DETAILED SOLUTIONS TO

Pattern Recognition and Machine Learning[*]

QUESTIONS

A Work of

Ziyue "Alan" Xiang

MSCS @ Syracuse University

`https://github.com/xziyue/PRML_Project`

[*] Bishop, Christopher M. *Pattern Recognition and Machine Learning.* Springer, 2006.

Contents

Chapter 1

1.1 Consider the sum-of-squares error function $E(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \boldsymbol{w}) - t_n\}^2$, in which the function $y(x_n, \boldsymbol{w})$ is given by $y(x_n, \boldsymbol{w}) = \sum_{j=0}^{M} w_j x^j$. Show that the coefficients $\boldsymbol{w} = \{w_i\}$ that minimize this error function are given by the solution to the following set of linear equations

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

where

$$A_{ij} = \sum_{n=1}^{N} (x_n)^{i+j}, \qquad T_i = \sum_{n=1}^{N} (x_n)^i t_n.$$

Here a suffix i or j denotes the index of a component, whereas $(x)^i$ denotes x raised to the power of i.

Answer . In order to find the minimum of E, we need to equate the partial derivative of E with respect to w_i to zero. Therefore we need to compute

$$\frac{\partial E}{\partial w_i} = \frac{\partial}{\partial w_i}\frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \boldsymbol{w}) - t_n\}^2. \tag{1.1}$$

Given a fixed n, it can be seen that

$$\frac{\partial}{\partial w_i}\frac{1}{2}\{y(x_n, \boldsymbol{w}) - t_n\}^2 \tag{1.2}$$

$$= \frac{\partial}{\partial w_i}\frac{1}{2}\left[\left(\sum_{j=0}^{M} w_j x_n^j\right) - t_n\right]^2 \tag{1.3}$$

$$= 2 \cdot \frac{1}{2} \cdot \left[\left(\sum_{j=0}^{M} w_j x_n^j\right) - t_n\right] \cdot \frac{\partial}{\partial w_i}\left[\left(\sum_{j=0}^{M} w_j x_n^j\right) - t_n\right] \tag{1.4}$$

$$= \left[\left(\sum_{j=0}^{M} w_j x_n^j\right) - t_n\right] \cdot x_n^i \tag{1.5}$$

$$= \sum_{j=0}^{M} w_j x_n^{i+j} - t_n x_n^i. \tag{1.6}$$

Hence we have

$$\frac{\partial E}{\partial w_i} = \frac{\partial}{\partial w_i}\frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\boldsymbol{w})-t_n\}^2 \tag{1.7}$$

$$= \frac{\partial}{\partial w_i}\sum_{n=1}^{N}\frac{1}{2}\{y(x_n,\boldsymbol{w})-t_n\}^2 \tag{1.8}$$

$$= \sum_{n=1}^{N}\left(\sum_{j=0}^{M} w_j x_n^{i+j} - t_n x_n^i\right) \tag{1.9}$$

$$= \sum_{j=0}^{M}\sum_{n=1}^{N} w_j x_n^{i+j} - \sum_{n=1}^{N} t_n x_n^i. \tag{1.10}$$

Equating it to zero yields

$$\sum_{j=0}^{M}\sum_{n=1}^{N} w_j x_n^{i+j} - \sum_{n=1}^{N} t_n x_n^i = 0 \tag{1.11}$$

$$\sum_{j=0}^{M}\sum_{n=1}^{N} w_j x_n^{i+j} = \sum_{n=1}^{N} t_n x_n^i \tag{1.12}$$

$$\sum_{j=0}^{M} w_j \cdot \sum_{n=1}^{N} x_n^{i+j} = \sum_{n=1}^{N} t_n x_n^i. \tag{1.13}$$

for $i = 0, 1, \ldots, M$.

By setting $A_{ij} = \sum_{n=1}^{N} x_n^{i+j}$ and $T_i = \sum_{n=1}^{N} t_n x_n^i$, it can be seen that the set of equations can be written as $\sum_{j=0}^{M} A_{ij} w_j = T_i$.

1.2 Write down the set of coupled linear equations, analogous to that in **1.1**, satisfied by the coefficients w_i which minimize the regularized sum-of-squares error function given by $\tilde{E}(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\boldsymbol{w})-t_n\}^2 + \frac{\lambda}{2}\|\boldsymbol{w}\|^2$.

Answer . *(using the notations in **1.1**)* Because $\tilde{E}(\boldsymbol{w}) = E(w) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2$, it is obvious that

$$\frac{\partial}{\partial w_i}\tilde{E}(\boldsymbol{w}) = \frac{\partial}{\partial w_i}E(\boldsymbol{w}) + \frac{\partial}{\partial w_i}\frac{\lambda}{2}\|\boldsymbol{w}\|^2 \tag{1.14}$$

$$= \sum_{j=0}^{M}\sum_{n=1}^{N} w_j x_n^{i+j} - \sum_{n=1}^{N} t_n x_n^i + \frac{\partial}{\partial w_i}\frac{\lambda}{2}\sum_{j=0}^{M} w_j^2 \tag{1.15}$$

$$= \sum_{j=0}^{M}\sum_{n=1}^{N} w_j x_n^{i+j} - \sum_{n=1}^{N} t_n x_n^i + \lambda w_i. \tag{1.16}$$

Similarly, the equations to be solved are

$$\sum_{j=0}^{M}\sum_{n=1}^{N} w_j x_n^{i+j} - \sum_{n=1}^{N} t_n x_n^i + \lambda w_i = 0 \tag{1.17}$$

$$\sum_{j=0}^{M}\sum_{n=1}^{N} w_j x_n^{i+j} + \lambda w_i = \sum_{n=1}^{N} t_n x_n^i. \tag{1.18}$$

for $i = 0, 1, \ldots, M$.

↑

Let I_{ij} be

$$I_{ij} = \begin{cases} 1, & i = j, \\ 0, & \text{otherwise.} \end{cases} \tag{1.19}$$

By setting $A_{ij} = \sum_{n=1}^{N} x_n^{i+j}$ and $T_i = \sum_{n=1}^{N} t_n x_n^i$, it can be seen that the set of equations can be written as $\sum_{j=0}^{M} \left(A_{ij} + I_{ij} \right) w_j = T_i$.

1.3 Suppose that we have three colored boxes r (red), b (blue) and g (green). Box r contains 3 apples, 4 oranges and 3 limes, box b contains 1 apple, 1 orange and 0 lime, and box g contains 3 apples, 3 oranges and 4 limes. If a box is chosen at random with probabilities $p(r) = 0.2$, $p(b) = 0.2$, $p(g) = 0.6$, and a piece of fruit is removed from the box (with equal probability of selecting any of the items in the box), then what is the probability of selecting an apple? If we observe that the second fruit is in fact an orange, what is the probability that it came from the green box?

Answer . Denote the event of selecting an apple, orange and lime by a, o and l, respectively; denote the set of all boxes by $B = \{r, b, g\}$. The probability of selecting an apple is $p(a)$, which by product rule, is equal to

$$\begin{aligned} p(a) &= \sum_{b \in B} p(a \mid b) p(b) && (1.20) \\ &= p(a \mid r)p(r) + p(a \mid b)p(b) + p(a \mid g)p(g) && (1.21) \\ &= \frac{3}{10} \times \frac{1}{5} + \frac{1}{2} \times \frac{1}{5} + \frac{3}{10} \times \frac{3}{5} && (1.22) \\ &= \frac{3+5+9}{50} = \frac{17}{50} = 0.34. && (1.23) \end{aligned}$$

Given the fruit selected is orange, the probability of it coming from the green box is given by $p(g \mid o)$. By *Bayes' Theorem*, it can be seen that

$$\begin{aligned} p(g \mid o) &= \frac{p(o \mid g)p(g)}{p(o)} && (1.24) \\ &= \frac{p(o \mid g)p(g)}{\sum_{b \in B} p(o \mid b)p(b)} && (1.25) \\ &= \frac{p(o \mid g)p(g)}{p(o \mid r)p(r) + p(o \mid b)p(b) + p(o \mid g)p(g)} && (1.26) \\ &= \frac{\frac{3}{10} \times \frac{3}{5}}{\frac{4}{10} \times \frac{1}{5} + \frac{1}{2} \times \frac{1}{5} + \frac{3}{10} \times \frac{3}{5}} && (1.27) \\ &= \frac{1}{2} = 0.5. && (1.28) \end{aligned}$$

1.4 Consider a probability $p_x(x)$ defined over a continuous variable x, and suppose that we make a nonlinear change of variable using $x = g(y)$, so that the density transforms according to $p_y(y) = p_x(g(y)) \, |g'(y)|$. By differentiating it, show that the location $\hat{y}$ of the maximum of the density in y is not in general related to the location $\hat{x}$ of the maximum of the

density over x by the simple functional relation $\hat{x} = g(\hat{y})$ as a consequence of the Jacobian factor. This shows that the maximum of a probability density (in contrast to a simple function) is dependent on the choice of variable. Verify that, in the case of a linear transformation, the location of the maximum transforms in the same way as the variable itself.

Answer . The location $\hat{x}$ of the maximum of the density over x should satisfy

$$p'_x(\hat{x}) = 0. \tag{1.29}$$

By differentiating $p_y(y) = p_x(g(y))\,|g'(y)|$, we have

$$\begin{aligned}
p'_y(y) &= \left[p_x(g(y))\,|g'(y)|\right]' && (1.30)\\
&= \left[p_x(g(y))\right]' \cdot |g'(y)| + p_x(g(y)) \cdot |g'(y)|' && (1.31)\\
&= p'_x(g(y)) \cdot g'(y) \cdot |g'(y)| + p_x(g(y)) \cdot \operatorname{sgn}\left(g'(y)\right) \cdot g''(y) && (1.32)\\
&= p'_x(g(y)) \cdot \operatorname{sgn}\left(g'(y)\right) \cdot \left[g'(y)\right]^2 + p_x(g(y)) \cdot \operatorname{sgn}\left(g'(y)\right) \cdot g''(y) && (1.33)\\
&= \operatorname{sgn}\left(g'(y)\right) \left\{p'_x(g(y)) \cdot \left[g'(y)\right]^2 + p_x(g(y)) \cdot g''(y)\right\}. \quad \left(g'(y) \neq 0\right) && (1.34)
\end{aligned}$$

Putting $\hat{x} = g(\hat{y})$ into the result above yields

$$\begin{aligned}
p'_y(\hat{y}) &= \operatorname{sgn}\left(g'(\hat{y})\right) \left\{p'_x(\hat{x}) \cdot \left[g'(\hat{y})\right]^2 + p_x(\hat{x}) \cdot g''(\hat{y})\right\} && (1.35)\\
&= \operatorname{sgn}\left(g'(\hat{y})\right) \left\{p_x(\hat{x}) \cdot g''(y)\right\}. && (1.36)
\end{aligned}$$

In general, it should be true that $p_x(\hat{x}) > 0$. Due to the fact that $g(y)$ is a nonlinear transformation, generally it should also be true that $g''(y) \neq 0$. That is to say $p'_y(\hat{y}) \neq 0$ in general. Hence we can conclude that the maximum of the density in y is not in general related to the location $\hat{x}$ of the maximum of the density over x.

However, when the transformation $g(y)$ is linear, we have $g''(y) = 0$, and therefore $p'_y(\hat{y}) = 0$ holds. That is, $\hat{y}$ is a maximum of $p_y(y)$. Because $\hat{x} = g(\hat{y})$, it can be seen that the location $\hat{y}$ of the maximum of the density over y is transformed in the same way as the variable itself.

1.5 Using the definition $\operatorname{var}[f] = \mathbb{E}\left[\left(f(x) - \mathbb{E}\left[f(x)\right]^2\right)^2\right]$, show that $\operatorname{var}[f(x)]$ satisfies $\operatorname{var}[f] = \mathbb{E}\left[f(x)^2\right] - \mathbb{E}\left[f(x)\right]^2$.

Answer . Assume that the density function of the expectation is $p(x)$. It should satisfy

$$\begin{aligned}
p(x) &\geq 0, && (1.37)\\
\int p(x)\,dx &= 1. && (1.38)
\end{aligned}$$

The expected value $\mathbb{E}\left[f(x)\right]$ of $f(x)$ is given by

$$\mathbb{E}\left[f(x)\right] = \int p(x) f(x)\,dx. \tag{1.39}$$

↑

It can be seen that when $f(x)$ is a constant function given by $f(x) = c$, then $\mathbb{E}[f(x)] = c$. We can also see that $\mathbb{E}[f(x)]$ is a constant itself because x is integrated out completely.

Given the definition of $\mathbb{E}[f(x)]$ and the properties of integral, it is not hard to conclude that $\mathbb{E}[af(x)] = a\mathbb{E}[f(x)]$ for some consant a, and $\mathbb{E}[f(x) \pm g(x)] = \mathbb{E}[f(x)] \pm \mathbb{E}[g(x)]$. This shows that expectation is a *linear* operator.

Another fact that needs to be shown is $\mathbb{E}[\mathbb{E}[f]] = \mathbb{E}[f]$, due to the fact that $\mathbb{E}[f]$ is a constant. Therefore, we have

$$\begin{aligned}
\operatorname{var}[f] &= \mathbb{E}\left[\left(f(x) - \mathbb{E}[f(x)]^2\right)^2\right] && (1.40)\\
&= \mathbb{E}\left[f(x)^2 - 2 \cdot f(x) \cdot \mathbb{E}[f(x)] + \mathbb{E}[f(x)]^2\right] && (1.41)\\
&= \mathbb{E}\left[f(x)^2\right] - \mathbb{E}[2 \cdot \mathbb{E}[f(x)] \cdot f(x)] + \mathbb{E}\left[\mathbb{E}[f(x)]^2\right] && (1.42)\\
&= \mathbb{E}\left[f(x)^2\right] - 2 \cdot \mathbb{E}[f(x)] \cdot \mathbb{E}[f(x)] + \mathbb{E}\left[f(x)^2\right] && (1.43)\\
&= \mathbb{E}\left[f(x)^2\right] - \mathbb{E}[f(x)]^2 . && (1.44)
\end{aligned}$$

1.6 Show that if two varaibles x and y are independent, then their covariance is zero.

Answer . The covariance of x and y is given by

$$\begin{aligned}
\operatorname{cov}[x, y] &= \mathbb{E}_{x,y}[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}] && (1.45)\\
&= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y] . && (1.46)
\end{aligned}$$

When two variables x and y are independent, then their joint distribution $p(x, y)$ satisfies $p(x, y) = p(x)p(y)$. It can be seen that

$$\begin{aligned}
\mathbb{E}_{x,y}[xy] &= \int\int p(x, y)xy \, dxdy && (1.47)\\
&= \int\int p(x)p(y)xy \, dxdy && (1.48)\\
&= \int p(x)x \, dx \int p(y)y \, dy && (1.49)\\
&= \mathbb{E}[x]\mathbb{E}[y] . && (1.50)
\end{aligned}$$

Therefore we can conclude that if x and y are independent, then $\operatorname{cov}[x, y] = 0$.

1.7 In this exercise, we prove the normalization condition for the univariate Gaussian. (i.e. $\int_{-\infty}^{+\infty} \mathcal{N}(x \mid \mu, \sigma^2) = 1$) To do this consider, the integral

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

which we can evaluate by first writing its square in the form

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

Now make the transformation from Cartesian coordinates (x, y) to polar coordinates (r, θ) and then substitute $u = r^2$. Show that, by performing the integrals over θ and u, and then taking the square root of both sides, we obtain

$$I = \left(2\pi\sigma^2\right)^{\frac{1}{2}}.$$

Finally, use this result to show that the Gaussian distribution $\mathcal{N}(x \mid \mu, \sigma^2)$ is normalized.

Answer . We can make following substitution

$$\begin{cases} x = \phi(r, \theta) = r\cos(\theta) \\ y = \psi(r, \theta) = r\sin(\theta) \end{cases} \tag{1.51}$$

to transform from Cartesian coordinates to polar coordinates. The Jacobian determinant J is

$$J = \left|\frac{\partial(\phi, \psi)}{\partial(r, \theta)}\right| = \begin{vmatrix} \cos(\theta) & -r\sin(\theta) \\ \sin(\theta) & r\cos(\theta) \end{vmatrix} = r. \tag{1.52}$$

Therefore we have

$$\begin{aligned}
I^2 &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \exp\left(-\frac{1}{2\sigma^2}x^2 - \frac{1}{2\sigma^2}y^2\right)\, dx\, dy && (1.53)\\
&= \int_0^{2\pi} \int_0^{+\infty} \exp\left(-\frac{1}{2\sigma^2}r^2\cos^2(\theta) - \frac{1}{2\sigma^2}r^2\sin^2(\theta)\right) |J|\, dr\, d\theta && (1.54)\\
&= \int_0^{2\pi} \int_0^{+\infty} \exp\left[-\frac{1}{2\sigma^2}r^2\left(\cos^2(\theta) + sin^2(\theta)\right)\right] r\, dr\, d\theta && (1.55)\\
&= \int_0^{2\pi} d\theta \int_0^{+\infty} \left(-\frac{1}{2\sigma^2}r^2\right) r\, dr && (1.56)\\
&= 2\pi \int_0^{+\infty} \left(-\frac{u}{2\sigma^2}\right) \frac{1}{2}\, du && (1.57)\\
&= \pi \cdot \left[(-2\sigma^2)\exp\left(-\frac{u}{2\sigma^2}\right)\Big|_0^{+\infty}\right] && (1.58)\\
&= 2\pi\sigma^2. && (1.59)
\end{aligned}$$

It is obvious that

$$\int_{-\infty}^{+\infty} \exp\left(-\frac{1}{2\sigma^2}x^2\right)\, dx = \int_{-\infty}^{+\infty} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right]\, dx = \left(2\pi\sigma^2\right)^{\frac{1}{2}}. \tag{1.60}$$

Therefore we can conclude that $\int_{-\infty}^{+\infty} \left(2\pi\sigma^2\right)^{\frac{1}{2}} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right]\, dx = 1$, that is to say, the Gaussian distribution is

normalized.

