Pattern Recognition and Machine Learning
Cristopher Bishop

Exercise Solutions

Stefan Stefanache

April 27, 2021

Chapter 1

Introduction

Exercise 1.1 ⋆

Consider the sum-of-squares error function given by (1.2) in which the function $y(x, \mathbf{w})$ is given by the polynomial (1.1). Show that the coefficients $\mathbf{w} = \{w_i\}$ that minimize this error function are given by the solution to the following set of linear equations

$$\sum_{j=0}^{M} A_{ij} w_j = T_i \tag{1.122}$$

where

$$A_{ij} = \sum_{n=1}^{N} (x_n)^{i+j}, \qquad T_i = \sum_{n=1}^{N} (x_n)^i t_n. \tag{1.123}$$

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

Proof. The function $y(x, \mathbf{w})$ is given by

$$y(x, \mathbf{w}) = \sum_{j=0}^{M} w_j x^j \tag{1.1}$$

and the error function is given by

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 \tag{1.2}$$

Since we want to find the coefficients $\mathbf{w}$ for which the error function is minimized, we compute its derivative with respect to $\mathbf{w}$:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\mathbf{w}} E(\mathbf{w}) =& \frac{\mathrm{d}}{\mathrm{d}\mathbf{w}} \left(\frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 \right) = \frac{1}{2} \sum_{n=1}^{N} \frac{\mathrm{d}}{\mathrm{d}\mathbf{w}} \{y(x_n, \mathbf{w})^2 - 2 t_n y(x_n, \mathbf{w}) + t_n^2\} \\ =& \sum_{n=1}^{N} y(x_n, \mathbf{w}) \frac{\mathrm{d}}{\mathrm{d}\mathbf{w}} y(x_n, \mathbf{w}) - \sum_{n=1}^{N} t_n \frac{\mathrm{d}}{\mathrm{d}\mathbf{w}} y(x_n, \mathbf{w}) \end{aligned} \tag{1.1.1}$$

We continue by computing the derivative of $y(x_n, \mathbf{w})$ separately and obtain that:

$$\frac{\mathrm{d}}{\mathrm{d}\mathbf{w}} y(x_n, \mathbf{w}) = \begin{bmatrix} \frac{\mathrm{d}}{\mathrm{d}w_0} y(x_n, \mathbf{w}) \\ \frac{\mathrm{d}}{\mathrm{d}w_1} y(x_n, \mathbf{w}) \\ \vdots \\ \frac{\mathrm{d}}{\mathrm{d}w_M} y(x_n, \mathbf{w}) \end{bmatrix} = \begin{bmatrix} x_n^0 \\ x_n^1 \\ \vdots \\ x_n^M \end{bmatrix} \tag{1.1.2}$$

By substituting the result of (1.1.2) into (1.1.1) we get that:

$$\frac{\mathrm{d}}{\mathrm{d}\mathbf{w}} E(\mathbf{w}) = B - T \tag{1.1.3}$$

where T is given by (1.123) and

$$B_i = \sum_{n=1}^{N} x_n^i y(x_n, \mathbf{w})$$

Now, we easily find that

$$B_i = \sum_{n=1}^{N} \left(x_n^i \sum_{j=0}^{M} w_j x_n^j \right) = \sum_{n=1}^{N} \sum_{j=0}^{M} x_n^{i+j} w_j = A_i \mathbf{w}$$

where A is given by (1.123). Now, the critical point of $E(\mathbf{w})$ is given by the equation:

$$A_i \mathbf{w} = T_i$$

which is equivalent with (1.122). □

Exercise 1.2 ⋆

Write down the set of coupled linear equations, analogous to (1.122), satisfied by the coefficients w_i which minimize the regularized sum-of-squares error function given by (1.4).

Proof. The regularized sum-of-squares error function is given by

$$\widetilde{E}(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2} ||\mathbf{w}||^2 \tag{1.4}$$

We'll have a similar approach to the previous exercise, i.e. we compute the derivative of the regularized error function and find the associated critical point. We notice that

$$\widetilde{E}(\mathbf{w}) = E(\mathbf{w}) + \frac{\lambda}{2} ||\mathbf{w}||^2$$

so

$$\frac{\mathrm{d}}{\mathrm{d}\mathbf{w}} \widetilde{E}(\mathbf{w}) = \frac{\mathrm{d}}{\mathrm{d}\mathbf{w}} E(\mathbf{w}) + \frac{\lambda}{2} \cdot \frac{\mathrm{d}}{\mathrm{d}\mathbf{w}} ||\mathbf{w}||^2$$

One could easily prove that

$$\frac{\mathrm{d}}{\mathrm{d}\mathbf{w}}||\mathbf{w}||^2 = 2\mathbf{w}$$

so by using this and (1.1.3) (where we substitute $B = A\mathbf{w}$), we have that:

$$\frac{\mathrm{d}}{\mathrm{d}\mathbf{w}}\widetilde{E}(\mathbf{w}) = A\mathbf{w} + \lambda\mathbf{w} - T = (A + \lambda I)\mathbf{w} - T$$

We obtain the critical point when the derivative is 0, so when

$$(A + \lambda I)\mathbf{w} = T$$

which is equivalent with the system of linear equations

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

where

$$C_{ij} = A_{ij} + \lambda I_{ij}$$

□

Exercise 1.3 ⋆⋆

Suppose that we have three coloured 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 limes, 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 picee 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 selected fruit is in fact an orange, what is the probability that it came from the green box?

Proof. The conditional probabilities of obtaining a fruit knowing that we are searching in a certain box are easily found since the fruits are equally likely to be extracted. We also now the probabilities of choosing a specific box, so we can simply apply the sum rule to obtain the probability of getting an apple:

$$p(\text{apple}) = p(\text{apple}|r)p(r) + p(\text{apple}|b)p(b) + p(\text{apple}|g)p(g) = \frac{3}{10}\cdot 0.2 + \frac{1}{2}\cdot 0.2 + \frac{3}{10}\cdot 0.6 = 34\%$$

If we know the selected fruit is an orange, the probability that it came from the green box is given by the Bayes' theorem:

$$p(g|\text{orange}) = \frac{p(g)p(\text{orange}|g)}{p(\text{orange})} \tag{1.3.1}$$

The probability of choosing the green box is known and the probability of getting an orange from the green box is also easily found. We only need to find the probability of extracting an orange in the general case:

$$p(\text{orange}) = p(\text{orange}|r)p(r) + p(\text{orange}|b)p(b) + p(\text{orange}|g)p(g) = \frac{4}{10}\cdot 0.2 + \frac{1}{2}\cdot 0.2 + \frac{3}{10}\cdot 0.6 = 36\%$$

The needed probability is now found by substituting the values in (1.3.1):

$$p(g|\text{orange}) = \frac{0.6 \cdot \frac{3}{10}}{\frac{36}{100}} = \frac{1}{2} = 50\%$$

□

Exercise 1.4 ⋆⋆

Consider a probability density $p_x(x)$ 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 (1.27). By differentiating (1.27), show that the location $\widehat{y}$ of the maximum of the density in y is not in general related to the location $\widehat{x}$ of the maximum of the density over x by the simple functional relation $\widehat{x} = g(\widehat{y})$ as a consequence of the Jacobian factor. This shows that the maximum of a probability density (in contrast to a simple function) is dependent of 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.

