1 Introduction

Problem 1.1

Given that:

$$\sum_{j=0}^{M} A_{ij} w_j = T_i 1 \tag{1}$$

Where:

$$A_{ij} = \sum_{n=1}^{N} (x_n)^{i+j} 2$$

$$T_i = \sum_{n=1}^{N} (x_n)^i t_n 3$$

$$y(x, w) = w_0 + w_1 x + w_2 x^2 + ... + w_M x^M = \sum_{j=0}^{M} w_j x^j 4$$

$$E(w) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, w) - t_n\}^2 5$$

Show that the coefficients $\mathbf{w} = \{w_i\}$ that minimizes (1) error equation. To solve this problem, we will derivative the sum-of-squares error function (5).

$$\begin{aligned}
&\frac{\partial E}{\partial w_i} = \frac{1}{2} \sum_{i=0}^{N} 2\{y(x_n, w) - t_n\} x_i = 0 \\
\iff &\sum_{i=0}^{N} \{y(x_n, w) x^i - t_n x^i\} = 0 \\
\iff &\sum_{i=0}^{N} \{y(x_n, w) x^i\} = \sum_{i=0}^{N} t_n x^i \\
\iff &\sum_{i=0}^{N} \sum_{j=0}^{M} w_j x_i^j x^i = \sum_{i=0}^{N} t_n x_i \\
\iff &\sum_{j=0}^{M} \sum_{i=0}^{N} w_j x_i^{i+j} = \sum_{i=0}^{N} x_n^i t_n
\end{aligned}$$

If we denote the above equation by (2) and (3), we get the answer. The problem is solved.

Problem 1.2

Given the regularized sum-of-squares equation:

$$\tilde{E}(w) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, w) - t_n\}^2 + \frac{\lambda}{2} \|w\|^2 \tag{2}$$

Write down the set of coupled linear equations, analogous to equation from previous excercise, satisfied by the coefficients $\{w_i\}$. To solve this problem, we derivative the given equation like ex 1.1:

$$\frac{\partial \tilde{E}}{\partial w_i} = \sum_{n=1}^{N} \{y(x_n, w) - t_n\} x_n^i + \lambda \|w\| = 0$$
$$\sum_{n=1}^{N} \sum_{n=1}^{M} (w_j x_n^j x_n^i - t_n x_n^i) + \lambda \|w\| = 0$$
$$\sum_{j=1}^{M} (A_{ij} - T_i) + \lambda \|w\| = 0$$

*Note: Incorrect, the theta variable called kronecker delta and the derivative of equation is incorrect

Problem 1.3

Based on Bayes's Theorem:

$$P(X) = \sum_{Y} P(X|Y)P(Y)$$

The probability of selecting an apple P(a) from all boxes:

$$P(a) = P(a|r) * P(r) + P(a|b) * P(b) + P(a|g) * P(g) = 0.3 * 0.2 + 0.5 * 0.2 + 0.3 * 0.6$$
$$= 0.34$$

The probabilty that orange came from green box P(g|o) and based on Bayes's Theorem, we have:

$$P(g|o) = \frac{P(o|g) * P(g)}{P(o)}$$

We will calculate P(o) like P(a):

$$P(o) = P(o|r) * P(r) + P(o|b) * P(b) + P(o|g) * P(g) = 0.4 * 0.2 + 0.5 * 0.2 + 0.3 * 0.6$$
$$= 0.36$$
$$P(g|o) = \frac{0.3 * 0.6}{0.36} = 0.5$$

Problem 1.4

Equation (1.27) from the change of varible theorem:

$$p_y(y) = p_x(x)|\frac{dx}{dy}|$$
$$= p_x(g(y))\|g'(y)\|$$

We will calculate the derivative of equation (1.27) with respect to y, we got:

$$\frac{p_y(y)}{dy} = \frac{dp_x(x)|g'(y)|}{dy} = \frac{dp_x(x)}{dy}|g'(y)| + p_x(x)\frac{g'(y)}{dy} \tag{3}$$

Applying Chain Rule into first term in equation (1):

$$\frac{dp_x(x)}{dy}|g'(y)| = \frac{p_x(g(y))}{dg(y)}\frac{dg(y)}{dy}|g'(y)| \tag{4}$$

if $\hat{x}$ is maximum of the density over x, we obtain:

$$\frac{dp_x(x)}{dx}|\hat{x} = 0$$

Therefore, the equation (2) will equal 0, leading to the first term of equation (1) equal 0. In case of linear traIn nsformation, the second term of equation (1) will be vanish (ex: y = ax + b) so equation(1) equal 0. In conclusion, in case of linear transformation, the location of the maximum transforms in the same way as the variable itself.

Problem 1.5

$$E[E[f(x)]] = E[f(x)] \qquad \text{cause E[f(x)] is constant}$$

$$\begin{aligned} var[f(x)] &= E[(f(x) - E[f(x)])^2] \\ &= E[f(x)^2 - 2f(x)E[f(x)] + E[f(x)]^2] \\ &= E[f(x)^2] - 2E[f(x)]^2 + E[f(x)]^2 \\ &= E[f(x)^2] - E[f(x)]^2 \end{aligned}$$

Problem 1.6

If x and y are independent, we obtain:

$$\begin{aligned} p_{xy}(x,y) &= p_x(x)p_y(y) \\ \int_{xy} p_{xy}(x,y) &= \int_x \int_y xyp_x(x)p_y(y) \\ &= \int_x xp_x(x) \int_y yp_y(y) \\ &= E[x]E[y] \end{aligned}$$

The result is leading to make the covariance of x and y equal 0. The problem is solved.

