Bishop Questions

David Ruhe

March 30, 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

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

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

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

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

1.5

$$var[f] = \mathbb{E}\left[(f(x) - \mathbb{E}\left[f(x)\right])^2\right] \tag{17}$$

$$= \mathbb{E}\left[f(x)^2 + \mathbb{E}\left[f(x)\right]^2 - 2f(x)\,\mathbb{E}[f(x)]\right] \tag{18}$$

$$= \mathbb{E}[f(x)^2] + \mathbb{E}[f(x)]^2 - 2\,\mathbb{E}[f(x)]\,\mathbb{E}[f(x)] \tag{19}$$

$$= \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2 \tag{20}$$

1.6

$$cov[x,y] = \mathbb{E}_{x,y}[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}] \tag{21}$$
$$= \int_{x,y} p(x,y)[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}] \tag{22}$$
$$= \int_x \int_y p(x)p(y)[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}] \tag{23}$$
$$= \int_x p(x)\{x - \mathbb{E}[x]\} \int_y p(y)\{y - \mathbb{E}[y]\} \tag{24}$$
$$= \mathbb{E}[x - \mathbb{E}[x]]\,\mathbb{E}[y - \mathbb{E}[y]] \tag{25}$$
$$= (\mathbb{E}[x] - \mathbb{E}[x])(\mathbb{E}[y] - \mathbb{E}[y]) \tag{26}$$

1.9

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{27}$$
$$\frac{\partial \mathcal{N}(x|\mu,\sigma^2)}{\partial \mu} = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \sigma^2(x-\mu) = 0 \tag{28}$$
$$x - \mu = 0 \tag{29}$$
$$x = \mu \tag{30}$$

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^2} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right] \tag{31}$$
$$\frac{\partial \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\Sigma)}{\partial \boldsymbol{\mu}} = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^2} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right] \left[-\frac{1}{2}(\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-1^T})(\boldsymbol{x}-\boldsymbol{\mu})\right] = 0 \tag{32}$$
$$\tag{33}$$

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

1.10

$$\mathbb{E}[x+z] = \int_{x+z} p(x+z)\,[x+z] \tag{34}$$
$$= \int_{x+z} p(x+z)x + \int_{x+z} p(x+z)z \tag{35}$$
$$= \int_{x,z} p(x,z)x + \int_{x,z} p(x,z)z \tag{36}$$
$$= \int_{x,z} p(x)p(z)x + \int_{x,z} p(x)p(z)z \tag{37}$$
$$= \int_x p(x)x + \int_z p(z)z \tag{38}$$
$$= \mathbb{E}[x] + \mathbb{E}[z] \tag{39}$$

$$
\begin{aligned}
var[x+z] &= \mathbb{E}\left[\left\{(x+z)-\mathbb{E}[x+z]\right)^2\right\}] && (40)\\
&= \mathbb{E}\left[(x+z)^2\right] - \mathbb{E}\left[x+z\right]^2 && (41)\\
&= \mathbb{E}\left[x^2+z^2+2xz\right] - \left(\mathbb{E}[x]+E[z]\right)^2 && (42)\\
&= \mathbb{E}[x^2] + \mathbb{E}[z^2] + \mathbb{E}[2xz] - \mathbb{E}[x]^2 - \mathbb{E}[z]^2 - 2\,\mathbb{E}[z]\,\mathbb{E}[x] && (43)\\
&= var[x] + var[y] + 2\int_{x,z} p(x,z)xz - 2\int_x p(x)\int_z p(z) && (44)\\
&= var[x] + var[y] + 2\int_{x,z} p(x)p(z)xz - 2\int_x p(x)\int_z p(z) && (45)\\
&= var[x] + var[y] + 2\int_x p(x)\int_z p(z) - 2\int_x p(x)\int_z p(z) && (46)
\end{aligned}
$$

1.11

Mean:

$$
\begin{aligned}
\ln p(\boldsymbol{x}|\mu,\sigma^2) &= -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln 2\pi && (47)\\
\frac{\partial \ln p(\boldsymbol{x}|\mu,\sigma^2)}{\partial\mu} &= \sigma^2\sum_{n=1}^{N}(x_n-\mu) = 0 && (48)\\
\sum_{n=1}^{N} x_n &= N\mu && (49)\\
\mu &= \frac{1}{N}\sum_{n=1}^{N} x_n && (50)
\end{aligned}
$$

Variance:

$$
\begin{aligned}
-\frac{1}{2\sigma^2} &= -(2\sigma^2)^{-1} && (51)\\
\frac{\partial \ln p(\boldsymbol{x}|\mu,\sigma^2)}{\partial\sigma^2} &= 2\cdot\frac{1}{4\sigma^4}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2\sigma^2} = 0 && (52)\\
N\cdot 2\sigma^4 &= 2\sigma^2\sum_{n=1}^{N}(x_n-\mu)^2 && (53)\\
\sigma^2 &= \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu)^2 && (54)
\end{aligned}
$$

1.13

From 1.56:

$$\mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}(x_n-\mu)^2\right] = \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[x^2] - \mathbb{E}[2x_n\mu] + \mathbb{E}[\mu^2] \tag{55}$$

$$= \mathbb{E}[x^2] - 2\mu\,\mathbb{E}[x] + \mu^2 \tag{56}$$

$$= \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2 \tag{57}$$

1.22

$$1 - I_{kj} = \begin{bmatrix} 0 & 1 & \dots & 1 \\ 1 & 0 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 0 \end{bmatrix} \tag{58}$$

$$\mathcal{R}_j^* = \arg\min_j L_{kj}p(C_k|\boldsymbol{x}) \tag{59}$$

$$= \arg\min_j \sum_{j\neq k} L_{kj}p(C_k|\boldsymbol{x}) \tag{60}$$

$$= \arg\max_j p(C_k|\boldsymbol{x}) \tag{61}$$

1.23

1.25

$$\mathbb{E}\left[L(\boldsymbol{t},\boldsymbol{y}(\boldsymbol{x}))\right] = \int\int \|\boldsymbol{y}(\boldsymbol{x})-\boldsymbol{t}\|^2\, p(\boldsymbol{x},\boldsymbol{t})d\boldsymbol{x}d\boldsymbol{t} \tag{62}$$

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

$$= \int\int \|\boldsymbol{y}(\boldsymbol{x})-\boldsymbol{t}\|^2\, p(\boldsymbol{x},\boldsymbol{t})d\boldsymbol{t}d\boldsymbol{x} \tag{63}$$

$$F[\boldsymbol{x},\boldsymbol{y},\boldsymbol{t}] = \int_{\boldsymbol{t}} \|\boldsymbol{y}(\boldsymbol{x})-\boldsymbol{t}\|^2\, p(\boldsymbol{x},\boldsymbol{t}) \tag{64}$$

Euler-Lagrange:

$$\frac{\partial F}{\partial \boldsymbol{y}} - \frac{d}{d\boldsymbol{x}}\left(\frac{\partial F}{\partial \boldsymbol{y}'}\right) = 0 \tag{65}$$

$$\frac{\partial F}{\partial \boldsymbol{y}'} = 0 \tag{66}$$

$$\frac{\partial F}{\partial \boldsymbol{y}} = \int_{\boldsymbol{t}} (\boldsymbol{y}(\boldsymbol{x}) - \boldsymbol{t}) p(\boldsymbol{x}, \boldsymbol{t}) = 0 \tag{67}$$

$$\int_{\boldsymbol{t}} \boldsymbol{y}(\boldsymbol{x}) p(\boldsymbol{x}, \boldsymbol{t}) - \int_{t} \boldsymbol{t} p(\boldsymbol{x}, \boldsymbol{t}) = 0 \tag{68}$$

$$\boldsymbol{y}(\boldsymbol{x}) \int_{\boldsymbol{t}} p(\boldsymbol{t}, \boldsymbol{x}) = \int_{t} \boldsymbol{t} p(\boldsymbol{x}, \boldsymbol{t}) \tag{69}$$

$$\boldsymbol{y}(\boldsymbol{x}) = \frac{1}{p(\boldsymbol{x})} \int_{t} \boldsymbol{t} p(\boldsymbol{x}, \boldsymbol{t}) = \mathbb{E}_t[\boldsymbol{t}|\boldsymbol{x}] \tag{70}$$

$$\tag{71}$$

1.26

First, let's derive the single value case.

$$\{y(\boldsymbol{x}) - t\}^2 = \{y(\boldsymbol{x}) - \mathbb{E}[t|\boldsymbol{x}] + \mathbb{E}[t|\boldsymbol{x}] - t\}^2 \tag{72}$$

$$= \{y(\boldsymbol{x}) - E[t|\boldsymbol{x}]\}^2 + \{E[t|\boldsymbol{x}] - t\}^2 + 2\left[y(\boldsymbol{x}) - \mathbb{E}[t|\boldsymbol{x}]\right]\{\mathbb{E}[t|\boldsymbol{x}] - t\} \tag{73}$$

$$\mathbb{E}[L] = \underbrace{\mathbb{E}_{x,t}\left[(y(\boldsymbol{x}) - \mathbb{E}[t|\boldsymbol{x}])^2\right]}_{(1)} + \underbrace{\mathbb{E}_{x,t}\left[(\mathbb{E}[t|\boldsymbol{x}] - t)^2\right]}_{(2)} + \underbrace{\mathbb{E}_{x,t}\left[2\left(y(\boldsymbol{x}) - \mathbb{E}[t|\boldsymbol{x}]\right)\left(\mathbb{E}[t|\boldsymbol{x}] - t\right)\right]}_{(3)} \tag{74}$$

(1)

$$\mathbb{E}_{x,t}\left[(y(\boldsymbol{x}) - \mathbb{E}[t|\boldsymbol{x}])^2\right] = \mathbb{E}_x\, \mathbb{E}_t\left[(y(\boldsymbol{x}) - \mathbb{E}[t|\boldsymbol{x}])^2\, |\boldsymbol{x}\right] \tag{75}$$

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

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

$$\tag{76}$$

(2)

$$\mathbb{E}_{x,t}\left[(\mathbb{E}[t|\boldsymbol{x}] - t)^2\right] = \mathbb{E}_x\left[\mathbb{E}_t\left[\mathbb{E}[t|x]^2 + t^2 - 2\,\mathbb{E}[t|x]t\right]|x\right] \tag{77}$$

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

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

$$= \mathbb{E}_x\left[var\left[t|x\right]\right] \tag{78}$$

(3)

$$\mathbb{E}_{x,t}\left[2\left(y(\boldsymbol{x}) - \mathbb{E}[t|\boldsymbol{x}]\right)\left(\mathbb{E}[t|\boldsymbol{x}] - t\right)\right] = \mathbb{E}_x\, \mathbb{E}_t\left[2(y(\boldsymbol{x}) - \mathbb{E}[t|x])(\mathbb{E}[t|\boldsymbol{x}] - t)|x\right] \tag{79}$$

$$= \mathbb{E}_x\left[2(y(\boldsymbol{x}) - \mathbb{E}[t|x])\right]\mathbb{E}_t\left[(\mathbb{E}[t|\boldsymbol{x}] - t)|x\right] \tag{80}$$

$$= \mathbb{E}_x\left[2(y(\boldsymbol{x}) - \mathbb{E}[t|x])\right] \cdot 0 \tag{81}$$

$$= 0 \tag{82}$$

Now we continue for the vector case.

$$\|\boldsymbol{y}(\boldsymbol{x}) - \boldsymbol{t}\|^2 = \|\boldsymbol{y}(\boldsymbol{x}) - \mathbb{E}[\boldsymbol{t}|\boldsymbol{x}] + \mathbb{E}[\boldsymbol{t}|\boldsymbol{x}] - \boldsymbol{t}\|^2 \tag{83}$$

$$= (\boldsymbol{y}(\boldsymbol{x}) - \mathbb{E}[\boldsymbol{t}|\boldsymbol{x}] + \mathbb{E}[\boldsymbol{t}|\boldsymbol{x}] - \boldsymbol{t})^T(\boldsymbol{y}(\boldsymbol{x}) - \mathbb{E}[\boldsymbol{t}|\boldsymbol{x}] + \mathbb{E}[\boldsymbol{t}|\boldsymbol{x}] - \boldsymbol{t}) \tag{84}$$

$$= (\boldsymbol{y}(\boldsymbol{x}) - \mathbb{E}[\boldsymbol{t}|\boldsymbol{x}]))^2 + (\mathbb{E}[\boldsymbol{t}|\boldsymbol{x}] - \boldsymbol{t})^2 + 2\left[(\boldsymbol{y}(\boldsymbol{x}) - \mathbb{E}[\boldsymbol{t}|\boldsymbol{x}])^T(\mathbb{E}[\boldsymbol{t}|\boldsymbol{x}] - \boldsymbol{t})\right] \tag{85}$$

$$\mathbb{E}\left[L(\boldsymbol{t}, \boldsymbol{y}(\boldsymbol{x})\right] = \mathbb{E}_{x,t}[\underbrace{(\boldsymbol{y}(\boldsymbol{x}) - \mathbb{E}[\boldsymbol{t}|\boldsymbol{x}]))^2}_{(1)} + \underbrace{(\mathbb{E}[\boldsymbol{t}|\boldsymbol{x}] - \boldsymbol{t})^2}_{(2)} + \underbrace{2\left[(\boldsymbol{y}(\boldsymbol{x}) - \mathbb{E}[\boldsymbol{t}|\boldsymbol{x}])^T(\mathbb{E}[\boldsymbol{t}|\boldsymbol{x}] - \boldsymbol{t})\right]}_{(3} \tag{86}$$

(1)

$$\mathbb{E}_{x,t}\left[(\boldsymbol{y}(\boldsymbol{x})) - \mathbb{E}[\boldsymbol{t}|\boldsymbol{x}])^2\right] = \mathbb{E}_x\,\mathbb{E}_t\left[(\boldsymbol{y}(\boldsymbol{x})) - \mathbb{E}[\boldsymbol{t}|\boldsymbol{x}])^2|\boldsymbol{x}\right] \tag{87}$$

$$= \mathbb{E}_x(\boldsymbol{y}(\boldsymbol{x})) - \mathbb{E}[\boldsymbol{t}|\boldsymbol{x}])^2 \tag{88}$$

(2)

$$\mathbb{E}_{x,t}\left[(\mathbb{E}[\boldsymbol{t}|\boldsymbol{x}] - \boldsymbol{t})^2\right] = \mathbb{E}_x\,\mathbb{E}_t\left[(\mathbb{E}[\boldsymbol{t}|\boldsymbol{x}] - \boldsymbol{t})^2|\boldsymbol{x}\right] \tag{89}$$

$$= \mathbb{E}_x\,\mathbb{E}_t\left[\mathbb{E}[\boldsymbol{t}|\boldsymbol{x}]^T\,\mathbb{E}[\boldsymbol{t}|\boldsymbol{x}] + \boldsymbol{t}^T\boldsymbol{t} - 2\,\mathbb{E}[\boldsymbol{t}|\boldsymbol{x}]^T\boldsymbol{t}|\boldsymbol{x}\right] \tag{90}$$

$$= \mathbb{E}_x\left[\mathbb{E}[\boldsymbol{t}|\boldsymbol{x}]^T\,\mathbb{E}[\boldsymbol{t}|\boldsymbol{x}] + \mathbb{E}_t\left[\boldsymbol{t}^T\boldsymbol{t}\right] - 2\,\mathbb{E}[\boldsymbol{t}|\boldsymbol{x}]^T\,\mathbb{E}[\boldsymbol{t}|\boldsymbol{x}]\right] \tag{91}$$

$$= \mathbb{E}_x\left[var[\boldsymbol{t}|\boldsymbol{x}]\right] \tag{92}$$

(3)

$$\mathbb{E}_{x,t}\left\{2\left[\boldsymbol{y}(\boldsymbol{x}) - \mathbb{E}[\boldsymbol{t}|\boldsymbol{x}]\right)^T(\mathbb{E}[\boldsymbol{t}|\boldsymbol{x}] - \boldsymbol{t})\right]\right\} = \mathbb{E}_x\,\mathbb{E}_t\left\{2(\left[\boldsymbol{y}\boldsymbol{x}) - \mathbb{E}[\boldsymbol{t}|\boldsymbol{x}]\right)^T(\mathbb{E}[\boldsymbol{t}|\boldsymbol{x}] - \boldsymbol{t})\right]|\boldsymbol{x}\right\} \tag{93}$$

$$= \mathbb{E}_x\,2(\boldsymbol{y}(\boldsymbol{x}) - \mathbb{E}[\boldsymbol{t}|\boldsymbol{x}])^T\,\mathbb{E}_t\left[\mathbb{E}[\boldsymbol{t}|\boldsymbol{x}] - \boldsymbol{t}\right]] \tag{94}$$

$$= \mathbf{0} \tag{95}$$

1.28

$n = k = 2$

$$h(p^n) = -\log(p(x)^2) = -2\log p(x) = 2h(p) \tag{96}$$

Induction:

$$h(p^{k+1}) = -\log p(x)^{k+1} = -(k+1)\log(p(x) = -(k+1)\log p(x) = (k+1)h(p) \tag{97}$$

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

$$H[x] = -\sum_{n=1}^{M} p(x) \ln p(x) \tag{98}$$

$$= \sum_{n=1}^{M} p(x) \ln \frac{1}{p(x)} \tag{99}$$

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

$$= \ln M \tag{100}$$

1.32

$$H[\boldsymbol{y}] = -\int p(\boldsymbol{y}) \ln p(\boldsymbol{y}) d\boldsymbol{y} \tag{101}$$

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

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

$$= -\int p(\boldsymbol{x}) \ln \frac{p(\boldsymbol{x})}{|\boldsymbol{A}|} d\boldsymbol{x} \tag{102}$$

$$= -\int p(\boldsymbol{x}) \ln p(\boldsymbol{x}) d\boldsymbol{x} + \int p(\boldsymbol{x}) \ln |\boldsymbol{A}| d\boldsymbol{x} \tag{103}$$

$$= H[\boldsymbol{x}] + \ln |\boldsymbol{A}| \int p(\boldsymbol{x}) d\boldsymbol{x} \tag{104}$$

$$= H[\boldsymbol{x}] + \ln |\boldsymbol{A}| \tag{105}$$

1.34

$$-H[x] = \int p(x) \ln p(x) dx \tag{106}$$

$$= \int p(x) \ln \frac{1}{\sqrt{2\pi\sigma^2}} dx - \int p(x) \frac{(x-\mu)^2}{2\sigma^2} \tag{107}$$

$$= \int p(x) \ln 1 dx - \int p(x) \ln \sqrt{2\pi\sigma^2} dx - \int p(x) \frac{(x-\mu)^2}{2\sigma^2} \tag{108}$$

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

$$= -\frac{1}{2} \ln 2\pi\sigma^2 - \frac{\sigma^2}{2\sigma^2} \tag{109}$$

$$= -\frac{1}{2} \ln 2\pi\sigma^2 - \frac{1}{2} \tag{110}$$

$$H[x] = \frac{1}{2} \left[\ln 2\pi\sigma^2 + 1 \right] \tag{111}$$

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

$$\begin{aligned}
-H[x, y] &= \int\int p(y, x) \ln p(x, y) dx dy && (112)\\
&= \int\int p(y, x) \ln[p(y|x)p(x)] dx dy && (113)\\
&= \int\int p(y, x) \ln p(y|x) + p(y, x) \ln p(x) dx dy && (114)\\
&= -H[y|x] + \int\int p(y, x) dy \ln p(x) dx && (115)\\
&= -H[y|x] - H[x] && (116)
\end{aligned}$$

