Bishop Questions

David Ruhe

July 20, 2020

Chapter 1

1.1

Differentiating to $\{w_i\}$ means differentiating to every single w_i in $\boldsymbol{w}$. Sum of squares: $E(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \boldsymbol{w}) - t_n\}^2$.

$$\frac{\partial E(\boldsymbol{w})}{\partial w_i} = \frac{\partial y(x_n, \boldsymbol{w})}{\partial w_i}\sum_{n=1}^{N}\{y(x_n, \boldsymbol{w}) - t_n\} \tag{1}$$

$$= x_n^i \sum_{n=1}^{N}\{\sum_{j=0}^{M} w_j x_n^j - t_n\} = 0 \tag{2}$$

$$= \sum_{n=1}^{N}\{\sum_{j=0}^{M} w_j (x_n)^{i+j} - (x_n)^i t_n\} = 0 \tag{3}$$

$$= \sum_{n=1}^{N}\sum_{j=0}^{M}\{w_j (x_n)^{i+j}\} - \sum_{n=1}^{N}(x_n)^i t_n = 0 \tag{4}$$

$$= \sum_{n=1}^{N}\sum_{j=0}^{M}\{w_j (x_n)^{i+j}\} = \sum_{n=1}^{N}(x_n)^i t_n \tag{5}$$

$$= \sum_{j=0}^{M} w_j \sum_{n=1}^{N}(x_n)^{i+j} = \sum_{n=1}^{N}(x_n)^i t_n \tag{6}$$

$$= \sum_{j=0}^{M} w_j A_{ij} = T_i \tag{7}$$

1.2

$\widetilde{E} = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \boldsymbol{w}) - t - N\}^2 + \frac{\lambda}{2}\boldsymbol{w}^T\boldsymbol{w}$

$$\frac{\partial \widetilde{E}(\boldsymbol{w})}{\partial w_i} = x_n^i \sum_{n=1}\{\sum_{j=1}^{M} w_j x_n^j - t_n\} + \lambda w_i \tag{8}$$

$$= \sum_{n=1}^{N}\{\sum_{j=0}^{M} w_j x_n^{j+i} - t_n + \frac{\lambda}{N} w_i\} = 0 \tag{9}$$

$$= \sum_{n=1}^{N}\{\sum_{j=0}^{M} w_j x_n^{j+i} + \frac{\lambda}{N} w_i\} = \sum_{n=1}^{N} x_n^i t_n \tag{10}$$

$$= \sum_{j=0}^{M} w_j\{\sum_{n=1}^{N} x_n^{j+i} + \frac{\lambda}{N} w_i\} = \sum_{n=1}^{N} x_n^i t_n \tag{11}$$

$$= \sum_{j=0}^{M} w_j A_{ij} = T_i \tag{12}$$

1.3

$$p(F = a) = \sum_b p(F = a|B = b)p(B = b) = 0.3 \cdot 0.2 + 0.2 \cdot 0.5 + 0.6 \cdot 0.3 = 0.34 \tag{13}$$

$$p(B = g|F = o) = \frac{p(F = o|B = g)p(B = g)}{p(F = o)} \tag{14}$$

$$= \frac{0.3 \cdot 0.6}{\sum_b p(F = o|B = b)p(B = b)} \tag{15}$$

$$= \frac{0.3 \cdot 0.6}{0.4 \cdot 0.2 + 0.5 \cdot 0.2 + 0.3 \cdot 0.6} = 0.5 \tag{16}$$

1.4

$$p'(y) = p'(g(y)) \cdot g'(y) \cdot |g'(y)| + p(g(y))\frac{d|g'(y)|}{dy} \tag{17}$$

If we maximize $p(x)$ we get that the first term is 0. In the nonlinear case, the second term is not equal to 0 however, so maximizing one distribution does not mean maximizing the other distribution.

Derivation of the change of variables formula. We have a random variable X with pdf f(x). We define Y=g(X) where $g(\cdot)$ is a *monotone* function, the pdf of Y is obtained as follows.

We start with the CDF of y.

$$P(Y \leq y) = P(g(X) \leq y) \tag{18}$$

$$= P(X \leq g^{-1}(y)) \qquad \text{Because } g \text{ is monotonically increasing.} \tag{19}$$

$$\tag{20}$$

This is equivalent to $F(y) = F(g^{-1}(y))$. Therefore, if we differentiate both sides and use the chain rule we get

$$f(y) = f(g^{-1}y)\frac{d}{dy}g^{-1}(y) \tag{21}$$

If the function is monotonically decreasing we get

$$\begin{aligned} P(Y \leq y) &= P(g(X) \leq y) && (22)\\ &= P(X \geq g^{-1}) && (23)\\ &= 1 - F(g^{-1}(y)) && (24) \end{aligned}$$

Differentiating both sides leads to

$$f(y) = f(g^{-1}y)\frac{d}{dy}g^{-1}(y) \tag{25}$$

Combining these formulas gives the formula with the absolute value. This works for both montonically decreasing and increasing. If g is not monotonic, we need to split the function into monotonic parts and do this transformation for them individually, then sum.

1.5

$$\begin{aligned} var[f] &= \mathbb{E}\left[(f(x) - \mathbb{E}\left[f(x)\right])^2\right] && (26)\\ &= \mathbb{E}\left[f(x)^2 + \mathbb{E}\left[f(x)\right]^2 - 2f(x)\,\mathbb{E}[f(x)]\right] && (27)\\ &= \mathbb{E}[f(x)^2] + \mathbb{E}[f(x)]^2 - 2\,\mathbb{E}[f(x)]\,\mathbb{E}[f(x)] && (28)\\ &= \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2 && (29) \end{aligned}$$

1.6

$$\begin{aligned} cov[x,y] &= \mathbb{E}_{x,y}[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}] && (30)\\ &= \int_{x,y} p(x,y)[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}] && (31)\\ &= \int_x \int_y p(x)p(y)[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}] && (32)\\ &= \int_x p(x)\{x - \mathbb{E}[x]\} \int_y p(y)\{y - \mathbb{E}[y]\} && (33)\\ &= \mathbb{E}[x - \mathbb{E}[x]]\,\mathbb{E}[y - \mathbb{E}[y]] && (34)\\ &= (\mathbb{E}[x] - \mathbb{E}[x])(\mathbb{E}[y] - \mathbb{E}[y]) && (35) \end{aligned}$$

1.7

We have:

$$I = \int \exp(-\frac{1}{2\sigma^2}x^2)dx \tag{36}$$

Making use of $\int f(x)dx \int g(y)dy = \int\int f(x)g(y)dxdy$:

$$I^2 = \int\int \exp(-\frac{1}{2\sigma^2}(x^2+y^2))dxdy \tag{37}$$

Using Pythagoras we know $x^2 + y^2 = r^2$.

$$I^2 = \int\int \exp(-\frac{1}{2\sigma^2}r^2)dxdy \tag{38}$$

Changing the integration to polar (i.e., over θ and r) works as follows. We now integrate over a tiny part of a circle ($d\theta$ and dr). The length of the corresponding arc is equal to $S = d\theta r$, where θ is expressed in radians. This follows directly from the fact that full circle circumference is $2\pi r$, where 2π is the 'angle' of a full circle expressed in radians. We approximate the area between the arc and dr as a rectangle, giving that the tiny parts we integrate with (corresponding to $dxdy$ in Cartesian plane) are of size $rd\theta dr$. Thus, the integral becomes:

$$I^2 = \int_0^\infty \int_0^{2\pi} \exp(-\frac{1}{2\sigma^2}r^2)rdrd\theta \tag{39}$$

Since one of the integrants is expressed in radians we integrate from 0 to 2π, and the radius is nonnegative.

$$\begin{aligned}
I^2 &= \int_0^{2\pi} d\theta \int_0^\infty \exp(-\frac{1}{2\sigma^2}r^2)rdr && (40)\\
&= 2\pi(-\sigma^2\exp(-\frac{1}{2\sigma^2}r^2))\Big|_0^\infty && (41)\\
&= 2\pi(-\sigma^2\exp(-\frac{1}{2\sigma^2}\infty^2)) - 2\pi(-\sigma^2\exp(-\frac{1}{2\sigma^2}0^2)) && (42)\\
&= 0 + 2\pi\sigma^2 && (43)
\end{aligned}$$

So

$$I = \sqrt{2\pi\sigma^2} \tag{44}$$

which exactly cancels the first factor in the Gaussian equation, so the result is 1.

1.8

$$\mathbb{E}[x] = \int \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2]xdx \tag{45}$$

Using $y = \frac{x-\mu}{\sigma}$ we get $\frac{dy}{dx} = \frac{1}{\sigma} \Rightarrow dx = dy\sigma$ and $x = y\sigma + \mu$.

$$\begin{aligned}
\mathbb{E}[x] &= \int \frac{1}{\sqrt{2\pi}\cancel{\sigma}}(y\sigma+\mu)\exp[-\frac{1}{2}y^2]\cancel{\sigma}dy && (46)\\
&= \int \frac{1}{\sqrt{2\pi}}y\sigma\exp[-\frac{1}{2}y^2] + \frac{1}{\sqrt{2\pi}}\mu\exp[-\frac{1}{2}y^2]dy && (47)\\
&= -\frac{\sigma}{\sqrt{2\pi}}\exp[-\frac{1}{2}y^2]\Big|_{-\infty}^{\infty} + \mu && (48)\\
&= 0 - 0 + \mu && (49)
\end{aligned}$$

Next,

$$\frac{1}{\sqrt{2\pi}\sigma}\int \exp[-\frac{1}{2}\frac{(x-mu)^2}{\sigma^2}]dx = 1 \tag{50}$$

$$\Rightarrow \int \exp[-\frac{1}{2}\frac{(x-mu)^2}{\sigma^2}]dx = \sqrt{2\pi\sigma^2} \tag{51}$$

$$\tag{52}$$

Now differentiating both sides.

$$\int \frac{1}{2}(x-\mu)^2\sigma^{-4}\exp[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}]dx = \frac{1}{2}(2\pi\sigma^2)^{-1/2}2\pi \tag{53}$$

$$= \frac{1}{2}(2\pi)^{1/2}\sigma^{-1} \tag{54}$$

$$\sigma^{-4}\int (x-\mu)^2\exp[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}] = (2\pi)^{1/2}\sigma^{-1} \tag{55}$$

$$\Rightarrow \sigma^{-1}(2\pi)^{-1/2}\int (x-\mu)^2\exp[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}] = \sigma^2 \tag{56}$$

$$\Rightarrow \mathbb{E}[(x-\mu)^2] = \mathrm{Var}[x] = \sigma^2 \tag{57}$$

$$\mathbb{E}[x^2] - 2\mu\,\mathbb{E}[x] + \mu^2 = \sigma^2 \tag{58}$$

$$\Rightarrow \mathbb{E}[x^2] - 2\mu^2 + \mu^2 = \sigma^2 \tag{59}$$

$$\Rightarrow \mathbb{E}[x^2] = \sigma^2 + \mu^2 \tag{60}$$

The last equality is easily found using the previous results.

1.9

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{61}$$

$$\frac{\partial\mathcal{N}(x|\mu,\sigma^2)}{\partial\mu} = \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}\sigma^2(x-\mu) = 0 \tag{62}$$

$$x-\mu = 0 \tag{63}$$

$$x = \mu \tag{64}$$

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^2}\exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right] \tag{65}$$

$$\frac{\partial\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\Sigma)}{\partial\boldsymbol{\mu}} = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^2}\exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]\left[-\frac{1}{2}(\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Sigma}^{-1^T})(\boldsymbol{x}-\boldsymbol{\mu})\right] = 0 \tag{66}$$

$$\tag{67}$$

So only when $\boldsymbol{x} = \boldsymbol{\mu}$. Used Matrix cookbook and diagonality of $\boldsymbol{\Sigma}^{-1}$.

1.10

$$\mathbb{E}[x+z] = \int_{x+z} p(x+z)\,[x+z] \tag{68}$$
$$= \int_{x+z} p(x+z)x + \int_{x+z} p(x+z)z \tag{69}$$
$$= \int_{x,z} p(x,z)x + \int_{x,z} p(x,z)z \tag{70}$$
$$= \int_{x,z} p(x)p(z)x + \int_{x,z} p(x)p(z)z \tag{71}$$
$$= \int_{x} p(x)x + \int_{z} p(z)z \tag{72}$$
$$= \mathbb{E}[x] + \mathbb{E}[z] \tag{73}$$

$$var[x+z] = \mathbb{E}\left[\left\{(x+z) - \mathbb{E}[x+z]\right)^2\right\}\Big] \tag{74}$$
$$= \mathbb{E}\left[(x+z)^2\right] - \mathbb{E}\left[x+z\right]^2 \tag{75}$$
$$= \mathbb{E}\left[x^2 + z^2 + 2xz\right] - \left(\mathbb{E}[x] + E[z]\right)^2 \tag{76}$$
$$= \mathbb{E}[x^2] + \mathbb{E}[z^2] + \mathbb{E}[2xz] - \mathbb{E}[x]^2 - \mathbb{E}[z]^2 - 2\,\mathbb{E}[z]\,\mathbb{E}[x] \tag{77}$$
$$= var[x] + var[y] + 2\int_{x,z} p(x,z)xz - 2\int_{x} p(x)\int_{z} p(z) \tag{78}$$
$$= var[x] + var[y] + 2\int_{x,z} p(x)p(z)xz - 2\int_{x} p(x)\int_{z} p(z) \tag{79}$$
$$= var[x] + var[y] + 2\int_{x} p(x)\int_{z} p(z) - 2\int_{x} p(x)\int_{z} p(z) \tag{80}$$

1.11

Mean:

$$\ln p(\boldsymbol{x}|\mu,\sigma^2) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln 2\pi \tag{81}$$
$$\frac{\partial \ln p(\boldsymbol{x}|\mu,\sigma^2)}{\partial\mu} = \sigma^2\sum_{n=1}^{N}(x_n-\mu) = 0 \tag{82}$$
$$\sum_{n=1}^{N} x_n = N\mu \tag{83}$$
$$\mu = \frac{1}{N}\sum_{n=1}^{N} x_n \tag{84}$$

Variance:

$$-\frac{1}{2\sigma^2} = -(2\sigma^2)^{-1} \tag{85}$$

$$\frac{\partial \ln p(\boldsymbol{x}|\mu,\sigma^2)}{\partial \sigma^2} = 2 \cdot \frac{1}{4\sigma^4} \sum_{n=1}^{N} (x_n - \mu)^2 - \frac{N}{2\sigma^2} = 0 \tag{86}$$

$$N \cdot 2\sigma^4 = 2\sigma^2 \sum_{n=1}^{N} (x_n - \mu)^2 \tag{87}$$

$$\sigma^2 = \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu)^2 \tag{88}$$

1.12

If $x_n = x_m$ we have $\mathbb{E}[x_n^2] = \mu^2 + \sigma^2$, otherwise we have $\mathbb{E}[x_n x_m] = \mathbb{E}[x_n]\,\mathbb{E}[x_m] = \mu^2$. So we only get σ^2 if $n = m$, which leads to the given equation.

$$\mathbb{E}[\mu_{ML}] = \frac{1}{N}\,\mathbb{E}[\sum_{n=1}^{N} x_n] = \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}[x_n] = \frac{1}{N} N\mu \tag{89}$$

$$\mathbb{E}[\sigma^2_{ML}] = \frac{1}{N}\,\mathbb{E}[\sum_{n=1}^{N} (x_n - \mu_{ML})^2] \tag{90}$$

$$= \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}[x_n^2 - 2x_n \frac{1}{N} \sum_{m=1}^{N} x_m + \frac{1}{N^2} \sum_{l=1}^{N} \sum_{k=1}^{N} x_l x_k] \tag{91}$$

$$= \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}[\mu^2 + \sigma^2 - 2\frac{1}{N} \sum_{m=1}^{N} x_m x_n + \frac{1}{N^2} \sum_{l=1}^{N} \sum_{k=1}^{N} x_l x_k] \tag{92}$$

$$= \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}[\mu^2 + \sigma^2] - 2\frac{1}{N} \sum_{m=1}^{N} \mathbb{E}[x_m x_n] + \frac{1}{N^2} \sum_{l=1}^{N} \sum_{k=1}^{N} \mathbb{E}[x_l x_k] \tag{93}$$

$$= \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}[\mu^2 + \sigma^2] - 2\frac{1}{N} \sum_{m=1}^{N} (\mu^2 + I_{nm}\sigma^2) + \frac{1}{N^2} \sum_{l=1}^{N} \sum_{k=1}^{N} (\mu^2 + I_{kl}\sigma^2) \tag{94}$$

$$= \mu^2 + \sigma^2 - 2\mu^2 - 2\frac{1}{N}\sigma^2 + \mu^2 + \frac{1}{N}\sigma^2 \tag{95}$$

$$= \frac{N}{N}\sigma^2 - \frac{1}{N}\sigma^2 \tag{96}$$

1.13

From 1.56:

$$\mathbb{E}\left[\frac{1}{N} \sum_{n=1}^{N} (x_n - \mu)^2\right] = \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}[x^2] - \mathbb{E}[2x_n\mu] + \mathbb{E}[\mu^2] \tag{97}$$

$$= \mathbb{E}[x^2] - 2\mu\,\mathbb{E}[x] + \mu^2 \tag{98}$$

$$= \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2 \tag{99}$$

1.14

We can write any matrix W as a sum of a symmetric matrix S where the off-diagonals $S_{ij} = \frac{W_{ij}+W_{ji}}{2}$ and the diagonals $S_{ii} = W_{ii}$. The antisymmetric matrix diagonal will be Diag(0). The off-diagonal entries will be $W_{ij} - S_{ij}$. Thus we can have

$$x^T W x = x^T (S + A)x = x^T S x + x^T A x = x^T S x + \sum_{i=1}^{D}\sum_{j=1}^{D} x_i x_j w_{ij} \tag{100}$$

The second term will be, because of the antisymmetry, 0. This means that only the symmetric part contributes.

1.17

$$\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} du \tag{101}$$

$$\Rightarrow \Gamma(x+1) = \int_0^\infty u^x e^{-u} du \tag{102}$$

Using $\int f(x)g'(x)dx = f(x)g(x) - \int f'(x)g(x)dx$ and using $f(x) = u^x, f'(x) = xu^{x-1}, g'(x) = e^{-u}, g(x) = -e^{-u}$ we get

$$\int_0^\infty u^x e^{-u} = \left(u^x e^{-u} + \int_0^\infty x u^{x-1} e^{-u}\right)\Big|_0^\infty \tag{103}$$

$$= \left(u^x e^{-u} + x\int_0^\infty u^{x-1} e^{-u} du\right)\Big|_0^\infty \tag{104}$$

$$= \left(u^x e^{-u} + x\int_0^\infty u^{x-1} e^{-u} du\right)\Big|_0^\infty \tag{105}$$

$$= \left(u^x e^{-u} + x\Gamma(x)\right)\Big|_0^\infty \tag{106}$$

$$= u^x e^{-u}\Big|_0^\infty + x\Gamma(x) \tag{107}$$

$$\tag{108}$$

Where the second term is no function of u anymore, so we don't need to evaluate it for u. Since $u^x e^{-u} = \frac{u^x}{e^u} = \frac{f(x)}{g(x)}$ with the indefinite form as we take the limit $\frac{\infty}{\infty}$ we can use l'Hopital.

Using l'Hopital we see that $\lim_{u\to\infty} = \frac{u^x}{e^u} = \lim_{u\to\infty} \frac{(x-1)u^{x-1}}{e^u}$

If we keep doing this recursively for any integer x we will obtain $\lim_{u\to\infty} \frac{x!}{e^u}$ which clearly converges to zero. However, even for a (negative) non integer y there will be a larger x, meaning a larger numerator, for which the limit still converges to zero, showing that the result still holds for y.

1.18

The right integral simplifies:

$$\int_0^\infty e^{-r^2} r^{D-1} dr = \frac{1}{2}\int_0^\infty e^{-u}\sqrt{D-1}\sqrt{u}^{-1} dr \qquad u = r^2 \Rightarrow \frac{du}{dr} = 2r = 2\sqrt{u} \Rightarrow dr = \frac{du}{2\sqrt{u}} \tag{109}$$

$$= \frac{1}{2}\int_0^\infty e^{-u}\sqrt{u}^{D-1}(\sqrt{u})^{-1} dr \tag{110}$$

$$= \frac{1}{2}\int_0^\infty e^{-u}(\sqrt{u})^{D-2} dr \tag{111}$$

$$= \frac{1}{2}\int_0^\infty e^{-u} u^{1/2D-1} dr \tag{112}$$

$$= \frac{1}{2}\Gamma(\frac{1}{2}D) \tag{113}$$

Using $\sigma = 1$ we see from 1.126 that $I^D = \prod_{i=1}^D I = (2\pi)^{D/2}$. Filling in these two results gives the answer.

The next part is a bit tricker. We need to integrate over the radius. The total volume will be a sum of 'slices' (spherical shells). E.g., the volume of an onion is the sum of the volume of its rings. For example, the volume of a sphere is $4/3\pi r^3$. What happens to the area if we increase r a bit? We get that $dV = \frac{4}{3}(r+dr)^3 - \frac{4}{3}\pi r^3$. Working this out gives $dV = 4r^2 dr + r dr^2 + dr^3$. We see that the second and third term become arbitrarily small as $dr \to 0$. Therefore, dV can be well approximated by $4r^2 dr$, using the formula for its surface area. Integrating dV will give V. This generalizes: $\int_0^1 S_D r^{D-1} dr$. Evaluating this integral gives the desired result.

1.19

$$\frac{S_D}{2^D D} = \frac{2\pi^{D/2}}{\Gamma(D/2)}\frac{1}{2^D D} = \frac{2\pi^{D/2}}{\Gamma(D/2)2^D D} = \frac{2^{1-D}\pi^{D/2}}{\Gamma(D/2)D} = \frac{\pi^{D/2}}{\Gamma(D/2)2^D D} \tag{114}$$

If we take $\Gamma(D/2) = \Gamma(D/2 - 1 + 1$ and apply Stirling we get $\Gamma(D/2) = 2\pi^{1/2}e^{-D/2-1}(D/2 - 1)^{D/2+1/2}$. Plugging this into the fraction and plotting shows that this function goes to infinity. An ambitious person could apply l'Hopital.

For the final problem: start with a square (D=2). The distance to the corner using Pythagoras will be $\sqrt{a^2+a^2} = \sqrt{2a^2} = \sqrt{Da^2}$ If we add a dimension the new distance to the corner will be applying pythagoras again, using the previous result. $\sqrt{\sqrt{a^2+a^2}^2 + a^2} = \sqrt{3a^2}$. Using induction you could show that this generalizes to $\sqrt{Da^2}$. Taking the square root and dividing over a gives the result.

1.20

1.21

$$a \leq b \tag{115}$$
$$\Rightarrow \sqrt{a} \leq \sqrt{b} \qquad \text{(Nonnegativity)} \tag{116}$$
$$\Rightarrow a \leq \sqrt{a}\sqrt{b} \tag{117}$$

We know that

$$\int_{R_1} p(x, C_2) \leq \left(\int_{R_1} p(x, C_1) dx \int_{R_1} p(x, C_2) dx \right)^{1/2} \tag{118}$$

Thus, by symmetry we get

$$P(\text{mistake}) = \left(\int_{R_1} p(x, C_1) dx \int_{R_1} p(x, C_2) dx \right)^{1/2} + \left(\int_{R_2} p(x, C_1) dx \int_{R_2} p(x, C_2) dx \right)^{1/2} \tag{119}$$

I do not see how this reduces to the required solution.

1.22

$$1 - I_{kj} = \begin{bmatrix} 0 & 1 & \dots & 1 \\ 1 & 0 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 0 \end{bmatrix} \tag{120}$$

$$\mathcal{R}_j^* = \arg\min_j L_{kj} p(C_k|\boldsymbol{x}) \tag{121}$$
$$= \arg\min_j \sum_{j \neq k} L_{kj} p(C_k|\boldsymbol{x}) \tag{122}$$
$$= \arg\max_j p(C_k|\boldsymbol{x}) \tag{123}$$

1.23

1.24

If the probability (output of our model) of vector x belonging to C_k times the loss we will occur if we select that class is smaller than λ we take the bet.

1.25

$$\mathbb{E}\left[L(\boldsymbol{t}, \boldsymbol{y}(\boldsymbol{x}))\right] = \int \int \|\boldsymbol{y}(\boldsymbol{x}) - \boldsymbol{t}\|^2 p(\boldsymbol{x}, \boldsymbol{t}) d\boldsymbol{x} d\boldsymbol{t} \tag{124}$$

We want to find $\mathbb{E}_t[\boldsymbol{t}|\boldsymbol{x}]$ so we switch the order of integrals.

$$= \int \int \|\boldsymbol{y}(\boldsymbol{x}) - \boldsymbol{t}\|^2 p(\boldsymbol{x}, \boldsymbol{t}) d\boldsymbol{t} d\boldsymbol{x} \tag{125}$$
$$F[\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{t}] = \int_{\boldsymbol{t}} \|\boldsymbol{y}(\boldsymbol{x}) - \boldsymbol{t}\|^2 p(\boldsymbol{x}, \boldsymbol{t}) \tag{126}$$

Euler-Lagrange:

$$\frac{\partial F}{\partial \boldsymbol{y}} - \frac{d}{d\boldsymbol{x}}\left(\frac{\partial F}{\partial \boldsymbol{y}'}\right) = 0 \tag{127}$$

$$\frac{\partial F}{\partial \boldsymbol{y}'} = 0 \tag{128}$$

$$\frac{\partial F}{\partial \boldsymbol{y}} = \int_{\boldsymbol{t}} (\boldsymbol{y}(\boldsymbol{x}) - \boldsymbol{t}) p(\boldsymbol{x}, \boldsymbol{t}) = 0 \tag{129}$$

$$\int_{\boldsymbol{t}} \boldsymbol{y}(\boldsymbol{x}) p(\boldsymbol{x}, \boldsymbol{t}) - \int_{t} \boldsymbol{t} p(\boldsymbol{x}, \boldsymbol{t}) = 0 \tag{130}$$

$$\boldsymbol{y}(\boldsymbol{x}) \int_{\boldsymbol{t}} p(\boldsymbol{t}, \boldsymbol{x}) = \int_{t} \boldsymbol{t} p(\boldsymbol{x}, \boldsymbol{t}) \tag{131}$$

$$\boldsymbol{y}(\boldsymbol{x}) = \frac{1}{p(\boldsymbol{x})} \int_{t} \boldsymbol{t} p(\boldsymbol{x}, \boldsymbol{t}) = \mathbb{E}_t[\boldsymbol{t}|\boldsymbol{x}] \tag{132}$$

$$\tag{133}$$

1.26

First, let's derive the single value case.

$$\{y(\boldsymbol{x}) - t\}^2 = \{y(\boldsymbol{x}) - \mathbb{E}[t|\boldsymbol{x}] + \mathbb{E}[t|\boldsymbol{x}] - t\}^2 \tag{134}$$

$$= \{y(\boldsymbol{x}) - E[t|\boldsymbol{x}]\}^2 + \{E[t|\boldsymbol{x}] - t\}^2 + 2\left[y(\boldsymbol{x}) - \mathbb{E}[t|\boldsymbol{x}]\right]\{\mathbb{E}[t|\boldsymbol{x}] - t\} \tag{135}$$

$$\mathbb{E}[L] = \underbrace{\mathbb{E}_{x,t}\left[(y(\boldsymbol{x}) - \mathbb{E}[t|\boldsymbol{x}])^2\right]}_{(1)} + \underbrace{\mathbb{E}_{x,t}\left[(\mathbb{E}[t|\boldsymbol{x}] - t)^2\right]}_{(2)} + \underbrace{\mathbb{E}_{x,t}\left[2\left(y(\boldsymbol{x}) - \mathbb{E}[t|\boldsymbol{x}]\right)\left(\mathbb{E}[t|\boldsymbol{x}] - t\right)\right]}_{(3)} \tag{136}$$

(1)

$$\mathbb{E}_{x,t}\left[(y(\boldsymbol{x}) - \mathbb{E}[t|\boldsymbol{x}])^2\right] = \mathbb{E}_x\, \mathbb{E}_t\left[(y(\boldsymbol{x}) - \mathbb{E}[t|\boldsymbol{x}])^2\, |\boldsymbol{x}\right] \tag{137}$$

$$= \mathbb{E}_x\, (y(\boldsymbol{x}) - \mathbb{E}[t|\boldsymbol{x}])^2$$

(Since the function is over x, the expectation over t vanishes)

$$\tag{138}$$

(2)

$$\mathbb{E}_{x,t}\left[(\mathbb{E}[t|\boldsymbol{x}] - t)^2\right] = \mathbb{E}_x\left[\mathbb{E}_t\left[\mathbb{E}[t|x]^2 + t^2 - 2\,\mathbb{E}[t|x]t\right]|x\right] \tag{139}$$

$$= \mathbb{E}_x\left[\mathbb{E}[t|x]^2 + \mathbb{E}[t^2|x] - 2\,\mathbb{E}_t[t|x]\,\mathbb{E}_t[t|x]\right]$$

($\mathbb{E}[t|x]$ can be taken out of the inner expectation since t is integrated out.)

$$= \mathbb{E}_x\left[var\left[t|x\right]\right] \tag{140}$$

(3)

$$\mathbb{E}_{x,t}\left[2\left(y(\boldsymbol{x}) - \mathbb{E}[t|\boldsymbol{x}]\right)\left(\mathbb{E}[t|\boldsymbol{x}] - t\right)\right] = \mathbb{E}_x\, \mathbb{E}_t\left[2(y(\boldsymbol{x}) - \mathbb{E}[t|x])(\mathbb{E}[t|\boldsymbol{x}] - t)|x\right] \tag{141}$$

$$= \mathbb{E}_x\left[2(y(\boldsymbol{x}) - \mathbb{E}[t|x])\right]\mathbb{E}_t\left[(\mathbb{E}[t|\boldsymbol{x}] - t)|x\right] \tag{142}$$

$$= \mathbb{E}_x\left[2(y(\boldsymbol{x}) - \mathbb{E}[t|x])\right] \cdot 0 \tag{143}$$

$$= 0 \tag{144}$$

Now we continue for the vector case.