Problem 1.7

$$I^2 = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \exp(\frac{-1}{2\sigma^2}x^2 - \frac{1}{2\sigma^2}y^2)dx$$

Transform Cartesian coordination (x,y) $\rightarrow$ polar coordination (r,θ):

$$x = rcos(\theta), y = rsin(\theta)$$

$$0 \leq r \leq +\infty, -\infty \leq \theta \leq 2\pi$$

$$I^2 = \int_{-\infty}^{2\pi}\int_{0}^{+\infty} \exp(\frac{-1}{2\sigma^2}(rcos\theta)^2 - \frac{1}{2\sigma^2}(rsin\theta))rdrd\theta$$

$$I^2 = \int_{-\infty}^{2\pi}\int_{0}^{+\infty} \exp(\frac{1}{2\theta^2}r^2)rdrd\theta$$

Subtituting u = r^2 and calculate integral over u firstly:

$$\int_0^{+\infty} \frac{1}{2} \exp(\frac{-1}{2\sigma^2}u)du$$
$$= -\sigma^2 \exp(\frac{-1}{2\sigma^2}u)_{|_0^{+\infty}}$$
$$= -\sigma^2(0-1) = \sigma^2$$

Integral over θ:

$$I^2 = \int_{-\infty}^{2\pi} \sigma^2 d\theta = \sigma^2\theta_{|_{-\infty}^{2\pi}}$$
$$I = \sqrt{2\pi\theta^2}$$

A probability distribution is called "Normalized" if the sum of all possible result equal one.

$$\int_{-\infty}^{+\infty} \mathcal{N}(x|\mu,\sigma)$$
$$= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(\frac{-1}{2\sigma^2}(x-\mu)^2)dx$$
$$= \frac{\sqrt{2\pi\sigma^2}}{\sqrt{2\pi\sigma^2}} = 1 \qquad \text{(Solved)}$$

Problem 1.8

$$E[x] = \int_{-\infty}^{+\infty} \mathcal{N}(x|\mu,\sigma^2)xdx$$
$$= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(\frac{-1}{2\sigma^2}(x-\mu)^2)xdx$$

Subtituting y = x - μ:

$$E[x] = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(\frac{-1}{2\sigma^2}y^2)(y+\mu)dx$$
$$= \frac{1}{\sqrt{2\pi\sigma^2}}(\int_{-\infty}^{+\infty} \exp(\frac{-1}{2\sigma^2}y^2)ydx + \int_{-\infty}^{+\infty} \exp(\frac{-1}{2\sigma^2}y^2)\mu dx)$$

Get the result from problem (1.7), we obtain:

$$E[x] = 0 + \mu = \mu$$

So the univariate Gauss Distribution given from (1.46) satisfies (1.49). Then we differentiate the equation (1.127) by using product rule.

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(\frac{(x-\mu)^2}{2\sigma^2})$$
$$\frac{d\mathcal{N}(x|\mu,\sigma^2)}{d\sigma^2} = (\frac{1}{\sqrt{2\pi\sigma^2}})' \exp(\frac{(x-\mu^2)}{2\sigma^2}) + \frac{1}{\sqrt{2\pi\sigma^2}} \exp(\frac{(x-\mu)^2}{2\sigma^2})'$$

The derivative of First term:

$$\frac{-1}{2\sqrt{2\pi}\sigma^3}\exp(\frac{(x-\mu^2)}{2\sigma^2})$$

The derivative of Second term:

$$\frac{(x-\mu)^2}{2\sqrt{2\pi}\sigma^5}\exp(\frac{(x-\mu^2)}{2\sigma^2})$$

Simplify the equation, we will obtain:

$$\int_{-\infty}^{+\infty}(\frac{-1}{2\sigma^2}+\frac{(x-\mu)^2}{2\sigma^4})\mathcal{N}(x|\mu,\sigma^2)=0$$
$$\int_{-\infty}^{+\infty}(x-\mu)^2\mathcal{N}(x|\mu,\sigma^2)=\sigma^2$$

Based on definition of variance, we proved:

$$var[x]=\sigma^2$$

Also we have:

$$E[x^2]=var[x]+E[x]^2=\sigma^2+\mu^2$$

Problem 1.9

The formula (1.52) is the general form of (1.46) so we will only prove this formula:

$$\frac{\partial(\frac{1}{2\pi^{\frac{D}{2}}}\frac{1}{|\Sigma|^{\frac{1}{2}}}\exp\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))}{\partial x}=0$$
$$\frac{1}{2}(\Sigma^{-1}+(\Sigma^{-1})^T)(x-\mu)\mathcal{N}(x|\mu,\Sigma)=0$$
$$\Sigma^{-1}(x-\mu)\mathcal{N}(x|\mu,\Sigma)=0$$
$$x=\mu$$

Using these formulas to get above result:

$$x^TA=A^Tx$$
$$\frac{\partial x^TAx}{\partial x}=(A+A^T)x$$

The Second Derivative of formula (1.52):

$$\Sigma^{-1}\mathcal{N}(x|\mu,\Sigma)+\Sigma^{-1}(x-\mu)\frac{1}{2}(\Sigma^{-1}+(\Sigma^{-1})^T)(x-\mu)\mathcal{N}(x|\mu,\Sigma)$$
$$=\Sigma^{-1}\mathcal{N}(x|\mu,\Sigma)(1+\Sigma^{-1}(x-\mu)^2)$$

As we can see, Σ is positive semi-definite, the distribution is always positive so the the above equation is always positive. Applying it to Hessian matrix, all the eigenvalues of matrix are positive. It mean the point $x=\mu$ is the maximum point.

