

So: natural parameters are $(\alpha, \beta)$

sufficient statistic is $(\log(x), \log(1-x))$

log-partition is $\log\left(\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}\right)$.

---

Binomial prob. distribution is:

$$p(x|\mu,N) = \binom{N}{x} \exp\left(x\log\frac{\mu}{1-\mu} - N\log\left(\frac{1}{1-\mu}\right)\right)$$

Beta prob distribution is:

$$p(x|\alpha,\beta) = \frac{1}{x(1-x)} \exp\left(\alpha\log x + \beta\log(1-x) - \log\left(\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}\right)\right)$$

Their product is:

$$p(x|\mu,N)\,p(x|\alpha,\beta)$$

$$= \frac{\binom{N}{x}}{x(1-x)} \exp\left(\log\left(\frac{\mu}{1-\mu}\right)x + \alpha\log x + \beta\log(1-x) - \left[N\log\left(\frac{1}{1-\mu}\right) + \log\left(\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}\right)\right]\right).$$

6.10 a). $p_1(x) \sim N(x|a, A)$, $p_2(x) \sim N(x|b, B)$.

$$(p_1 p_2)(x) = \underbrace{\frac{1}{(2\pi)^D\sqrt{|A||B|}}}_{\text{call this } \lambda} \exp\left(-\frac{1}{2}\left[(x-a)^T A^{-1}(x-a) + (x-b)^T B^{-1}(x-b)\right]\right)$$

$$= \lambda \exp\left(-\frac{1}{2}\left[x^T A^{-1} x - x^T A^{-1} a - a^T A^{-1} x + a^T A^{-1} a + x^T B^{-1} x - x^T B^{-1} b - b^T B^{-1} x + b^T B^{-1} b\right]\right).$$

$$= \lambda \exp\left(-\frac{1}{2}\left[x^T (A^{-1}+B^{-1}) x - x^T(A^{-1}a + B^{-1}b) - (a^T A^{-1} + b^T B^{-1})x + a^T A^{-1} a + b^T B^{-1} b\right]\right)$$

$$\left(\text{Recall: } \begin{array}{l} C = (A^{-1}+B^{-1})^{-1} \Leftrightarrow C^{-1} = A^{-1} + B^{-1} \\ c = C(A^{-1}a + B^{-1}b) \Leftrightarrow C^{-1}c = (A^{-1}a + B^{-1}b) \end{array}\right)$$

$$= \lambda \exp\left(-\frac{1}{2}\left[x^T C^{-1} x - x^T C^{-1} c \underbrace{-(C^{-1}c)^T x}_{= c^T C^{-1} x} + a^T A^{-1} a + b^T B^{-1} b\right]\right)$$

$$= \lambda \exp\left(-\frac{1}{2}\left[(x-c)^T C^{-1}(x-c) - c^T C^{-1} c + a^T A^{-1} a + b^T B^{-1} b\right]\right)$$

sub back in terms of A, a, B, b

$$= \lambda \exp\left(-\frac{1}{2}\left[(x-c)^T C^{-1}(x-c) - (a^T A^{-1} + b^T B^{-1})(A^{-1}+B^{-1})^{-1}(A^{-1}a + B^{-1}b) + a^T A^{-1} a + b^T B^{-1} b\right]\right)$$

$$= \lambda \exp\left(-\frac{1}{2}(x-c)^T C^{-1}(x-c)\right) \exp\left(-\frac{1}{2}\left[-(a^T A^{-1} + b^T B^{-1})(A^{-1}+B^{-1})^{-1}(A^{-1}a + B^{-1}b) + a^T A^{-1} a + b^T B^{-1} b\right]\right)$$

$$= \lambda (2\pi)^{D/2} |C|^{1/2} \exp\left(-\frac{1}{2}\left[-(a^T A^{-1} + b^T B^{-1})(A^{-1}+B^{-1})^{-1}(A^{-1}a + B^{-1}b) + a^T A^{-1} a + b^T B^{-1} b\right]\right) \cdot N(x|c, C).$$

This is $\frac{1}{(2\pi)^{D/2}|C|^{1/2}} \exp(\ldots)$

It remains to show that this line is $c = N(a|b, A+B)$.