$$\|\boldsymbol{y}(\boldsymbol{x}) - \boldsymbol{t}\|^2 = \|\boldsymbol{y}(\boldsymbol{x}) - \mathbb{E}[\boldsymbol{t}|\boldsymbol{x}] + \mathbb{E}[\boldsymbol{t}|\boldsymbol{x}] - \boldsymbol{t}\|^2 \tag{145}$$

$$= (\boldsymbol{y}(\boldsymbol{x}) - \mathbb{E}[\boldsymbol{t}|\boldsymbol{x}] + \mathbb{E}[\boldsymbol{t}|\boldsymbol{x}] - \boldsymbol{t})^T(\boldsymbol{y}(\boldsymbol{x}) - \mathbb{E}[\boldsymbol{t}|\boldsymbol{x}] + \mathbb{E}[\boldsymbol{t}|\boldsymbol{x}] - \boldsymbol{t}) \tag{146}$$

$$= (\boldsymbol{y}(\boldsymbol{x}) - \mathbb{E}[\boldsymbol{t}|\boldsymbol{x}]))^2 + (\mathbb{E}[\boldsymbol{t}|\boldsymbol{x}] - \boldsymbol{t})^2 + 2\left[(\boldsymbol{y}(\boldsymbol{x}) - \mathbb{E}[\boldsymbol{t}|\boldsymbol{x}])^T(\mathbb{E}[\boldsymbol{t}|\boldsymbol{x}] - \boldsymbol{t})\right] \tag{147}$$

$$\mathbb{E}\left[L(\boldsymbol{t}, \boldsymbol{y}(\boldsymbol{x})\right] = \mathbb{E}_{x,t}[\underbrace{(\boldsymbol{y}(\boldsymbol{x}) - \mathbb{E}[\boldsymbol{t}|\boldsymbol{x}]))^2}_{(1)} + \underbrace{(\mathbb{E}[\boldsymbol{t}|\boldsymbol{x}] - \boldsymbol{t})^2}_{(2)} + \underbrace{2\left[(\boldsymbol{y}(\boldsymbol{x}) - \mathbb{E}[\boldsymbol{t}|\boldsymbol{x}])^T(\mathbb{E}[\boldsymbol{t}|\boldsymbol{x}] - \boldsymbol{t})\right]}_{(3} \tag{148}$$

(1)

$$\mathbb{E}_{x,t}\left[(\boldsymbol{y}(\boldsymbol{x})) - \mathbb{E}[\boldsymbol{t}|\boldsymbol{x}])^2\right] = \mathbb{E}_x\,\mathbb{E}_t\left[(\boldsymbol{y}(\boldsymbol{x})) - \mathbb{E}[\boldsymbol{t}|\boldsymbol{x}])^2|\boldsymbol{x}\right] \tag{149}$$

$$= \mathbb{E}_x(\boldsymbol{y}(\boldsymbol{x})) - \mathbb{E}[\boldsymbol{t}|\boldsymbol{x}])^2 \tag{150}$$

(2)

$$\mathbb{E}_{x,t}\left[(\mathbb{E}[\boldsymbol{t}|\boldsymbol{x}] - \boldsymbol{t})^2\right] = \mathbb{E}_x\,\mathbb{E}_t\left[(\mathbb{E}[\boldsymbol{t}|\boldsymbol{x}] - \boldsymbol{t})^2|\boldsymbol{x}\right] \tag{151}$$

$$= \mathbb{E}_x\,\mathbb{E}_t\left[\mathbb{E}[\boldsymbol{t}|\boldsymbol{x}]^T\,\mathbb{E}[\boldsymbol{t}|\boldsymbol{x}] + \boldsymbol{t}^T\boldsymbol{t} - 2\,\mathbb{E}[\boldsymbol{t}|\boldsymbol{x}]^T\boldsymbol{t}|\boldsymbol{x}\right] \tag{152}$$

$$= \mathbb{E}_x\left[\mathbb{E}[\boldsymbol{t}|\boldsymbol{x}]^T\,\mathbb{E}[\boldsymbol{t}|\boldsymbol{x}] + \mathbb{E}_t\left[\boldsymbol{t}^T\boldsymbol{t}\right] - 2\,\mathbb{E}[\boldsymbol{t}|\boldsymbol{x}]^T\,\mathbb{E}[\boldsymbol{t}|\boldsymbol{x}]\right] \tag{153}$$

$$= \mathbb{E}_x\left[var[\boldsymbol{t}|\boldsymbol{x}]\right] \tag{154}$$

(3)

$$\mathbb{E}_{x,t}\left\{2\left[\boldsymbol{y}(\boldsymbol{x}) - \mathbb{E}[\boldsymbol{t}|\boldsymbol{x}])^T(\mathbb{E}[\boldsymbol{t}|\boldsymbol{x}] - \boldsymbol{t})\right]\right\} = \mathbb{E}_x\,\mathbb{E}_t\left\{2(\left[\boldsymbol{y}\boldsymbol{x}) - \mathbb{E}[\boldsymbol{t}|\boldsymbol{x}])^T(\mathbb{E}[\boldsymbol{t}|\boldsymbol{x}] - \boldsymbol{t})\right]|\boldsymbol{x}\right\} \tag{155}$$

$$= \mathbb{E}_x\,2(\boldsymbol{y}(\boldsymbol{x}) - \mathbb{E}[\boldsymbol{t}|\boldsymbol{x}])^T\,\mathbb{E}_t\left[\mathbb{E}[\boldsymbol{t}|\boldsymbol{x}] - \boldsymbol{t}\right]] \tag{156}$$

$$= \mathbf{0} \tag{157}$$

You observe from the first term that the function that minimizes this expression is $y(\boldsymbol{x}) = \mathbb{E}[t|\boldsymbol{x}]$

1.27

Using a graphing calculator one can show that derivatives of absolute value polynomials require an additional `sign` component to account for the absolute parts.

Therefore

$$\frac{d}{dy(\boldsymbol{x})} = q\int[y(\boldsymbol{x} - t)]^{q-1}\text{sign}(y(\boldsymbol{x}) - t)p(\boldsymbol{x}, t) = 0 \tag{158}$$

We see that the `sign` expression is negative for $t > t$ and vice versa. Therefore we can split the integral like so:

$$q\int[y(\boldsymbol{x} - t)]^{q-1}\text{sign}(y(\boldsymbol{x}) - t)p(\boldsymbol{x}, t) = q\int_{-infty}^{y(x)}[y(\boldsymbol{x} - t)]^{q-1}\text{sign}(y(\boldsymbol{x}) - t)p(\boldsymbol{x}, t) \tag{159}$$

$$+ q\int_{y(x)}^{\infty}[y(\boldsymbol{x} - t)]^{q-1}\text{sign}(y(\boldsymbol{x}) - t)p(\boldsymbol{x}, t) \tag{160}$$

$$\tag{161}$$

The sign of the second term will always be negative and the first term positive:

$$q\int[y(\boldsymbol{x}-t)]^{q-1}\text{sign}(y(\boldsymbol{x})-t)p(\boldsymbol{x},t) = q\int_{-infty}^{y(x)}[y(\boldsymbol{x}-t)]^{q-1}(y(\boldsymbol{x})-t)p(\boldsymbol{x},t) \tag{162}$$

$$-q\int_{y(x)}^{\infty}[y(\boldsymbol{x}-t)]^{q-1}(y(\boldsymbol{x})-t)p(\boldsymbol{x},t) \tag{163}$$

$$=0 \Rightarrow \int_{y(x)}^{\infty}|y(\boldsymbol{x})-t|^{q-1}p(t|x)dt = \int_{y(x)}^{\infty}|y(\boldsymbol{x})-t|^{q-1}p(t|x)dt \tag{164}$$

In the case of $q=1$ this reduces to an equality of two integrals over the pdf. Since the total area is 1 it must be that the both terms are 0.5, which shows that $y(x)$ is the median of the distribution.

1.28

$n=k=2$

$$h(p^n) = -\log(p(x)^2) = -2\log p(x) = 2h(p) \tag{165}$$

Induction:

$$h(p^{k+1}) = -\log p(x)^{k+1} = -(k+1)\log(p(x) = -(k+1)\log p(x) = (k+1)h(p) \tag{166}$$

So it works for $k+1$. $\frac{n}{m}$ is straightforward.

$$h(p) = -\log_z(p) = -\frac{\ln p}{\ln z} \propto \ln p \tag{???}$$

1.29

$$H[x] = -\sum_{n=1}^{M}p(x)\ln p(x) \tag{167}$$

$$=\sum_{n=1}^{M}p(x)\ln\frac{1}{p(x)} \tag{168}$$

$$\leq \ln\sum_{n=1}^{M}p(x)\frac{1}{p(x)} \qquad \text{(Concavity of } \ln\cdot\text{)}$$

$$=\ln M \tag{169}$$

1.30

$$KL[p||z] = \int p(x)\ln\frac{q(x)}{p(x)}dx \tag{170}$$

$$\ln \frac{q(x)}{p(x)} = \ln q(x) - \ln p(x) \tag{171}$$

$$= \ln\left(\frac{1}{\sqrt{2\pi\sigma}}\exp[-\frac{1}{2\sigma^2}(x-\mu)^2]\right) - \ln\left(\frac{1}{\sqrt{2\pi s^2}}\exp[-\frac{1}{2s^2}(x-m)^2]\right) \tag{172}$$

$$= -\ln(\sqrt{2\pi}) - \ln\sigma + \left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) + \ln(\sqrt{2\pi}) + \ln s - \left(-\frac{1}{2s^2}(x-m)^2\right) \tag{173}$$

$$= \ln\frac{s}{\sigma} - \frac{1}{2\sigma^2}(x^2 - 2x\mu + \mu^2) + \frac{1}{2s^2}(x^2 - 2xm + m^2) \tag{174}$$

$$= x^2\left(\frac{1}{2s^2} - \frac{1}{2\sigma^2}\right) + x\left(\frac{\mu}{\sigma^2} - \frac{m}{s^2}\right) + \left(\ln\frac{s}{\sigma} - \frac{\mu^2}{2\sigma^2} + \frac{m^2}{2s^2}\right) \tag{175}$$

The integral now becomes

$$\left(\frac{1}{2s^2} - \frac{1}{2\sigma^2}\right)\int p(x)x^2dx + \left(\frac{\mu}{\sigma^2} - \frac{m}{s^2}\right)\int xp(x)dx + \left(\ln\frac{s}{\sigma} - \frac{\mu^2}{2\sigma^2} + \frac{m^2}{2s^2}\right)\int p(x)dx \tag{176}$$

$$= \left(\frac{1}{2s^2} - \frac{1}{2\sigma^2}\right)(\mu^2 + \sigma^2) + \left(\frac{\mu}{\sigma^2} - \frac{m}{s^2}\right)\mu + \ln\frac{s}{\sigma} - \frac{\mu^2}{2\sigma^2} + \frac{m^2}{2s^2} \tag{177}$$

Which reduces to

$$\ln\frac{s}{\sigma} + \frac{\sigma^2 + (\mu - m)^2}{2s^2} - \frac{1}{2} \tag{178}$$

1.31

$$H[x,y] = -\int\int p(x,y)\log p(x,y)dxdy \tag{179}$$

$$= -\int\int p(x,y)\log p(x)p(y|x)dxdy \tag{180}$$

$$= -\int\int p(x,y)\log p(x)dxdy - \int\int p(x,y)\log p(y|x)dxdy \tag{181}$$

$$= -\int \log p(x)\int p(x,y)dydx + H[y|x] \tag{182}$$

$$= -\int \log p(x)p(x) + H[y|x] \qquad = H[x] + H[y|x] \tag{183}$$

$$H[x] + H[y] - H[x,y] = H[x] + H[y] - H[x] - H[y|x] \tag{184}$$

$$= H[y] - H[y|x] = I(x;y) \geq 0 \tag{185}$$

1.32

$$H[\boldsymbol{y}] = -\int p(\boldsymbol{y}) \ln p(\boldsymbol{y}) d\boldsymbol{y} \tag{186}$$

$$= -\int p(\boldsymbol{A}\boldsymbol{x}) \ln p(\boldsymbol{A}\boldsymbol{x}) |\frac{d\boldsymbol{y}}{d\boldsymbol{x}}| d\boldsymbol{x} \qquad \text{(Integration by substitution)}$$

$$= -\int \frac{p(\boldsymbol{x})}{|\boldsymbol{A}|} \ln \frac{p(\boldsymbol{x})}{|\boldsymbol{A}|} |\boldsymbol{A}| d\boldsymbol{x} \qquad (p(\boldsymbol{x}) = p(\boldsymbol{y})|\boldsymbol{A}| \text{ for a function } f(\boldsymbol{x}) \to \boldsymbol{x})$$

$$= -\int p(\boldsymbol{x}) \ln \frac{p(\boldsymbol{x})}{|\boldsymbol{A}|} d\boldsymbol{x} \tag{187}$$

$$= -\int p(\boldsymbol{x}) \ln p(\boldsymbol{x}) d\boldsymbol{x} + \int p(\boldsymbol{x}) \ln |\boldsymbol{A}| d\boldsymbol{x} \tag{188}$$

$$= H[\boldsymbol{x}] + \ln |\boldsymbol{A}| \int p(\boldsymbol{x}) d\boldsymbol{x} \tag{189}$$

$$= H[\boldsymbol{x}] + \ln |\boldsymbol{A}| \tag{190}$$

1.33

$$H[y|x] = 0 \tag{191}$$

$$= -\int \int p(x, y) \log p(y|x) dy dx \tag{192}$$

$$= -\int \int p(x, y) \log p(x, y) dy dx + \int \int p(x, y) \log p(x) dx dy \tag{193}$$

$$\Rightarrow H(x) = H(y, x) \tag{194}$$

From this we see that the information gained when we obtain x is equal to the information we gain when we see both x and y. Therefore y must be a function of x. We can also directly see it from the definition. If we use the product rule $p(x, y) = p(y|x)p(x)$ and $\lim_{p(y|x)\to 0} \ln p(y|x) = 0$ we see that the equality will only hold if $p(y|x) = 0$ or $p(y|x) = 1$ if $p(x) > 1$ which is given in the exercise. This, together with the requirement that $p(y|x)$ should integrate to 1 means that the only solution is there's 1 $p(y|x) = 1$ and all the others are $p(y|x) = 0$.

1.34

1.35

$$-H[x] = \int p(x) \ln p(x) dx \tag{195}$$

$$= \int p(x) \ln \frac{1}{\sqrt{2\pi\sigma^2}} dx - \int p(x) \frac{(x-\mu)^2}{2\sigma^2} \tag{196}$$

$$= \int p(x) \ln 1 dx - \int p(x) \ln \sqrt{2\pi\sigma^2} dx - \int p(x) \frac{(x-\mu)^2}{2\sigma^2} \tag{197}$$

$$= -\frac{1}{2} \ln 2\pi\sigma^2 - \frac{1}{2\sigma^2} \int p(x)(x-\mu)^2 dx \qquad \text{(Since } x \text{ is integrated out of } \sigma^2\text{)}$$

$$= -\frac{1}{2} \ln 2\pi\sigma^2 - \frac{\sigma^2}{2\sigma^2} \tag{198}$$

$$= -\frac{1}{2} \ln 2\pi\sigma^2 - \frac{1}{2} \tag{199}$$

$$H[x] = \frac{1}{2} \left[\ln 2\pi\sigma^2 + 1 \right] \tag{200}$$

1.36

For $(a, b) \in \mathbb{R}$ and $b \geq a$, convexity means $f'(b) \geq f'(a)$. Therefore, We have

$$f''(x) = \frac{f'(b) - f'(a)}{b - a} \geq 0$$

1.37

$$-H[x, y] = \int \int p(y, x) \ln p(x, y) dx dy \tag{201}$$

$$= \int \int p(y, x) \ln[p(y|x)p(x)] dx dy \tag{202}$$

$$= \int \int p(y, x) \ln p(y|x) + p(y, x) \ln p(x) dx dy \tag{203}$$

$$= -H[y|x] + \int \int p(y, x) dy \ln p(x) dx \tag{204}$$

$$= -H[y|x] - H[x] \tag{205}$$

1.38

We know from 1.114 that $k = 2$ is true:

$$f(\sum_{i=1}^{M} \lambda_i x_i) \leq \sum_{k=1}^{M} \lambda_i f(x_i) \tag{206}$$

With $\{x_i\} = \{a, b\}$ and λ_1 must be $1 - \lambda$ due to the constraint. So $M = k = 2$, the bsse case, checks out. Now $M = k + 1$:

$$f(\sum_{i=1}^{k+1}\lambda_i x_i) = f(\sum_{i=1}^{k}\lambda_i x_i + \lambda_{k+1}x_{k+1}) \tag{207}$$

$$= f((1-\lambda_{k+1})\frac{1}{1-\lambda_{k+1}}\sum_{i=1}^{K}\lambda_i x_i + \lambda_{k+1}x_{k+1} \tag{208}$$

This is of the form $M = 2$ so we can apply the rule:

$$\leq (1-\lambda_{k+1})f(\frac{1}{1-\lambda_{k+1}}\sum_{i=1}^{K}\lambda_i x_i) + \lambda_{k+1}x_{k+1} \tag{209}$$

$$= (1-\lambda_{k+1})f(\sum_{i=1}^{K}\frac{\lambda_i}{1-\lambda_{k+1}}x_i) + \lambda_{k+1}x_{k+1} \tag{210}$$

Now we have to observe that when $\sum_{i=1}^{K+1}\lambda_i = \sum_{i=1}^{K}\lambda_i + \lambda_{k+1} = 1 \iff \sum_{i=1}^{K}\lambda_i = 1-\lambda_{k+1}$ and then divide both sides by $1-\lambda_{k+1}$ to see that the factor in the sum adds up to one and therefore we can apply the inequality again.

$$\leq (1-\lambda_{k+1})\sum_{i=1}^{K}\frac{\lambda_i}{1-\lambda_{k+1}}f(x_i) + \lambda_{k+1}x_{k+1} \tag{211}$$

$$= \sum_{i=1}^{k+1}\lambda_i f(x_i) \tag{212}$$

1.40

$$\ln\prod_{i=1}^{n}{a_i}^{1/n} = \frac{1}{n}\ln\prod_{i=1}^{n}a_i \tag{213}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\ln a_i \tag{214}$$

$$\geq \ln\frac{1}{n}\sum_{i=1}^{n}a_i \qquad \text{(Jensen)} \tag{215}$$

By monotonicity of logarithm we have $\prod_{i=1}^{n}a^{1/n} \geq \frac{1}{n}\sum_{i=1}^{n}a_i$.

1.41

$$-I[x,y] = \int\int \Big[p(x,y)\ln p(x) + p(x,y)\ln p(y) - p(x,y)\ln p(x|y) - p(x,y)\ln p(y)\Big]dxdy \tag{216}$$

$$= -H[x] + H[x|y] \tag{217}$$

2.1

$$p(x|\mu) = \mu^x(1-\mu)^{1-x} \tag{218}$$

$$\sum_{x=0}^{1} p(x|\mu) = 1 - \mu + \mu = 1 \tag{219}$$

$$\mathbb{E}[x] = \sum_{x=0}^{1} x\mu^x(1-\mu)^{1-x} \tag{220}$$

$$= \mu(1-\mu)^0 \qquad = \mu \tag{221}$$

$$\mathrm{Var}[x] = \left[(x - \mathbb{E}[x])^2\right[\tag{222}$$

$$= \left[(x-\mu)^2\right[\tag{223}$$

$$= \sum_{i=0}^{1}(x-\mu)^2\mu^x(1-\mu)^{1-x} \tag{224}$$

$$\tag{225}$$

Writing out this sum and the resulting squares results in the desired result.

$$H[x] = \sum i = 0^1 p(x)\ln p(x) = \mu\ln\mu + (1-\mu)\ln(1-\mu) \tag{226}$$

2.2

$$\sum_x p(x|\mu) = \sum_x \left(\frac{1-\mu}{2}\right)^{(1-x)/2}\left(\frac{1+\mu}{2}\right)^{(1+x)/2} \tag{227}$$

$$= \frac{1-\mu}{2} + \frac{1+\mu}{2} = 1 \tag{228}$$

where $\mathcal{X} = \{0, 1\}$

$$\mathbb{E}[x] = \sum_x xp(x) = \frac{-1+\mu}{2} + \frac{1+\mu}{2} = \mu \tag{229}$$

$$\mathbb{E}[(x - \mathbb{E}[x])^2] = \mathbb{E}[x^2] - 2\,\mathbb{E}[x\,\mathbb{E}[x]] + \mathbb{E}[\mathbb{E}[x]^2] \tag{230}$$

$$= 1 - 2\mu\,\mathbb{E}[x] + \mu^2 \tag{231}$$

$$= 1 - 2\mu^2 + \mu^2 = 1 - \mu^2 \tag{232}$$

$$H[x] = -\sum_x p(x)\ln p(x) = \frac{1+\mu}{2}\ln\frac{1+\mu}{2} + \frac{1-\mu}{2}\ln\frac{1-\mu}{2} \tag{233}$$

2.3

$$\frac{N!}{(N-m)!m!} + \frac{N!}{(N-(m-1))!(m-1)!} \tag{234}$$

$$= \frac{N!}{(N-m+1)!m!(N-m+1)^{-1})} + \frac{N!}{(N-m+1)!m!m^{-1}} \tag{235}$$

$$= \frac{N!(N-m+1)}{(N-m+1)!m!} + \frac{N!m}{(N-m+1)!m!)} \tag{236}$$

$$= \frac{N!(N+1)}{(N-m+1)!m!} = \frac{(N+1)!}{(N-m+1)!m!} \tag{237}$$

Now the induction part. The general form is:

$$(x+y)^N = \sum_{m=0}^{N} \binom{N}{m} x^{N-m} y^m \tag{238}$$

Base case: $N = 0 \Rightarrow 1 = 1$ is trivially true.

$$(x+y)^{N+1} = \sum_{m=0}^{N+1} \binom{N+1}{m} x^{N+1-m} y^m \tag{239}$$

Continuing with the LHS.

$$(x+y)^{N+1} = (x+y) \sum_{m=0}^{N} \binom{N}{m} x^{N-m} y^m \quad \text{Using the base case} \tag{240}$$

$$= \sum_{m=0}^{N} \binom{N}{m} x^{N-m+1} y^m + \sum_{m=0}^{N} \binom{N}{m} x^{N-m} y^{m+1} \tag{241}$$

We can equalize the exponents by changing the sum range.

$$= \sum_{m=0}^{N} \binom{N}{m} x^{N-m+1} y^m + \sum_{m=1}^{N+1} \binom{N}{m-1} x^{N-(m-1)} y^{(m-1)+1} \tag{242}$$

Now let's take out the terms that they not have in common. I.e., $m = 0$ from the first term and $m = N+1$ from the second.

$$= \binom{N}{0} x^{N-0+1} y^0 + \binom{N}{N} x^{N-N} y^{N+1} + \sum_{m=1}^{N} \binom{N}{m} x^{N-m+1} y^m + \sum_{m=1}^{N} \binom{N}{m-1} x^{N-m+1} y^m \tag{243}$$

$$= x^{N+1} + y^{N+1} + \sum_{m=1}^{N} (x^{N-m+1} y^m) \left(\binom{N}{m} + \binom{N}{m-1} \right) \tag{244}$$

Now we can use our deriven formula.

$$= x^{N+1} + y^{N+1} + \sum_{m=1}^{N} x^{N-m+1} y^m \binom{N+1}{m} \tag{245}$$

Now, see that $x^{N+1} = \binom{N+1}{m} x^{N-m+1} y^m \iff m = 0$ which can be plugged into the sum.

$$= y^{N+1} + \sum_{m=0}^{N} x^{N-m+1} y^m \binom{N+1}{m} \tag{246}$$

Similary, $y^{N+1} = x^{N-m+1} y^m \binom{N+1}{N+1} \iff m = N+1$

$$= \sum_{m=0}^{N+1} x^{N-m+1} y^m \binom{N+1}{m} \tag{247}$$

Which is equal to the RHS.

Now for the normalization. Using the binomial theorem the LHS is $(\mu + 1 - \mu)^N$

$$(\mu + 1 - \mu)^N = 1^N = 1 \tag{248}$$

2.4

$$\sum_{m=0}^{N} \binom{N}{m} \frac{m}{\mu} \mu^m (1-\mu)^{N-m} - \frac{N-m}{1-\mu} (1-\mu)^{N-m} \mu^m = 0 \tag{249}$$

$$\iff \sum_{m=0}^{N} \frac{m}{\mu} \binom{N}{m} \mu^m (1-\mu)^{N-m} = \sum_{m=0}^{N} \frac{N-m}{1-\mu} \binom{N}{m} (1-\mu)^{N-m} \mu^m \tag{250}$$

$$\iff \frac{1}{\mu} \mathbb{E}[m] = \frac{N}{1-\mu} \sum_{m=0}^{N} \binom{N}{m} (1-\mu)^{N-m} \mu^m - \frac{1}{1-\mu} \sum_{m=0}^{N} m \binom{N}{m} (1-\mu)^{N-m} \mu^m \tag{251}$$

$$= \frac{N}{1-\mu} \sum_{m=0}^{N} Bin(m|N,\mu) - \frac{1}{1-\mu} \mathbb{E}[m] \tag{252}$$

$$\iff \mathbb{E}[m] = \frac{\mu N}{1-\mu} \sum_{m=0}^{N} Bin(m|N,\mu) - \frac{\mu}{1-\mu} \mathbb{E}[m] \tag{253}$$

$$\iff 1 = \frac{\mu N}{\mathbb{E}[m](1-\mu)} \sum_{m=0}^{N} Bin(m|N,\mu) - \frac{\mu}{1-\mu} \tag{254}$$

$$\iff \frac{1}{1-\mu} = \frac{\mu N}{\mathbb{E}[m](1-\mu)} \sum_{m=0}^{N} Bin(m|N,\mu) \tag{255}$$

2.4

$$\sum_{m=0}^{N} \binom{N}{m} \frac{m}{\mu} \mu^m (1-\mu)^{N-m} - \frac{N-m}{1-\mu}(1-\mu)^{N-m}\mu^m = 0 \tag{256}$$

$$= \sum_{m=0}^{N} \binom{N}{m} \mu^m (1-\mu)^{N-m} \left(\frac{m}{\mu} - \frac{N-m}{1-\mu} \right) = 0 \tag{257}$$

$$= \sum_{m=0}^{N} \binom{N}{m} \mu^m (1-\mu)^{N-m} \left(\frac{(1-\mu)m}{(1-\mu)\mu} - \frac{\mu(N-m)}{\mu(1-\mu)} \right) = 0 \tag{258}$$

$$= \sum_{m=0}^{N} \binom{N}{m} \mu^m (1-\mu)^{N-m} (m - \mu N) = 0 \tag{259}$$

$$= \mathbb{E}[m] - \mu N \sum_{m=0}^{N} \binom{N}{m} \mu^m (1-\mu)^{N-m} = 0 \tag{260}$$

The sum in the second term is the integral of the binomial distribution (2.264 in the book), and therefore is equal to 1.

2.5

Moving the integral inside

$$\Gamma(a)\Gamma(b) = \int_0^\infty \int_0^\infty \exp(-x) x^{a-1} \exp(-y) y^{b-1} dx dy \tag{261}$$

And using $t = y + x \iff dy = dt$

$$= \int_0^\infty \int_{t=x=0}^\infty \exp(-x) x^{a-1} \exp(x-t)(t-x)^{b-1} dt dx \tag{262}$$

$$= \int_0^\infty \int_{t=x=0}^{x=t=\infty} x^{a-1} \exp(-t)(t-x)^{b-1} dt dx \tag{263}$$

Using the second CoV $x = t\mu \iff dx = d\mu t$

$$= \int_{t=0}^\infty \int_0^1 \exp(-t) t^{a-1} \mu^{a-1} (t - t\mu)^{b-1} t d\mu dt \tag{264}$$

$$= \int_{t=0}^\infty \int_0^1 \exp(-t) t^{a-1} \mu^{a-1} (1-\mu)^{b-1} t^{b-1} t d\mu dt \tag{265}$$

$$= \int_{t=0}^\infty \exp(-t) t^{a-1} t^{b-1} t dt \int_0^1 \mu^{a-1} (1-\mu)^{b-1} d\mu \tag{266}$$

$$= \Gamma(a+b) \int_0^1 \mu^{a-1} (1-\mu)^{b-1} d\mu \tag{267}$$

$$\tag{268}$$

2.6

$$p(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \tag{269}$$

$$\int_0^1 \mu^{a-1}(1-\mu)^{b-1}d\mu = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \tag{270}$$

$$\Gamma(x) = \int_0^\infty \mu^{x-1}e^{-u}du \tag{271}$$

$$\mathbb{E}[\mu] = \int_0^1 p(\mu|a,b)\mu d\mu \tag{272}$$

$$= \int_0^1 \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}\mu d\mu \tag{273}$$

$$= \int_0^1 \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a}(1-\mu)^{b-1}d\mu \tag{274}$$

Since added 1 in the exponent this now looks like a beta distribution with $p(\mu, a+1, b)$

$$. = \frac{a}{a+b}\int_0^1 \frac{\Gamma(a+1+b)}{\Gamma(a+1)\Gamma(b)}\mu^{a}(1-\mu)^{b-1}d\mu \tag{275}$$

$$= \frac{a}{a+b} \tag{276}$$

Where we used

$$\Gamma(a+1) = a\Gamma(a) \text{ and } \Gamma(a+b+1) = (a+b)\Gamma(a+b) \tag{277}$$

$$\text{Var}[\mu] = \mathbb{E}[\mu^2] - \mathbb{E}[\mu]^2 \tag{278}$$

$$\mathbb{E}[\mu] = \int_0^1 p(\mu|a,b)\mu d\mu^2 \tag{279}$$

$$= \int_0^1 \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a+1}(1-\mu)^{b-1}\mu d\mu \tag{280}$$

Which looks like Beta with a+2

$$= \frac{a}{a+b}\int_0^1 \frac{\Gamma(a+1+b)}{\Gamma(a+1)\Gamma(b)}\mu^{a+2}(1-\mu)^{b-1}d\mu \tag{281}$$

$$= \frac{a(a+1)}{(a+b)(a+1+b)}\int_0^1 \frac{\Gamma(a+2+b)}{\Gamma(a+2)\Gamma(b)}\mu^{a+2}(1-\mu)^{b-1}d\mu \tag{282}$$

$$= \frac{a(a+1)}{(a+b)(a+1+b)} \tag{283}$$

$$\text{Var}[\mu] = \frac{a(a+1)}{(a+b)(a+1+b)} - \frac{a^2}{(a+b)^2} \tag{284}$$

$$= \frac{(a+b)a(a+1)}{(a+b)^2(a+1+b)} - \frac{(a+b+1)a^2}{(a+b+1)(a+b)^2} \tag{285}$$

$$= \frac{(a^2+ab)(a+1)}{(a+b)^2(a+1+b)} - \frac{(a+b+1)a^2}{(a+b+1)(a+b)^2} \tag{286}$$

$$= \frac{a^3+a^2b+a^2+ab}{(a+b)^2(a+1+b)} - \frac{a^3+ba^2+a^2}{(a+b+1)(a+b)^2} \tag{287}$$

Which gives you the desired result.