Problem 1.10

Proof of First Term:

$$
\begin{aligned}
E[x+z] &= E[x] + E[z] \\
\int p(x)dx &= 1 \\
p(x,z) &= p(x)p(z) \qquad \text{(cause x,z are independent)} \\
E[x+z] &= \int\int (x+z)p(x,z)dxdz \\
&= \int\int xp(x)p(z)dxdz + \int\int zp(x)p(z)dxdz \\
&= E[x] + E[z]
\end{aligned}
$$

Proof of Second Term:

$$
\begin{aligned}
var[x+z] &= E[(x+z)^2] - E[x+z]^2 \\
&= \int\int (x+z)^2 p(x,z) - (\int\int (x+z)p(x,z))^2 \\
&= E[x^2] + E[z^2] + \int\int 2xzp(x)p(z)dxdz - (E[x]^2 + E[z]^2 + 2\int\int 2xzp(x)p(z)dxdz) \\
&= E[x^2] - E[x]^2 + E[z^2] - E[z]^2 \\
&= var[x] + var[z]
\end{aligned}
$$

Problem 1.11

$$
\ln \prod_{n=1}^{N} f(x) = \sum_{n=1}^{N} ln(f(x))
$$

From above equation, we obtain:

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

Setting the derivative of the log likelihood function with respect to μ:

$$
\begin{aligned}
\frac{dln(p(x|\mu,\sigma^2))}{d\mu} &= \frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n-\mu) = 0 \\
\sum_{n=1}^{N} x_n - N\mu &= 0 \\
\mu_{ML} &= \frac{1}{N}\sum_{n=1}^{N} x_n
\end{aligned}
$$

Setting the derivative of the log likelihood function with respect to σ^2:

$$\frac{dp(x|\mu,\sigma^2)}{d\sigma^2} = \frac{1}{2\sigma^4}\sum_{n=1} N(x_n-\mu)^2 - \frac{N}{2\sigma^2} = 0$$

$$\sigma^2_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{ML})^2$$

Problem 1.12

If n $\neq$ m then x_n and x_m are independent.Hence E[$x_n x_m$] = E[x_n][x_m] = $\mu^2 + \sigma^2$. In case of n = m, E[$x_n x_m$] = E[x] = μ^2. Combine those results, we obtain the equation (1.130).

Easy to prove the equation (1.49):

$$E[\mu_{ML}] = E[\frac{1}{N}\sum_{n=1}^{N} x_n] = \frac{1}{N}\sum_{n=1}^{N} E[x] = \frac{1}{N}N\mu = \mu$$

With equation (1.50):

$$E[\sigma^2_{ML}] = E[\frac{1}{N}\sum_{n=1}^{N}(x-\mu_{ML})^2] = \frac{1}{N}\sum_{n=1}^{N} E[(x^2 - 2x\mu_{ML} + \mu^2_{ML}]$$

$$= \frac{1}{N}\sum_{n=1}^{N}(E[x^2] - 2E[x\mu_{ML}] + E[\mu^2_{ML}])$$

$$= \mu^2 + \sigma^2 - \frac{2}{N^2}E[\sum_{n=1}^{N}\sum_{m=1}^{M} x^2_{nm}] + \frac{1}{N^2}E[(\sum_{p=1}^{P} x_p)^2]$$

Because x = m = p, hence:

$$E[\mu_{ML}] = \mu^2 + \sigma^2 - \frac{1}{N^2}E[(\sum_{n=1}^{N} x_n)^2]$$

Based on equation (1.130), we will obtain

$$: E[\mu_{ML}] = \mu^2 + \sigma^2 - \mu^2 - \frac{1}{N}\sigma^2$$

$$= \frac{N-1}{N}\sigma^2$$

Problem 1.13

We will solve similarly like problem (1.12):

$$E[\sigma^2_{ML}] = E[\frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^2]$$
$$= \frac{1}{N}\sum_{n=1}^{N}(E[x^2] - 2E[x\mu] + E[\mu^2])$$
$$= \frac{1}{N}\sum_{n=1}^{N}(\mu^2 + \sigma^2 - 2\mu^2 + \mu^2)$$
$$= \sigma^2$$

In case of Maximum Likelihood Mean:

$$E[\mu^2_{ML}] = E[(\frac{1}{N}\sum_{n=1}^{N} x_n)^2]$$
$$= \frac{1}{N^2}(N(N\mu^2 + \sigma^2)) = \mu^2 + \frac{\sigma^2}{N}$$

Problem 1.17

The derivative of $\Gamma(x+1)$:

$$\Gamma(x+1) = \int_0^\infty u^x e^{-u} du$$
$$= -u^x e^{-u}|_0^\infty + x\Gamma(x)$$

Through L'hopital rule, we obtain:

$$\lim_{u->\infty} \frac{u^x}{e^u} = \lim_{u->\infty} \frac{x!}{e^\infty} = 0$$

So we have proved $\Gamma(x+1) = x\Gamma(x)$.

$$\Gamma(1) = \int_0^\infty e^{-u} du = e^{-u}|_0^\infty = (1-0) = 1$$

Based on two above proved equation, we will obtain:

$$\Gamma(x+1) = x\Gamma(x) = x!\Gamma(1) = x!$$

Problem 1.35

The Differential Entropy of Gaussian Distribution:

$$\begin{aligned} H(x) &= -\int p(x) lnp(x) dx \\ &= -\int p(x) ln\{\frac{1}{(2\pi\sigma^2)^{1/2}} \exp -\frac{(x-\mu)^2}{2\sigma^2}\} dx \\ &= -\int p(x) ln\{\frac{1}{(2\pi\sigma^2)^{1/2}}\} dx + \int p(x)(\frac{(x-\mu)^2}{2\sigma^2}) dx \\ &= \frac{1}{2} ln(2\pi\sigma^2) + \frac{1}{2} = \frac{1}{2}(1 + ln(2\pi\sigma^2)) \end{aligned}$$

