Pattern Recognition and Machine Learning Exercises: Chapter 1

Paul Scemama

August 18, 2023

Contents

1 Introduction

1.1 Error minimization

We have the error function $E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2$, where $y(x, \mathbf{w}) = \sum_{j=0}^{M} w_j x^j$. We'd like to find the optimal setting for $\mathbf{w}$ in the sense that it minimizes $E(\mathbf{w})$. We will show that we can cast this minimization problem as solving a system of linear equations.

Claim: The w_i in $\mathbf{w} = (w_0, w_1, w_2..., w_M)$ that minimize the error function $E(\mathbf{w})$ are given by the solution to the following set of linear equations,

$$\sum_{j=0}^{M} A_{ij} w_j = T_i \quad \text{where} \quad A_{ij} = \sum_{n=1}^{N} (x_n)^{(i+j)}, \; T_i = \sum_{n=1}^{N} (x_n)^i t_n$$

Approach: We will take the derivative of $E(\mathbf{w})$ with respect to $\mathbf{w}$, set it to zero, and then rearrange terms to prove the claim above.

Proof: By the chain rule,

$$\frac{\partial E(\mathbf{w})}{\partial w_i} = \frac{\partial E(\mathbf{w})}{\partial y(x_n, \mathbf{w})} \frac{\partial y(x_n, \mathbf{w})}{\partial w_i} \tag{1}$$

Solving the two terms on the right hand side yields

$$\frac{\partial E(\mathbf{w})}{\partial y(x_n, \mathbf{w})} = \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}$$
$$\frac{\partial y(x_n, \mathbf{w})}{\partial w_i} = \frac{\partial}{\partial w_i}(w_0 + w_1 x_1 + \ldots w_i x_n^i \cdots + w_m x_n^M) = x_n^i$$

Substituting back into (1) results in

$$\begin{aligned}
\frac{\partial E(\mathbf{w})}{\partial w_i} &= \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\} x_n^i \\
&\overset{(i)}{=} \sum_{n=1}^{N} (\sum_{j=0}^{M} w_j x_n^j - t_n) x_n^i \\
&\overset{(ii)}{=} \sum_{n=1}^{N} (\sum_{j=0}^{M} w_j x_n^i x_n^j - t_n x_n^i) \\
&\overset{(iii)}{=} \sum_{n=1}^{N} (\sum_{j=0}^{M} w_j x_n^{(i+j)} - t_n x_n^i)
\end{aligned}$$

where in (i) we use the definition of $y(x_n, \mathbf{w})$, in (ii) we distribute x_n^i into the parentheses, and in (iii) we use the exponent rule. Setting the derivative to 0 and rearranging,

$$\begin{aligned}
\sum_{n=1}^{N} (\sum_{j=0}^{M} w_j x_n^{(i+j)} - t_n x_n^i) &= 0 \\
\sum_{n=1}^{N} \sum_{j=0}^{M} w_j x_n^{(i+j)} &= \sum_{n=1}^{N} t_n x_n^i \\
\sum_{j=0}^{M} A_{ij} w_j &= T_i
\end{aligned}$$

■

1.2 Error minimization with regularization

We have the error function $\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w})\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$, where $\|\mathbf{w}\|^2 = \mathbf{w}^\intercal\mathbf{w} = w_0^2 + w_1^2 + \cdots + w_M^2$ and the parameter λ controls the strength of regularization. We'd like to find the optimal setting for $\mathbf{w}$ in the sense that it minimizes $\widetilde{E}(\mathbf{w})$. We will show that we can cast this minimization problem as solving a system of linear equations.

Claim: The w_i in $\mathbf{w} = (w_1, w_2, ..., w_M)$ that minimize the error function $\widetilde{E}(\mathbf{w})$ are given by the solution to the following set of linear equations,

$$\sum_{j=0}^{M} A_{ij}w_j + \lambda w_i = T_i \quad \text{where} \quad A_{ij} = \sum_{n=1}^{N}(x_n)^{(i+j)},\ T_i = \sum_{n=1}^{N}(x_n)^i t_n$$

Approach: We will take the derivative of $\widetilde{E}(\mathbf{w})$ with respect to $\mathbf{w}$, set it to zero, and then rearrange terms to prove the claim above.