Mode is given where the derivative w.r.t. μ is zero:

$$\frac{\partial p(\mu|a,b)}{\partial \mu} = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\left[(a-1)\mu^{a-2} - \mu^{a-1}(b-1)(1-\mu)^{b-2}\right] = 0 \tag{288}$$

$$(a-1)\mu^{a-2}(1-\mu)^{b-1} = \mu^{a-1}(b-1)(1-\mu)^{b-2} \tag{289}$$

$$\frac{a-1}{\mu} = \frac{b-1}{1-\mu} \tag{290}$$

$$\mu = \frac{-a+1}{-a-b+2} = \frac{a-1}{a+b-2} \tag{291}$$

2.7

The ML estimate is $\mu_{ML} = \frac{m}{N} = \frac{m}{N}$ and the prior mean is $\frac{a}{a+b}$. The posterior is $p(x=1|D) = \frac{m+a}{m+a+n+b}$.

Following the tip in the exercise.

$$\lambda(\frac{a}{a+b}) + (1-\lambda)\frac{m}{m+n} = \frac{m+a}{m+a+n+b} \tag{292}$$

Solving for lambda gives

$$\lambda = \frac{1}{1+(n+m)/(a+b)} \tag{293}$$

Which must lie in (0,1).

2.8

$$\mathbb{E}[x] = \int p(x)xdx \tag{294}$$
$$= \int\int p(x,y)xdxdy \tag{295}$$
$$= \int\int p(x|y)p(y)xdxdy \tag{296}$$
$$= \int p(y)\int p(x|y)xdxdy \tag{297}$$
$$= \int p(y)\,\mathbb{E}_x[x|y]dy \tag{298}$$
$$= \mathbb{E}_y\,\mathbb{E}_x[x|y] \tag{299}$$

$$\mathbb{E}_y[\mathrm{Var}_x[x|y]] = \mathbb{E}_y\,\mathbb{E}_{x|y}[(x-\mathbb{E}[x|y])^2] \tag{300}$$
$$= \mathbb{E}_y\,\mathbb{E}_{x|y}[(x^2+\mathbb{E}[x|y]^2-2x\,\mathbb{E}[x|y])] \tag{301}$$
$$= \mathbb{E}_y\,\mathbb{E}_{x|y}\,x^2+\mathbb{E}_y\,\mathbb{E}_{x|y}\,\mathbb{E}[x|y]^2-2\,\mathbb{E}_y\,\mathbb{E}_{x|y}\,x\,\mathbb{E}[x|y])] \tag{302}$$
$$= \mathbb{E}_y\int p(x|y)x^2dx+\mathbb{E}_y\,\mathbb{E}[x|y]^2-2\int p(y)\int p(x|y)x\,\mathbb{E}[x|y]dxdy \tag{303}$$
$$= \int p(y)\int p(x|y)x^2dxdy+\mathbb{E}_y\,\mathbb{E}[x|y]^2-2\int \mathbb{E}[x|y]\int p(x|y)xdxdy \tag{304}$$
$$= \int\int p(x,y)x^2dxdy+\mathbb{E}_y\,\mathbb{E}[x|y]^2-2\int \mathbb{E}[x|y]^2dy)] \tag{305}$$
$$= \int\int p(x,y)x^2dxdy+\mathbb{E}_y\,\mathbb{E}[x|y]^2-2\,\mathbb{E}_y\,\mathbb{E}[x|y]^2)] \tag{306}$$
$$= \mathbb{E}[x^2]-\mathbb{E}_y\,\mathbb{E}[x|y]^2 \tag{307}$$
$$\tag{308}$$

$$\mathrm{Var}_y[\mathbb{E}_x[x|y]] = \mathbb{E}_y\left[(\mathbb{E}_x[x|y]-\mathbb{E}_y\,\mathbb{E}_x[x|y])^2\right] \tag{309}$$
$$= \mathbb{E}_y\left[(\mathbb{E}_x[x|y]-\mathbb{E}_x[x])^2\right] \tag{310}$$
$$= \mathbb{E}_y\left[(\mathbb{E}_x[x|y]^2+\mathbb{E}_x[x]^2-2\,\mathbb{E}_x[x]\,\mathbb{E}_x[x|y]\right] \tag{311}$$
$$= \mathbb{E}_y[\mathbb{E}_x[x|y]^2]+\mathbb{E}_y[\mathbb{E}_x[x]^2]-2\,\mathbb{E}_y\,\mathbb{E}_x[x]\,\mathbb{E}_x[x|y] \tag{312}$$
$$= \mathbb{E}_y[\mathbb{E}_x[x|y]^2]+\mathbb{E}_x[x]^2-2\,\mathbb{E}_x[x]\,\mathbb{E}_y\,\mathbb{E}_x[x|y] \tag{313}$$
$$= \mathbb{E}_y[\mathbb{E}_x[x|y]^2]+\mathbb{E}_x[x]^2-2\,\mathbb{E}_x[x]^2 \tag{314}$$
$$= \mathbb{E}_y[\mathbb{E}_x[x|y]^2]-\mathbb{E}_x[x]^2 \tag{315}$$
$$\tag{316}$$

Sum of these expressions gives desired result.

2.10

$$\mathbb{E}[\mu_j] = \int \mu_j \frac{\Gamma(a_0)}{\Gamma(a_1)\dots\Gamma(a_K)} \prod_{k=1}^{K} \mu_k^{a_k-1} d\boldsymbol{\mu} \tag{317}$$

$$= \frac{\Gamma(a_0)}{\Gamma(a_1)\dots\Gamma(a_K)} \int \prod_{k=1}^{j-1} \mu_k^{a_k-1} \mu_j^{a_j} \prod_{k=j+1}^{K} \mu_k^{a_k-1} d\boldsymbol{\mu} \tag{318}$$

Now we use $a_j := b_j - 1$ and $a_k := b_k$, then we get $b_0 = \sum b_k = \sum a_k + 1 = a_0 + 1$

$$= \frac{1/a_0\Gamma(a_0+1)}{1/a_j\Gamma(a_1)\dots\Gamma(a_j+1)\dots\Gamma(a_k)} \tag{319}$$

$$= \frac{a_j}{a_0} \frac{\Gamma(b_0)}{\Gamma(b_0)\dots\Gamma(b_K)} \int \prod_{k=1}^{K} \mu_k^{b_k-1} d\boldsymbol{\mu} \tag{320}$$

$$= \frac{a_j}{a_0} \int \mathrm{Dir}(\boldsymbol{\mu}, \boldsymbol{b}) d\boldsymbol{\mu} \tag{321}$$

$$= \frac{a_j}{a_0} \tag{322}$$

$$var[\mu_j] = \mathbb{E}[(\mu_j - \mathbb{E}[\mu_j])^2] \tag{323}$$

$$= \mathbb{E}[\mu^2] - \mathbb{E}[\mu]^2 \tag{324}$$

$$= \mathbb{E}[\mu^2] - \left(\frac{a_j}{a_0}\right)^2 \tag{325}$$

$$\mathbb{E}[\mu^2] = \int \mu_j^2 \frac{\Gamma(a_0)}{\Gamma(a_1)\dots\Gamma(a_K)} \prod_{k=1}^{K} \mu^{a_k-1} d\boldsymbol{\mu} \tag{326}$$

$$= \frac{\Gamma(a_0)}{\Gamma(a_1)\dots\Gamma(a_K)} \int \prod_{k=1}^{j-1} \mu_k^{a_k-1} \mu^{a_j+1} \prod_{k=j+1}^{K} \mu_k^{a_k-1} d\boldsymbol{\mu} \tag{327}$$

We apply the same trick: $a_j + 1 = b_j - 1$ and $a_k = b_k$, thus $b_0 = \sum b_k = \sum a_k + 2 = a_0 + 2$

$$= \frac{1/a_0\Gamma(a_0+1)}{1/a_j\Gamma(a_1)\dots\Gamma(a_j+1)\dots\Gamma(a_k)} \int \prod_{k=1}^{j-1} \mu_k^{a_k-1} \mu^{a_j+1} \prod_{k=j+1}^{K} \mu_k^{a_k-1} d\boldsymbol{\mu} \tag{328}$$

$$= \frac{\frac{1}{a_0(a_0+1)}\Gamma(a_0+2)}{\frac{1}{a_j(a_j+1)}\Gamma(a_1)\dots\Gamma(a_j+2)\dots\Gamma(a_k)} \int \prod_{k=1}^{j-1} \mu_k^{a_k-1} \mu^{a_j+1} \prod_{k=j+1}^{K} \mu_k^{a_k-1} d\boldsymbol{\mu} \tag{329}$$

$$= \frac{a_j(a_j+1)}{a_0(a_0+1)} \int Dir(\boldsymbol{\mu}, \boldsymbol{b}) d\mu \tag{330}$$

$$= \frac{a_j(a_j+1)}{a_0(a_0+1)} \tag{331}$$

$$\begin{aligned}
\mathrm{Var}[\mu_j] &= \frac{a_j(a_j+1)}{a_0(a_0+1)} - \frac{a_j^2}{a_0^2} && (332)\\
&= \frac{a_0^2 a_j(a_j+1)}{a_0^2(a_0^2+a_0)} - \frac{a_j^2(a_0^2+a_0)}{a_0^2(a_0^2+a_0)} && (333)\\
&= \frac{a_0^2 a_j(a_j+1) - a_j^2(a_0^2+a_0)}{a_0^2(a_0^2+a_0)} && (334)\\
&= \frac{a_0^2 a_j^2 + a_0^2 a_j - a_j^2 a_0^2 - a_j^2 a_0}{a_0^2(a_0^2+a_0)} && (335)\\
&= \frac{a_0^2 a_j - a_j^2 a_0}{a_0^2(a_0^2+a_0)} && (336)\\
&= \frac{a_0 a_j(a_0 - a_j)}{a_0^2(a_0^2+a_0)} && (337)\\
&= \frac{a_j(a_0 - a_j)}{a_0(a_0^2+a_0)} && (338)\\
& && (339)
\end{aligned}$$

2.11

2.12

$$\int_a^b \frac{1}{b-a} dx = \frac{x}{b-a} - \frac{x}{b-a}\Big|_a^b = \frac{b-a}{b-a} = 1 \tag{340}$$

$$\begin{aligned}
\mathbb{E}[x] &= \int_a^b p(x) x dx && (341)\\
&= \int_a^b \frac{x}{b-a} dx && (342)\\
&= \frac{x^2}{2(b-a)}\Big|_a^b && (343)\\
&= \frac{b^2-a^2}{2(b-a)} = \frac{(b-a)(b+a)}{2(b-a)} = \frac{b+a}{2} && (344)
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}[x] &= \mathbb{E}[x^2] - \mathbb{E}[x]^2 && (345)\\
&= \int_a^b \frac{x^2}{b-a} dx - \frac{(b+a)^2}{4} && (346)\\
&= \frac{\frac{1}{3}x^3}{b-a}\Big|_a^b - \frac{(b+a)^2}{4} && (347)\\
&= \frac{\frac{1}{3}b^3}{b-a} - \frac{\frac{1}{3}a^3}{b-a} - \frac{(b+a)^2}{4} && (348)\\
&= \frac{(b-a)(b^2+a^2+ab)}{3(b-a)} - \frac{(b+a)^2}{4} && (349)\\
&= \frac{b^2+a^2+ab}{3} - \frac{b^2+a^2+2ab}{4} && (350)\\
&= \frac{4b^2+4a^2+4ab}{12} - \frac{3b^2+3a^2+6ab}{12} && (351)\\
&= \frac{b^2+a^2-2ab}{12} && (352)\\
&= \frac{(b-a)^2}{12} && (353)\\
& && (354)
\end{aligned}$$

2.13

$$\begin{aligned}
KL[p||q] &= \int p(\boldsymbol{x}) \ln \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} && (355)\\
&= \int p(\boldsymbol{x}) \ln \frac{|\boldsymbol{\Sigma}|^{-1/2} \exp[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{|\boldsymbol{L}|^{-1/2} \exp[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})^T \boldsymbol{L}^{-1}(\boldsymbol{x}-\boldsymbol{m})} && (356)\\
&= \int p(\boldsymbol{x}) \ln \frac{|\boldsymbol{L}|^{1/2}}{|\boldsymbol{\Sigma}|^{1/2}} + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})^T \boldsymbol{L}^{-1}(\boldsymbol{x}-\boldsymbol{m}) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) && (357)\\
&= \ln \frac{|\boldsymbol{L}|^{1/2}}{|\boldsymbol{\Sigma}|^{1/2}} + \frac{1}{2}\int p(\boldsymbol{x})(\boldsymbol{x}-\boldsymbol{m})^T \boldsymbol{L}^{-1}(\boldsymbol{x}-\boldsymbol{m}) - \frac{1}{2}\int p(\boldsymbol{x})(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) && (358)\\
&= \ln \frac{|\boldsymbol{L}|^{1/2}}{|\boldsymbol{\Sigma}|^{1/2}} + \frac{1}{2}\,\mathbb{E}[(\boldsymbol{x}-m)^T \boldsymbol{L}^{-1}(\boldsymbol{x}-\boldsymbol{m})] - \frac{1}{2}\,\mathbb{E}[(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})] && (359)
\end{aligned}$$

Now we use Matrix Cookook (15-11-2012) equation 318

$$\begin{aligned}
&= \ln \frac{|\boldsymbol{L}|^{1/2}}{|\boldsymbol{\Sigma}|^{1/2}} + \frac{1}{2}(\mathrm{Tr}(\boldsymbol{L}^{-1}\boldsymbol{L}) + \mathbb{E}[\boldsymbol{x}]^T \boldsymbol{L}^{-1}\,\mathbb{E}[\boldsymbol{x}]) - \frac{1}{2}(\mathrm{Tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}) + \mathbb{E}[\boldsymbol{x}]^T \boldsymbol{\Sigma}^{-1}\,\mathbb{E}[\boldsymbol{x}]) && (360)\\
&= \ln \frac{|\boldsymbol{L}|^{1/2}}{|\boldsymbol{\Sigma}|^{1/2}} + \boldsymbol{m}^T \boldsymbol{L}^{-1}\boldsymbol{m} + \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} && (361)
\end{aligned}$$

2.14

I don't see where the Tr comes from in the solution.

2.15

$$H[\boldsymbol{x}] = -\int p(\boldsymbol{x}) \ln p(\boldsymbol{x}) d\boldsymbol{x} \tag{362}$$

$$= -\int p(\boldsymbol{x}) \left[-\ln \left[(2\pi)^{D/2} \right] - \ln \left[|\boldsymbol{\Sigma}|^{1/2} \right] - \frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right] \tag{363}$$

$$\tag{364}$$

$$\int p(\boldsymbol{x}) \ln \left[(2\pi)^{D/2} \right]] = \frac{D}{2} \ln 2\pi \tag{365}$$

$$\int p(\boldsymbol{x}) \ln \left[|\Sigma|^{1/2} \right] = \frac{1}{2} \int p(\boldsymbol{x}) \ln |\boldsymbol{\Sigma}| = \frac{1}{2} \ln |\boldsymbol{\Sigma}| \tag{366}$$

$$\frac{1}{2} \int p(\boldsymbol{x}) (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) = \frac{1}{2} \mathbb{E} \left[(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right] \tag{367}$$

$$= \frac{1}{2} \mathrm{Tr}\big(\Sigma^{-1} \Sigma\big) + \frac{1}{2} \mathbb{E}[\boldsymbol{x} - \mu]]^T \Sigma^{-1} \mathbb{E}[\boldsymbol{x} - \boldsymbol{\mu}] \tag{368}$$

$$= \frac{1}{2} D + \frac{1}{2} (\boldsymbol{\mu} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}) \tag{369}$$

$$= \frac{D}{2} \tag{370}$$

Adding the terms gives the desired result.

2.17

$$A = \frac{1}{2} \underbrace{A + A^T}_{Symmetric} + \frac{1}{2} \underbrace{A - A^T}_{Non-symmetric} \tag{371}$$

$$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto \exp -\frac{1}{2} (x - \mu)^T \frac{1}{2} (\Sigma^{-1} + \Sigma^{-T} + \Sigma^{-1} - \Sigma^{-T}) (x - \mu) \tag{372}$$

$$= \exp -\frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{M} (x_n - \mu_n) (\Sigma_{nm}^{-1} + \Sigma_{nm}^{-T}) (x_m - \mu_m) \tag{373}$$

$$- \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{M} (x_n - \mu_n) \underbrace{(\Sigma_{nm}^{-1} - \Sigma_{nm}^{-T})}_{=0} (x_m - \mu_m) \tag{374}$$

Which leaves us only with a symmetric covariance matrix.

2.19

From (2.45) we have

$$\boldsymbol{\Sigma}\boldsymbol{u}_i = \lambda_i \boldsymbol{u}_i \tag{375}$$
$$\Rightarrow \boldsymbol{u}_i^T \boldsymbol{\Sigma}\boldsymbol{u}_i = \boldsymbol{u}_i^T \lambda_i \boldsymbol{u}_i \tag{376}$$
$$\Rightarrow \boldsymbol{U}^T \boldsymbol{\Sigma}\boldsymbol{U} = \boldsymbol{U}^T \boldsymbol{\Lambda}\boldsymbol{U} \tag{377}$$
$$\Rightarrow \boldsymbol{\Sigma} = \boldsymbol{\Lambda} \tag{378}$$

Therefore we have

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda} = \boldsymbol{\Lambda}\boldsymbol{U}\boldsymbol{U}^T = \sum_i \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^T \tag{379}$$

$$\boldsymbol{\Sigma}^{-1} = (\boldsymbol{\Lambda}\boldsymbol{U}\boldsymbol{U}^T)^{-1} = \boldsymbol{U}\boldsymbol{U}^T\boldsymbol{\Lambda}^{-1} = \sum_i \frac{1}{\lambda_i} \boldsymbol{u}_i \boldsymbol{u}_i^T \tag{380}$$

2.20

Since the eigenvectors of $\boldsymbol{\Sigma}$ form an orthonormal set, they form a basis of $\mathbb{R}^D$, and therefore $\boldsymbol{a} = \sum_i a_i \boldsymbol{u}_i$

$$\boldsymbol{a}^T \boldsymbol{\Sigma}\boldsymbol{a} = (\sum_i a_i \boldsymbol{u}_i)^T \boldsymbol{\Sigma} (\sum_i a_i \boldsymbol{u}_i) \tag{381}$$
$$= (\sum_i a_i \boldsymbol{u}_i)^T \sum_i a_i \boldsymbol{\Sigma}\boldsymbol{u}_i \tag{382}$$
$$= (\sum_i a_i \boldsymbol{u}_i)^T \sum_i a_i \lambda_i \boldsymbol{u}_i \tag{383}$$
$$\tag{384}$$

Example:

$$(a_1\boldsymbol{u}_1 + a_2\boldsymbol{u}_2)^T (a_1\lambda_1\boldsymbol{u}_1 + a_2\lambda_2\boldsymbol{u}_2) = a_1^2\lambda_1\boldsymbol{u}_1^T\boldsymbol{u}_1 + a_1a_2\lambda_2\boldsymbol{u}_1^T\boldsymbol{u}_2 + a_2a_1\lambda_1\boldsymbol{u}_2^T\boldsymbol{u}_1 + a_2^2\lambda_2\boldsymbol{u}_2^T\boldsymbol{u}_2 \tag{385}$$
$$= a_1^2\lambda_1 + a_2^2\lambda_2 \tag{386}$$

Since $\boldsymbol{u}_i^T\boldsymbol{u}_j = 0$ for $i \neq j$

So the cross terms vanish, and therefore we are left with an expression that will be positive for all $\boldsymbol{a}$ only if $\lambda_i > 0$.

2.21

Symmetry means that if the lower triangular half of the matrix is defined, the other half is defined too. That means that we have to count the amount of entries in a triangular matrix of size D.

$$\#\text{parameters} = D + (D-1) + (D-2) + \cdots + D - (D-2) + D - (D-1) \tag{387}$$
$$= (D+1) + (D+1) + \cdots + (D+1) \tag{388}$$
$$= \frac{D}{2}(D+1) \tag{389}$$

2.22

$$A^{-1}A = A^{-1}A^T \quad \text{(Symmetry)}$$

$$A^{-1}AA^{T\,-1} = A^{-1}A^TA^{T\,-1} \tag{390}$$

$$A^{-1\,T} = A^{-1}A^TA^{T\,-1} \tag{391}$$

$$= A^{-1} \tag{392}$$

2.23

2.24

The left side will return the identity matrix, but it's not trivial to show as block matrix inversion cannot be done block by block.

The RHS can be multiplied block by block.

Upper left:

$$AM - BD^{-1}CM = (A - BD^{-1}C)M = I \tag{393}$$

Upper right:

$$-AMBD^{-1} + BD^{-1} + BD^{-1}CMBD^{-1} \tag{394}$$

$$= (-AM + I + BD^{-1}CM)BD^{-1} \tag{395}$$

$$= (-AM + AM - BD^{-1}CM + BD^{-1}CM)BD^{-1} \qquad = 0 \tag{396}$$

Bottom left:

$$CM - DD^{-1}CM = 0 \tag{397}$$

Bottom right:

$$-CMBD^{-1} + D(D^{-1} + D^{-1}CMBD^{-1}) \tag{398}$$

$$= -CMBD^{-1} + I + CMBD^{-1} = I \tag{399}$$

2.25

We have $p(x_a|x_b) = \frac{1}{p(x_b)} \int p(x_a, x_b, x_c)dx_c$ From section 2.3.2 we know that after marginalizing out we have $\mathbb{E}\begin{bmatrix}\boldsymbol{x}_a \\ \boldsymbol{x}_b\end{bmatrix} = \begin{pmatrix}\boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b\end{pmatrix}$

And then by section 2.3.1 we have $\boldsymbol{\mu}_{a|b}$ equal to (2.81) and $\boldsymbol{\Sigma}_{a|b}$ is equal to (2.82).

2.26

The LHS trivially becomes the identity matrix.

$$(A + BCD)(A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1}) \tag{400}$$
$$= I - B(C^{-1} + DA^{-1}B)^{-1}DA^{-1} + BCDA^{-1} - BCDA^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1} \tag{401}$$
$$\tag{402}$$

$$BCDA^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1} \tag{403}$$
$$= BC(C^{-1} - C^{-1} + DA^{-1}B)(C^{-1} + DA^{-1}B)^{-1}DA^{-1} \tag{404}$$
$$= BC(C^{-1} + DA^{-1}B)^{-1}DA^{-1} - BC(C^{-1} + DA^{-1}B)(C^{-1} + DA^{-1}B)^{-1}DA^{-1} \tag{405}$$
$$= BC(C^{-1} + DA^{-1}B)^{-1}DA^{-1} - BCDA^{-1} \tag{406}$$

Plugging this into the previous result gives I.

2.27

$$\mathbb{E}[\boldsymbol{y}] = \mathbb{E}[\boldsymbol{x} + \boldsymbol{z}] \tag{407}$$
$$= \int\int p(\boldsymbol{x}, \boldsymbol{z})(\boldsymbol{x} + \boldsymbol{z})d\boldsymbol{x}d\boldsymbol{z} \tag{408}$$
$$= \int\int p(\boldsymbol{x}, \boldsymbol{z})\boldsymbol{x} + p(\boldsymbol{x}, \boldsymbol{z})\boldsymbol{z}d\boldsymbol{x}d\boldsymbol{z} \tag{409}$$
$$= \int\int p(\boldsymbol{x})p(\boldsymbol{z})\boldsymbol{x} + p(\boldsymbol{z})p(\boldsymbol{x})\boldsymbol{z}d\boldsymbol{x}d\boldsymbol{z} \tag{410}$$
$$= \int p(\boldsymbol{z}) \int p(\boldsymbol{x})\boldsymbol{x}d\boldsymbol{x}d\boldsymbol{z} + \int p(\boldsymbol{x}) \int p(\boldsymbol{z})\boldsymbol{z}d\boldsymbol{x}d\boldsymbol{z} \tag{411}$$
$$= \mathbb{E}[\boldsymbol{x}] + \mathbb{E}[\boldsymbol{z}] \tag{412}$$

2.29

It's quite straightforward.

$$M = (A - BD^{-1}C)^{-1} = (\Lambda + A^T LA - ALL^{-1}LA)^{-1} \tag{413}$$
$$= (\Lambda + A^T LA - A^T LA)^{-1} \tag{414}$$
$$= \Lambda^{-1} \tag{415}$$

$$-MBD^{-1} = \Lambda^{-1}A^T LL^{-1} = \Lambda^{-1}A^T \tag{416}$$

$$D^{-1}CM = -L^{-1}LA^{\Lambda^{-1}} = A\lambda^{-1} \tag{417}$$

$$D^{-1} + D^{-1}CMBD^{-1} = L^{-1} + L^{-1}LA\Lambda^{-1}A^TLL^{-1} = L^{-1} + A\Lambda^{-1}A^T \tag{418}$$

2.30

$$\mathbb{E}[z] = R^{-1}\begin{pmatrix}\Lambda\mu - A^TLb\\ Lb\end{pmatrix} \tag{419}$$

$$= \begin{pmatrix}\Lambda^{-1} & \Lambda^{-1}A^T\\ A\Lambda^{-1} & L^{-1} + A\Lambda^{-1}A^T\end{pmatrix}\begin{pmatrix}\Lambda\mu - A^TLb\\ Lb\end{pmatrix} \tag{420}$$

$$= \begin{pmatrix}\Lambda^{-1}(\Lambda\mu - A^TLb) + \Lambda^{-1}A^TLb\\ A\Lambda^{-1}(\Lambda\mu - A^TLb) + \mathcal{L}^{-1}Lb + A\Lambda^{-1}A^TLb\end{pmatrix} \tag{421}$$

$$= \begin{pmatrix}\mu - \Lambda^{-1}A^TLb + \Lambda^{-1}A^TLb\\ A\mu - A\Lambda^{-1}A^TLb + b + A\Lambda^{-1}A^TLb\end{pmatrix} \tag{422}$$

Which gives the desired result.

2.31

You plug in the values according to equation 2.115.

2.34

The derivative splits into two terms. The first one:

$$-\frac{N}{2}\frac{d}{d\Sigma}\ln|\boldsymbol{\Sigma}| = -\frac{N}{2}\boldsymbol{\Sigma}^{-T} = \frac{N}{2}\boldsymbol{\Sigma}^{-1} \tag{423}$$

The second part:

$$-\frac{1}{2}\sum_{n=1}^{N}\frac{d}{d\boldsymbol{\Sigma}}(\boldsymbol{x}_n - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}) \tag{424}$$

By Matrix Cookbook November 15, 2012 eq. 61 we get

$$= \frac{1}{2}\sum_{n=1}^{N}\boldsymbol{\Sigma}^{-T}(\boldsymbol{x}_n - \boldsymbol{\mu})(\boldsymbol{x}_n - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-T} \tag{425}$$

$$= \frac{1}{2}\sum_{n=1}^{N}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu})(\boldsymbol{x}_n - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1} \tag{426}$$

$$= \frac{N}{2}\boldsymbol{\Sigma}^{-1}\left(\frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu})(\boldsymbol{x}_n - \boldsymbol{\mu})^T\right)\boldsymbol{\Sigma}^{-1} \tag{427}$$

$$\tag{428}$$

Setting the sum of the terms to zero gives

$$\frac{N}{2}\mathbf{\Sigma}^{-1}\left(\frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n-\boldsymbol{\mu})(\boldsymbol{x}_n-\boldsymbol{\mu})^T\right)\mathbf{\Sigma}^{-1}=\frac{N}{2}\mathbf{\Sigma}^{-1} \tag{429}$$

$$\iff \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n-\boldsymbol{\mu})(\boldsymbol{x}_n-\boldsymbol{\mu})^T=\mathbf{\Sigma} \tag{430}$$

2.35

First we use

$$xx^T=(x-\mathbb{E}[x])(x-\mathbb{E}[x])^T-\mathbb{E}[x]\,\mathbb{E}[x]^T+x\,\mathbb{E}[x]^T+\mathbb{E}[x]x^T \tag{431}$$

Then we have

$$\mathbb{E}[xx^T]=\Sigma-\mu\mu^T+\mathbb{E}[x]\mu^T+\mu\,\mathbb{E}[x]^T=\Sigma+\mu\mu^T \tag{432}$$

The result 2.291 follows from the proof in 2.61 where we see that Σ only appear when $i=j$.

$$\mathbb{E}[\Sigma_{ML}]=\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[(\boldsymbol{x}_n-\boldsymbol{\mu}_{ML})(\boldsymbol{x}_n-\boldsymbol{\mu}_{ML})^T] \tag{433}$$

$$=\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[(\boldsymbol{x}_n-\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n)(\boldsymbol{x}_n-\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n)^T] \tag{434}$$

$$=\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[\boldsymbol{x}_n\boldsymbol{x}_n^T-\boldsymbol{x}_n\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n^T-\left(\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n\right)\boldsymbol{x}_n^T+\frac{1}{N^2}\sum_{n=1}^{N}\boldsymbol{x}_n\boldsymbol{x}_n^T] \tag{435}$$

$$=\mu\mu^T+\Sigma-\frac{2}{N^2}\sum_{n=1}^{N}\sum_{n=1}^{N}\mathbb{E}[\boldsymbol{x}_n\boldsymbol{x}_n^T]+\frac{1}{N}(\mu\mu^T+\Sigma) \tag{436}$$

$$=\mu\mu^T+\Sigma-2\mu\mu^T-\frac{2}{N}\Sigma+\mu\mu^T+\frac{1}{N}\Sigma \tag{437}$$

From which the result follows. We get $\frac{1}{N}$ before the variances since it only appears when $i=j$ in the sum.