1.40

$$\begin{aligned}
\ln \prod_{i=1}^{n} {a_i}^{1/n} &= \frac{1}{n} \ln \prod_{i=1}^{n} a_i && (117)\\
&= \frac{1}{n} \sum_{i=1}^{n} \ln a_i && (118)\\
&\geq \ln \frac{1}{n} \sum_{i=1}^{n} a_i && \text{(Jensen)}\\
& && (119)
\end{aligned}$$

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

$$\begin{aligned}
-I[x, y] &= \int\int \Big[p(x, y) \ln p(x) + p(x, y) \ln p(y) - p(x, y) \ln p(x|y) - p(x, y) \ln p(y) \Big] dx dy && (120)\\
&= -H[x] + H[x|y] && (121)
\end{aligned}$$

2.1

$$p(x|\mu) = \mu^x (1 - \mu)^{1-x} \tag{122}$$

$$\sum_{x=0}^{1} p(x|\mu) = 1 - \mu + \mu = 1 \tag{123}$$

$$\mathbb{E}[x] = \sum_{x=0}^{1} x\mu^x(1-\mu)^{1-x} \tag{124}$$
$$= \mu(1-\mu)^0 \qquad = \mu \tag{125}$$

$$\mathrm{Var}[x] = \big[(x - \mathbb{E}[x])^2\big[\tag{126}$$
$$= \big[(x-\mu)^2\big[\tag{127}$$
$$= \sum_{i=0}^{1}(x-\mu)^2\mu^x(1-\mu)^{1-x} \tag{128}$$
$$\tag{129}$$

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

$$H[x] = \sum i = 0^1 p(x)\ln p(x) = \mu\ln\mu + (1-\mu)\ln(1-\mu) \tag{130}$$

2.6

$$p(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \tag{131}$$

$$\int_0^1 \mu^{a-1}(1-\mu)^{b-1}d\mu = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \tag{132}$$

$$\Gamma(x) = \int_0^\infty \mu^{x-1}e^{-u}du \tag{133}$$

$$\mathbb{E}[\mu] = \int_0^1 p(\mu|a,b)\mu d\mu \tag{134}$$
$$= \int_0^1 \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}\mu d\mu \tag{135}$$
$$= \int_0^1 \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a}(1-\mu)^{b-1}d\mu \tag{136}$$

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

$$. = \frac{a}{a+b}\int_0^1 \frac{\Gamma(a+1+b)}{\Gamma(a+1)\Gamma(b)}\mu^{a}(1-\mu)^{b-1}d\mu \tag{137}$$
$$= \frac{a}{a+b} \tag{138}$$

Where we used

$$\Gamma(a+1) = a\Gamma(a) \text{ and } \Gamma(a+b+1) = (a+b)\Gamma(a+b) \tag{139}$$

$$\mathrm{Var}[\mu] = \mathbb{E}[\mu^2] - \mathbb{E}[\mu]^2 \tag{140}$$

$$\mathbb{E}[\mu] = \int_0^1 p(\mu|a,b)\mu d\mu^2 \tag{141}$$

$$= \int_0^1 \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a+1}(1-\mu)^{b-1}\mu d\mu \tag{142}$$

Which looks like Beta with a+2

$$= \frac{a}{a+b}\int_0^1 \frac{\Gamma(a+1+b)}{\Gamma(a+1)\Gamma(b)}\mu^{a+2}(1-\mu)^{b-1}d\mu \tag{143}$$

$$= \frac{a(a+1)}{(a+b)(a+1+b)}\int_0^1 \frac{\Gamma(a+2+b)}{\Gamma(a+2)\Gamma(b)}\mu^{a+2}(1-\mu)^{b-1}d\mu \tag{144}$$

$$= \frac{a(a+1)}{(a+b)(a+1+b)} \tag{145}$$

$$\mathrm{Var}[\mu] = \frac{a(a+1)}{(a+b)(a+1+b)} - \frac{a^2}{(a+b)^2} \tag{146}$$

$$= \frac{(a+b)a(a+1)}{(a+b)^2(a+1+b)} - \frac{(a+b+1)a^2}{(a+b+1)(a+b)^2} \tag{147}$$

$$= \frac{(a^2+ab)(a+1)}{(a+b)^2(a+1+b)} - \frac{(a+b+1)a^2}{(a+b+1)(a+b)^2} \tag{148}$$

$$= \frac{a^3+a^2b+a^2+ab}{(a+b)^2(a+1+b)} - \frac{a^3+ba^2+a^2}{(a+b+1)(a+b)^2} \tag{149}$$

Which gives you the desired result.

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

$$\frac{\partial p(\mu|a,b)}{\partial \mu} = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\left[(a-1)\mu^{a-2} - \mu^{a-1}(b-1)(1-\mu)^{b-2}\right] = 0 \tag{150}$$

$$(a-1)\mu^{a-2}(1-\mu)^{b-1} = \mu^{a-1}(b-1)(1-\mu)^{b-2} \tag{151}$$

$$\frac{a-1}{\mu} = \frac{b-1}{1-\mu} \tag{152}$$

$$\mu = \frac{-a+1}{-a-b+2} = \frac{a-1}{a+b-2} \tag{153}$$

2.8

$$\mathbb{E}[x] = \int p(x)xdx \tag{154}$$
$$= \int\int p(x,y)xdxdy \tag{155}$$
$$= \int\int p(x|y)p(y)xdxdy \tag{156}$$
$$= \int p(y)\int p(x|y)xdxdy \tag{157}$$
$$= \int p(y)\,\mathbb{E}_x[x|y]dy \tag{158}$$
$$= \mathbb{E}_y\,\mathbb{E}_x[x|y] \tag{159}$$

$$\mathbb{E}_y[\mathrm{Var}_x[x|y]] = \mathbb{E}_y\,\mathbb{E}_{x|y}[(x-\mathbb{E}[x|y])^2] \tag{160}$$
$$= \mathbb{E}_y\,\mathbb{E}_{x|y}[(x^2+\mathbb{E}[x|y]^2-2x\,\mathbb{E}[x|y])] \tag{161}$$
$$= \mathbb{E}_y\,\mathbb{E}_{x|y}\,x^2+\mathbb{E}_y\,\mathbb{E}_{x|y}\,\mathbb{E}[x|y]^2-2\,\mathbb{E}_y\,\mathbb{E}_{x|y}\,x\,\mathbb{E}[x|y])] \tag{162}$$
$$= \mathbb{E}_y\int p(x|y)x^2dx+\mathbb{E}_y\,\mathbb{E}[x|y]^2-2\int p(y)\int p(x|y)x\,\mathbb{E}[x|y]dxdy \tag{163}$$
$$= \int p(y)\int p(x|y)x^2dxdy+\mathbb{E}_y\,\mathbb{E}[x|y]^2-2\int \mathbb{E}[x|y]\int p(x|y)xdxdy \tag{164}$$
$$= \int\int p(x,y)x^2dxdy+\mathbb{E}_y\,\mathbb{E}[x|y]^2-2\int \mathbb{E}[x|y]^2dy)] \tag{165}$$
$$= \int\int p(x,y)x^2dxdy+\mathbb{E}_y\,\mathbb{E}[x|y]^2-2\,\mathbb{E}_y\,\mathbb{E}[x|y]^2)] \tag{166}$$
$$= \mathbb{E}[x^2]-\mathbb{E}_y\,\mathbb{E}[x|y]^2 \tag{167}$$
$$\tag{168}$$

$$\mathrm{Var}_y[\mathbb{E}_x[x|y]] = \mathbb{E}_y\left[(\mathbb{E}_x[x|y]-\mathbb{E}_y\,\mathbb{E}_x[x|y])^2\right] \tag{169}$$
$$= \mathbb{E}_y\left[(\mathbb{E}_x[x|y]-\mathbb{E}_x[x])^2\right] \tag{170}$$
$$= \mathbb{E}_y\left[(\mathbb{E}_x[x|y]^2+\mathbb{E}_x[x]^2-2\,\mathbb{E}_x[x]\,\mathbb{E}_x[x|y]\right] \tag{171}$$
$$= \mathbb{E}_y[\mathbb{E}_x[x|y]^2]+\mathbb{E}_y[\mathbb{E}_x[x]^2]-2\,\mathbb{E}_y\,\mathbb{E}_x[x]\,\mathbb{E}_x[x|y] \tag{172}$$
$$= \mathbb{E}_y[\mathbb{E}_x[x|y]^2]+\mathbb{E}_x[x]^2-2\,\mathbb{E}_x[x]\,\mathbb{E}_y\,\mathbb{E}_x[x|y] \tag{173}$$
$$= \mathbb{E}_y[\mathbb{E}_x[x|y]^2]+\mathbb{E}_x[x]^2-2\,\mathbb{E}_x[x]^2 \tag{174}$$
$$= \mathbb{E}_y[\mathbb{E}_x[x|y]^2]-\mathbb{E}_x[x]^2 \tag{175}$$
$$\tag{176}$$

Sum of these expressions gives desired result.

2.11

2.12

$$\int_a^b \frac{1}{b-a}dx = \frac{x}{b-a} - \frac{x}{b-a}\Big|_a^b = \frac{b-a}{b-a} = 1 \tag{177}$$

$$\begin{aligned}
\mathbb{E}[x] &= \int_a^b p(x)x dx && (178)\\
&= \int_a^b \frac{x}{b-a}dx && (179)\\
&= \frac{x^2}{2(b-a)}\Big|_a^b && (180)\\
&= \frac{b^2-a^2}{2(b-a)} = \frac{(b-a)(b+a)}{2(b-a)} = \frac{b+a}{2} && (181)
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}[x] &= \mathbb{E}[x^2] - \mathbb{E}[x]^2 && (182)\\
&= \int_a^b \frac{x^2}{b-a}dx - \frac{(b+a)^2}{4} && (183)\\
&= \frac{\frac{1}{3}x^3}{b-a}\Big|_a^b - \frac{(b+a)^2}{4} && (184)\\
&= \frac{\frac{1}{3}b^3}{b-a} - \frac{\frac{1}{3}a^3}{b-a} - \frac{(b+a)^2}{4} && (185)\\
&= \frac{(b-a)(b^2+a^2+ab)}{3(b-a)} - \frac{(b+a)^2}{4} && (186)\\
&= \frac{b^2+a^2+ab}{3} - \frac{b^2+a^2+2ab}{4} && (187)\\
&= \frac{4b^2+4a^2+4ab}{12} - \frac{3b^2+3a^2+6ab}{12} && (188)\\
&= \frac{b^2+a^2-2ab}{12} && (189)\\
&= \frac{(b-a)^2}{12} && (190)\\
& && (191)
\end{aligned}$$

2.17

$$A = \frac{1}{2}\underbrace{A + A^T}_{Symmetric} + \frac{1}{2}\underbrace{A - A^T}_{Non-symmetric} \tag{192}$$

$$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto \exp -\frac{1}{2}(x-\mu)^T \frac{1}{2}(\Sigma^{-1} + \Sigma^{-T} + \Sigma^{-1} - \Sigma^{-T})(x-\mu) \tag{193}$$

$$= \exp -\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{M}(x_n-\mu_n)(\Sigma_{nm}^{-1} + \Sigma_{nm}^{-T})(x_m-\mu_m) \tag{194}$$

$$-\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{M}(x_n-\mu_n)\underbrace{(\Sigma_{nm}^{-1} - \Sigma_{nm}^{-T})}_{=0}(x_m-\mu_m) \tag{195}$$

Which leaves us only with a symmetric covariance matrix.

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.

$$\#\text{parameters} = D + (D-1) + (D-2) + \cdots + D - (D-2) + D - (D-1) \tag{196}$$

$$= (D+1) + (D+1) + \cdots + (D+1) \tag{197}$$

$$= \frac{D}{2}(D+1) \tag{198}$$

2.22

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

$$A^{-1}AA^{T\,-1} = A^{-1}A^TA^{T\,-1} \tag{199}$$

$$A^{-1\,T} = A^{-1}A^TA^{T\,-1} \tag{200}$$

$$= A^{-1} \tag{201}$$

2.27

$$\mathbb{E}[\boldsymbol{y}] = \mathbb{E}[\boldsymbol{x} + \boldsymbol{z}] \tag{202}$$

$$= \int\int p(\boldsymbol{x}, \boldsymbol{z})(\boldsymbol{x} + \boldsymbol{z}) d\boldsymbol{x} d\boldsymbol{z} \tag{203}$$

$$= \int\int p(\boldsymbol{x}, \boldsymbol{z})\boldsymbol{x} + p(\boldsymbol{x}, \boldsymbol{z})\boldsymbol{z} d\boldsymbol{x} d\boldsymbol{z} \tag{204}$$

$$= \int\int p(\boldsymbol{x})p(\boldsymbol{z})\boldsymbol{x} + p(\boldsymbol{z})p(\boldsymbol{x})\boldsymbol{z} d\boldsymbol{x} d\boldsymbol{z} \tag{205}$$

$$= \int p(\boldsymbol{z}) \int p(\boldsymbol{x})\boldsymbol{x} d\boldsymbol{x} d\boldsymbol{z} + \int p(\boldsymbol{x}) \int p(\boldsymbol{z})\boldsymbol{z} d\boldsymbol{x} d\boldsymbol{z} \tag{206}$$

$$= \mathbb{E}[\boldsymbol{x}] + \mathbb{E}[\boldsymbol{z}] \tag{207}$$

2.30

$$\mathbb{E}[z] = R^{-1}\begin{pmatrix}\Lambda\mu - A^T Lb \\ Lb\end{pmatrix} \tag{208}$$

$$= \begin{pmatrix}\Lambda^{-1} & \Lambda^{-1}A^T \\ A\Lambda^{-1} & L^{-1} + A\Lambda^{-1}A^T\end{pmatrix}\begin{pmatrix}\Lambda\mu - A^T Lb \\ Lb\end{pmatrix} \tag{209}$$

$$= \begin{pmatrix}\Lambda^{-1}(\Lambda\mu - A^T Lb) + \Lambda^{-1}A^T Lb \\ A\Lambda^{-1}(\Lambda\mu - A^T Lb) + \mathcal{L}^{-1}Lb + A\Lambda^{-1}A^T Lb\end{pmatrix} \tag{210}$$

$$= \begin{pmatrix}\mu - \Lambda^{-1}A^T Lb + \Lambda^{-1}A^T Lb \\ A\mu - A\Lambda^{-1}A^T Lb + b + A\Lambda^{-1}A^T Lb\end{pmatrix} \tag{211}$$

Which gives the desired result.

2.41

$$\Gamma(x) = \int_0^\infty u^{x-1}\exp(-u)du \tag{212}$$

$$\int_0^\infty \mathrm{Gam}(\lambda|a,b)d\lambda = \int_0^\infty \frac{1}{\Gamma(a)}b^a\lambda^{a-1}\exp(-b\lambda)d\lambda \tag{213}$$

$$= \frac{1}{\Gamma(a)}b^a\int_0^\infty \lambda^{a-1}\exp(-b\lambda)d\lambda \tag{214}$$

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

$$= \frac{\Gamma(a)}{\Gamma(a)} = 1 \tag{215}$$

2.46

$$a = \nu/2 \quad b = \frac{\nu}{2\lambda} \tag{216}$$

$$p(x|\mu,a,b) = \frac{b^a}{\Gamma(a)}\left(\frac{1}{2a}\right)^{1/2}\left[b + \frac{1}{2}(x-\mu)^2\right]^{-a-\frac{1}{2}}\Gamma(a+\frac{1}{2}) \tag{217}$$

$$= \frac{\Gamma(\nu/2+1/2)}{\Gamma(\nu/2)}\left(\frac{1}{2\pi}\right)^{1/2}\left[\frac{\nu}{2\lambda} + \frac{1}{2}(x-\mu)^2\right]^{-\frac{\nu}{2}-\frac{1}{2}}\left(\frac{\nu}{2\lambda}\right)^{\nu/2} \tag{218}$$

$$= \frac{\Gamma(\nu/2+1/2)}{\Gamma(\nu/2)}\left[\frac{2\lambda}{\nu}\left(\frac{\nu}{2\lambda} + \frac{1}{2}(x-\mu)^2\right)\right]^{-\frac{\nu}{2}}\left[2\pi\left(\frac{\nu}{2\lambda} + \frac{1}{2}(x-\mu)^2\right)\right]^{-\frac{1}{2}} \tag{219}$$

$$= \frac{\Gamma(\nu/2+1/2)}{\Gamma(\nu/2)}\left[1 + \frac{\lambda(x-\mu)^2}{\nu}\right]^{-\frac{\nu}{2}}\left[\frac{\pi\nu}{\lambda} + \pi(x-\mu)^2\right]^{-\frac{1}{2}} \tag{220}$$

$$= \frac{\Gamma(\nu/2+1/2)}{\Gamma(\nu/2)}\left[1 + \frac{\lambda(x-\mu)^2}{\nu}\right]^{-\frac{\nu}{2}}\left[1 + \frac{\lambda}{\nu}(x-\mu)^2\right]^{-\frac{1}{2}}\left(\frac{\lambda}{\pi\nu}\right)^{1/2} \tag{221}$$

$$\tag{222}$$

2.47

$$\mathrm{St}(x|\mu,\lambda,\nu) \propto \left[1+\frac{\lambda(x-\mu)^2}{\nu}\right]^{-v/2-1/2} \tag{223}$$

$$= \exp\ln\left[1+\frac{\lambda(x-\mu)^2}{\nu}\right]^{-v/2-1/2} \tag{224}$$

$$= \exp\left[\frac{-v-1}{2}\ln\left(1+\frac{\lambda(x-\mu)^2}{\nu}\right)\right] \tag{225}$$

$$= \exp\left[\frac{-v-1}{2}\left(\frac{\lambda(x-\mu)^2}{\nu}-\mathcal{O}\left[\frac{\lambda(x-\mu)^2}{2\nu}\right]^2\right)\right] \tag{226}$$

$$= \exp\left[\frac{-v-1}{2}\left(\frac{\lambda(x-\mu)^2}{\nu}-\mathcal{O}\left[\frac{\lambda^2(x-\mu)^4}{4\nu^2}\right]\right)\right] \tag{227}$$

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

2.48

$$\Gamma(x) = \int \mu^{x-1}e^{-u}du \tag{228}$$

$$|(\eta\Lambda)^{-1}|^{1/2} = \left[\eta^{-D}|\Lambda^{-1}|\right]^{1/2} = \eta^{-D/2}|\Lambda|^{-1/2} \tag{229}$$

$$\Gamma(D/2+\nu/2) = \int \eta^{D/2+\nu/2-1}\exp(-\eta)d\eta \tag{230}$$

$$z = \eta/2(\Delta^2+\nu) \iff \eta = 2\frac{z}{\Delta^2+\nu} \tag{231}$$

$$d\eta = \frac{d\eta}{dz}dz = \frac{z}{\Delta^2+\nu}dz \tag{232}$$

$$\int \mathcal{N}(x|\mu,(\eta\Lambda)^{-1})\mathrm{Gam}\left(\eta|\frac{\nu}{2},\frac{\nu}{2}\right)d\eta \tag{233}$$

$$= \int \frac{1}{(2\pi)^{D/2}}\frac{1}{|(\eta\Lambda)^{-1}|^{1/2}}\exp\{-\frac{1}{2}(\boldsymbol{x}-\mu)^T(\eta\Lambda)(\boldsymbol{x}-\mu)\}\frac{1}{\Gamma(\nu/2)}(\nu/2)^{\nu/2}\eta^{\frac{\nu}{2}-1}\exp(-\nu/2\eta)d\eta \tag{234}$$

$$= \int \frac{1}{(2\pi)^{D/2}}\frac{1}{|(\eta\Lambda)^{-1}|^{1/2}}\exp\{-\frac{\eta}{2}\Delta^2\}\frac{1}{\Gamma(\nu/2)}(\nu/2)^{\nu/2}\eta^{\frac{\nu}{2}-1}\exp(-\nu/2\eta)d\eta \tag{235}$$

$$= \int \frac{1}{(2\pi)^{D/2}}\eta^{D/2}|\Lambda|^{1/2}\exp\{-\frac{\eta}{2}\Delta^2\}\frac{1}{\Gamma(\nu/2)}(\nu/2)^{\nu/2}\eta^{\frac{\nu}{2}-1}\exp(-\nu/2\eta)d\eta \tag{236}$$

$$= \frac{1}{(2\pi)^{D/2}}|\Lambda|^{1/2}\frac{1}{\Gamma(\nu/2)}(\nu/2)^{\nu/2}\int \eta^{D/2}\exp\{-\frac{\eta}{2}\Delta^2\}\eta^{\frac{\nu}{2}-1}\exp(-\nu/2\eta)d\eta \tag{237}$$

$$= \frac{1}{(2\pi)^{D/2}}|\Lambda|^{1/2}\frac{1}{\Gamma(\nu/2)}(\nu/2)^{\nu/2}\int \eta^{D/2+v/2-1}\exp\{-\frac{\eta}{2}(\Delta^2-\nu)\}d\eta \tag{238}$$

$$= \frac{1}{(2\pi)^{D/2}}|\Lambda|^{1/2}\frac{1}{\Gamma(\nu/2)}(\nu/2)^{\nu/2}2(\Delta^2+\nu)^{-D/2-\nu/2}\int z^{1/2D+\nu/2}\exp\{-z\}d\eta \tag{239}$$

$$= \frac{1}{(2\pi)^{D/2}}|\Lambda|^{1/2}\frac{1}{\Gamma(\nu/2)}(\nu/2)^{\nu/2}2(\Delta^2+\nu)^{-D/2-\nu/2}\Gamma(D/2+\nu/2) \tag{240}$$

$$\tag{241}$$

Which works out algebraically to the desired result.