Proof: By the chain rule,

$$\frac{\partial\widetilde{E}(\mathbf{w})}{\partial w_i} = \frac{\partial E(\mathbf{w})}{\partial y(x_n,\mathbf{w})}\frac{\partial y(x_n,\mathbf{w})}{\partial w_i} + \frac{\lambda}{2}\frac{\partial \mathbf{w}^\intercal\mathbf{w}}{\partial w_i} \tag{1}$$

Solving the two terms on the right hand side yields

$$\begin{aligned}
\frac{\partial\widetilde{E}(\mathbf{w})}{\partial y(x_n,\mathbf{w})} &= \sum_{n=1}^{N}\{y(x_n,\mathbf{w}) - t_n\} \\
\frac{\partial y(x_n,\mathbf{w})}{\partial w_i} &= \frac{\partial}{\partial w_i}(w_0 + w_1x_1 + \ldots w_ix_n^i \cdots + w_mx_n^M) = x_n^i \\
\frac{\partial \mathbf{w}^\intercal\mathbf{w}}{\partial w_i} &= \frac{\partial}{\partial w_i}(w_0^2 + w_1^2 + \ldots w_i^2 + \ldots w_M^2) = 2w_i
\end{aligned}$$

Substituting back into (1) yields

$$\begin{aligned}
\frac{\partial\widetilde{E}(\mathbf{w})}{\partial w_i} &= \sum_{n=1}^{N}\{y(x_n,\mathbf{w}) - t_n\}x_n^i + \frac{\lambda}{2}2w_i \\
&\overset{(i)}{=} \sum_{n=1}^{N}(\sum_{j=0}^{M} w_jx_n^j - t_n)x_n^i + \lambda w_i \\
&\overset{(ii)}{=} \sum_{n=1}^{N}(\sum_{j=0}^{M} w_jx_n^ix_n^j - t_nx_n^i) + \lambda w_i \\
&\overset{(iii)}{=} \sum_{n=1}^{N}(\sum_{j=0}^{M} w_jx_n^{(i+j)} - t_nx_n^i) + \lambda w_i
\end{aligned}$$

where in (i) we use the definition of $y(x_n,\mathbf{w})$ and $\frac{\lambda}{2}\cdot 2 = \lambda$, in (ii) we distribute x_n^i into the parentheses, and in (iii) we use the exponent rule. Setting the derivative to 0 and rearranging,

$$\begin{aligned}
\sum_{n=1}^{N}(\sum_{j=0}^{M} w_jx_n^{(i+j)} - t_nx_n^i) + \lambda w_i &= 0 \\
\sum_{n=1}^{N}\sum_{j=0}^{M} w_jx_n^{(i+j)} + \lambda w_i &= \sum_{n=1}^{N} t_nx_n^i \\
\sum_{j=0}^{M} A_{ij}w_j + \lambda w_i &= T_i
\end{aligned}$$

■

1.3 Sum & product rules of probability

We have boxes filled with fruit:

A red (r) box $\rightarrow$ 3 apples, 4 oranges, 3 limes
A blue (b) box $\rightarrow$ 1 apple, 1 oranges, 0 limes
A green (g) box $\rightarrow$ 3 apples, 3 oranges, 4 limes

A box is chosen at random with probabilities

$$p(r) = 0.2$$
$$p(b) = 0.2$$
$$p(g) = 0.6,$$

and then a fruit is chosen at random.

Question: What is the probability of selecting an apple?

Answer: By the sum rule,

$$p(fruit = a) = \sum_{box \in Boxes} p(a, box) = p(a, r) + p(a, b) + p(a, g) \tag{1}$$

By the product rule on each term in (1),

$$\begin{aligned} p(a) &= p(a|r)p(r) + p(a|b)p(b) + p(a|g)p(g) \\ &= \frac{3}{10} \cdot \frac{1}{5} + \frac{1}{2} \cdot \frac{1}{5} + \frac{3}{10} \cdot \frac{3}{5} \\ &= \frac{17}{50} \end{aligned}$$

Question: Given you select an orange, what is the probability that it came from the green box?