Proof. If we make a nonlinear change of variable $x = g(y)$ in the probbability density $p_x(x)$, it transforms according to

$$p_y(y) = p_x(g(y))|g'(y)| \tag{1.27}$$

We assume that the mode of $p_x(x)$ is given by an unique $\widehat{x}$, i.e.

$$p'_x(x) = 0 \iff x = \widehat{x}$$

Now, let $s \in \{-1, 1\}$ such that $g'(y) = sg'(y)$. The derivative of (1.27) with respect to y is given by:

$$p'_y(y) = sp'x(g(y))\{g'(y)\}^2 + sp_x(g(y))g''(y)$$

For a linear change of variable, we have that $g''(y) = 0$, so the mode of $p_y(y)$ is given by $g'(y) = 0$ and since $x = g(y)$, respectively $x' = g'(y)$ we have that $\widehat{x} = g(\widehat{y})$. Therefore, for a linear change of variable, the location of the maximum transforms in the same way as the variable itself.

For a nonlinear change of variable, the second derivative will not be generally 0, so the mode is not given by $g'(y) = 0$ anymore. As a result, in general $\widehat{x} \neq g(\widehat{y})$, so the location of the mode will transform differently from the variable itself. □

Exercise 1.5 ⋆

Using the definition (1.38) show that $\text{var}[f(x)]$ satisfies (1.39).

Proof. The variance is defined by

$$\text{var}[f] = \mathbb{E}\big[(f(x) - \mathbb{E}[f(x)])^2\big] \tag{1.38}$$

We expand the square and then use the linearity of expectation to obtain:

$$\text{var}[f] = \mathbb{E}\big[f(x)^2 - 2f(x)\mathbb{E}[f(x)] + \mathbb{E}[f(x)]^2\big] = \mathbb{E}[f(x)^2] - 2\mathbb{E}\big[f(x)\mathbb{E}[f(x)]\big] + \mathbb{E}\big[\mathbb{E}[f(x)]^2\big]$$

Since $\mathbb{E}[f(x)]$ is a constant, the expression of the variance becomes:

$$\text{var}[f] = \mathbb{E}[f(x)^2] - 2\mathbb{E}[f(x)]^2 + \mathbb{E}[f(x)]^2 = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2 \tag{1.39}$$

□

Exercise 1.6 ⋆

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

Proof. The covariance of two random variables is given by:

$$\text{cov}[x, y] = \mathbb{E}_{x,y}[xy] - E[x]E[y] \tag{1.41}$$

We assume that the variables are continuous, but the discrete case result is similarly obtained. If x and y are independent, we have that $p(x, y) = p(x)p(y)$, so

$$E_{x,y}[xy] = \iint p(x, y)xy\, dxdy = \iint p(x)p(y)xy\, dxdy = \left(\int p(x)x\, dx \right)\left(\int p(y)y\, dy \right) = E[x]E[y]$$

and (1.41) becomes 0. □

Exercise 1.7 ⋆⋆

In this exercise, we prove the normalization condition (1.48) for the univariate Gaussian. To do this consider the integral

$$I = \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}x^2 \right) dx \tag{1.124}$$

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

$$I^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}x^2 - \frac{1}{2\sigma^2}y^2 \right) dxdy \tag{1.125}$$

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 takinng the square root of both sides, we obtain

$$I = (2\pi\sigma^2)^{1/2} \tag{1.126}$$

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

Proof. We transform (1.125) from Cartesian coordinates to polar coordinates and obtain:

$$I^2 = \int_0^{2\pi}\int_0^{\infty} \exp\left(-\frac{r^2\sin^2\theta + r^2\cos^2\theta}{2\sigma^2} \right) r dr d\theta = \int_0^{2\pi}\int_0^{\infty} \exp\left(-\frac{r^2}{2\sigma^2} \right) r dr d\theta$$

We use the substitution $u = r^2$ and then compute the integral to get:

$$I^2 = \frac{1}{2}\int_0^{2\pi}\int_0^{\infty} \exp\left(-\frac{u}{2\sigma^2} \right) du d\theta = \frac{1}{2}\int_0^{2\pi} -2\sigma^2 \exp\left(-\frac{u}{2\sigma^2} \right)\Big|_0^{\infty} d\theta = \sigma^2 \int_0^{2\pi} d\theta = 2\pi\sigma^2$$

If we take the square root of this we see that

$$I = (2\pi\sigma^2)^{1/2} \tag{1.126}$$

We can assume without loss of generality that the mean of the Gaussian is 0, as we could make the change of variable $y = x - \mu$. Therefore, by using (1.126) we obtain

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{x}{2\sigma^2} \right) dx = \frac{I}{\sqrt{2\pi\sigma^2}} = 1$$

which shows that the Gaussian distribution is normailized. □

Exercise 1.8 ⋆⋆

By using a change of variables, verify that the univariate Gaussian given by (1.46) satisfies (1.49). Next, by differentiating both sides of the normalization condition

$$\int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) dx = 1 \tag{1.127}$$

with respect to σ^2, verify that the Gaussian satisfies (1.50). Finally, show that (1.51) holds.

Proof. We start by computing the expected value of the Gaussian:

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) x \, \mathrm{d}x = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\} x \, \mathrm{d}x$$

We do a little trick to prepare for the substitution $u = (x - \mu)^2$:

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

Since the Gaussian is normalized, the second term of the expression will be μ. By using the substitution $u = (x - \mu)^2$, the expected value becomes:

$$\mathbb{E}[x] = \frac{1}{2\sqrt{2\pi\sigma^2}} \int_{\infty}^{\infty} \exp\left(-\frac{u}{2\sigma^2} \right) \mathrm{d}u + \mu$$

We notice that the endpoints of the integral are "equal" (one could rewrite it as a limit of an integral with actual equal endpoints), so its value is 0. Therefore,

$$\mathbb{E}[x] = \mu \tag{1.49}$$

Now, we take the derivative of (1.127) with respect to σ^2 and obtain:

$$\begin{aligned}
\frac{\partial}{\partial \sigma^2} \left(\frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\} \mathrm{d}x \right) &= 0 \\
-\frac{I}{2\sigma^3\sqrt{2\pi}} + \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \frac{\partial}{\partial \sigma^2} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\} \mathrm{d}x &= 0 \\
-\frac{1}{2\sigma^2} + \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \frac{(x-\mu)^2}{2\sigma^4} \exp\left\{ -\frac{(x+\mu)^2}{2\sigma^2} \right\} \mathrm{d}x &= 0
\end{aligned}$$

We let J be the integral term and compute it separately:

$$J = \frac{1}{2\sigma^4}\int_{-\infty}^{\infty}(x-\mu)^2 \exp\left\{-\frac{(x+\mu)^2}{2\sigma^2}\right\}\mathrm{d}x$$
$$= \frac{1}{2\sigma^4}\int_{-\infty}^{\infty}x^2 \exp\left\{-\frac{(x+\mu)^2}{2\sigma^2}\right\}\mathrm{d}x - \frac{2\mu}{2\sigma^4}\int_{-\infty}^{\infty}x \exp\left\{-\frac{(x+\mu)^2}{2\sigma^2}\right\}\mathrm{d}x + \frac{\mu^2}{2\sigma^4}I$$