2.36

$$\sigma_{ML}^{2(N)}=\frac{1}{N}\sum_{n=1}^{N}(x_n-\mu^{(N)})^2 \tag{438}$$

$$=\frac{1}{N}(x_n-\mu^{(N)})^2+\frac{1}{N}\sum_{n=1}^{N}(x_n-\mu^{(N)})^2 \tag{439}$$

$$=\frac{1}{N}(x_n-\mu^{(N)})^2+\frac{N-1}{N}\sigma_{ML}^{2(N-1)} \tag{440}$$

$$=\sigma_{ML}^{2(ML)}+\frac{1}{N}\left((x_n-\mu^{(N)})^2-\sigma_{ML}^{2(N-1)}\right) \tag{441}$$

2.37

$$\boldsymbol{\Sigma}_{ML} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu}_{ML})(\boldsymbol{x}_n - \boldsymbol{\mu}_{ML})^T \tag{442}$$

$$= \frac{1}{N}(\boldsymbol{x}_n - \boldsymbol{\mu}_{ML})(\boldsymbol{x}_n - \boldsymbol{\mu}_{ML})^T + \frac{1}{N}\sum_{n=1}^{N-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_{ML})(\boldsymbol{x}_n - \boldsymbol{\mu}_{ML})^T \tag{443}$$

$$= \frac{1}{N}(\boldsymbol{x}_n - \boldsymbol{\mu}_{ML})(\boldsymbol{x}_n - \boldsymbol{\mu}_{ML})^T + \frac{N-1}{N}\boldsymbol{\Sigma}_{ML}^{(N-1)} \tag{444}$$

$$= \boldsymbol{\Sigma}_{ML}^{(N-1)} + \frac{1}{N}\left((\boldsymbol{x}_n - \boldsymbol{\mu}_{ML})(\boldsymbol{x}_n - \boldsymbol{\mu}_{ML})^T - \boldsymbol{\Sigma}_{ML}^{(N-1)}\right) \tag{445}$$

2.39

2.40

The easiest way is to deduce the parameters by examining the proportionality of the log-posterior. The exponent becomes

$$\sum_{n=1}^{N} -\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu) - \frac{1}{2}(\mu-\mu_0)^T\Sigma_0^{-1}(\mu-\mu_0) \tag{446}$$

We multiply everything through to get

$$= -\frac{N}{2}\mu^T\Sigma^{-1} - \frac{1}{2}\sum_{i}^{N} x^T\Sigma^{-1}x + \sum_{i}^{N} x^T\Sigma^{-1}\mu - \frac{1}{2}\mu^T\Sigma_0^{-1}\mu + \mu^T\Sigma_0^{-1}\mu_0 - \frac{1}{2}\mu_0^T\Sigma_0^{-1}\mu_0 \tag{447}$$

Analogously to 2.71 and the text thereafter we exclude the terms not depending on μ and group them together

$$= -\frac{1}{2}\mu^T(N\Sigma^{-1} + \Sigma_0^{-1})\mu + \mu^T(\Sigma_0^{-1}\mu_0 + \Sigma^{-1}\sum_{i}^{N} x_i) + C \tag{448}$$

Now we use the completing-the-square formula $x^TAx. + x^Tb + c = (x-h)^TA(x-h) + k$ with $h = \frac{1}{2}A^{-1}b$ to see that the mean of the posterior becomes

$$\tag{449}$$

$$h = -\frac{1}{2}A^{-1}b = (N\Sigma^{-1} + \Sigma_0^{-1})^{-1}(\Sigma_0^{-1}\mu_0 + \Sigma^{-1}\sum_{i} x_i) \tag{450}$$

and the variance

$$A = N\Sigma^{-1} + \Sigma_0^{-1} \tag{451}$$

2.41

$$\Gamma(x) = \int_0^\infty u^{x-1} \exp(-u) du \tag{452}$$

$$\int_0^\infty \text{Gam}(\lambda|a,b) d\lambda = \int_0^\infty \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda) d\lambda \tag{453}$$

$$= \frac{1}{\Gamma(a)} b^a \int_0^\infty \lambda^{a-1} \exp(-b\lambda) d\lambda \tag{454}$$

$$= \frac{1}{\Gamma(a)} b^a \int_0^\infty u^{a-1} b^{-a+1} \exp(-u) b^{-1} du \qquad (\lambda = \tfrac{u}{b}) \tag{455}$$

$$= \frac{\Gamma(a)}{\Gamma(a)} = 1$$

2.42

$$\mathbb{E}[\lambda] = \int \lambda \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda) d\lambda \tag{456}$$

$$= \int \frac{1}{\Gamma(a)} b^a \lambda^a \exp(-b\lambda) d\lambda \tag{457}$$

$$= \frac{1}{\Gamma(a)} b^a \int \lambda^a \exp(-b\lambda) d\lambda \tag{458}$$

Which looks a lot like the gamma function $\Gamma(x) := \int u^{x-1} e^{-u} du$, so we do the change of variables $u = b\lambda \iff d\lambda = \frac{du}{b}$

$$\tag{459}$$

$$= \frac{1}{\Gamma(a)} b^a \int b^{-a} u^a \exp(-u) \frac{1}{b} du \tag{460}$$

$$= \frac{1}{b\Gamma(a)} \Gamma(a+1) \tag{461}$$

$$= \frac{a}{b} \tag{462}$$

$$\mathbb{E}[\lambda^2] = \frac{b^a}{\Gamma(a)} \int \lambda^{a+1} \exp(-b\lambda) d\lambda \tag{463}$$

$$= \frac{b^a}{\Gamma(a)} \int u^{a+1} b^{-a-1} \exp(-u) \frac{1}{b} du \tag{464}$$

$$= \frac{1}{\Gamma(a) b^2} \int u^{a+1} \exp(-u) du \tag{465}$$

$$= \frac{1}{\Gamma(a) b^2} \Gamma(a+2) \tag{466}$$

$$= \frac{a+1}{\Gamma(a) b^2} \Gamma(a+1) \tag{467}$$

$$= \frac{a(a+1)}{b^2} \tag{468}$$

And therefore

$$\mathbb{E}[\lambda^2] - \mathbb{E}[\lambda]^2 = \frac{a(a+1) - a^2}{b} = \frac{a}{b^2} \tag{469}$$

$$\frac{d}{d\lambda}\frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda) \tag{470}$$

$$= \frac{b^a}{\Gamma(a)}(a-1)\lambda^{a-2}\exp(-b\lambda) - b\frac{b^a}{\Gamma(a)}\lambda^{a-1}\exp(-b\lambda) = 0 \tag{471}$$

$$\iff (a-1)\lambda^{a-2} = b\lambda^{a-1} \tag{472}$$

$$\iff \lambda = \frac{a-1}{b} \tag{473}$$

$$\tag{474}$$

2.44

I think τ^{-1} should be λ?

2.46

$$a = \nu/2 \quad b = \frac{\nu}{2\lambda} \tag{475}$$

$$p(x|\mu, a, b) = \frac{b^a}{\Gamma(a)}\left(\frac{1}{2a}\right)^{1/2}\left[b + \frac{1}{2}(x-\mu)^2\right]^{-a-\frac{1}{2}}\Gamma(a+\frac{1}{2}) \tag{476}$$

$$= \frac{\Gamma(\nu/2+1/2)}{\Gamma(\nu/2)}\left(\frac{1}{2\pi}\right)^{1/2}\left[\frac{\nu}{2\lambda} + \frac{1}{2}(x-\mu)^2\right]^{-\frac{\nu}{2}-\frac{1}{2}}\left(\frac{\nu}{2\lambda}\right)^{\nu/2} \tag{477}$$

$$= \frac{\Gamma(\nu/2+1/2)}{\Gamma(\nu/2)}\left[\frac{2\lambda}{\nu}\left(\frac{\nu}{2\lambda} + \frac{1}{2}(x-\mu)^2\right)\right]^{-\frac{\nu}{2}}\left[2\pi\left(\frac{\nu}{2\lambda} + \frac{1}{2}(x-\mu)^2\right)\right]^{-\frac{1}{2}} \tag{478}$$

$$= \frac{\Gamma(\nu/2+1/2)}{\Gamma(\nu/2)}\left[1 + \frac{\lambda(x-\mu)^2}{\nu}\right]^{-\frac{\nu}{2}}\left[\frac{\pi\nu}{\lambda} + \pi(x-\mu)^2\right]^{-\frac{1}{2}} \tag{479}$$

$$= \frac{\Gamma(\nu/2+1/2)}{\Gamma(\nu/2)}\left[1 + \frac{\lambda(x-\mu)^2}{\nu}\right]^{-\frac{\nu}{2}}\left[1 + \frac{\lambda}{\nu}(x-\mu)^2\right]^{-\frac{1}{2}}\left(\frac{\lambda}{\pi\nu}\right)^{1/2} \tag{480}$$

$$\tag{481}$$

2.47

$$\mathrm{St}(x|\mu,\lambda,\nu) \propto \left[1+\frac{\lambda(x-\mu)^2}{\nu}\right]^{-v/2-1/2} \tag{482}$$

$$= \exp\ln\left[1+\frac{\lambda(x-\mu)^2}{\nu}\right]^{-v/2-1/2} \tag{483}$$

$$= \exp\left[\frac{-v-1}{2}\ln\left(1+\frac{\lambda(x-\mu)^2}{\nu}\right)\right] \tag{484}$$

$$= \exp\left[\frac{-v-1}{2}\left(\frac{\lambda(x-\mu)^2}{\nu}-\mathcal{O}\left[\frac{\lambda(x-\mu)^2}{2\nu}\right]^2\right)\right] \tag{485}$$

$$= \exp\left[\frac{-v-1}{2}\left(\frac{\lambda(x-\mu)^2}{\nu}-\mathcal{O}\left[\frac{\lambda^2(x-\mu)^4}{4\nu^2}\right]\right)\right] \tag{486}$$

$$\approx \exp\left[\frac{-1}{2}\left(\lambda(x-\mu)^2\right)\right] \qquad (v\to\infty) \tag{}$$

2.48

$$\Gamma(x) = \int \mu^{x-1}e^{-u}du \tag{487}$$

$$|(\eta\Lambda)^{-1}|^{1/2} = \left[\eta^{-D}|\Lambda^{-1}|\right]^{1/2} = \eta^{-D/2}|\Lambda|^{-1/2} \tag{488}$$

$$\Gamma(D/2+\nu/2) = \int \eta^{D/2+\nu/2-1}\exp(-\eta)d\eta \tag{489}$$

$$z = \eta/2(\Delta^2+\nu) \iff \eta = 2\frac{z}{\Delta^2+\nu} \tag{490}$$

$$d\eta = \frac{d\eta}{dz}dz = \frac{z}{\Delta^2+\nu}dz \tag{491}$$

$$\int \mathcal{N}(x|\mu,(\eta\Lambda)^{-1})\text{Gam}\left(\eta|\frac{\nu}{2},\frac{\nu}{2}\right)d\eta \tag{492}$$

$$= \int \frac{1}{(2\pi)^{D/2}}\frac{1}{|(\eta\Lambda)^{-1}|^{1/2}}\exp\{-\frac{1}{2}(\boldsymbol{x}-\mu)^T(\eta\Lambda)(\boldsymbol{x}-\mu)\}\frac{1}{\Gamma(\nu/2)}(\nu/2)^{\nu/2}\eta^{\frac{\nu}{2}-1}\exp(-\nu/2\eta)d\eta \tag{493}$$

$$= \int \frac{1}{(2\pi)^{D/2}}\frac{1}{|(\eta\Lambda)^{-1}|^{1/2}}\exp\{-\frac{\eta}{2}\Delta^2\}\frac{1}{\Gamma(\nu/2)}(\nu/2)^{\nu/2}\eta^{\frac{\nu}{2}-1}\exp(-\nu/2\eta)d\eta \tag{494}$$

$$= \int \frac{1}{(2\pi)^{D/2}}\eta^{D/2}|\Lambda|^{1/2}\exp\{-\frac{\eta}{2}\Delta^2\}\frac{1}{\Gamma(\nu/2)}(\nu/2)^{\nu/2}\eta^{\frac{\nu}{2}-1}\exp(-\nu/2\eta)d\eta \tag{495}$$

$$= \frac{1}{(2\pi)^{D/2}}|\Lambda|^{1/2}\frac{1}{\Gamma(\nu/2)}(\nu/2)^{\nu/2}\int \eta^{D/2}\exp\{-\frac{\eta}{2}\Delta^2\}\eta^{\frac{\nu}{2}-1}\exp(-\nu/2\eta)d\eta \tag{496}$$

$$= \frac{1}{(2\pi)^{D/2}}|\Lambda|^{1/2}\frac{1}{\Gamma(\nu/2)}(\nu/2)^{\nu/2}\int \eta^{D/2+v/2-1}\exp\{-\frac{\eta}{2}(\Delta^2-\nu)\}d\eta \tag{497}$$

$$= \frac{1}{(2\pi)^{D/2}}|\Lambda|^{1/2}\frac{1}{\Gamma(\nu/2)}(\nu/2)^{\nu/2}2(\Delta^2+\nu)^{-D/2-\nu/2}\int z^{1/2D+\nu/2}\exp\{-z\}d\eta \tag{498}$$

$$= \frac{1}{(2\pi)^{D/2}}|\Lambda|^{1/2}\frac{1}{\Gamma(\nu/2)}(\nu/2)^{\nu/2}2(\Delta^2+\nu)^{-D/2-\nu/2}\Gamma(D/2+\nu/2) \tag{499}$$

$$\tag{500}$$

Which works out algebraically to the desired result.

2.49

$$\mathbb{E}_x[\int_0^\infty \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},(\eta\boldsymbol{\Lambda})^{-1})\Gamma(\eta|\nu/2,\nu/2)d\eta] = \int_0^\infty \Gamma(\eta|\nu/2,\nu/2)d\eta\,\mathbb{E}_x[\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},(\eta\boldsymbol{\Lambda})^{-1})] \tag{501}$$

$$= 1\cdot\boldsymbol{\mu} \tag{502}$$

$$\text{cov}[\boldsymbol{x}] = \mathbb{E}[xx^T] - \mathbb{E}[x]\,\mathbb{E}[x]^T = \mathbb{E}[xx^T] - \mu\mu^T \tag{503}$$

$$\mathbb{E}[xx^T] = \int_0^\infty \text{Gam}(\eta|\nu/2,\nu/2)\int xx^T\mathcal{N}(x|\mu,(\eta\Lambda)^{-1})dxd\eta \tag{504}$$

We use 2.62 for the inner expectation

$$= \int_0^\infty \text{Gam}(\eta|\nu/2, \nu 2)(\mu\mu^T + (\eta\Lambda)^{-1})d\eta \tag{505}$$

$$= \mu\mu^T + \Lambda^{-1}\int_0^\infty \text{Gam}(\eta|\nu/2, \nu/2)\eta^{-1}d\eta \tag{506}$$

$$= \mu\mu^T + \Lambda^{-1}\int_0^\infty \frac{1}{\Gamma(\nu/2)}(\nu/2)^{\nu/2}\eta^{\nu/2-1}\eta^{-1}\exp(-\nu/2\eta)d\eta \tag{507}$$

$$= \mu\mu^T + \Lambda^{-1}\int_0^\infty \frac{1}{\Gamma(\nu/2)}(\nu/2)^{\nu/2}\eta^{\nu/2-2}\exp(-\nu/2\eta)d\eta \tag{508}$$

$$= \mu\mu^T + \Lambda^{-1}\int_0^\infty \frac{1}{(\nu/2-1)\Gamma(\nu/2-1)}(\nu/2)^{\nu/2}\eta^{\nu/2-2}\exp(-\nu/2\eta)d\eta \tag{509}$$

$$= \mu\mu^T + \Lambda^{-1}\frac{\nu/2}{\nu/2-1}\int_0^\infty \frac{1}{\Gamma(\nu/2-1)}(\nu/2)^{\nu/2-1}\eta^{\nu/2-2}\exp(-\nu/2\eta)d\eta \tag{510}$$

$$= \mu\mu^T + \Lambda^{-1}\frac{\nu/2}{\nu/2-1}\int_0^\infty \text{Gam}(\nu/2-1, \nu/2)d\nu \tag{511}$$

$$= \mu\mu^T + \Lambda^{-1}\frac{\nu}{\nu-2} \tag{512}$$

Plugging this into the covariance formula gives the desired result.
Finding the mode is done by setting the derivative of 2.162 and setting it to zero.

2.50

$$\text{St}(\boldsymbol{x}|\mu, \Lambda, \nu) \propto \left[1 + \frac{1}{\nu}\Delta^2\right]^{-D/2-v/2} \tag{513}$$

$$= \exp\frac{-D-\nu}{2}\ln\left[1 + \frac{1}{\nu}\Delta^2\right] \tag{514}$$

$$= \exp\frac{-D-\nu}{2}\left[\frac{1}{\nu}\Delta^2 - \boldsymbol{\mathcal{O}}(1/\nu^2)\right] \qquad \text{(Taylor)}$$

$$= \exp\frac{-D-\nu}{2}\frac{1}{\nu}\Delta^2 - \frac{-D-\nu}{2}\boldsymbol{\mathcal{O}}(1/\nu^2) \tag{515}$$

$$= \exp -1/2\Delta^2 \qquad (\nu \to \infty)$$

2.51

$$1 = [\cos(A) + i\sin(A)][\cos(A) - i\sin(A)] \tag{516}$$

$$= \cos^2(A) - \sin^2(A) \tag{517}$$

$$\cos(A-B) = \Re \exp[i(A-B)] \tag{518}$$
$$= \Re \exp(iA)\exp(-iB) \tag{519}$$
$$= \Re[\cos(A) + i\sin(A)][\cos(B) - i\sin(B)] \tag{520}$$
$$= \cos(A)\cos(B) - \sin(A)\sin(B) \tag{521}$$

The final question is exactly the same but considering the $\Im$ part.

2.52

Plugging in the Taylor expansion in the exponent, we see that the exponent becomes proportional to $-\frac{m}{2}(\theta-\theta_0)^2 = -\frac{1}{2\sigma^2}(\theta-\theta_0)$ if we define $m = 1/\sigma^2$ the precision.

2.54

$$0 = \sum_{n=1}^{N} \cos(\theta_0)\sin(\theta_n) - \cos(\theta_n)\sin(\theta_0) \tag{522}$$
$$= \cos(\theta_0)\sum_{n=1}^{N}\sin(\theta_n) - \sin(\theta_0)\sum_{n=1}^{N}\cos(\theta_n) \tag{523}$$
$$\tag{524}$$
$$\frac{\sin(\theta_0)}{\cos(\theta_0} = \frac{\sum_{n=1}^{N}\sin(\theta_n)}{\sum_{n=1}^{N}\cos(\theta_n)} \tag{525}$$
$$\tan(\theta_0) = \frac{\sum_{n=1}^{N}\sin(\theta_n)}{\sum_{n=1}^{N}\cos(\theta_n)} \tag{526}$$
$$\theta_0 = \arctan\left\{\frac{\sum_{n=1}^{N}\sin(\theta_n)}{\sum_{n=1}^{N}\cos(\theta_n)}\right\} \tag{527}$$

2.54

$$\frac{\partial p}{\partial \theta} = -(2\pi I_0(m))^{-1}\exp\{m\cos(\theta-\theta_0\}(m\sin(\theta-\theta_0)) \tag{528}$$

Setting to 0 and solving gives $\sin(\theta-\theta_0) = 0$ which resolutes to $\theta^* = \theta_0 + n\pi \quad n \in \mathbb{Z}$ where $\mathbb{Z}$ are the positive integers.

$$\frac{\partial\partial p}{\partial \theta} = \frac{1}{2\pi I_0(m)}\left[-\exp(m\cos(\theta-\theta_0)(m\sin(\theta-\theta_0)^2 - m\cos(\theta-\theta_0)\exp(m\cos(\theta-\theta_0))\right] \tag{529}$$

Left term vanishes since $sin(\theta^*) = 0$. Right term is positive (so maximal) for $\theta^* = \theta_0 + 0\pi \pmod{2\pi}$, negative (minimal) for $\theta^* = \theta_0 + \pi \pmod{2\pi}$.

2.55

$$A(m_{ML}) = \left(\frac{1}{N}\sum_{n=1}^{N}\cos\theta_n\right)\cos\theta_0^{ML} - \left(\frac{1}{N}\sum_{n=1}^{N}\sin\theta_n\right)\sin\theta_0^{ML} \tag{530}$$

$$= \bar{r}\left(\cos\bar{\theta}\cos\theta_0^{ML} - \sin\bar{\theta}\sin\theta_0^{ML}\right) \tag{531}$$

$$= \bar{r}\left(\cos(\bar{\theta} - \theta_0^{ML})\right) \tag{532}$$

$$= \bar{r} \tag{533}$$

2.56

$$\beta(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \tag{534}$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\exp\log[\mu^{a-1}(1-\mu)^{b-1}] \tag{535}$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\left(\exp[(a-1)\log\mu + (b-1)\log[1-\mu]]\right) \tag{536}$$

$$\tag{537}$$

And so we have $h(\mu) = 1, g(\eta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}, \eta^T = [a-1, b-1]$ and $u(\mu) = [\ln(\mu), \ln(1-\mu)]^T$

$$\mathrm{Gam}(\lambda|a,b) = \frac{1}{\Gamma(a)}b^a\lambda^{a-1}\exp(-b\lambda) \tag{538}$$

$$= \Gamma(a)^{-1}b^a\exp\log[\lambda^{a-1}]\exp(-b\lambda) \tag{539}$$

$$= \Gamma(a)^{-1}b^a\exp(a-1)\log[\lambda]\exp(-b\lambda) \tag{540}$$

$$= \Gamma(a)^{-1}b^a\exp[(a-1)\log[\lambda] + -b\lambda)] \tag{541}$$

$$\tag{542}$$

And so we have $h(\lambda) = 1, g(\eta) = \Gamma(a)^{-1}b^a, \eta^T = [a-1, -b]$ and $u(\lambda) = [\log\lambda, \lambda]^T$

The von Mises distribution is obvious. We only have to expand the cos in the exponent: $\cos(\theta - \theta_0) = \cos(\theta)\cos(\theta_0) + \sin(\theta)\sin(\theta_0)$ We then have $u(\theta) = [\cos(\theta), \sin(\theta)]^T$ and $\eta^T = [m\cos(\theta_0), m\sin(\theta_0)]$

2.57

For this question, the following knowledge is necessary:

$$a^TBa = \underbrace{B : aa^T}_{\text{Frobenius product (Hadamard \& sum)}} = \mathrm{vec}(B)^T\mathrm{vec}(aa^T) \tag{543}$$

where $vec(\cdot)$ is the vectorization operation.

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = (2\pi)^{-D/2}|\boldsymbol{\Sigma}|^{1/2}\exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right] \tag{544}$$

$$= (2\pi)^{-D/2}|\boldsymbol{\Sigma}|^{1/2}\exp\left[-\frac{1}{2}\left(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)\right] \tag{545}$$

$$= (2\pi)^{-D/2}|\boldsymbol{\Sigma}|^{1/2}\exp(-1/2\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\exp\left[\begin{pmatrix}-1/2\mathrm{vec}(\boldsymbol{\Sigma}^{-1})\\ -\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\end{pmatrix}\begin{pmatrix}\mathrm{vec}(\boldsymbol{x}\boldsymbol{x}^T) & \boldsymbol{x}\end{pmatrix}\right] \tag{546}$$

$$= (2\pi)^{-D/2}|\boldsymbol{\Sigma}|^{1/2}\exp(-1/2\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\exp\left[\eta^T u(\boldsymbol{x})\right] \tag{547}$$

$$\tag{548}$$

Now we only need to rewrite the remaining factors in terms of η. which gives:

$$g(\eta) = |\mathrm{vec}^{-1}(-2\eta_1)|^{-1/2}\eta_2\mathrm{vec}^{-1}(-2\eta_1)\eta_2^T \tag{549}$$

and

$$h(x) = (2\pi)^{-D/2} \tag{550}$$

2.59

$$\int_0^\infty \frac{1}{x}f(\frac{1}{x})dx = \int_0^\infty \frac{1}{x}f(y)xdy = \int_0^\infty f(y)dy = 1 \tag{551}$$

2.60

Maybe I'm mistaken, but shouldn't 'volume' be domain here? I think the volume of a region i is equal to the domain Δ_i times it's density h_i. We then have the constraint $\sum_i \Delta_i h_i = 1$

Solving $\sum_{n=1}^N \log h(\boldsymbol{x}) + \lambda(\sum_i \Delta_i h_i - 1)$ will give the solution.

2.61

3.1

$$2\sigma(2a) - 1 = \frac{2}{1+e^{-2a}} - 1 \tag{552}$$

$$= \frac{e^a}{e^a}\frac{2}{1+e^{-2a}} - 1 \tag{553}$$

$$= \frac{2e^a}{e^a+e^{-a}} - 1 \tag{554}$$

$$= \frac{2e^a}{e^a+e^{-a}} - \frac{e^a+e^{-a}}{e^a+e^{-a}} \tag{555}$$

$$= \tanh(a) \tag{556}$$

$$a = \frac{x - \mu_j}{s} \tag{557}$$

$$y(x, \boldsymbol{w}) = w_0 + \sum_{j=1}^{M} w_j \sigma(a) \tag{558}$$

$$= w_0 + \sum_{j=1}^{M} \frac{w_j}{2} \tanh(\frac{1}{2}a) + w_j \tag{559}$$

$$= w_0 + \sum_{j=1}^{M} + \sum_{j=1}^{M} \frac{w_j}{2} \tanh(a') \tag{560}$$

$$= u_0 + \sum_{j=1}^{M} u_j \tanh(a') \tag{561}$$

3.2

Proving that some vector is in the column space of a matrix comes down to proving that that vector the result of a multiplication with that matrix. We have $\Phi(\Phi^T\Phi)^{-1}\Phi^T v$ which can be written as Φw and so the result of the projection operator $(\Phi(\Phi^T\Phi)^{-1}\Phi^T)$ will lie in $\mathcal{S}$. Now you can see that $y = \Phi w_{ML} = \Phi(\Phi^T\Phi)^{-1}\Phi^T t$ is a projection of t onto $\mathcal{S}$. Also note that a projection of a vector from that same matrix is the same vector (projection onto its own column space doesn't change anything): $\Phi(\Phi^T\Phi)^{-1}\Phi^T\Phi = \Phi I = \Phi$. Now note by looking at the figure that connects the heads of y and t is the vector $y - t$. If the projection is orthogonal, this vector must be orthogonal to any vector in $\mathcal{S}$.

$$(y - t)^T\Phi = y^T\Phi - t^T\Phi \tag{562}$$

$$= (\Phi w_{ML})^T\Phi - t^T\Phi \tag{563}$$

$$= w_{ML}^T\Phi^T\Phi - t^T\Phi \tag{564}$$

$$= ((\Phi^T\Phi)^{-1}\Phi^T t)^T\Phi^T\Phi - t^T\Phi \tag{565}$$

$$= t^T\Phi(\Phi^T\Phi)^{-1}\Phi^T\Phi - t^T\Phi \tag{566}$$

$$= t^T\Phi - t^T\Phi \tag{567}$$

which shows that the projection is orthogonal.

3.3

$$\frac{dE_D(\boldsymbol{w})}{d\boldsymbol{w}} = \sum_{n=1}^{N} r_n[t_n - \boldsymbol{w}^T\phi(\boldsymbol{x}_n)]\phi(\boldsymbol{x}_n) = 0 \tag{568}$$

$$\sum_{n=1}^{N} r_n t_n \phi(\boldsymbol{x}_n) = \sum_{n=1}^{N} \boldsymbol{w}^T \phi(\boldsymbol{x}_n)\phi(x_n) \tag{569}$$

$$= \sum_{n=1}^{N} \phi(\boldsymbol{x}_n)^T \boldsymbol{w}\phi(\boldsymbol{x}_n) \tag{570}$$

$$= \sum_{n=1}^{N} \phi(\boldsymbol{x}_n)\phi(\boldsymbol{x}_n)^T \boldsymbol{w} \tag{571}$$

$$= \boldsymbol{\Phi}^T \boldsymbol{\Phi} \boldsymbol{w} \tag{572}$$

$$\left([\boldsymbol{r} \cdot \boldsymbol{t}]^T \boldsymbol{\Phi}\right)^T = \boldsymbol{\Phi}^T \boldsymbol{\Phi} \boldsymbol{w} \tag{573}$$

$$\boldsymbol{\Phi}^T [\boldsymbol{r} \cdot \boldsymbol{t}] = \boldsymbol{\Phi}^T \boldsymbol{\Phi} \boldsymbol{w} \tag{574}$$

$$(\boldsymbol{\Phi}^T \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^T [\boldsymbol{r} \cdot \boldsymbol{t}] = \boldsymbol{w} \tag{575}$$

(i) ?
(ii) $r_n > 0$ essentially replicates data-points that otherwise would have been summed.

3.4

$$\mathbb{E}[E_d(\boldsymbol{w})] = \mathbb{E}[\frac{1}{2} \sum_{n=1}^{N} (y(\boldsymbol{x}_n, \boldsymbol{w}) - t_n)^2] \tag{576}$$

$$= \mathbb{E}[\frac{1}{2} \sum_{n=1}^{N} (w_0 + \sum_{i=1}^{D} w_i(x_i + \epsilon_i) - t_n)^2] \tag{577}$$

$$= \mathbb{E}[\frac{1}{2} \sum_{n=1}^{N} (y(\boldsymbol{x}_n, \boldsymbol{w}) + \sum_{i=1}^{D} w_i \epsilon_i) - t_n)^2] \tag{578}$$

$$= \mathbb{E}[\frac{1}{2} \sum_{n=1}^{N} (y(\boldsymbol{x}_n, \boldsymbol{w}) - t_n])^2 + (\sum_{i=1}^{D} w_i \epsilon_i)^2 + 2 \underbrace{(y(\boldsymbol{x}_n, \boldsymbol{w}) - t_n) \sum_{i=1}^{D} w_i \epsilon_i]}_{=0(\mathbb{E}[\epsilon_i]=0))} \tag{579}$$

Since $\mathbb{E}[w_i w_j] = \mathbb{E}[w_i]\,\mathbb{E}[w_j] = 0$ for all $i \neq j$:

$$\mathbb{E}(\sum_{i=1}^{D} w_i \epsilon_i)^2 = \mathbb{E}[\sum_{i=1}^{D} \sum_{j=1}^{D} w_j \epsilon_j w_i \epsilon_i] = \sum_{i=1}^{D} w_i^2\, \mathbb{E}\, \epsilon_i^2 = \sum_{i=1}^{D} w_i^2\, \mathbb{E}\, \epsilon_i^2 = \sum_{i=1}^{D} w_i^2 (\sigma^2 + \underbrace{E[\epsilon_i]^2}_{=0}) \tag{580}$$

Which gives us our desired result.

3.5

$$\mathcal{L}(\boldsymbol{w}, \lambda) = \frac{1}{2} \sum_{n=1}^{N} \left[t_n - \boldsymbol{w}^T \phi(\boldsymbol{x}_n) \right]^2 + \lambda \left[\sum_{j=1}^{M} |w_j|^q - \eta \right] \tag{581}$$

Which has the same dependence on $\boldsymbol{w}$ up to a scaling factor.