Answer: By Bayes' theorem,

$$\begin{aligned} p(box = g|fruit = o) &= \frac{p(o|g)p(g)}{p(fruit = o)} \\ &= \frac{3/10 \cdot 6/10}{p(fruit = o)} \end{aligned}$$

By the sum rule,

$$p(fruit = o) = \sum_{box \in Boxes} p(o|box) = p(o, r) + p(o, b) + p(o, g) \tag{1}$$

By the product rule on each term in (1),

$$\begin{aligned} p(o) &= p(o|r)p(r) + p(o|b)p(b) + p(o|g)p(g) \\ &= \frac{4}{10} \cdot \frac{2}{10} + \frac{1}{2} \cdot \frac{2}{10} + \frac{3}{10} \cdot \frac{6}{10} \\ &= \frac{9}{25} \end{aligned}$$

So $p(box = g|fruit = o) = \frac{18/100}{36/100} = \frac{1}{2}$.

1.4 Change of variables and probability densities

Claim: If we have two variables x, y related by $x = g(y)$, then the modes of the probability densities $\widehat{x}, \widehat{y}$ will be related by $\widehat{x} = g(\widehat{y})$ *only* if the g is linear. If g is instead nonlinear, then the relation does not hold. This is in contrast to regular functions (not probability densities), where the relation holds even if g is nonlinear.

Approach: Calculate the respective expressions that the modes $\widehat{x}, \widehat{y}$ must satisfy in terms of their densities. First show that if g is linear, the expressions are related by $\widehat{x} = g(\widehat{y})$. Then show that if g is nonlinear, those expressions are *not* related by $\widehat{x} = g(\widehat{y})$.

Proof: According to the relation $x = g(y)$, the density $p_x(x)$ turns into

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

If $\widehat{x}$ is the mode of p_x it must satisfy $p_x'(\widehat{x}) = 0$. If $\widehat{y}$ is the mode of p_y it must satisfy

$$\begin{aligned} p_y'(\widehat{y}) &\overset{(i)}{=} p_x'(g(\widehat{y}))g'(\widehat{y}))|g'(\widehat{y})| + p_x'(g(\widehat{y}))\frac{g'(\widehat{y})}{|g'(\widehat{y})|}g''(\widehat{y}) \\ &\overset{(ii)}{=} p_x'(g(\widehat{y}))(g'(\widehat{y}))^2 + p_x'(g(\widehat{y}))\frac{g'(\widehat{y})}{|g'(\widehat{y})|}g''(\widehat{y}) \\ &= 0 \end{aligned} \tag{4}$$

Where we've used in the chain rule in (i) and simplified terms in (ii). Assuming g is linear, then it must be that $g''(y) = 0\ \forall y$ and the right hand side of (4) vanishes. We are then left with

$$p_y'(\widehat{y}) = p_x'(g(\widehat{y}))(g'(\widehat{y}))^2 = 0$$

Assuming $g'(\widehat{y}) \neq 0$ (i.e. that g' is non-zero at the mode $\widehat{y}$ of the *density* p_y) means it must be that $p_x'(g(\widehat{y})) = 0$. Thus,

$$\begin{aligned} &\widehat{x} \text{ must satisfy } p_x'(\widehat{x}) = 0. \\ &\widehat{y} \text{ must satisfy } p_x'(g(\widehat{y})) = 0. \end{aligned}$$

And so the locations of the modes are related by $\widehat{x} = g(\widehat{y})$.

However if g is *nonlinear*, then the right hand side of (4) does *not* vanish and the relation $\widehat{x} = g(\widehat{y})$ no longer holds.

■

Conclusion: If, in the relation $x = g(y)$, g is *linear*, then the location of $\widehat{x}$ by maximizing p_x will be the same value obtained by finding $\widehat{y}$ from transforming to p_y and maximizing with respect to y, and then finally transforming back to x. If g is *nonlinear*, however, this is not the case. And so in general we cannot do this.

1.5 Variance Theorem

Claim: Given a random variable x, the variance of a function $f(x)$ is defined by

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

and represents the variability of $f(x)$ around its mean. We claim that the variance can alternatively be written as

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

Approach: Expand the squared expression within the outermost expectation. Then use the fact that the expectation is a linear operator to arrive at (1.39).

Proof: We begin with the definition (1.38):

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

where in (i) we used the fact that the expectation is a linear operator, and in (ii) we used the fact that an expectation of a constant is that constant. ■

1.6 Independence and Covariance

Claim: If two random variables x and y are independent, then their covariance is zero.