If we multiply by the normalization constants, the integrals become expected values and the I factor vanishes. Therefore:

$$J = \sqrt{2\pi\sigma^2}\left(\frac{1}{2\sigma^4}\mathbb{E}[x^2] - \frac{2\mu}{2\sigma^4}\mathbb{E}[x] + \frac{\mu^2}{2\sigma^4}\right)$$

We substitute J back in the initial expression to obtain:

$$-\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbb{E}[x^2] - 2\mu^2 + \mu^2) = 0$$

from which is straightforard to show that

$$E[x^2] = \sigma^2 + \mu^2 \tag{1.50}$$

Finally, one can easily see that:

$$\mathrm{var}[x] = E[x^2] - E[x]^2 = \sigma^2 \tag{1.51}$$

□

Exercise 1.9 ⋆

Show that the mode (i.e. the maximum) of the Gaussian distribution (1.46) is given by μ. Similarly, show that the mode of the multivariate Gaussian (1.52) is given by $\boldsymbol{\mu}$.

Proof. In the univariate case, we start by taking the derivative of (1.46) with respect to x :

$$\frac{\partial}{\partial x}\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\left(\frac{\partial}{\partial x}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}\right) = \frac{1}{\sqrt{2\pi\sigma^2}}\frac{(x-\mu)^2}{2\sigma^4}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

We notice that the derivative is 0, for $x = \mu$, so the mode of the univariate Gaussian is given by the mean.

Analogously, we take the derivative of (1.52) with respect to $\mathbf{x}$ and get:

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

The covariance matrix $\boldsymbol{\Sigma}$ is both nonsingular and symmetric, so one can easily show that $\boldsymbol{\Sigma}^{-1}$ will be symmetric too. Therefore, we have that (see matrix cookbook):

$$\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = 2\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$$

As a result, our derivative becomes

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

and is 0 for $\mathbf{x} = \boldsymbol{\mu}$, so like in the case of the univariate distribution, the mode of the multivariate distribution is given by the mean $\boldsymbol{\mu}$. □

Exercise 1.10 ⋆

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

$$\mathbb{E}[x+z] = \mathbb{E}[x] + \mathbb{E}[z] \tag{1.128}$$

$$\text{var}[x+z] = \text{var}[x] + \text{var}[z] \tag{1.129}$$

Proof. Since the variables are independent, we have that $p(x,z) = p(x)p(z)$. Therefore, by using this, the expression of the expected value and the fact that the distributions are normalized, we have that

$$\begin{aligned}
\mathbb{E}[x+z] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x,z)(x+z)\,\mathrm{d}x\,\mathrm{d}z \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x)p(z)x + p(x)p(z)z\,\mathrm{d}x\,\mathrm{d}z \\
&= \int_{-\infty}^{\infty} p(z)\left(\int_{-\infty}^{\infty} p(x)x\,\mathrm{d}x\right) + p(z)z\left(\int_{-\infty}^{\infty} p(x)\,\mathrm{d}x\right)\mathrm{d}z \\
&= \int_{-\infty}^{\infty} p(z)\mathbb{E}[x] + p(z)z\,\mathrm{d}z \\
&= \mathbb{E}[x]\int_{-\infty}^{\infty} p(z)\,\mathrm{d}z + \int_{-\infty}^{\infty} p(z)z\,\mathrm{d}z \\
&= \mathbb{E}[x] + \mathbb{E}[z]
\end{aligned} \tag{1.128}$$

Analogously, we can solve the discrete case. Now, by using all the available tools, i.e. (1.39) and (1.128), the linearity of the expectation and the independence of variables, we have that the variance of the sum is given by:

$$\begin{aligned}
\text{var}[x+z] &= \mathbb{E}[(x+z)^2] - \mathbb{E}[x+z]^2 = \mathbb{E}[x^2+2xz+z^2] - (\mathbb{E}[x]+\mathbb{E}[z])^2 \\
&= \mathbb{E}[x^2] + 2\mathbb{E}[x]\mathbb{E}[z] + \mathbb{E}[z^2] - \mathbb{E}[x]^2 - \mathbb{E}[x^2+2xz+z^2] - E[z]^2 \\
&= E[x^2] - E[x]^2 + E[z^2] - E[z]^2 \\
&= \text{var}[x] + \text{var}[z]
\end{aligned} \tag{1.129}$$

□

Exercise 1.11 ⋆

By setting the derivatives of the log likelihood function (1.54) with respect to μ and σ^2 equal to zero, verify the results (1.55) and (1.56).

Proof. The log likelihood of the Gaussian is given by:

$$\ln p(\mathbf{x}|\mu,\sigma^2) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi) \tag{1.54}$$

By taking the derivative of (1.54) with respect to μ we get that:

$$\frac{\partial}{\partial \mu} \ln p(\mathbf{x}|\mu, \sigma^2) = -\frac{1}{2\sigma^2}\left\{ \frac{\partial}{\partial \mu} \sum_{n=1}^{N} (x_n - \mu)^2 \right\} = -\frac{1}{2\sigma^2}\left\{ \frac{\partial}{\partial \mu}\left(\sum_{n=1}^{N} x_n^2 - 2\sum_{n=1}^{N} x_n \mu + N\mu^2 \right) \right\}$$
$$= \frac{1}{\sigma^2}\left(\sum_{n=1}^{N} x_n - N\mu \right)$$

which is 0 for the maximum point:

$$\mu_{ML} = \frac{1}{N} \sum_{n=1}^{N} x_n \tag{1.55}$$

Now, we want the variance that maximizes the log likelihood, so we take the derivative of (1.54) (by using μ_{ML}) with respect to σ^2:

$$\frac{\partial}{\partial \sigma^2} \ln p(\mathbf{x}|\mu_{ML}, \sigma^2) = \frac{1}{2\sigma^4} \sum_{n=1}^{N} (x_n - \mu_{ML})^2 - \frac{N}{2\sigma^2} = \frac{1}{2\sigma^4}\left(\sum_{n=1}^{N} (x_n - \mu_{ML})^2 - N\sigma^2 \right)$$

The derivative is 0 for the maximum point

$$\sigma^2_{ML} = \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu_{ML})^2 \tag{1.56}$$

□

Exercise 1.12 ⋆⋆

Using the results (1.49) and (1.50), show that

$$\mathbb{E}[x_n x_m] = \mu^2 + I_{nm}\sigma^2 \tag{1.130}$$

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 the results (1.57) and (1.58).