Firstly, $\lambda (2\pi)^{D/2}|C|^{1/2} = \frac{(2\pi)^{D/2}|C|^{1/2}}{(2\pi)^{D}|A|^{1/2}|B|^{1/2}} = \frac{1}{(2\pi)^{D/2}|AC^{-1}B|^{1/2}}$

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

So, just need to show that:

$$-(a^TA^{-1}+b^TB^{-1})(A^{-1}+B^{-1})^{-1}(A^{-1}a+B^{-1}b)+a^TA^{-1}a+b^TB^{-1}b = (a-b)^T(A+B)^{-1}(a-b)$$

We have: $\left(A^{-1}(A^{-1}+B^{-1})^{-1}B^{-1} = (A+B)^{-1}\right.$ is used throughout).
& same with $A, B$ swapped

$$-(a^TA^{-1}+b^TB^{-1})(A^{-1}+B^{-1})^{-1}(A^{-1}a+B^{-1}b) + a^TA^{-1}a + b^TB^{-1}b$$

$$= -a^T(A+B)^{-1}b - b^T(A+B)^{-1}a - a^TA^{-1}(A^{-1}+B^{-1})^{-1}A^{-1}a - b^TB^{-1}(A^{-1}+B^{-1})^{-1}B^{-1}b + a^TA^{-1}a + b^TB^{-1}b$$

$$= -a^T(A+B)^{-1}b - b^T(A+B)^{-1}a + a^T(A^{-1}-A^{-1}(A^{-1}+B^{-1})^{-1}A^{-1})a + b^T(B^{-1}-B^{-1}(A^{-1}+B^{-1})^{-1}B^{-1})b.$$

So our problem reduces to showing that:

i) $A^{-1}-A^{-1}(A^{-1}+B^{-1})^{-1}A^{-1} = (A+B)^{-1}$

ii) $B^{-1}-B^{-1}(A^{-1}+B^{-1})^{-1}B^{-1} = (A+B)^{-1}$.

Firstly, note that i) holds $\Leftrightarrow$ ii) holds, as the statements are symmetric in $A$ & $B$. We show that i) is true.

$$\begin{aligned}
A^{-1}-A^{-1}(A^{-1}+B^{-1})^{-1}A^{-1} &= A^{-1}-A^{-1}(A^{-1}+B^{-1})^{-1}B^{-1}BA^{-1} \\
&= A^{-1}-(A+B)^{-1}BA^{-1} \\
&= (A+B)^{-1}\left[(A+B)A^{-1}-BA^{-1}\right] \\
&= (A+B)^{-1}\left[I+BA^{-1}-BA^{-1}\right] \\
&= (A+B)^{-1}, \text{ as required.}
\end{aligned}$$

b) If $x \sim N(x|\mu, \Sigma)$ then $p(x) = \frac{1}{(2\pi)^{D/2}\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$.

Notation: write $\langle v_1, v_2 \rangle_A := v_1^T A v_2$ for inner product w.r.t $A$.

If $\Sigma$ symmetric (e.g our covariance matrices) then

$$\langle v_1, v_2 \rangle_\Sigma = v_1^T \Sigma v_2 = \langle v_1, \Sigma v_2 \rangle$$
$$= (\Sigma^T v_1)^T v_2 = (\Sigma v_1)^T v_2 = \langle \Sigma v_1, v_2 \rangle$$

Fact: $\langle x, x \rangle_{\Sigma^{-1}} = Tr(\Sigma^{-1} x x^T)$

proof: $\langle x, x \rangle_{\Sigma^{-1}} = x^T \Sigma^{-1} x = Tr(x^T \Sigma^{-1} x) = Tr(\Sigma^{-1} x x^T)$.

$$p(x) = \frac{1}{(2\pi)^{D/2}\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}\langle x, x \rangle_{\Sigma^{-1}} + \langle \mu, x \rangle_{\Sigma^{-1}} - \frac{1}{2}\langle \mu, \mu \rangle_{\Sigma^{-1}}\right).$$

Now let $p_1(x) \sim N(x|a, A)$, $p_2(x) \sim N(x|b, B)$

then
$$(p_1 p_2)(x) = \frac{1}{(2\pi)^D \sqrt{|A||B|}} \exp\Big(-\frac{1}{2} Tr(A^{-1} x x^T) + \langle a, x \rangle_{A^{-1}} - \frac{1}{2}\langle a, a \rangle_{A^{-1}}$$
$$- \frac{1}{2} Tr(B^{-1} x x^T) + \langle b, x \rangle_{B^{-1}} - \frac{1}{2}\langle b, b \rangle_{B^{-1}}\Big)$$

$$\left(\text{Note: } \frac{\sqrt{|C|}}{\sqrt{|A||B|}} = \frac{1}{\sqrt{|A||C^{-1}||B|}} = \frac{1}{\sqrt{|AC^{-1}B|}} = \frac{1}{\sqrt{|A+B|}} \iff \frac{1}{\sqrt{|A||B|}} = \frac{1}{\sqrt{|C|}\sqrt{|A+B|}}.\right)$$

Used $AC^{-1}B = A + B$.