Problem 1.37

We already have:

$$\begin{aligned} H[x] &= -\int p(x) lnp(x) dx \\ H[y|x] &= -\int\int p(x,y) lnp(y|x) dxdy \end{aligned}$$

It's straightforward by adding two above equations:

$$\begin{aligned} H[x] + H[y|x] &= -(\int p(x) lnp(x) dx + \int\int p(x,y) lnp(y|x) dxdy) \\ &= -(\int\int p(x,y) lnp(x) dxdy + \int\int p(x,y) lnp(y|x) dxdy) \\ &= -(\int\int p(x,y) ln(p(x)p(y|x)) dxdy) \\ &= -\int\int p(x,y) lnp(x,y) dxdy = H[x,y] \end{aligned}$$ (Solved)

2 Probability Distributions

Problem 2.1

Bernoulli Distribution:

$$\begin{aligned} Bern(x|\mu) = \mu^x (1-\mu)^{1-x} \sum_{x=0,1} Bern(x|\mu) &= \sum_{x=0,1} \mu^x (1-\mu)^{1-x} = (1-\mu) + \mu = 1 \\ E[x] = \sum_{x=0,1} Bern(x|\mu) x &= \sum x = 0, 1 \mu^x (1-\mu)^{1-x} x = 0 + \mu = \mu \\ var[x] = \sum_{x=0,1} (E[x^2] - E[x]^2) &= \sum_{x=0,1} (\sum_{x=0,1} \mu^x (1-\mu)^(1-x) x^2) + \mu^2 \\ &= \mu + \mu^2 = \mu(1+\mu) \end{aligned}$$

The entropy H(x) of Bernoulli Distribution:

$$H(x) = -\sum_{x=0,1} Bern(x|mu) \ln Bern(x|mu) = -\sum_{x=0,1} \mu^x (1-\mu)^{1-x} ln(\mu^x (1-\mu)^{1-x})$$
$$= -(1-\mu)\ln 1 - \mu - \mu \ln \mu$$

Problem 2.2

This problem is the same as problem (2.1), so we will solve similarly:

$$\sum_{x=-1,1} p(x|\mu) = \sum_{x=-1,1} (\frac{1-\mu}{2})^{(1-x)/2} (\frac{1+\mu}{2})^{(1+x)/2}$$
$$= \frac{1-\mu}{2} + \frac{1+\mu}{2} = 1$$

So the distribution (2.261) is normalized

$$E[x] = \sum_{x=-1,1} (\frac{1-\mu}{2})^{(1-x)/2} (\frac{1+\mu}{2})^{(1+x)/2} x$$
$$= -\frac{1-\mu}{2} + \frac{1+\mu}{2} = \frac{2\mu}{2} = \mu var[x] = E[x^2] - E[x]^2 = \sum_{x=-1,1} (\frac{1-\mu}{2})^{(1-x)/2} (\frac{1+\mu}{2})^{(1+x)/2} x^2 - \mu^2$$
$$= \frac{1-\mu}{2} + \frac{1+\mu}{2} - \mu^2 = \mu - \mu^2 = \mu(1-\mu)$$
$$H[x] = -\sum_{x=-1,1} p(x|\mu) ln(p(x|\mu)) = \sum_{x=-1,1} (\frac{1-\mu}{2})^{(1-x)/2} (\frac{1+\mu}{2})^{(1+x)/2} \ln (\frac{1-\mu}{2})^{(1-x)/2} (\frac{1+\mu}{2})^{(1+x)/2}$$
$$= -(\frac{1-\mu}{2}) \ln \frac{1-\mu}{2} - (\frac{1+\mu}{2}) \ln \frac{1+\mu}{2}$$

Problem 2.3

Transforming the equation (2.262) based on the definition (2.10) of the number of combinations, we will obtain:

$$\frac{N!}{(N-m)!m!} + \frac{N!}{(N-m+1)!(m-1)!} = \frac{(N+1)!}{(N-m+1)!m!}$$

So we will solve the problem with above equation:

$$\frac{N!}{(N-m)!m!} + \frac{N!}{(N-m+1)!(m-1)!} = \frac{N!(N-m+1)}{(N-m+1)!m!} + \frac{N!m}{(N-m+1)!m!}$$
$$= \frac{N!(N+1)}{(N-m+1)!m!} = \frac{(N+1)!}{(N-m+1)!m!} \qquad (Solved)$$

We will prove equation (2.263) by induction and easy to see that the equation is true if N = 1. Assuming the equation (2.263) holds true on N then we will prove that equation also holds:

$$
\begin{aligned}
(1+x)^{N+1} = (1+x)\sum_{m=0}^{N} C_N^m x^m &= x\sum_{m=0}^{N} C_N^m x^m + \sum_{m=0}^{N} C_N^m x^m \\
= \sum_{m=0}^{N} C_N^m x^{m+1} + \sum_{m=0}^{N} C_N^m x^m &= \sum_{m=1}^{N+1} C_N^{m-1} x^{m+1} + \sum_{m=0}^{N} C_N^m x^m \\
&= \sum_{m=1}^{N} (C_N^m + C_N^{m-1}) x^m + x^{N+1} + x^0 \\
&= \sum_{m=1}^{N} C_{N+1}^m x^m + x^{N+1} + x^0 \\
&= \sum_{m=0}^{N+1} C_{N+1}^m x^m \qquad (Solved)
\end{aligned}
$$