2.50

$$\mathrm{St}(\boldsymbol{x}|\mu,\Lambda,\nu) \propto \left[1+\frac{1}{\nu}\Delta^2\right]^{-D/2-v/2} \tag{242}$$

$$= \exp\frac{-D-\nu}{2}\ln\left[1+\frac{1}{\nu}\Delta^2\right] \tag{243}$$

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

$$= \exp\frac{-D-\nu}{2}\frac{1}{\nu}\Delta^2-\frac{-D-\nu}{2}\boldsymbol{\mathcal{O}}(1/\nu^2) \tag{244}$$

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

2.51

$$1 = [\cos(A)+i\sin(A)][\cos(A)-i\sin(A)] \tag{245}$$

$$= \cos^2(A)-\sin^2(A) \tag{246}$$

$$\cos(A-B) = \Re \exp[i(A-B)] \tag{247}$$
$$= \Re \exp(iA)\exp(-iB) \tag{248}$$
$$= \Re[\cos(A) + i\sin(A)][\cos(B) - i\sin(B)] \tag{249}$$
$$= \cos(A)\cos(B) - \sin(A)\sin(B) \tag{250}$$

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

2.54

$$0 = \sum_{n=1}^{N} \cos(\theta_0)\sin(\theta_n) - \cos(\theta_n)\sin(\theta_0) \tag{251}$$
$$= \cos(\theta_0)\sum_{n=1}^{N}\sin(\theta_n) - \sin(\theta_0)\sum_{n=1}^{N}\cos(\theta_n) \tag{252}$$
$$\tag{253}$$
$$\frac{\sin(\theta_0)}{\cos(\theta_0} = \frac{\sum_{n=1}^{N}\sin(\theta_n)}{\sum_{n=1}^{N}\cos(\theta_n)} \tag{254}$$
$$\tan(\theta_0) = \frac{\sum_{n=1}^{N}\sin(\theta_n)}{\sum_{n=1}^{N}\cos(\theta_n)} \tag{255}$$
$$\theta_0 = \arctan\left\{\frac{\sum_{n=1}^{N}\sin(\theta_n)}{\sum_{n=1}^{N}\cos(\theta_n)}\right\} \tag{256}$$

2.54

$$\frac{\partial p}{\partial \theta} = -(2\pi I_0(m))^{-1}\exp\{m\cos(\theta-\theta_0\}(m\sin(\theta-\theta_0)) \tag{257}$$

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.

$$\frac{\partial\partial p}{\partial\theta} = \frac{1}{2\pi I_0(m)}\left[-\exp(m\cos(\theta-\theta_0)(m\sin(\theta-\theta_0)^2 - m\cos(\theta-\theta_0)\exp(m\cos(\theta-\theta_0))\right] \tag{258}$$

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

$$A(m_{ML}) = \left(\frac{1}{N}\sum_{n=1}^{N}\cos\theta_n\right)\cos\theta_0^{ML} - \left(\frac{1}{N}\sum_{n=1}^{N}\sin\theta_n\right)\sin\theta_0^{ML} \tag{259}$$

$$= \bar{r}\left(\cos\bar{\theta}\cos\theta_0^{ML} - \sin\bar{\theta}\sin\theta_0^{ML}\right) \tag{260}$$

$$= \bar{r}\left(\cos(\bar{\theta} - \theta_0^{ML})\right) \tag{261}$$

$$= \bar{r} \tag{262}$$

2.57

For this question, the following knowledge is necessary:

$$a^T B a = \underbrace{B : aa^T}_{\text{Frobenius product (Hadamard \& sum)}} = \text{vec}(B)^T \text{vec}(aa^T) \tag{263}$$

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

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-D/2}|\boldsymbol{\Sigma}|^{1/2}\exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right] \tag{264}$$

$$= (2\pi)^{-D/2}|\boldsymbol{\Sigma}|^{1/2}\exp\left[-\frac{1}{2}\left(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)\right] \tag{265}$$

$$= (2\pi)^{-D/2}|\boldsymbol{\Sigma}|^{1/2}\exp(-1/2\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\exp\left[\begin{pmatrix}-1/2\text{vec}(\boldsymbol{\Sigma}^{-1})\\ -\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\end{pmatrix}\begin{pmatrix}\text{vec}(\boldsymbol{x}\boldsymbol{x}^T) & \boldsymbol{x}\end{pmatrix}\right] \tag{266}$$

$$= (2\pi)^{-D/2}|\boldsymbol{\Sigma}|^{1/2}\exp(-1/2\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\exp\left[\eta^T u(\boldsymbol{x})\right] \tag{267}$$

$$\tag{268}$$

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

$$g(\eta) = |\text{vec}^{-1}(-2\eta_1)|^{-1/2}\eta_2\text{vec}^{-1}(-2\eta_1)\eta_2^T \tag{269}$$

and

$$h(x) = (2\pi)^{-D/2} \tag{270}$$

2.59

$$\int_0^\infty \frac{1}{x}f(\frac{1}{x})dx = \int_0^\infty \frac{1}{x}f(y)xdy = \int_0^\infty f(y)dy = 1 \tag{271}$$

2.61

3.1

$$2\sigma(2a) - 1 = \frac{2}{1 + e^{-2a}} - 1 \tag{272}$$
$$= \frac{e^a}{e^a} \frac{2}{1 + e^{-2a}} - 1 \tag{273}$$
$$= \frac{2e^a}{e^a + e^{-a}} - 1 \tag{274}$$
$$= \frac{2e^a}{e^a + e^{-a}} - \frac{e^a + e^{-a}}{e^a + e^{-a}} \tag{275}$$
$$= \tanh(a) \tag{276}$$

$$a = \frac{x - \mu_j}{s} \tag{277}$$

$$y(x, \boldsymbol{w}) = w_0 + \sum_{j=1}^{M} w_j \sigma(a) \tag{278}$$
$$= w_0 + \sum_{j=1}^{M} \frac{w_j}{2} \tanh(\frac{1}{2}a) + w_j \tag{279}$$
$$= w_0 + \sum_{j=1}^{M} + \sum_{j=1}^{M} \frac{w_j}{2} \tanh(a') \tag{280}$$
$$= u_0 + \sum_{j=1}^{M} u_j \tanh(a') \tag{281}$$

3.3

$$\frac{dE_D(\boldsymbol{w})}{d\boldsymbol{w}} = \sum_{n=1}^{N} r_n [t_n - \boldsymbol{w}^T \phi(\boldsymbol{x}_n)] \phi(\boldsymbol{x}_n) = 0 \tag{282}$$

$$\sum_{n=1}^{N} r_n t_n \phi(\boldsymbol{x}_n) = \sum_{n=1}^{N} \boldsymbol{w}^T \phi(\boldsymbol{x}_n) \phi(x_n) \tag{283}$$
$$= \sum_{n=1}^{N} \phi(\boldsymbol{x}_n)^T \boldsymbol{w} \phi(\boldsymbol{x}_n) \tag{284}$$
$$= \sum_{n=1}^{N} \phi(\boldsymbol{x}_n) \phi(\boldsymbol{x}_n)^T \boldsymbol{w} \tag{285}$$
$$= \boldsymbol{\Phi}^T \boldsymbol{\Phi} \boldsymbol{w} \tag{286}$$

$$\left([\boldsymbol{r}\cdot\boldsymbol{t}]^T\boldsymbol{\Phi}\right)^T = \boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{w} \tag{287}$$

$$\boldsymbol{\Phi}^T[\boldsymbol{r}\cdot\boldsymbol{t}] = \boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{w} \tag{288}$$

$$(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T[\boldsymbol{r}\cdot\boldsymbol{t}] = \boldsymbol{w} \tag{289}$$

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

3.4

$$\mathbb{E}[E_d(\boldsymbol{w})] = \mathbb{E}[\frac{1}{2}\sum_{n=1}^{N}(y(\boldsymbol{x}_n,\boldsymbol{w}) - t_n)^2] \tag{290}$$

$$= \mathbb{E}[\frac{1}{2}\sum_{n=1}^{N}(w_0 + \sum_{i=1}^{D} w_i(x_i+\epsilon_i) - t_n)^2] \tag{291}$$

$$= \mathbb{E}[\frac{1}{2}\sum_{n=1}^{N}(y(\boldsymbol{x}_n,\boldsymbol{w}) + \sum_{i=1}^{D} w_i\epsilon_i) - t_n)^2] \tag{292}$$

$$= \mathbb{E}[\frac{1}{2}\sum_{n=1}^{N}(y(\boldsymbol{x}_n,\boldsymbol{w}) - t_n])^2 + (\sum_{i=1}^{D} w_i\epsilon_i)^2 + \underbrace{2\,(y(\boldsymbol{x}_n,\boldsymbol{w}) - t_n)\sum_{i=1}^{D} w_i\epsilon_i]}_{=0(\mathbb{E}[\epsilon_i]=0))} \tag{293}$$

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

$$\mathbb{E}(\sum_{i=1}^{D} w_i\epsilon_i)^2 = \mathbb{E}[\sum_{i=1}^{D}\sum_{j=1}^{D} w_j\epsilon_j w_i\epsilon_i] = \sum_{i=1}^{D} w_i^2\,\mathbb{E}\,\epsilon_i^2 = \sum_{i=1}^{D} w_i^2\,\mathbb{E}\,\epsilon_i^2 = \sum_{i=1}^{D} w_i^2(\sigma^2 + \underbrace{E[\epsilon_i]^2}_{=0}) \tag{294}$$

Which gives us our desired result.

3.5

$$\mathcal{L}(\boldsymbol{w},\lambda) = \frac{1}{2}\sum_{n=1}^{N}\left[t_n - \boldsymbol{w}^T\phi(\boldsymbol{x}_n)\right]^2 + \lambda\left[\sum_{j=1}^{M}|w_j|^q - \eta\right] \tag{295}$$

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

3.6

$$p(\boldsymbol{T}|\boldsymbol{W},\boldsymbol{\Sigma}) = \prod_{n=1}^{N} p(\boldsymbol{t}_n|\boldsymbol{W},\boldsymbol{\Sigma}) \tag{296}$$

$$\ln p(\boldsymbol{T}|\boldsymbol{W},\boldsymbol{\Sigma}) = \ln \prod_{n=1}^{N} p(\boldsymbol{t}_n|\boldsymbol{W},\boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln p(\boldsymbol{t}_n|\boldsymbol{W},\boldsymbol{\Sigma}) \tag{297}$$

$$= \sum_{n=1}^{N} \ln \left((2\pi)^{-D/2} |\boldsymbol{\Sigma}|^{-1/2} \right) + \sum_{n=1}^{N} \frac{1}{2} (\boldsymbol{y}(\boldsymbol{x},\boldsymbol{W}) - \boldsymbol{t})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}(\boldsymbol{x},\boldsymbol{W}) - \boldsymbol{t})] \tag{298}$$

$$\frac{\partial p}{\partial \boldsymbol{W}} = \frac{1}{2} \sum_{n=1}^{N} \frac{\partial}{\partial \boldsymbol{W}} \left[(\boldsymbol{W}^T \phi(\boldsymbol{x}) - \boldsymbol{t})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{w}^T \phi(\boldsymbol{x}) - \boldsymbol{t}) \right] \tag{299}$$

$$= \frac{1}{2} \sum_{n=1}^{N} (\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-1^T}) (\boldsymbol{W}^T \phi(\boldsymbol{x}) - \boldsymbol{t}) \phi(\boldsymbol{x})^T = 0 \tag{300}$$

$$= \sum_{n=1}^{N} \boldsymbol{\Sigma}^{-1} (\boldsymbol{W}^T \phi(\boldsymbol{x}) - \boldsymbol{t}) \phi(\boldsymbol{x})^T \tag{301}$$

$$\boldsymbol{\Sigma}^{-1} \sum_{n=1}^{N} \boldsymbol{W}^T \phi(\boldsymbol{x}) \phi(x)^T = \boldsymbol{\Sigma}^{-1} \sum_{n=1}^{N} \boldsymbol{t} \phi(\boldsymbol{x})^T \tag{302}$$

$$\boldsymbol{W}^T \boldsymbol{\Phi}^T \boldsymbol{\Phi} = \boldsymbol{T} \boldsymbol{\Phi}^T \tag{303}$$

$$\boldsymbol{W}^T = \boldsymbol{T} \boldsymbol{\Phi}^T (\boldsymbol{\Phi}^T \boldsymbol{\Phi})^{-1} \tag{304}$$

$$\boldsymbol{W} = (\boldsymbol{\Phi}^T \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^T \boldsymbol{T} \tag{305}$$

$$\frac{d \ln p(\boldsymbol{T}|\boldsymbol{x},\boldsymbol{W},\boldsymbol{\Sigma}}{d\boldsymbol{\Sigma}} = -\frac{N}{2} \frac{d}{d\boldsymbol{\Sigma}} \ln |\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{n=1}^{N} \frac{d}{d\boldsymbol{\Sigma}} (\boldsymbol{t}_n - \boldsymbol{W}_{ML}^T \phi(\boldsymbol{x}_n))^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{t}_n - \boldsymbol{W}_{ML}^T \phi(\boldsymbol{x}_n)) \tag{306}$$

$$= \frac{N}{2} \frac{d}{d\boldsymbol{\Sigma}} \ln |\boldsymbol{\Sigma}|^{-1} - \frac{1}{2} \sum_{n=1}^{N} (\boldsymbol{t}_n - \boldsymbol{W}_{ML}^T \phi(\boldsymbol{x}_n)) (\boldsymbol{t}_n - \boldsymbol{W}_{ML}^T \phi(\boldsymbol{x}_n))^T \tag{307}$$

$$= \frac{N}{2} \frac{d}{d\boldsymbol{\Sigma}} \ln |\boldsymbol{\Sigma}^{-1}| - \frac{1}{2} \sum_{n=1}^{N} (\boldsymbol{t}_n - \boldsymbol{W}_{ML}^T \phi(\boldsymbol{x}_n)) (\boldsymbol{t}_n - \boldsymbol{W}_{ML}^T \phi(\boldsymbol{x}_n))^T \tag{308}$$

$$= \frac{N}{2} \boldsymbol{\Sigma} - \frac{1}{2} \sum_{n=1}^{N} (\boldsymbol{t}_n - \boldsymbol{W}_{ML}^T \phi(\boldsymbol{x}_n)) (\boldsymbol{t}_n - \boldsymbol{W}_{ML}^T \phi(\boldsymbol{x}_n))^T \tag{309}$$

$$\tag{310}$$

$$\boldsymbol{\Sigma} = \frac{1}{N} \sum_{n=1}^{N} (\boldsymbol{t}_n - \boldsymbol{W}_{ML}^T \phi(\boldsymbol{x}_n)) (\boldsymbol{t}_n - \boldsymbol{W}_{ML}^T \phi(\boldsymbol{x}_n))^T \tag{311}$$

3.7

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

$$p(\boldsymbol{w}|\boldsymbol{t}) \propto p(\boldsymbol{w})p(\boldsymbol{t}|,\boldsymbol{x},\boldsymbol{w},\beta^{-1}) \tag{312}$$

$$= \mathcal{N}(\boldsymbol{w}|_0,\boldsymbol{S}_0)\prod_{n=1}^{N}\mathcal{N}(t_n|\boldsymbol{w}^T\phi(\boldsymbol{x}_n),\beta^{-1}) \tag{313}$$

$$= (2\pi)^{-D/2}|\boldsymbol{S}_0|^{-1/2}\exp[-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{m}_0)^T\boldsymbol{S}_0^{-1}(\boldsymbol{w}-\boldsymbol{m}_0)]\prod_{n=1}^{N}\left(\frac{\beta}{2\pi}\right)^{\frac{1}{2}}\exp[-\frac{\beta}{2}(t_n-\boldsymbol{w}^T\phi(\boldsymbol{x}_n))^2] \tag{314}$$

$$= (2\pi)^{-D/2}|\boldsymbol{S}_0|^{-1/2}\left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}}\exp[-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{m}_0)^T\boldsymbol{S}_0^{-1}(\boldsymbol{w}-\boldsymbol{m}_0)+\sum_{n=1}^{N}-\frac{\beta}{2}(t_n-\boldsymbol{w}^T\phi(\boldsymbol{x}_n))^2] \tag{315}$$

$$\tag{316}$$

Now we have to get the exponent right.