Proof. We assume that the data points are i.i.d, so we have that the variables x_n and x_m are not independent for $n \neq m$ and independent for $n = m$. Therefore,

$$\mathbb{E}[x_n x_m] = \begin{cases} \mu^2 & n \neq m \\ \mu^2 + \sigma^2 & n = m \end{cases}$$

which is equivalent with (1.130). Now, the expectation of μ_{ML} is given by:

$$\mathbb{E}[\mu_{ML}] = \mathbb{E}\left[\frac{1}{N} \sum_{n=1}^{N} x_n \right] = \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}[x_n] = \mu \tag{1.57}$$

Similarly, the expectation of σ^2_{ML} is given by:

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

We compute each expectation separately and get:

$$E[\mu^2_{ML}] = \frac{1}{N^2}\mathbb{E}\left[\sum_{n=1}^{N}x_n^2 + 2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}x_ix_j\right] = \frac{1}{N^2}\sum_{n=1}^{N}\mathbb{E}[x_n^2] + \frac{2}{N^2}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\mathbb{E}[x_ix_j] = \frac{\sigma^2}{N} + \mu^2$$

$$E[x_n\mu_{ML}] = \frac{1}{N}\mathbb{E}\left[x_n\sum_{i=1}^{N}x_i\right] = \frac{1}{N}(\sigma^2 + N\mu^2) = \frac{\sigma^2}{N} + \mu^2$$

By putting everything together, we obtain

$$\mathbb{E}[\sigma^2_{ML}] = \left(\frac{N-1}{N}\right)\sigma^2 \tag{1.58}$$

□

Exercise 1.13 ⋆

Suppose that the variance of a Gaussian is estimated using the result (1.56) 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.

Proof. Let

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

be the estimator described in the hypothesis. It's straightforward to show that the expectation of the estimator is the actual variance:

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

□

Exercise 1.14 ⋆⋆

Show that an arbitrary square matrix with elements w_{ij} can be written in the form $w_{ij} = w^S_{ij} + w^A_{ij}$ where w^S_{ij} and w^A_{ij} are symmetric and anti-symmetric matrices, respectively, satisfying $w^S_{ij} = w^S_{ji}$

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

$$\sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}x_i x_j \tag{1.131}$$

Show that

$$\sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}x_i x_j = \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}^S x_i x_j \tag{1.132}$$

so that the contribution from the anti-symmetric 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 is given by $D(D+1)/2$.

Proof. If we consider the system of equations

$$w_{ij} = w_{ij}^S + w_{ij}^A \qquad w_{ji} = w_{ij}^S - w_{ij}^A$$

we quickly reach the conclusion that the solutions are given by

$$w_{ij}^S = \frac{w_{ij} + w_{ji}}{2} \qquad w_{ij}^A = \frac{w_{ij} - w_{ji}}{2} \tag{1.14.1}$$

such that for all i and j,

$$w_{ij} = w_{ij}^S + w_{ij}^A$$

The coefficient matrix w associated with the second order higher order polynomial in D dimensions is actually a $D \times D$ *symmetric* matrix. Therefore, from (1.14.1) we'd have that $w^S = w$ and $w_A = 0_D$, where 0_D is the null matrix of dimension D, so (1.132) definitely holds as the anti-symmetric contribution vanishes.

We consider as independent parameters of the matrix w the elements on and above the diagonal, since the ones under the diagonal are reflections of the ones above. There are

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

such independent parameters □

Exercise 1.15 ⋆⋆⋆

In this exercise and the next, we explore how the number of independent parameters in a polynomial grows with the order M of the polynomial and with the dimensionality D of the input space. We start by writing down the M^{th} order term for a polynomial in D dimensions in the form

$$\sum_{i_1=1}^{D}\sum_{i_2=1}^{D}\cdots\sum_{i_M=1}^{D} w_{i_1,i_2,\ldots,i_M}x_{i_1}x_{i_2}\cdot\ldots x_{i_M} \tag{1.133}$$

The coefficients $w_{i_1,i_2,\ldots i_M}$ compromise D^M elements, but the number of independent parameters is significantly fewer due to the many interchange symmetries of the factor $x_{i_1}, x_{i_2}\ldots x_{i_M}$. Begin

by showing that the redundancy in the coefficients can be removed by rewriting the M^{th} order term in the form

$$\sum_{i_1=1}^{D}\sum_{i_2=1}^{i_1}\cdots\sum_{i_M=1}^{i_{M-1}}\widetilde{w}_{i_1,i_2,\ldots,i_M}x_{i_1}x_{i_2}\cdot\ldots x_{i_M} \tag{1.134}$$

Note that the precise relationship between the $\widetilde{w}$ coefficients and 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

$$n(D, M) = \sum_{i=1}^{D} n(i, M-1) \tag{1.135}$$

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

$$\sum_{i=1}^{D}\frac{(i+M-2)!}{(i-1)!(M-1)!} = \frac{(D+M-1)!}{(D-1)!M!} \tag{1.136}$$

which can be done by first proving the result 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 it is correct for dimension $D + 1$. Finally, use the two previous results, together with proof by induction, to show

$$n(D, M) = \frac{(D+M-1)!}{(D-1)!M!} \tag{1.137}$$

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 (1.135), together with (1.136), to show that, if the result holds at order $M - 1$, then it will also hold at order M.

Proof. □

Exercise 1.17 ⋆⋆

The gamma function is defined by

$$\Gamma(x) = \int_0^\infty u^{x-1}e^{-u}\,\mathrm{d}u \tag{1.141}$$

Using 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.

Proof. Knowing that $-u^x e^{-u} \to 0$ as $u \to \infty$, we integrate $\Gamma(x + 1)$ by parts and obtain:

$$\Gamma(x+1) = \int_0^\infty u^x(-e^{-u})'\,\mathrm{d}u = -u^x e^{-u}\Big|_0^\infty + x\int_0^\infty u^{x-1}e^{-u}\,\mathrm{d}u = x\Gamma(x)$$

Computing $\Gamma(1)$ is also easily done by integrating by parts:

$$\Gamma(1) = \int_0^\infty ue^{-u}\,\mathrm{d}u = \int_0^\infty u(-e^{-u})'\,\mathrm{d}u = -ue^{-u}\Big|_0^\infty + \int_0^\infty e^{-u}\,\mathrm{d}u = 1$$

We can prove by induction that $\Gamma(x+1) = x!$ when x is an integer. This is obviously valid for $x = 0$, since $0! = 1$. Now, assume that $\Gamma(k) = (k-1)!$, for $k \in \mathbb{N}$. Then,

$$\Gamma(k+1) = k\Gamma(k) = k \cdot (k-1)! = k!$$

Therefore, $\Gamma(n+1) = n!$ for all $n \in \mathbb{N}$. □

Exercise 1.18 ⋆⋆

We can use the result (1.126) 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

$$\prod_{i=1}^{D} \int_{-\infty}^{\infty} e^{-x_i^2} \, \mathrm{d}x_i = S_D \int_0^{\infty} e^{-r^2} r^{D-1} \, \mathrm{d}r \tag{1.142}$$

Using the definition (1.141) of the Gamma function, together with (1.126), evaluate both sides of this equation, and hence show that

$$S_D = \frac{2\pi^{D/2}}{\Gamma(D/2)} \tag{1.143}$$

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

$$V_D = \frac{S_D}{D} \tag{1.144}$$

Finally, use the results $\Gamma(1) = 1$ and $\Gamma(3/2) = \sqrt{\pi}/2$ to show that (1.143) and (1.144) reduce to the usual expressions for $D = 2$ and $D = 3$.