so
$$(p_1 p_2)(x) = \frac{1}{(2\pi)^{D/2}\sqrt{|A+B|}} \cdot \frac{1}{(2\pi)^{D/2}\sqrt{|C|}} \exp\Big(-\frac{1}{2} Tr\left((A^{-1}+B^{-1}) x x^T\right) + \langle A^{-1}a + B^{-1}b, x \rangle$$
$$- \frac{1}{2}\left(\langle A^{-1}a, a \rangle + \langle B^{-1}b, b \rangle\right)\Big)$$

$$= \frac{1}{(2\pi)^{D/2}\sqrt{|A+B|}} \cdot \frac{1}{(2\pi)^{D/2}\sqrt{|C|}} \exp\Big(-\frac{1}{2} Tr\left(C^{-1} x x^T\right) + \langle C(A^{-1}a + B^{-1}b), x \rangle_{C^{-1}}$$
$$- \frac{1}{2}\left(\langle A^{-1}, a \rangle + \langle B^{-1}b, b \rangle\right)\Big).$$

It remains to prove that:

$$\langle A^{-1}a, a\rangle + \langle B^{-1}b, b\rangle = \langle c, c\rangle_{C^{-1}} + \underbrace{(a-b)^T (A+B)^{-1}(a-b)}_{= \langle a-b,\, a-b\rangle_{(A+B)^{-1}}}.$$

We have:

$$\langle A^{-1}a, a\rangle + \langle B^{-1}b, b\rangle - \langle c, c\rangle_{C^{-1}}$$

$$= \langle A^{-1}a, a\rangle + \langle B^{-1}b, b\rangle - \left[\langle C(A^{-1}a + B^{-1}b),\ CA^{-1}a + CB^{-1}b\rangle_{C^{-1}}\right]$$

$$= \langle A^{-1}a, a\rangle + \langle B^{-1}b, b\rangle - \langle A^{-1}a + B^{-1}b,\ CA^{-1}a + CB^{-1}b\rangle$$

$$= \langle A^{-1}a, a\rangle + \langle B^{-1}b, b\rangle - \left[\langle A^{-1}a, CA^{-1}a\rangle + 2\langle A^{-1}a, CB^{-1}b\rangle + \langle B^{-1}b, CB^{-1}b\rangle\right]$$

$$= \langle a, A^{-1}a\rangle + \langle b, B^{-1}b\rangle - \left[\langle a, A^{-1}CA^{-1}a\rangle + 2\langle a, (A+B)^{-1}b\rangle + \langle b, B^{-1}CB^{-1}b\rangle\right]$$

$$= \langle a, (A^{-1} - A^{-1}CA^{-1})a\rangle + \langle b, (B^{-1} - B^{-1}CB^{-1})b\rangle - 2\langle a, (A+B)^{-1}b\rangle$$

$$= \langle a, (A+B)^{-1}a\rangle + \langle b, (A+B)^{-1}b\rangle - 2\langle a, (A+B)^{-1}b\rangle$$

$$= \langle (a-b), (A+B)^{-1}(a-b)\rangle = \langle a-b, a-b\rangle_{(A+B)^{-1}}, \text{ as required}$$

Note: In the second-to-last equality we used

i) $A^{-1} - A^{-1}CA^{-1} = (A+B)^{-1}$

ii) $B^{-1} - B^{-1}CB^{-1} = (A+B)^{-1}$

which were proven as part of part 6.10a).

6.11) $\mathbb{E}_x[x] = \int x\, p(x)\, dx$

$$= \int x \left( \int p(x|y) p(y)\, dy \right) dx$$

$$= \int \left( \int x\, p(x|y)\, dx \right) p(y)\, dy$$

$$= \int \mathbb{E}_x[x|y]\, p(y)\, dy$$

$$= \mathbb{E}_y\left[\mathbb{E}_x[x|y]\right].$$

6.12a) $p(y|x) = p\left(Ax+b+w \,|\, x\right)$ whee $w \sim \mathcal{N}(w|0, Q)$

$$\sim \mathcal{N}(Ax+b, Q).$$

b) $y = Ax+b+w$ , $w \sim \mathcal{N}(w|0,Q)$. , $x \sim \mathcal{N}(x|\mu_x, \Sigma_x)$.

As $w \sim \mathcal{N}(w|0,Q)$, $w+b \sim \mathcal{N}(w|b, Q)$

(i.e addition of b shifts the mean but leaves covariance alone.)

Since $x \sim \mathcal{N}(x|\mu_x, \Sigma_x)$, $Ax \sim \mathcal{N}(A\mu_x, A\Sigma_x A^T)$

Finally, $Ax$ and $b+w$ are independent. Therefore their sum is Gaussian distributed (cf. Eq 6.78) and we have

$$p(y) = p\left(Ax + (b+w)\right) \sim \mathcal{N}\left(A\mu_x + b, A\Sigma_x A^T + Q\right).$$

c) $z = Cy + v, \quad y \sim N(A\mu_x + b, A\Sigma_x A^T + Q)$

$v \sim N(0, R)$

$p(z|y) = p(Cy + v|y) \sim N(Cy, R)$ (y is no longer random, the only random part is v).