Then we will show that the equation (2.264) holds true by induction like above. Subtituting $y = (1-x)$, we obtain:

$$
\begin{aligned}
(x+y)^{N+1} &= (x+y)\sum_{m=0}^{N} C_N^m x^m y^{N-m} \\
&= x\sum_{m=0}^{N} C_N^m x^m y^{N-m} + y\sum_{m=0}^{N} C_N^m x^m y^{N-m} \\
&= \sum_{m=0}^{N} C_N^m x^{m+1} y^{N-m} + y\sum_{m=0}^{N} C_N^m x^m y^{N-m+1} \\
&= \sum_{m=1}^{N+1} C_{N+1}^{m-1} x^m y^{N-m} + \sum_{m=0}^{N} C_N^m x^m y^{N-m+1} \\
&= \sum_{m=1}^{N} (C_{N+1}^{m-1} + C_N^m) x^m y^{N-m} + x^{N+1} + y^{N+1} \\
&= \sum_{m=1}^{N} C_{N+1}^m x^m y^{N-m} + x^{N+1} + y^{N+1} \\
&= \sum_{m=0}^{N+1} C_{N+1}^m x^{m+1} y^{N-m+1}
\end{aligned}
$$

We have solved the equation by inducntion then subtitute again $y = (1-x)$, we will prove the binomial distribution is normalized.

$$
\sum_{m=0}^{N} C_N^m x^m y^{N-m} = (x+1-x)^N = 1^N = 1
$$

Problem 2.5

let t = y + x and x = $t\mu$, we obtain:

$$x = t\mu \quad y = t(1-\mu) \quad t = x + y \quad \mu = \frac{x}{x+y}$$

Using Jacobian Determinant for multiple integral:

$$\frac{\partial(x,y)}{\partial(\mu,t)} = \begin{bmatrix} \frac{\partial x}{\partial \mu} & \frac{\partial x}{\partial t} \\ \frac{\partial y}{\partial \mu} & \frac{\partial y}{\partial t} \end{bmatrix} = \begin{bmatrix} t & \mu \\ -t & 1-\mu \end{bmatrix} = t$$

$$\begin{aligned}
\Gamma(a)\Gamma(b) &= \int_0^\infty \exp(-x)x^{a-1}dx \int_0^\infty \exp(-y)y^{b-1}dy \\
&= \int_0^\infty \int_0^\infty \exp(-x-y)x^{a-1}y^{b-1}dxdy \\
&= \int_0^\infty \int_0^1 \exp(-t)(t\mu)^{a-1}(t(1-\mu))^{b-1}tdtd\mu \\
&= \int_0^\infty \exp(-t)t^{a+b-1}dt \int_0^1 \mu^{a-1}(1-\mu)^{b-1}d\mu \\
&= \Gamma(a+b)\int_0^1 \mu^{a-1}(1-\mu)^{b-1}d\mu \\
&\to \int_0^1 \mu^{a-1}(1-\mu)^{b-1}d\mu = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}
\end{aligned}$$

Problem 2.6

$$\begin{aligned}
E[\mu] &= \int_0^1 \mu^{a-1}(1-\mu)^{b-1}\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}\mu d\mu \\
&= \int_0^1 \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^a(1-\mu)^{b-1}d\mu
\end{aligned}$$

We have equation $\Gamma(x+1) = x\Gamma(x)$ that have proved in problem (1.17). Hence we obtain:

$$\begin{aligned}
&= \frac{\Gamma(a+b+1)a}{\Gamma(a)\Gamma(b)(a+b)}\int_0^1 \mu^a(1-\mu)^{b-1}d\mu \\
&= \frac{a}{a+b}\frac{\Gamma(a+b+1)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} \\
&= \frac{a}{a+b}
\end{aligned}$$

In case of variance of beta distribution, we will solve it similarly like $E[\mu]$:

$$\begin{aligned} var[\mu] = E[x^2] - E[x]^2 &= \int_0^1 \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a+1}(1-\mu)^{b-1} d\mu - (\frac{a}{a+b})^2 \\ &= \frac{(a+1)a}{(a+b)(a+b+1)} \frac{\Gamma(a+b+2)}{\Gamma(a+2)(\Gamma(b))} \int_0^1 \mu^{a+1} 1 - \mu^{b-1} - (\frac{a}{a+b})^2 \\ &= \frac{(a+1)a}{(a+b)(a+b+1)} - (\frac{a}{a+b}^2) = \frac{ab}{(a+b)^2(a+b+1)} \end{aligned}$$

Differentating the beta distribution and give it equal zero:

$$\begin{aligned} \frac{\partial Beta(\mu|a,b)}{\partial \mu} = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}((a-1)\mu^{a-2}(1-\mu)^{b-1} - (b-1)\mu^{a-1}(1-\mu)^{b-2}) = 0 \\ \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}(\mu^{a-2}(1-\mu)^{b-2}((1-\mu)(a-1) - (b-1)\mu)) = 0 \\ (1-\mu)(a-1) - (b-1)\mu = 0 \\ \mu = \frac{a-1}{a+b-2} \end{aligned}$$

So the mode of the beta distribution is $\mu = \frac{a-1}{a+b-2}$ when $a > 1$ and $b > 1$.