Another way of determining I is to make use of the Gamma function $\Gamma(z) = \int_0^{+\infty} x^{z-1} e^{-x}\, dx$. We know that $\Gamma(\frac{1}{2}) = \pi^{\frac{1}{2}}$, that is,

$$\int_0^{+\infty} x^{-\frac{1}{2}} e^{-x}\, dx = \pi^{-\frac{1}{2}}. \tag{1.61}$$

By applying change of variable $v = x^{\frac{1}{2}}$, we have $dv = \frac{1}{2} x^{-\frac{1}{2}} dx$. It can be seen that

$$\int_0^{+\infty} e^{-x} x^{-\frac{1}{2}}\, dx = \int_0^{+\infty} e^{-v^2} 2\, dv = \int_{-\infty}^{+\infty} e^{-v^2}\, dv. \tag{1.62}$$

Further substituting $v = \frac{1}{\sqrt{2}\sigma} w$ yields

$$\int_{-\infty}^{+\infty} e^{-v^2}\, dv = \int_{-\infty}^{+\infty} \exp\left(-\frac{1}{2\sigma^2} w^2\right) \frac{1}{\sqrt{2}\sigma}\, dw = \frac{1}{\sqrt{2}\sigma} \int_{-\infty}^{+\infty} \exp\left(-\frac{1}{2\sigma^2} w^2\right) dw = \pi^{\frac{1}{2}}. \tag{1.63}$$

We can also conclude that

$$I = \int_{-\infty}^{+\infty} \exp\left(-\frac{1}{2\sigma^2} w^2\right) dw = \sqrt{2\pi}\sigma. \tag{1.64}$$

1.8 By using a change of variables, verify that the univariate Gaussian distribution given by $\mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$ satisfies $\mathbb{E}[x] = \mu$. Next, by differentiating both sides of the normalization condition

$$\int_{-\infty}^{+\infty} \mathcal{N}(x \mid \mu, \sigma^2)\, dx = 1$$

with respect to σ^2, verify that the Gaussian satisfies $\mathbb{E}[x^2] = \mu^2 + \sigma^2$. Finally, show that $\mathrm{var}[x] = \sigma^2$ holds.

Answer . We know that

$$\mathbb{E}[x] = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} x\, dx \tag{1.65}$$

$$(\text{let } u = (x-\mu),\ dx = du)$$

$$= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2} u^2\right\} (u+\mu)\, du \tag{1.66}$$

$$= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2} u^2\right\} u\, du + \mu \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2} u^2\right\} du \tag{1.67}$$

Because $\frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2} u^2\right\} u$ is an odd function, the result of the first term is zero. Hence we can conclude that $\mathbb{E}[x] = \mu$.

To compute $\mathbb{E}[x^2]$, we apply differentiation with respect to σ^2 on both sides of the normalization condition. In order to

↑

do this, let $v = \sigma^2$, then $v^{\frac{1}{2}} = \sigma$, and we have

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} dx = 1 \tag{1.68}$$

$$\int_{-\infty}^{+\infty} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} dx = \sqrt{2\pi}\sigma \tag{1.69}$$

$$\int_{-\infty}^{+\infty} \exp\left\{-\frac{1}{2v}(x-\mu)^2\right\} dx = \sqrt{2\pi}v^{\frac{1}{2}} \tag{1.70}$$

$$\frac{\partial}{\partial v}\int_{-\infty}^{+\infty} \exp\left\{-\frac{1}{2v}(x-\mu)^2\right\} dx = \frac{\partial}{\partial v}\sqrt{2\pi}v^{\frac{1}{2}} \tag{1.71}$$

$$\int_{-\infty}^{+\infty} \frac{\partial}{\partial v}\exp\left\{-\frac{1}{2v}(x-\mu)^2\right\} dx = \frac{1}{2}\sqrt{2\pi}v^{-\frac{1}{2}} \tag{1.72}$$

$$\int_{-\infty}^{+\infty} \exp\left\{-\frac{1}{2v}(x-\mu)^2\right\} \frac{(x-\mu)^2}{2v^2} dx = \frac{1}{2}\sqrt{2\pi}v^{-\frac{1}{2}} \tag{1.73}$$

$$\int_{-\infty}^{+\infty} \exp\left\{-\frac{1}{2v}(x-\mu)^2\right\} (x-\mu)^2 dx = \sqrt{2\pi}v^{\frac{3}{2}} \tag{1.74}$$

$$\int_{-\infty}^{+\infty} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} (x-\mu)^2 dx = \sqrt{2\pi}\sigma^3 \tag{1.75}$$

$$\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{+\infty} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} (x-\mu)^2 dx = \sigma^2. \tag{1.76}$$

That is to say, $\mathbb{E}\left[(x-\mu)^2\right] = \mathbb{E}\left[x^2 - 2\mu x + \mu^2\right] = \mathbb{E}\left[x^2\right] - 2\mu\mathbb{E}\left[x\right] + \mu^2 = \mathbb{E}\left[x^2\right] - \mu^2 = \sigma^2$. Therefore we can conclude that $\mathbb{E}\left[x^2\right] = \sigma^2 + \mu^2$.

By definition, $\text{var}\left[x\right] = \mathbb{E}\left[(x - \mathbb{E}\left[x\right])^2\right] = \mathbb{E}\left[(x-\mu)^2\right] = \sigma^2$.

1.9 Show that the mode (i.e. the maximum) of the Gaussian distribution

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

is given by μ.

Similarly, show that the mode of the D-dimensional multivariate Gaussian

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

is given by $\boldsymbol{\mu}$.

Answer . The univariate Gaussian distribution is obviously a special case of the multivariate Gaussian distribution, therefore we only need to prove the multivariate situation.

Denote the element on ith row, jth column of some matrix $\boldsymbol{A}$ by A_{ij}; denote the ith element of some vector $\boldsymbol{a}$ by a_i. It is important to point out that $\boldsymbol{\Sigma}$ is symmetric, which indicates that $\boldsymbol{\Sigma}^{-1}$ is also symmetric. That is to say, $\Sigma_{ij}^{-1} = \Sigma_{ji}^{-1}$.

To prove that the mode of D-dimensional multivariate Gaussian is given by $\boldsymbol{\mu}$, we simply apply partial derivative with respect to the lth element x_l and equate it to zero. By setting $l = 1, 2, \ldots, D$, we will get a set of linear equations that allows

us to solve for each x_l. The partial derivative is

$$\frac{\partial}{\partial x_l}\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\partial}{\partial x_l}\frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \tag{1.77}$$

$$= \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}\frac{\partial}{\partial x_l}\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}. \tag{1.78}$$

For it to be equal to zero, it is clear that

$$\frac{\partial}{\partial x_l}\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\} = 0 \tag{1.79}$$

$$\frac{\partial}{\partial x_l}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = 0 \tag{1.80}$$

Expanding the result of $\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$, we have

$$\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = \begin{bmatrix} \sum_j \Sigma_{1j}^{-1}\left(x_j-\mu_j\right) \\ \vdots \\ \sum_j \Sigma_{Dj}^{-1}\left(x_j-\mu_j\right) \end{bmatrix}, \tag{1.81}$$

which allows us to write

$$(\boldsymbol{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = \sum_k (x_k-\mu_k)\left[\sum_j \Sigma_{kj}^{-1}\left(x_j-\mu_j\right)\right] \tag{1.82}$$

$$= \sum_k\sum_j \Sigma_{kj}^{-1}(x_k-\mu_k)\left(x_j-\mu_j\right) \tag{1.83}$$

$$= \sum_k\sum_j \Sigma_{kj}^{-1}\left(x_kx_j-\mu_kx_j-\mu_jx_k+\mu_k\mu_j\right). \tag{1.84}$$

It can be seen that

$$\frac{\partial}{\partial x_l}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \tag{1.85}$$

$$= \frac{\partial}{\partial x_l}\sum_k\sum_j \Sigma_{kj}^{-1}\left(x_kx_j-\mu_kx_j-\mu_jx_k+\mu_k\mu_j\right) \tag{1.86}$$

$$= \sum_k\sum_j \frac{\partial}{\partial x_l}\Sigma_{kj}^{-1}x_kx_j - \sum_k\sum_j \frac{\partial}{\partial x_l}\Sigma_{kj}^{-1}\mu_kx_j - \sum_k\sum_j \frac{\partial}{\partial x_l}\Sigma_{kj}^{-1}\mu_jx_k + \sum_k\sum_j \frac{\partial}{\partial x_l}\Sigma_{kj}^{-1}\mu_k\mu_j. \tag{1.87}$$

There are four terms in (1.87), which we will simplify one by one. It is obvious that the fourth term equals to zero. For the second term, it is not hard to see that

$$\sum_k\sum_j \frac{\partial}{\partial x_l}\Sigma_{kj}^{-1}\mu_kx_j = \sum_k \Sigma_{kl}^{-1}\mu_k. \tag{1.88}$$

↑

Similarly, the third term can be simplified as

$$\sum_k \sum_j \frac{\partial}{\partial x_l} \Sigma_{kj}^{-1} \mu_j x_k = \sum_j \Sigma_{lj}^{-1} \mu_j. \tag{1.89}$$

Because $\mathbf{\Sigma}^{-1}$ is symmetric, we have

$$\sum_k \Sigma_{kl}^{-1} \mu_k = \sum_j \Sigma_{lj}^{-1} \mu_j. \tag{1.90}$$

The first term is a bit more complicated. Due to the symmetry of $\mathbf{\Sigma}^{-1}$, it can be seen that

$$\sum_k \sum_j \frac{\partial}{\partial x_l} \Sigma_{kj}^{-1} x_k x_j \tag{1.91}$$

$$= \sum_k \sum_j \Sigma_{kj}^{-1} \frac{\partial x_k}{\partial x_l} x_j + \sum_k \sum_j \Sigma_{kj}^{-1} \frac{\partial x_j}{\partial x_l} x_k \tag{1.92}$$

$$= \sum_j \Sigma_{lj}^{-1} x_j + \sum_k \Sigma_{kl}^{-1} x_k \tag{1.93}$$

$$= 2 \sum_j \Sigma_{lj}^{-1} x_j. \tag{1.94}$$

Hence (1.87) yields

$$\sum_k \sum_j \frac{\partial}{\partial x_l} \Sigma_{kj}^{-1} x_k x_j - \sum_k \sum_j \frac{\partial}{\partial x_l} \Sigma_{kj}^{-1} \mu_k x_j - \sum_k \sum_j \frac{\partial}{\partial x_l} \Sigma_{kj}^{-1} \mu_j x_k + \sum_k \sum_j \frac{\partial}{\partial x_l} \Sigma_{kj}^{-1} \mu_k \mu_j \tag{1.95}$$

$$= 2 \sum_j \Sigma_{lj}^{-1} x_j - 2 \sum_j \Sigma_{lj}^{-1} \mu_j. \tag{1.96}$$

Equating it to zero, we have

$$\sum_j \Sigma_{lj}^{-1} x_j = \sum_j \Sigma_{lj}^{-1} \mu_j, \tag{1.97}$$

for $l = 1, 2, \ldots D$. By observing this equation, we find out that it is exactly the expansion of the equation

$$\mathbf{\Sigma}^{-1} \boldsymbol{x}' = \mathbf{\Sigma}^{-1} \boldsymbol{\mu}. \tag{1.98}$$

Multiplying both sides by $\mathbf{\Sigma}$ yields

$$\boldsymbol{x}' = \boldsymbol{\mu}. \tag{1.99}$$

1.10 Suppose that the two variables x and z are statistically independent. Show that the mean and variance of their sum

satisfies

$$\mathbb{E}[x+z] = \mathbb{E}[x] + evz$$
$$\text{var}[x+z] = \text{var}[x] + \text{var}[z].$$

Answer . Because these two variables are statistically independent, we know that $p(x,z) = p(x)p(z)$. Therefore

$$\mathbb{E}[x+z] = \int\int (x+z)p(x,z)\,dx\,dz \tag{1.100}$$
$$= \int\int (x+z)p(x)p(z)\,dx\,dz \tag{1.101}$$
$$= \int\int xp(x)p(z) + zp(x)p(z)\,dx\,dz \tag{1.102}$$
$$= \int \mathbb{E}[x]\,p(z) + zp(z)\,dz \tag{1.103}$$
$$= \mathbb{E}[x] + \mathbb{E}[z]. \tag{1.104}$$

Similarly, we have

$$\text{var}[x+z] = \int\int (x+z-\mathbb{E}[x+z])^2 p(x,z)\,dx\,dz \tag{1.105}$$
$$= \int\int (x+z-\mathbb{E}[x]-\mathbb{E}[z])^2 p(x)p(z)\,dx\,dz \tag{1.106}$$
$$= \int\int [(x-\mathbb{E}[x]) + (z-\mathbb{E}[z])]^2\, p(x)p(z)\,dx\,dz \tag{1.107}$$
$$= \int\int (x-\mathbb{E}[x])^2\, p(x)p(z)\,dx\,dz + \int\int (z-\mathbb{E}[z])^2\, p(x)p(z)\,dx\,dz + \int\int 2\,(x-\mathbb{E}[x])\,(z-\mathbb{E}[z])\, p(x)p(z)\,dx\,dz \tag{1.108}$$
$$= \text{var}[x] + \text{var}[z] + 2\text{cov}[x,z]. \tag{1.109}$$

By the conclusions of question **1.6**, we have already known that $\text{cov}[x,z] = 0$ if x and z are independent. Therefore $\text{var}[x+z] = \text{var}[x] + \text{var}[z]$ holds.

1.11 By setting the derivatives of log likelihood function

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

verify the results of

$$\mu_{\text{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n,$$

↑

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

Answer . By setting the derivative with respect to μ to zero, we have

$$\frac{\partial}{\partial \mu}\left[-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)\right] = 0 \tag{1.110}$$

$$\frac{\partial}{\partial \mu}\left[-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\right] = 0 \tag{1.111}$$

$$\frac{\partial}{\partial \mu}\sum_{n=1}^{N}(x_n^2 - 2\mu x_n + \mu^2) = 0 \tag{1.112}$$

$$\frac{\partial}{\partial \mu}\left[\sum_{n=1}^{N}x_n^2 - 2\sum_{n=1}^{N}\mu x_n + N\mu^2\right] = 0 \tag{1.113}$$

$$-2\sum_{n=1}^{N}x_n + 2N\mu = 0 \tag{1.114}$$

$$\mu = \frac{1}{N}\sum_{n=1}^{N}x_n. \tag{1.115}$$

We can conclude that the formula for μ_{ML} is correct.

Let $u = \sigma^2$, by setting the derivative with respect to u to zero, we have

$$\frac{\partial}{\partial \sigma^2}\left[-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)\right] = 0$$

$$\frac{\partial}{\partial u}\left[-\frac{1}{2u}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln u - \frac{N}{2}\ln(2\pi)\right] = 0$$

$$\frac{1}{2u^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2u} = 0$$

$$u = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu)^2.$$

By substituting μ with μ_{ML}, we can conclude that the formula for σ^2_{ML} is correct.

1.12 Using the fact that

$$\mathbb{E}[x] = \mu$$

$$\mathbb{E}\left[x^2\right] = \mu^2 + \sigma^2,$$

show that

$$\mathbb{E}[x_n x_m] = \mu^2 + I_{nm}\sigma^2,$$

↑

where x_n and x_m denote data points sampled from a Gaussian distribution with mean μ and variance σ^2, and I_{nm} satisfies $I_{nm} = 1$ if $n = m$ and $I_{nm} = 0$ otherwise. Hence prove

$$\mathbb{E}\left[\mu_{\mathrm{ML}}\right] = \mu$$
$$\mathbb{E}\left[\sigma^2_{\mathrm{ML}}\right] = \left(\frac{N-1}{N}\right)\sigma^2.$$

Answer . If $n = m$, clearly that $\mathbb{E}\left[x_n x_m\right] = \mathbb{E}\left[x_n^2\right] = \mu^2 + \sigma^2$. Otherwise, because x_n and x_m are independent, it can be seen that $p(x_n, x_m) = p(x_n)p(x_m)$. Therefore we have

$$\begin{aligned}
\mathbb{E}\left[x_n x_m\right] &= \int\int x_n x_m p(x_n, x_m)\, dx_n\, dx_m && (1.116)\\
&= \int\int x_n x_m p(x_n) p(x_m)\, dx_n\, dx_m && (1.117)\\
&= \int x_n p(x_n)\, dx_n \int x_m p(x_m)\, dx_m && (1.118)\\
&= \mathbb{E}\left[x_n\right]\mathbb{E}\left[x_m\right] && (1.119)\\
&= \mu^2. && (1.120)
\end{aligned}$$

Hence we can conclude that $\mathbb{E}\left[x_n x_m\right] = \mu^2 + I_{nm}\sigma^2$.