Approach: Use the definition of expectation to expand the terms in the definition of the covariance. Then use the fact that if $x \perp y$, then their joint distribution factorizes like so $p(x, y) = p(x)p(y)$ to illustrate that the covariance is zero in such a case.

Proof: We begin with the definition of covariance ({1.41),

$$\begin{aligned}
\text{cov}[x, y] &= \mathbb{E}[\{\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}] \\
&= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y] \\
&\overset{(i)}{=} \int\int xyp(x, y)\mathrm{d}x\mathrm{d}y - \int xp(x)\mathrm{d}x \int yp(y)\mathrm{d}y \\
&\overset{(ii)}{=} \int\int xyp(x)p(y)\mathrm{d}x\mathrm{d}y - \int xp(x)\mathrm{d}x \int yp(y)\mathrm{d}y \\
&\overset{(iii)}{=} \int xp(x)\mathrm{d}x \int yp(y)\mathrm{d}y - \int xp(x)\mathrm{d}x \int yp(y)\mathrm{d}y \\
&= 0
\end{aligned}$$

where in (i) we use the definition of expectation, in (ii) we use the factorization of the joint of two independent variables, and in (iii) we see that the integrand contains know dependent parts between x and y. ■

1.7 Univariate Gaussian's Normalizing Constant

Claim: The normalization condition

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

holds where

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

Approach: We first note: suppose we want to show $\frac{a}{b} = 1$. Through some algebra, we can instead show $a = b$. With this in mind, we are going to show that

$$\int_{-\infty}^{\infty} \exp\{-\frac{1}{2\sigma^2}(x - \mu)^2\}\mathrm{d}x = (2\pi\sigma^2)^{1/2},$$

by evaluating the integral on the left hand side (in a slightly more general form).

Proof: Consider

$$I = \int_{-\infty}^{\infty} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}\mathrm{d}x \tag{1}$$

$$\stackrel{(i)}{=} \int_{-\infty}^{\infty} \exp\{-\frac{1}{2\sigma^2}y^2\}\mathrm{d}y \tag{2}$$

where in (i) we've made the change of variables defined by the relation $y = x - \mu$. To evaluate the integral (2) we are going to

1. Square it.
2. Transform from the Cartesian coordinates to Polar coordinates $(x, y) \to (r, \theta)$.
3. Make the change of variables defined by the relation $u = r^2$.
4. Compute the integral over θ and u.
5. Take the square root of both sides.

Let's first write the square of (2) in the form

$$I^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\{-\frac{1}{2\sigma^2}x^2 - \frac{1}{2\sigma^2}y^2\}\mathrm{d}x\mathrm{d}y.$$

We then make the change to Polar coordinates (which is considered a change of variables),

$$\begin{aligned}
I^2 &\stackrel{(i)}{=} \int_0^{2\pi}\int_0^{\infty} \exp\{-\frac{1}{2\pi\sigma^2}(r\cos\theta)^2 - \frac{1}{2\sigma^2}(r\sin\theta)^2\}r\mathrm{d}r\mathrm{d}\theta \\
&= \int_0^{2\pi}\int_0^{\infty} \exp\{-\frac{1}{2\pi\sigma^2}(r^2\cos^2\theta + r^2\sin^2\theta)\}r\mathrm{d}r\mathrm{d}\theta \\
&= \int_0^{2\pi}\int_0^{\infty} \exp\{-\frac{r^2}{2\pi\sigma^2}(\cos^2\theta + \sin^2\theta)\}r\mathrm{d}r\mathrm{d}\theta \\
&\stackrel{(ii)}{=} \int_0^{2\pi}\int_0^{\infty} \exp\{-\frac{r^2}{2\pi\sigma^2}\}r\mathrm{d}r\mathrm{d}\theta
\end{aligned}$$

where in (i) we use $x = r\cos\theta$ and $y = r\sin\theta$ resulting from the right triangle trigonometric identity $r^2 = x^2 + y^2$, and in (ii) we use another trigonometric identity $\cos^2\theta + \sin^2\theta = 1$. We will now make the change of variables defined by the relation $u = r^2$,