Problem 2.8

Proof of equation (2.270):

$$\begin{aligned} E_y[E_x[x|y]] &= E_y[\int xp(x|y)dx] \\ &= \int\int xp(y)p(x|y)dxdy \\ &= \int\int xp(x,y)dxdy \\ &= \int xp(x)dx = E[x] \qquad \text{(Solved)} \end{aligned}$$

Proof of equation (2.271):

$$var_x[x] = E_x[x^2] - E_x[x]^2$$

Separating the above equation into two terms. The First Term:

$$\begin{aligned} E_x[x^2] &= E_y[E_x[x^2|y]] \\ &= E_y[var_x[x|y] + E_x[x|y]^2] \\ &= E_y[var_x[x|y]] + E_y[E_x[x|y]^2] \\ &= E_y[var_x[x|y]] + var_y[E_x[x|y]] + E_y[E_x[x|y]]^2 \end{aligned}$$

The Second Term:

$E_x[x]^2 = E_y[E_x[x|y]]^2$ (Based on equation 2.270)

Combining two terms, we will obtain:

$$\begin{aligned} var[x] &= E_y[var_x[x|y]] + var_y[E_x[x|y]] + E_y[E_x[x|y]]^2 - E_y[E_x[x|y]]^2 \\ &= E_y[var_x[x|y]] + var_y[E_x[x|y]] \end{aligned}$$ (Solved)

Problem 2.10

Based on Dirichlet Distribution given by 2.38, we have:

$$\begin{aligned} E[\mu_j] &= \int Dir(\mu|\alpha)\mu_j d\mu \\ &= \int \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_{k-1}} \mu_j du \\ &= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \int \prod_{k=1}^{K} \mu_k^{\alpha_{k-1}} \mu_j du \\ &= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \frac{\Gamma(\alpha_1)\cdots\Gamma(\alpha_j+1)\cdots\Gamma(\alpha_K)}{\Gamma(\alpha_0+1)} \\ &= \frac{\Gamma(\alpha_0)\Gamma(\alpha_j+1)}{\Gamma(\alpha_0+1)\Gamma(\alpha_j)} \end{aligned}$$

We also have $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$. Hence:

$$E[\mu_j] = \frac{\alpha_j}{\alpha_0}$$

We will solve as above with $E[\mu_j^2]$:

$$\begin{aligned} E[\mu_j^2] &= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \frac{\Gamma(\alpha_1)\cdots\Gamma(\alpha_j+2)\cdots\Gamma(\alpha_K)}{\Gamma(\alpha_0+2)} \\ &= \frac{\Gamma(\alpha_0)\Gamma(\alpha_j+2)}{\Gamma(\alpha_0+2)\Gamma(\alpha_j)} \\ &= \frac{\alpha_j(\alpha_j+1)}{\alpha_0(\alpha_0+1)} \end{aligned}$$

Bring the above result into equation $var[\mu_j]$:

$$\begin{aligned} var[\mu_j] &= E[\mu_j^2] - E[\mu_j]^2 \\ &= \frac{\alpha_j(\alpha_j+1)}{\alpha_0(\alpha_0+1)} - \frac{\alpha_j}{\alpha_0}^2 \\ &= \frac{\alpha_j(\alpha_0+1)}{\alpha_0^2(\alpha_0+1)} \end{aligned}$$

In case of Covariance:

$$
\begin{aligned}
cov[\mu_j\mu_l] &= E[(\mu_j - E[\mu_j])(\mu_l - E[\mu_l])] \\
&= \int (\mu_j - E[\mu_j])(\mu_l - E[\mu_l]) Dir(x|\mu) d\mu \\
&= \int (\mu_j\mu_l - E[\mu_l]\mu_j - E[\mu_j]\mu_l + E[\mu_j]E[\mu_l]) Dir(x|mu) d\mu \\
&= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \frac{\Gamma(\alpha_1)\cdots\Gamma(\alpha_j+1)\Gamma(\alpha_l+1)\cdots\Gamma(\alpha_K)}{\Gamma(\alpha_0+2)} - E[\mu_l]E[\mu_j] - E[\mu_j]E[\mu_l] + E[\mu_j]E[\mu_l] \\
&= \frac{\alpha_j\alpha_l}{\alpha_0(\alpha_0+1)} - E[\mu_j]E[\mu_l] \\
&= \frac{\alpha_j\alpha_l}{\alpha_0(\alpha_0+1)} - \frac{\alpha_j\alpha_l}{\alpha_0^2} \\
&= -\frac{\alpha_j\alpha_l}{\alpha_0^2(\alpha_0+1)}
\end{aligned}
$$

Problem 2.11

Differentating the Dirichlet Distribution:

$$
\begin{aligned}
\frac{\partial Dir(\mu|\alpha)}{\partial\alpha_j} &= \partial(\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k-1})/\partial\alpha_j \\
&= \frac{\partial\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)}}{\partial\alpha_j} \prod_{k=1}^{K} \mu_k^{\alpha_k-1} + \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \frac{\partial\prod_{k=1}^{K}\mu_k^{\alpha_k-1}}{\partial\alpha_j}
\end{aligned}
$$

We will handle with second term:

$$
\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \frac{\partial\prod_{k=1}^{K}\mu_k^{\alpha_k-1}}{\partial\alpha_j} = ln(\alpha_j) Dir(\mu|\alpha)
$$

With above result, we integrate both sides of equation and handle the left side firstly:

$$
\int \frac{\partial Dir(\mu|\alpha)}{\partial\alpha_j} d\mu = \int \frac{\partial 1}{\partial\alpha_j} d\mu = 0
$$

The Right Side:

$$
\begin{aligned}
&\int \frac{\partial\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)}}{\partial\alpha_j} \prod_{k=1}^{K} \mu_k^{\alpha_k-1} d\mu + \int ln(\alpha_j) Dir(\mu|\alpha) d\mu \\
&= \frac{\partial\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)}}{\partial\alpha_j} \frac{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)}{\Gamma(\alpha_0)} + E[ln(\alpha_j)] \\
&= \frac{\partial ln(\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)})}{\partial\alpha_j} + E[ln(\alpha_j)]
\end{aligned}
$$