3.6

$$p(\boldsymbol{T}|\boldsymbol{W}, \boldsymbol{\Sigma}) = \prod_{n=1}^{N} p(\boldsymbol{t}_n|\boldsymbol{W}, \boldsymbol{\Sigma}) \tag{582}$$

$$\ln p(\boldsymbol{T}|\boldsymbol{W}, \boldsymbol{\Sigma}) = \ln \prod_{n=1}^{N} p(\boldsymbol{t}_n|\boldsymbol{W}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln p(\boldsymbol{t}_n|\boldsymbol{W}, \boldsymbol{\Sigma}) \tag{583}$$

$$= \sum_{n=1}^{N} \ln \left((2\pi)^{-D/2} |\boldsymbol{\Sigma}|^{-1/2} \right) + \sum_{n=1}^{N} \frac{1}{2} (\boldsymbol{y}(\boldsymbol{x}, \boldsymbol{W}) - \boldsymbol{t})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}(\boldsymbol{x}, \boldsymbol{W}) - \boldsymbol{t})] \tag{584}$$

$$\frac{\partial p}{\partial \boldsymbol{W}} = \frac{1}{2} \sum_{n=1}^{N} \frac{\partial}{\partial \boldsymbol{W}} \left[(\boldsymbol{W}^T \phi(\boldsymbol{x}) - \boldsymbol{t})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{w}^T \phi(\boldsymbol{x}) - \boldsymbol{t}) \right] \tag{585}$$

$$= \frac{1}{2} \sum_{n=1}^{N} (\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-1^T}) (\boldsymbol{W}^T \phi(\boldsymbol{x}) - \boldsymbol{t}) \phi(\boldsymbol{x})^T = 0 \tag{586}$$

$$= \sum_{n=1}^{N} \boldsymbol{\Sigma}^{-1} (\boldsymbol{W}^T \phi(\boldsymbol{x}) - \boldsymbol{t}) \phi(\boldsymbol{x})^T \tag{587}$$

$$\boldsymbol{\Sigma}^{-1} \sum_{n=1}^{N} \boldsymbol{W}^T \phi(\boldsymbol{x}) \phi(x)^T = \boldsymbol{\Sigma}^{-1} \sum_{n=1}^{N} \boldsymbol{t} \phi(\boldsymbol{x})^T \tag{588}$$

$$\boldsymbol{W}^T \boldsymbol{\Phi}^T \boldsymbol{\Phi} = \boldsymbol{T} \boldsymbol{\Phi}^T \tag{589}$$

$$\boldsymbol{W}^T = \boldsymbol{T} \boldsymbol{\Phi}^T (\boldsymbol{\Phi}^T \boldsymbol{\Phi})^{-1} \tag{590}$$

$$\boldsymbol{W} = (\boldsymbol{\Phi}^T \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^T \boldsymbol{T} \tag{591}$$

$$\frac{d \ln p(\boldsymbol{T}|\boldsymbol{x}, \boldsymbol{W}, \boldsymbol{\Sigma}}{d\boldsymbol{\Sigma}} = -\frac{N}{2}\frac{d}{d\boldsymbol{\Sigma}} \ln |\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N} \frac{d}{d\boldsymbol{\Sigma}} (\boldsymbol{t}_n - \boldsymbol{W}_{ML}^T \phi(\boldsymbol{x}_n))^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{t}_n - \boldsymbol{W}_{ML}^T \phi(\boldsymbol{x}_n)) \quad (592)$$

$$= \frac{N}{2}\frac{d}{d\boldsymbol{\Sigma}} \ln |\boldsymbol{\Sigma}|^{-1} - \frac{1}{2}\sum_{n=1}^{N} (\boldsymbol{t}_n - \boldsymbol{W}_{ML}^T \phi(\boldsymbol{x}_n))(\boldsymbol{t}_n - \boldsymbol{W}_{ML}^T \phi(\boldsymbol{x}_n))^T \quad (593)$$

$$= \frac{N}{2}\frac{d}{d\boldsymbol{\Sigma}} \ln |\boldsymbol{\Sigma}^{-1}| - \frac{1}{2}\sum_{n=1}^{N} (\boldsymbol{t}_n - \boldsymbol{W}_{ML}^T \phi(\boldsymbol{x}_n))(\boldsymbol{t}_n - \boldsymbol{W}_{ML}^T \phi(\boldsymbol{x}_n))^T \quad (594)$$

$$= \frac{N}{2}\boldsymbol{\Sigma} - \frac{1}{2}\sum_{n=1}^{N} (\boldsymbol{t}_n - \boldsymbol{W}_{ML}^T \phi(\boldsymbol{x}_n))(\boldsymbol{t}_n - \boldsymbol{W}_{ML}^T \phi(\boldsymbol{x}_n))^T \quad (595)$$

$$(596)$$

$$\boldsymbol{\Sigma} = \frac{1}{N}\sum_{n=1}^{N} (\boldsymbol{t}_n - \boldsymbol{W}_{ML}^T \phi(\boldsymbol{x}_n))(\boldsymbol{t}_n - \boldsymbol{W}_{ML}^T \phi(\boldsymbol{x}_n))^T \quad (597)$$

3.7

Hint 1: use $x^T A x + x^T b + c = (x-h)^T A (x-h) + k$ where $h = -(1/2)A^{-1}b$ and $k = c - \frac{1}{4} b^T A^{-1} b$ if A is symmetric.

$$p(\boldsymbol{w}|\boldsymbol{t}) \propto p(\boldsymbol{w}) p(\boldsymbol{t}|, \boldsymbol{x}, \boldsymbol{w}, \beta^{-1}) \quad (598)$$

$$= \mathcal{N}(\boldsymbol{w}|_0, \boldsymbol{S}_0) \prod_{n=1}^{N} \mathcal{N}(t_n|\boldsymbol{w}^T \phi(\boldsymbol{x}_n), \beta^{-1}) \quad (599)$$

$$= (2\pi)^{-D/2} |\boldsymbol{S}_0|^{-1/2} \exp[-\frac{1}{2}(\boldsymbol{w} - \boldsymbol{m}_0)^T \boldsymbol{S}_0^{-1} (\boldsymbol{w} - \boldsymbol{m}_0)] \prod_{n=1}^{N} \left(\frac{\beta}{2\pi}\right)^{\frac{1}{2}} \exp[-\frac{\beta}{2}(t_n - \boldsymbol{w}^T \phi(\boldsymbol{x}_n))^2] \quad (600)$$

$$= (2\pi)^{-D/2} |\boldsymbol{S}_0|^{-1/2} \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} \exp[-\frac{1}{2}(\boldsymbol{w} - \boldsymbol{m}_0)^T \boldsymbol{S}_0^{-1} (\boldsymbol{w} - \boldsymbol{m}_0) + \sum_{n=1}^{N} -\frac{\beta}{2}(t_n - \boldsymbol{w}^T \phi(\boldsymbol{x}_n))^2] \quad (601)$$

$$(602)$$

Now we have to get the exponent right.

$$-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{m}_0)^T\boldsymbol{S}_0^{-1}(\boldsymbol{w}-\boldsymbol{m}_0)+\sum_{n=1}^{N}-\frac{\beta}{2}(t_n-\boldsymbol{w}^T\phi(\boldsymbol{x}_n))^2 \tag{603}$$

$$=-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{m}_0)^T\boldsymbol{S}_0^{-1}(\boldsymbol{w}-\boldsymbol{m}_0)-\frac{\beta}{2}\sum_{n=1}^{N}(t_n^2+\boldsymbol{w}^T\phi(\boldsymbol{x}_n)\boldsymbol{w}^T\phi(\boldsymbol{x}_n)-2\boldsymbol{w}^T\phi(\boldsymbol{x}_n)t_n) \tag{604}$$

$$=-\frac{1}{2}(\boldsymbol{w}^T\boldsymbol{S}_0^{-1}\boldsymbol{w}-\boldsymbol{w}^T\boldsymbol{S}_0^{-1}\boldsymbol{m}_0-\boldsymbol{m}_0^T\boldsymbol{S}_0^{-1}\boldsymbol{w}+\boldsymbol{m}_0^T\boldsymbol{S}_0^{-1}\boldsymbol{m}_0)-\frac{\beta}{2}(\boldsymbol{t}^T\boldsymbol{t}+\boldsymbol{w}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{w}-2\boldsymbol{w}^T\boldsymbol{\Phi}^T\boldsymbol{t}) \tag{605}$$

$$=-\frac{1}{2}\left(\boldsymbol{w}^T\boldsymbol{S}_0^{-1}\boldsymbol{w}-2\boldsymbol{w}^T\boldsymbol{S}_0^{-1}\boldsymbol{m}_0+\boldsymbol{m}_0^T\boldsymbol{S}_0^{-1}\boldsymbol{m}_0-\beta\boldsymbol{t}^T\boldsymbol{t}+\beta\boldsymbol{w}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{w}-2\beta\boldsymbol{w}^T\boldsymbol{\Phi}^T\boldsymbol{t}\right) \tag{606}$$

$$=\boldsymbol{w}^T(-\frac{1}{2}(\boldsymbol{S}_0^{-1}+\beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}))\boldsymbol{w}+\boldsymbol{w}(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0+\boldsymbol{\Phi}^T\boldsymbol{t})+c \tag{607}$$

$$=\boldsymbol{w}^T(-\frac{1}{2}\boldsymbol{S}_N^{-1})\boldsymbol{w}+\boldsymbol{w}(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0+\boldsymbol{\Phi}^T\boldsymbol{t})+c \tag{608}$$

Now completing the square.

$$\boldsymbol{w}^T(-\frac{1}{2}\boldsymbol{S}_N^{-1})\boldsymbol{w}+\boldsymbol{w}(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0+\boldsymbol{\Phi}^T\boldsymbol{t})+c \tag{609}$$

$$=(\boldsymbol{w}+\frac{1}{2}(-\frac{1}{2}S_N^{-1})^{-1}(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0+\boldsymbol{\Phi}^T\boldsymbol{t}))^T(-\frac{1}{2}\boldsymbol{S}_n^{-1})(\boldsymbol{w}+\frac{1}{2}(-\frac{1}{2}S_N^{-1})^{-1}(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0+\boldsymbol{\Phi}^T\boldsymbol{t})+k \tag{610}$$

$$=-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{m}_N)^T\boldsymbol{S}_n^{-1}(\boldsymbol{w}-\boldsymbol{m}_N)+k \tag{611}$$

3.8

$$\ln p(w|t)^{(N)}\propto\ln p(t|X,w)^{(N)}+\ln p(w|t)^{(N-1)} \tag{612}$$

$$=\ln\mathcal{N}(t|w^T\phi(x),\beta^{-1}I)+\ln\mathcal{N}(w|m_N,S_N) \tag{613}$$

$$=C-\frac{\beta}{2}(t-w^T\phi(x))^T(t-w^T\phi(x))-\frac{1}{2}(w-m_n)^TS_N^{-1}(w-m_n) \tag{614}$$

$$=C-\frac{1}{2}w^T(\beta\phi(x)\phi(x)^T+S_N^{-1})w-w^T(\beta\phi(x)t+S_N^{-1}m_N \tag{615}$$

We added terms not dependent on w to C

$$\tag{616}$$

From the completing-the-square formula we can read off the result:

$$S_{N+1}^{-1}=S_N^{-1}+\beta\phi(x)\phi(x)^T \tag{617}$$

$$m_{N+1}=S_{N+1}(S_N^{-1}m_N+\beta\phi(x)t) \tag{618}$$

3.9

We have

$$p(x|y) = \mathcal{N}(x|\Sigma[A^T L(y-b) + \Lambda\mu], \Sigma) \tag{619}$$

with $x = w, \mu = m_N, \Lambda^{-1} = S_N y = t, (Ax+b) = \phi(x)^T w, L^{-1} = \beta^{-1} I$, therefore $\Sigma = (S_N^{-1} + \beta\phi(x)\phi(x)^T)$

$$\begin{aligned} p(x|y) &= \mathcal{N}(w|(S_N^{-1} + \beta\phi(x)\phi(x)^T)[\phi(x)^T \beta I t) + S_N^{-1} m_N], (S_N^{-1} + \beta\phi\phi^T) && (620)\\ &= \mathcal{N}(w|S_{N+1}^{-1}(\beta\Phi I t + S_N^{-1} m_N, S_{N+1}^{-1}) && (621)\\ &= \mathcal{N}(w|m_{N+1}, S_{N+1}^{-1}) && (622) \end{aligned}$$

3.10

We have

1. $Ax + b \Rightarrow y(x,w) = \phi(x)^T w$
2. $L^{-1} \Rightarrow \beta^{-1}$
3. $y \Rightarrow t$
4. $\mu \Rightarrow m_N$
5. $\Lambda^{-1} \Rightarrow S_N$

Directly filling in these values gives

$$p(t|x, \boldsymbol{t}, \alpha, \beta) = \mathcal{N}(t|\phi(x)^T m_N, \beta^{-1} + \phi(x)^T S_N \phi(x)) \tag{623}$$

3.11

$$\begin{aligned} S_{N+1} &= \left(S_N^{-1} + \beta(\phi(x_{N+1})\phi(x_{N+1})^T\right)^{-1} && (624)\\ &= S_N - \left(\frac{S_N\phi(x_{N+1})(\phi(x_{N+1}^T S_N)}{1 + \phi(x)^T S_N \phi(x)}\right) && (625) \end{aligned}$$

thus

$$\phi(x)^T S_{N+1}\phi(x) = \phi(x)^T S_N \phi(x) - \phi(x)^T A \phi(x) \tag{626}$$

with A as the fraction.

Then we have

$$\sigma_N^2 - \sigma_{N+1}^2 = \phi(x)^T A \phi(x) \geq 0 \tag{627}$$

since (denominator) S_N is positive-semidefinite (i.e., all its quadratic forms are nonnegative) and we get $(\phi(x)^T S_N \phi(x))^2$ in the numerator.

3.12

3.13

The predictive distribution is the integrated posterior multiplied with the likelihood for a new datapoint. In our case the posterior is $p(w, \beta|t)$ and the likelihood $p(t|w, x, \beta)$ (3.113) is our posterior and the likelihood used in 3.12 was $\mathcal{N}(w^T\phi(x), \beta^{-1})$. The predictive distribution thus becomes

$$\int\int \mathcal{N}(w|m_N, \beta^{-1}S_N)\text{Gam}(\beta|a_N, b_N)\,\mathcal{N}(w^T\phi(x), \beta^{-1})dwd\beta \tag{628}$$

Noticing that this is the convolution of two Gaussians we can use (2.115). Then we're left with something that looks much like 2.161, and thus a student t.

3.14

3.15

Hint 1: Use 3.95 and 3.92

$$E(\boldsymbol{m}_N) = \frac{\beta}{2}\|\boldsymbol{t} - \boldsymbol{\Phi}\boldsymbol{m}_N\|^2 + \frac{\alpha}{2}\boldsymbol{m}_N^T\boldsymbol{m}_N \tag{629}$$

$$= \frac{\beta}{2}\|\boldsymbol{t} - \boldsymbol{\Phi}\boldsymbol{m}_N\|^2 + \frac{\gamma(\boldsymbol{m}_N^T\boldsymbol{m}_N)^{-1}}{2}\boldsymbol{m}_N^T\boldsymbol{m}_N \tag{630}$$

$$= \frac{\beta}{2}\|\boldsymbol{t} - \boldsymbol{\Phi}\boldsymbol{m}_N\|^2 + \frac{\gamma}{2} \tag{631}$$

$$\tag{632}$$

$$\beta = (N - \gamma)\|\boldsymbol{t} - \boldsymbol{\Phi}^T\boldsymbol{m}_n\|^{-2} \tag{633}$$

$$E(\boldsymbol{m}_N) = \frac{1}{2}(\mathcal{N} - \gamma + \gamma) = \frac{N}{2} \tag{634}$$

3.17

Evidence function:

$$p(D|M_i) = \int p(D|\boldsymbol{w}, M_i)p(\boldsymbol{w}, M_i)d\boldsymbol{w} \tag{635}$$

For Linear regression we have $M_i = (\alpha, \beta)$

$$p(D|\alpha, \beta) = \int p(\boldsymbol{w}|\alpha^{-1}I)\prod_{n=1}^{N} p(t_n|\boldsymbol{w}^T\phi(\boldsymbol{x}_n), \beta^{-1}) \tag{636}$$

$$= \int (2\pi)^{-M/2}|\alpha^{-1}I|^{-1/2}\exp[-1/2(\boldsymbol{w}^T(\alpha^{-1}I)^{-1}\boldsymbol{w}]\prod_{n=1}^{N}\frac{1}{\sqrt{2\pi\beta^{-1}}}\exp[\beta/2(t_n - \boldsymbol{w}^T\phi(\boldsymbol{x}_n))^2] \tag{637}$$

$$= \int (2\pi)^{-M/2}(\alpha^{-1})^{-M/2}\frac{\beta}{2\pi}^{N/2}\exp[-\alpha/2\boldsymbol{w}^T\boldsymbol{w} - \beta/2\|\boldsymbol{t} - \boldsymbol{\Phi}\boldsymbol{w}\|^2] \tag{638}$$

3.18

$$\frac{\beta}{2}(\boldsymbol{t}-\boldsymbol{\Phi w})^T(\boldsymbol{t}-\boldsymbol{\Phi w})+\frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w} \tag{639}$$

$$=\frac{\beta}{2}\boldsymbol{t}^T\boldsymbol{t}-\beta\boldsymbol{w}^T\boldsymbol{\Phi}^T\boldsymbol{t}+\frac{\beta}{2}\boldsymbol{w}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi w}+\boldsymbol{w}^T\frac{a}{2}\boldsymbol{I w} \tag{640}$$

$$=\frac{1}{2}\boldsymbol{w}^T(\beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}+a\boldsymbol{I})\boldsymbol{w}+\beta\boldsymbol{w}^T\boldsymbol{\Phi}^T\boldsymbol{t}+\frac{\beta}{2}\boldsymbol{t}^T\boldsymbol{t} \tag{641}$$

$$\tag{642}$$

Using the symmetry of $\boldsymbol{A}=\Phi^T\Phi+\alpha I$ we can use the matrix completing the square formula with h $=-\frac{1}{2}(A^{-1}\boldsymbol{\Phi}^T\boldsymbol{t})$ and $k=\frac{\beta}{2}\boldsymbol{t}^Tt-\frac{1}{4}\boldsymbol{t}^T\boldsymbol{\Phi}A^{-1}\boldsymbol{\Phi}^T\boldsymbol{t}$ to get

$$(\boldsymbol{w}-h)^TA(\boldsymbol{w}-h)+k \tag{643}$$

and so we only have to show that $k=E(m_N)$

3.19

The integrand in the RHS of the first equation, together with $A=S_N^{-1}$, is part of a MV Gaussian.

$$\int\exp\left\{-\frac{1}{2}(w-m_N)^TA(w-m_N)\right\}dw \tag{644}$$

$$=(2\pi)^{M/2}\left|A\right|^{-1/2}\int(2\pi)^{M/2}\left|A\right|^{-1/2}\exp\left\{-\frac{1}{2}(w-m_N)^TA(w-m_N)\right\}dw \tag{645}$$

$$=(2\pi)^{M/2}\left|A\right|^{-1/2} \tag{646}$$

3.20

I think the most interesting part here is to realise that the determinant of a matrix is the product of its eigenvalues and that we have

$$\beta\Phi^T\Phi=\lambda_iu_i \tag{647}$$

Since $\Phi^T\Phi$ is square

$$\iff\beta\Phi^T\Phi u_i+\alpha Iu_i=\lambda_iu_i+\alpha Iu_i \tag{648}$$

$$\iff(\beta\Phi^T\Phi\alpha I)u_i=\lambda_iu_i+\alpha u_i=(\lambda_i+\alpha)u_i \tag{649}$$

$$\tag{650}$$

And so the eigenvalues can quite easily be computed.

3.21

Considering the result in the previous exercise (eigenvalue is $\lambda_i + \alpha$, we get

$$\frac{d}{d\alpha} \ln |A| = \frac{d}{d\alpha} \ln \prod_{i=1}^{M} (\alpha + \lambda_i) \tag{651}$$

$$= \frac{d}{d\alpha} \sum_{i=1}^{M} \ln(\alpha + \lambda_i) \tag{652}$$

$$= \sum_{i=1}^{M} \frac{1}{\alpha + \lambda_i} \tag{653}$$

$$= \mathrm{Tr} A^{-1} \qquad \text{By spectral theorem, C.46 and C.48} \tag{654}$$

$$= \mathrm{Tr}(A^{-1} \frac{d}{d\alpha} A) \tag{655}$$

Where the last equation is since $\frac{d}{d\alpha} A = \frac{d}{d\alpha} \alpha I + \beta \Phi^T \Phi = I$

3.22

The solutions is pretty detailed in the book. However, I had to realise that if ϵ_i is an eigenvalue of $\Phi^T \Phi$, then $\lambda_i = \beta \epsilon_i$, and therefore $\frac{d\lambda_i}{d\beta} = \epsilon_i = \lambda_i / \beta$

3.23

Like 3.12, this involves multiplying the distributions and then integrating, making use of the fact that the resulting distribution is Gaussian and thus normalized.

3.24

4.4

$$\mathcal{L} = \boldsymbol{w}^T (\boldsymbol{m}_2 - \boldsymbol{m}_1) + \lambda (\boldsymbol{w}^T \boldsymbol{w} - 1) \tag{656}$$

$$\nabla_{\boldsymbol{w}} \mathcal{L} = (\boldsymbol{m}_2 - \boldsymbol{m}_1) + 2\lambda \boldsymbol{w} = 0 \tag{657}$$

$$\boldsymbol{w} = -\frac{1}{2\lambda} (\boldsymbol{m}_2 - \boldsymbol{m}_1) \propto (\boldsymbol{m}_2 - \boldsymbol{m}_1) \tag{658}$$

4.5

It's quite straightforward if we fill in the given equations. Numerator:

$$(m_2 - m_1)^2 = (\boldsymbol{w}^T\boldsymbol{m}_2 - \boldsymbol{w}^T\boldsymbol{m}_1)^2 \tag{659}$$
$$= \boldsymbol{w}^T\boldsymbol{m}_2\boldsymbol{w}^T\boldsymbol{m}_2 - \boldsymbol{w}^T\boldsymbol{m}_2\boldsymbol{w}^T\boldsymbol{m}_1 - \boldsymbol{w}^T\boldsymbol{m}_1\boldsymbol{w}^T\boldsymbol{m}_2 + \boldsymbol{w}^T\boldsymbol{m}_1\boldsymbol{w}^T\boldsymbol{m}_1 \tag{660}$$
$$= \boldsymbol{w}^T(\boldsymbol{m}_2\boldsymbol{m}_2^T - \boldsymbol{m}_2\boldsymbol{m}_1^T - \boldsymbol{m}_1\boldsymbol{m}_2^T + \boldsymbol{m}_1\boldsymbol{m}_1^T)\boldsymbol{w} \tag{661}$$
$$= \boldsymbol{w}^T(\boldsymbol{m}_2 - \boldsymbol{m}_1)(\boldsymbol{m}_2 - \boldsymbol{m}_1)^T\boldsymbol{w} \tag{662}$$

The denominator uses exactly the same approach. I'll show for s_1^2.

$$s_1^2 = \sum_{n\in C_1}(\boldsymbol{w}^T\boldsymbol{x}_n - \boldsymbol{w}^T\boldsymbol{m}_1)^2 \tag{663}$$
$$= \sum_{n\in C_1}(\boldsymbol{w}^T\boldsymbol{x}_n\boldsymbol{w}^T\boldsymbol{x}_n - \boldsymbol{w}^T\boldsymbol{x}_n\boldsymbol{w}^T\boldsymbol{m}_1 - \boldsymbol{w}^T\boldsymbol{m}_1\boldsymbol{w}^T\boldsymbol{x}_n + \boldsymbol{w}^T\boldsymbol{m}_1\boldsymbol{w}^T\boldsymbol{m}_1) \tag{664}$$
$$= \sum_{n\in C_1}\boldsymbol{w}^T(\boldsymbol{x}_n\boldsymbol{x}_n^T - \boldsymbol{x}_n\boldsymbol{m}_1^T - \boldsymbol{m}_1\boldsymbol{x}_n^T + \boldsymbol{m}_1\boldsymbol{m}_1^T)\boldsymbol{w} \tag{665}$$
$$= \sum_{n\in C_1}\boldsymbol{w}^T(\boldsymbol{x}_n - \boldsymbol{m}_1)(\boldsymbol{x}_n - \boldsymbol{m}_1)^T\boldsymbol{w} \tag{666}$$

4.6

$$\sum_{n=1}^{N}(\boldsymbol{w}^T\boldsymbol{x}_n + w_0 - t_n)\boldsymbol{x}_n = 0 \tag{667}$$
$$= \sum_{n=1}^{N}(\boldsymbol{w}^T\boldsymbol{x}_n - \boldsymbol{w}^T\boldsymbol{m} - t_n)\boldsymbol{x}_n \tag{668}$$
$$\boldsymbol{w}^T\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{m})\boldsymbol{x}_n = \sum_{n=1}^{N}t_n\boldsymbol{x}_n \tag{669}$$

This gives us the three terms that we have to solve for.

$$\sum_{n=1}^{N}t_n\boldsymbol{x}_n = \sum_{n\in C_1}^{N_1}t_n\boldsymbol{x}_n - \sum_{n\in C_2}^{N_2}t_n\boldsymbol{x}_n \tag{670}$$
$$= \sum_{n\in C_1}^{N_1}\frac{N}{N_1}\boldsymbol{x}_n - \sum_{n\in C_2}^{N_2}\frac{N}{N_2}\boldsymbol{x}_n \tag{671}$$
$$= N\left(\sum_{n\in C_1}^{N_1}\frac{1}{N_1}\boldsymbol{x}_n - \sum_{n\in C_2}^{N_2}\frac{1}{N_2}\boldsymbol{x}_n\right) \tag{672}$$
$$= N(\boldsymbol{m}_1 - \boldsymbol{m}_2) \tag{673}$$

$$-\boldsymbol{w}^T\boldsymbol{m}\sum_{n=1}^{N}\boldsymbol{x}_n = -\frac{1}{N}\boldsymbol{w}^T(N_1\boldsymbol{m}_1 + N_2\boldsymbol{m}_2)(N_1\boldsymbol{m}_1 + N_2\boldsymbol{m}_2) \tag{674}$$

$$= -\frac{1}{N}(N_1\boldsymbol{m}_1 + N_2\boldsymbol{m}_2)(N_1\boldsymbol{m}_1 + N_2\boldsymbol{m}_2)^T\boldsymbol{w} \tag{675}$$

$$= -\frac{1}{N}(N_1^2\boldsymbol{m}_1\boldsymbol{m}_1^T + N_1N_2\boldsymbol{m}_1\boldsymbol{m}_2^T + N_2N_1\boldsymbol{m}_2\boldsymbol{m}_1^T + N_2^2\boldsymbol{m}_2\boldsymbol{m}_2^T)\boldsymbol{w} \tag{676}$$

$$= -\frac{1}{N}\left((N-N_2)N_1\boldsymbol{m}_1\boldsymbol{m}_1^T + N_1N_2\boldsymbol{m}_1\boldsymbol{m}_2^T + N_2N_1\boldsymbol{m}_2\boldsymbol{m}_1^T + (N-N_1)N_2\boldsymbol{m}_2\boldsymbol{m}_2^T\right)\boldsymbol{w} \tag{677}$$

$$= \left(-N_1\boldsymbol{m}_1\boldsymbol{m}_1^T + \frac{N_1N_2}{N}\boldsymbol{m}_1\boldsymbol{m}_1^T - \frac{N_1N_2}{N}\boldsymbol{m}_1\boldsymbol{m}_2^T - \frac{N_2N_1}{N}\boldsymbol{m}_2\boldsymbol{m}_1^T - N_2\boldsymbol{m}_2\boldsymbol{m}_2^T + \frac{N_1N_2}{N}\boldsymbol{m}_2\boldsymbol{m}_2^T\right)\boldsymbol{w} \tag{678}$$

$$= \left(-N_1\boldsymbol{m}_1\boldsymbol{m}_1^T - N_2\boldsymbol{m}_2\boldsymbol{m}_2^T + \frac{N_1N_2}{N}(\boldsymbol{m}_2-\boldsymbol{m}_1)(\boldsymbol{m}_2-\boldsymbol{m}_1)^T\right)\boldsymbol{w} \tag{679}$$

$$= \left(-N_1\boldsymbol{m}_1\boldsymbol{m}_1^T - N_2\boldsymbol{m}_2\boldsymbol{m}_2^T + \frac{N_1N_2}{N}\boldsymbol{S}_B\right)\boldsymbol{w} \tag{680}$$

$$\tag{681}$$

We add the remaining terms to the final term:

$$\left(-N_1\boldsymbol{m}_1\boldsymbol{m}_1^T - N_2\boldsymbol{m}_2\boldsymbol{m}_2^T + \sum_{n=1}^{N}\boldsymbol{x}_n\boldsymbol{x}_n^T\right)\boldsymbol{w} \tag{682}$$

$$= \left(N_1\boldsymbol{m}_1\boldsymbol{m}_1^T - 2N_1\boldsymbol{m}_1\boldsymbol{m}_1^T + N_2\boldsymbol{m}_2\boldsymbol{m}_2^T - 2N_2\boldsymbol{m}_2\boldsymbol{m}_2^T + \sum_{n=1}^{N}\boldsymbol{x}_n\boldsymbol{x}_n^T\right)\boldsymbol{w} \tag{683}$$

$$= \left(N_1\boldsymbol{m}_1\boldsymbol{m}_1^T - \sum_{n\in C_1}\boldsymbol{x}_n\boldsymbol{m}_1^T - \boldsymbol{m}_1\sum_{n\in C_1}\boldsymbol{x}_1^T + ... + \sum_{n=1}^{N}\boldsymbol{x}_n\boldsymbol{x}_n^T\right)\boldsymbol{w} \tag{684}$$

$$= \left(\sum_{n\in C_1}\boldsymbol{m}_1\boldsymbol{m}_1^T - \boldsymbol{x}_n\boldsymbol{m}_1^T - \boldsymbol{m}_1\boldsymbol{x}_n^T + \boldsymbol{x}_n\boldsymbol{x}_n^T + ...\right)\boldsymbol{w} \tag{685}$$

Which gives the product we need. "..." denotes symmetric steps but for the class 2.