$$-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{m}_0)^T\boldsymbol{S}_0^{-1}(\boldsymbol{w}-\boldsymbol{m}_0)+\sum_{n=1}^{N}-\frac{\beta}{2}(t_n-\boldsymbol{w}^T\phi(\boldsymbol{x}_n))^2 \tag{317}$$

$$= -\frac{1}{2}(\boldsymbol{w}-\boldsymbol{m}_0)^T\boldsymbol{S}_0^{-1}(\boldsymbol{w}-\boldsymbol{m}_0)-\frac{\beta}{2}\sum_{n=1}^{N}(t_n^2+\boldsymbol{w}^T\phi(\boldsymbol{x}_n)\boldsymbol{w}^T\phi(\boldsymbol{x}_n)-2\boldsymbol{w}^T\phi(\boldsymbol{x}_n)t_n) \tag{318}$$

$$= -\frac{1}{2}(\boldsymbol{w}^T\boldsymbol{S}_0^{-1}\boldsymbol{w}-\boldsymbol{w}^T\boldsymbol{S}_0^{-1}\boldsymbol{m}_0-\boldsymbol{m}_0^T\boldsymbol{S}_0^{-1}\boldsymbol{w}+\boldsymbol{m}_0^T\boldsymbol{S}_0^{-1}\boldsymbol{m}_0)-\frac{\beta}{2}(\boldsymbol{t}^T\boldsymbol{t}+\boldsymbol{w}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{w}-2\boldsymbol{w}^T\boldsymbol{\Phi}^T\boldsymbol{t}) \tag{319}$$

$$= -\frac{1}{2}\left(\boldsymbol{w}^T\boldsymbol{S}_0^{-1}\boldsymbol{w}-2\boldsymbol{w}^T\boldsymbol{S}_0^{-1}\boldsymbol{m}_0+\boldsymbol{m}_0^T\boldsymbol{S}_0^{-1}\boldsymbol{m}_0-\beta\boldsymbol{t}^T\boldsymbol{t}+\beta\boldsymbol{w}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{w}-2\beta\boldsymbol{w}^T\boldsymbol{\Phi}^T\boldsymbol{t}\right) \tag{320}$$

$$= \boldsymbol{w}^T(-\frac{1}{2}(\boldsymbol{S}_0^{-1}+\beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}))\boldsymbol{w}+\boldsymbol{w}(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0+\boldsymbol{\Phi}^T\boldsymbol{t})+c \tag{321}$$

$$= \boldsymbol{w}^T(-\frac{1}{2}\boldsymbol{S}_N^{-1})\boldsymbol{w}+\boldsymbol{w}(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0+\boldsymbol{\Phi}^T\boldsymbol{t})+c \tag{322}$$

Now completing the square.

$$\boldsymbol{w}^T(-\frac{1}{2}\boldsymbol{S}_N^{-1})\boldsymbol{w}+\boldsymbol{w}(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0+\boldsymbol{\Phi}^T\boldsymbol{t})+c \tag{323}$$

$$= (\boldsymbol{w}+\frac{1}{2}(-\frac{1}{2}S_N^{-1})^{-1}(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0+\boldsymbol{\Phi}^T\boldsymbol{t}))^T(-\frac{1}{2}\boldsymbol{S}_n^{-1})(\boldsymbol{w}+\frac{1}{2}(-\frac{1}{2}S_N^{-1})^{-1}(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0+\boldsymbol{\Phi}^T\boldsymbol{t})+k \tag{324}$$

$$= -\frac{1}{2}(\boldsymbol{w}-\boldsymbol{m}_N)^T\boldsymbol{S}_n^{-1}(\boldsymbol{w}-\boldsymbol{m}_N)+k \tag{325}$$

3.15

Hint 1: Use 3.95 and 3.92

$$E(\boldsymbol{m}_N) = \frac{\beta}{2}\|\boldsymbol{t} - \boldsymbol{\Phi}\boldsymbol{m}_N\|^2 + \frac{\alpha}{2}\boldsymbol{m}_N^T\boldsymbol{m}_N \tag{326}$$

$$= \frac{\beta}{2}\|\boldsymbol{t} - \boldsymbol{\Phi}\boldsymbol{m}_N\|^2 + \frac{\gamma(\boldsymbol{m}_N^T\boldsymbol{m}_N)^{-1}}{2}\boldsymbol{m}_N^T\boldsymbol{m}_N \tag{327}$$

$$= \frac{\beta}{2}\|\boldsymbol{t} - \boldsymbol{\Phi}\boldsymbol{m}_N\|^2 + \frac{\gamma}{2} \tag{328}$$

$$\tag{329}$$

$$\beta = (N - \gamma)\|\boldsymbol{t} - \boldsymbol{\Phi}^T\boldsymbol{m}_n\|^{-2} \tag{330}$$

$$E(\boldsymbol{m}_N) = \frac{1}{2}(\mathcal{N} - \gamma + \gamma) = \frac{N}{2} \tag{331}$$

3.17

Evidence function:

$$p(D|M_i) = \int p(D|\boldsymbol{w}, M_i)p(\boldsymbol{w}, M_i)d\boldsymbol{w} \tag{332}$$

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

$$p(D|\alpha, \beta) = \int p(\boldsymbol{w}|\alpha^{-1}I)\prod_{n=1}^{N} p(t_n|\boldsymbol{w}^T\phi(\boldsymbol{x}_n), \beta^{-1}) \tag{333}$$

$$= \int (2\pi)^{-M/2}|\alpha^{-1}I|^{-1/2}\exp[-1/2(\boldsymbol{w}^T(\alpha^{-1}I)^{-1}\boldsymbol{w}]\prod_{n=1}^{N}\frac{1}{\sqrt{2\pi\beta^{-1}}}\exp[\beta/2(t_n - \boldsymbol{w}^T\phi(\boldsymbol{x}_n))^2] \tag{334}$$

$$= \int (2\pi)^{-M/2}(\alpha^{-1})^{-M/2}\frac{\beta}{2\pi}^{N/2}\exp[-\alpha/2\boldsymbol{w}^T\boldsymbol{w} - \beta/2\|\boldsymbol{t} - \boldsymbol{\Phi}\boldsymbol{w}\|^2] \tag{335}$$

4.4

$$\mathcal{L} = \boldsymbol{w}^T(\boldsymbol{m}_2 - \boldsymbol{m}_1) + \lambda(\boldsymbol{w}^T\boldsymbol{w} - 1) \tag{336}$$

$$\nabla_{\boldsymbol{w}}\mathcal{L} = (\boldsymbol{m}_2 - \boldsymbol{m}_1) + 2\lambda\boldsymbol{w} = 0 \tag{337}$$

$$\boldsymbol{w} = -\frac{1}{2\lambda}(\boldsymbol{m}_2 - \boldsymbol{m}_1) \propto (\boldsymbol{m}_2 - \boldsymbol{m}_1) \tag{338}$$

4.5

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

$$(m_2 - m_1)^2 = (\boldsymbol{w}^T\boldsymbol{m}_2 - \boldsymbol{w}^T\boldsymbol{m}_1)^2 \tag{339}$$
$$= \boldsymbol{w}^T\boldsymbol{m}_2\boldsymbol{w}^T\boldsymbol{m}_2 - \boldsymbol{w}^T\boldsymbol{m}_2\boldsymbol{w}^T\boldsymbol{m}_1 - \boldsymbol{w}^T\boldsymbol{m}_1\boldsymbol{w}^T\boldsymbol{m}_2 + \boldsymbol{w}^T\boldsymbol{m}_1\boldsymbol{w}^T\boldsymbol{m}_1 \tag{340}$$
$$= \boldsymbol{w}^T(\boldsymbol{m}_2\boldsymbol{m}_2^T - \boldsymbol{m}_2\boldsymbol{m}_1^T - \boldsymbol{m}_1\boldsymbol{m}_2^T + \boldsymbol{m}_1\boldsymbol{m}_1^T)\boldsymbol{w} \tag{341}$$
$$= \boldsymbol{w}^T(\boldsymbol{m}_2 - \boldsymbol{m}_1)(\boldsymbol{m}_2 - \boldsymbol{m}_1)^T\boldsymbol{w} \tag{342}$$

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

$$s_1^2 = \sum_{n\in C_1}(\boldsymbol{w}^T\boldsymbol{x}_n - \boldsymbol{w}^T\boldsymbol{m}_1)^2 \tag{343}$$
$$= \sum_{n\in C_1}(\boldsymbol{w}^T\boldsymbol{x}_n\boldsymbol{w}^T\boldsymbol{x}_n - \boldsymbol{w}^T\boldsymbol{x}_n\boldsymbol{w}^T\boldsymbol{m}_1 - \boldsymbol{w}^T\boldsymbol{m}_1\boldsymbol{w}^T\boldsymbol{x}_n + \boldsymbol{w}^T\boldsymbol{m}_1\boldsymbol{w}^T\boldsymbol{m}_1) \tag{344}$$
$$= \sum_{n\in C_1}\boldsymbol{w}^T(\boldsymbol{x}_n\boldsymbol{x}_n^T - \boldsymbol{x}_n\boldsymbol{m}_1^T - \boldsymbol{m}_1\boldsymbol{x}_n^T + \boldsymbol{m}_1\boldsymbol{m}_1^T)\boldsymbol{w} \tag{345}$$
$$= \sum_{n\in C_1}\boldsymbol{w}^T(\boldsymbol{x}_n - \boldsymbol{m}_1)(\boldsymbol{x}_n - \boldsymbol{m}_1)^T\boldsymbol{w} \tag{346}$$

4.6

$$\sum_{n=1}^{N}(\boldsymbol{w}^T\boldsymbol{x}_n + w_0 - t_n)\boldsymbol{x}_n = 0 \tag{347}$$
$$= \sum_{n=1}^{N}(\boldsymbol{w}^T\boldsymbol{x}_n - \boldsymbol{w}^T\boldsymbol{m} - t_n)\boldsymbol{x}_n \tag{348}$$
$$\boldsymbol{w}^T\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{m})\boldsymbol{x}_n = \sum_{n=1}^{N}t_n\boldsymbol{x}_n \tag{349}$$

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

$$\sum_{n=1}^{N}t_n\boldsymbol{x}_n = \sum_{n\in C_1}^{N_1}t_n\boldsymbol{x}_n - \sum_{n\in C_2}^{N_2}t_n\boldsymbol{x}_n \tag{350}$$
$$= \sum_{n\in C_1}^{N_1}\frac{N}{N_1}\boldsymbol{x}_n - \sum_{n\in C_2}^{N_2}\frac{N}{N_2}\boldsymbol{x}_n \tag{351}$$
$$= N\left(\sum_{n\in C_1}^{N_1}\frac{1}{N_1}\boldsymbol{x}_n - \sum_{n\in C_2}^{N_2}\frac{1}{N_2}\boldsymbol{x}_n\right) \tag{352}$$
$$= N(\boldsymbol{m}_1 - \boldsymbol{m}_2) \tag{353}$$

$$-\boldsymbol{w}^T\boldsymbol{m}\sum_{n=1}^{N}\boldsymbol{x}_n = -\frac{1}{N}\boldsymbol{w}^T(N_1\boldsymbol{m}_1 + N_2\boldsymbol{m}_2)(N_1\boldsymbol{m}_1 + N_2\boldsymbol{m}_2) \tag{354}$$

$$= -\frac{1}{N}(N_1\boldsymbol{m}_1 + N_2\boldsymbol{m}_2)(N_1\boldsymbol{m}_1 + N_2\boldsymbol{m}_2)^T\boldsymbol{w} \tag{355}$$

$$= -\frac{1}{N}(N_1^2\boldsymbol{m}_1\boldsymbol{m}_1^T + N_1N_2\boldsymbol{m}_1\boldsymbol{m}_2^T + N_2N_1\boldsymbol{m}_2\boldsymbol{m}_1^T + N_2^2\boldsymbol{m}_2\boldsymbol{m}_2^T)\boldsymbol{w} \tag{356}$$

$$= -\frac{1}{N}\left((N-N_2)N_1\boldsymbol{m}_1\boldsymbol{m}_1^T + N_1N_2\boldsymbol{m}_1\boldsymbol{m}_2^T + N_2N_1\boldsymbol{m}_2\boldsymbol{m}_1^T + (N-N_1)N_2\boldsymbol{m}_2\boldsymbol{m}_2^T\right)\boldsymbol{w} \tag{357}$$

$$= \left(-N_1\boldsymbol{m}_1\boldsymbol{m}_1^T + \frac{N_1N_2}{N}\boldsymbol{m}_1\boldsymbol{m}_1^T - \frac{N_1N_2}{N}\boldsymbol{m}_1\boldsymbol{m}_2^T - \frac{N_2N_1}{N}\boldsymbol{m}_2\boldsymbol{m}_1^T - N_2\boldsymbol{m}_2\boldsymbol{m}_2^T + \frac{N_1N_2}{N}\boldsymbol{m}_2\boldsymbol{m}_2^T\right)\boldsymbol{w} \tag{358}$$

$$= \left(-N_1\boldsymbol{m}_1\boldsymbol{m}_1^T - N_2\boldsymbol{m}_2\boldsymbol{m}_2^T + \frac{N_1N_2}{N}(\boldsymbol{m}_2-\boldsymbol{m}_1)(\boldsymbol{m}_2-\boldsymbol{m}_1)^T\right)\boldsymbol{w} \tag{359}$$

$$= \left(-N_1\boldsymbol{m}_1\boldsymbol{m}_1^T - N_2\boldsymbol{m}_2\boldsymbol{m}_2^T + \frac{N_1N_2}{N}\boldsymbol{S}_B\right)\boldsymbol{w} \tag{360}$$

$$\tag{361}$$

We add the remaining terms to the final term:

$$\left(-N_1\boldsymbol{m}_1\boldsymbol{m}_1^T - N_2\boldsymbol{m}_2\boldsymbol{m}_2^T + \sum_{n=1}^{N}\boldsymbol{x}_n\boldsymbol{x}_n^T\right)\boldsymbol{w} \tag{362}$$

$$= \left(N_1\boldsymbol{m}_1\boldsymbol{m}_1^T - 2N_1\boldsymbol{m}_1\boldsymbol{m}_1^T + N_2\boldsymbol{m}_2\boldsymbol{m}_2^T - 2N_2\boldsymbol{m}_2\boldsymbol{m}_2^T + \sum_{n=1}^{N}\boldsymbol{x}_n\boldsymbol{x}_n^T\right)\boldsymbol{w} \tag{363}$$

$$= \left(N_1\boldsymbol{m}_1\boldsymbol{m}_1^T - \sum_{n\in C_1}\boldsymbol{x}_n\boldsymbol{m}_1^T - \boldsymbol{m}_1\sum_{n\in C_1}\boldsymbol{x}_1^T + ... + \sum_{n=1}^{N}\boldsymbol{x}_n\boldsymbol{x}_n^T\right)\boldsymbol{w} \tag{364}$$

$$= \left(\sum_{n\in C_1}\boldsymbol{m}_1\boldsymbol{m}_1^T - \boldsymbol{x}_n\boldsymbol{m}_1^T - \boldsymbol{m}_1\boldsymbol{x}_n^T + \boldsymbol{x}_n\boldsymbol{x}_n^T + ...\right)\boldsymbol{w} \tag{365}$$

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

4.7

$$\sigma(-a) = \frac{1}{e^x+1} = \frac{e^{-x}}{1+e^{-x}} = \frac{e^{-x}+1}{e^{-x}+1} - \frac{1}{e^{-x}+1} \tag{366}$$

$$y = \frac{1}{1+e^{-x}} \tag{367}$$

$$e^{-x} = \frac{1-y}{y} \tag{368}$$

$$y = e^x(1-y) \tag{369}$$

$$e^x = y/(1-y) \tag{370}$$

$$x = \ln\left[y/(1-y)\right] \tag{371}$$

4.8

$$p(C_1|\boldsymbol{x}) = \sigma(a) \tag{372}$$

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

$$a = \ln\frac{p(\boldsymbol{x}|C_1)}{p(\boldsymbol{x}|C_2)} + \ln\frac{p(C_1)}{p(C_2)} \tag{373}$$

$$\ln\frac{p(\boldsymbol{x}|C_1)}{p(\boldsymbol{x}|C_2)} = \ln\left[\frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)\right)\right] \tag{374}$$

$$-\ln\left[\frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)\right)\right] \tag{375}$$

$$= -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\boldsymbol{x}-\mu_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2) \tag{376}$$

$$= \frac{1}{2}\left[2\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - 2\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 + \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right] \tag{377}$$

$$\tag{378}$$

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.

$$\ln p(\boldsymbol{X}|\boldsymbol{T}) = \sum_{n=1}^N \ln\prod_{j=1}^K (\pi_j\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma})_j^t) \tag{379}$$

$$= \sum_{n=1}^N\sum_{j=1}^K t_j \ln(\pi_j\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k\boldsymbol{\Sigma})) \tag{380}$$

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

$$\mathcal{L}(\boldsymbol{\pi},\lambda) = \ln(p(\boldsymbol{X},\boldsymbol{T})) - \lambda(\sum_{j=1}^K \pi_j - 1) \tag{381}$$

$$\frac{\partial\mathcal{L}}{\partial\pi_j} = \sum_{n=1}^N \frac{t_j}{\pi_j} - \lambda = 0 \tag{382}$$

$$\lambda = \sum_{n=1}^N \frac{t_j}{\pi_j} = N\frac{t_j}{\pi_j} = \frac{N_j}{\pi_j} \tag{383}$$

$$\pi_j = \frac{N_j}{\lambda} \tag{384}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = \sum_{j=1}^{K} \pi_j = 1 \tag{385}$$

Plugging this into 380.

$$\sum_{j=1}^{K} \frac{N_j}{\lambda} = 1 \iff \lambda = N \tag{386}$$

Gives us the desired result.

4.12

$$\frac{d\sigma}{da} = (1 + e^{-a})^{-2} e^{-a} \tag{387}$$

$$= \frac{1}{1 + e^{-a}} \frac{1}{1 + e^{-a}} e^{-a} \tag{388}$$

$$= \sigma(a) \frac{e^{-a}}{1 + e^{-a}} \tag{389}$$

$$= \sigma(a) \left[\frac{1 + e^{-a}}{1 + e^{-a}} - \frac{1}{1 + e^{-a}} \right] \tag{390}$$

4.13

$$\nabla_{\boldsymbol{w}} E(\boldsymbol{w}) = -\sum_{n=1}^{N} \frac{t_n}{y_n} \nabla_{\boldsymbol{w}} \sigma(\boldsymbol{w}^T \boldsymbol{\phi}_n) - \frac{1 - t_n}{1 - y_n} \nabla_{\boldsymbol{w}} \sigma(\boldsymbol{w}^T \boldsymbol{\phi}_n) \tag{391}$$

$$= -\sum_{n=1}^{N} \frac{t_n}{y_n} \sigma(\boldsymbol{w}^T \boldsymbol{\phi}_n)(1 - \sigma(\boldsymbol{w}^T \boldsymbol{\phi}_n)) \boldsymbol{\phi}_n - \frac{1 - t_n}{1 - y_n} \sigma(\boldsymbol{w}^T \boldsymbol{\phi}_n)(1 - \sigma(\boldsymbol{w}^T \boldsymbol{\phi}_n)) \boldsymbol{\phi}_n \tag{392}$$

$$= -\sum_{n=1}^{N} t_n (1 - y_n) \boldsymbol{\phi}_n - (1 - t_n) y_n \boldsymbol{\phi}_n \tag{393}$$

$$= -\sum_{n=1}^{N} (t_n - y_n) \boldsymbol{\phi}_n \tag{394}$$

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

$$p(\boldsymbol{t}, \boldsymbol{w}) = \prod_{n=1}^{N} y_n^{\pi}[1 - y_n]^{1-t_n} \tag{395}$$

$$\ln p = \sum_{n=1}^{N} \pi_n \ln y_n + (1 - \pi_n) \ln(1 - y_n) \tag{396}$$

4.17

$$p(C_k|\boldsymbol{\phi}) = y_k = \frac{\exp a_k}{\sum_{j=1} \exp a_j} \tag{397}$$

$$\frac{\partial y_k}{\partial a_j} = -\exp a_k (\sum_j \exp(a_j)^{-2} \exp(a_j) \tag{398}$$

$$= \begin{cases} y_k(0 - y_j) & j \neq k \\ y_k(1 - y_j) & j = k \end{cases} \tag{399}$$

$$= y_k(I_{kj} - y_j) \tag{400}$$

4.18

$$\nabla_{\boldsymbol{w}_j} E(\boldsymbol{W}) = -\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \nabla_{\boldsymbol{w}_j} \ln y_n k \tag{401}$$

$$\nabla_{\boldsymbol{w}_j} \ln(y_{nk}) = -(I_{kj} - y_j)\boldsymbol{\phi}_n \tag{402}$$

$$\nabla_{\boldsymbol{w}_j} E(\boldsymbol{W}) = -\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk}(I_{kj} - y_j n)\boldsymbol{\phi}_n \tag{403}$$

$$= \boldsymbol{\phi} \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} y_{jn} \boldsymbol{\phi} - t_{nk} I_{kj} \tag{404}$$

$$= \boldsymbol{\phi} \sum_{n=1}^{N} t_{nj} - y_{jn} \boldsymbol{\phi} \underbrace{\sum_{k=1}^{K} t_{nk}}_{=1} \tag{405}$$

$$= \sum_{n=1}^{N} \boldsymbol{\phi}(y_{jn} - t_{nj}) \tag{406}$$

4.19

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

$$p(\boldsymbol{t}, \boldsymbol{w}) = \sum_{n=1}^{N} [t_n \ln y_n + (1 - t_n \ln(1 - y_n] \tag{407}$$

$$\nabla_{\boldsymbol{w}} \sum_{n=1}^{N} [t_n \ln \Phi(a) + (1 - t_n) \ln(1 - \Phi(a))] = \sum_{n=1}^{N} (\frac{t_n}{\Phi(a)} - \frac{1 - t_n}{1 - \Phi(a)}) \Phi(a) \boldsymbol{\phi}_n \tag{408}$$