We know that

$$\begin{aligned}
\mu_{\mathrm{ML}} &= \frac{1}{N}\sum_{n=1}^{N} x_n && (1.121)\\
\sigma^2_{\mathrm{ML}} &= \frac{1}{N}\sum_{n=1}^{N}\left(x_n - \mu_{\mathrm{ML}}\right)^2. && (1.122)
\end{aligned}$$

It is obvious that

$$\begin{aligned}
\mathbb{E}\left[\mu_{\mathrm{ML}}\right] &= \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N} x_n\right] && (1.123)\\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[x_n\right] && (1.124)\\
&= \mu. && (1.125)
\end{aligned}$$

To compute the expectation of σ^2_{ML}, we need to compute the expectation of $\mu_{\mathrm{ML}} x_m$ and μ^2_{ML}. The first one writes

$$\begin{aligned}
\mathbb{E}\left[\mu_{\mathrm{ML}} x_m\right] &= \mathbb{E}\left[\frac{x_m}{N}\sum_{n=1}^{N} x_n\right] && (1.126)\\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[x_n x_m\right]. && (1.127)
\end{aligned}$$

There are N terms in total, where only one term satisfies $n = m$, therefore we have

$$\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[x_n x_m\right] \tag{1.128}$$

$$= \frac{1}{N}\left(N\mu^2 + \sigma^2\right) \tag{1.129}$$

$$= \mu^2 + \frac{1}{N}\sigma^2. \tag{1.130}$$

The second one writes

$$\mathbb{E}\left[\mu_{\mathrm{ML}}^2\right] = \mathbb{E}\left[\left(\frac{1}{N}\sum_{n=1}^{N}x_n\right)\left(\frac{1}{N}\sum_{n=1}^{N}x_n\right)\right] \tag{1.131}$$

$$= \frac{1}{N^2}\mathbb{E}\left[\sum_{i=1}^{N}\sum_{j=1}^{N}x_i x_j\right]. \tag{1.132}$$

There are N^2 terms in total, among which N terms satisfies $i = j$. Therefore, we can write

$$\frac{1}{N^2}\mathbb{E}\left[\sum_{i=1}^{N}\sum_{j=1}^{N}x_i x_j\right] \tag{1.133}$$

$$= \frac{1}{N^2}(N^2\mu^2 + N\sigma^2) \tag{1.134}$$

$$= \mu^2 + \frac{1}{N}\sigma^2. \tag{1.135}$$

Hence we have

$$\mathbb{E}\left[\sigma_{\mathrm{ML}}^2\right] = \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}\left(x_n - \mu_{\mathrm{ML}}\right)^2\right] \tag{1.136}$$

$$= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[\left(x_n - \mu_{\mathrm{ML}}\right)^2\right] \tag{1.137}$$

$$= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[x_n^2 - 2\mu_{\mathrm{ML}}x_n + \mu_{\mathrm{ML}}^2\right] \tag{1.138}$$

$$= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[x_n^2\right] - \frac{2}{N}\sum_{n=1}^{N}\mathbb{E}\left[\mu_{\mathrm{ML}}x_n\right] + \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[\mu_{\mathrm{ML}}^2\right] \tag{1.139}$$

$$= \frac{1}{N}\cdot N\cdot(\mu^2 + \sigma^2) - \frac{2}{N}\cdot N\cdot(\mu^2 + \frac{1}{N}\sigma^2) + \frac{1}{N}\cdot N\cdot(\mu^2 + \frac{1}{N}\sigma^2) \tag{1.140}$$

$$= \sigma^2 - \frac{1}{N}\sigma^2 \tag{1.141}$$

$$= \frac{N-1}{N}\sigma^2. \tag{1.142}$$

1.13 Suppose that the variance of a Gaussian is estimated using $\sigma_{\mathrm{ML}}^2 = \frac{1}{N}\sum_{n=1}^{N}\left(x_n - \mu_{\mathrm{ML}}\right)^2$, but with the maximum likelihood estimate μ_{ML} replaced with the true value μ of the mean. Show that this estimator has the property that its expectation is given by the true variance σ^2.

↑

Answer . In this scenario, the estimator $\hat{\sigma^2}$ of σ^2 is given by $\hat{\sigma^2} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^2$, it can be seen that

$$\mathbb{E}\left[\hat{\sigma^2}\right] = \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^2\right] \tag{1.143}$$

$$= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[(x_n - \mu)^2\right] \tag{1.144}$$

$$= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[x_n^2 - 2\mu x_n + \mu^2\right] \tag{1.145}$$

$$= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[x_n^2\right] - \frac{2\mu}{N}\sum_{n=1}^{N}\mathbb{E}\left[x_n\right] + \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[\mu^2\right] \tag{1.146}$$

$$= \frac{1}{N}\cdot N\cdot(\mu^2 + \sigma^2) - \frac{2\mu}{N}\cdot N\cdot\mu + \frac{1}{N}\cdot N\cdot\mu^2 \tag{1.147}$$

$$= \sigma^2. \tag{1.148}$$

The difference between $\hat{\sigma^2}$ and σ^2_{ML} is raised by the fact that μ is a constant. When it is multiplied to another variable, it only changes the expected value linearly. However, μ_{ML} is a statistic, which is a function of the observations. When it is multiplied to another variable, it might not change the expected value in a linear fashion.

1.14 Show that an arbitrary square matrix with elements w_{ij} can be written in the form $w_{ij} = w_{ij}^S + w_{ij}^A$ where w_{ij}^S and w_{ij}^A are symmetric and anti-symmetric matrices, respectively, satisfying $w_{ij}^S = w_{ji}^S$ and $w_{ij}^A = -w_{ji}^A$ for all i and j. Now consider the second order term in a higher order polynomial in D dimensions, given by

$$\sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}x_i x_j.$$

Show that

$$\sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}x_i x_j = \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}^S x_i x_j$$

so that the contribution from the anti-symmetric matrix vanishes. We therefore see that without loss of generality, the matrix of coefficients w_{ij} can be chosen to be symmetric, and so not all of the D^2 elements of this matrix can be chosen independently. Show that the number of independent parameters in the matrix w_{ij}^S can be given by $D(D+1)/2$.

Answer . Denote the original matrix, symmetric matrix and anti-symmetric matrix by W, W^S and W^A, respectively. It is clear that we have

$$W = W^S + W^A \tag{1.149}$$

$$W^S = \left(W^S\right)^{\mathsf{T}} \tag{1.150}$$

$$W^A = -\left(W^A\right)^{\mathsf{T}}. \tag{1.151}$$

↑

By transposing 1.150 we have

$$W^{\mathsf{T}} = \left(W^S\right)^{\mathsf{T}} + \left(W^A\right)^{\mathsf{T}} \tag{1.152}$$

$$= W^S - W^A. \tag{1.153}$$

Immediately we can write

$$W^S = \frac{W + W^{\mathsf{T}}}{2}, \tag{1.154}$$

and that

$$W^A = \frac{W - W^{\mathsf{T}}}{2}. \tag{1.155}$$

Therefore we have

$$w_{ij}^S = \frac{1}{2}\left(w_{ij} + w_{ij}^{\mathsf{T}}\right) \tag{1.156}$$

$$= \frac{1}{2}\left(w_{ij} + w_{ji}\right) \tag{1.157}$$

which allows us to write

$$\sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}^S x_i x_j \tag{1.158}$$

$$= \sum_{i=1}^{D}\sum_{j=1}^{D} \frac{1}{2}\left(w_{ij} + w_{ji}\right) x_i x_j \tag{1.159}$$

$$= \frac{1}{2}\sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij} x_i x_j + \frac{1}{2}\sum_{i=1}^{D}\sum_{j=1}^{D} w_{ji} x_j x_i \tag{1.160}$$

$$= \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij} x_i x_j. \tag{1.161}$$

Let $\boldsymbol{x} = (x_1, \ldots, x_D)^{\mathsf{T}}$, this actually proves $\boldsymbol{x}^{\mathsf{T}} W \boldsymbol{x} = \boldsymbol{x}^{\mathsf{T}} W^S \boldsymbol{x}$. Now we have verified that the the contribution from the anti-symmetric matrix vanishes.

In symmetric matrix W^S, there are D^2 elements in total. D elements are on the diagonal, and there are $D^2 - D$ elements above and below the diagonal, which are symmetric. Therefore we can only choose among

$$\frac{D^2 - D}{2} + D = \frac{D(D+1)}{2} \tag{1.162}$$

parameters.

1.15 In this exercise and next, we explore how te number of independent parameters in a polynomial grows with the

order M of the polynomial and with the dimensionality M of the output space. We start by writing down the M^{th} order term for a polynomial in D dimensions in the form

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

The coefficients $w_{i_1 i_2 \cdots i_M}$ comprise D^M elements, but the number of independent parameters is significantly fewer due to many interchange symmetries of the factor $x_{i_1} x_{i_2} \cdots x_{i_M}$. Begin by showing that the redundancy in the coefficients can be removed by rewriting this M^{th} order in the form

$$\sum_{i_1=1}^{D}\sum_{i_2=1}^{i_1}\cdots\sum_{i_M=1}^{i_{M-1}} \tilde{w}_{i_1 i_2 \cdots i_M} x_{i_1} x_{i_2} \cdots x_{i_M}.$$

Note that the precise relationship between the $\tilde{w}$ coefficients and the w coefficients need not be made explicit. Use this result to show that the number of *independent* parameters $n(D, M)$, which appear at order M, satisfies the following recursion relation

$$n(D, M) = \sum_{i=1}^{D} n(i, M-1).$$

Next use proof by induction to show that the following result holds

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

which can be done by first proving the results for $D = 1$ and arbitrary M by making use of the result $0! = 1$, then assuming it is correct for dimension D and verifying that is is correct for dimension $D+1$. Finally, use the two previous results, together with proof by induction, show

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

To do this, first show that the result is true for $M = 2$, and any value of $D \geq 1$, by comparison with the result of Exercise **1.14**. Then make use of $n(D, M) = \sum_{i=1}^{D} n(i, M-1)$, to show that, if the result holds at order $M-1$, then it will also hold at order M.

Answer . To find out how the redundancy forms and how to eliminate it, let us appreciate i_a and i_b, where $a < b$. The M^{th} order term writes

$$\sum_{i_1=1}^{D}\cdots\sum_{i_a=1}^{D}\cdots\sum_{i_b=1}^{D}\cdots\sum_{i_M=1}^{D} w_{i_1 i_2 \cdots i_M} x_{i_1} x_{i_2} \cdots x_{i_M} \tag{1.163}$$

$$= \sum_{i_a=1}^{D}\sum_{i_b=1}^{D}\cdots\sum\cdots\left(w_{\cdots i_a \cdots i_b \cdots}\right)\left(x_{i_a} x_{i_b} \cdots\right). \tag{1.164}$$

↑

Temporarily ignoring other variables, now the term has an identical form to the expression in question **1.14**. Now the square coefficient matrix W is given by $w_{ij} = w_{\cdots i_a \cdots i_b \cdots}\big|_{i_a=i, i_b=j}$. By its conclusion, we can make W a symmetric matrix, and the independent elements are those that locate on the diagonal and below. That is to say, when $i_a \geq i_b$, the coefficients are independent. This holds for any arbitrary pairs of i_a and i_b where $a < b$. If we guarantee

$$i_1 \geq i_2 \geq \cdots \geq i_M, \tag{1.165}$$

then the resulting terms have no redundancy. Therefore the expression

$$\sum_{i_1=1}^{D} \sum_{i_2=1}^{i_1} \cdots \sum_{i_M=1}^{i_{M-1}} \tilde{w}_{i_1 i_2 \cdots i_M} x_{i_1} x_{i_2} \cdots x_{i_M} \tag{1.166}$$

has no redundancy.

The number of independent parameters $n(D, M)$ can be given by

$$n(D, M) = \sum_{i_1=1}^{D} \sum_{i_2=1}^{i_1} \cdots \sum_{i_M=1}^{i_{M-1}} 1 \tag{1.167}$$

$$= \sum_{i_1=1}^{D} \left(\sum_{i_2=1}^{i_1} \cdots \sum_{i_M=1}^{i_{M-1}} 1 \right) \tag{1.168}$$

$$= \sum_{i_1=1}^{D} n(i_1, M-1). \tag{1.169}$$

Now we continue to prove

$$\sum_{i=1}^{D} \frac{(i+M-2)!}{(i-1)!(M-1)!} = \frac{(D+M-1)!}{(D-1)!M!} \tag{1.170}$$

by induction on D.

Basis: when $D = 1$, it can be seen that

$$\sum_{i=1}^{D} \frac{(i+M-2)!}{(i-1)!(M-1)!} = \frac{(M-1)!}{(M-1)!} = 1. \tag{1.171}$$

For the right side, we have

$$\frac{(D+M-1)!}{(D-1)!M!} = \frac{M!}{M!} = 1. \tag{1.172}$$

Therefore the basis holds.

Induction Hypothesis: suppose for some $D' \geq 1$, the claim holds. That is,

$$\sum_{i=1}^{D'} \frac{(i+M-2)!}{(i-1)!(M-1)!} = \frac{(D'+M-1)!}{(D'-1)!M!} \tag{1.173}$$

Induction Step: Now we need to prove that the claim holds for $D'+1$. It can be seen that

$$\sum_{i=1}^{D'+1} \frac{(i+M-2)!}{(i-1)!(M-1)!} \tag{1.174}$$

$$= \frac{(D'+M-1)!}{(D'-1)!M!} + \frac{(D'+M-1)!}{(D')!(M-1)!} \tag{1.175}$$

$$= \frac{D' \cdot (D'+M-1)! + M \cdot (D'+M-1)!}{D'!M!} \tag{1.176}$$

$$= \frac{(D'+M)(D'+M-1)!}{D'!M!} \tag{1.177}$$

$$= \frac{(D'+M)!}{D'!M!}. \tag{1.178}$$

By induction, we can conclude that the claim holds.

If we let $n(D,M) = \frac{(D+M-1)!}{(D-1)!M!}$, what (1.170) suggest is that we have found a possible expression that satisfies equation (1.169). If it indeed equals to the true value of $n(D,M)$ for some value of M, then it will hold for all subsequent M's. When $M=0$, it is a constant term, so there is only one independent parameter; when $M=1$, there are D independent parameters; by the result of **1.14**, we know that when $M=2$, there are $D(D+1)/2$ independent parameters. It can be seen that

$$n(D,0) = 1 \tag{1.179}$$

$$n(D,1) = D \tag{1.180}$$

$$n(D,2) = \frac{D(D+1)}{2}. \tag{1.181}$$

Therefore we can conclude that

$$n(D,M) = \frac{(D+M-1)!}{(D-1)!M!} \tag{1.182}$$

holds for all possible D's and M's.

1.16 In exercise **1.15**, we proved the result

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

for the number of independent parameters in the M^{th} order term of a D-dimensional polynomial. We now find an expression for the total number $N(D,M)$ of independent parameters in all of the terms up to and including the M^{th} order. First show that $N(D,M)$ satisfies

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

↑

where $n(D, m)$ is the number of independent parameters in the term of order m. Now make use of

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

together with proof by induction, to show that

$$N(D, M) = \frac{(D + M)!}{D!M!}.$$

This can be done by first proving that the result holds for $M = 0$ and arbitrary $D \geq 1$, then assuming that it holds at order M, and hence showing that it holds at order $M + 1$. Finally, make use of Stirling's approximation in the form

$$n! \simeq n^n e^{-n}$$

for large n to show that, for $D \gg M$, the quantity $N(D, M)$ grows like D^M, and for $M \gg D$, it grows like M^D. Consider a cubic ($M = 3$) polynomial in D dimensions, and evaluate numerically the total number of independent parameters for (i) $D = 10$ and (ii) $D = 100$, which correspond to typical small-scale and medium-scale machine learning applications.

Answer . Since $n(D, m)$ is the number of independent parameters in the term of order m, it is intuitive to say that the total number of independent parameters is the sum of each term, from order 0 to order M. That is to say,

$$N(D, M) = \sum_{m=0}^{M} n(D, m). \tag{1.183}$$

This is true, as long as there is no redundancy across any two terms, which is also true because the order between any two terms will be different, and that their coefficients will not depend on each other's.

Now we need to prove that $N(D, M) = \frac{(D+M)!}{D!M!}$. It can be seen that

$$N(D, M) = \sum_{m=0}^{M} n(D, m) \tag{1.184}$$

$$= \sum_{m=0}^{M} \frac{(m + D - 1)!}{(D - 1)!m!}. \tag{1.185}$$

We shall prove this claim by mathematical induction on M.

Basis: when $M = 0$, we have

$$N(D, 0) = n(D, 0) = \frac{(D - 1)!}{(D - 1)!} = 1, \tag{1.186}$$

while for the right side, we have

$$\frac{(D + 0)!}{D!0!} = 1. \tag{1.187}$$

Therefore the claim is true.