$$\begin{aligned}
u &= r^2 \\
\mathrm{d}u &= 2r\mathrm{d}r \\
\frac{1}{2}\mathrm{d}u &= r\mathrm{d}r
\end{aligned}$$

so that we have

$$\begin{aligned}
I^2 &= \frac{1}{2}\int_0^{2\pi}\int_0^{\infty} \exp\{-\frac{u}{2\pi\sigma^2}\}\mathrm{d}u\mathrm{d}\theta \\
&\overset{(i)}{=} \frac{2\pi}{2}\int_0^{\infty} \exp\{-\frac{u}{2\pi\sigma^2}\}\mathrm{d}u \\
&\overset{(ii)}{=} \pi\int_0^{\infty} \exp\{-\frac{u}{2\pi\sigma^2}\}(-2\sigma^2)\Big|_0^{\infty} \\
&\overset{(iii)}{=} (-2\pi\sigma^2)\lim_{b\to\infty}\exp\{-\frac{u}{2\pi\sigma^2}\}\Big|_0^{b} \\
&\overset{(iv)}{=} (-2\pi\sigma^2)\lim_{b\to\infty}\exp\{-\frac{b}{2\pi\sigma^2}\} - \exp\{-\frac{0}{2\pi\sigma^2}\} \\
&\overset{(v)}{=} (-2\pi\sigma^2)(0-1) \\
&= 2\pi\sigma^2
\end{aligned} \tag{1}$$

where in (i) we evaluate the integral with respect to θ which is 2π since $\int_a^b \mathrm{d}z = b - a$. In (ii) we simplify the constant fraction and take the integral of the exponential which includes using the chain rule. In (iii) we begin evaluating the improper integral. In (iv) we compute the newly expressed definite proper integral. And in (v) we first note that $\exp\{-g(b)\} = 1/e^{g(b)}$ and then take the limit $b \to \infty$.

Finally, taking the square root of (1) yields $(2\pi\sigma^2)^{1/2}$. We have shown that

$$\int_{-\infty}^{\infty} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}\mathrm{d}x = (2\pi\sigma^2)^{1/2}$$

which directly implies that

$$\frac{1}{(2\pi\sigma^2)^{1/2}}\int_{-\infty}^{\infty} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}\mathrm{d}x = \int_{-\infty}^{\infty} \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}\mathrm{d}x = 1,$$

and so the normalization condition holds. ■

1.8 The Expectation and Variance of the Univariate Gaussian

1.8.1 The Expectation

Claim: The expectation of a univariate Gaussian distribution is μ.

Approach: We will use the definition of expectation which results in an integral we need to evaluate. To evaluate it we split it into two different parts, and (i) notice that one of the parts is an odd function and thus evaluates to 0, and (ii) notice that the other of the parts is a univariate Gaussian pdf which must integrate to 1 (normalization condition). The end result is that we're left with μ as the expectation.

Proof: Consider a random variable x that follows a univariate Gaussian distribution $\mathcal{N}(x|\mu, \sigma^2)$. The definition of expectation yields,

$$\begin{aligned}
\mathbb{E}[x] &\overset{(i)}{=} \int_{-\infty}^{\infty} xp(x)\mathrm{d}x \\
&\overset{(ii)}{=} \int_{-\infty}^{\infty} x\frac{1}{(2\pi\sigma^2)^{1/2}}\exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}\mathrm{d}x \\
&= \frac{1}{(2\pi\sigma^2)^{1/2}}\int_{-\infty}^{\infty} x\exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}\mathrm{d}x
\end{aligned}$$

where in (i) we use the definition of expectation and in (ii) we use the definition of the univariate Gaussian pdf. We next use a change of variables $y = x - \mu$ (alternatively $x = y + \mu$),

$$\begin{aligned}\mathbb{E}[x] &= \frac{1}{(2\pi\sigma^2)^{1/2}} \int_{-\infty}^{\infty} (y+\mu) \exp\{-\frac{1}{2\sigma^2} y^2\} \mathrm{d}y \\ &\overset{(i)}{=} \frac{1}{(2\pi\sigma^2)^{1/2}} \underbrace{\int_{-\infty}^{\infty} y \exp\{-\frac{1}{2\sigma^2} y^2\} \mathrm{d}y}_{(1)} + \underbrace{\frac{1}{(2\pi\sigma^2)^{1/2}} \int_{-\infty}^{\infty} \mu \exp\{-\frac{1}{2\sigma^2} y^2\} \mathrm{d}y}_{(2)}\end{aligned}$$