4.21

$$\Phi(a) = \int_0^a \mathcal{N}(0, 1) d\theta \tag{409}$$

$$= \int_0^a 1/\sqrt{2\pi} \exp(-\frac{1}{2}\theta^2) d\theta \tag{410}$$

$$= \frac{1}{2} + \int_{-\infty}^a \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}\theta^2) d\theta \tag{411}$$

$$= \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \frac{\sqrt{\pi}}{2} \int_{-\infty}^a \frac{2}{\sqrt{\pi}} \exp(-\frac{1}{2}\theta^2) d\theta \tag{412}$$

$$= \frac{1}{2} \left(1 + \frac{1}{\sqrt{2}} \text{erf}(a) \right) \tag{413}$$

4.22

$$\ln p(D) = \ln \left[f(z_0) \frac{(2\pi)^{M/2}}{|A|^{1/2}} \right] \tag{414}$$

$$= \ln f(z_0) = \frac{M}{2} \ln 2\pi - \frac{1}{2} \ln |A| \tag{415}$$

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

$$\ln f(\boldsymbol{\theta}) \Big| z_0 = \ln f(\boldsymbol{\theta}_{MAP}) = \ln p(D|\boldsymbol{\theta}_{MAP}) + \ln p(\boldsymbol{\theta}_{MAP}) \tag{416}$$

5.2

$$p(\boldsymbol{T}, \boldsymbol{X}, \boldsymbol{w}) = \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{t}_n | y(\boldsymbol{x}_n, \boldsymbol{w}), \beta^{-1} I) \tag{417}$$

$$= \prod_{n=1}^{N} \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\beta^{-1} I|^{\frac{1}{2}}} \exp(-\frac{1}{2} (\boldsymbol{t}_n - y(\boldsymbol{x}_n, \boldsymbol{w}))^T \beta^{-1} I (\boldsymbol{t}_n - y(\boldsymbol{x}_n, \boldsymbol{w}))) \tag{418}$$

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

$$(-\frac{1}{2}(\boldsymbol{t}_n - y(\boldsymbol{x}_n, \boldsymbol{w}))^T \beta^{-1} I(\boldsymbol{t}_n - y(\boldsymbol{x}_n, \boldsymbol{w}))) = -\frac{1}{2\beta}\|y(\boldsymbol{x}_n, \boldsymbol{w}) - \boldsymbol{t}_n\|^2. \tag{419}$$

5.5

$$p(\boldsymbol{T}, \boldsymbol{X}, \boldsymbol{w}) = \prod_{n=1}^{N}\prod_{k=1}^{K} y_k(\boldsymbol{x}_n, \boldsymbol{w})^{t_n^k}(1 - y_k)(\boldsymbol{x}_n, \boldsymbol{w})^{1-t_n^k} \tag{420}$$

Taking the log becomes the cross-entropy function.

5.6

$$\frac{\partial E(\boldsymbol{w})}{\partial a_k} = \sum_{n=1}^{N}\left(\frac{\partial}{\partial a_k} t_n \ln y_n + \frac{\partial}{\partial a_k}(1 - t_n)\ln(1 - y_n)\right) \tag{421}$$

$$= -\frac{t_k}{y_k}\frac{\partial}{\partial a_k} y_k - \frac{1 - t_k}{1 - y_k}\frac{\partial}{\partial a_k}(1 - y_k) \tag{422}$$

$$= -t_k(1 - y_k) + (1 - t_k)y_k \tag{423}$$

$$= y_k - t_k \tag{424}$$

5.7

$$E(\boldsymbol{w}) = \sum_{k=1}^{K} t_k \ln y_k(\boldsymbol{x}, \boldsymbol{w}) \tag{425}$$

$$\frac{\partial E}{\partial a_j} = -\sum_{k=1}^{K}\frac{t_k}{y_k} y_k(I_{kj} - y_j) \tag{426}$$

$$= -\sum_{k=1}^{K} t_k(I_{kj} - y_j) \tag{427}$$

$$= -t_j + y_j \tag{428}$$

$$\frac{d\tanh}{da} = (e^a - e^{-a})\frac{d}{da}(e^a + e^{-a})^{-1} + (e^a + e^{-a})^{-1}\frac{d}{da}(e^a - e^{-a}) \tag{429}$$

$$= -(e^a - e^{-a})(e^a + e^{-a})^{-2}\frac{d}{da}(e^a + e^{-a}) + (e^a + e^{-a})^{-1}(e^a + e^{-a}) \tag{430}$$

$$= -(e^a - e^{-a})(e^a + e^{-a})^{-2}(e^a - e^{-a}) + 1 \tag{431}$$

$$= -h^2(a) + 1 \tag{432}$$

5.9

Still Bernoulli, so

$$p(t|\boldsymbol{x}, \boldsymbol{w}) = (\frac{1+y}{2})^{\frac{1+t}{2}}(\frac{1-y}{2})^{\frac{1-t}{2}} \tag{433}$$

$$p(\boldsymbol{t}|\boldsymbol{X}, \boldsymbol{w}) = \prod_{n=1}^{N} p(t_n|\boldsymbol{x}_n, \boldsymbol{w}) \tag{434}$$

$$-\ln(p(\boldsymbol{t}|\boldsymbol{X}, \boldsymbol{w}) = -\sum_{n=1}^{N} \ln p(t_n|\boldsymbol{x}_n, \boldsymbol{w}) \tag{435}$$

$$= -\sum_{n=1}^{N} (\frac{1+t}{2}) \ln(\frac{1+y}{2}) + (\frac{1-t}{2}) \ln(\frac{1-y}{2}) \tag{436}$$

We can use tanh as activation function.

5.10

$$u_i^T H u_i = u_i^T \lambda u_i = \delta_i i \lambda_i = \lambda_i \tag{437}$$

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.

$$\frac{N(N+1)}{2} + N = \frac{N(N+3)}{2} \tag{438}$$

5.14

Hint 1: Taylor expansion on both terms in numerator.

Taylor:

$$E_n(w_j + \epsilon) = E_n(w_{ji}) + \epsilon \frac{\partial E_n}{\partial w_{ji}} + \frac{\epsilon^2}{2} \frac{\partial \partial E_n}{\partial w_{ji}} + O(\epsilon^3) \tag{439}$$

$$E_n(w_j - \epsilon) = E_n(w_{ji}) - \epsilon \frac{\partial E_n}{\partial w_{ji}} + \frac{\epsilon^2}{2} \frac{\partial \partial E_n}{\partial w_{ji}} - O(\epsilon^3) \tag{440}$$

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

5.16

$$E = \frac{1}{2}\sum_{n=1}^{N}\|\boldsymbol{y}_n - \boldsymbol{t}_n\|^2 = \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{M}(\boldsymbol{y}_n - \boldsymbol{t}_n)^2 \tag{441}$$

$$\nabla E = \sum_{n=1}^{N}\sum_{m=1}^{M}(\boldsymbol{y}_n - \boldsymbol{t}_n)\nabla \boldsymbol{y}_n \tag{442}$$

$$\nabla\nabla E = \sum_{n=1}^{N}\sum_{m=1}^{M}\nabla \boldsymbol{y}_n \nabla(\boldsymbol{y}_n - \boldsymbol{t}_n) + (\boldsymbol{y}_n - \boldsymbol{t}_n)\nabla\nabla \boldsymbol{y}_n \approx \sum_{n=1}^{N}\sum_{m=1}^{M}\nabla_{\boldsymbol{y}_n}\nabla \boldsymbol{y}_n^T \tag{443}$$

5.17

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

$$\frac{\partial E}{\partial w_r} = \int\int (y(x,w) - t)\frac{\partial y}{\partial w_r}p(x,t)dxdt \tag{444}$$

$$\frac{\partial^2 E}{\partial w_r \partial w_s} = \int\int \left[(y(x,w) - t)\frac{\partial^2 y(x,w)}{\partial w_r \partial w_s} + \frac{\partial y(x,w)}{\partial w_r}\frac{\partial y}{\partial w_s}\right]p(x,t)dxdt \tag{445}$$

$$\int\int (y(x,w) - t)p(x,t)dxdt = \int\int (y(x,w) - t)p(t|x)p(x)dxdt \tag{446}$$

$$= \int p(x)\left(y(x,w) - \underbrace{\int tp(t|x)}_{=y(x,w)}\right)dxdt \qquad = 0 \tag{447}$$

The remaining integral is the answer.

5.18

$$y_k(\boldsymbol{x}, \boldsymbol{w}) = \sigma(\sum_{j=1}^{M} w_{kj}h(\sum_{i=1}^{D} w_{ji}x_i + w_{j0}) + \sum_{l=1}^{D} w_l i x_i) \tag{448}$$

Finding the derivatives to these skip weights is straightforward.

5.19

$$E = -\sum_{n=1}^{N} t_n \ln y_n + (1 - t_n)\ln(1 - y_n) \tag{449}$$

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

$$\nabla\nabla E = \sum_{n=1}^{N} y_n(1 - y_n)\nabla a_n \nabla a_n^T + (y_n - t_n)\nabla\nabla a_n \tag{450}$$

5.20

$$\nabla_{w_j} E(\boldsymbol{W}) = \sum_{n=1}^{N} (y_{n_j} - bt_{n_j}) \nabla a_j \tag{451}$$

$$\nabla\nabla \boldsymbol{w}_j \approx \sum_{n=1} y_k (I - y_j) \nabla a_j \nabla a_j^T \tag{452}$$

5.24

$$\sum_i \frac{1}{a} w_{ji}(ax_i + b) + w_{j0} - \frac{b}{a} \sum_i w_{ji} = \sum_i w_{ji} x + \frac{b}{a} w_{ji} + w_{j0} - \frac{b}{a} \sum_i w_{ji} \tag{453}$$

$$= \sum_i w_{ji} x_i + w_{j0} \tag{454}$$

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

$$\tilde{E}(\boldsymbol{w}) = E(\boldsymbol{w}) + \lambda \Omega(\boldsymbol{w}) \tag{455}$$

$$\frac{\partial \tilde{E}}{\partial w_i} = \frac{\partial E}{\partial w_i} + \frac{\partial}{\partial w_i} \lambda \Omega(\boldsymbol{w}) \tag{456}$$

$$\frac{\partial}{\partial w_i} \lambda \Omega(\boldsymbol{w}) = -\lambda \frac{\partial}{\partial w_i} \ln(\sum_{j=1}^{M} \pi_j \mathcal{N}(w_i | \mu_j, \sigma_j^2)) \tag{457}$$

$$\frac{\partial}{\partial w_i} = \frac{1}{\sum_{j=1}^{M} \pi_j \mathcal{N}(w_i | \mu_j, \sigma_j^2)} \frac{\partial}{\partial w_i} \ln(\sum_{j=1}^{M} \pi_j \mathcal{N}(w_i | \mu_j, \sigma_j^2)) \tag{458}$$

$$\frac{\partial}{\partial w_i} \ln(\sum_{j=1}^{M} \pi_j \mathcal{N}(w_i | \mu_j, \sigma_j^2)) = \sum_{j=1}^{M} \pi_j \frac{\partial}{\partial w_i} \mathcal{N}(w_i | \mu_j, \sigma_j^2) \tag{459}$$

$$\frac{\partial}{\partial w_i} \mathcal{N}(w_i | \mu_j, \sigma_j^2) = \frac{1}{\sqrt{2\pi\sigma_j^2} \exp(-\frac{1}{2} \frac{(w_i - \mu_j)^2}{\sigma_j^2})} \frac{\partial}{\partial w_i} (-\frac{1}{2} \frac{(w_i - \mu_j)^2}{\sigma_j^2}) \tag{460}$$

$$\frac{\partial}{\partial w_i} = -\sigma_j^{-2} (w_i - \mu_j) \tag{461}$$

Plugging everything in gives

$$\frac{\partial}{\partial w_i} = \frac{\sum_{j=1}^{M} \pi_j \mathcal{N}(w_i|\mu_j, \sigma_j)}{\sum_{k=1}^{K} \mathcal{N}(w_i|\mu_j, \sigma_j^2)} \tag{462}$$

5.30

$$\frac{\partial \tilde{E}}{\partial \mu_j} = -\lambda \frac{\partial}{\partial \mu_j} \sum_i \ln \sum_{j=1}^{M} \pi_j \mathcal{N}(w_i|\mu_j, \sigma_j) \tag{463}$$

$$= -\lambda \sum_i \frac{1}{\sum_j \pi_j \mathcal{N}(w_i} \tag{464}$$

$$= -\lambda \sum_i \frac{1}{\sum_j \pi_j N(w_i|\mu_j, \sigma_j)} \sum_j \pi_j \frac{d}{d\mu_j} N(w_i|\mu_j, \sigma_j) \tag{465}$$

$$\frac{d}{d\mu_j} N(w_i|\mu_j, \sigma_j) = -\frac{1}{\sqrt{2\pi\sigma_i^2}} \exp(-\frac{1}{2\sigma_i^2}(w_i - \mu_j)^2)[\frac{1}{\sigma_j^2}(w_i - \mu_j] \tag{466}$$

Plugging in gives the result.

5.31

$$\frac{\partial}{\partial \sigma_j} \pi_j \, \mathcal{N}(w_i|\mu_j, \sigma_j) \tag{467}$$

$$= \pi \left[\exp\left(-\frac{(\mu_j - w_i)^2)}{2\sigma_j^2} \right) \frac{\partial}{\partial \sigma_j} \frac{1}{\sqrt{2\pi\sigma_j^2}} + \frac{1}{\sqrt{2\pi\sigma_j^2}} \frac{\partial}{\partial \sigma_j} \exp\left(-\frac{(\mu_j - w_i)^2}{2\sigma_j^2} \right) \right] \tag{468}$$

$$\frac{\partial}{\partial \sigma_j} (2\pi\sigma_j^2)^{-\frac{1}{2}} = -\frac{1}{2} (2\pi\sigma_j^2)^{\frac{3}{2}} \frac{\partial}{\partial \sigma_j} (2\pi\sigma_j^2) \tag{469}$$

$$= -\frac{1}{2} (2\pi\sigma_j^2)^{-\frac{3}{2}} 4\pi\sigma_j \tag{470}$$

$$= -\frac{1}{\sigma_j} \frac{1}{\sqrt{2\pi\sigma_j^2}} \tag{471}$$

$$\frac{\partial}{\partial \sigma_j} \exp\left(-\frac{(\mu_j - w_i)^2}{2\sigma_j^2} \right) = \exp\left(-\frac{(\mu_j - w_i)^2}{2\sigma_j^2} \right) \frac{\partial}{\partial \sigma_j} \left(-\frac{1}{2\sigma_j^2} (\mu_j - w_i)^2 \right) \tag{472}$$

$$= \exp\left(-\frac{(\mu_j - w_i)^2}{2\sigma_j^2} \right) \frac{1}{\sigma_j^3} (\mu_j - w_i)^2 \qquad = \tag{473}$$

Plugging these values in gives the result.

5.33

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

$$\boldsymbol{y} = \begin{pmatrix} \cos(\pi - \theta)L_1 + v_1 \\ \sin(\pi - \theta)L_1 + v_2 \end{pmatrix} \tag{474}$$

$$\boldsymbol{x} = \begin{pmatrix} y_1 + \cos(\theta_1 + \theta_2 - \pi)L_2 \\ y_2 + \sin(\theta_1 + \theta_2 - \pi)L_2 \end{pmatrix} \tag{475}$$

5.34

$$\frac{\partial E_n}{\partial a_{kl}^{\pi}} = -\frac{1}{\sum_{l=1} \pi_l(\boldsymbol{x}_n, \boldsymbol{w})\mathcal{N}(\boldsymbol{t}_n|\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}), \sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w}))} \frac{\partial}{\partial a_k^{\pi}} \sum_{l=1} \pi_l(\boldsymbol{x}_n, \boldsymbol{w})\mathcal{N}(\boldsymbol{t}_n|\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}), \sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w})) \tag{476}$$

$$\frac{\partial}{\partial a_k^{\pi}} \sum_{l=1} \pi_l(\boldsymbol{x}_n, \boldsymbol{w})\mathcal{N}(\boldsymbol{t}_n|\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}), \sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w})) = \mathcal{N}(\boldsymbol{t}_n|\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}), \sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w})) \frac{\partial}{\partial a_k^{\pi}} \sum_{l=1} \pi_k \tag{477}$$

$$\frac{\partial}{\partial a_k^{\pi}} \sum_{l=1} \pi_k = \sum_{l=1} \pi_k(I_{kl} - \pi_l) \tag{478}$$

$$= \pi_k - \pi_k \sum_{l=1} \pi_l \tag{479}$$

$$\frac{\partial E_n}{\partial a_{kl}^{\pi}} = -\frac{\pi_k \mathcal{N}(\boldsymbol{t}_n|\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}), \sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w})) - \pi_k \sum_{l=1} \pi_l \mathcal{N}(\boldsymbol{t}_n|\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}), \sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w}))}{\sum_{l=1} \pi_l(\boldsymbol{x}_n, \boldsymbol{w})\mathcal{N}(\boldsymbol{t}_n|\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}), \sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w}))} \tag{480}$$

$$= -\gamma_k(w) + \pi_k \tag{481}$$

5.35

$$\frac{\partial E_n}{\partial a_{kl}^{\mu}} = -\frac{1}{\sum_{l=1} \pi_l(\boldsymbol{x}_n, \boldsymbol{w})\mathcal{N}(\boldsymbol{t}_n|\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}), \sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w}))} \frac{\partial}{\partial a_k^{\mu}} \sum_{l=1} \pi_l(\boldsymbol{x}_n, \boldsymbol{w})\mathcal{N}(\boldsymbol{t}_n|\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}), \sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w})) \tag{482}$$

$$\frac{\partial}{\partial a_k^{\mu}} \sum_{l=1} \pi_l(\boldsymbol{x}_n, \boldsymbol{w})\mathcal{N}(\boldsymbol{t}_n|\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}), \sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w})) \tag{483}$$

$$= \pi_k(\boldsymbol{x}_n, \boldsymbol{w})\mathcal{N}(\boldsymbol{t}_n|\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}), \sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w})) \frac{\partial}{\partial a_k^{\mu}} - \frac{1}{2}(\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}) - \boldsymbol{t}))^T \sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w})^{-1}(\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}) - \boldsymbol{t})) \tag{484}$$

$$\tag{485}$$

$$\frac{\partial}{\partial a_k^\mu} - \frac{1}{2}(\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}) - \boldsymbol{t}))^T \sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w})^{-1}(\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}) - \boldsymbol{t})) \tag{486}$$

$$= -\sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w})^{-1}(\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}) - \boldsymbol{t}) \underbrace{\frac{\partial \boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w})}{\partial a_k^\mu}}_{=1} \tag{487}$$

$$\frac{\partial E_n}{\partial a_{kl}^\mu} = \gamma_{nk}\sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w})^{-1}(\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}) - \boldsymbol{t}) \tag{488}$$