Proof. We observe that the left side factor of (1.142) looks like (1.126) for $\sigma^2 = 1/2$. Therefore,

$$\prod_{i=1}^{D} \int_{-\infty}^{\infty} e^{-x_i^2} \, \mathrm{d}x_i = \prod_{i=1}^{D} \pi^{1/2} = \pi^{D/2}$$

One can easily notice that the integral in the right side of (1.142) can be written as:

$$\int_0^{\infty} e^{-r^2} r^{D-1} \, \mathrm{d}r = \int_0^{\infty} e^{-r^2} (r^2)^{(D-2)/2} r \, \mathrm{d}r = \frac{1}{2} \int_0^{\infty} e^{-u} u^{(D-2)/2} \, \mathrm{d}u = \frac{1}{2}\Gamma(D/2) \, \mathrm{d}u$$

where we made the substitution $u = r^2$.

Therefore, from those results and from (1.142), we find that

$$S_D = \frac{2\pi^{D/2}}{\Gamma(D/2)} \tag{1.143}$$

The volume of the unit hypersphere is now given by the integral

$$V_D = \int_0^1 S_D r^{D-1} \, \mathrm{d}r = \frac{S_D}{D} \tag{1.144}$$

Now, we get the expected results for $D = 2$ and $D = 3$:

$$S_2 = \frac{2\pi}{\Gamma(1)} = 2\pi \qquad V_2 = \pi \qquad S_3 = \frac{2\pi^{3/2}}{\Gamma(\frac{3}{2})} = 4\pi \qquad V_3 = \frac{4\pi}{3}$$

□

Exercise 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 centres 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

$$\frac{\text{volume of sphere}}{\text{volume of cube}} = \frac{\pi^{D/2}}{D2^{D-1}\Gamma(D/2)} \tag{1.145}$$

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

$$\Gamma(x+1) \simeq (2\pi)^{1/2} e^{-x} x^{x+1/2} \tag{1.146}$$

which is valid for $x \gg 1$, to show that, as $D \to \infty$, the ratio (1.145) goes to zero. Show also that the ratio of the distance from the centre 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 a large number of corners, which themselves become very lone 'spikes'!

Proof. Using the results of Exercise 1.18, we have that the volume of D-dimensional hypersphere of radius a is

$$V_{D_{\text{sphere}}}(a) = \frac{2\pi^{D/2} a^D}{D\Gamma(D/2)}$$

We also know that the volume of the $D-$hypercube of size $2a$ is given by:

$$V_{D_{\text{cube}}(2a)} = (2a)^D = 2^D a^D$$

Therefore the ratio of the volumes is given by

$$\frac{V_{D_{\text{sphere}}(a)}}{V_{D_{\text{cube}}}(a)} = \frac{\pi^{D/2}}{D2^{D-1}\Gamma(D/2)} \tag{1.145}$$

By using Stirling's approximation, we have that

$$\begin{aligned}
\lim_{D\to\infty} \frac{\pi^{D/2}}{D2^{D-1}\Gamma(D/2)} &= \lim_{D\to\infty} \frac{\pi^{D/2}}{D2^{D-1}(2\pi)^{1/2} e^{1-D/2}(D/2-1)^{D/2-1/2}} \\
&= \lim_{D\to\infty} \left\{ \left(\frac{\pi}{4}\right)^{D/2} \cdot \left(\frac{e}{D/2-1}\right)^{D/2-1} \cdot \frac{\sqrt{D-2}}{D\sqrt{\pi}} \right\} = 0
\end{aligned}$$

Now, we want to find the ratio between the distance from the centre of the hypercube to one of the corners and the distance from the centre to a side. We can consider without loss of generality a D-dimensional hypercube of length 2α, centered in the origin 0_D of the $\mathbb{R}^D$ Cartesian system. The center of a hypercube side takes the form $\mathbf{s} = (\alpha_1, \alpha_2, \ldots, \alpha_D)$, where $\alpha_i \in \{0, a\}$ such that $||\mathbf{s}|| = a$, i.e. only one coordinate is equal to a and the rest are 0. On the other hand, the corners of the hypercube take the form $\mathbf{c} = (\beta_1, \beta_2, \ldots, \beta_D)$, where $\beta_i \in \{\pm a\}$. We'll then have that $||\mathbf{c}|| = a\sqrt{D}$. As a result, our ratio looks like expected:

$$\frac{\text{distance from center to corner}}{\text{distance from center to side}} = \frac{||\mathbf{s}||}{||\mathbf{c}||} = \frac{a\sqrt{D}}{a} = \sqrt{D}$$

□

Exercise 1.21 ⋆⋆

Consider two nonnegative numbers a and b, and show that, if $a \leq b$, then $a \leq (ab)^{1/2}$. 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

$$p(\text{mistake}) \leq \int \{p(\mathbf{x}, \mathcal{C}_1)p(\mathbf{x}, \mathcal{C}_2)\}^{1/2} \, \mathrm{d}\mathbf{x} \tag{1.150}$$

Proof. We start by proving the identity. We have that

$$a \leq (ab)^{1/2} \iff a^2 \leq ab \iff a^2 - ab \leq 0 \iff a(a-b) \leq 0$$

which is true since $a \leq b$.

Now, since the regions are chosen to minimize the probability of misclassification, for an individual value of $\mathbf{x}$, the region $\mathcal{R}_k$ with the higher joint/posterior probability associated to $\mathcal{C}_k$ is chosen, so:

$$p(\mathbf{x}, \mathcal{C}_2) \leq p(\mathbf{x}, \mathcal{C}_1), \forall \mathbf{x} \in \mathcal{R}_1 \qquad p(\mathbf{x}, \mathcal{C}_1) \leq p(\mathbf{x}, \mathcal{C}_2), \forall \mathbf{x} \in \mathcal{R}_2$$

By applying the $a \leq (ab)^{1/2}$ identity above, we get that

$$p(\mathbf{x}, \mathcal{C}_2) \leq \{p(\mathbf{x}, \mathcal{C}_1)p(\mathbf{x}, \mathcal{C}_2)\}^{1/2}, \forall \mathbf{x} \in \mathcal{R}_1 \qquad p(\mathbf{x}, \mathcal{C}_1) \leq \{p(\mathbf{x}, \mathcal{C}_1)p(\mathbf{x}, \mathcal{C}_2)\}^{1/2}, \forall \mathbf{x} \in \mathcal{R}_2$$

If we integrate the inequalities over the associated regions, we have that:

$$\int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2) \, \mathrm{d}\mathbf{x} \leq \int_{\mathcal{R}_1} \{p(\mathbf{x}, \mathcal{C}_1)p(\mathbf{x}, \mathcal{C}_2)\}^{1/2} \, \mathrm{d}\mathbf{x}$$

$$\int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1) \, \mathrm{d}\mathbf{x} \leq \int_{\mathcal{R}_2} \{p(\mathbf{x}, \mathcal{C}_1)p(\mathbf{x}, \mathcal{C}_2)\}^{1/2} \, \mathrm{d}\mathbf{x}$$

By summing the above inequalities, we find that:

$$\int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2) \, \mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1) \, \mathrm{d}\mathbf{x} \leq \int \{p(\mathbf{x}, \mathcal{C}_1)p(\mathbf{x}, \mathcal{C}_2)\}^{1/2}$$

which is equivalent to (1.150). □

Exercise 1.22 ⋆

Given a loss matrix with elements L_{kj}, the expected risk is minimized, if for each $\mathbf{x}$, we choose the class that minimizes (1.81). 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?