Induction Hypothesis: Suppose for some arbitrary M', the claim is true. That is,

$$N(D, M') = \frac{(D + M')!}{D!M'!}. \tag{1.188}$$

Induction Step: We need to prove that the claim holds for $M' + 1$. It can be seen that

$$\begin{aligned}
N(D, M' + 1) &= \sum_{m=0}^{M'+1} n(D, m) && (1.189)\\
&= \frac{(D + M')!}{D!M'!} + n(D, M' + 1) && (1.190)\\
&= \frac{(D + M')!}{D!M'!} + \frac{(D + M')!}{(D - 1)!(M' + 1)!} && (1.191)\\
&= \frac{(M' + 1)(D + M')! + D(D + M')!}{D!(M + 1)!} && (1.192)\\
&= \frac{(D + M' + 1)(D + M')!}{D!(M' + 1)!} && (1.193)\\
&= \frac{(D + M' + 1)!}{D!(M' + 1)!}. && (1.194)
\end{aligned}$$

By induction, we know that the claim is true.

Using the Stirling's formula, we can get

$$\begin{aligned}
N(D, M) &\simeq \frac{(D + M)^{D+M} e^{-(D+M)}}{D^D e^{-D} M^M e^{-M}} && (1.195)\\
&\simeq \frac{(D + M)^{D+M}}{D^D M^M} && (1.196)\\
&\simeq \frac{(D + M)^D (D + M)^M}{D^D M^M} && (1.197)\\
&\simeq \frac{D + M}{D}^D \cdot \frac{D + M}{M}^M && (1.198)\\
&\simeq M^D \cdot D^M. && (1.199)
\end{aligned}$$

Consider the case when $D \gg M$, we assume that

$$\frac{M}{D} \simeq 0. \tag{1.200}$$

To compare the scale of M^D and D^M, we study their quotient, which is

$$\frac{M^D}{D^M} = \frac{M^{D-M} M^M}{D^M} = M^{D-M} \cdot \left(\frac{M}{D}\right)^M \simeq 0. \tag{1.201}$$

That is to say, though being a positive integer, the scale of M^D is almost insignificant when it is compared to the scale of D^M. Therefore we can conclude that when $D \gg M$, it grows like D^M. The case when $M \gg D$ can be reasoned in the same manner.

↑

When $M = 3$, $D = 10$, we have

$$N(10,3) = \frac{13!}{10!3!} = 286, \tag{1.202}$$

when $M = 3$, $D = 100$, we have

$$N(100,3) = \frac{(103)!}{100!3!} = 176851. \tag{1.203}$$

1.17 The gamma function is defined by

$$\Gamma(x) \equiv \int_0^{+\infty} u^{x-1} e^{-u}\, du.$$

Use integration by parts, prove the relation $\Gamma(x+1) = x\Gamma(x)$. Show also that $\Gamma(1) = 1$ and hence that $\Gamma(x+1) = x!$ when x is an integer.

Answer . By using integration by parts, it can be seen that

$$\begin{aligned}
\Gamma(x+1) &= \int_0^{+\infty} u^x e^{-u}\, du && (1.204)\\
&= u^x e^{-u}\big|_0^{+\infty} + \int_0^{+\infty} x u^{x-1} e^{-u}\, du && (1.205)\\
&= 0 + x\int_0^{+\infty} u^{x-1} e^{-u}\, du && (1.206)\\
&= x\Gamma(x). && (1.207)
\end{aligned}$$

When $x = 1$, we have

$$\begin{aligned}
\Gamma(1) &= \int_0^{+\infty} e^{-u}\, du && (1.208)\\
&= -e^{-u}\big|_0^{\infty} && (1.209)\\
&= 1. && (1.210)
\end{aligned}$$

Let us prove $\Gamma(x+1) = x! \quad (x \in [0,\infty) \cap \mathbb{Z})$ by induction.

Basis: when $x = 0$, $\Gamma(1) = 0!$ holds.

Induction Hypothesis: suppose for some arbitrary x', $\Gamma(x'+1) = x'!$ holds.

Induction Step: since $\Gamma(x'+1) = x'!$, it can be seen that $\Gamma(x'+2) = (x'+1)\Gamma(x'+1) = (x'+1)x'! = (x'+1)!$. That is to say, the claim holds for $x'+1$. Therefore we can conclude that the claim is true.

↑

1.18 We can use the result

$$\int_{-\infty}^{+\infty} e^{-x^2}\, dx = \sqrt{\pi}$$

to derive an expression for the surface area S_D, and the volume V_D, of a sphere of unit radius in D dimensions. To do this, consider the following result, which is obtained by transforming from Cartesian to polar coordinates

$$\prod_{i=1}^{D} \int_{-\infty}^{+\infty} e^{-x_i^2}\, dx_i = S_D \int_0^{+\infty} e^{-r^2} r^{D-1}\, dr.$$

Using the definition of the Gamma function, together with the result above, evaluate both sides of this equation, and hence show that

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

Next, by integrating with respect to radius form 0 to 1, show that the volume of the unit sphere in D dimensions is given by

$$V_D = \frac{S_D}{D}.$$

Finally, use the results $\Gamma(1) = 1$ and $\Gamma(3/2) = \sqrt{\pi}/2$ to show that the expression of S_D and V_D reduce to the usual expression for $D = 2$ and $D = 3$.

Answer . We begin by appreciating the integral of function $f(x) = \exp(\sum_{i=1}^{D} -x_i^2)$, which can help us derive the formula for the surface area of a n-sphere. It can be seen that

$$\int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \exp(\sum_{i=1}^{D} -x_i^2)\, dx_1 \cdots\, dx_D \tag{1.211}$$

$$= \prod_{i=1}^{D} \int_{-\infty}^{+\infty} e^{-x_i^2}\, dx_i \tag{1.212}$$

$$= \pi^{D/2}. \tag{1.213}$$

Now we transform from Cartesian coordinates $(x_1, x_2, \ldots, x_D)$ to spherical coordinates $(r, \phi_1, \phi_2, \ldots, \phi_{D-1})$, it can be seen

↑

that

$$\begin{cases} x_1 = r\cos(\phi_1) \\ x_2 = r\sin(\phi_1)\cos(\phi_2) \\ x_3 = r\sin(\phi_1)\sin(\phi_2)\cos(\phi) \\ \quad\vdots \\ x_{D-1} = r\sin(\phi_1)\sin(\phi_2)\cdots\sin(\phi_{D-2})\cos(\phi_{D-1}) \\ x_D = r\sin(\phi_1)\sin(\phi_2)\cdots\sin(\phi_{D-2})\sin(\phi_{D-1}) \end{cases}, \tag{1.214}$$

where $r \geq 0, \phi_1, \ldots, \phi_{D-2} \in [0, \pi]$, and $\phi_{D-1} \in [0, 2\pi)$. The property of spherical coordinate ensures that

$$\sum_{i=1}^{D} x_i^2 = r^2. \tag{1.215}$$

The Jacobian of the transformation is

$$\frac{\partial(x_1, \ldots, x_D)}{\partial(r, \phi_1, \ldots, \phi_{D-1})} = \begin{bmatrix} \cos(\phi_1) & -r\sin(\phi_1) & 0 & \cdots & 0 \\ \sin(\phi_1)\cos(\phi_2) & r\cos(\phi_1)\cos(\phi_2) & -r\sin(\phi_1)\sin(\phi_2) & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sin(\phi_1)\cdots\sin(\phi_{D-2})\cos(\phi_{D-1}) & \cdots & \cdots & \cdots & -r\sin(\phi_1)\cdots\sin(\phi_{D-2})\sin(\phi_{D-1}) \\ \sin(\phi_1)\cdots\sin(\phi_{D-2})\sin(\phi_{D-1}) & \cdots & \cdots & \cdots & r\sin(\phi_1)\cdots\sin(\phi_{D-2})\cos(\phi_{D-1}) \end{bmatrix}. \tag{1.216}$$

Denote this Jacobian of the D-dimension by JV_D, to transform from Cartesian coordinates to spherical coordinates, we need to compute the determinant of JV_D. It can be seen that

$$\begin{aligned} |JV_D| &= \begin{vmatrix} \cos(\phi_1) & -r\sin(\phi_1) & 0 & \cdots & 0 \\ \sin(\phi_1)\cos(\phi_2) & r\cos(\phi_1)\cos(\phi_2) & -r\sin(\phi_1)\sin(\phi_2) & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sin(\phi_1)\cdots\sin(\phi_{D-2})\cos(\phi_{D-1}) & \cdots & \cdots & \cdots & -r\sin(\phi_1)\cdots\sin(\phi_{D-2})\sin(\phi_{D-1}) \\ \sin(\phi_1)\cdots\sin(\phi_{D-2})\sin(\phi_{D-1}) & \cdots & \cdots & \cdots & r\sin(\phi_1)\cdots\sin(\phi_{D-2})\cos(\phi_{D-1}) \end{vmatrix} \\ &= r\sin(\phi_1)\cdots\sin(\phi_{D-2})\sin(\phi_{D-1})\times \\ &\quad \begin{vmatrix} \cos(\phi_1) & -r\sin(\phi_1) & 0 & \cdots & 0 \\ \sin(\phi_1)\cos(\phi_2) & r\cos(\phi_1)\cos(\phi_2) & -r\sin(\phi_1)\sin(\phi_2) & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sin(\phi_1)\cdots\sin(\phi_{D-3})\cos(\phi_{D-2}) & \cdots & \cdots & \cdots & -r\sin(\phi_1)\cdots\sin(\phi_{D-3})\sin(\phi_{D-2}) \\ \sin(\phi_1)\cdots\sin(\phi_{D-2})\sin(\phi_{D-1}) & \cdots & \cdots & \cdots & r\sin(\phi_1)\cdots\cos(\phi_{D-2})\sin(\phi_{D-1}) \end{vmatrix} \\ &\quad + r\sin(\phi_1)\cdots\sin(\phi_{D-2})\cos(\phi_{D-1})\times \\ &\quad \begin{vmatrix} \cos(\phi_1) & -r\sin(\phi_1) & 0 & \cdots & 0 \\ \sin(\phi_1)\cos(\phi_2) & r\cos(\phi_1)\cos(\phi_2) & -r\sin(\phi_1)\sin(\phi_2) & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sin(\phi_1)\cdots\sin(\phi_{D-3})\cos(\phi_{D-2}) & \cdots & \cdots & \cdots & -r\sin(\phi_1)\cdots\sin(\phi_{D-3})\sin(\phi_{D-2}) \\ \sin(\phi_1)\cdots\sin(\phi_{D-2})\cos(\phi_{D-1}) & \cdots & \cdots & \cdots & r\sin(\phi_1)\cdots\cos(\phi_{D-2})\cos(\phi_{D-1}) \end{vmatrix} \end{aligned} \tag{1.217, 1.218}$$

↑

$$= r\sin(\phi_1)\cdots\sin(\phi_{D-2})\sin^2(\phi_{D-1})\times$$

$$\begin{vmatrix} \cos(\phi_1) & -r\sin(\phi_1) & 0 & \cdots & 0 \\ \sin(\phi_1)\cos(\phi_2) & r\cos(\phi_1)\cos(\phi_2) & -r\sin(\phi_1)\sin(\phi_2) & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sin(\phi_1)\cdots\sin(\phi_{D-3})\cos(\phi_{D-2}) & \cdots & \cdots & \cdots & -r\sin(\phi_1)\cdots\sin(\phi_{D-3})\sin(\phi_{D-2}) \\ \sin(\phi_1)\cdots\sin(\phi_{D-3})\sin(\phi_{D-2}) & \cdots & \cdots & \cdots & r\sin(\phi_1)\cdots\sin(\phi_{D-3})\cos(\phi_{D-2}) \end{vmatrix} \tag{1.219}$$

$$+ r\sin(\phi_1)\cdots\sin(\phi_{D-2})\cos^2(\phi_{D-1})\times$$

$$\begin{vmatrix} \cos(\phi_1) & -r\sin(\phi_1) & 0 & \cdots & 0 \\ \sin(\phi_1)\cos(\phi_2) & r\cos(\phi_1)\cos(\phi_2) & -r\sin(\phi_1)\sin(\phi_2) & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sin(\phi_1)\cdots\sin(\phi_{D-3})\cos(\phi_{D-2}) & \cdots & \cdots & \cdots & -r\sin(\phi_1)\cdots\sin(\phi_{D-3})\sin(\phi_{D-2}) \\ \sin(\phi_1)\cdots\sin(\phi_{D-3})\sin(\phi_{D-2}) & \cdots & \cdots & \cdots & r\sin(\phi_1)\cdots\sin(\phi_{D-3})\cos(\phi_{D-2}) \end{vmatrix}$$

$$= r\sin(\phi_1)\cdots\sin(\phi_{D-2})\sin^2(\phi_{D-1})\cdot|JV_{D-1}| + r\sin(\phi_1)\cdots\sin(\phi_{D-2})\cos^2(\phi_{D-1})\cdot|JV_{D-1}| \tag{1.220}$$

$$= r\sin(\phi_1)\cdots\sin(\phi_{D-2})\cdot|JV_{D-1}|\cdot\left[\sin^2(\phi_{D-1})+\cos^2(\phi_{D-1})\right] \tag{1.221}$$

$$= r\sin(\phi_1)\cdots\sin(\phi_{D-2})\cdot|JV_{D-1}|. \tag{1.222}$$

Now we have acquired a recursion relation for $|JV_D|$, which is

$$|JV_D| = r\left[\prod_{i=1}^{D-2}\sin(\phi_i)\right]|JV_{D-1}|. \tag{1.223}$$

When $D = 2$, it is just the case of polar coordinates, and we know that $|JV_2| = r$. Therefore we can conclude that

$$|JV_D| = r^{D-1}\sin^{D-2}(\phi_1)\sin^{D-3}(\phi_2)\cdots\sin(\phi_{D-2}). \tag{1.224}$$

By the domain of $\phi_1,\ldots,\phi_{D-2}$ we know that JV_D is always greater or equal to zero. It can be seen that

$$dx_1\cdots\ dx_D = dV = \|JV_D\|dr\ d\phi_1\cdots\ d\phi_{D-1} \tag{1.225}$$

$$= r^{D-1}\sin^{D-2}(\phi_1)\sin^{D-3}(\phi_2)\cdots\sin(\phi_{D-2})\ dr\ d\phi_1\cdots\ d\phi_{D-1}, \tag{1.226}$$

where dV is the D-dimensional volume element. If we encapsulate all angular parameters $\phi_1,\ldots,\phi_{D-1}$ with a set Ω, it can be seen that

$$dV = r^{D-1}\ dr\ d\Omega. \tag{1.227}$$

As a matter of fact, $d\Omega$ is just the surface element in D-dimensional spherical coordinate. If we denote the domain of the D-dimensional unit sphere by $\odot$, we would have $\int_{\odot} d\Omega = S_D$.

Back to (1.211), whose actual value is already known. By transforming from Cartesian coordinates to spherical coordinates, we have

$$\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}\exp(\sum_{i=1}^{D}-x_i^2)\ dx_1\cdots\ dx_D \tag{1.228}$$

↑

$$= \int_0^{+\infty} \int_{\odot} e^{-r^2} r^{D-1} \, dr \, d\Omega \tag{1.229}$$

$$= \int_{\odot} d\Omega \int_0^{+\infty} e^{-r^2} r^{D-1} \, dr \tag{1.230}$$

$$= \frac{1}{2} S_D \int_0^{+\infty} r^{D-2} e^{-r^2} 2r \, dr \tag{1.231}$$

$$\text{(let } u = r^2\text{, then } du = 2r \, dr\text{)} \tag{1.232}$$

$$= \frac{1}{2} \cdot S_D \int_0^{+\infty} u^{\frac{D}{2}-1} e^{-u} \, du \tag{1.233}$$

$$= \frac{1}{2} \cdot S_D \cdot \Gamma(D/2) = \pi^{D/2}. \tag{1.234}$$

Therefore we can conclude that

$$S_D = \frac{2\pi^{D/2}}{\Gamma(D/2)}. \tag{1.235}$$

The volume of the unit sphere is given by

$$V_D = \int_0^1 \int_{\odot} r^{D-1} \, dr \, d\Omega \tag{1.236}$$

$$= \int_{\odot} d\Omega \int_0^1 r^{D-1} \, dr \tag{1.237}$$

$$= S_D \left(\frac{r^D}{D} \Big|_0^1 \right) \tag{1.238}$$

$$= \frac{S_D}{D}. \tag{1.239}$$

We know that when $D = 2$, the circumference of the unit circle is 2π, and the area is π; when $D = 3$, the surface area of the unit sphere is 4π, and the volume is $\frac{4}{3}\pi$. According to (1.235) and (1.239), it can be seen that

$$S_2 = \frac{2\pi}{\Gamma(1)} = 2\pi \tag{1.240}$$

$$V_2 = \frac{S_2}{2} = \pi \tag{1.241}$$

$$S_3 = \frac{2\pi^{3/2}}{\Gamma(3/2)} = 4\pi \tag{1.242}$$

$$V_3 = \frac{S_3}{3} = \frac{4\pi}{3}. \tag{1.243}$$

We can see that our formulae accord with actual observations.