5.36

$$\frac{\partial E_n}{\partial a_{kl}^\sigma} = -\frac{1}{\sum_{l=1} \pi_l(\boldsymbol{x}_n, \boldsymbol{w})\mathcal{N}(\boldsymbol{t}_n|\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}), \sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w}))} \frac{\partial}{\partial a_k^\sigma} \sum_{l=1} \pi_l(\boldsymbol{x}_n, \boldsymbol{w})\mathcal{N}(\boldsymbol{t}_n|\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}), \sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w})) \tag{489}$$

$$\frac{\partial}{\partial a_k^\sigma} \sum_{l=1} \pi_l(\boldsymbol{x}_n, \boldsymbol{w})\mathcal{N}(\boldsymbol{t}_n|\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}), \sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w})) = \pi_k(\boldsymbol{x}_n, \boldsymbol{w}) \frac{\partial}{\partial a_k^\sigma} \mathcal{N}(\boldsymbol{t}_n|\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}) \tag{490}$$

$$\frac{\partial}{\partial a_k^\sigma}\mathcal{N}(\boldsymbol{t}_n|\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}) = \frac{\partial}{\partial a_k^\sigma} \frac{1}{(2\pi)^{\frac{L}{2}}} \frac{1}{|\sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w})|^{\frac{1}{2}}} \exp(-\frac{1}{2}(\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}) - \boldsymbol{t}))^T \sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w})^{-1}(\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}) - \boldsymbol{t})) \tag{491}$$

$$= \frac{1}{(2\pi)^{\frac{L}{2}}} \Bigg(\exp(-\frac{1}{2}(\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}) - \boldsymbol{t}))^T \sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w})^{-1}(\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}) - \boldsymbol{t})) \frac{\partial}{\partial a_k^\sigma} \frac{1}{|\sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w})|^{\frac{1}{2}}} \tag{492}$$

$$+ \frac{1}{|\sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w})|^{\frac{1}{2}}} \frac{\partial}{\partial a_k^\sigma} \exp\Big(-\frac{1}{2}(\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}) - \boldsymbol{t}))^T \sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w})^{-1}(\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}) - \boldsymbol{t})\Big)\Bigg) \tag{493}$$

$$\frac{\partial}{\partial a_k^\sigma} \frac{1}{|\sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w})|^{\frac{1}{2}}} = \frac{\partial}{\partial a_k^\sigma} \frac{1}{\sigma_k(\boldsymbol{x}, \boldsymbol{w})^L} = -\frac{L}{\sigma^2(\boldsymbol{x}, \boldsymbol{w})^{L-1}} = \frac{L}{\sigma_k} \frac{1}{\sigma_k(\boldsymbol{x}, \boldsymbol{w})^L} = \frac{L}{\sigma_k} \frac{1}{|\sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w})|^{\frac{1}{2}}} \tag{494}$$

$$\frac{\partial}{\partial a_k^\sigma}(-\frac{1}{2}(\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}) - \boldsymbol{t}))^T \sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w})^{-1}(\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}) - \boldsymbol{t}) = \frac{\|\boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}) - \boldsymbol{t}\|^2}{\sigma_k^3} \tag{495}$$

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

5.37

5.38

5.39

Using the approximation:

$$p(D|\alpha,\beta) = \int p(D|w,\beta)p(w|\alpha)dw \approx p(D,w_{map})p(w_{map})\frac{(2\pi)^{W/2}}{|A|^{\frac{1}{2}}} \tag{496}$$

Taking the log gives the result.

5.40

Use a softmax activation.

6.3

$$\|x - x_n\| = x^T x + x_n^T x - 2x_n^T x \tag{497}$$
$$= k(x,x) + k(x_n,x) - 2k(x_n,x) \tag{498}$$

6.4

$$\begin{pmatrix} 2 & 0 \\ -1 & 3 \end{pmatrix} \tag{499}$$

6.5

$$ck(\boldsymbol{x},\boldsymbol{x}') = c\psi^T\psi = (c^{\frac{1}{2}}\psi)^T(c^{\frac{1}{2}}\psi) = \phi^T\phi \tag{500}$$
$$f(\boldsymbol{x})k(\boldsymbol{x},\boldsymbol{x}')f(\boldsymbol{x}') = f(\boldsymbol{x})\psi^T\psi f(\boldsymbol{x}') = (f(\boldsymbol{x})\psi^T)^T(\psi f(\boldsymbol{x}')) = \phi^T\phi \tag{501}$$

6.6

$$q(k(x,x')) = \sum_{n=1}^{N} \beta_n k(x,x')^n \tag{502}$$
$$= \sum_{n=1}^{N} \beta_0 k(x,x') \qquad 6.18 \tag{503}$$
$$= \sum_{n=1}^{N} k(x,x') \qquad 6.13 \tag{504}$$
$$= k(x,x') \qquad 6.17 \tag{505}$$

6.16 follows from the previuos proof and the definition of the exponential function.

6.7

$$\psi(x)^T\psi(x') + \theta(x)^T\theta(x') = \sum_{n=1}^{N}(\psi(x_n) + \theta(x_n))(\psi(x'_n) + \theta(x'_n)) - \psi(x_n)\theta(x'_n) - \theta(x_n)\psi(x'_n) \quad (506)$$

$$= \begin{bmatrix}\dots & \psi(x_n) + \theta(x_n) & -\psi(x_n) & -\theta(x_n) & \dots\end{bmatrix} \begin{bmatrix}\vdots \\ \psi(x'_n) + \theta(x'_n) \\ \theta(x'_n) \\ \psi(x'_n) \\ \vdots\end{bmatrix} \quad (507)$$

$$= \phi(x)^T\phi(x') \quad (508)$$

$$k_1(x, x')k_2(x, x') = \theta(x)^T\theta(x')(\psi(x)^T\psi(x') \quad (509)$$

$$= \sum_{n=1}^{N}\theta(x_n)\theta(x'_n)\psi(x_n)\psi(x'_n) \quad (510)$$

$$= \sum_{n=1}^{N}\theta(x_n)\psi(x_n)\theta(x'_n)\psi(x'_n) \quad (511)$$

$$= \phi(x)^T\phi(x') \quad (512)$$

$$= k(x, x') \quad (513)$$

6.8

$$k_3(\phi(x), \phi(x')) = \psi(\phi(x))^T\psi(\phi(x')) = k(x, x') \quad (514)$$

$$x^TAx' = x^TU^TUx' \quad \text{(Symmetry)}$$

$$= (Ux)^TUx' \quad (515)$$

$$= \phi(x)^T\phi(x') \quad (516)$$

$$= k(x, x') \quad (517)$$

6.9

$$\sum_{n=1}^{N} \psi(x_{na})\psi(x'_{na}) + \sum_{m=1}^{M} \theta(x_{mb})\theta(x'_{mb}) \tag{518}$$

$$= \begin{bmatrix} \psi(x_{a1}) & \dots & \psi(x_{aN}) & \theta(x_{b1}) & \dots & \theta(x_{bM}) \end{bmatrix} \begin{bmatrix} \psi(x'_{a1}) \\ \vdots \\ \psi(x'_{aN}) \\ \theta(x'_{b1}) \\ \vdots \\ \theta(x'_{bM}) \end{bmatrix} = k(x, x') \tag{519}$$

$$x \in \mathbb{R}^n \tag{520}$$

g bijective, $f : \mathbb{R}^n \to \mathbb{R}$

$$y = f(g(x)) \tag{521}$$

$$y'(x) = \nabla_{g(x)} f(g(x)) J_x(g(x)) \tag{522}$$

$$k_a(x_a, x'_a)k(x_b, x'_b) = \sum_{n=1}^{N} \theta(x_{an})\theta(x'_{an}) \sum_{m=1}^{M} \psi(x_{bm})\psi(x'_{bm}) \tag{523}$$

$$= \sum_{n=1}^{N} \sum_{m=1}^{M} \theta(x_{an})\theta(x'_{an})\psi(x_{bm})\psi(x'_{bm}) \tag{524}$$

$$= \sum_{n=1}^{N} \sum_{m=1}^{M} \theta(x_{an})\psi(x_{bm})\theta(x'_{an})\psi(x'_{bm}) \tag{525}$$

$$= \phi(x)^T \phi(x') = k(x, x') \tag{526}$$

6.10

$$y(x) = \sum_{n=1}^{N} f(x_n) f(x_n)(K + \lambda I_N)^{-1} t \tag{527}$$

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 $\boldsymbol{\theta}$.

$$g(\boldsymbol{\theta}, \boldsymbol{x}) = J_\theta(\phi) \nabla_\phi \ln p(\boldsymbol{x}, \boldsymbol{\phi}) = J_\theta(\boldsymbol{\phi}) g(\boldsymbol{\phi}, \boldsymbol{x}) \tag{528}$$

$$g(\boldsymbol{\phi}, \boldsymbol{x}) = J_\theta^{-1}(\boldsymbol{\phi}) g(\boldsymbol{\theta}, \boldsymbol{x}) \tag{529}$$

$$\boldsymbol{F}' = J_\theta^{-1}(\boldsymbol{\phi})\, \mathbb{E}_x[g(\boldsymbol{\theta}, \boldsymbol{x}) g(\boldsymbol{\theta}, \boldsymbol{x})^T] J_\theta^{-T}(\boldsymbol{\phi}) = J_\theta^{-1}(\boldsymbol{\phi}) \boldsymbol{F} J_\theta^{-T}(\boldsymbol{\phi}) \tag{530}$$

$$\begin{aligned}
k'(\boldsymbol{x}, \boldsymbol{x}') &= g(\boldsymbol{\phi}, \boldsymbol{x})^T \boldsymbol{F}'^{-1} g(\boldsymbol{\phi}, \boldsymbol{x}) && (531)\\
&= (J_\theta^{-1}(\boldsymbol{\phi}) g(\boldsymbol{\theta}, \boldsymbol{x}))^T (J_\theta^{-1}(\boldsymbol{\phi}) \boldsymbol{F} J_\theta^{-T}(\boldsymbol{\phi}))^{-1} J_\theta^{-1}(\boldsymbol{\phi}) g(\boldsymbol{\theta}, \boldsymbol{x}) && (532)\\
&= g(\boldsymbol{\theta}, \boldsymbol{x})^T J_\theta^{-T}(\boldsymbol{\phi}) J_\theta^T(\boldsymbol{\phi}) \boldsymbol{F} J_\theta(\boldsymbol{\phi}) J_\theta^{-1}(\boldsymbol{\phi}) g(\boldsymbol{\theta}, \boldsymbol{x}) && (533)\\
&= g(\boldsymbol{\theta}, \boldsymbol{x})^T \boldsymbol{F} g(\boldsymbol{\theta}, \boldsymbol{x}) && (534)\\
&= k(\boldsymbol{x}, \boldsymbol{x}') && (535)
\end{aligned}$$

Therefore, the Fisher kernel is invariant.

6.14

$$\begin{aligned}
\nabla_\theta(\ln p(\boldsymbol{x}, \boldsymbol{\theta})) &= \nabla_\mu \ln p(\boldsymbol{x}, \boldsymbol{\mu}) && (536)\\
&= S^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) && (537)
\end{aligned}$$

$$\begin{aligned}
F &= \mathbb{E}[S^{-1}(\boldsymbol{x} - \boldsymbol{\mu})(S^{-1}(\boldsymbol{x} - \boldsymbol{\mu}))^T] && (538)\\
&= S^{-1}\, \mathbb{E}_x[(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^T] S^{-1} && (539)\\
&= S^{-1} S S^{-1} = S^{-1} && (540)
\end{aligned}$$

$$\begin{aligned}
k(\boldsymbol{x}, \boldsymbol{x}') &= (S^{-1}(\boldsymbol{x} - \mu))^T F^{-1} (S^{-1}(\boldsymbol{x} - \boldsymbol{\mu})) && (541)\\
&= (\boldsymbol{x} - \boldsymbol{\mu})^T S^{-1} S S^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) && (542)\\
&= (\boldsymbol{x} - \boldsymbol{\mu})^T S^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) && (543)
\end{aligned}$$

6.15

Since the gram matrix is positive semi-definite we have:

$$|K| = k(x_1, x_1) k(x_2, x_2) - k(x_2, x_1) k(x_1, x_2) \geq 0 \tag{544}$$

from which the result follows.

6.18

$$\boldsymbol{z} = (x, t) \tag{545}$$

$$p(\boldsymbol{z}) = \frac{\sum_n \mathcal{N}(\boldsymbol{z} - \boldsymbol{z}_n, \sigma^2 I)}{\sum_m \int \mathcal{N}(\boldsymbol{z} - \boldsymbol{z}_n, \sigma^2 I) dt} = \frac{\sum_n \mathcal{N}(\boldsymbol{z} - \boldsymbol{z}_n, \sigma^2 I)}{\sum_m \mathcal{N}(x - x_n, \sigma^2 I)} \tag{546}$$

$$\frac{\sum_n \mathcal{N}(\boldsymbol{z} - \boldsymbol{z}_n, \sigma^2 I)}{\sum_m \mathcal{N}(x - x_n, \sigma^2 I)} = \frac{\sum_n \frac{1}{2\pi\sigma^2} \exp[-\frac{1}{2\sigma^2}(\boldsymbol{z} - \boldsymbol{z}_n)^T(\boldsymbol{z} - \boldsymbol{z}_n)]}{\sum_m \frac{1}{\sqrt{2\pi}\sigma} \exp[-\frac{1}{2\sigma^2}(x - x_n)^2]} \tag{547}$$

$$= \frac{\sum_n \frac{1}{2\pi\sigma^2} \exp[-\frac{1}{2\sigma^2}(x - x_n)^2 + -\frac{1}{2\sigma^2}(t - t_n)^2]}{\sum_m \frac{1}{\sqrt{2\pi}\sigma} \exp[-\frac{1}{2\sigma^2}(x - x_n)^2]} \tag{548}$$

$$= \sum_n \frac{\frac{1}{\sqrt{2\pi\sigma^2}} \exp[-\frac{1}{2\sigma^2}(x - x_n)^2}{\sum_m \frac{1}{\sqrt{2\pi}\sigma} \exp[-\frac{1}{2\sigma^2}(x - x_n)^2]} \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-\frac{1}{2\sigma^2}(t - t_n)^2] \tag{549}$$

$$= \sum_n \pi_n \mathcal{N}(t | t_n, \sigma^2) \tag{550}$$

6.24

$$u^T W u = u^T \sqrt{W}\sqrt{W} u \tag{551}$$

$$= (\sqrt{W}u)^T \sqrt{W} u \tag{552}$$

$$> 0 \quad \forall u \in \mathbb{R} \setminus \mathbf{0} \tag{553}$$

$$u^T(W + V)u = u^T W u + u^T V u > 0 \tag{554}$$

6.25

$$a_N = a_N - \nabla\nabla\Psi(a_N)\nabla\Psi(a_n) \tag{555}$$

$$= a_N + (W_N + C^{-1})^{-1}[t_N - \sigma_N - C^{-1}a_N] \tag{556}$$

$$= (W_N + C_N^{-1})[W_N a_N - \sigma_N + t_N] \tag{557}$$

$$= C_N((W_N + C_N^{-1})C_N)^{-1}[W_N a_N - \sigma_N + t_N] \tag{558}$$

$$= C_N(C_N W_N + I)^{-1}[W_N a_N - \sigma_N + t_N] \tag{559}$$

6.26

7.2

$$t_n(\boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}) + b) = \gamma \tag{560}$$

$$L(\boldsymbol{w}, b, \boldsymbol{a}) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{n=1}^{N} a_n(t_n(\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n) + b) - \gamma) \tag{561}$$

$$\nabla_{\boldsymbol{w}} = \boldsymbol{w} - \sum_{n=1}^{N} \nabla a_n t_n \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}_n) \tag{562}$$

$$= \boldsymbol{w} - \sum_{n=1}^{N} a_n t_n \boldsymbol{\phi}(\boldsymbol{x}_n) = 0 \tag{563}$$

$$\iff \boldsymbol{w} = \sum_{n=1}^{N} a_n t_n \boldsymbol{\phi}(\boldsymbol{x}_n) \tag{564}$$

$$\frac{\partial}{\partial b} = -\sum_{n=1}^{N} a_n t_n = 0 \tag{565}$$

$$= \sum_{n=1} a_n t_n \tag{566}$$

$$\tilde{L} = \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n t_n a_m t_m \phi(\boldsymbol{x}_n)^T \boldsymbol{\phi}(\boldsymbol{x}_n) - \sum_{n=1}^{N} a_n(t_n(\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n) + b) - \gamma) \tag{567}$$

$$\sum_{n=1}^{N} a_n(t_n(\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n) + b) - \gamma) = \sum_{n=1}^{N} 0 - a_n\gamma \tag{568}$$

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

$$-\ln p(t_n) = -\sum_{n=1} \ln \sigma(t_n y_n) \tag{569}$$

which is the cross-entropy error function.

7.7

It is quite straightforward if you use the following steps.

$$\sum_{n=1}^{N} \xi_n(C - \mu_n - a_n) + \sum_{n=1}^{N} \hat{\xi}_n(C - \hat{\mu}_n - \hat{a}_n) = 0 \tag{570}$$

and

$$-\sum_{n=1}^{N} a_n(y_n) - \sum_{n=1}^{N} \hat{a}_n(-y_n) = -\sum_{n=1}^{N}(a_n - \hat{a_n})y_n = 0 \tag{571}$$

7.8

This follows from 7.67 and 7.68.

7.9

$$p(\boldsymbol{w}, \boldsymbol{t}) = \mathcal{N}(\boldsymbol{w}|_N, \boldsymbol{S}_N) \tag{572}$$

$$_N = \boldsymbol{S}_N(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + \beta\boldsymbol{\Phi}^T\boldsymbol{t}) \tag{573}$$

$$\boldsymbol{S}_N^{-1} = \boldsymbol{S}_0^{-1} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi} \tag{574}$$

We have

$$\boldsymbol{S}_N^{-1} = \boldsymbol{S}_0^{-1} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi} = A + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi} = \Sigma^{-1} \tag{575}$$

Considering $\boldsymbol{m}_0 = \boldsymbol{0}$, $\boldsymbol{m}_N$ follows directly.