4.7

$$\sigma(-a) = \frac{1}{e^x+1} = \frac{e^{-x}}{1+e^{-x}} = \frac{e^{-x}+1}{e^{-x}+1} - \frac{1}{e^{-x}+1} \tag{686}$$

$$y = \frac{1}{1+e^{-x}} \tag{687}$$

$$e^{-x} = \frac{1-y}{y} \tag{688}$$

$$y = e^x(1-y) \tag{689}$$

$$e^x = y/(1-y) \tag{690}$$
$$x = \ln\left[y/(1-y)\right] \tag{691}$$

4.8

$$p(C_1|\boldsymbol{x}) = \sigma(a) \tag{692}$$

So we have to show: $a = \boldsymbol{w}^T\boldsymbol{x} + w_0$

$$a = \ln\frac{p(\boldsymbol{x}|C_1)}{p(\boldsymbol{x}|C_2)} + \ln\frac{p(C_1)}{p(C_2)} \tag{693}$$

$$\ln\frac{p(\boldsymbol{x}|C_1)}{p(\boldsymbol{x}|C_2)} = \ln\left[\frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)\right)\right] \tag{694}$$
$$- \ln\left[\frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)\right)\right] \tag{695}$$
$$= -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\boldsymbol{x}-\mu_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2) \tag{696}$$
$$= \frac{1}{2}\left[2\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - 2\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 + \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right] \tag{697}$$
$$\tag{698}$$

From which the result easily follows.

4.9

Hint 1: Using a Lagrange multiplier, make sure $\sum_{j=1}^k \pi_j = 1$ before optimizing.

$$\ln p(\boldsymbol{X}|\boldsymbol{T}) = \sum_{n=1}^N \ln\prod_{j=1}^K (\pi_j\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma})_j^t) \tag{699}$$
$$= \sum_{n=1}^N\sum_{j=1}^K t_j \ln(\pi_j\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k\boldsymbol{\Sigma})) \tag{700}$$

Now we optimize with constraint $\sum_{j=1}^K \pi_j = 1$.

$$\mathcal{L}(\boldsymbol{\pi},\lambda) = \ln(p(\boldsymbol{X},\boldsymbol{T})) - \lambda(\sum_{j=1}^K \pi_j - 1) \tag{701}$$

$$\frac{\partial\mathcal{L}}{\partial\pi_j} = \sum_{n=1}^N \frac{t_j}{\pi_j} - \lambda = 0 \tag{702}$$

$$\lambda = \sum_{n=1}^N \frac{t_j}{\pi_j} = N\frac{t_j}{\pi_j} = \frac{N_j}{\pi_j} \tag{703}$$

$$\pi_j = \frac{N_j}{\lambda} \tag{704}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = \sum_{j=1}^{K} \pi_j = 1 \tag{705}$$

Plugging this into 380.

$$\sum_{j=1}^{K} \frac{N_j}{\lambda} = 1 \iff \lambda = N \tag{706}$$

Gives us the desired result.

4.12

$$\frac{d\sigma}{da} = (1 + e^{-a})^{-2} e^{-a} \tag{707}$$

$$= \frac{1}{1 + e^{-a}} \frac{1}{1 + e^{-a}} e^{-a} \tag{708}$$

$$= \sigma(a) \frac{e^{-a}}{1 + e^{-a}} \tag{709}$$

$$= \sigma(a) \left[\frac{1 + e^{-a}}{1 + e^{-a}} - \frac{1}{1 + e^{-a}} \right] \tag{710}$$

4.13

$$\nabla_{\boldsymbol{w}} E(\boldsymbol{w}) = -\sum_{n=1}^{N} \frac{t_n}{y_n} \nabla_{\boldsymbol{w}} \sigma(\boldsymbol{w}^T \boldsymbol{\phi}_n) - \frac{1 - t_n}{1 - y_n} \nabla_{\boldsymbol{w}} \sigma(\boldsymbol{w}^T \boldsymbol{\phi}_n) \tag{711}$$

$$= -\sum_{n=1}^{N} \frac{t_n}{y_n} \sigma(\boldsymbol{w}^T \boldsymbol{\phi}_n)(1 - \sigma(\boldsymbol{w}^T \boldsymbol{\phi}_n)) \boldsymbol{\phi}_n - \frac{1 - t_n}{1 - y_n} \sigma(\boldsymbol{w}^T \boldsymbol{\phi}_n)(1 - \sigma(\boldsymbol{w}^T \boldsymbol{\phi}_n)) \boldsymbol{\phi}_n \tag{712}$$

$$= -\sum_{n=1}^{N} t_n (1 - y_n) \boldsymbol{\phi}_n - (1 - t_n) y_n \boldsymbol{\phi}_n \tag{713}$$

$$= -\sum_{n=1}^{N} (t_n - y_n) \boldsymbol{\phi}_n \tag{714}$$

4.14

Hint 1: approach it with an argument, using that we have a perfect decision boundary at $\boldsymbol{w}^T \boldsymbol{\phi} = 0$.

We know that if C_1 is labelled with $t_{C_1} = 1$ and C_2 is labelled with $t_{C_2} = 0$ then we want $p(C_1|\phi) = \sigma(\boldsymbol{w}^T \phi) > 0.5$ and $p(C_2|\phi) = \sigma(\boldsymbol{w}^T \phi) < 0.5$ which happens if the decision boundary perfectly separates them at $\boldsymbol{w}^T \phi = 0$. Now the binary cross entropy will be minimal as $p(C_1|\phi) \to 1$ which happens when $\boldsymbol{w} \to \infty$. And vice versa.

4.16

$$p(\boldsymbol{t}, \boldsymbol{w}) = \prod_{n=1}^{N} y_n^{\pi}[1 - y_n]^{1-t_n} \tag{715}$$

$$\ln p = \sum_{n=1}^{N} \pi_n \ln y_n + (1 - \pi_n) \ln(1 - y_n) \tag{716}$$

4.17

$$p(C_k|\boldsymbol{\phi}) = y_k = \frac{\exp a_k}{\sum_{j=1} \exp a_j} \tag{717}$$

$$\frac{\partial y_k}{\partial a_j} = -\exp a_k (\sum_j \exp(a_j)^{-2} \exp(a_j) \tag{718}$$

$$= \begin{cases} y_k(0 - y_j) & j \neq k \\ y_k(1 - y_j) & j = k \end{cases} \tag{719}$$

$$= y_k(I_{kj} - y_j) \tag{720}$$

4.18

$$\nabla_{\boldsymbol{w}_j} E(\boldsymbol{W}) = -\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \nabla_{\boldsymbol{w}_j} \ln y_n k \tag{721}$$

$$\nabla_{\boldsymbol{w}_j} \ln(y_{nk}) = -(I_{kj} - y_j)\boldsymbol{\phi}_n \tag{722}$$

$$\nabla_{\boldsymbol{w}_j} E(\boldsymbol{W}) = -\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk}(I_{kj} - y_j n)\boldsymbol{\phi}_n \tag{723}$$

$$= \boldsymbol{\phi} \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} y_{jn} \boldsymbol{\phi} - t_{nk} I_{kj} \tag{724}$$

$$= \boldsymbol{\phi} \sum_{n=1}^{N} t_{nj} - y_{jn} \boldsymbol{\phi} \underbrace{\sum_{k=1}^{K} t_{nk}}_{=1} \tag{725}$$

$$= \sum_{n=1}^{N} \boldsymbol{\phi}(y_{jn} - t_{nj}) \tag{726}$$

4.19

Hint 1: Use binary cross netropy and 4.114 as the activation function. Use the fundamental theorem of calculus.

$$p(\boldsymbol{t}, \boldsymbol{w}) = \sum_{n=1}^{N} [t_n \ln y_n + (1 - t_n \ln(1 - y_n] \tag{727}$$

$$\nabla_{\boldsymbol{w}} \sum_{n=1}^{N} [t_n \ln \Phi(a) + (1 - t_n) \ln(1 - \Phi(a))] = \sum_{n=1}^{N} (\frac{t_n}{\Phi(a)} - \frac{1 - t_n}{1 - \Phi(a)}) \Phi(a) \boldsymbol{\phi}_n \tag{728}$$

4.21

$$\Phi(a) = \int_0^a \mathcal{N}(0, 1) d\theta \tag{729}$$

$$= \int_0^a 1/\sqrt{2\pi} \exp(-\frac{1}{2}\theta^2) d\theta \tag{730}$$

$$= \frac{1}{2} + \int_{-\infty}^a \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}\theta^2) d\theta \tag{731}$$

$$= \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \frac{\sqrt{\pi}}{2} \int_{-\infty}^a \frac{2}{\sqrt{\pi}} \exp(-\frac{1}{2}\theta^2) d\theta \tag{732}$$

$$= \frac{1}{2} \left(1 + \frac{1}{\sqrt{2}} \text{erf}(a) \right) \tag{733}$$

4.22

$$\ln p(D) = \ln \left[f(z_0) \frac{(2\pi)^{M/2}}{|A|^{1/2}} \right] \tag{734}$$

$$= \ln f(z_0) = \frac{M}{2} \ln 2\pi - \frac{1}{2} \ln |A| \tag{735}$$

z_0 is the location of the $\boldsymbol{\theta}_{MAP}$ estimate so

$$\ln f(\boldsymbol{\theta}) \Big| z_0 = \ln f(\boldsymbol{\theta}_{MAP}) = \ln p(D|\boldsymbol{\theta}_{MAP}) + \ln p(\boldsymbol{\theta}_{MAP}) \tag{736}$$

5.2

$$p(\boldsymbol{T}, \boldsymbol{X}, \boldsymbol{w}) = \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{t}_n | y(\boldsymbol{x}_n, \boldsymbol{w}), \beta^{-1} I) \tag{737}$$

$$= \prod_{n=1}^{N} \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\beta^{-1} I|^{\frac{1}{2}}} \exp(-\frac{1}{2} (\boldsymbol{t}_n - y(\boldsymbol{x}_n, \boldsymbol{w}))^T \beta^{-1} I (\boldsymbol{t}_n - y(\boldsymbol{x}_n, \boldsymbol{w}))) \tag{738}$$

Now it's obvious that if we take the the log likelihood it cancels the exp and we end up with

$$(-\frac{1}{2}(\boldsymbol{t}_n - y(\boldsymbol{x}_n, \boldsymbol{w}))^T \beta^{-1} I(\boldsymbol{t}_n - y(\boldsymbol{x}_n, \boldsymbol{w}))) = -\frac{1}{2\beta}\|y(\boldsymbol{x}_n, \boldsymbol{w}) - \boldsymbol{t}_n\|^2. \tag{739}$$

5.5

$$p(\boldsymbol{T}, \boldsymbol{X}, \boldsymbol{w}) = \prod_{n=1}^{N}\prod_{k=1}^{K} y_k(\boldsymbol{x}_n, \boldsymbol{w})^{t_n^k}(1 - y_k)(\boldsymbol{x}_n, \boldsymbol{w})^{1-t_n^k} \tag{740}$$

Taking the log becomes the cross-entropy function.

5.6

$$\frac{\partial E(\boldsymbol{w})}{\partial a_k} = \sum_{n=1}^{N}\left(\frac{\partial}{\partial a_k} t_n \ln y_n + \frac{\partial}{\partial a_k}(1 - t_n)\ln(1 - y_n)\right) \tag{741}$$

$$= -\frac{t_k}{y_k}\frac{\partial}{\partial a_k} y_k - \frac{1 - t_k}{1 - y_k}\frac{\partial}{\partial a_k}(1 - y_k) \tag{742}$$

$$= -t_k(1 - y_k) + (1 - t_k)y_k \tag{743}$$

$$= y_k - t_k \tag{744}$$

5.7

$$E(\boldsymbol{w}) = \sum_{k=1}^{K} t_k \ln y_k(\boldsymbol{x}, \boldsymbol{w}) \tag{745}$$

$$\frac{\partial E}{\partial a_j} = -\sum_{k=1}^{K}\frac{t_k}{y_k} y_k(I_{kj} - y_j) \tag{746}$$

$$= -\sum_{k=1}^{K} t_k(I_{kj} - y_j) \tag{747}$$

$$= -t_j + y_j \tag{748}$$

$$\frac{d\tanh}{da} = (e^a - e^{-a})\frac{d}{da}(e^a + e^{-a})^{-1} + (e^a + e^{-a})^{-1}\frac{d}{da}(e^a - e^{-a}) \tag{749}$$

$$= -(e^a - e^{-a})(e^a + e^{-a})^{-2}\frac{d}{da}(e^a + e^{-a}) + (e^a + e^{-a})^{-1}(e^a + e^{-a}) \tag{750}$$

$$= -(e^a - e^{-a})(e^a + e^{-a})^{-2}(e^a - e^{-a}) + 1 \tag{751}$$

$$= -h^2(a) + 1 \tag{752}$$

5.9

Still Bernoulli, so

$$p(t|\boldsymbol{x}, \boldsymbol{w}) = (\frac{1+y}{2})^{\frac{1+t}{2}}(\frac{1-y}{2})^{\frac{1-t}{2}} \tag{753}$$

$$p(\boldsymbol{t}|\boldsymbol{X}, \boldsymbol{w}) = \prod_{n=1}^{N} p(t_n|\boldsymbol{x}_n, \boldsymbol{w}) \tag{754}$$

$$-\ln(p(\boldsymbol{t}|\boldsymbol{X}, \boldsymbol{w}) = -\sum_{n=1}^{N} \ln p(t_n|\boldsymbol{x}_n, \boldsymbol{w}) \tag{755}$$

$$= -\sum_{n=1}^{N} (\frac{1+t}{2}) \ln(\frac{1+y}{2}) + (\frac{1-t}{2}) \ln(\frac{1-y}{2}) \tag{756}$$

We can use tanh as activation function.

5.10

$$u_i^T H u_i = u_i^T \lambda u_i = \delta_i i \lambda_i = \lambda_i \tag{757}$$

We started with 5.37, so this will always be positive.

The converse direction is the one in the book.

5.13

Hint 1: $\boldsymbol{b}$ has W parameters and H is $W \times W$ We already know H has $\frac{N(N+1)}{2}$ parameters and b has N parameters.

$$\frac{N(N+1)}{2} + N = \frac{N(N+3)}{2} \tag{758}$$

5.14

Hint 1: Taylor expansion on both terms in numerator.

Taylor:

$$E_n(w_j + \epsilon) = E_n(w_{ji}) + \epsilon \frac{\partial E_n}{\partial w_{ji}} + \frac{\epsilon^2}{2} \frac{\partial \partial E_n}{\partial w_{ji}} + O(\epsilon^3) \tag{759}$$

$$E_n(w_j - \epsilon) = E_n(w_{ji}) - \epsilon \frac{\partial E_n}{\partial w_{ji}} + \frac{\epsilon^2}{2} \frac{\partial \partial E_n}{\partial w_{ji}} - O(\epsilon^3) \tag{760}$$

Subtracting these and solving for the partial derivative shows that the second order terms cancel.

5.16

$$E = \frac{1}{2}\sum_{n=1}^{N}\|\boldsymbol{y}_n - \boldsymbol{t}_n\|^2 = \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{M}(\boldsymbol{y}_n - \boldsymbol{t}_n)^2 \tag{761}$$

$$\nabla E = \sum_{n=1}^{N}\sum_{m=1}^{M}(\boldsymbol{y}_n - \boldsymbol{t}_n)\nabla \boldsymbol{y}_n \tag{762}$$

$$\nabla\nabla E = \sum_{n=1}^{N}\sum_{m=1}^{M}\nabla \boldsymbol{y}_n \nabla(\boldsymbol{y}_n - \boldsymbol{t}_n) + (\boldsymbol{y}_n - \boldsymbol{t}_n)\nabla\nabla \boldsymbol{y}_n \approx \sum_{n=1}^{N}\sum_{m=1}^{M}\nabla_{\boldsymbol{y}_n}\nabla \boldsymbol{y}_n^T \tag{763}$$

5.17

Hint 1: Use $y(\boldsymbol{x}, \boldsymbol{w}) = \int tp(t|x)dt$

$$\frac{\partial E}{\partial w_r} = \int\int (y(x,w) - t)\frac{\partial y}{\partial w_r}p(x,t)dxdt \tag{764}$$

$$\frac{\partial^2 E}{\partial w_r \partial w_s} = \int\int \left[(y(x,w) - t)\frac{\partial^2 y(x,w)}{\partial w_r \partial w_s} + \frac{\partial y(x,w)}{\partial w_r}\frac{\partial y}{\partial w_s}\right]p(x,t)dxdt \tag{765}$$

$$\int\int (y(x,w) - t)p(x,t)dxdt = \int\int (y(x,w) - t)p(t|x)p(x)dxdt \tag{766}$$

$$= \int p(x)\left(y(x,w) - \underbrace{\int tp(t|x)}_{=y(x,w)}\right)dxdt \qquad = 0 \tag{767}$$

The remaining integral is the answer.

5.18

$$y_k(\boldsymbol{x}, \boldsymbol{w}) = \sigma(\sum_{j=1}^{M} w_{kj}h(\sum_{i=1}^{D} w_{ji}x_i + w_{j0}) + \sum_{l=1}^{D} w_l i x_i) \tag{768}$$

Finding the derivatives to these skip weights is straightforward.

5.19

$$E = -\sum_{n=1}^{N} t_n \ln y_n + (1 - t_n)\ln(1 - y_n) \tag{769}$$

$$\nabla E = \sum_{n=1}^{N}(y_n - t_n)\nabla a_n \qquad \text{(Taken from earlier solutions)}$$

$$\nabla\nabla E = \sum_{n=1}^{N} y_n(1 - y_n)\nabla a_n \nabla a_n^T + (y_n - t_n)\nabla\nabla a_n \tag{770}$$

5.20

$$\nabla_{w_j} E(\boldsymbol{W}) = \sum_{n=1}^{N} (y_{n_j} - bt_{n_j}) \nabla a_j \tag{771}$$

$$\nabla\nabla \boldsymbol{w}_j \approx \sum_{n=1} y_k (I - y_j) \nabla a_j \nabla a_j^T \tag{772}$$

5.24

$$\sum_i \frac{1}{a} w_{ji}(ax_i + b) + w_{j0} - \frac{b}{a} \sum_i w_{ji} = \sum_i w_{ji} x + \frac{b}{a} w_{ji} + w_{j0} - \frac{b}{a} \sum_i w_{ji} \tag{773}$$

$$= \sum_i w_{ji} x_i + w_{j0} \tag{774}$$

y_k scaling is similar.

5.28

If we normally have $y = \sum_{j=0}^{M} w_{kj} z_j$ we now have $y_k = \sum_{j=0}^{M} w_k z_j$. Therefore the backprop becomes $\frac{\partial}{\partial w_k} = \sum_{j=0}^{M} z_j$. I.e., the weights are updated according to the outputs that the generated for all receptive fields and summed.

5.29

$$\tilde{E}(\boldsymbol{w}) = E(\boldsymbol{w}) + \lambda \Omega(\boldsymbol{w}) \tag{775}$$

$$\frac{\partial \tilde{E}}{\partial w_i} = \frac{\partial E}{\partial w_i} + \frac{\partial}{\partial w_i} \lambda \Omega(\boldsymbol{w}) \tag{776}$$

$$\frac{\partial}{\partial w_i} \lambda \Omega(\boldsymbol{w}) = -\lambda \frac{\partial}{\partial w_i} \ln(\sum_{j=1}^{M} \pi_j \mathcal{N}(w_i | \mu_j, \sigma_j^2)) \tag{777}$$

$$\frac{\partial}{\partial w_i} = \frac{1}{\sum_{j=1}^{M} \pi_j \mathcal{N}(w_i | \mu_j, \sigma_j^2)} \frac{\partial}{\partial w_i} \ln(\sum_{j=1}^{M} \pi_j \mathcal{N}(w_i | \mu_j, \sigma_j^2)) \tag{778}$$

$$\frac{\partial}{\partial w_i} \ln(\sum_{j=1}^{M} \pi_j \mathcal{N}(w_i | \mu_j, \sigma_j^2)) = \sum_{j=1}^{M} \pi_j \frac{\partial}{\partial w_i} \mathcal{N}(w_i | \mu_j, \sigma_j^2) \tag{779}$$

$$\frac{\partial}{\partial w_i} \mathcal{N}(w_i | \mu_j, \sigma_j^2) = \frac{1}{\sqrt{2\pi\sigma_j^2} \exp(-\frac{1}{2} \frac{(w_i - \mu_j)^2}{\sigma_j^2})} \frac{\partial}{\partial w_i} (-\frac{1}{2} \frac{(w_i - \mu_j)^2}{\sigma_j^2}) \tag{780}$$

$$\frac{\partial}{\partial w_i} = -\sigma_j^{-2} (w_i - \mu_j) \tag{781}$$

Plugging everything in gives

$$\frac{\partial}{\partial w_i} = \frac{\sum_{j=1}^{M} \pi_j \mathcal{N}(w_i|\mu_j, \sigma_j)}{\sum_{k=1}^{K} \mathcal{N}(w_i|\mu_j, \sigma_j^2)} \tag{782}$$

5.30

$$\frac{\partial \tilde{E}}{\partial \mu_j} = -\lambda \frac{\partial}{\partial \mu_j} \sum_i \ln \sum_{j=1}^{M} \pi_j \mathcal{N}(w_i|\mu_j, \sigma_j) \tag{783}$$

$$= -\lambda \sum_i \frac{1}{\sum_j \pi_j \mathcal{N}(w_i} \tag{784}$$

$$= -\lambda \sum_i \frac{1}{\sum_j \pi_j N(w_i|\mu_j, \sigma_j)} \sum_j \pi_j \frac{d}{d\mu_j} N(w_i|\mu_j, \sigma_j) \tag{785}$$

$$\frac{d}{d\mu_j} N(w_i|\mu_j, \sigma_j) = -\frac{1}{\sqrt{2\pi\sigma_i^2}} \exp(-\frac{1}{2\sigma_i^2}(w_i - \mu_j)^2)[\frac{1}{\sigma_j^2}(w_i - \mu_j] \tag{786}$$

Plugging in gives the result.

5.31

$$\frac{\partial}{\partial \sigma_j} \pi_j \, \mathcal{N}(w_i|\mu_j, \sigma_j) \tag{787}$$

$$= \pi \left[\exp\left(-\frac{(\mu_j - w_i)^2)}{2\sigma_j^2} \right) \frac{\partial}{\partial \sigma_j} \frac{1}{\sqrt{2\pi\sigma_j^2}} + \frac{1}{\sqrt{2\pi\sigma_j^2}} \frac{\partial}{\partial \sigma_j} \exp\left(-\frac{(\mu_j - w_i)^2}{2\sigma_j^2} \right) \right] \tag{788}$$

$$\frac{\partial}{\partial \sigma_j} (2\pi\sigma_j^2)^{-\frac{1}{2}} = -\frac{1}{2} (2\pi\sigma_j^2)^{\frac{3}{2}} \frac{\partial}{\partial \sigma_j} (2\pi\sigma_j^2) \tag{789}$$

$$= -\frac{1}{2} (2\pi\sigma_j^2)^{-\frac{3}{2}} 4\pi\sigma_j \tag{790}$$

$$= -\frac{1}{\sigma_j} \frac{1}{\sqrt{2\pi\sigma_j^2}} \tag{791}$$

$$\frac{\partial}{\partial \sigma_j} \exp\left(-\frac{(\mu_j - w_i)^2}{2\sigma_j^2} \right) = \exp\left(-\frac{(\mu_j - w_i)^2}{2\sigma_j^2} \right) \frac{\partial}{\partial \sigma_j} \left(-\frac{1}{2\sigma_j^2} (\mu_j - w_i)^2 \right) \tag{792}$$

$$= \exp\left(-\frac{(\mu_j - w_i)^2}{2\sigma_j^2} \right) \frac{1}{\sigma_j^3} (\mu_j - w_i)^2 \qquad = \tag{793}$$

Plugging these values in gives the result.

5.33

If we start with $\boldsymbol{v} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ then using trigonometry:

$$\boldsymbol{y} = \begin{pmatrix} \cos(\pi - \theta)L_1 + v_1 \\ \sin(\pi - \theta)L_1 + v_2 \end{pmatrix} \tag{794}$$

$$\boldsymbol{x} = \begin{pmatrix} y_1 + \cos(\theta_1 + \theta_2 - \pi)L_2 \\ y_2 + \sin(\theta_1 + \theta_2 - \pi)L_2 \end{pmatrix} \tag{795}$$

5.34

$$\frac{\partial E_n}{\partial a_{kl}^{\pi}} = -\frac{1}{\sum_{l=1} \pi_l(\boldsymbol{x}_n, \boldsymbol{w})\mathcal{N}(\boldsymbol{t}_n|\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}), \sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w}))} \frac{\partial}{\partial a_k^{\pi}} \sum_{l=1} \pi_l(\boldsymbol{x}_n, \boldsymbol{w})\mathcal{N}(\boldsymbol{t}_n|\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}), \sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w})) \tag{796}$$

$$\frac{\partial}{\partial a_k^{\pi}} \sum_{l=1} \pi_l(\boldsymbol{x}_n, \boldsymbol{w})\mathcal{N}(\boldsymbol{t}_n|\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}), \sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w})) = \mathcal{N}(\boldsymbol{t}_n|\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}), \sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w})) \frac{\partial}{\partial a_k^{\pi}} \sum_{l=1} \pi_k \tag{797}$$

$$\frac{\partial}{\partial a_k^{\pi}} \sum_{l=1} \pi_k = \sum_{l=1} \pi_k(I_{kl} - \pi_l) \tag{798}$$

$$= \pi_k - \pi_k \sum_{l=1} \pi_l \tag{799}$$

$$\frac{\partial E_n}{\partial a_{kl}^{\pi}} = -\frac{\pi_k \mathcal{N}(\boldsymbol{t}_n|\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}), \sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w})) - \pi_k \sum_{l=1} \pi_l \mathcal{N}(\boldsymbol{t}_n|\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}), \sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w}))}{\sum_{l=1} \pi_l(\boldsymbol{x}_n, \boldsymbol{w})\mathcal{N}(\boldsymbol{t}_n|\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}), \sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w}))} \tag{800}$$

$$= -\gamma_k(w) + \pi_k \tag{801}$$

5.35

$$\frac{\partial E_n}{\partial a_{kl}^{\mu}} = -\frac{1}{\sum_{l=1} \pi_l(\boldsymbol{x}_n, \boldsymbol{w})\mathcal{N}(\boldsymbol{t}_n|\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}), \sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w}))} \frac{\partial}{\partial a_k^{\mu}} \sum_{l=1} \pi_l(\boldsymbol{x}_n, \boldsymbol{w})\mathcal{N}(\boldsymbol{t}_n|\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}), \sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w})) \tag{802}$$

$$\frac{\partial}{\partial a_k^{\mu}} \sum_{l=1} \pi_l(\boldsymbol{x}_n, \boldsymbol{w})\mathcal{N}(\boldsymbol{t}_n|\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}), \sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w})) \tag{803}$$

$$= \pi_k(\boldsymbol{x}_n, \boldsymbol{w})\mathcal{N}(\boldsymbol{t}_n|\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}), \sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w})) \frac{\partial}{\partial a_k^{\mu}} - \frac{1}{2}(\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}) - \boldsymbol{t}))^T \sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w})^{-1}(\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}) - \boldsymbol{t})) \tag{804}$$

$$\tag{805}$$

$$\frac{\partial}{\partial a_k^\mu} - \frac{1}{2}(\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}) - \boldsymbol{t}))^T \sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w})^{-1}(\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}) - \boldsymbol{t})) \tag{806}$$

$$= -\sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w})^{-1}(\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}) - \boldsymbol{t}) \underbrace{\frac{\partial \boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w})}{\partial a_k^\mu}}_{=1} \tag{807}$$

$$\frac{\partial E_n}{\partial a_{kl}^\mu} = \gamma_{nk}\sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w})^{-1}(\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}) - \boldsymbol{t}) \tag{808}$$

5.36

$$\frac{\partial E_n}{\partial a_{kl}^\sigma} = -\frac{1}{\sum_{l=1} \pi_l(\boldsymbol{x}_n, \boldsymbol{w})\mathcal{N}(\boldsymbol{t}_n|\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}), \sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w}))} \frac{\partial}{\partial a_k^\sigma} \sum_{l=1} \pi_l(\boldsymbol{x}_n, \boldsymbol{w})\mathcal{N}(\boldsymbol{t}_n|\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}), \sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w})) \tag{809}$$

$$\frac{\partial}{\partial a_k^\sigma} \sum_{l=1} \pi_l(\boldsymbol{x}_n, \boldsymbol{w})\mathcal{N}(\boldsymbol{t}_n|\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}), \sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w})) = \pi_k(\boldsymbol{x}_n, \boldsymbol{w}) \frac{\partial}{\partial a_k^\sigma} \mathcal{N}(\boldsymbol{t}_n|\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}) \tag{810}$$

$$\frac{\partial}{\partial a_k^\sigma}\mathcal{N}(\boldsymbol{t}_n|\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}) = \frac{\partial}{\partial a_k^\sigma} \frac{1}{(2\pi)^{\frac{L}{2}}} \frac{1}{|\sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w})|^{\frac{1}{2}}} \exp(-\frac{1}{2}(\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}) - \boldsymbol{t}))^T \sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w})^{-1}(\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}) - \boldsymbol{t})) \tag{811}$$

$$= \frac{1}{(2\pi)^{\frac{L}{2}}} \Bigg(\exp(-\frac{1}{2}(\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}) - \boldsymbol{t}))^T \sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w})^{-1}(\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}) - \boldsymbol{t})) \frac{\partial}{\partial a_k^\sigma} \frac{1}{|\sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w})|^{\frac{1}{2}}} \tag{812}$$

$$+ \frac{1}{|\sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w})|^{\frac{1}{2}}} \frac{\partial}{\partial a_k^\sigma} \exp\Big(-\frac{1}{2}(\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}) - \boldsymbol{t}))^T \sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w})^{-1}(\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}) - \boldsymbol{t})\Big) \Bigg) \tag{813}$$

$$\frac{\partial}{\partial a_k^\sigma} \frac{1}{|\sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w})|^{\frac{1}{2}}} = \frac{\partial}{\partial a_k^\sigma} \frac{1}{\sigma_k(\boldsymbol{x}, \boldsymbol{w})^L} = -\frac{L}{\sigma^2(\boldsymbol{x}, \boldsymbol{w})^{L-1}} = \frac{L}{\sigma_k} \frac{1}{\sigma_k(\boldsymbol{x}, \boldsymbol{w})^L} = \frac{L}{\sigma_k} \frac{1}{|\sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w})|^{\frac{1}{2}}} \tag{814}$$

$$\frac{\partial}{\partial a_k^\sigma}(-\frac{1}{2}(\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}) - \boldsymbol{t}))^T \sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w})^{-1}(\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}) - \boldsymbol{t}) = \frac{\|\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}) - \boldsymbol{t}\|^2}{\sigma_k^3} \tag{815}$$

Plugging in these values gives the correct result. The book's answer has a typo.