$p(z)$ is Gaussian, as it is the sum of two Gaussians which are independent.

$p(Cy) \sim N(CA\mu_x + Cb, CA\Sigma_x A^T C^T + CQC^T)$

$p(v) \sim N(0, R)$

$\Rightarrow p(z) \sim N(CA\mu_x + Cb, CA\Sigma_x (CA)^T + CQC^T + R)$.

d) We have:

$$p(x,y) = p\left(\begin{pmatrix} x \\ y \end{pmatrix}\right) = p\left(\begin{pmatrix} x \\ Axb+w \end{pmatrix}\right)$$

$$= p\left(\begin{pmatrix} 1 & 0 \\ A & 1 \end{pmatrix}\begin{pmatrix} x \\ w \end{pmatrix} + \begin{pmatrix} 0 \\ b \end{pmatrix}\right)$$

Here, 1 & 0 denote appropiately sized identity & zero matrices.

As x and w are independent, we know the joint distribution immediately (Eq 6.64):

$$p\left(\begin{pmatrix} x \\ w \end{pmatrix}\right) \sim N\left(\begin{pmatrix} \mu_x \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_{xx} & 0 \\ 0 & R \end{pmatrix}\right)$$

Now apply affine transformation $\begin{pmatrix} 1 & 0 \\ A & 1 \end{pmatrix}\begin{pmatrix} x \\ w \end{pmatrix} + \begin{pmatrix} 0 \\ b \end{pmatrix}$ to obtain distribution of $\begin{pmatrix} x \\ y \end{pmatrix}$.

$$p\left(\begin{pmatrix} x \\ y \end{pmatrix}\right) \sim N\left(\begin{pmatrix} 1 & 0 \\ A & 1 \end{pmatrix}\begin{pmatrix} \mu_x \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ b \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ A & 1 \end{pmatrix}\begin{pmatrix} \Sigma_{xx} & 0 \\ 0 & R \end{pmatrix}\begin{pmatrix} 1 & A^T \\ 0 & 1 \end{pmatrix}\right)$$

$$= N\left(\begin{pmatrix} \mu_x \\ A\mu_x \end{pmatrix} + \begin{pmatrix} 0 \\ b \end{pmatrix}, \begin{pmatrix} \Sigma_{xx} & 0 \\ A\Sigma_{xx} & R \end{pmatrix}\begin{pmatrix} 1 & A^T \\ 0 & 1 \end{pmatrix}\right)$$

$$= N\left(\begin{pmatrix} \mu_x \\ A\mu_x + b \end{pmatrix}, \begin{pmatrix} \Sigma_{xx} & \Sigma_{xx}A^T \\ A\Sigma_{xx} & A\Sigma_{xx}A^T + R \end{pmatrix}\right)$$

Following Eq (6.65) – (6.67) we have

$$p(x|\hat{y}) \sim N(\mu_{x|\hat{y}}, \Sigma_{x|\hat{y}}) \quad \text{with}$$

$$\mu_{x|\hat{y}} = \mu_x + (\Sigma_{xx}A^T)(A\Sigma_{xx}A^T + R)^{-1}(\hat{y} - A\mu_x - b)$$

$$\Sigma_{x|\hat{y}} = \Sigma_{xx} - (\Sigma_{xx}A^T)(A\Sigma_{xx}A^T + R)^{-1}A\Sigma_{xx}$$

6.13) $X$ has cdf $F_X(x)$. Let $Y := F_X(X)$

i.e., if $X : \Omega \to \mathcal{T}$, $\omega \mapsto t(\omega)$, then $Y : \Omega \to \mathcal{T} \to [0,1]$, $\omega \mapsto t(\omega) \to F_X(t(\omega)) = \mathbb{P}(X \leq t(\omega))$.

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(X \leq F_X^{-1}(y))$$

(Note: State space of $Y$ is $[0,1]$. So (little) $y \in [0,1]$).

so $F_Y(y) = \int_{-\infty}^{F_X^{-1}(y)} f_X(x)\, dx$, let $u = F_X(x) \Rightarrow du = f_X(x)\, dx$

limit $x = -\infty \to u = 0$, $x = F_X^{-1}(y) \to u = y$

$$= \int_0^y du.$$

Differentiating gives $f_Y(y) = \frac{d}{dy}\left(\int_0^y du\right) = 1$. So $Y \sim \text{Uniform}(0,1)$.