7.16

7.18

$$\nabla \ln p(\boldsymbol{w}|\boldsymbol{t}, \alpha) = \nabla \sum_{n=1}^{N} t_n \ln y_n + (1 - t_n)\ln(1 - y_n) - \boldsymbol{A}\boldsymbol{w} \tag{576}$$

$$\nabla \sum_{n=1}^{N} t_n \ln y_n + (1 - t_n)\ln(1 - y_n) = \sum_{n=1}^{N} t_n \nabla \ln y_n + (1 - t_n)\nabla(1 - y_n) \tag{577}$$

$$\nabla \ln y_n = (1 - y_n)\phi(\boldsymbol{x}_n) \tag{578}$$

$$\nabla \ln(1 - y_n) = -y_n\phi(\boldsymbol{x}_n) \tag{579}$$

$$\nabla \ln p(\boldsymbol{w}|\boldsymbol{t}, \alpha) = \sum_{n=1}^{N} \phi(\boldsymbol{x}_n)[t_n - y_n t_n - y_n + y_n t_n] - \boldsymbol{A}\boldsymbol{w} \tag{580}$$

$$= \boldsymbol{\Phi}^T[\boldsymbol{t} - \boldsymbol{y}] - \boldsymbol{A}\boldsymbol{w} \tag{581}$$

$$\nabla\nabla \ln p(\boldsymbol{w}|\boldsymbol{t}, \alpha) = \nabla\boldsymbol{\Phi}^T[\boldsymbol{t} - \boldsymbol{y}] - \boldsymbol{A} \tag{582}$$

$$= \nabla \sum_{n=1}^{N} [t_n - y_n]\boldsymbol{\phi}_n - \boldsymbol{A} \tag{583}$$

$$= -\sum_{n=1}^{N} \boldsymbol{\phi}_n \nabla^T y_n - \boldsymbol{A} \tag{584}$$

$$= -\sum_{n=1}^{N} \boldsymbol{\phi}_n y_n(1 - y_n)\boldsymbol{\phi}_n^T - \boldsymbol{A} \tag{585}$$

$$= -\boldsymbol{\Phi}^T\boldsymbol{B}\boldsymbol{\Phi} - \boldsymbol{A} \tag{586}$$

Which gives the result.

8.1

$$\int p(\boldsymbol{x})d\boldsymbol{x} = \int\int p(x_K|pa_K) \prod_{k=1}^{K-1} p(x_k|pa_k)dx_K dx_1 \dots dx_{K-1} \tag{587}$$

$$= \int\int p(x_K|pa_K)dx_K \prod_{k=1}^{K-1} p(x_k|pa_k)dx_1 \dots dx_{K-1} \tag{588}$$

$$= \int \prod_{k=1}^{K-1} p(x_k|pa_k)dx_1 \dots dx_{K-1} \tag{589}$$

$$\vdots \tag{590}$$

$$= \int p(x_1)dx_1 = 1 \tag{591}$$

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.

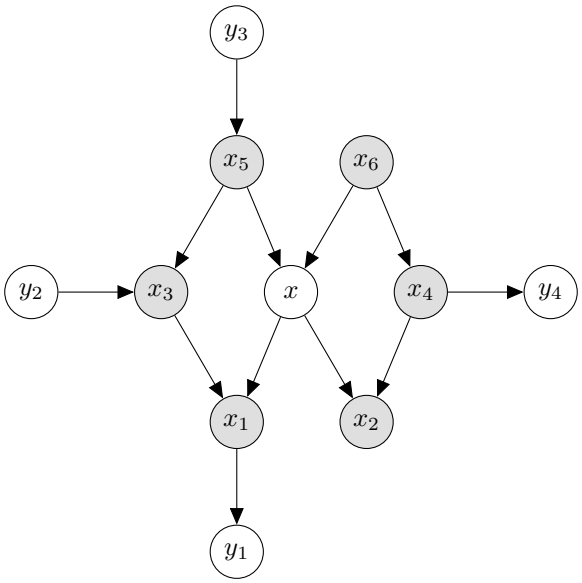
8.8

$$p(a, b, c|d) = \frac{p(a, b, c)}{p(d)} = \frac{p(a|b, c)p(b)p(c)}{p(d)} = \frac{p(a)p(b)p(c)}{p(d)} \tag{592}$$

$$\int \frac{p(a)p(b)p(c)}{p(d)}dc = p(a|d)p(b|d) \tag{593}$$

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

$$p(a, b, c, d) = p(a)p(b)p(c|a, b)p(d|c) \tag{594}$$

$$p(a, b) = \int \int p(a, b, c, d) dc dd = p(a)p(b) \tag{595}$$

Second part:

$$p(a, b, c|d) = \frac{p(a, b, c, d)}{p(d)} = \frac{p(a)p(b)p(c|a, b)p(d|c)}{p(d)} \tag{596}$$

$$p(a, b|d) = \int p(a, b, c|d) dc = \frac{p(a)p(b)p(d|c)}{p(d)} \tag{597}$$

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

$$E(x,y)_{x_k=1} - E(x,y)_{x_k=-1} = h\sum_i x_i - \beta\sum_{i,j} x_i x_j - \eta\sum_i x_i y_i + h - \beta_x k\sum_j x_j - \eta x_k \sum_i y_i \tag{598}$$

$$- h\sum_i x_i - \beta\sum_{i,j} x_i x_j - \eta\sum_i x_i y_i + h - \beta x_k \sum_j x_j - \eta x_k \sum_i y_i \tag{599}$$

$$= 2h - 2\beta x_k \sum_j x_i - \eta x_k \sum_i y_i \tag{600}$$

8.14

$$p(x,y) = \frac{1}{Z}\exp[-E(x,y)] \tag{601}$$

$$\ln p(x,y) = -\ln(Z) - E(x,y) = -\ln(Z) + \eta\sum_i x_i y_i \tag{602}$$

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

$$p(\boldsymbol{x},\boldsymbol{z}) = \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \Sigma_k)^{\boldsymbol{z}_k} \pi_k^{\boldsymbol{z}_k} \tag{603}$$

$$p(\boldsymbol{z}) = \sum_z \prod_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{\boldsymbol{z}_k} \tag{604}$$

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:

$$\boldsymbol{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}_{old})_{MAP} = \sum_z p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta}_{old}) \ln p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{x}) \tag{605}$$

$$\propto \sum_z p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta}_{old}) \ln p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{x}) \tag{606}$$

$$= \sum_z p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta}_{old}) \ln p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta}) + \ln p(\theta) \sum_z p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta}_{old}) \tag{607}$$

$$= \sum_z p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta}_{old}) \ln p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta}) + \ln p(\theta) \tag{608}$$

9.5

This is obvious, as there simply is no connection between z_m and x_n

9.7

$$\nabla_\mu \ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \pi) = \sum_{n=1}^{N} \sum_{k=1}^{k} \boldsymbol{z}_{kn} \nabla_\mu \ln \mathcal{N}(\boldsymbol{x}_n, \mu_k, \Sigma_k) \tag{609}$$

$$\nabla \ln \mathcal{N}(\boldsymbol{x}_n|\mu_k, \Sigma_k) = -\frac{1}{2} \nabla (\boldsymbol{x}_n - \mu_k)^T \sigma_k^{-1} (\boldsymbol{x}_n - \mu_k) = \Sigma_k^{-1} (\boldsymbol{x}_n - \mu_k) \tag{610}$$

Since z is 1-of-k this gradient only concerns distribution k for every n.

$$\mathcal{L}(\pi_k, \lambda) \ln p(\boldsymbol{X}, \boldsymbol{Z}|\mu_k \Sigma_k \pi_k) + \lambda[-1 + \sum_k \pi_k] \tag{611}$$

$$\frac{\partial \mathcal{L}}{\partial \pi_k} = \sum_{n=1}^{N} z_{nk} \frac{1}{\pi_k} - \lambda \tag{612}$$

$$= \sum_{n=1}^{N} z_{nk} - \pi_k \lambda \tag{613}$$

$$= \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} - \pi_k \lambda \tag{614}$$

$$= N - \lambda = 0 \iff \lambda = N \tag{615}$$

Substituting back:

$$\sum_{n=1}^{N} z_{nk} \frac{1}{\pi_k} = N \iff \pi_k = \frac{1}{N} \sum_{n=1} z_{nk} = \frac{N_k}{N} \tag{616}$$

$$\nabla_{\mu_k} \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk})[\ln \pi_k + \ln \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)] \tag{617}$$

$$= \sum_{n=1}^{N} \gamma(z_n k) \nabla \ln \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{618}$$

$$= \sum_{n=1}^{N} \gamma(z_n k) \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_k) = 0 \tag{619}$$

$$\Rightarrow \sum_{n=1}^{N} \gamma(z_n k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k) \tag{620}$$

$$= \sum_{n=1}^{N} \gamma(z_n k) \boldsymbol{x}_n - \underbrace{\sum_{n=1}^{N} \gamma(\boldsymbol{z}_{nk})}_{N_k} \boldsymbol{\mu}_k = 0 \tag{621}$$

$$\tag{622}$$

Which leads to the correct answer.

9.9

$$\frac{\partial}{\partial \boldsymbol{\Sigma}_k} = \sum_{n=1}^{N} \gamma(z_{nk}[\frac{\partial}{\partial \boldsymbol{\Sigma}_k} \frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_k) - \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \ln |\boldsymbol{\Sigma}|^{-\frac{1}{2}}] \tag{623}$$

$$= \sum_{n=1}^{N} \gamma(z_{nk}[\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T - \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \ln |\boldsymbol{\Sigma}^{-1}|^{\frac{1}{2}}] \tag{624}$$

$$= \sum_{n=1}^{N} \gamma(z_{nk}[\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T - \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \frac{1}{2} \ln |\boldsymbol{\Sigma}^{-1}|] \tag{625}$$

$$= \sum_{n=1}^{N} \gamma(z_{nk}[\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T - \frac{1}{2}\boldsymbol{\Sigma}] = 0 \tag{626}$$

$$\iff \sum_{n=1}^{N} \gamma(z_{nk})[\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T = \sum_{n=1}^{N} \gamma(z_{nk}) \frac{1}{2} \boldsymbol{\Sigma}] = N_k \frac{1}{2} \boldsymbol{\Sigma}_k \tag{627}$$

From which the answer is easily seen.

9.11

$$\mathbb{E}_{\boldsymbol{z}}[\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})] = \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(\boldsymbol{z}_{nk})[\ln \pi_k + \ln \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)] \tag{628}$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(\boldsymbol{z}_{nk})[\ln \pi_k + \ln \frac{1}{(2\pi\epsilon)^{M/2}} - \frac{1}{2\epsilon}\|\boldsymbol{x}_n - \boldsymbol{\mu}_k\|^2] \tag{629}$$

$$\propto \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(\boldsymbol{z}_{nk})[\epsilon \ln \pi_k + \epsilon \ln \frac{1}{(2\pi\epsilon)^{M/2}} - \frac{1}{2}\|\boldsymbol{x}_n - \boldsymbol{\mu}_k\|^2] \tag{630}$$

$$= \lim_{\epsilon \to 0} -\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\|\boldsymbol{x}_n - \boldsymbol{\mu}_k\|^2 + C \tag{631}$$

9.12

$$\mathbb{E}[\boldsymbol{x}] = \int \boldsymbol{x} p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\pi}) d\boldsymbol{x} \tag{632}$$

$$= \int \boldsymbol{x} \sum_{k=1}^{K} \pi_k p(\boldsymbol{x}|\boldsymbol{\mu}_k) d\boldsymbol{x} \tag{633}$$

$$= \sum_{k=1}^{K} \pi_k \int \boldsymbol{x} p(\boldsymbol{x}|\boldsymbol{\mu}_k) d\boldsymbol{x} \tag{634}$$

$$= \sum_{k=1}^{K} \pi_k \boldsymbol{\mu}_k \tag{635}$$

$$\operatorname{cov}[\boldsymbol{x}] = \mathbb{E}[(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])^T] \tag{636}$$

$$= \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T] - \mathbb{E}[\boldsymbol{x}]\,\mathbb{E}[\boldsymbol{x}]^T \tag{637}$$

$$\tag{638}$$

$$\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T] = \int \boldsymbol{x}\boldsymbol{x}^T \sum_{k=1}^{K} \pi_k p(\boldsymbol{x}|\boldsymbol{\mu}_k) d\boldsymbol{x} \tag{639}$$

$$= \sum_{k=1}^{K} \pi_k \int \boldsymbol{x}\boldsymbol{x}^T p(\boldsymbol{x}|\boldsymbol{\mu}_k) d\boldsymbol{x} = \sum_{k=1}^{K} \pi_k\, \mathbb{E}_k[\boldsymbol{x}\boldsymbol{x}^T] \tag{640}$$

$$\tag{641}$$

$$\mathbb{E}_k[\boldsymbol{x}\boldsymbol{x}^T] = \mathbb{E}_k[(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}] + \mathbb{E}[\boldsymbol{x}])(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}] + \mathbb{E}[\boldsymbol{x}])^T] \tag{642}$$

$$= \mathbb{E}\left[(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])^T + (\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])\,\mathbb{E}[\boldsymbol{x}]^T + \mathbb{E}[\boldsymbol{x}][\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}]]^T + \mathbb{E}[\boldsymbol{x}]\,\mathbb{E}[\boldsymbol{x}]^T\right] \tag{643}$$

$$= \underbrace{\operatorname{cov}_k(\boldsymbol{x})}_{=\boldsymbol{\Sigma}_k} + \boldsymbol{\mu}\boldsymbol{\mu}^T \tag{644}$$

9.14

$$p(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\mu})p(\boldsymbol{z}|\boldsymbol{\pi}) = \prod_{k=1}^{K} p(\boldsymbol{x}|\boldsymbol{\mu}_k)^{z_k}\pi_k^{z_k} \tag{645}$$

$$= \sum_{\boldsymbol{z}} \prod_{k=1}^{K} p(\boldsymbol{x}|\boldsymbol{\mu}_k)^{\boldsymbol{z}_k}\pi_k^{\boldsymbol{z}_k} \tag{646}$$

$$\tag{647}$$

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

$$\frac{\partial}{\partial \mu_{ki}} = \sum_{n=1}^{N} \gamma(z_{nk})[\frac{\partial}{\partial \mu_{ki}} x_{ni} \ln \mu_{ki} + \frac{\partial}{\partial \mu_{ki}}(1 - x_{ni}) \ln(1 - \mu_{ki})] \tag{648}$$

$$= \sum_{n=1}^{N} \gamma(z_{nk}) \left[\frac{x_{ni}}{\mu_{ki}} - \frac{1 - x_{ni}}{1 - \mu_{ki}}\right] \tag{649}$$

$$= \sum_{n=1}^{N} \gamma(z_{nk}) \left[\frac{x_{ni}(1 - \mu_{ki}) - \mu_{ki}(1 - x_{ni})}{\mu_{ki}(1 - \mu_{ki})}\right] = 0 \tag{650}$$

$$\Rightarrow \sum_{n=1}^{N} \gamma(z_{nk})\mu_{ki} = \sum_{n=1}^{N} \gamma(z_{nk})x_n \tag{651}$$

$$= N_k\mu_{ki} = \sum_{n=1}^{N} \gamma(z_{nk})x_n \tag{652}$$

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

$$\frac{\partial}{\partial \alpha} = \frac{M}{2}\frac{1}{\alpha} - \frac{1}{2}\mathbb{E}[\boldsymbol{w}^T\boldsymbol{w}] = 0 \tag{653}$$

$$\Rightarrow \frac{M}{\alpha} = \mathbb{E}[\boldsymbol{w}^T\boldsymbol{w}] \tag{654}$$

$$\alpha = \frac{M}{\mathbb{E}[\boldsymbol{w}^T\boldsymbol{w}]} \tag{655}$$

9.24

$$\ln p(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_z q(\boldsymbol{z}) \ln p(\boldsymbol{x}|\boldsymbol{\theta}) \tag{656}$$

$$= \sum_z q(\boldsymbol{z}) \ln \frac{p(\boldsymbol{z}, \boldsymbol{x}|\boldsymbol{\theta})}{p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta})} \tag{657}$$

$$= \sum_z q(\boldsymbol{z}) \ln \frac{p(\boldsymbol{z}, \boldsymbol{x}|\boldsymbol{\theta})}{p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta})} \frac{q(\boldsymbol{z})}{q(\boldsymbol{z})} \tag{658}$$

$$= \sum_z q(\boldsymbol{z}) \ln \frac{p(\boldsymbol{z}, \boldsymbol{x}|\boldsymbol{\theta})}{q(\boldsymbol{z})} + \sum_z q(\boldsymbol{z}) \ln \frac{q(\boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta})} \tag{659}$$

$$= \sum_z q(\boldsymbol{z}) \ln \frac{p(\boldsymbol{z}, \boldsymbol{x}|\boldsymbol{\theta})}{q(\boldsymbol{z})} - \sum_z q(\boldsymbol{z}) \ln \frac{p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta})}{q(\boldsymbol{z})} \tag{660}$$

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.

$$\ln p(\boldsymbol{x}) = \int q(\boldsymbol{z}) \ln p(\boldsymbol{x}) d\boldsymbol{z} \tag{661}$$

$$= \int q(\boldsymbol{z}) \ln \frac{p(\boldsymbol{x}, \boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{x})} \tag{662}$$

$$= \int q(\boldsymbol{z}) \ln \frac{p(\boldsymbol{x}, \boldsymbol{z}) q(\boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{x}) q(\boldsymbol{z})} \tag{663}$$

$$= \int q(\boldsymbol{z}) \ln \frac{p(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z})} - \int q(\boldsymbol{z}) \ln \frac{q(\boldsymbol{z})}{p(\boldsymbol{x}|\boldsymbol{z})} \tag{664}$$

$$\tag{665}$$

10.2

This is easily seen by simply filing in the values. The result is a solution to the equations.

10.8

Filling in:

$$\mathbb{E}[\tau] = \frac{a}{b} = (a_0 + \frac{N}{2})(b_0 + \frac{1}{2}\,\mathbb{E}_\mu[\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu})^2 + \lambda_0(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^2]^{-1} \underset{N\to\infty}{\approx} b_N^{-1} \tag{666}$$

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

$$\ln p(\boldsymbol{x}) = \sum_m \sum_{\boldsymbol{z}} q(\boldsymbol{z}, m) \ln p(\boldsymbol{x}) \tag{667}$$

$$= \sum_m \sum_z q(\boldsymbol{z}, m) \ln \frac{p(\boldsymbol{x}, m, \boldsymbol{z})}{p(m, \boldsymbol{z}|\boldsymbol{x})} \frac{q(\boldsymbol{z}, m)}{q(\boldsymbol{z}, m)} \tag{668}$$

Which results in 10.35 after rearrangement.

10.15

We have:

$$\mathbb{E}[\pi_k] = \frac{\alpha_k}{\sum_{k=1}^{K} \alpha_k} \tag{669}$$

and $\alpha_k = \alpha_0 + N_k$.

$$\mathbb{E}[\pi_k] = \frac{\alpha_0 + N_k}{\sum_{k=1}^{K} \alpha_0 + N_k} \tag{670}$$

$$= K\alpha_0 + \underbrace{\sum_{k=1}^{K} N_k}_{=N} \tag{671}$$

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

$$\frac{dd\ln(x)}{dx} = -x^{-2} \tag{672}$$

which is negative everywhere so $\ln(x)$ is concave.