5.37

5.38

5.39

Using the approximation:

$$p(D|\alpha,\beta) = \int p(D|w,\beta)p(w|\alpha)dw \approx p(D,w_{map})p(w_{map})\frac{(2\pi)^{W/2}}{|A|^{\frac{1}{2}}} \tag{816}$$

Taking the log gives the result.

5.40

Use a softmax activation.

6.3

$$\|x - x_n\| = x^T x + x_n^T x - 2x_n^T x \tag{817}$$
$$= k(x,x) + k(x_n,x) - 2k(x_n,x) \tag{818}$$

6.4

$$\begin{pmatrix} 2 & 0 \\ -1 & 3 \end{pmatrix} \tag{819}$$

6.5

$$ck(\boldsymbol{x},\boldsymbol{x}') = c\psi^T\psi = (c^{\frac{1}{2}}\psi)^T(c^{\frac{1}{2}}\psi) = \phi^T\phi \tag{820}$$
$$f(\boldsymbol{x})k(\boldsymbol{x},\boldsymbol{x}')f(\boldsymbol{x}') = f(\boldsymbol{x})\psi^T\psi f(\boldsymbol{x}') = (f(\boldsymbol{x})\psi^T)^T(\psi f(\boldsymbol{x}')) = \phi^T\phi \tag{821}$$

6.6

$$q(k(x,x')) = \sum_{n=1}^{N} \beta_n k(x,x')^n \tag{822}$$
$$= \sum_{n=1}^{N} \beta_0 k(x,x') \qquad 6.18 \tag{823}$$
$$= \sum_{n=1}^{N} k(x,x') \qquad 6.13 \tag{824}$$
$$= k(x,x') \qquad 6.17 \tag{825}$$

6.16 follows from the previuos proof and the definition of the exponential function.

6.7

$$\psi(x)^T\psi(x') + \theta(x)^T\theta(x') = \sum_{n=1}^{N}(\psi(x_n) + \theta(x_n))(\psi(x'_n) + \theta(x'_n)) - \psi(x_n)\theta(x'_n) - \theta(x_n)\psi(x'_n) \quad (826)$$

$$= \begin{bmatrix}\dots & \psi(x_n) + \theta(x_n) & -\psi(x_n) & -\theta(x_n) & \dots\end{bmatrix} \begin{bmatrix}\vdots \\ \psi(x'_n) + \theta(x'_n) \\ \theta(x'_n) \\ \psi(x'_n) \\ \vdots\end{bmatrix} \quad (827)$$

$$= \phi(x)^T\phi(x') \quad (828)$$

$$k_1(x, x')k_2(x, x') = \theta(x)^T\theta(x')(\psi(x)^T\psi(x') \quad (829)$$

$$= \sum_{n=1}^{N}\theta(x_n)\theta(x'_n)\psi(x_n)\psi(x'_n) \quad (830)$$

$$= \sum_{n=1}^{N}\theta(x_n)\psi(x_n)\theta(x'_n)\psi(x'_n) \quad (831)$$

$$= \phi(x)^T\phi(x') \quad (832)$$

$$= k(x, x') \quad (833)$$

6.8

$$k_3(\phi(x), \phi(x')) = \psi(\phi(x))^T\psi(\phi(x')) = k(x, x') \quad (834)$$

$$x^TAx' = x^TU^TUx' \quad \text{(Symmetry)}$$

$$= (Ux)^TUx' \quad (835)$$

$$= \phi(x)^T\phi(x') \quad (836)$$

$$= k(x, x') \quad (837)$$

6.9

$$\sum_{n=1}^{N} \psi(x_{na})\psi(x'_{na}) + \sum_{m=1}^{M} \theta(x_{mb})\theta(x'_{mb}) \tag{838}$$

$$= \begin{bmatrix} \psi(x_{a1}) & \dots & \psi(x_{aN}) & \theta(x_{b1}) & \dots & \theta(x_{bM}) \end{bmatrix} \begin{bmatrix} \psi(x'_{a1}) \\ \vdots \\ \psi(x'_{aN}) \\ \theta(x'_{b1}) \\ \vdots \\ \theta(x'_{bM}) \end{bmatrix} = k(x, x') \tag{839}$$

$$x \in \mathbb{R}^n \tag{840}$$

g bijective, $f : \mathbb{R}^n \to \mathbb{R}$

$$y = f(g(x)) \tag{841}$$

$$y'(x) = \nabla_{g(x)} f(g(x)) J_x(g(x)) \tag{842}$$

$$k_a(x_a, x'_a)k(x_b, x'_b) = \sum_{n=1}^{N} \theta(x_{an})\theta(x'_{an}) \sum_{m=1}^{M} \psi(x_{bm})\psi(x'_{bm}) \tag{843}$$

$$= \sum_{n=1}^{N} \sum_{m=1}^{M} \theta(x_{an})\theta(x'_{an})\psi(x_{bm})\psi(x'_{bm}) \tag{844}$$

$$= \sum_{n=1}^{N} \sum_{m=1}^{M} \theta(x_{an})\psi(x_{bm})\theta(x'_{an})\psi(x'_{bm}) \tag{845}$$

$$= \phi(x)^T \phi(x') = k(x, x') \tag{846}$$

6.10

$$y(x) = \sum_{n=1}^{N} f(x_n) f(x_n)(K + \lambda I_N)^{-1} t \tag{847}$$

which is proportional to f?

6.11

Just by plugging in we see that we observe an infinite amount of terms in the kernel dot-product, and therefore the vectors are of infinite dimensionality.

6.13

If $\boldsymbol{\phi}$ is a invertible differentiable transformation of $\boldsymbol{\theta}$.

$$g(\boldsymbol{\theta}, \boldsymbol{x}) = J_\theta(\phi) \nabla_\phi \ln p(\boldsymbol{x}, \boldsymbol{\phi}) = J_\theta(\boldsymbol{\phi}) g(\boldsymbol{\phi}, \boldsymbol{x}) \tag{848}$$

$$g(\boldsymbol{\phi}, \boldsymbol{x}) = J_\theta^{-1}(\boldsymbol{\phi}) g(\boldsymbol{\theta}, \boldsymbol{x}) \tag{849}$$

$$\boldsymbol{F}' = J_\theta^{-1}(\boldsymbol{\phi}) \, \mathbb{E}_x[g(\boldsymbol{\theta}, \boldsymbol{x}) g(\boldsymbol{\theta}, \boldsymbol{x})^T] J_\theta^{-T}(\boldsymbol{\phi}) = J_\theta^{-1}(\boldsymbol{\phi}) \boldsymbol{F} J_\theta^{-T}(\boldsymbol{\phi}) \tag{850}$$

$$\begin{aligned}
k'(\boldsymbol{x}, \boldsymbol{x}') &= g(\boldsymbol{\phi}, \boldsymbol{x})^T \boldsymbol{F}'^{-1} g(\boldsymbol{\phi}, \boldsymbol{x}) && (851)\\
&= (J_\theta^{-1}(\boldsymbol{\phi}) g(\boldsymbol{\theta}, \boldsymbol{x}))^T (J_\theta^{-1}(\boldsymbol{\phi}) \boldsymbol{F} J_\theta^{-T}(\boldsymbol{\phi}))^{-1} J_\theta^{-1}(\boldsymbol{\phi}) g(\boldsymbol{\theta}, \boldsymbol{x}) && (852)\\
&= g(\boldsymbol{\theta}, \boldsymbol{x})^T J_\theta^{-T}(\boldsymbol{\phi}) J_\theta^T(\boldsymbol{\phi}) \boldsymbol{F} J_\theta(\boldsymbol{\phi}) J_\theta^{-1}(\boldsymbol{\phi}) g(\boldsymbol{\theta}, \boldsymbol{x}) && (853)\\
&= g(\boldsymbol{\theta}, \boldsymbol{x})^T \boldsymbol{F} g(\boldsymbol{\theta}, \boldsymbol{x}) && (854)\\
&= k(\boldsymbol{x}, \boldsymbol{x}') && (855)
\end{aligned}$$

Therefore, the Fisher kernel is invariant.

6.14

$$\begin{aligned}
\nabla_\theta(\ln p(\boldsymbol{x}, \boldsymbol{\theta})) &= \nabla_\mu \ln p(\boldsymbol{x}, \boldsymbol{\mu}) && (856)\\
&= S^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) && (857)
\end{aligned}$$

$$\begin{aligned}
F &= \mathbb{E}[S^{-1}(\boldsymbol{x} - \boldsymbol{\mu})(S^{-1}(\boldsymbol{x} - \boldsymbol{\mu}))^T] && (858)\\
&= S^{-1} \, \mathbb{E}_x[(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^T] S^{-1} && (859)\\
&= S^{-1} S S^{-1} = S^{-1} && (860)
\end{aligned}$$

$$\begin{aligned}
k(\boldsymbol{x}, \boldsymbol{x}') &= (S^{-1}(\boldsymbol{x} - \mu))^T F^{-1} (S^{-1}(\boldsymbol{x} - \boldsymbol{\mu})) && (861)\\
&= (\boldsymbol{x} - \boldsymbol{\mu})^T S^{-1} S S^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) && (862)\\
&= (\boldsymbol{x} - \boldsymbol{\mu})^T S^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) && (863)
\end{aligned}$$

6.15

Since the gram matrix is positive semi-definite we have:

$$|K| = k(x_1, x_1) k(x_2, x_2) - k(x_2, x_1) k(x_1, x_2) \geq 0 \tag{864}$$

from which the result follows.

6.18

$$\boldsymbol{z} = (x, t) \tag{865}$$

$$p(\boldsymbol{z}) = \frac{\sum_n \mathcal{N}(\boldsymbol{z} - \boldsymbol{z}_n, \sigma^2 I)}{\sum_m \int \mathcal{N}(\boldsymbol{z} - \boldsymbol{z}_n, \sigma^2 I) dt} = \frac{\sum_n \mathcal{N}(\boldsymbol{z} - \boldsymbol{z}_n, \sigma^2 I)}{\sum_m \mathcal{N}(x - x_n, \sigma^2 I)} \tag{866}$$

$$\frac{\sum_n \mathcal{N}(\boldsymbol{z} - \boldsymbol{z}_n, \sigma^2 I)}{\sum_m \mathcal{N}(x - x_n, \sigma^2 I)} = \frac{\sum_n \frac{1}{2\pi\sigma^2} \exp[-\frac{1}{2\sigma^2}(\boldsymbol{z} - \boldsymbol{z}_n)^T(\boldsymbol{z} - \boldsymbol{z}_n)]}{\sum_m \frac{1}{\sqrt{2\pi}\sigma} \exp[-\frac{1}{2\sigma^2}(x - x_n)^2]} \tag{867}$$

$$= \frac{\sum_n \frac{1}{2\pi\sigma^2} \exp[-\frac{1}{2\sigma^2}(x - x_n)^2 + -\frac{1}{2\sigma^2}(t - t_n)^2]}{\sum_m \frac{1}{\sqrt{2\pi}\sigma} \exp[-\frac{1}{2\sigma^2}(x - x_n)^2]} \tag{868}$$

$$= \sum_n \frac{\frac{1}{\sqrt{2\pi\sigma^2}} \exp[-\frac{1}{2\sigma^2}(x - x_n)^2}{\sum_m \frac{1}{\sqrt{2\pi}\sigma} \exp[-\frac{1}{2\sigma^2}(x - x_n)^2]} \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-\frac{1}{2\sigma^2}(t - t_n)^2] \tag{869}$$

$$= \sum_n \pi_n \mathcal{N}(t | t_n, \sigma^2) \tag{870}$$

6.24

$$u^T W u = u^T \sqrt{W}\sqrt{W} u \tag{871}$$

$$= (\sqrt{W}u)^T \sqrt{W} u \tag{872}$$

$$> 0 \quad \forall u \in \mathbb{R} \setminus \mathbf{0} \tag{873}$$

$$u^T(W + V)u = u^T W u + u^T V u > 0 \tag{874}$$

6.25

$$a_N = a_N - \nabla\nabla\Psi(a_N)\nabla\Psi(a_n) \tag{875}$$

$$= a_N + (W_N + C^{-1})^{-1}[t_N - \sigma_N - C^{-1}a_N] \tag{876}$$

$$= (W_N + C_N^{-1})[W_N a_N - \sigma_N + t_N] \tag{877}$$

$$= C_N((W_N + C_N^{-1})C_N)^{-1}[W_N a_N - \sigma_N + t_N] \tag{878}$$

$$= C_N(C_N W_N + I)^{-1}[W_N a_N - \sigma_N + t_N] \tag{879}$$

6.26

7.2

$$t_n(\boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}) + b) = \gamma \tag{880}$$

$$L(\boldsymbol{w}, b, \boldsymbol{a}) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{n=1}^{N} a_n(t_n(\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n) + b) - \gamma) \tag{881}$$

$$\nabla_{\boldsymbol{w}} = \boldsymbol{w} - \sum_{n=1}^{N} \nabla a_n t_n \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}_n) \tag{882}$$

$$= \boldsymbol{w} - \sum_{n=1}^{N} a_n t_n \boldsymbol{\phi}(\boldsymbol{x}_n) = 0 \tag{883}$$

$$\iff \boldsymbol{w} = \sum_{n=1}^{N} a_n t_n \boldsymbol{\phi}(\boldsymbol{x}_n) \tag{884}$$

$$\frac{\partial}{\partial b} = -\sum_{n=1}^{N} a_n t_n = 0 \tag{885}$$

$$= \sum_{n=1} a_n t_n \tag{886}$$

$$\tilde{L} = \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n t_n a_m t_m \phi(\boldsymbol{x}_n)^T \boldsymbol{\phi}(\boldsymbol{x}_n) - \sum_{n=1}^{N} a_n(t_n(\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n) + b) - \gamma) \tag{887}$$

$$\sum_{n=1}^{N} a_n(t_n(\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n) + b) - \gamma) = \sum_{n=1}^{N} 0 - a_n\gamma \tag{888}$$

7.6

Since we have $p(t = 1) = \sigma(y)$ and $p(t = -1) = \sigma(-y)$ and $t \in \{-1, 1\}$ we have $p(t|y) = \sigma(ty)$. Therefore

$$-\ln p(t_n) = -\sum_{n=1} \ln \sigma(t_n y_n) \tag{889}$$

which is the cross-entropy error function.

7.7

It is quite straightforward if you use the following steps.

$$\sum_{n=1}^{N} \xi_n(C - \mu_n - a_n) + \sum_{n=1}^{N} \hat{\xi}_n(C - \hat{\mu}_n - \hat{a}_n) = 0 \tag{890}$$

and

$$-\sum_{n=1}^{N} a_n(y_n) - \sum_{n=1}^{N} \hat{a}_n(-y_n) = -\sum_{n=1}^{N}(a_n - \hat{a_n})y_n = 0 \tag{891}$$

7.8

This follows from 7.67 and 7.68.

7.9

$$p(\boldsymbol{w}, \boldsymbol{t}) = \mathcal{N}(\boldsymbol{w}|_N, \boldsymbol{S}_N) \tag{892}$$

$$_N = \boldsymbol{S}_N(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + \beta\boldsymbol{\Phi}^T\boldsymbol{t}) \tag{893}$$

$$\boldsymbol{S}_N^{-1} = \boldsymbol{S}_0^{-1} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi} \tag{894}$$

We have

$$\boldsymbol{S}_N^{-1} = \boldsymbol{S}_0^{-1} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi} = A + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi} = \Sigma^{-1} \tag{895}$$

Considering $\boldsymbol{m}_0 = \boldsymbol{0}$, $\boldsymbol{m}_N$ follows directly.

7.16

7.18

$$\nabla \ln p(\boldsymbol{w}|\boldsymbol{t}, \alpha) = \nabla \sum_{n=1}^{N} t_n \ln y_n + (1 - t_n)\ln(1 - y_n) - \boldsymbol{A}\boldsymbol{w} \tag{896}$$

$$\nabla \sum_{n=1}^{N} t_n \ln y_n + (1 - t_n)\ln(1 - y_n) = \sum_{n=1}^{N} t_n \nabla \ln y_n + (1 - t_n)\nabla(1 - y_n) \tag{897}$$

$$\nabla \ln y_n = (1 - y_n)\phi(\boldsymbol{x}_n) \tag{898}$$

$$\nabla \ln(1 - y_n) = -y_n\phi(\boldsymbol{x}_n) \tag{899}$$

$$\nabla \ln p(\boldsymbol{w}|\boldsymbol{t}, \alpha) = \sum_{n=1}^{N} \phi(\boldsymbol{x}_n)[t_n - y_n t_n - y_n + y_n t_n] - \boldsymbol{A}\boldsymbol{w} \tag{900}$$

$$= \boldsymbol{\Phi}^T[\boldsymbol{t} - \boldsymbol{y}] - \boldsymbol{A}\boldsymbol{w} \tag{901}$$

$$\nabla\nabla \ln p(\boldsymbol{w}|\boldsymbol{t}, \alpha) = \nabla\boldsymbol{\Phi}^T[\boldsymbol{t} - \boldsymbol{y}] - \boldsymbol{A} \tag{902}$$

$$= \nabla \sum_{n=1}^{N} [t_n - y_n]\boldsymbol{\phi}_n - \boldsymbol{A} \tag{903}$$

$$= -\sum_{n=1}^{N} \boldsymbol{\phi}_n \nabla^T y_n - \boldsymbol{A} \tag{904}$$

$$= -\sum_{n=1}^{N} \boldsymbol{\phi}_n y_n(1 - y_n)\boldsymbol{\phi}_n^T - \boldsymbol{A} \tag{905}$$

$$= -\boldsymbol{\Phi}^T\boldsymbol{B}\boldsymbol{\Phi} - \boldsymbol{A} \tag{906}$$

Which gives the result.

8.1

$$\int p(\boldsymbol{x})d\boldsymbol{x} = \int\int p(x_K|pa_K) \prod_{k=1}^{K-1} p(x_k|pa_k)dx_K dx_1 \dots dx_{K-1} \tag{907}$$

$$= \int\int p(x_K|pa_K)dx_K \prod_{k=1}^{K-1} p(x_k|pa_k)dx_1 \dots dx_{K-1} \tag{908}$$

$$= \int \prod_{k=1}^{K-1} p(x_k|pa_k)dx_1 \dots dx_{K-1} \tag{909}$$

$$\vdots \tag{910}$$

$$= \int p(x_1)dx_1 = 1 \tag{911}$$

8.2

For an acyclic graph, if you number any graph (backwards or forwards) and (by accident) encounter a connection that connects a node to a lower-numbered node you can always swap the numbers (and adjust the rest of the graph accordingly) and continue.

8.5

8.6

The constraint $\sum_i \mu_i = 1$ ensures that any $\mu_i : i > 0$ increases the probability. This means that setting the cutoff point at which we predict $y = 1$ at μ_0 ensures the OR-function. The $\mu_i :> 0$ control the increase in probability for that x_i.

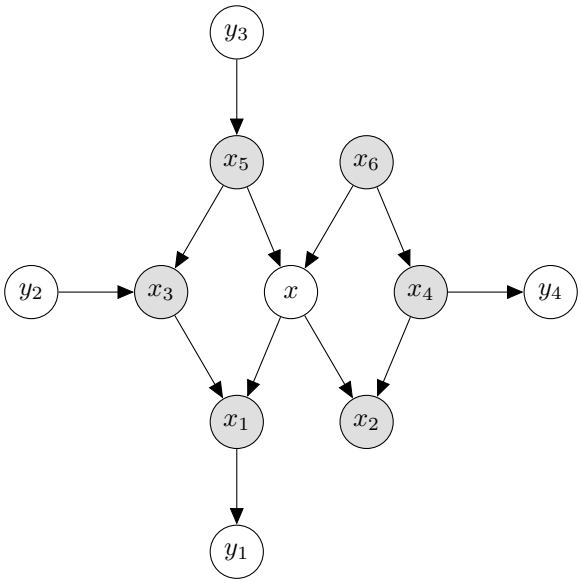
8.8

$$p(a, b, c|d) = \frac{p(a, b, c)}{p(d)} = \frac{p(a|b, c)p(b)p(c)}{p(d)} = \frac{p(a)p(b)p(c)}{p(d)} \tag{912}$$

$$\int \frac{p(a)p(b)p(c)}{p(d)}dc = p(a|d)p(b|d) \tag{913}$$

Which is what we needed to prove.

8.9



$\{y_1, y_2, y_3 y_4\}$ are all the possible connections for the Markov blanket. The path from x to y_1 via x_1 is blocked since x_1 is in C and the path meets head to tail. Same path through x_5 is blocked since they meet tail to tail and x_5 is in C. The path from x to y_2 is blocked by x_3 since the arrows meet head to tail. Same for y_3 to x. All paths have either a head tot tail or tail to tail node with an observed variable in it.

8.10

$$p(a, b, c, d) = p(a)p(b)p(c|a, b)p(d|c) \tag{914}$$

$$p(a, b) = \int \int p(a, b, c, d) dc dd = p(a)p(b) \tag{915}$$

Second part:

$$p(a, b, c|d) = \frac{p(a, b, c, d)}{p(d)} = \frac{p(a)p(b)p(c|a, b)p(d|c)}{p(d)} \tag{916}$$

$$p(a, b|d) = \int p(a, b, c|d) dc = \frac{p(a)p(b)p(d|c)}{p(d)} \tag{917}$$

This does not factor into $p(a|d)p(b|d)$.

8.12

In an undirected graph we can remove or add a link between each node and every other node and that will create a new graph. So $2^{\#\text{Links}}$ graphs. Any node can connect with any other node, giving $N(N-1)$ pairs, but we cannot cound the reverse paths so we divide by two.

8.13

$$E(x,y)_{x_k=1} - E(x,y)_{x_k=-1} = h\sum_i x_i - \beta\sum_{i,j} x_i x_j - \eta\sum_i x_i y_i + h - \beta_x k\sum_j x_j - \eta x_k \sum_i y_i \tag{918}$$

$$- h\sum_i x_i - \beta\sum_{i,j} x_i x_j - \eta\sum_i x_i y_i + h - \beta x_k \sum_j x_j - \eta x_k \sum_i y_i \tag{919}$$

$$= 2h - 2\beta x_k \sum_j x_i - \eta x_k \sum_i y_i \tag{920}$$

8.14

$$p(x,y) = \frac{1}{Z}\exp[-E(x,y)] \tag{921}$$

$$\ln p(x,y) = -\ln(Z) - E(x,y) = -\ln(Z) + \eta\sum_i x_i y_i \tag{922}$$

This is maximized when $x_i = y_i$, i.e. $-1 \cdot -1 = 1$ or $1 \cdot 1 = 1$

8.20

8.22

8.26

9.1

The loss function clearly is convex. Moreover, 9.2 and 9.4 both are guaranteed to lower the function (arg min and an analytical solution for $\boldsymbol{\mu}_k$. Therefore, it will always converge.

9.2

9.3

$$p(\boldsymbol{x},\boldsymbol{z}) = \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \Sigma_k)^{\boldsymbol{z}_k} \pi_k^{\boldsymbol{z}_k} \tag{923}$$

$$p(\boldsymbol{z}) = \sum_z \prod_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{\boldsymbol{z}_k} \tag{924}$$

Since the product is 1 for every $z_k \neq k$ this becomes $\sum_z \pi_{z_k} \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_{z_k}, \boldsymbol{\Sigma}_{z_k}$, which is equal to the required equation 9.7.

9.4

Log posterior: $\ln p(\boldsymbol{\theta}|\boldsymbol{x}) \propto \ln p(\boldsymbol{x}|\boldsymbol{\theta}) + \ln p(\theta)$. For the e-step: Evaluate $p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta}_{old})$, this function only depends on the likelihood part of the objective, so by definition will be the same. For the m-step:

$$\boldsymbol{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}_{old})_{MAP} = \sum_z p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta}_{old}) \ln p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{x}) \tag{925}$$

$$\propto \sum_z p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta}_{old}) \ln p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{x}) \tag{926}$$

$$= \sum_z p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta}_{old}) \ln p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta}) + \ln p(\theta) \sum_z p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta}_{old}) \tag{927}$$

$$= \sum_z p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta}_{old}) \ln p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta}) + \ln p(\theta) \tag{928}$$

9.5

This is obvious, as there simply is no connection between z_m and x_n

9.7

$$\nabla_\mu \ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \pi) = \sum_{n=1}^{N} \sum_{k=1}^{k} \boldsymbol{z}_{kn} \nabla_\mu \ln \mathcal{N}(\boldsymbol{x}_n, \mu_k, \Sigma_k) \tag{929}$$

$$\nabla \ln \mathcal{N}(\boldsymbol{x}_n|\mu_k, \Sigma_k) = -\frac{1}{2} \nabla (\boldsymbol{x}_n - \mu_k)^T \sigma_k^{-1} (\boldsymbol{x}_n - \mu_k) = \Sigma_k^{-1} (\boldsymbol{x}_n - \mu_k) \tag{930}$$

Since z is 1-of-k this gradient only concerns distribution k for every n.

$$\mathcal{L}(\pi_k, \lambda) \ln p(\boldsymbol{X}, \boldsymbol{Z}|\mu_k \Sigma_k \pi_k) + \lambda[-1 + \sum_k \pi_k] \tag{931}$$

$$\frac{\partial \mathcal{L}}{\partial \pi_k} = \sum_{n=1}^{N} z_{nk} \frac{1}{\pi_k} - \lambda \tag{932}$$

$$= \sum_{n=1}^{N} z_{nk} - \pi_k \lambda \tag{933}$$

$$= \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} - \pi_k \lambda \tag{934}$$

$$= N - \lambda = 0 \iff \lambda = N \tag{935}$$

Substituting back:

$$\sum_{n=1}^{N} z_{nk} \frac{1}{\pi_k} = N \iff \pi_k = \frac{1}{N} \sum_{n=1} z_{nk} = \frac{N_k}{N} \tag{936}$$

$$\nabla_{\mu_k} \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk})[\ln \pi_k + \ln \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)] \tag{937}$$

$$= \sum_{n=1}^{N} \gamma(z_n k) \nabla \ln \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{938}$$

$$= \sum_{n=1}^{N} \gamma(z_n k) \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_k) = 0 \tag{939}$$

$$\Rightarrow \sum_{n=1}^{N} \gamma(z_n k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k) \tag{940}$$

$$= \sum_{n=1}^{N} \gamma(z_n k) \boldsymbol{x}_n - \underbrace{\sum_{n=1}^{N} \gamma(\boldsymbol{z}_{nk})}_{N_k} \boldsymbol{\mu}_k = 0 \tag{941}$$

$$\tag{942}$$

Which leads to the correct answer.

9.9

$$\frac{\partial}{\partial \boldsymbol{\Sigma}_k} = \sum_{n=1}^{N} \gamma(z_{nk}[\frac{\partial}{\partial \boldsymbol{\Sigma}_k} \frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_k) - \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \ln |\boldsymbol{\Sigma}|^{-\frac{1}{2}}] \tag{943}$$

$$= \sum_{n=1}^{N} \gamma(z_{nk}[\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T - \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \ln |\boldsymbol{\Sigma}^{-1}|^{\frac{1}{2}}] \tag{944}$$

$$= \sum_{n=1}^{N} \gamma(z_{nk}[\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T - \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \frac{1}{2} \ln |\boldsymbol{\Sigma}^{-1}|] \tag{945}$$

$$= \sum_{n=1}^{N} \gamma(z_{nk}[\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T - \frac{1}{2}\boldsymbol{\Sigma}] = 0 \tag{946}$$

$$\iff \sum_{n=1}^{N} \gamma(z_{nk})[\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T = \sum_{n=1}^{N} \gamma(z_{nk}) \frac{1}{2} \boldsymbol{\Sigma}] = N_k \frac{1}{2} \boldsymbol{\Sigma}_k \tag{947}$$

From which the answer is easily seen.