1.19 Consider a sphere of radius a in D-dimensions together with the concentric hypercube of side $2a$, so that the sphere touches the hypercube at the centers of each of its sides. By using the results of exercise **1.18**, show that the ratio of the volume of the sphere to the volume of the cube is given by

$$\frac{\text{volume of sphere}}{\text{volume of cube}} = \frac{\pi^{D/2}}{D 2^{D-1} \Gamma(D/2)}.$$

↑

Now make use of the Stirling's formula in the form

$$\Gamma(x+1) \simeq (2\pi)^{1/2} e^{-x} x^{x+(1/2)}$$

which is valid for $x \gg 1$, to show that as $D \to \infty$, the ratio above goes to zero. Show also that the ratio of the distance from the center of the hypercube to one of the corners, divided by the perpendicular distance to one of the sides, is $\sqrt{D}$, which therefore goes to ∞ as $D \to \infty$. From these results we see that, in a space of high dimensionality, most of the volume of a cube is concentrated in the large number of corners, which themselves become very long 'spikes'!

Answer . The volume of the cube in D-dimensions is

$$\int_{-a}^{a} \cdots \int_{-a}^{a} dx_1 \cdots dx_D \tag{1.244}$$

$$= \prod_{i=1}^{D} \int_{-a}^{a} dx_i \tag{1.245}$$

$$= (2a)^D. \tag{1.246}$$

Using the similar approach in **1.18**, we know that the volume of the sphere is

$$\int_0^a \int_{\odot} r^{D-1} \, dr \, d\Omega \tag{1.247}$$

$$= \int_{\odot} d\Omega \int_0^a r^{D-1} \, dr \tag{1.248}$$

$$= S_D \cdot \left(\frac{r^D}{D} \Big|_0^a \right) \tag{1.249}$$

$$= \frac{a^D S_D}{D} \tag{1.250}$$

$$= \frac{2a^D \pi^{D/2}}{D\Gamma(D/2)}. \tag{1.251}$$

Now we can clearly see that

$$\frac{\text{volume of sphere}}{\text{volume of cube}} = \frac{\pi^{D/2}}{D 2^{D-1} \Gamma(D/2)}. \tag{1.252}$$

Consider putting $D+2$ instead of D in (1.252), when $D \to \infty$, there limit should be the same. By doing so, the ratio becomes

$$\frac{\pi^{D/2+1}}{(D+2) 2^{D+1} \Gamma(D/2+1)}. \tag{1.253}$$

Inserting the Stirling's formula, we have

$$\lim_{D \to \infty} \frac{\pi^{D/2+1}}{(D+2) 2^{D+1} \Gamma(D/2+1)} \tag{1.254}$$

↑

$$\simeq \lim_{D\to\infty} \frac{\pi^{D/2+1}}{(D+2)2^{D+1}(2\pi)^{1/2}e^{-D/2}(D/2)^{(D+1)/2}} \tag{1.255}$$

$$= \lim_{D\to\infty} \frac{e^{D/2}\pi^{D/2+1}}{(D+2)2^{D+1}(2\pi)^{1/2}(D/2)^{(D+1)/2}} \tag{1.256}$$

$$= \lim_{D\to\infty} \frac{\pi}{(D+2)2^{D+1}(2\pi)^{1/2}(D/2)^{1/2}} \cdot \frac{(e\pi)^{D/2}}{(D/2)^{D/2}} \tag{1.257}$$

$$= \lim_{D\to\infty} \frac{\pi}{(D+2)2^{D+1}(2\pi)^{1/2}(D/2)^{1/2}} \cdot \lim_{D\to\infty} \left(\frac{2e\pi}{D}\right)^{D/2} \tag{1.258}$$

It is obvious that $\lim_{D\to\infty} \frac{\pi}{(D+2)2^{D+1}(2\pi)^{1/2}(D/2)^{1/2}} = 0$, now we need to prove that $\lim_{D\to\infty} \left(\frac{2e\pi}{D}\right)^{D/2} = 0$. It can be seen that when $D > 100$, the following inequality holds

$$\frac{1}{D} \leq \frac{2e\pi}{D} \leq \frac{1}{2}, \tag{1.259}$$

and therefore

$$\left(\frac{1}{D}\right)^{D/2} \leq \left(\frac{2e\pi}{D}\right)^{D/2} \leq \left(\frac{1}{2}\right)^{D/2}. \tag{1.260}$$

It is very easy to see that

$$\lim_{D\to\infty} \left(\frac{1}{D}\right)^{D/2} = 0 \tag{1.261}$$

$$\lim_{D\to\infty} \left(\frac{1}{2}\right)^{D/2} = 0, \tag{1.262}$$

and by the squeeze theorem we can conclude that $\lim_{D\to\infty} \left(\frac{2e\pi}{D}\right)^{D/2} = 0$. Hence we can see that the ratio in (1.254) goes to zero when $D \to \infty$.

The center of the hypercube is the origin. It is obvious that one of the vertices of the hypercube is $\overbrace{(-a, -a, \ldots, -a)}^{\text{D } a\text{'s}}{}^{\mathsf{T}}$. Therefore the distance between it and the origin is $a\sqrt{D}$. A hyperplane in D-dimensional space can be constructed by $D-1$ vectors. Let us consider one of the planes where one side of the hypercube lies on, which is spanned by the following group of $D-1$ vectors

$$\begin{cases} (2a, 0, \ldots, 0, 0)^{\mathsf{T}} \\ (0, 2a, \ldots, 0, 0)^{\mathsf{T}} \\ (0, 0, \ldots, 2a, 0)^{\mathsf{T}} \end{cases} \tag{1.263}$$

It is not hard to find one of its normal vector $\boldsymbol{n} = (0, 0, \ldots, 0, 1)^{\mathsf{T}}$. The distance between the origin and the plain, which equals to the distance between the origin and the side that lies on the plain is

$$\left| \left(\mathbf{0} - (-a, -a, \ldots, -a)^{\mathsf{T}} \right) \bullet \boldsymbol{n} \right| = a, \tag{1.264}$$

where $(\bullet)$ denotes dot product. Therefore we know the ratio is

$$\frac{\text{distance from origin to a corner}}{\text{distance from origin to a side}} = \frac{a\sqrt{D}}{a} = \sqrt{D}. \tag{1.265}$$

When $D \to \infty$, the ratio (1.254) goes to zero, which means the volume of the insphere of the hypercube is significantly smaller to the volume of the hypercube itself. Also, the ratio (1.265) goes to infinity, which indicates that in high dimensions, the distance from the center of the hypercube to a corner is significantly greater to the distance from the center of the hypercube to a side. That is to say, more volume is concentrated around the corners, which is right 'above' the insphere. It is very imaginative to describe this phenomenon as 'spikes'.

Comment. Question **1.18** and **1.19** may have little to do with machine learning, but it is still worth the effort to derive the formulae and study the property of high dimensionality. One very useful lesson we can learn from these questions is that due to the oddities of high dimensions, our intuition in low dimensions will almost always fail. That is to say, when we need to reason a problem using the properties of high dimensionality (even the simplest), instead of trusting our intuition, we should always refer to mathematical tools to verify them.

1.20 In this exercise, we explore the behavior of the Gaussian distribution in high-dimensional spaces. Consider a Gaussian distribution in D dimensions given by

$$p(\boldsymbol{x}) = \frac{1}{(2\pi\sigma^2)^{D/2}} \exp\left(-\frac{\|\boldsymbol{x}\|^2}{2\sigma^2}\right).$$

We wish to find the density with respect to radius in polar coordinates in which the direction variables have been integrated out. To do this, show that the integral of the probability density over a thin shell of radius r and thickness ϵ, where $|\epsilon| \ll 1$, is given by $p(r)\epsilon$ where

$$p(r) = \frac{S_D r^{D-1}}{(2\pi\sigma^2)^{D/2}} \exp\left(\frac{-r^2}{2\sigma^2}\right)$$

where S_D is the surface area of a unit sphere in D dimensions. Show that the function $p(r)$ has a single stationary point located, for large D, at $\hat{r} \simeq \sqrt{D}\sigma$. By considering $p(\hat{r}) + \epsilon$ where $|\epsilon| \ll \hat{r}$, show that for large D,

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

which shows that $\hat{r}$ is a maximum of the radial probability density and also that $p(r)$ decays exponentially away from its maximum at $\hat{r}$ with length scale σ. We have already seen that $|\sigma| \ll \hat{r}$ for large D, and so we see that most of the probability mass is concentrated in a thin shell at large radius. Finally, show that the probability density $p(\boldsymbol{x})$ is larger at the origin than at the radius $\hat{r}$ by a factor of $\exp(D/2)$. We therefore see that most of the probability mass in a high-dimensional Gaussian distribution is located at a different radius from the origin of high probability density. This property of distributions in spaces of high dimensionality will have important consequences when we consider Bayesian inference

of model parameters in later chapters.

Answer . It can be seen that $\boldsymbol{x} = (x_1, x_2, \ldots, x_D)^{\mathsf{T}}$. The normalization property of a probability distribution ensures that

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p(\boldsymbol{x})\, dx_1 \cdots\, dx_D = 1. \tag{1.266}$$

To transform from Cartesian coordinates to spherical coordinates, we use the results from exercise **1.18**, namely

$$dx_1 \cdots\, dx_D = r^{D-1}\, dr\, d\Omega, \tag{1.267}$$

where Ω denotes the set of all angular variables. It can be seen that $d\Omega$ is the surface element of D-dimensional sphere. Directly substituting the variables in $p(\boldsymbol{x})$ into spherical coordinates to get $\tilde{p}(r)$, we get

$$\tilde{p}(r) = \frac{1}{(2\pi\sigma^2)^{D/2}} \exp\left(-\frac{r^2}{2\sigma^2}\right). \tag{1.268}$$

By transforming (1.266), we have

$$\int_0^{\infty} \int_{\odot} \tilde{p}(r) r^{D-1}\, dr\, d\Omega = 1, \tag{1.269}$$

where $\odot$ denotes the domain of a unit D-sphere. We can further derive that

$$\int_0^{\infty} \int_{\odot} \tilde{p}(r) r^{D-1}\, dr\, d\Omega \tag{1.270}$$

$$= \int_{\odot} d\Omega \int_0^{\infty} \tilde{p}(r) r^{D-1}\, dr \tag{1.271}$$

$$= S_D \int_0^{\infty} \tilde{p}(r) r^{D-1}\, dr = 1. \tag{1.272}$$

That is to say, the normalized probability density function $p(r)$ is

$$p(r) = S_D r^{D-1} \tilde{p}(r) \tag{1.273}$$

$$= \frac{S_D r^{D-1}}{(2\pi\sigma^2)^{D/2}} \exp\left(-\frac{r^2}{2\sigma^2}\right). \tag{1.274}$$

To derive the stationary point of the function, we compute

$$\frac{dp(r)}{dr} = \frac{S_D}{(2\pi\sigma^2)^{D/2}} \left[(D-1) r^{D-2} \exp\left(-\frac{r^2}{2\sigma^2}\right) - r^{D-1} \exp\left(-\frac{r^2}{2\sigma^2}\right) \frac{r}{\sigma^2}\right] \tag{1.275}$$

$$= \frac{S_D}{(2\pi\sigma^2)^{D/2}} \exp\left(-\frac{r^2}{2\sigma^2}\right) r^{D-2} \left[(D-1) - \frac{r^2}{\sigma^2}\right]. \tag{1.276}$$

All terms are nonnegative except for the last one, which is surrounded by square brackets. When $r \geq 0$, there is only one zero for $\left[(D-1) - \frac{r^2}{\sigma^2}\right]$, namely $\hat{r} = \sigma\sqrt{D-1}$. In the range $[0, \hat{r}]$, $p(r)$ is increasing; in the range $(\hat{r}, +\infty)$, $p(r)$ is decreasing. Therefore $p(\hat{r})$ is the global maximum of function $p(r)$. It can be seen that when D is large, $\hat{r} \simeq \sigma\sqrt{D}$.

↑

Consider the expression

$$\frac{p(\hat{r}+\epsilon)}{p(\hat{r})} = \frac{\frac{S_D(\hat{r}+\epsilon)^{D-1}}{(2\pi\sigma^2)^{D/2}} \exp\left(-\frac{(\hat{r}+\epsilon)^2}{2\sigma^2}\right)}{\frac{S_D\hat{r}^{D-1}}{(2\pi\sigma^2)^{D/2}} \exp\left(-\frac{\hat{r}^2}{2\sigma^2}\right)} \tag{1.277}$$

$$= \frac{(\hat{r}+\epsilon)^{D-1} \exp\left(-\frac{(\hat{r}+\epsilon)^2}{2\sigma^2}\right)}{\hat{r}^{D-1} \exp\left(-\frac{\hat{r}^2}{2\sigma^2}\right)} \tag{1.278}$$

$$= \left(1+\frac{\epsilon}{\hat{r}}\right)^{D-1} \exp\left(-\frac{(\hat{r}+\epsilon)^2}{2\sigma^2} + \frac{\hat{r}^2}{2\sigma^2}\right) \tag{1.279}$$

$$= \left(1+\frac{\epsilon}{\hat{r}}\right)^{D-1} \exp\left(-\frac{2\epsilon\hat{r}+\epsilon^2}{2\sigma^2}\right) \tag{1.280}$$

$$= \exp\left[-\frac{2\epsilon\hat{r}+\epsilon^2}{2\sigma^2} + (D-1)\ln\left(1+\frac{\epsilon}{\hat{r}}\right)\right]. \tag{1.281}$$

When D is large, $D \simeq (D-1)$. The Maclaurin series of $\ln(x+1)$ suggests that

$$\ln(x+1) = x - \frac{x^2}{2} + O(x^3). \tag{1.282}$$

Combining these facts, we have

$$\exp\left[-\frac{2\epsilon\hat{r}+\epsilon^2}{2\sigma^2} + (D-1)\ln\left(1+\frac{\epsilon}{\hat{r}}\right)\right] \tag{1.283}$$

$$\simeq \exp\left[-\frac{2\epsilon\hat{r}+\epsilon^2}{2\sigma^2} + D\left(\frac{\epsilon}{\hat{r}} - \frac{\epsilon^2}{2\hat{r}^2}\right)\right]. \tag{1.284}$$

Substitute $\hat{r}$ by $\sigma\sqrt{D}$, we have

$$\exp\left[-\frac{2\epsilon\hat{r}+\epsilon^2}{2\sigma^2} + D\left(\frac{\epsilon}{\hat{r}} - \frac{\epsilon^2}{2\hat{r}^2}\right)\right] \tag{1.285}$$

$$= \exp\left[-\frac{2\epsilon\sigma\sqrt{D}+\epsilon^2}{2\sigma^2} + D\left(\frac{\epsilon}{\sigma\sqrt{D}} - \frac{\epsilon^2}{2\sigma^2 D}\right)\right] \tag{1.286}$$

$$= \exp\left(-\frac{\epsilon\sqrt{D}}{\sigma} - \frac{\epsilon^2}{2\sigma^2} + \frac{\epsilon\sqrt{D}}{\sigma} - \frac{\epsilon^2}{2\sigma^2}\right) \tag{1.287}$$

$$= \exp\left(-\frac{\epsilon^2}{\sigma^2}\right). \tag{1.288}$$

Therefore we can conclude that $p(\hat{r}+\epsilon) = p(\hat{r})\exp\left(-\frac{\epsilon^2}{\sigma^2}\right)$, which indicates that $p(r)$ decays exponentially around $\hat{r}$.

The mode of the distribution under Cartesian coordinates is

$$p(\boldsymbol{x})|_{\boldsymbol{x}=\boldsymbol{0}} = \frac{1}{(2\pi\sigma^2)^{D/2}}. \tag{1.289}$$

The density at radius $\hat{r}$ is

$$p(\boldsymbol{x})|_{\|\boldsymbol{x}\|=\hat{r}} = \frac{1}{(2\pi\sigma^2)^{D/2}} \exp\left(-\frac{D}{2}\right). \tag{1.290}$$

Because $D \gg 1$, we can see that the mode of the distribution is greater than the density at $\hat{r}$ by a factor of $\exp\left(\frac{D}{2}\right)$. However, the analysis of $p(r)$ suggests that the majority of the probability mass is actually concentrated around $\hat{r}$, where the density is significantly smaller. This is obviously, another oddity when it comes to high dimensionality.