Combining both sides, we obtain:

$$\begin{aligned} E[ln(\mu)] &= -\frac{\partial ln(\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)})}{\partial \alpha_j} \\ &= -\frac{\partial ln(\Gamma(\alpha_0)) - ln(\Gamma(\alpha_1)\cdots\Gamma(\alpha_K))}{\partial \alpha_j} \\ &= \frac{\partial ln(\Gamma(\alpha_j))}{\partial \alpha_j} - \frac{\partial ln(\Gamma(\alpha_0))}{\partial \alpha_j} \\ &= \frac{\partial ln(\Gamma(\alpha_j))}{\partial \alpha_j} - \frac{\partial ln(\Gamma(\alpha_0))}{\partial \alpha_0}\frac{\partial \alpha_0}{\partial \alpha_j} = \psi(\alpha_j) - \psi(\alpha_0) \end{aligned}$$

Problem 2.12

Based on equation (2.278):

$$\int_a^b U(x|a,b)dx = \int_a^b \frac{1}{b-a}dx = \frac{b}{b-a} - \frac{a}{b-a} = 1$$

Hence the Uniform Distribution is normalized. Next is mean of the distribution:

$$E[x] = \int_a^b xU(x|a,b)dx = \frac{1}{2(b-a)}x^2|_a^b = \frac{b^2}{2(b-a)} - \frac{a^2}{2(b-a)} = \frac{(a+b)}{2}$$

In case of variance of the Uniform Distribution:

$$\begin{aligned} var[x] = E[x^2] - E[x]^2 &= \int_a^b x^2 U(x|a,b)dx - (\frac{(a+b)}{2})^2 \\ &= \frac{b^3-a^3}{3(b-a)} - \frac{(a+b)^2}{4} \\ &= \frac{a^2+b^2-2ab}{12} = \frac{(a+b)^2}{12} \end{aligned}$$

Problem 2.13

The Gaussian Distribution:

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

KL Divergence:

$$KL(p(x)||q(x)) = \int p(x)ln\{\frac{p(x)}{q(x)}\}dx$$

With two above equations, we obtain:

$$KL(p(x)||q(x)) = \int p(x)\frac{1}{2}(ln(\frac{|L|}{|\Sigma|}) - (x-\mu)^T\Sigma^{-1}(x-\mu) + (x-m)^TL^{-1}(x-m))dx$$

We have property of expectation:

$$E[(x-\mu)^T A(x-\mu)] = Tr(A\Sigma) + (x-\mu)^T A(x-\mu)$$
$$E[x] = \int p(x)xdx$$
$$\int p(x)dx = 1$$

Hence:

$$\begin{aligned}
KL(p(x)||q(x)) &= \frac{1}{2}(ln(\frac{|L|}{|\Sigma|}) + E[(x-\mu)^T\Sigma^{-1}(x-\mu)] - E[(x-m)^T L^{-1}(x-m)]) \\
&= \frac{1}{2}(ln(\frac{|L|}{|\Sigma|}) + Tr(\Sigma^{-1}\Sigma) + (x-\mu)^T\Sigma^{-1}(x-\mu) - Tr(L^{-1}\Sigma) - (x-m)^T L^{-1}(x-m)) \\
&= \frac{1}{2}(ln(\frac{|L|}{|\Sigma|}) + Tr(I_D) + (\mu-\mu)^T\Sigma^{-1}(\mu-\mu) - Tr(L^{-1}\Sigma) - (\mu-m)^T L^{-1}(\mu-m)) \\
&= \frac{1}{2}(ln(\frac{|L|}{|\Sigma|}) + D - Tr(L^{-1}\Sigma) - (\mu-m)^T L^{-1}(\mu-m))
\end{aligned}$$

Problem 2.30

Mean of z:

$$E[z] = R^{-1}\begin{pmatrix}\Lambda\mu - A^T Lb \\ Lb\end{pmatrix}$$

Covariance of z:

$$cov[z] = R^{-1} = \begin{pmatrix}\Lambda^{-1} & \Lambda^{-1}A^T \\ A\Lambda^{-1} & L^{-1} + A\Lambda^{-1}A^T\end{pmatrix}$$

Hence:

$$E[z] = \begin{pmatrix}\Lambda^{-1} & \Lambda^{-1}A^T \\ A\Lambda^{-1} & L^{-1} + A\Lambda^{-1}A^T\end{pmatrix}\begin{pmatrix}\Lambda\mu - A^T Lb \\ Lb\end{pmatrix} = \begin{pmatrix}\mu \\ A\mu + b\end{pmatrix}$$

Problem 2.35

The proof of 2.62 is in book so we don't prove it. Consequently, we will prove the equation (2.124)

$$\begin{aligned}
E[\Sigma_{ML}] &= E[\frac{1}{N}\sum_{N=1}^{N}(x-\mu_{ML})^T(x-\mu_{ML})] \\
&= \frac{1}{N}E[\sum_{n=1}^{N}(x_n x_m^T - 2x_n\mu_{ML} + \mu_{ML}^T\mu_{ML})] \\
= \frac{1}{N}\sum_{n=1}^{N}E[x_n x_n^T] - \frac{2}{N}\sum_{n=1}^{N}E[x_n\mu_{ML}] &+ \frac{1}{N}\sum_{n=1}^{N}E[\mu_{ML}^T\mu_{ML}]
\end{aligned}$$

For the first term:

$$\frac{1}{N}\sum_{n=1}^{N} E[x_n x_n^T] = \frac{1}{N}(N(\mu\mu^T + \Sigma)) = \mu\mu^T + \Sigma$$