Proof. The expectation is minimized if for each $\mathbf{x}$ we choose the class $\mathcal{C}_j$ such that the quantity

$$\sum_k L_{kj} p(\mathcal{C}_k | \mathbf{x}) \tag{1.81}$$

is minimized. For $L_{kj} = 1 - I_{kj}$ the quantity becomes

$$\sum_k (1 - I_{kj}) p(\mathcal{C}_k|\mathbf{x}) = \sum_k p(\mathcal{C}_k|\mathbf{x}) - p(\mathcal{C}_j|\mathbf{x}) = 1 - p(\mathcal{C}_j|\mathbf{x})$$

and it's obviously minimised by choosing the class $\mathcal{C}_j$ having the largest posterior probability $p(\mathcal{C}_j|\mathbf{x})$

This form of loss matrix makes each mistake have the same "weight", no mistake is worse than another. □

Exercise 1.23 ⋆

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

Proof. Minimizing the expected loss

$$\mathbb{E}[L] = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_j) \, \mathrm{d}\mathbf{x} \tag{1.80}$$

is equivalent with minimizing

$$\sum_k L_{kj} p(\mathbf{x}, \mathcal{C}_k)$$

for each $\mathbf{x}$. Therefore, by using Bayes' theorem, we have that the criterion of minimizing the expected loss is the class $\mathcal{C}_j$ for each $\mathbf{x}$ such that

$$\sum_k L_{kj} p(\mathbf{x}|\mathcal{C}_k)$$

is minimized. □

Exercise 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 selecting the reject 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 when the loss matrix is given by $L_{kj} = 1 - I_{kj}$. What is the relationship between λ and the rejection threshold θ?

Proof. The decision criterion reduces to choosing the minimum between the loss of choosing the best class and the reject loss λ. Therefore, if

$$\alpha = \operatorname*{argmin}_j \sum_k L_{kj} p(\mathbf{x}|\mathcal{C}_k)$$

we choose the class α if the above quantity is less than λ and use the reject option otherwise. If the loss matrix is given by $L_{kj} = 1 - I_{kj}$, then

$$\alpha = \operatorname*{argmin}_j \{1 - p(\mathcal{C}_j|\mathbf{x})\}$$

which makes $\mathcal{C}_\alpha$ the class with the highest posterior probability. Therefore the criterion reduces to the one discussed in Section 1.5.3. If the highest posterior probability is smaller than $1 - \lambda$, then we use the reject option. This is equivalent with using $\theta = 1 - \lambda$ in Section 1.5.3. □

Exercise 1.25 ⋆ TODO

Consider the generalization of the squared loss function (1.87) for a single target variable t to the case of multiple target variables described by the vector $\mathbf{t}$ given by

$$\mathbb{E}[L(\mathbf{t}, \mathbf{y}(\mathbf{x})] = \iint ||\mathbf{y}(\mathbf{x}) - \mathbf{t}||^2 p(\mathbf{x}, \mathbf{t}) \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{t} \tag{1.151}$$

Using the calculus of the variations, show that the function $\mathbf{y}(\mathbf{x})$ for which this expected loss is minimized is given by $\mathbf{y}(\mathbf{x}) = \mathbb{E}_\mathbf{t}[\mathbf{t}|\mathbf{x}]$. Show that this result reduces to (1.89) for the case of a single target variable t.

Exercise 1.26 ⋆ TODO

By expansion of the square in (1.151), derive a result analogous to (1.90), and hence show that the function $\mathbf{y}(\mathbf{x})$ that minimizes the expected square loss for the case of a vector $\mathbf{t}$ of target variables is again given by the conditional expectation of $\mathbf{t}$.

Exercise 1.28 ⋆ TODO

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 that 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 $h(p^{n/m}) = n/mh(p)$ 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 $h(p) \propto \ln p$.

Exercise 1.30 ⋆⋆

Evaluate the Kullback-Leibler divergence (1.113) between two Gaussians $p(x) = \mathcal{N}(x|\mu, \sigma^2)$ and $q(x) = \mathcal{N}(x|m, s^2)$.

Proof. The Kullback-Leibler divergence is given by

$$\mathrm{KL}(p||q) = -\int p(x) \ln \left\{ \frac{q(x)}{p(x)} \right\} \mathrm{d}x \tag{1.113}$$

We start by splitting the integral into:

$$\mathrm{KL}(p||q) = -\int p(x) \ln q(x) \, \mathrm{d}x + \int p(x) \ln p(x) \, \mathrm{d}x$$

The negation of the second term will be equal to the entropy of the Gaussian, that is:

$$H_p[x] = \frac{1}{2}\{1 + \ln\big(2\pi\sigma^2\big)\} \tag{1.110}$$

We have that

$$\ln q(x) = \ln \mathcal{N}(x|m, s^2) = \frac{1}{2}\ln\big(2\pi s^2\big) - \frac{(x-m)^2}{s^2}$$

so by using the fact that the Gaussian is normalized and by noticing the expected values, the KL divergence becomes:

$$\begin{aligned}
\mathrm{KL}(p||q) &= \frac{1}{s^2}\int p(x)x^2\,\mathrm{d}x - \frac{2m}{s^2}\int p(x)x\,\mathrm{d}x + \left\{\frac{1}{2}\ln\big(2\pi s^2\big) + \frac{m^2}{s^2}\right\}\int p(x)\,\mathrm{d}x - H_p[x] \\
&= \frac{1}{s^2}\mathbb{E}[x^2] - \frac{2m}{s^2}E[x] + \frac{m^2}{s^2} + \ln\frac{s}{\sigma} + \frac{1}{2} \\
&= \frac{1}{2} + \ln\frac{s}{\sigma} + \frac{\sigma^2 + (\mu - m)^2}{s^2}
\end{aligned}$$

□

Exercise 1.31 ⋆⋆

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

$$H[\mathbf{x}, \mathbf{y}] \leq H[\mathbf{x}] + H[\mathbf{y}] \tag{1.152}$$

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

Proof. The differential entropy of two variables $\mathbf{x}$ and $\mathbf{y}$ is given by

$$H[\mathbf{x}, \mathbf{y}] = H[\mathbf{y}|\mathbf{x}] + H[\mathbf{x}] \tag{1.112}$$

so (1.152) becomes equivalent with

$$H[\mathbf{y}|\mathbf{x}] - H[\mathbf{y}] \leq 0 \tag{1.31.1}$$

which we're going to prove now.