1.21 Consider two nonnegative numbers a and b, and show that, if $a \leq b$, then $a \leq (ab)^{\frac{1}{2}}$. Use this result to show that, if the decision regions of a two-class classification problem are chosen to minimize the probability of misclassification, this probability will satisfy

$$p(\text{mistake}) \leq \int \{p(\boldsymbol{x}, \mathcal{C}_1)p(\boldsymbol{x}, \mathcal{C}_2)\}^{1/2} \, d\boldsymbol{x}.$$

Answer . Because a, b are nonnegative and $a \leq b$, it can be seen that $a^2 \leq ab$. Applying square root on both sides yields

$$a \leq (ab)^{1/2}. \tag{1.291}$$

We know that for a two-class classification problem, the probability of misclassification is

$$\begin{aligned} p(\text{mistake}) &= p(\boldsymbol{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\boldsymbol{x} \in \mathcal{R}_2, \mathcal{C}_1) && (1.292)\\ &= \int_{\mathcal{R}_1} p(\boldsymbol{x}, \mathcal{C}_2) \, d\boldsymbol{x} + \int_{\mathcal{R}_2} p(\boldsymbol{x}, \mathcal{C}_1) \, d\boldsymbol{x}. && (1.293) \end{aligned}$$

Because the decision regions are chosen to minimize $p(\text{mistake})$, in $\mathcal{R}_1$ we should have $p(\boldsymbol{x}, \mathcal{C}_2) \leq p(\boldsymbol{x}, \mathcal{C}_1)$; in $\mathcal{R}_2$ we should have $p(\boldsymbol{x}, \mathcal{C}_1) < p(\boldsymbol{x}, \mathcal{C}_2)$. Using the inequality above, we have

$$\begin{aligned} p(\text{mistake}) &= \int_{\mathcal{R}_1} p(\boldsymbol{x}, \mathcal{C}_2) \, d\boldsymbol{x} + \int_{\mathcal{R}_2} p(\boldsymbol{x}, \mathcal{C}_1) \, d\boldsymbol{x} && (1.294)\\ &\leq \int_{\mathcal{R}_1} \{p(\boldsymbol{x}, \mathcal{C}_1)p(\boldsymbol{x}, \mathcal{C}_2)\}^{1/2} + \int_{\mathcal{R}_2} \{p(\boldsymbol{x}, \mathcal{C}_1)p(\boldsymbol{x}, \mathcal{C}_2)\}^{1/2} && (1.295)\\ &= \int \{p(\boldsymbol{x}, \mathcal{C}_1)p(\boldsymbol{x}, \mathcal{C}_2)\}^{1/2} \, d\boldsymbol{x}. && (1.296) \end{aligned}$$

1.22 Given a loss matrix with elements L_{kj}, the expected risk is minimized if, for each $\boldsymbol{x}$, we choose the class that minimizes

$$\sum_k L_{kj} p(\mathcal{C}_k \mid \boldsymbol{x}).$$

Verify that, when the loss matrix is given by $L_{kj} = 1 - I_{kj}$, where I_{kj} are the elements of the identity matrix, this reduces to the criterion of choosing the class having the largest posterior probability. What is the interpretation of this form of loss matrix?

Answer . It can be seen that $L = \mathbf{1} - I$, that is to say

$$L_{kj} = \begin{cases} 0, & k = j \\ 1, & \text{otherwise.} \end{cases} \tag{1.297}$$

We know that $\sum_k p(\mathcal{C}_k \mid \boldsymbol{x}) = 1$. Given a fixed j, it can be seen that

$$\sum_k L_{kj} p(\mathcal{C}_k \mid \boldsymbol{x}) = 1 - p(\mathcal{C}_j \mid \boldsymbol{x}). \tag{1.298}$$

That is to say, we wish to find the j such that the quantity $1 - p(\mathcal{C}_j \mid \boldsymbol{x})$ is the smallest. It is equivalent to finding the largest $p(\mathcal{C}_j \mid \boldsymbol{x})$, namely the posterior probability.

This form of loss matrix assigns identical loss (here, 1) for any type of misclassification, while giving 0 loss for correct classifications.

1.23 Derive the criterion for minimizing the expected loss when there is a general loss matrix and general prior probabilities for the classes.

Answer . The minimal-loss criterion in terms of posterior probabilities writes

$$\sum_k L_{kj} p(\mathcal{C}_k \mid \boldsymbol{x}). \tag{1.299}$$

Using the Bayes' theorem, it can be seen that

$$\sum_k L_{kj} p(\mathcal{C}_k \mid \boldsymbol{x}) = \sum_k L_{kj} \frac{p(\boldsymbol{x} \mid \mathcal{C}_k) p(\mathcal{C}_k)}{p(\boldsymbol{x})} \tag{1.300}$$

$$= \frac{1}{p(\boldsymbol{x})} \sum_k L_{kj} p(\boldsymbol{x} \mid \mathcal{C}_k) p(\mathcal{C}_k). \tag{1.301}$$

1.24 Consider a classification problem in which the loss incurred when an input vector from class $\mathcal{C}_k$ is classified as belonging to class $\mathcal{C}_j$ is given by the loss matrix L_{kj}, and for which the loss incurred in the selecting the rejection option is λ. Find the decision criterion that will give the minimum expected loss. Verify that this reduces to the reject criterion discussed in Section 1.5.3 where the loss matrix is given by $L_{kj} = 1 - I_{kj}$. What is the relationship between λ and the rejection threshold θ?

Answer . Taking the rejection option into account, it can be seen that the expected loss now writes

$$\mathbb{E}[L] = \begin{cases} \lambda, & \text{if a rejection is made,} \\ \sum_k \sum_j \int_{R_j} L_{kj} p(\mathcal{C}_k \mid \boldsymbol{x}) \, d\boldsymbol{x}, & \text{otherwise,} \end{cases} \tag{1.302}$$

where λ is a manually determined parameter to determine the trade-off between committing to a class or rejection. We know that in general classification problems, the optimal decision is to assign $\boldsymbol{x}$ to class j' that minimizes

$$\sum_k L_{kj'} p(\mathcal{C}_k \mid \boldsymbol{x}), \tag{1.303}$$

which is equivalent to minimizing the second equation in (1.302). The value of (1.303) is exactly the loss incurred by this decision.

In order to minimize the expected loss, we choose the smaller quantity between (1.303) and λ, that is, we choose to commit to $\mathcal{C}'_j$, if $\sum_k L_{kj'} p(\mathcal{C}_k \mid \boldsymbol{x}) \leq \lambda$. Otherwise, we choose to reject.

In question **1.22**, we have concluded that if the loss matrix is given by $L_{kj} = \mathbf{1} - I_{kj}$, the loss of committing to class $\mathcal{C}_j$ is given by $1 - p(\mathcal{C}_j \mid \boldsymbol{x})$, and that the decision rule is equivalent to finding j' such that $p(\mathcal{C}'_j \mid \boldsymbol{x})$ is the greatest. Combined with the rejection option, the decision principle now can be expressed as follows: we choose to commit to $\mathcal{C}'_j$ if $1 - p(\mathcal{C}'_j \mid \boldsymbol{x}) \leq \lambda$. Otherwise, we choose to reject.

That is to say, we would choose a class if and if only

$$1 - \max_j \left[p(\mathcal{C}_j \mid \boldsymbol{x}) \right] \leq \lambda \tag{1.304}$$

$$\max_j \left[p(\mathcal{C}_j \mid \boldsymbol{x}) \right] \geq 1 - \lambda, \tag{1.305}$$

which is identical to setting the rejection threshold θ to be $1 - \lambda$. Therefore we can conclude that $\theta = 1 - \lambda$.

1.25 Consider the generalization of the squared loss function

$$\mathbb{E}[L] = \int \int \{y(\boldsymbol{x}) - t\}^2 \, p(\boldsymbol{x}, t) \, d\boldsymbol{x} \, dt$$

for a single variable t to the case of multiple target variables described by the vector $\boldsymbol{t}$ given by

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

Using the calculus of variations, show that the function $\boldsymbol{y}(\boldsymbol{x})$ for which this expected loss is minimized is given by $\boldsymbol{y}(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{t}}[\boldsymbol{t} \mid \boldsymbol{x}]$. Show that this reduces to the case of a single target variable t.

Answer . It is important to point out that the expressions of the squared loss function are definite integrals. Because their boundaries are trivial, they are not pointed out explicitly. In the following discussion, we also ignore the boundaries of integral signs, if possible. But keep in mind that all integrals are definite integrals with specific boundaries.

For simplicity, let the functional $\boldsymbol{f}(\boldsymbol{y})$ be $\boldsymbol{f}[\boldsymbol{y}(\boldsymbol{x})] = \mathbb{E}[L(\boldsymbol{t}, \boldsymbol{y}(\boldsymbol{x}))]$. Denote the functional $\boldsymbol{\Phi}(\epsilon)$ by

$$\boldsymbol{\Phi}(\epsilon) = \boldsymbol{f}(\boldsymbol{y} + \epsilon \boldsymbol{\eta}), \tag{1.306}$$

where ϵ is a small variation, and $\boldsymbol{\eta}$ is a flexible function that vanishes (equals to zero) at the boundaries of the integrals, then for any ϵ close to 0. If $\boldsymbol{y}$ is a local minima of $\boldsymbol{f}$, the derivative around $\boldsymbol{y}$ should be close to zero. Therefore it is intuitive to see

$$\left.\frac{\partial \boldsymbol{\Phi}}{\partial \epsilon}\right|_{\epsilon=0} = \mathbf{0}. \tag{1.307}$$

Let us prove a fact that is vital for our proof. If the one-dimensional integral satisfies

$$\int_a^b \eta(x)h(x)\,dx = 0, \tag{1.308}$$

where $\eta(a) = \eta(b) = 0$, then we must have $h(x) \equiv 0$ on (a, b). We can prove this claim by contradiction. Assume that $h(x) \not\equiv 0$ on (a, b). Create a function $\zeta(x)$ that satisfies $\zeta(a) = \zeta(b) = 0$, and that $\zeta(x) > 0$ on (a, b). Let $h(x) = \zeta(x)$. Now we have

$$\int_a^b \zeta(x)h^2(x)\,dx = 0. \tag{1.309}$$

Because $h^2(x)$ is nonnegative, it must be true that $h(x) \equiv 0$ on (a, b). This can very easily generalize to our vectorized case. That is to say, if

$$\int \boldsymbol{\eta}(\boldsymbol{x})^{\mathsf{T}}\boldsymbol{h}(\boldsymbol{x})\,d\boldsymbol{x} = \mathbf{0}, \tag{1.310}$$

then we can conclude that $\boldsymbol{h}(\boldsymbol{x}) \equiv \mathbf{0}$.

By the rules of differentiation under the integral sign, it can be seen that

$$\frac{\partial \boldsymbol{\Phi}}{\partial \epsilon} = \frac{\partial}{\partial \epsilon}\int\int \|\boldsymbol{y}(\boldsymbol{x}) + \epsilon\boldsymbol{\eta} - \boldsymbol{t}\|^2 p(\boldsymbol{x}, \boldsymbol{t})\,d\boldsymbol{x}\,d\boldsymbol{t} \tag{1.311}$$

$$= \int\int \frac{\partial}{\partial \epsilon}\|\boldsymbol{y}(\boldsymbol{x}) + \epsilon\boldsymbol{\eta} - \boldsymbol{t}\|^2 p(\boldsymbol{x}, \boldsymbol{t})\,d\boldsymbol{t}\,d\boldsymbol{x}. \tag{1.312}$$

We know that

$$\|\boldsymbol{y}(\boldsymbol{x}) + \epsilon\boldsymbol{\eta} - \boldsymbol{t}\|^2 = \|\boldsymbol{y}(\boldsymbol{x}) + \epsilon\boldsymbol{\eta}\|^2 - [\boldsymbol{y}(\boldsymbol{x}) + \epsilon\boldsymbol{\eta}]^{\mathsf{T}}\boldsymbol{t} - \boldsymbol{t}^{\mathsf{T}}[\boldsymbol{y}(\boldsymbol{x}) + \epsilon\boldsymbol{\eta}] + \boldsymbol{t}^2, \tag{1.313}$$

and given the fact all these matrix multiplications above are yielding single numbers (1×1 matrices), which indicates that all products are equal to themselves transposed, we can derive that

$$\frac{\partial}{\partial \epsilon}\|\boldsymbol{y}(\boldsymbol{x}) + \epsilon\boldsymbol{\eta} - \boldsymbol{t}\|^2 p(\boldsymbol{x}, \boldsymbol{t}) = \left[2\boldsymbol{\eta}^{\mathsf{T}}\boldsymbol{y}(\boldsymbol{x}) - 2\boldsymbol{\eta}^{\mathsf{T}}\boldsymbol{t} + 2\epsilon\boldsymbol{\eta}^2\right] p(\boldsymbol{x}, \boldsymbol{t}). \tag{1.314}$$

Putting the condition (1.307) into consideration, we have

$$\int\int \frac{\partial}{\partial \epsilon} \|\boldsymbol{y}(\boldsymbol{x}) + \epsilon\boldsymbol{\eta} - \boldsymbol{t}\|^2 \, p(\boldsymbol{x}, \boldsymbol{t}) \, d\boldsymbol{t} \, d\boldsymbol{x} \tag{1.315}$$

$$= \int\int \left[2\boldsymbol{\eta}^{\mathsf{T}}\boldsymbol{y}(\boldsymbol{x}) - 2\boldsymbol{\eta}^{\mathsf{T}}\boldsymbol{t} + 2\epsilon\boldsymbol{\eta}^2\right] p(\boldsymbol{x}, \boldsymbol{t}) \, d\boldsymbol{t} \, d\boldsymbol{x} \tag{1.316}$$

$$= \int \boldsymbol{\eta}^{\mathsf{T}} \int 2\left[\boldsymbol{y}(\boldsymbol{x}) - \boldsymbol{t}\right] p(\boldsymbol{x}, \boldsymbol{t}) \, d\boldsymbol{t} \, d\boldsymbol{x}. \tag{1.317}$$

By (1.310) we know that

$$\int 2\left[\boldsymbol{y}(\boldsymbol{x}) - \boldsymbol{t}\right] p(\boldsymbol{x}, \boldsymbol{t}) \, d\boldsymbol{t} = \boldsymbol{0}. \tag{1.318}$$

It can be seen that

$$\int 2\left[\boldsymbol{y}(\boldsymbol{x}) - \boldsymbol{t}\right] p(\boldsymbol{x}, \boldsymbol{t}) \, d\boldsymbol{t} = \boldsymbol{0} \tag{1.319}$$

$$\int \boldsymbol{y}(\boldsymbol{x}) p(\boldsymbol{x}, \boldsymbol{t}) \, d\boldsymbol{t} = \int \boldsymbol{t} p(\boldsymbol{x}, \boldsymbol{t}) \, d\boldsymbol{t} \tag{1.320}$$

$$\int \boldsymbol{y}(\boldsymbol{x}) p(\boldsymbol{x}) p(\boldsymbol{t} \mid \boldsymbol{x}) \, d\boldsymbol{t} = \int \boldsymbol{t} p(\boldsymbol{x}, \boldsymbol{t}) \, d\boldsymbol{t} \tag{1.321}$$

$$\boldsymbol{y}(\boldsymbol{x}) p(\boldsymbol{x}) = \int \boldsymbol{t} p(\boldsymbol{x}, \boldsymbol{t}) \, d\boldsymbol{t} \tag{1.322}$$

$$\boldsymbol{y}(\boldsymbol{x}) = \frac{1}{p(\boldsymbol{x})} \int \boldsymbol{t} p(\boldsymbol{x}, \boldsymbol{t}) \, d\boldsymbol{t} = \int \boldsymbol{t} p(\boldsymbol{t} \mid \boldsymbol{x}) \, d\boldsymbol{t} = \mathbb{E}_{\boldsymbol{t}}\left[\boldsymbol{t} \mid \boldsymbol{x}\right]. \tag{1.323}$$

In one-dimensional case, it is just equivalent to $\mathbb{E}_t\left[t \mid x\right]$.

Comment. The derivation of calculus of variation is actually provided in Appendix D of the textbook, which was completely ignored by me when I was solving this question. Next time we can directly make use of the Euler-Lagrange equations.

1.26 By expansion of the square in

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

derive a result analogous to

$$\mathbb{E}\left[L\right] = \int \left\{y(\boldsymbol{x}) - \mathbb{E}\left[t \mid \boldsymbol{x}\right]\right\}^2 p(\boldsymbol{x}) \, d\boldsymbol{x} + \int \mathrm{var}\left[t \mid \boldsymbol{x}\right] p(\boldsymbol{x}) \, d\boldsymbol{x}$$

and hence show that the function $\boldsymbol{y}(\boldsymbol{x})$ that minimizes the expected squared loss for the case of a vector $\boldsymbol{t}$ of target variables is again given by the conditional expectation of $\boldsymbol{t}$.