9.11

$$\mathbb{E}_{\boldsymbol{z}}[\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})] = \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(\boldsymbol{z}_{nk})[\ln \pi_k + \ln \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)] \tag{948}$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(\boldsymbol{z}_{nk})[\ln \pi_k + \ln \frac{1}{(2\pi\epsilon)^{M/2}} - \frac{1}{2\epsilon}\|\boldsymbol{x}_n - \boldsymbol{\mu}_k\|^2] \tag{949}$$

$$\propto \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(\boldsymbol{z}_{nk})[\epsilon \ln \pi_k + \epsilon \ln \frac{1}{(2\pi\epsilon)^{M/2}} - \frac{1}{2}\|\boldsymbol{x}_n - \boldsymbol{\mu}_k\|^2] \tag{950}$$

$$= \lim_{\epsilon \to 0} -\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\|\boldsymbol{x}_n - \boldsymbol{\mu}_k\|^2 + C \tag{951}$$

9.12

$$\mathbb{E}[\boldsymbol{x}] = \int \boldsymbol{x} p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\pi}) d\boldsymbol{x} \tag{952}$$

$$= \int \boldsymbol{x} \sum_{k=1}^{K} \pi_k p(\boldsymbol{x}|\boldsymbol{\mu}_k) d\boldsymbol{x} \tag{953}$$

$$= \sum_{k=1}^{K} \pi_k \int \boldsymbol{x} p(\boldsymbol{x}|\boldsymbol{\mu}_k) d\boldsymbol{x} \tag{954}$$

$$= \sum_{k=1}^{K} \pi_k \boldsymbol{\mu}_k \tag{955}$$

$$\operatorname{cov}[\boldsymbol{x}] = \mathbb{E}[(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])^T] \tag{956}$$

$$= \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T] - \mathbb{E}[\boldsymbol{x}]\,\mathbb{E}[\boldsymbol{x}]^T \tag{957}$$

$$\tag{958}$$

$$\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T] = \int \boldsymbol{x}\boldsymbol{x}^T \sum_{k=1}^{K} \pi_k p(\boldsymbol{x}|\boldsymbol{\mu}_k) d\boldsymbol{x} \tag{959}$$

$$= \sum_{k=1}^{K} \pi_k \int \boldsymbol{x}\boldsymbol{x}^T p(\boldsymbol{x}|\boldsymbol{\mu}_k) d\boldsymbol{x} = \sum_{k=1}^{K} \pi_k\, \mathbb{E}_k[\boldsymbol{x}\boldsymbol{x}^T] \tag{960}$$

$$\tag{961}$$

$$\mathbb{E}_k[\boldsymbol{x}\boldsymbol{x}^T] = \mathbb{E}_k[(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}] + \mathbb{E}[\boldsymbol{x}])(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}] + \mathbb{E}[\boldsymbol{x}])^T] \tag{962}$$

$$= \mathbb{E}\left[(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])^T + (\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])\,\mathbb{E}[\boldsymbol{x}]^T + \mathbb{E}[\boldsymbol{x}][\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}]]^T + \mathbb{E}[\boldsymbol{x}]\,\mathbb{E}[\boldsymbol{x}]^T\right] \tag{963}$$

$$= \underbrace{\operatorname{cov}_k(\boldsymbol{x})}_{=\boldsymbol{\Sigma}_k} + \boldsymbol{\mu}\boldsymbol{\mu}^T \tag{964}$$

9.14

$$p(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\mu})p(\boldsymbol{z}|\boldsymbol{\pi}) = \prod_{k=1}^{K} p(\boldsymbol{x}|\boldsymbol{\mu}_k)^{z_k}\pi_k^{z_k} \tag{965}$$

$$= \sum_{\boldsymbol{z}} \prod_{k=1}^{K} p(\boldsymbol{x}|\boldsymbol{\mu}_k)^{\boldsymbol{z}_k}\pi_k^{\boldsymbol{z}_k} \tag{966}$$

$$\tag{967}$$

Since $\boldsymbol{z}$ is 1-of-K and the inner product only returns when $k = \boldsymbol{z}_k$ this becomes $\sum_{k=1}^{K} p(\boldsymbol{x}|\boldsymbol{\mu}_k)^{\boldsymbol{z}_k}\pi_k^{\boldsymbol{z}_k}$

9.15

$$\frac{\partial}{\partial \mu_{ki}} = \sum_{n=1}^{N} \gamma(z_{nk})[\frac{\partial}{\partial \mu_{ki}} x_{ni} \ln \mu_{ki} + \frac{\partial}{\partial \mu_{ki}}(1 - x_{ni}) \ln(1 - \mu_{ki})] \tag{968}$$

$$= \sum_{n=1}^{N} \gamma(z_{nk}) \left[\frac{x_{ni}}{\mu_{ki}} - \frac{1 - x_{ni}}{1 - \mu_{ki}}\right] \tag{969}$$

$$= \sum_{n=1}^{N} \gamma(z_{nk}) \left[\frac{x_{ni}(1 - \mu_{ki}) - \mu_{ki}(1 - x_{ni})}{\mu_{ki}(1 - \mu_{ki})}\right] = 0 \tag{970}$$

$$\Rightarrow \sum_{n=1}^{N} \gamma(z_{nk})\mu_{ki} = \sum_{n=1}^{N} \gamma(z_{nk})x_n \tag{971}$$

$$= N_k\mu_{ki} = \sum_{n=1}^{N} \gamma(z_{nk})x_n \tag{972}$$

9.16

This solution is exactly the same as exercise 9.7's.

9.17

By observing equation 9.51 we see that since $p(\boldsymbol{x}_n|\boldsymbol{\mu}_k) \leq 1$ and $\sum_k \pi_k = 1$ the maximum of the ln is 0.

9.20

$$\frac{\partial}{\partial \alpha} = \frac{M}{2}\frac{1}{\alpha} - \frac{1}{2}\mathbb{E}[\boldsymbol{w}^T\boldsymbol{w}] = 0 \tag{973}$$

$$\Rightarrow \frac{M}{\alpha} = \mathbb{E}[\boldsymbol{w}^T\boldsymbol{w}] \tag{974}$$

$$\alpha = \frac{M}{\mathbb{E}[\boldsymbol{w}^T\boldsymbol{w}]} \tag{975}$$

9.24

$$\ln p(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_z q(\boldsymbol{z}) \ln p(\boldsymbol{x}|\boldsymbol{\theta}) \tag{976}$$

$$= \sum_z q(\boldsymbol{z}) \ln \frac{p(\boldsymbol{z}, \boldsymbol{x}|\boldsymbol{\theta})}{p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta})} \tag{977}$$

$$= \sum_z q(\boldsymbol{z}) \ln \frac{p(\boldsymbol{z}, \boldsymbol{x}|\boldsymbol{\theta})}{p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta})} \frac{q(\boldsymbol{z})}{q(\boldsymbol{z})} \tag{978}$$

$$= \sum_z q(\boldsymbol{z}) \ln \frac{p(\boldsymbol{z}, \boldsymbol{x}|\boldsymbol{\theta})}{q(\boldsymbol{z})} + \sum_z q(\boldsymbol{z}) \ln \frac{q(\boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta})} \tag{979}$$

$$= \sum_z q(\boldsymbol{z}) \ln \frac{p(\boldsymbol{z}, \boldsymbol{x}|\boldsymbol{\theta})}{q(\boldsymbol{z})} - \sum_z q(\boldsymbol{z}) \ln \frac{p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta})}{q(\boldsymbol{z})} \tag{980}$$

9.25

This directly follows from the fat that the KL-divergence reduces to 0 if the distribution are equal.

9.26

10.1

We've shown this in the previous chapter.

$$\ln p(\boldsymbol{x}) = \int q(\boldsymbol{z}) \ln p(\boldsymbol{x}) d\boldsymbol{z} \tag{981}$$

$$= \int q(\boldsymbol{z}) \ln \frac{p(\boldsymbol{x}, \boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{x})} \tag{982}$$

$$= \int q(\boldsymbol{z}) \ln \frac{p(\boldsymbol{x}, \boldsymbol{z}) q(\boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{x}) q(\boldsymbol{z})} \tag{983}$$

$$= \int q(\boldsymbol{z}) \ln \frac{p(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z})} - \int q(\boldsymbol{z}) \ln \frac{q(\boldsymbol{z})}{p(\boldsymbol{x}|\boldsymbol{z})} \tag{984}$$

$$\tag{985}$$

10.2

This is easily seen by simply filing in the values. The result is a solution to the equations.

10.8

Filling in:

$$\mathbb{E}[\tau] = \frac{a}{b} = (a_0 + \frac{N}{2})(b_0 + \frac{1}{2}\,\mathbb{E}_\mu[\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu})^2 + \lambda_0(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^2]^{-1} \underset{N\to\infty}{\approx} b_N^{-1} \tag{986}$$

Which follows from the fact that the denominator grows proportionally to the numerator.

Similarly, the variance $\mathrm{Var}[\tau] = \frac{a}{b^2}$ is derived but this time the denominator grows quadratically and therefore goes to zero in the limit $N \to \infty$.

10.10

$$\ln p(\boldsymbol{x}) = \sum_m \sum_{\boldsymbol{z}} q(\boldsymbol{z}, m) \ln p(\boldsymbol{x}) \tag{987}$$

$$= \sum_m \sum_z q(\boldsymbol{z}, m) \ln \frac{p(\boldsymbol{x}, m, \boldsymbol{z})}{p(m, \boldsymbol{z}|\boldsymbol{x})} \frac{q(\boldsymbol{z}, m)}{q(\boldsymbol{z}, m)} \tag{988}$$

Which results in 10.35 after rearrangement.

10.15

We have:

$$\mathbb{E}[\pi_k] = \frac{\alpha_k}{\sum_{k=1}^{K} \alpha_k} \tag{989}$$

and $\alpha_k = \alpha_0 + N_k$.

$$\mathbb{E}[\pi_k] = \frac{\alpha_0 + N_k}{\sum_{k=1}^{K} \alpha_0 + N_k} \tag{990}$$

$$= K\alpha_0 + \underbrace{\sum_{k=1}^{K} N_k}_{=N} \tag{991}$$

10.21

This is easily observed by the fact that if we fix 1 distribution we still have $K - 1$ combinations. If we continue this cascade you end up with $K \cdot K - 1 \ldots 2 \cdot 1$ distributions, which is $K!$

10.29

$$\frac{dd\ln(x)}{dx} = -x^{-2} \tag{992}$$

which is negative everywhere so $\ln(x)$ is concave.

Taylor approximation:

$$y(x) = \ln(\xi) + \frac{1}{\xi}(x - \xi) = \lambda x - \ln(\lambda) - 1 \quad \lambda = \frac{1}{\xi} \tag{993}$$

$$g(\lambda) = \min_x [\lambda x - f(x)] \tag{994}$$

$$\frac{d}{dx} = \lambda - \frac{1}{x} = 0 \iff x = \frac{1}{\lambda} \tag{995}$$

Therefore $g(\lambda) = 1 - \ln \frac{1}{\lambda} = 1 + \ln \lambda$.

$$\frac{d}{d\lambda}\lambda x - g(\lambda) = x - \frac{1}{\lambda} \iff x = \frac{1}{\lambda} \tag{996}$$

Plugging this into $y(x)$ gives the result $\ln x$

10.30

$$\frac{d}{dx} = \frac{1}{1 + e^{-x}} e^{-x} = \sigma(x) e^{-x} \tag{997}$$

$$\frac{dd}{dx} = \sigma(x)(1 - \sigma(x))e^{-x} - e^{-x}\sigma(x) = -\sigma^2(x)e^{-x} \tag{998}$$

Both functions are positive everywhere, showing that the function itself is negative, therefore concave.

$$\frac{d}{dx} = \frac{1}{1 + e^{-x}} e^{-x} \tag{999}$$

$$\frac{dd}{dx} = \sigma(x)(1 - \sigma(x))e^{-x} - e^{-x}\sigma(x) = -\sigma^2(x)e^{-x} \tag{1000}$$

Taylor:

$$f(x) = f(\xi) + f'(\xi)(x - \xi) + \mathcal{O}(\xi) \tag{1001}$$

Since the approximation is linear and $f(x)$ is concave it must be that the LHS is smaller-equal to the RHS.

$$f(x) \leq -\ln(1 + e^{-xi}) + \sigma(\xi)e^{-\xi}(x - \xi) + \mathcal{O}(\xi^2) \tag{1002}$$

$$= -\ln(1 + e^{-xi}) + \sigma(\xi)e^{-\xi}x - \sigma(\xi)e^{-\xi}\xi + \mathcal{O}(\xi^2) \tag{1003}$$

$$= \lambda x - g(\lambda) \quad \lambda = \sigma(\xi)e^{-\xi} \tag{1004}$$

10.33

$$\frac{d}{d\xi_n} = (1 - \sigma(\xi_n)) - \frac{1}{2} - \boldsymbol{\phi}_n^T \mathbb{E}[\boldsymbol{w}\boldsymbol{w}^T]\boldsymbol{\phi}_n \lambda'(\xi_n) + 2\xi_n\lambda(\xi_n) + \lambda'(\xi_n)\xi_n^2 \tag{1005}$$

$$= -2\xi_n\lambda(\xi_n) - \boldsymbol{\phi}_n^T \mathbb{E}[\boldsymbol{w}\boldsymbol{w}^T]\boldsymbol{\phi}_n \lambda'(\xi_n) + 2\xi_n\lambda(\xi_n) + \lambda'(\xi_n)\xi_n^2 \tag{1006}$$

$$= -\lambda'(\xi_n)(\xi_n^2 - \boldsymbol{\phi}_n^T \mathbb{E}[\boldsymbol{w}\boldsymbol{w}^T]\boldsymbol{\phi}_n) = 0 \tag{1007}$$

$$\iff \xi_n^2 = \boldsymbol{\phi}_n^T \mathbb{E}[\boldsymbol{w}\boldsymbol{w}^T]\boldsymbol{\phi}_n \tag{1008}$$

10.37

$$q^{\setminus j}(\boldsymbol{\theta}) = \frac{q(\boldsymbol{\theta})}{\tilde{f}_j(\boldsymbol{\theta}} \tag{1009}$$

$$Z_j = \int q^{\setminus j}(\boldsymbol{\theta}) f_j(\boldsymbol{\theta}) d\theta) = \int q^{\setminus j}(\boldsymbol{\theta}) \tilde{f}_j(\boldsymbol{\theta}) d\theta) = \int (q(\boldsymbol{\theta}) d\theta = 1 \tag{1010}$$

$$\tilde{f}_j(\boldsymbol{\theta}) = \frac{q^{new}}{q^{\setminus j}(\boldsymbol{\theta})} = \tilde{f}_j = f_j \tag{1011}$$

11.1

$$\mathbb{E}[\hat{f} = \mathbb{E}[\frac{1}{L} \sum_{l=1}^{L} f(z^l) \tag{1012}$$

$$= \frac{1}{L} \sum_{p} (\boldsymbol{z}) f(\boldsymbol{z}^l) d\boldsymbol{z} \tag{1013}$$

$$= \frac{1}{L} \sum_{l=1}^{L} \mathbb{E}[f] = \mathbb{E}[f] \tag{1014}$$

For the variance, we have to note that

$$\mathbb{E}[f(\boldsymbol{z}^l), f(\boldsymbol{z}^k)] = \mathrm{Var}(f(\boldsymbol{z})) + \mathbb{E}[f(\boldsymbol{z})]^2 \tag{1015}$$

And $\mathrm{Var}(f(\boldsymbol{z})) = 0$ for $k \neq l$.

$$\mathrm{Var}[\hat{f}] = \mathbb{E}[\hat{f}^2] - \mathbb{E}[\hat{f}]^2 \tag{1016}$$

$$= \frac{1}{L^2} (\sum_{l=1}^{L} \sum_{k=1}^{L} \mathbb{E}[f(\boldsymbol{z}^l) f(\boldsymbol{z}^k)]) - \mathbb{E}[f]^2 \tag{1017}$$

$$= \frac{1}{L^2} (\sum_{l=1}^{L} \sum_{k=1}^{L} \mathbb{E}[f(\boldsymbol{z}^l), f(\boldsymbol{z}^k)]) - \mathbb{E}[f]^2 \tag{1018}$$

$$= \frac{1}{L^2} L \mathrm{Var}(f(\boldsymbol{z})) + L^2 \, \mathbb{E}[f]^2 - \mathbb{E}[f]^2 \tag{1019}$$

$$= \frac{1}{L} \mathrm{Var}(f(\boldsymbol{z})) \tag{1020}$$

11.2

What we need to show is that we can transform z into any distribution if we use 11.6.

$$p(y) = 1 \cdot |\frac{dz}{dy}| = 1 \cdot p(y) \tag{1021}$$

11.3

$$z = (\frac{1}{\pi}\int_{-\infty}^{y}\frac{1}{1+\hat{y}^2}) \tag{1022}$$
$$= (\frac{1}{\pi}[\arctan(y) - \arctan(-\infty)]) \tag{1023}$$
$$= (\frac{1}{\pi}\arctan(y) + \frac{1}{2}) \tag{1024}$$

$$y = h^{-1}(z) \tag{1025}$$
$$\iff \pi z - \frac{\pi}{2} = \arctan(y) \tag{1026}$$
$$\iff h^{-1}(z) = \tan(\pi z - \frac{\pi}{2}) \tag{1027}$$

11.5

Hint 1: Show that the expectation and covariance are equal to μ and Σ.

$$\mathbb{E}[y] = \mathbb{E}_z[\boldsymbol{\mu}] + \mathbb{E}_z[\boldsymbol{Lz}] \tag{1028}$$
$$= \mathbb{E}_z[\boldsymbol{\mu}] + \boldsymbol{L}\,\mathbb{E}_z[\boldsymbol{z}] \tag{1029}$$
$$= \mathbb{E}_z[\boldsymbol{\mu}] + \boldsymbol{L0} \tag{1030}$$
$$= \boldsymbol{\mu} \tag{1031}$$

$$\text{cov}[\boldsymbol{y}] = \mathbb{E}[(\boldsymbol{y}-\boldsymbol{\mu})(\boldsymbol{y}-\boldsymbol{\mu})^T] \tag{1032}$$
$$= \mathbb{E}[\boldsymbol{y}\boldsymbol{y}^T - \boldsymbol{\mu}\boldsymbol{y}^T - \boldsymbol{y}\boldsymbol{\mu}^T + \boldsymbol{\mu}\boldsymbol{\mu}^T] \tag{1033}$$
$$= \mathbb{E}[(\boldsymbol{\mu}+\boldsymbol{Lz})(\boldsymbol{\mu}+\boldsymbol{Lz})^T - \boldsymbol{\mu}(\boldsymbol{\mu}+\boldsymbol{Lz})^T - (\boldsymbol{\mu}+\boldsymbol{Lz})\boldsymbol{\mu}^T + \boldsymbol{\mu}\boldsymbol{\mu}^T] \tag{1034}$$
$$= \mathbb{E}[\boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\mu}(\boldsymbol{Lz})^T + \boldsymbol{Lz}\boldsymbol{\mu}^T + \boldsymbol{Lz}(\boldsymbol{Lz})^T - \boldsymbol{\mu}\boldsymbol{\mu}^T - \boldsymbol{\mu}(\boldsymbol{Lz})^T - \boldsymbol{\mu}\boldsymbol{\mu}^T - \boldsymbol{Lz}\boldsymbol{\mu}^T + \boldsymbol{\mu}\boldsymbol{\mu}^T] \tag{1035}$$
$$= \mathbb{E}[\boldsymbol{Lz}(\boldsymbol{Lz})^T] \tag{1036}$$
$$= \boldsymbol{L}\,\mathbb{E}[\boldsymbol{z}\boldsymbol{z}^T]\boldsymbol{L}^T \tag{1037}$$

Now, we know that $\text{Var}[\boldsymbol{z}] = \mathbb{E}[\boldsymbol{z}\boldsymbol{z}^T] + \boldsymbol{\mu}\boldsymbol{\mu}^T$, therefore $\mathbb{E}[\boldsymbol{z}\boldsymbol{z}^T] = \boldsymbol{I}$. Then the result follows.

11.7

$$p(y) = p(z)|\frac{dz}{dy}| \tag{1038}$$

We have $z \sim \mathcal{U}(0,1)$ and therefore $p(z) = \frac{1}{1-0} = 1$. Inverting the given equation:

$$y = b\tan z + c \iff \frac{y-c}{b} = \tan z \iff \arctan\frac{y-c}{b} = z \tag{1039}$$

$$\frac{dz}{dy} = \frac{1}{1 + \left(\frac{y-c}{b}\right)^2} \frac{1}{b} \tag{1040}$$

$$\tag{1041}$$

Multiplying these two gives the desired result without the scaling constant k.

11.10

11.12

Since there are regions with zero conditional probability, Gibbs sampling will not be ergodic.

11.14

Hint: calculate $\mathbb{E}[z_i']$ and $E[(z_i' - \mu_i)^2$.

$$\mathbb{E}_{z,\nu}[z_i'] = \mathbb{E}[\mu_i] + \mathbb{E}[\alpha(z_i - \mu_i) + \mathbb{E}[\sigma_i(1-\alpha_i^2)^{\frac{1}{2}}\nu] \tag{1042}$$

$$= \mu_i \tag{1043}$$

$$\mathrm{Var}_{z,\nu} = \mathbb{E}[z_i'^2] - \mathbb{E}[z_i']^2 \tag{1044}$$

$$= \mathbb{E}[(\mu_i + \alpha(z_i - \mu_i) + \sigma_i(1-\alpha_i^2)^{\frac{1}{2}}\nu)(\mu_i + \alpha(z_i - \mu_i) + \sigma_i(1-\alpha_i^2)^{\frac{1}{2}}\nu)] - \mu_i^2 \tag{1045}$$

$$= \mathbb{E}[(\mu_i + \alpha(z_i - \mu_i))(\mu_i + \alpha(z_i - \mu_i) + \sigma_i(1-\alpha_i^2)^{\frac{1}{2}}\nu)(\mu_i + \alpha(z_i - \mu_i) \tag{1046}$$

$$+ (\mu_i + \alpha(z_i - \mu_i)(\sigma_i(1-\alpha_i^2)^{\frac{1}{2}}\nu) + (\sigma_i(1-\alpha_i^2)^{\frac{1}{2}}\nu)(\sigma_i(1-\alpha_i^2)^{\frac{1}{2}}\nu)] - \mu_i^2 \tag{1047}$$

$$\vdots \tag{1048}$$

$$\tag{1049}$$

11.15

$$\frac{dH}{dr_i} = \frac{1}{Z}\frac{d}{dr_i} r_i^2 = r_i \tag{1050}$$

$$\frac{dr_i}{d\tau} = -\frac{dE(z)}{dz_i} = -\frac{dH}{dz_i} \tag{1051}$$

11.16

$$p(\boldsymbol{r}|\boldsymbol{z}) = \frac{p(\boldsymbol{r},\boldsymbol{z})}{p(\boldsymbol{z})} = \frac{Z_p}{Z_H}\exp(-K(\boldsymbol{r})) \tag{1052}$$

11.17

12.3

$$\|u_i\|^2 = [\frac{1}{(N\lambda_i)^{1/2}} X^T v_i]^T [\frac{1}{(N\lambda_i)^{1/2}} X^T v_i] \tag{1053}$$
$$= (N\lambda_i)^{-1} [X^T v_i]^T [X^T v_i] \tag{1054}$$
$$= (N\lambda_i)^{-1} v^T X X^T v_i \tag{1055}$$
$$= \lambda_i^{-1} v^T \lambda_i v_i \tag{1056}$$
$$= 1 \tag{1057}$$

And therefore $\|u_i\| = 1$

12.4

This problem can be solved by using 2.113-2.115 and just filling in the variables.

12.9

$$\frac{\partial \ln p(\boldsymbol{X}|\boldsymbol{\mu}, \boldsymbol{W}, \sigma^2)}{\partial \boldsymbol{\mu}} = -\frac{1}{2} \frac{\partial}{\partial \boldsymbol{\mu}} \sum_{n=1}^{N} (\boldsymbol{x}_n - \boldsymbol{\mu})^T \boldsymbol{C}^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}) \tag{1058}$$
$$= \sum_{n=1}^{N} \boldsymbol{C}^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}) = 0 \tag{1059}$$
$$\Rightarrow 0 = -N\boldsymbol{\mu} + \sum_{n=1}^{N} \boldsymbol{x}_n \tag{1060}$$
$$\Rightarrow \boldsymbol{\mu} = \bar{\boldsymbol{x}} \tag{1061}$$

12.14

$$D(D-1) + 1 - (D-1)(D-2)/2 = D^2 - D + 1 - \frac{1}{2}D^2 + \frac{3}{2}D - 1 = \frac{1}{2}D^2 + \frac{1}{2}D = D(D+1)/2 \tag{1062}$$

12.18

$$M \times D + 1 \tag{1063}$$

where $W \in \mathbb{R}^{M \times D}$ and sigma is 1-dimensional.

13.1

x_{n+2} is d-separated from x_n Since x_{n+1} is observed and the nodes meet head-to-tail.

13.3

Using the d-separation criterion, we see that there is always a path connecting any two observed variables xn and xm via the latent variables, and that this path is never blocked. Thus the predictive distribution $p(x_{n+1}|x_1, \ldots, x_n)$ for observation x_{n+1} given all previous observations does not exhibit any conditional independence properties, and so our predictions for x_{n+1} depends on all previous observations.

13.6

If π_k is 0, three is no probability density for latent variable z_k. This means that z_k will always be 0 and therefore the update steps $\gamma(z_{nk})$ and $\xi(z_{n-1}, z_{nk})$ will also be 0.

13.7

$$\frac{\partial Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}}{\boldsymbol{\mu}_k} = \sum_{n=1}^{N} \gamma(z_{nk}) \frac{1}{p(\boldsymbol{x}_n|\boldsymbol{\phi}_k)} \frac{\partial}{\partial \boldsymbol{\mu}^k} p(\boldsymbol{x}_n|\boldsymbol{\phi}_k) \tag{1064}$$

$$= \sum_{n=1}^{N} \gamma(z_{nk}) \frac{1}{p(\boldsymbol{x}_n|\boldsymbol{\phi}_k)} p(\boldsymbol{x}_n|\boldsymbol{\phi}_k) \frac{\partial}{\partial \boldsymbol{\mu}_k} - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \tag{1065}$$

$$= 0 \iff \sum_{n=1}^{N} \gamma(z_{nk}) \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_k) = 0 \tag{1066}$$

$$\iff \boldsymbol{\mu} = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) \boldsymbol{x}_n}{\sum_{n=1}^{N} \gamma(z_{nk})} \tag{1067}$$

$$\tag{1068}$$

$$\frac{\partial Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})}{\boldsymbol{\Sigma}_k} = \sum_{n=1}^{N} \gamma(z_{nk}) \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \ln p(\boldsymbol{x}|\boldsymbol{\phi}_k) \tag{1069}$$

$$\tag{1070}$$

$$\frac{\partial}{\partial \boldsymbol{\Sigma}_k} \ln p(\boldsymbol{x}|\boldsymbol{\phi}_k) = \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \left[\ln \frac{1}{(2\pi)^{D/2}} + \ln \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} - \frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \right] \tag{1071}$$

$$= \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \frac{1}{2} \ln |\boldsymbol{\Sigma}^{-1}| - \frac{1}{2} \frac{\partial}{\partial \boldsymbol{\Sigma}_k} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}) \tag{1072}$$

$$= \frac{1}{2}\boldsymbol{\Sigma} - \frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T \tag{1073}$$

$$\frac{\partial}{\partial \boldsymbol{\Sigma}_k} = \sum_{n=1}^{N} \gamma(z_{nk}) \left(\frac{1}{2}\boldsymbol{\Sigma} - \frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T \right) = 0 \tag{1074}$$

$$\iff \boldsymbol{\Sigma}_k = \frac{\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T}{\sum_{n=1}^{N}\gamma(z_{nk})} \tag{1075}$$

13.17

$$h(z_1) = p(z_1|u_1)p(x_1|z_1, u_1) \tag{1076}$$

$$f_n(z_{n-1}, z_n) = p(z_n|z_{n-1}, u_n)p(x_n|z_n, u_n) \tag{1077}$$

13.19

This follows from:

$$\frac{d}{dz}\ln p(\boldsymbol{Z}) = \frac{d}{dz}\ln p(z_0)p(z_1|z_0) \times \cdots \times p(z_n|z_{n-1}) = \begin{bmatrix} \frac{\partial}{z_0}\ln p(z_0) \\ \frac{\partial}{z_1}\ln p(z_1|z_0) \\ \vdots \\ \frac{\partial}{z_n} p(z_n|z_{n-1}) \end{bmatrix} \tag{1078}$$

So individually optimizing the conditional distributions will optimize the total distribution.

13.27

If the noise goes to zero, $\boldsymbol{\Sigma} = \mathbf{0}$. Considering $C = I$: Note that $K_1 = V_0(V_0 + \Sigma) = I$ Therefore $\mu_1 = \mu_0 + x_1 - \mu_0 = x_1$

$V_n = (I - K_n C)P_{n-1} = 0$. Therefore $P_{n-1} = \Gamma$. Now every term in 13.86 and 13.87 is directly dependent on x, except for Γ. I don't see how this term doesn't influence the distribution.

14.2

$$\mathbb{E}\left[\left(\frac{1}{M}\sum_{m=1}^{M}\epsilon_m(x)\right)^2\right] = \mathbb{E}\left[\left(\frac{1}{M^2}\sum_{m=1}^{M}\sum_{l=1}^{M}\epsilon_m(x)\epsilon_l(x)\right)\right] \tag{1079}$$

$$= \mathbb{E}\left[\frac{1}{M^2}\left(\sum_{m=1}^{M}\sum_{\substack{l=1 \\ l\neq m}}^{M}\epsilon_m(x)\epsilon_l(x) + \sum_{l=1}^{M}\epsilon_l(x)^2\right)\right] \tag{1080}$$

$$= \frac{1}{M^2}\sum_{l=1}^{M}\mathbb{E}[\epsilon_l(x)^2] \tag{1081}$$

$$= \frac{1}{M}E_{AV} \tag{1082}$$

14.3

Through Jensen's inequality:

$$\sum_{m=1}^{M} \frac{1}{M} \epsilon_m(x)^2 \geq \left(\sum_{m=1}^{M} \frac{1}{M} \epsilon_m(x) \right)^2 \tag{1083}$$

Therefore:

$$\mathbb{E}_x[\sum_{m=1}^{M} \frac{1}{M} \epsilon_m(x)^2] \geq \mathbb{E}\left[\left(\sum_{m=1}^{M} \frac{1}{M} \epsilon_m(x) \right)^2 \right] = E_{com} \tag{1084}$$

14.6

$$\frac{d}{d\alpha_m} = (e^{\alpha_m/2} + e^{-\alpha/2}) \sum_{n=1}^{N} w_n^{(m)} I(y_m(x_n) = t_n) = e^{\alpha_m/2} \sum_{n=1}^{N} w_n^{(m)} = 0 \tag{1085}$$

$$\frac{e^{-\alpha_m/2}}{e^{\alpha_m/2} + e^{-\alpha_m/2}} = \epsilon_m \tag{1086}$$

Rearranging this results in the desired formula.

14.7 14.8

14.9

$$\mathcal{L}_m(\boldsymbol{x}_n) = (y_n - \sum_{l=1}^{M} \alpha_l \hat{y}_l)^2 \tag{1087}$$

$$= (\underbrace{y - \sum_{l=1}^{M-1} \alpha_l \hat{y}_l}_{\text{residual}} + \alpha_M y_{\hat{M}})^2 \tag{1088}$$

$$\tag{1089}$$

14.10

$$L = \frac{1}{2} \sum_{n=1}^{N} (t_n - t)^2 \tag{1090}$$

$$\frac{dL}{dt} = -\sum_{n=1}^{N} (t_n - t) = 0 \iff Nt = \sum_{n=1}^{N} t_n \iff t = \frac{1}{N} \sum_{n=1}^{N} t_n \tag{1091}$$

14.13

$$p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) = \prod_{n=1}^{N} \prod_{k=1}^{K} (\pi_k \mathcal{N}(t_n|\boldsymbol{w}_k^T \boldsymbol{\phi}_n, \beta^{-1})^{z_{nk}} \tag{1092}$$

Taking the log of this will yield the result.

14.14

$$\frac{dL}{d\pi_k} = \sum_{n=1}^{N} \gamma_{nk}/\pi_k - \lambda = 0 \tag{1093}$$

$$\iff \sum_k \sum_{n=1}^{N} \gamma_{nk} = \sum_k \pi_k \lambda \tag{1094}$$

$$\tag{1095}$$

Since summing over k is equal to marginalizing $p(z|\theta_k)$ we get $\sum_{n=1}^{N} 1 = \lambda \iff \lambda = N$ Substituting back results in $\pi_k = \frac{1}{N} \sum_{N=1} \gamma_{nk}$.

14.15

$$\mathbb{E}[t|\hat{\boldsymbol{\phi}}, \boldsymbol{\theta}] = \sum_{k=1}^{K} \pi_k \, \mathbb{E}[t|\hat{\boldsymbol{\phi}}, \boldsymbol{w}_k, \beta] \tag{1096}$$