We start by rewriting the entropy $H[\mathbf{y}]$ as

$$H[\mathbf{y}] = -\int p(\mathbf{y})\ln p(\mathbf{y})\,\mathrm{d}\mathbf{y} = -\iint p(\mathbf{x}, \mathbf{y})\ln p(\mathbf{y})\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y}$$

Therefore, since the differential entropy is given by

$$H[\mathbf{y}|\mathbf{x}] = \iint p(\mathbf{x}, \mathbf{y})\ln p(\mathbf{y}|\mathbf{x})\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} \tag{1.111}$$

we have that

$$\begin{aligned}
H[\mathbf{y}|\mathbf{x}] - H[\mathbf{y}] &= -\iint p(\mathbf{x}, \mathbf{y})\ln p(\mathbf{y}|\mathbf{x})\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} + \iint p(\mathbf{x}, \mathbf{y})\ln p(\mathbf{y})\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} \\
&= \iint p(\mathbf{x}, \mathbf{y})\ln\left\{\frac{p(\mathbf{y})}{p(\mathbf{y}|\mathbf{x})}\right\}\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y}
\end{aligned}$$

By using the inequality $\ln \alpha \leq \alpha - 1$, for all $\alpha > 0$, we obtain:

$$\begin{aligned} H[\mathbf{y}|\mathbf{x}] - H[\mathbf{y}] &\leq \iint p(\mathbf{x})p(\mathbf{y}|\mathbf{x})\left\{\frac{p(\mathbf{y})}{p(\mathbf{y}|\mathbf{x})} - 1\right\} \mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} \\ &\leq \iint p(\mathbf{x})p(\mathbf{y})\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} - \iint p(\mathbf{x})p(\mathbf{y})\,\mathrm{d}\mathbf{x}\mathbf{y} \\ &\leq \iint p(\mathbf{x})p(\mathbf{y})\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} - 1 \\ &\leq 0 \end{aligned}$$

which proves (1.31.1), respectively (1.152). □

Exercise 1.32 ⋆

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

Proof. By generalizing (1.27) for the multivariate case, we have that:

$$p_{\mathbf{y}}(\mathbf{y}) = p_{\mathbf{x}}(\mathbf{x})\left|\frac{\partial \mathbf{x}}{\partial \mathbf{y}}\right| = p_{\mathbf{x}}(\mathbf{x})\left|\frac{\partial \mathbf{A}^{-1}\mathbf{y}}{\partial \mathbf{y}}\right| = p_{\mathbf{x}}(\mathbf{x})|\mathbf{A}^{-1}| = p_{\mathbf{x}}(\mathbf{x})|\mathbf{A}|^{-1}$$

where $J = \left|\frac{\partial \mathbf{x}}{\partial \mathbf{y}}\right| = |\mathbf{A}|^{-1}$ is the Jacobian determinant.

Now, the entropy of $\mathbf{y}$ is given by:

$$\begin{aligned} H[\mathbf{y}] &= -\int p_{\mathbf{y}}(\mathbf{y}) \ln p_{\mathbf{y}}(\mathbf{y})\,\mathrm{d}\mathbf{y} = -\int \frac{p_{\mathbf{x}}(\mathbf{x})}{|\mathbf{A}|} \ln \frac{p_{\mathbf{x}}(\mathbf{x})}{|\mathbf{A}|}\left|\frac{\mathrm{d}\mathbf{y}}{\mathrm{d}\mathbf{x}}\right| \mathrm{d}\mathbf{x} = -\int p_{\mathbf{x}}(\mathbf{x}) \ln \frac{p_{\mathbf{x}}(\mathbf{x})}{|\mathbf{A}|}\,\mathrm{d}\mathbf{x} \\ &= -\int p_{\mathbf{x}}(\mathbf{x}) \ln p_{\mathbf{x}}\,\mathrm{d}\mathbf{x} + \ln|\mathbf{A}| \int p_{\mathbf{x}}(\mathbf{x})\,\mathrm{d}\mathbf{x} \\ &= H[\mathbf{y}] + \ln|\mathbf{A}| \end{aligned}$$

□

Exercise 1.33 ⋆⋆ TODO

Suppose that the conditional entropy $H[y|x]$ 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 $p(y|x) \neq 0$.

Exercise 1.35 ⋆

Use the results (1.106) and (1.107) to show that the entropy of the univariate Gaussian (1.109) is given by (1.110).

Proof. The entropy of the univariate Gaussian is given by:

$$H[x] = -\int \mathcal{N}(x|\mu,\sigma^2)\ln\mathcal{N}(x|\mu,\sigma^2)\,\mathrm{d}x = -\frac{1}{2}\ln\big(2\pi\sigma^2\big)\int \mathcal{N}(x|\mu,\sigma^2)\,\mathrm{d}x + \int \mathcal{N}(x|\mu,\sigma^2)\frac{(x-\mu)^2}{\sigma^2}\,\mathrm{d}x$$
$$= \Big\{ -\frac{1}{2}\ln\big(2\pi\sigma^2\big) + \frac{\mu^2}{\sigma^2}\Big\}\int \mathcal{N}(x|\mu,\sigma^2)\,\mathrm{d}x + \frac{1}{2\sigma^2}\int \mathcal{N}(x|\mu,\sigma^2)x^2\,\mathrm{d}x - \frac{2\mu}{2\sigma^2}\int \mathcal{N}(x|\mu,\sigma^2)x\,\mathrm{d}x$$

By using the fact that the Gaussian is normalized and by noticing the expression of the expected value, we have that

$$H[x] = -\frac{1}{2}\ln\big(2\pi\sigma^2\big) + \frac{\mu^2}{2\sigma^2} + \frac{1}{2\sigma^2}\mathbb{E}[x^2] - \frac{2\mu}{2\sigma^2}\mathbb{E}[x] = \frac{1}{2}\Big\{1 - \frac{1}{2}\ln\big(2\pi\sigma^2\big)\Big\} \tag{1.110}$$

□

Exercise 1.37 ⋆

Using the definition (1.111) together with the product rule of probability, prove the result (1.112).

Proof. Using the product rule of probability, one could rewrite the entropy of $\mathbf{x}$ as:

$$H[\mathbf{x}] = -\int p(\mathbf{x})\ln p(\mathbf{x})\,\mathrm{d}\mathbf{x} = -\iint p(\mathbf{x},\mathbf{y})\ln p(\mathbf{x})\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y}$$

Now, by summing this with (1.111) we see that:

$$\begin{aligned} H[\mathbf{y}|\mathbf{x}] + H[\mathbf{x}] &= -\iint p(\mathbf{x},\mathbf{y})\ln p(\mathbf{y}|\mathbf{x})\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} - \iint p(\mathbf{x},\mathbf{y})\ln p(\mathbf{x})\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} \\ &= -\iint p(\mathbf{x},\mathbf{y})\ln\{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})\}\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} \\ &= -\iint p(\mathbf{x},\mathbf{y})\ln p(\mathbf{x},\mathbf{y})\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} \\ &= H[\mathbf{x},\mathbf{y}] \end{aligned} \tag{1.112}$$

□

Exercise 1.38 ⋆⋆

Using proof by induction, show that the inequality (1.114) for convex functions implies the result (1.115).