Answer . The two cases are very similar. It can be seen that

$$\|\boldsymbol{y}(\boldsymbol{x}) - \boldsymbol{t}\|^2 = \|\boldsymbol{y}(\boldsymbol{x}) - \mathbb{E}\left[\boldsymbol{t} \mid \boldsymbol{x}\right] + \mathbb{E}\left[\boldsymbol{t} \mid \boldsymbol{x}\right] - \boldsymbol{t}\|^2 \tag{1.324}$$

$$
\begin{aligned}
&= \|\boldsymbol{y}(\boldsymbol{x}) - \mathbb{E}\left[\boldsymbol{t} \mid \boldsymbol{x}\right]\|^2 \\
&\quad + \left\{\boldsymbol{y}(\boldsymbol{x}) - \mathbb{E}\left[\boldsymbol{t} \mid \boldsymbol{x}\right]\right\}^{\mathsf{T}} \left\{\mathbb{E}\left[\boldsymbol{t} \mid \boldsymbol{x}\right] - \boldsymbol{t}\right\} \\
&\quad + \left\{\mathbb{E}\left[\boldsymbol{t} \mid \boldsymbol{x}\right] - \boldsymbol{t}\right\}^{\mathsf{T}} \left\{\boldsymbol{y}(\boldsymbol{x}) - \mathbb{E}\left[\boldsymbol{t} \mid \boldsymbol{x}\right]\right\} \\
&\quad + \|\mathbb{E}\left[\boldsymbol{t} \mid \boldsymbol{x}\right] - \boldsymbol{t}\|^2 .
\end{aligned} \tag{1.325}
$$

Because the cross terms are essentially numbers (1×1 matrices), they equal to their own transpose. Therefore, the two cross terms are equal to each other, and we only consider the first term. In order to deal with the cross terms, consider the expectation $\mathbb{E}_{x,t}\left[g\right]$. It can be seen that

$$
\begin{aligned}
\mathbb{E}_{x,t}\left[g\right] &= \int\int g(\boldsymbol{x}, \boldsymbol{t}) p(\boldsymbol{x}, \boldsymbol{t}) \, d\boldsymbol{x} \, d\boldsymbol{t} && (1.326)\\
&= \int p(\boldsymbol{x}) \, d\boldsymbol{x} \int g(\boldsymbol{x}, \boldsymbol{t}) p(\boldsymbol{t} \mid \boldsymbol{x}) \, d\boldsymbol{t} && (1.327)\\
&= \int \mathbb{E}_{\boldsymbol{t}}\left[g \mid \boldsymbol{x}\right] p(\boldsymbol{x}) \, d\boldsymbol{x} && (1.328)\\
&= \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_{\boldsymbol{t}}\left[g \mid \boldsymbol{x}\right]\right] . && (1.329)
\end{aligned}
$$

For the cross term, we have

$$
\begin{aligned}
&\int\int \left\{\boldsymbol{y}(\boldsymbol{x}) - \mathbb{E}\left[\boldsymbol{t} \mid \boldsymbol{x}\right]\right\}^{\mathsf{T}} \left\{\mathbb{E}\left[\boldsymbol{t} \mid \boldsymbol{x}\right] - \boldsymbol{t}\right\} p(\boldsymbol{x}, \boldsymbol{t}) \, d\boldsymbol{x} \, d\boldsymbol{t} && (1.330)\\
&= \mathbb{E}_{\boldsymbol{x},\boldsymbol{t}}\left[\left\{\boldsymbol{y}(\boldsymbol{x}) - \mathbb{E}\left[\boldsymbol{t} \mid \boldsymbol{x}\right]\right\}^{\mathsf{T}} \left\{\mathbb{E}\left[\boldsymbol{t} \mid \boldsymbol{x}\right] - \boldsymbol{t}\right\}\right] && (1.331)\\
&= \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_{\boldsymbol{t}}\left[\left\{\boldsymbol{y}(\boldsymbol{x}) - \mathbb{E}\left[\boldsymbol{t} \mid \boldsymbol{x}\right]\right\}^{\mathsf{T}} \left\{\mathbb{E}\left[\boldsymbol{t} \mid \boldsymbol{x}\right] - \boldsymbol{t}\right\} \mid \boldsymbol{x}\right]\right] . && (1.332)
\end{aligned}
$$

Using subscript to denote a particular element of a vector, it can be seen that

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_{\boldsymbol{t}}\left[\left\{\boldsymbol{y}(\boldsymbol{x}) - \mathbb{E}\left[\boldsymbol{t} \mid \boldsymbol{x}\right]\right\}^{\mathsf{T}} \left\{\mathbb{E}\left[\boldsymbol{t} \mid \boldsymbol{x}\right] - \boldsymbol{t}\right\} \mid \boldsymbol{x}\right]\right] && (1.333)\\
&= \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_{\boldsymbol{t}}\left[\sum_i \left\{\boldsymbol{y}(\boldsymbol{x}) - \mathbb{E}\left[\boldsymbol{t} \mid \boldsymbol{x}\right]\right\}_i \left\{\mathbb{E}\left[\boldsymbol{t} \mid \boldsymbol{x}\right] - \boldsymbol{t}\right\}_i \mid \boldsymbol{x}\right]\right] && (1.334)\\
&= \mathbb{E}_{\boldsymbol{x}}\left[\sum_i \left\{\boldsymbol{y}(\boldsymbol{x}) - \mathbb{E}\left[\boldsymbol{t} \mid \boldsymbol{x}\right]\right\}_i \mathbb{E}_{\boldsymbol{t}}\left[\left\{\mathbb{E}\left[\boldsymbol{t} \mid \boldsymbol{x}\right] - \boldsymbol{t}\right\}_i \mid \boldsymbol{x}\right]\right] && (1.335)\\
&= \mathbb{E}_{\boldsymbol{x}}\left[\sum_i \left\{\boldsymbol{y}(\boldsymbol{x}) - \mathbb{E}\left[\boldsymbol{t} \mid \boldsymbol{x}\right]\right\}_i \left\{\mathbb{E}_{\boldsymbol{t}}\left[\mathbb{E}\left[\boldsymbol{t} \mid \boldsymbol{x}\right]_i \mid \boldsymbol{x}\right] - \mathbb{E}_{\boldsymbol{t}}\left[\boldsymbol{t}_i \mid \boldsymbol{x}\right]\right\}\right] && (1.336)\\
&= \mathbb{E}_{\boldsymbol{x}}\left[\sum_i \left\{\boldsymbol{y}(\boldsymbol{x}) - \mathbb{E}\left[\boldsymbol{t} \mid \boldsymbol{x}\right]\right\}_i \left\{\mathbb{E}_{\boldsymbol{t}}\left[\boldsymbol{t} \mid \boldsymbol{x}\right]_i - \mathbb{E}_{\boldsymbol{t}}\left[\boldsymbol{t}_i \mid \boldsymbol{x}\right]\right\}\right] && (1.337)\\
&= \mathbb{E}_{\boldsymbol{x}}\left[\sum_i \left\{\boldsymbol{y}(\boldsymbol{x}) - \mathbb{E}\left[\boldsymbol{t} \mid \boldsymbol{x}\right]\right\}_i \cdot 0\right] && (1.338)\\
&= 0. && (1.339)
\end{aligned}
$$

Therefore we have proved that the two cross term vanish after integration. We can now write

$$\mathbb{E}\left[L(\boldsymbol{t}, \boldsymbol{y}(\boldsymbol{x}))\right] = \int\int \|\boldsymbol{y}(\boldsymbol{x}) - \boldsymbol{t}\|^2 p(\boldsymbol{x}, \boldsymbol{t})\, d\boldsymbol{x}\, d\boldsymbol{t} \tag{1.340}$$

$$= \int\int \|\boldsymbol{y}(\boldsymbol{x}) - \mathbb{E}\left[\boldsymbol{t} \mid \boldsymbol{x}\right]\|^2 p(\boldsymbol{x}, \boldsymbol{t})\, d\boldsymbol{x}\, d\boldsymbol{t} + \int\int \|\mathbb{E}\left[\boldsymbol{t} \mid \boldsymbol{x}\right] - \boldsymbol{t}\|^2 p(\boldsymbol{x}, \boldsymbol{t})\, d\boldsymbol{x}\, d\boldsymbol{t} \tag{1.341}$$

$$\begin{aligned} = &\int p(\boldsymbol{t} \mid \boldsymbol{x})\, d\boldsymbol{t} \int \|\boldsymbol{y}(\boldsymbol{x}) - \mathbb{E}\left[\boldsymbol{t} \mid \boldsymbol{x}\right]\|^2 p(\boldsymbol{x})\, d\boldsymbol{x} \\ &+ \int \|\boldsymbol{t} - \mathbb{E}\left[\boldsymbol{t} \mid \boldsymbol{x}\right]\|^2 p(\boldsymbol{t} \mid \boldsymbol{x})\, d\boldsymbol{t} \int p(\boldsymbol{x})\, d\boldsymbol{x} \end{aligned} \tag{1.342}$$

$$= \int \|\boldsymbol{y}(\boldsymbol{x}) - \mathbb{E}\left[\boldsymbol{t} \mid \boldsymbol{x}\right]\|^2 p(\boldsymbol{x})\, d\boldsymbol{x} + \int \operatorname{var}\left[\boldsymbol{t} \mid \boldsymbol{x}\right] p(\boldsymbol{x})\, d\boldsymbol{x} \tag{1.343}$$

$$= \int \|\boldsymbol{y}(\boldsymbol{x}) - \mathbb{E}\left[\boldsymbol{t} \mid \boldsymbol{x}\right]\|^2 p(\boldsymbol{x})\, d\boldsymbol{x} + \operatorname{var}\left[\boldsymbol{t} \mid \boldsymbol{x}\right] \tag{1.344}$$

The second term does not depend on the choice of $\boldsymbol{y}$, while selecting

$$\boldsymbol{y}(\boldsymbol{x}) = \mathbb{E}\left[\boldsymbol{t} \mid \boldsymbol{x}\right] \tag{1.345}$$

minimizes the first term. Therefore we can conclude that (1.345) minimizes the loss function.

Comment. Here we are using the conclusion $\mathbb{E}_{\boldsymbol{t}}\left[\boldsymbol{t} \mid \boldsymbol{x}\right]_i = \mathbb{E}_{\boldsymbol{t}}\left[\boldsymbol{t}_i \mid \boldsymbol{x}\right]$ without proving explicitly.

1.27 Consider the expected loss for regression problems under the L_q loss function given by

$$\mathbb{E}\left[L_q\right] = \int\int |y(\boldsymbol{x}) - t|^q p(\boldsymbol{x}, t)\, d\boldsymbol{x}\, dt.$$

Write down the condition that $y(\boldsymbol{x})$ must satisfy in order to minimize $\mathbb{E}\left[L_q\right]$. Show that, for $q = 1$, this solution represents the conditional median, i.e., the function $y(\boldsymbol{x})$ such that the probability mass for $t < y(\boldsymbol{x})$ is the same as for $t \geq y(\boldsymbol{x})$. Also show that the minimum expected L_q loss for $q \to 0$ is given by the conditional mode, i.e. by the function $y(\boldsymbol{x})$ equal to the value of t that maximizes $p(t \mid \boldsymbol{x})$ for each $\boldsymbol{x}$.

Answer . Our goal is to find expressions of $y(\boldsymbol{x})$ in order to minimize $\mathbb{E}\left[L_q\right]$. It can be seen that

$$\mathbb{E}\left[L_q\right] = \int\int |y(\boldsymbol{x}) - t|^q p(\boldsymbol{x}, t)\, dt\, d\boldsymbol{x} \tag{1.346}$$

$$= \int p(\boldsymbol{x}) \int |y(\boldsymbol{x}) - t|^q p(t \mid \boldsymbol{x})\, dt\, d\boldsymbol{x} \tag{1.347}$$

All quantities in the integral above are nonnegative, and we can try to minimize the integrand, namely

$$\int |y(\boldsymbol{x}) - t|^q p(t \mid \boldsymbol{x})\, dt. \tag{1.348}$$

The optimal $y(\boldsymbol{x})$ can be found by setting the partial derivative with respect to itself to be zero. We know that $|a| = \left(x^2\right)^{1/2}$,

therefore we have

$$\frac{\partial}{\partial y(\boldsymbol{x})} \int |y(\boldsymbol{x}) - t|^q p(t \mid \boldsymbol{x}) \, dt \tag{1.349}$$

$$= \frac{\partial}{\partial y(\boldsymbol{x})} \int \left\{ [y(\boldsymbol{x}) - t]^2 \right\}^{\frac{q}{2}} p(t \mid \boldsymbol{x}) \, dt \tag{1.350}$$

$$= \int \frac{q}{2} \left\{ [y(\boldsymbol{x}) - t]^2 \right\}^{\frac{q}{2}-1} \cdot 2 \left[y(\boldsymbol{x}) - t \right] p(t \mid \boldsymbol{x}) \, dt \tag{1.351}$$

$$= \int q \left\{ [y(\boldsymbol{x}) - t]^2 \right\}^{\frac{q-1}{2}} \cdot \frac{y(\boldsymbol{x}) - t}{|y(\boldsymbol{x}) - t|} p(t \mid \boldsymbol{x}) \, dt \tag{1.352}$$

$$= q \int |y(\boldsymbol{x}) - t|^{q-1} \cdot \operatorname{sgn} \left(y(\boldsymbol{x}) - t \right) p(t \mid \boldsymbol{x}) \, dt \tag{1.353}$$

$$= q \int_{-\infty}^{y(\boldsymbol{x})} [y(\boldsymbol{x}) - t]^{q-1} \, p(t \mid \boldsymbol{x}) \, dt - q \int_{y(\boldsymbol{x})}^{+\infty} [y(\boldsymbol{x}) - t]^{q-1} \, p(t \mid \boldsymbol{x}) \, dt = 0. \tag{1.354}$$

Therefore we can see that the stationary condition for $y(\boldsymbol{x})$ is

$$q \int_{-\infty}^{y(\boldsymbol{x})} [y(\boldsymbol{x}) - t]^{q-1} \, p(t \mid \boldsymbol{x}) \, dt - q \int_{y(\boldsymbol{x})}^{+\infty} [y(\boldsymbol{x}) - t]^{q-1} \, p(t \mid \boldsymbol{x}) \, dt = 0 \tag{1.355}$$

$$\int_{-\infty}^{y(\boldsymbol{x})} [y(\boldsymbol{x}) - t]^{q-1} \, p(t \mid \boldsymbol{x}) \, dt = \int_{y(\boldsymbol{x})}^{+\infty} [y(\boldsymbol{x}) - t]^{q-1} \, p(t \mid \boldsymbol{x}) \, dt. \tag{1.356}$$

When $q = 1$, the condition reduces to

$$\int_{-\infty}^{y(\boldsymbol{x})} p(t \mid \boldsymbol{x}) \, dt = \int_{y(\boldsymbol{x})}^{+\infty} p(t \mid \boldsymbol{x}) \, dt, \tag{1.357}$$

which indicates that $y(\boldsymbol{x})$ is the conditional median of t.

When $q \to 0$, we need some other method to derive the minimizer. Consider the limit

$$\lim_{q \to 0} a^q = 1, \tag{1.358}$$

where $a \neq 0$. However, the function $|y(\boldsymbol{x}) - t|^q$ does not equal to zero unless for some region where $y(\boldsymbol{x}) = t$. Suppose that this happens only when $t = t_0$, and on other regions we should always have $|y(\boldsymbol{x}) - t| \ > 0$. During the integration, we can exclude a small vicinity around t_0, which is denoted by $(t_0 - \tau, t_0 + \tau)$, where $\tau > 0$. Under this setting, it can be seen that

$$\lim_{q \to 0} \int_{-\infty}^{+\infty} |y(\boldsymbol{x}) - t|^q p(t \mid \boldsymbol{x}) \, dt \tag{1.359}$$

$$\begin{aligned} &= \lim_{q \to 0} \int_{-\infty}^{t_0 - \tau} |y(\boldsymbol{x}) - t|^q p(t \mid \boldsymbol{x}) \, dt \\ &\quad + \lim_{q \to 0} \int_{t_0 - \tau}^{t_0 + \tau} |y(\boldsymbol{x}) - t|^q p(t \mid \boldsymbol{x}) \, dt \\ &\quad + \lim_{q \to 0} \int_{t_0 + \tau}^{+\infty} |y(\boldsymbol{x}) - t|^q p(t \mid \boldsymbol{x}) \, dt \end{aligned} \tag{1.360}$$

$$= \int_{-\infty}^{t_0 - \tau} p(t \mid \boldsymbol{x}) \, dt + \int_{t_0 + \tau}^{+\infty} p(t \mid \boldsymbol{x}) \, dt + \lim_{q \to 0} \int_{t_0 - \tau}^{t_0 + \tau} |y(\boldsymbol{x}) - t|^q p(t \mid \boldsymbol{x}) \, dt \tag{1.361}$$

$$= 1 - \int_{t_0-\tau}^{t_0+\tau} p(t \mid \boldsymbol{x})\, dt + \lim_{q\to 0} \int_{t_0-\tau}^{t_0+\tau} |y(\boldsymbol{x}) - t|^q p(t \mid \boldsymbol{x})\, dt. \tag{1.362}$$

It is intuitive that we can select τ such that

$$|y(\boldsymbol{x}) - t| < 1 \tag{1.363}$$

holds on $(t_0 - \tau, t_0 + \tau)$. Under proper choice of τ (which can be arbitrarily small), if we select a q' that is very close to zero, we would have

$$\lim_{q\to 0} \int_{-\infty}^{+\infty} |y(\boldsymbol{x}) - t|^q p(t \mid \boldsymbol{x})\, dt \tag{1.364}$$

$$= 1 - \int_{t_0-\tau}^{t_0+\tau} p(t \mid \boldsymbol{x})\, dt + \lim_{q\to 0} \int_{t_0-\tau}^{t_0+\tau} |y(\boldsymbol{x}) - t|^q p(t \mid \boldsymbol{x})\, dt \tag{1.365}$$

$$\approx 1 - \int_{t_0-\tau}^{t_0+\tau} p(t \mid \boldsymbol{x})\, dt + \int_{t_0-\tau}^{t_0+\tau} |y(\boldsymbol{x}) - t|^{q'} p(t \mid \boldsymbol{x})\, dt. \tag{1.366}$$

By the mean theorem of definite integrals, we can see that there exists $t_1 \in (t_0 - \tau, t_0 + \tau)$ that satisfies

$$\int_{t_0-\tau}^{t_0+\tau} |y(\boldsymbol{x}) - t|^{q'} p(t \mid \boldsymbol{x})\, dt = |y(\boldsymbol{x}) - t_1|^{q'} \int_{t_0-\tau}^{t_0+\tau} p(t \mid \boldsymbol{x})\, dt. \tag{1.367}$$

This allows us to write

$$1 - \int_{t_0-\tau}^{t_0+\tau} p(t \mid \boldsymbol{x})\, dt + \int_{t_0-\tau}^{t_0+\tau} |y(\boldsymbol{x}) - t|^{q'} p(t \mid \boldsymbol{x})\, dt \tag{1.368}$$

$$= 1 - \int_{t_0-\tau}^{t_0+\tau} p(t \mid \boldsymbol{x})\, dt + |y(\boldsymbol{x}) - t_1|^{q'} \int_{t_0-\tau}^{t_0+\tau} p(t \mid \boldsymbol{x})\, dt \tag{1.369}$$

$$= 1 - (1 - |y(\boldsymbol{x}) - t_1|^{q'}) \int_{t_0-\tau}^{t_0+\tau} p(t \mid \boldsymbol{x})\, dt. \tag{1.370}$$

Denote the constant $1 - |y(\boldsymbol{x}) - t_1|^{q'}$ by k, it can be seen that

$$0 \leq k < 1. \tag{1.371}$$

Now we can write

$$\lim_{q\to 0} \int_{-\infty}^{+\infty} |y(\boldsymbol{x}) - t|^q p(t \mid \boldsymbol{x})\, dt \tag{1.372}$$

$$\approx 1 - k \int_{t_0-\tau}^{t_0+\tau} p(t \mid \boldsymbol{x})\, dt. \tag{1.373}$$

Using the mean value theorem again, we know that there exists $t' \in (t_0 - \tau, t_0 + \tau)$ that satisfies

$$\int_{t_0-\tau}^{t_0+\tau} p(t \mid \boldsymbol{x})\, dt = 2\tau p(t' \mid \boldsymbol{x}), \tag{1.374}$$

↑

which allows us to conclude that

$$\lim_{q\to 0}\int_{-\infty}^{+\infty}|y(\boldsymbol{x})-t|^q p(t\mid\boldsymbol{x})\,dt\approx 1-2k\tau p(t'\mid\boldsymbol{x}). \tag{1.375}$$

Because τ can be arbitrarily small, we can assume that $t'\approx t_0$. In order to minimize (1.359), we would like to choose the t' such that $p(t'\mid\boldsymbol{x})$ is the greatest. It is equivalent to choosing a $y(\boldsymbol{x})$ such that $p(t_0\mid\boldsymbol{x})$ is the greatest. Therefore we can say that the conditional mode is the minimizer at this point.