Taylor approximation:

$$y(x) = \ln(\xi) + \frac{1}{\xi}(x - \xi) = \lambda x - \ln(\lambda) - 1 \quad \lambda = \frac{1}{\xi} \tag{673}$$

$$g(\lambda) = \min_x [\lambda x - f(x)] \tag{674}$$

$$\frac{d}{dx} = \lambda - \frac{1}{x} = 0 \iff x = \frac{1}{\lambda} \tag{675}$$

Therefore $g(\lambda) = 1 - \ln \frac{1}{\lambda} = 1 + \ln \lambda$.

$$\frac{d}{d\lambda}\lambda x - g(\lambda) = x - \frac{1}{\lambda} \iff x = \frac{1}{\lambda} \tag{676}$$

Plugging this into $y(x)$ gives the result $\ln x$

10.30

$$\frac{d}{dx} = \frac{1}{1 + e^{-x}} e^{-x} = \sigma(x) e^{-x} \tag{677}$$

$$\frac{dd}{dx} = \sigma(x)(1 - \sigma(x))e^{-x} - e^{-x}\sigma(x) = -\sigma^2(x)e^{-x} \tag{678}$$

Both functions are positive everywhere, showing that the function itself is negative, therefore concave.

$$\frac{d}{dx} = \frac{1}{1 + e^{-x}} e^{-x} \tag{679}$$

$$\frac{dd}{dx} = \sigma(x)(1 - \sigma(x))e^{-x} - e^{-x}\sigma(x) = -\sigma^2(x)e^{-x} \tag{680}$$

Taylor:

$$f(x) = f(\xi) + f'(\xi)(x - \xi) + \mathcal{O}(\xi) \tag{681}$$

Since the approximation is linear and $f(x)$ is concave it must be that the LHS is smaller-equal to the RHS.

$$f(x) \leq -\ln(1 + e^{-xi}) + \sigma(\xi)e^{-\xi}(x - \xi) + \mathcal{O}(\xi^2) \tag{682}$$

$$= -\ln(1 + e^{-xi}) + \sigma(\xi)e^{-\xi}x - \sigma(\xi)e^{-\xi}\xi + \mathcal{O}(\xi^2) \tag{683}$$

$$= \lambda x - g(\lambda) \quad \lambda = \sigma(\xi)e^{-\xi} \tag{684}$$

10.33

$$\frac{d}{d\xi_n} = (1 - \sigma(\xi_n)) - \frac{1}{2} - \boldsymbol{\phi}_n^T \mathbb{E}[\boldsymbol{w}\boldsymbol{w}^T]\boldsymbol{\phi}_n \lambda'(\xi_n) + 2\xi_n\lambda(\xi_n) + \lambda'(\xi_n)\xi_n^2 \tag{685}$$

$$= -2\xi_n\lambda(\xi_n) - \boldsymbol{\phi}_n^T \mathbb{E}[\boldsymbol{w}\boldsymbol{w}^T]\boldsymbol{\phi}_n \lambda'(\xi_n) + 2\xi_n\lambda(\xi_n) + \lambda'(\xi_n)\xi_n^2 \tag{686}$$

$$= -\lambda'(\xi_n)(\xi_n^2 - \boldsymbol{\phi}_n^T \mathbb{E}[\boldsymbol{w}\boldsymbol{w}^T]\boldsymbol{\phi}_n) = 0 \tag{687}$$

$$\iff \xi_n^2 = \boldsymbol{\phi}_n^T \mathbb{E}[\boldsymbol{w}\boldsymbol{w}^T]\boldsymbol{\phi}_n \tag{688}$$

10.37

$$q^{\setminus j}(\boldsymbol{\theta}) = \frac{q(\boldsymbol{\theta})}{\tilde{f}_j(\boldsymbol{\theta}} \tag{689}$$

$$Z_j = \int q^{\setminus j}(\boldsymbol{\theta}) f_j(\boldsymbol{\theta}) d\theta) = \int q^{\setminus j}(\boldsymbol{\theta}) \tilde{f}_j(\boldsymbol{\theta}) d\theta) = \int (q(\boldsymbol{\theta}) d\theta = 1 \tag{690}$$

$$\tilde{f}_j(\boldsymbol{\theta}) = \frac{q^{new}}{q^{\setminus j}(\boldsymbol{\theta})} = \tilde{f}_j = f_j \tag{691}$$

11.1

$$\mathbb{E}[\hat{f} = \mathbb{E}[\frac{1}{L} \sum_{l=1}^{L} f(z^l) \tag{692}$$

$$= \frac{1}{L} \sum_{p} (\boldsymbol{z}) f(\boldsymbol{z}^l) d\boldsymbol{z} \tag{693}$$

$$= \frac{1}{L} \sum_{l=1}^{L} \mathbb{E}[f] = \mathbb{E}[f] \tag{694}$$

For the variance, we have to note that

$$\mathbb{E}[f(\boldsymbol{z}^l), f(\boldsymbol{z}^k)] = \mathrm{Var}(f(\boldsymbol{z})) + \mathbb{E}[f(\boldsymbol{z})]^2 \tag{695}$$

And $\mathrm{Var}(f(\boldsymbol{z})) = 0$ for $k \neq l$.

$$\mathrm{Var}[\hat{f}] = \mathbb{E}[\hat{f}^2] - \mathbb{E}[\hat{f}]^2 \tag{696}$$

$$= \frac{1}{L^2} (\sum_{l=1}^{L} \sum_{k=1}^{L} \mathbb{E}[f(\boldsymbol{z}^l) f(\boldsymbol{z}^k)]) - \mathbb{E}[f]^2 \tag{697}$$

$$= \frac{1}{L^2} (\sum_{l=1}^{L} \sum_{k=1}^{L} \mathbb{E}[f(\boldsymbol{z}^l), f(\boldsymbol{z}^k)]) - \mathbb{E}[f]^2 \tag{698}$$

$$= \frac{1}{L^2} L \mathrm{Var}(f(\boldsymbol{z})) + L^2 \, \mathbb{E}[f]^2 - \mathbb{E}[f]^2 \tag{699}$$

$$= \frac{1}{L} \mathrm{Var}(f(\boldsymbol{z})) \tag{700}$$

11.2

What we need to show is that we can transform z into any distribution if we use 11.6.

$$p(y) = 1 \cdot |\frac{dz}{dy}| = 1 \cdot p(y) \tag{701}$$

11.3

$$z = (\frac{1}{\pi}\int_{-\infty}^{y}\frac{1}{1+\hat{y}^2}) \tag{702}$$
$$= (\frac{1}{\pi}[\arctan(y) - \arctan(-\infty)]) \tag{703}$$
$$= (\frac{1}{\pi}\arctan(y) + \frac{1}{2}) \tag{704}$$

$$y = h^{-1}(z) \tag{705}$$
$$\iff \pi z - \frac{\pi}{2} = \arctan(y) \tag{706}$$
$$\iff h^{-1}(z) = \tan(\pi z - \frac{\pi}{2}) \tag{707}$$

11.5

Hint 1: Show that the expectation and covariance are equal to μ and Σ.

$$\mathbb{E}[y] = \mathbb{E}_z[\boldsymbol{\mu}] + \mathbb{E}_z[\boldsymbol{L}\boldsymbol{z}] \tag{708}$$
$$= \mathbb{E}_z[\boldsymbol{\mu}] + \boldsymbol{L}\,\mathbb{E}_z[\boldsymbol{z}] \tag{709}$$
$$= \mathbb{E}_z[\boldsymbol{\mu}] + \boldsymbol{L}\boldsymbol{0} \tag{710}$$
$$= \boldsymbol{\mu} \tag{711}$$

$$\text{cov}[\boldsymbol{y}] = \mathbb{E}[(\boldsymbol{y}-\boldsymbol{\mu})(\boldsymbol{y}-\boldsymbol{\mu})^T] \tag{712}$$
$$= \mathbb{E}[\boldsymbol{y}\boldsymbol{y}^T - \boldsymbol{\mu}\boldsymbol{y}^T - \boldsymbol{y}\boldsymbol{\mu}^T + \boldsymbol{\mu}\boldsymbol{\mu}^T] \tag{713}$$
$$= \mathbb{E}[(\boldsymbol{\mu}+\boldsymbol{L}\boldsymbol{z})(\boldsymbol{\mu}+\boldsymbol{L}\boldsymbol{z})^T - \boldsymbol{\mu}(\boldsymbol{\mu}+\boldsymbol{L}\boldsymbol{z})^T - (\boldsymbol{\mu}+\boldsymbol{L}\boldsymbol{z})\boldsymbol{\mu}^T + \boldsymbol{\mu}\boldsymbol{\mu}^T] \tag{714}$$
$$= \mathbb{E}[\boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\mu}(\boldsymbol{L}\boldsymbol{z})^T + \boldsymbol{L}\boldsymbol{z}\boldsymbol{\mu}^T + \boldsymbol{L}\boldsymbol{z}(\boldsymbol{L}\boldsymbol{z})^T - \boldsymbol{\mu}\boldsymbol{\mu}^T - \boldsymbol{\mu}(\boldsymbol{L}\boldsymbol{z})^T - \boldsymbol{\mu}\boldsymbol{\mu}^T - \boldsymbol{L}\boldsymbol{z}\boldsymbol{\mu}^T + \boldsymbol{\mu}\boldsymbol{\mu}^T] \tag{715}$$
$$= \mathbb{E}[\boldsymbol{L}\boldsymbol{z}(\boldsymbol{L}\boldsymbol{z})^T] \tag{716}$$
$$= \boldsymbol{L}\,\mathbb{E}[\boldsymbol{z}\boldsymbol{z}^T]\boldsymbol{L}^T \tag{717}$$

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

$$p(y) = p(z)|\frac{dz}{dy}| \tag{718}$$

We have $z \sim \mathcal{U}(0,1)$ and therefore $p(z) = \frac{1}{1-0} = 1$. Inverting the given equation:

$$y = b\tan z + c \iff \frac{y-c}{b} = \tan z \iff \arctan\frac{y-c}{b} = z \tag{719}$$

$$\frac{dz}{dy} = \frac{1}{1 + \left(\frac{y-c}{b}\right)^2} \frac{1}{b} \tag{720}$$

$$\tag{721}$$

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 $E[(z_i' - \mu_i)^2$.

$$\mathbb{E}_{z,\nu}[z_i'] = \mathbb{E}[\mu_i] + \mathbb{E}[\alpha(z_i - \mu_i) + \mathbb{E}[\sigma_i(1-\alpha_i^2)^{\frac{1}{2}}\nu] \tag{722}$$

$$= \mu_i \tag{723}$$

$$\begin{aligned}
\mathrm{Var}_{z,\nu} &= \mathbb{E}[z_i'^2] - \mathbb{E}[z_i']^2 && (724)\\
&= \mathbb{E}[(\mu_i + \alpha(z_i - \mu_i) + \sigma_i(1-\alpha_i^2)^{\frac{1}{2}}\nu)(\mu_i + \alpha(z_i - \mu_i) + \sigma_i(1-\alpha_i^2)^{\frac{1}{2}}\nu)] - \mu_i^2 && (725)\\
&= \mathbb{E}[(\mu_i + \alpha(z_i - \mu_i))(\mu_i + \alpha(z_i - \mu_i) + \sigma_i(1-\alpha_i^2)^{\frac{1}{2}}\nu)(\mu_i + \alpha(z_i - \mu_i) && (726)\\
&\quad + (\mu_i + \alpha(z_i - \mu_i)(\sigma_i(1-\alpha_i^2)^{\frac{1}{2}}\nu) + (\sigma_i(1-\alpha_i^2)^{\frac{1}{2}}\nu)(\sigma_i(1-\alpha_i^2)^{\frac{1}{2}}\nu)] - \mu_i^2 && (727)\\
&\vdots && (728)\\
& && (729)
\end{aligned}$$

11.15

$$\frac{dH}{dr_i} = \frac{1}{Z}\frac{d}{dr_i} r_i^2 = r_i \tag{730}$$

$$\frac{dr_i}{d\tau} = -\frac{dE(z)}{dz_i} = -\frac{dH}{dz_i} \tag{731}$$

11.16

$$p(\boldsymbol{r}|\boldsymbol{z}) = \frac{p(\boldsymbol{r},\boldsymbol{z})}{p(\boldsymbol{z})} = \frac{Z_p}{Z_H}\exp(-K(\boldsymbol{r})) \tag{732}$$

11.17

12.3

$$\|u_i\|^2 = [\frac{1}{(N\lambda_i)^{1/2}} X^T v_i]^T [\frac{1}{(N\lambda_i)^{1/2}} X^T v_i] \tag{733}$$
$$= (N\lambda_i)^{-1} [X^T v_i]^T [X^T v_i] \tag{734}$$
$$= (N\lambda_i)^{-1} v^T X X^T v_i \tag{735}$$
$$= \lambda_i^{-1} v^T \lambda_i v_i \tag{736}$$
$$= 1 \tag{737}$$

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

$$\frac{\partial \ln p(\boldsymbol{X}|\boldsymbol{\mu}, \boldsymbol{W}, \sigma^2)}{\partial \boldsymbol{\mu}} = -\frac{1}{2} \frac{\partial}{\partial \boldsymbol{\mu}} \sum_{n=1}^{N} (\boldsymbol{x}_n - \boldsymbol{\mu})^T \boldsymbol{C}^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}) \tag{738}$$
$$= \sum_{n=1}^{N} \boldsymbol{C}^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}) = 0 \tag{739}$$
$$\Rightarrow 0 = -N\boldsymbol{\mu} + \sum_{n=1}^{N} \boldsymbol{x}_n \tag{740}$$
$$\Rightarrow \boldsymbol{\mu} = \bar{\boldsymbol{x}} \tag{741}$$

12.14

$$D(D-1) + 1 - (D-1)(D-2)/2 = D^2 - D + 1 - \frac{1}{2}D^2 + \frac{3}{2}D - 1 = \frac{1}{2}D^2 + \frac{1}{2}D = D(D+1)/2 \tag{742}$$

12.18

$$M \times D + 1 \tag{743}$$

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

$$\frac{\partial Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}}{\boldsymbol{\mu}_k} = \sum_{n=1}^{N} \gamma(z_{nk}) \frac{1}{p(\boldsymbol{x}_n|\boldsymbol{\phi}_k)} \frac{\partial}{\partial \boldsymbol{\mu}^k} p(\boldsymbol{x}_n|\boldsymbol{\phi}_k) \tag{744}$$

$$= \sum_{n=1}^{N} \gamma(z_{nk}) \frac{1}{p(\boldsymbol{x}_n|\boldsymbol{\phi}_k)} p(\boldsymbol{x}_n|\boldsymbol{\phi}_k) \frac{\partial}{\partial \boldsymbol{\mu}_k} - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \tag{745}$$

$$= 0 \iff \sum_{n=1}^{N} \gamma(z_{nk}) \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_k) = 0 \tag{746}$$

$$\iff \boldsymbol{\mu} = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) \boldsymbol{x}_n}{\sum_{n=1}^{N} \gamma(z_{nk})} \tag{747}$$

$$\tag{748}$$

$$\frac{\partial Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})}{\boldsymbol{\Sigma}_k} = \sum_{n=1}^{N} \gamma(z_{nk}) \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \ln p(\boldsymbol{x}|\boldsymbol{\phi}_k) \tag{749}$$

$$\tag{750}$$

$$\frac{\partial}{\partial \boldsymbol{\Sigma}_k} \ln p(\boldsymbol{x}|\boldsymbol{\phi}_k) = \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \left[\ln \frac{1}{(2\pi)^{D/2}} + \ln \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} - \frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \right] \tag{751}$$

$$= \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \frac{1}{2} \ln |\boldsymbol{\Sigma}^{-1}| - \frac{1}{2} \frac{\partial}{\partial \boldsymbol{\Sigma}_k} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}) \tag{752}$$

$$= \frac{1}{2}\boldsymbol{\Sigma} - \frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T \tag{753}$$

$$\frac{\partial}{\partial \boldsymbol{\Sigma}_k} = \sum_{n=1}^{N} \gamma(z_{nk}) \left(\frac{1}{2}\boldsymbol{\Sigma} - \frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T \right) = 0 \tag{754}$$

$$\iff \boldsymbol{\Sigma}_k = \frac{\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T}{\sum_{n=1}^{N}\gamma(z_{nk})} \tag{755}$$

13.17

$$h(z_1) = p(z_1|u_1)p(x_1|z_1, u_1) \tag{756}$$

$$f_n(z_{n-1}, z_n) = p(z_n|z_{n-1}, u_n)p(x_n|z_n, u_n) \tag{757}$$

13.19

This follows from:

$$\frac{d}{dz}\ln p(\boldsymbol{Z}) = \frac{d}{dz}\ln p(z_0)p(z_1|z_0) \times \cdots \times p(z_n|z_{n-1}) = \begin{bmatrix} \frac{\partial}{z_0}\ln p(z_0) \\ \frac{\partial}{z_1}\ln p(z_1|z_0) \\ \vdots \\ \frac{\partial}{z_n} p(z_n|z_{n-1}) \end{bmatrix} \tag{758}$$

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

$$\mathbb{E}\left[\left(\frac{1}{M}\sum_{m=1}^{M}\epsilon_m(x)\right)^2\right] = \mathbb{E}\left[\left(\frac{1}{M^2}\sum_{m=1}^{M}\sum_{l=1}^{M}\epsilon_m(x)\epsilon_l(x)\right)\right] \tag{759}$$

$$= \mathbb{E}\left[\frac{1}{M^2}\left(\sum_{m=1}^{M}\sum_{\substack{l=1 \\ l\neq m}}^{M}\epsilon_m(x)\epsilon_l(x) + \sum_{l=1}^{M}\epsilon_l(x)^2\right)\right] \tag{760}$$

$$= \frac{1}{M^2}\sum_{l=1}^{M}\mathbb{E}[\epsilon_l(x)^2] \tag{761}$$

$$= \frac{1}{M}E_{AV} \tag{762}$$

14.3

Through Jensen's inequality:

$$\sum_{m=1}^{M} \frac{1}{M} \epsilon_m(x)^2 \geq \left(\sum_{m=1}^{M} \frac{1}{M} \epsilon_m(x) \right)^2 \tag{763}$$

Therefore:

$$\mathbb{E}_x[\sum_{m=1}^{M} \frac{1}{M} \epsilon_m(x)^2] \geq \mathbb{E} \left[\left(\sum_{m=1}^{M} \frac{1}{M} \epsilon_m(x) \right)^2 \right] = E_{com} \tag{764}$$

14.6

$$\frac{d}{d\alpha_m} = (e^{\alpha_m/2} + e^{-\alpha/2}) \sum_{n=1}^{N} w_n^{(m)} I(y_m(x_n) = t_n) = e^{\alpha_m/2} \sum_{n=1}^{N} w_n^{(m)} = 0 \tag{765}$$

$$\frac{e^{-\alpha_m/2}}{e^{\alpha_m/2} + e^{-\alpha_m/2}} = \epsilon_m \tag{766}$$

Rearranging this results in the desired formula.

14.7 14.8

14.9

$$\mathcal{L}_m(\boldsymbol{x}_n) = (y_n - \sum_{l=1}^{M} \alpha_l \hat{y}_l)^2 \tag{767}$$

$$= (\underbrace{y - \sum_{l=1}^{M-1} \alpha_l \hat{y}_l}_{\text{residual}} + \alpha_M y_{\hat{M}})^2 \tag{768}$$

$$\tag{769}$$

14.10

$$L = \frac{1}{2} \sum_{n=1}^{N} (t_n - t)^2 \tag{770}$$

$$\frac{dL}{dt} = -\sum_{n=1}^{N} (t_n - t) = 0 \iff Nt = \sum_{n=1}^{N} t_n \iff t = \frac{1}{N} \sum_{n=1}^{N} t_n \tag{771}$$

14.13

$$p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) = \prod_{n=1}^{N} \prod_{k=1}^{K} (\pi_k \mathcal{N}(t_n|\boldsymbol{w}_k^T \boldsymbol{\phi}_n, \beta^{-1})^{z_{nk}} \tag{772}$$

Taking the log of this will yield the result.

14.14

$$\frac{dL}{d\pi_k} = \sum_{n=1}^{N} \gamma_{nk}/\pi_k - \lambda = 0 \tag{773}$$

$$\iff \sum_k \sum_{n=1}^{N} \gamma_{nk} = \sum_k \pi_k \lambda \tag{774}$$

$$\tag{775}$$

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

$$\mathbb{E}[t|\hat{\boldsymbol{\phi}}, \boldsymbol{\theta}] = \sum_{k=1}^{K} \pi_k \, \mathbb{E}[t|\hat{\boldsymbol{\phi}}, \boldsymbol{w}_k, \beta] \tag{776}$$