Proof. We'll prove Jensen's inequality by induction, i.e. if we have N points $x_1,\dots x_n$, f is a convex function and $\lambda_i \geq 0$, $\sum_{i=1}^{N}\lambda_i = 1$, then

$$f\Big(\sum_{i=1}^{N}\lambda_i x_i\Big) \leq \sum_{i=1}^{N}\lambda_i f(x_i) \tag{1.115}$$

We consider the base case of the induction to be given by

$$f(\lambda a + (1-\lambda)b) \leq \lambda f(a) + (1-\lambda)f(b) \tag{1.114}$$

for $N = 2, x_1 = a, x_2 = b, \lambda_1 = \lambda, \lambda_2 = 1 - \lambda$, which is valid for any convex function f. Now, we assume that the general case with K elements, where $\lambda_i \geq 0, \sum_{i=1}^{K} \lambda_i = 1$ is true. Therefore,

$$f\left(\sum_{i=1}^{K} \lambda_i x_i\right) \leq \sum_{i=1}^{K} \lambda_i f(x_i)$$

We want to deduce that this is true for $K+1$ elements too. Therefore, let $\lambda_{k+1} \geq 0$ such that

$$\lambda_i' = \lambda_i - \frac{1}{K}\lambda_{k+1}$$

and

$$\sum_{i=1}^{K} \lambda_i x_i = (1 - \lambda_{k+1}) \sum_{i=1}^{K} \lambda_i' x_i$$

$\lambda_2 = 1 - \lambda_1$ such that $\lambda_1 + \lambda_2 = 1$. By Jensens' inequality we have that:

$$f\left(\sum_{i=1}^{K} \lambda_i x_i\right) \leq \sum_{i=1}^{K} \lambda_i f(x_i)$$

□

Exercise 1.39 ⋆⋆⋆

Consider two binary variables x and y having the joint distribution given in Table 1.3. Evaluate the following quantities:

(a) $H[x]$ (c) $H[y|x]$ (e) $H[x, y]$

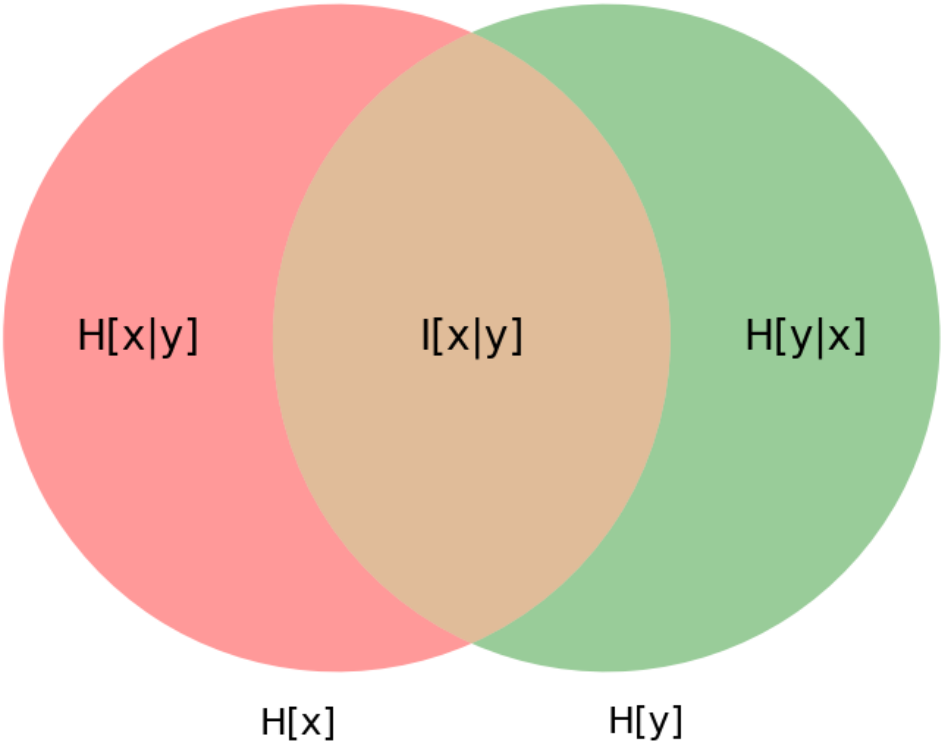
(b) $H[y]$ (d) $H[x|y]$ (f) $I[x, y]$

Draw a diagram to show the relationship between these various quantities.

		y 0	y 1
x	0	1/3	1/3
x	1	0	1/3

Table 1.3 The joint distribution $p(x, y)$ used in Exercise 1.39.

Proof. Through straightforward computations using the discrete formula for the entropy, we have



Exercise 1.39 Diagram

(a) $H[x] = -2/3 \ln 2 + \ln 3$ (c) $H[x|y] = 2/3 \ln 2$ (e) $H[x, y] = \ln 3$

(b) $H[y] = -2/3 \ln 2 + \ln 3$ (d) $H[y|x] = 2/3 \ln 2$ (f) $I[x, y] = -4/3 \ln 2 + \ln 3$

The diagram shows the relationship between the entropies. Note that the joint entropy $H[x, y]$ occupies all three colored areas.

□

Exercise 1.40 ⋆

By applying Jensen's inequality (1.115) with $f(x) = \ln x$, show that the arithmetic mean of a set of real numbers is never less than their geometric mean.

Proof. Let N be the cardinality of the considered set of real numbers. By considering $f(x) = \ln x$ (which is convex) and $\lambda_i = 1/N$, we use Jensen's inequality to obtain:

$$\ln\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right) = \frac{1}{N}\sum_{i=1}^{N} \ln x_i = \frac{1}{N} \ln\left(\prod_{i=1}^{N} x_i\right) = \ln\left\{\left(\prod_{i=1}^{N} x_i\right)^{1/N}\right\}$$

Since $\ln x$ is increasing, the above inequality is equivalent with:

$$\frac{1}{N}\sum_{i=1}^{N} x_i \leq \left(\prod_{i=1}^{n} x_i\right)^{1/N}$$

which proves that the arithmetic mean of a set of real numebrs is never less than their geometric mean. □

Exercise 1.41 ⋆

Using the sum and product rules of probability, show that the mutual information $I(\mathbf{x}, \mathbf{y})$ satisfies the relation (1.121).

Proof. The mutual information between the variables $\mathbf{x}$ and $\mathbf{y}$ is given by:

$$I[\mathbf{x}, \mathbf{y}] = -\iint p(\mathbf{x}, \mathbf{y}) \ln \left(\frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})} \right) \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y} \tag{1.120}$$

We split the integral and use the product and sum rules of probability to obtain the desired result:

$$\begin{aligned} I[\mathbf{x}, \mathbf{y}] &= -\iint p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{x})\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y} + \iint p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{x}|\mathbf{y})\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y} \\ &= -\int p(\mathbf{x}) \ln p(\mathbf{x})\, \mathrm{d}\mathbf{x} + \iint p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{x}|\mathbf{y})\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y} \\ &= H[\mathbf{x}] - H[\mathbf{x}|\mathbf{y}] \end{aligned} \tag{1.121}$$

Analogously, one could easily show that also $I[\mathbf{x}, \mathbf{y}] = H[\mathbf{y}] - H[\mathbf{y}|\mathbf{x}]$ □