For the second term:

$$\begin{aligned}
-\frac{2}{N}\sum_{n=1}^{N} E[x_n \mu_{ML}] &= -\frac{2}{N}\sum_{n=1}^{N} E[x_n(\frac{1}{N}\sum_{m=1}^{M} x_m)] \\
&= -\frac{2}{N^2}\sum_{n=1}^{N}\sum_{m=1}^{M} E[x_n x_m] \\
&= -\frac{2}{N^2}\sum_{n=1}^{N}\sum_{m=1}^{M}(\mu\mu^T + I_{nm}\Sigma) \\
= -\frac{2}{N^2}N^2(\mu\mu^T + \frac{1}{N}\Sigma) &= -2(\mu\mu^T + \frac{1}{N}\Sigma)
\end{aligned}$$

For the third term:

$$\begin{aligned}
\frac{1}{N}\sum_{n=1}^{N} E[\mu_{ML}\mu_{ML}^T] &= \frac{1}{N}E[\sum_{n=1}^{N}\sum_{m=1}^{M} x_n x_m] \\
&= \frac{1}{N^2}N^2(\mu\mu^T + \frac{1}{N}\Sigma)
\end{aligned}$$

Combine all above results, we obtain:

$$\begin{aligned}
E[\Sigma_{ML}] &= \mu\mu^T + \Sigma - 2(\mu\mu^T + \frac{1}{N}\Sigma) + \mu\mu^T + \frac{1}{N}\Sigma \\
&= \Sigma - \frac{1}{N}\Sigma = \frac{N-1}{N}\Sigma
\end{aligned}$$

Problem 2.41

$$\begin{aligned}
\int_0^\infty Gamma(\lambda|a,b) = 1 \int_0^\infty Gamma(\lambda|a,b) &= \int_0^\infty \frac{1}{\Gamma(a)} b^a \lambda^{a-1} e^{-b\lambda} d\lambda \\
&= \frac{b^a}{\Gamma(a)}\int_0^\infty \lambda^{a-1} e^{-b\lambda} d\lambda
\end{aligned}$$

Subtituting $u = b\lambda$, we obtain:

$$\begin{aligned}
&= \frac{b^a}{\Gamma(a)}\int_0^\infty \frac{1}{b}(\frac{u}{b})^{a-1} e^u du \\
&= \frac{b^a}{\Gamma(a)}\int_0^\infty (\frac{1}{b})^a u^{a-1} e^u du \\
&= \frac{1}{\Gamma(a)}\Gamma(a) = 1 \qquad \text{Solved}
\end{aligned}$$

Problem 2.42

I will get result from problem (2.41):

$$\begin{aligned} E[\lambda] &= \int_0^\infty Gamma(\lambda|a,b)\lambda d\lambda \\ &= \frac{b^a}{\Gamma(a)}\int_0^\infty \int_0^\infty \lambda^a e^{-b\lambda} d\lambda \\ &= \frac{b^a}{\Gamma(a)}\frac{1}{b^{a+1}}\Gamma(a+1) \end{aligned}$$

We have proved $\Gamma(x+1) = x\Gamma(a)$. Hence:

$$\begin{aligned} E[\lambda] &= b^a \frac{1}{b^{a+1}}\frac{1}{\Gamma(a)}a\Gamma(a) = \frac{a}{b}(\text{Solved}) \\ var[\lambda] &= E[\lambda^2] - E[\lambda]^2 \\ &= \int_0^\infty Gamma(\lambda|a,b)\lambda^2 d\lambda - (\frac{a}{b})^2 \\ &= \frac{b^a}{\Gamma(a)}\frac{1}{b^{a+2}}\Gamma(a+1) - \frac{a^2}{b^2} \\ &= \frac{a(a+1)}{b^2} - \frac{a^2}{b^2} = \frac{a}{b^2}(\text{Solved}) \\ mode[\lambda] &= \frac{\partial}{\partial\lambda}\frac{1}{\Gamma(a)}b^a\lambda^{a-1}e^{-b\lambda} \\ &= \frac{1}{\Gamma(a)}b^a((a-1)\lambda^{a-2}e^{-b\lambda} - b\lambda^{a-1}e^{-b\lambda}) \\ &= \frac{1}{\Gamma(a)}b^a\lambda^{a-2}e^{-b\lambda}((a-1) - b\lambda) \end{aligned}$$

Give it equal zero so we obtain:

$$mode[\lambda] = \frac{a-1}{b}$$

When $b \neq 0$

Linear Models for Regression

Problem 3.1

We have formula of tanh(a):

$$\begin{aligned} tanh(a) = \frac{e^a - e^{-a}}{e^a + e^{-a}} &= \frac{e^a(1 - e^{-2a})}{e^a(1 + e^{-2a})} \\ = \frac{(1 - e^{-2a})}{(1 + e^{-2a})} &= \frac{-1 - e^{-2a} + 2}{1 + e^{-2a}} \\ &= -1 + 2\sigma(-2a) \end{aligned}$$

Linear combination of logistic sigmoid function:

$$\begin{aligned} y(x, w) &= w_0 + \sum_{j=1}^{M} w_j \sigma(\frac{x - \mu_j}{s}) \\ &= w_0 + \sum_{j=1}^{M} w_j (\frac{tanh(\frac{x-\mu_j}{s}) + 1}{2}) \\ = w_0 + \frac{1}{2}\sum_{j=1}^{M} w_j + \sum_{j=1}^{M} \frac{w_j}{2} tanh(\frac{x - \mu_j}{s}) \end{aligned}$$

So the linear combination of sigmoid function will be equivalent to a linear combination of tanh function when:

$$\mu_0 = w_0 + \frac{1}{2}\sum_{j=1}^{M} w_j \qquad\qquad \mu_j = \frac{w_j}{2}$$