Comment. I tried very hard to solve the case when $q\to 0$, but I am still not satisfied with the proof. Because given $0<q_1<q_2$, if we consider the ratio

$$\frac{|y(\boldsymbol{x})-t|^{q_1}p(t\mid\boldsymbol{x})}{|y(\boldsymbol{x})-t|^{q_2}p(t\mid\boldsymbol{x})}=|y(\boldsymbol{x})-t|^{q_1-q_2}. \tag{1.376}$$

Under the proper choice of τ, we can see that this value is no less than one. That is to say, (1.375) is actually a lower bound of the original limit, for it is almost always smaller than the limit. In this case, because both q' and τ are very small, I would just assume that the approximation is extremely close to the true value, and that the term $p(t'\mid\boldsymbol{x})$ will affect the value in a much more fundamental way.

1.28 In section 1.6, we introduced the idea of entropy $h(x)$ as the information gained on observing the value of a random variable x having distribution $p(x)$. We saw that, for independent variables x and y for which $p(x,y)=p(x)p(y)$, the entropy functions are additive, so that $h(x,y)=h(x)+h(y)$. In this exercise, we derive the relation between h and p in the form of a function $h(p)$. First show that $h(p^2)=2h(p)$, and hence by induction that $h(p^n)=nh(p)$ where n is a positive integer. Hence show that $h(p^{n/m})=(n/m)h(p)$ where m is also a positive integer. This implies that $h(p^x)=xh(p)$ where x is a positive rational number, and hence by continuity when it is a positive real number. Finally, show that this implies $h(p)$ must take the form $h(p)\propto\ln p$.

Answer . Based on the additive assumption, it can be seen that

$$h(p^2)=h(p)+h(p)=2h(p). \tag{1.377}$$

We can prove $h(p^n)=nh(p)$ for positive integer by induction.

Basis: $h(p^1)=1\cdot h(p)$ holds.

Induction Hypothesis: for any $k>0$, $h(p^k)=kh(p)$ holds.

Induction Step: by the additive assumption, it can be seen that $h(p^{k+1})=h(p^k\cdot p)=kh(p)+h(p)=(k+1)h(p)$, which indicates that the claim holds for $k+1$.

By induction, we can conclude that the claim holds true for any positive integer.

↑

Now we are proving the second claim. It can be seen that $p^n = \left\{p^{n/m}\right\}^m$. Therefore we can write

$$nh(p) \equiv h(p^n) = h\left(\left\{p^{n/m}\right\}^m\right) \equiv mh(p^{n/m}). \tag{1.378}$$

This equation yields $h(p^{n/m}) = \frac{n}{m}h(p)$. That is to say $h(p^x) = xh(p)$ where x is a positive rational number, and hence by continuity when it is a positive real number.

Given a positive real number p, for arbitrary positive real number q, we can always find a real number s such that $p = q^s$. It can be seen that

$$\frac{h(p)}{ln(p)} = \frac{h(q^s)}{\ln q^s} = \frac{sh(q)}{s\ln q} = \frac{h(q)}{\ln q}, \tag{1.379}$$

which is a constant. Therefore we can conclude that $h(p) \propto \ln p$.

Comment. In the book, we try to deduce the form of the information content by defining $h(x, y) = h(x) + h(y)$ for *independent* variables x and y. In this question, it is extended to the case of $h(p^2) = h(p) + h(p) = 2h(p)$. However, we know that p and p itself are not independent, so this conclusion might be *wrong*. I am not sure why we can generalize from an independent case to a dependent case. Nonetheless if the form of information content is indeed logarithmic, everything problem will vanish.

1.29 Consider an M-state discrete random variable x, and use Jensen's inequality in the form

$$f\left(\sum_{i=1}^{M}\lambda_i x_i\right) \leq \sum_{i=1}^{M}\lambda_i f(x_i) \quad (f \text{ is convex})$$

to show that the entropy of this distribution $p(x)$ satisfies $H[x] \leq \ln M$.

Answer . We know that

$$\mathrm{H}[x] = -\sum_{i=1}^{M} p(x_i)\ln p(x_i) = \sum_{i=1}^{M} p(x_i)\ln\frac{1}{p(x_i)}. \tag{1.380}$$

In the form of Jensen's inequality, it can be seen that $f(x_i) = \ln x_i$. We can easily derive that $f''(x_i) = -1/x_i^2$, which is nonpositive. That is to say, f is a concave function. Also, it is clear that $p(x_i) \geq 0$ and $\sum_{i=1}^{M} p(x_i) = 1$, which meets the requirements of the inequality. Because the function is concave, when using Jensen's inequality, we need to reverse it, namely t use the form

$$f\left(\sum_{i=1}^{M}\lambda_i x_i\right) \geq \sum_{i=1}^{M}\lambda_i f(x_i) \quad (f \text{ is concave}). \tag{1.381}$$

Now it can be seen that

$$\ln\left(\sum_{i=1}^{M} p(x_i)\frac{1}{p(x_i)}\right) \geq \sum_{i=1}^{M} p(x_i)\ln\frac{1}{p(x_i)}, \tag{1.382}$$

↑

which indicates that $\mathrm{H}[x] \leq \ln M$.

1.30 Evaluate the Kullback-Leibler divergence

$$\mathrm{KL}(p \,||\, q) = -\int p(x) \ln \left\{ \frac{q(x)}{p(x)} \right\} dx$$

between two Gaussians $p(x) = \mathcal{N}(x \mid \mu, \sigma^2)$ and $q(x) = \mathcal{N}(x \mid m, s^2)$.

Answer . The density function of p is given by

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2} \right], \tag{1.383}$$

the density function of q shares a similar form. Putting them into the KL divergence formula, we have

$$\begin{aligned}
\mathrm{KL}(p \,||\, q) &= -\int p(x) \ln \left\{ \frac{\frac{1}{\sqrt{2\pi}s} \exp\left[-\frac{(x-m)^2}{2s^2} \right]}{\frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2} \right]} \right\} dx && (1.384)\\
&= -\int p(x) \ln \left\{ \frac{\sigma}{s} \exp\left[\frac{(x-\mu)^2}{2\sigma^2} - \frac{(x-m)^2}{2s^2} \right] \right\} dx && (1.385)\\
&= -\int p(x) \left\{ \ln \frac{\sigma}{s} + \frac{(x-\mu)^2}{2\sigma^2} - \frac{(x-m)^2}{2s^2} \right\} dx && (1.386)\\
&= -\int p(x) \ln \frac{\sigma}{s} \, dx \\
&\quad - \int p(x) \frac{(x-\mu)^2}{2\sigma^2} \, dx && (1.387)\\
&\quad + \int p(x) \frac{(x-m)^2}{2s^2} \, dx \\
&= -\mathbb{E}\left[\ln \frac{\sigma}{s} \right] - \frac{1}{2\sigma^2} \mathrm{var}\,[x] + \frac{1}{2s^2} \mathbb{E}\left[x^2 - 2mx + m^2 \right]. && (1.388)
\end{aligned}$$

All expectation here are with respect to the distribution of p. Given the fact that $\mathbb{E}\left[x^2\right] = \mu^2 + \sigma^2$ for Gaussians, we have

$$\begin{aligned}
&-\mathbb{E}\left[\ln \frac{\sigma}{s} \right] - \frac{1}{2\sigma^2} \mathrm{var}\,[x] + \frac{1}{2s^2} \mathbb{E}\left[x^2 - 2mx + m^2 \right] && (1.389)\\
&= -\ln \frac{\sigma}{s} - \frac{\sigma^2}{2\sigma^2} + \frac{1}{2s^2} \left[\mu^2 + \sigma^2 - 2m\mu + m^2 \right] && (1.390)\\
&= \ln \frac{s}{\sigma} + \frac{1}{2s^2} \left[\mu^2 + \sigma^2 - 2m\mu + m^2 \right] - \frac{1}{2}. && (1.391)
\end{aligned}$$

Therefore we can conclude that

$$\mathrm{KL}(p \,||\, q) = \ln \frac{s}{\sigma} + \frac{1}{2s^2} \left[\mu^2 + \sigma^2 - 2m\mu + m^2 \right] - \frac{1}{2}. \tag{1.392}$$

1.31 Consider two variables $\boldsymbol{x}$ and $\boldsymbol{y}$ having joint distribution $p(\boldsymbol{x}, \boldsymbol{y})$. Show that the differential entropy of this pair of

variables satisfies

$$\mathrm{H}[\boldsymbol{x}, \boldsymbol{y}] \leq \mathrm{H}[\boldsymbol{x}] + \mathrm{H}[\boldsymbol{y}] \tag{1.393}$$

with equality if, and only if, $\boldsymbol{x}$ and $\boldsymbol{y}$ are statistically independent.

Answer . Recall that mutual information of a distribution $p(\boldsymbol{x}, \boldsymbol{y})$ (denoted by $\mathrm{I}[\boldsymbol{x}, \boldsymbol{y}]$) is essentially the KL divergence between the joint distribution and the product of the marginals, which means it is always greater than or equal to zero. The equality stands only when $\boldsymbol{x}$ and $\boldsymbol{y}$ are statistically independent. We also know that

$$\mathrm{I}[\boldsymbol{x}, \boldsymbol{y}] = \mathrm{H}[\boldsymbol{y}] - \mathrm{H}[\boldsymbol{y} \mid \boldsymbol{x}] \geq 0. \tag{1.394}$$

That is to say, $\mathrm{H}[\boldsymbol{y}] \geq \mathrm{H}[\boldsymbol{y} \mid \boldsymbol{x}]$, where the equality holds only when $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent.

According to the property of entropy, we know that

$$\mathrm{H}[\boldsymbol{x}, \boldsymbol{y}] = \mathrm{H}[\boldsymbol{y} \mid \boldsymbol{x}] + \mathrm{H}[\boldsymbol{x}] \leq \mathrm{H}[\boldsymbol{y}] + \mathrm{H}[\boldsymbol{x}]. \tag{1.395}$$

The equality holds only when $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent.

1.32 Consider a vector $\boldsymbol{x}$ of continuous variables with distribution $p(\boldsymbol{x})$ and corresponding entropy $\mathrm{H}[\boldsymbol{x}]$. Suppose that we make a nonsingular linear transformation of $\boldsymbol{x}$ to obtain a new variable $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$. Show that the corresponding entropy is given by $\mathrm{H}[\boldsymbol{y}] = \mathrm{H}[\boldsymbol{x}] + \ln|\boldsymbol{A}|$ where $|\boldsymbol{A}|$ denotes the determinant of $\boldsymbol{A}$.

Answer . Assume that both $\boldsymbol{x}$ and $\boldsymbol{y}$ are N-dimensional vector. Applying the change of variables over density functions, we have

$$p(\boldsymbol{x}) = p(\boldsymbol{y}) \left| \det \left[\frac{\partial(y_1, \ldots, y_N)}{\partial(x_1, \ldots, x_N)} \right] \right| = p(\boldsymbol{y}) |\det(\boldsymbol{A})| \tag{1.396}$$

$$p(\boldsymbol{y}) = |\det(\boldsymbol{A})|^{-1} p(\boldsymbol{x}), \tag{1.397}$$

where $|\cdot|$ denotes absolute value. Because $\boldsymbol{A}$ is nonsingular, we know that $\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{y}$. The Jacobian of this transformation (denoted by $\boldsymbol{J}$) writes

$$\boldsymbol{J} = \det \left[\frac{\partial(x_1, \ldots, x_N)}{\partial(y_1, \ldots, y_N)} \right] = \det(\boldsymbol{A}^{-1}) = \det(\boldsymbol{A})^{-1}. \tag{1.398}$$

We can now see that $d\boldsymbol{x} = |\det(\boldsymbol{J})|\, d\boldsymbol{y} = |\det(\boldsymbol{A})|^{-1}\, d\boldsymbol{y}$. It can be seen that

$$\mathrm{H}[\boldsymbol{y}] = -\int p(\boldsymbol{y}) \ln p(\boldsymbol{y})\, d\boldsymbol{y} \tag{1.399}$$

$$= -\int p(\boldsymbol{x}) \ln \left[|\det(\boldsymbol{A})|^{-1} p(\boldsymbol{x}) \right] |\det(\boldsymbol{A})|^{-1}\, d\boldsymbol{y} \tag{1.400}$$

$$= -\int p(\boldsymbol{x}) \ln \left[|\det(\boldsymbol{A})|^{-1} p(\boldsymbol{x}) \right] \, d\boldsymbol{x} \tag{1.401}$$

$$= -\int p(\boldsymbol{x}) \ln p(\boldsymbol{x}) \, d\boldsymbol{x} - \int p(\boldsymbol{x}) \ln |\det(\boldsymbol{A})|^{-1} \, d\boldsymbol{x} \tag{1.402}$$

$$= \mathrm{H}[\boldsymbol{x}] - \ln |\det(\boldsymbol{A})|^{-1} \tag{1.403}$$

$$= \mathrm{H}[\boldsymbol{x}] + \ln |\det(\boldsymbol{A})|. \tag{1.404}$$

Therefore we can conclude that

$$\mathrm{H}[\boldsymbol{y}] = \mathrm{H}[\boldsymbol{x}] + \ln |\det(\boldsymbol{A})|. \tag{1.405}$$

1.33 Suppose that the conditional entropy $\mathrm{H}[y \mid x]$ between two discrete random variables x and y is zero. Show that, for all values of x such that $p(x) > 0$, the variable y must be a function of x, in other words for each x there is only one value of y such that $p(y \mid x) \neq 0$.

Answer . By the definition of entropy, we have

$$\mathrm{H}[y \mid x] = -\sum_i \sum_j p(x_i, y_j) \ln p(y_j \mid x_i) \tag{1.406}$$

$$= -\sum_i \sum_j p(x_i) p(y_j \mid x_i) \ln p(y_j \mid x_i) \tag{1.407}$$

$$= \sum_i \sum_j p(x_i) p(y_j \mid x_i) \ln \frac{1}{p(y_j \mid x_i)}. \tag{1.408}$$

By L'Hôpital's rule, it can be seen that

$$\lim_{x \to 0} x \ln \frac{1}{x} = \lim_{x \to 0} -x \ln x = -\lim_{x \to 0} \frac{\ln x}{\frac{1}{x}} = \lim_{x \to 0} \frac{\frac{1}{x}}{\frac{1}{x^2}} = \lim_{x \to 0} x = 0. \tag{1.409}$$

Given the fact that $0 \leq p(y_j \mid x_i) \leq 1$, we can conclude that $p(y_j \mid x_i) \ln \frac{1}{p(y_j|x_i)} \geq 0$, with equality holds only when $p(y_j \mid x_i) = 0$ or $p(y_j \mid x_i) = 1$.

Because (1.408) equals to zero, and that every quantity in the product is nonnegative, we know that for those is where $p(x_i) > 0$, we would have $p(y_j \mid x_i) \ln \frac{1}{p(y_j|x_i)} = 0$. Therefore in this case, we would either have $p(y_j \mid x_i) = 0$ or $p(y_j \mid x_i) = 1$. However, $p(y_j \mid x_i)$ is a probability distribution that satisfies

$$\sum_j p(y_j \mid x_i) = 1. \tag{1.410}$$

That is to say, for each x_i that has $p(x_i) > 0$, there exists one and only one j such that $p(y_j \mid x_i) = 1$, while for all other j values we have $p(y_j \mid x_i) = 0$.