where in (i) we've distributed the $(y + \mu)$. We now need to notice that the integrand (1) is an *odd* function, meaning $f(-a) = -f(a)$. Let's first show this,

$$\begin{aligned} f(-y) &= -y \exp\{-\frac{1}{2\sigma^2}(-y)^2\} \mathrm{d}y \\ &= -y \exp\{-\frac{1}{2\sigma^2} y^2\} \mathrm{d}y \\ &= -f(y) \end{aligned}$$

Because the integrand (1) is an odd function we know, when integrated from $-\infty$ to ∞, it must be equal to 0. We are then left with (2),

$$\begin{aligned}\mathbb{E}[x] &= \frac{1}{(2\pi\sigma^2)^{1/2}} \int_{-\infty}^{\infty} \mu \exp\{-\frac{1}{2\sigma^2} y^2\} \mathrm{d}y \\ &= \mu \int_{-\infty}^{\infty} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\{-\frac{1}{2\sigma^2} y^2\} \mathrm{d}y \end{aligned} \tag{3}$$

We now need to notice that the integrand in (3) is the pdf of the univariate Gaussian (with the change of variables defined by the relation $y = x - \mu$). From the normalization condition of pdfs we know, when integrated from $-\infty$ to ∞, this must be equal to 1. So finally we have,

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

1.8.2 The Second moment

Claim: The second order moment of a Gaussian random variable x is the following,

$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) x^2 \mathrm{d}x = \mu^2 + \sigma^2 \tag{1.50}$$

Approach: As per the exercise instructions, we will differentiate both sides of the normalization condition

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

with respect to σ^2. Then we will rearrange terms to show (1.50). Perhaps some motivation behind differentiating both sides of the *normalization condition* is that the Gaussian pdf has an expression $e^{f(\sigma^2)h(x)^2}$ and so its derivative will lead to a multiplicative term with x^2 in it. Then if we we take the expectation of both sides, perhaps it is easy to rearrange so we have $\mathbb{E}[x^2]$ on one side.

Proof: Beginning with the normalization condition,

$$\begin{aligned}
\int_{-\infty}^{\infty} \mathcal{N}(x|\mu,\sigma^2)\mathrm{d}x &= 1 \\
\int_{-\infty}^{\infty} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}\mathrm{d}x &= 1 \\
\frac{1}{(2\pi\sigma^2)^{1/2}} \int_{-\infty}^{\infty} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}\mathrm{d}x &= 1 \\
\int_{-\infty}^{\infty} \exp\{-\frac{1}{2\pi\sigma^2}(x-\mu)^2\}\mathrm{d}x &= (2\pi\sigma^2)^{1/2} \qquad (1)
\end{aligned}$$

First we substitute $\sigma^2 = z$ for notational simplicity, and then we differentiate both sides of the equation (1),

$$\begin{aligned}
\frac{\partial}{\partial z}\int_{-\infty}^{\infty} \exp\{-\frac{1}{2\pi z}(x-\mu)^2\}\mathrm{d}x &= \frac{\partial}{\partial z}(2\pi z)^{1/2} \\
\overset{(i)}{\Rightarrow} \quad \int_{-\infty}^{\infty} \frac{\partial}{\partial z}\exp\{-\frac{1}{2\pi z}(x-\mu)^2\}\mathrm{d}x &= \frac{\partial}{\partial z}(2\pi z)^{1/2} \\
\overset{(ii)}{\Rightarrow} \quad \int_{-\infty}^{\infty} \frac{\partial}{\partial z}\exp\{-\frac{1}{2\pi z}(x-\mu)^2\}\mathrm{d}x &= \frac{\pi}{(2\pi z)^{1/2}} \\
\overset{(iii)}{\Rightarrow} \quad \int_{-\infty}^{\infty} \exp\{-\frac{1}{2\pi z}(x-\mu)^2\}\frac{(x-\mu)^2}{2z^2}\mathrm{d}x &= \frac{\pi}{(2\pi z)^{1/2}} \\
\frac{1}{2z^2}\int_{-\infty}^{\infty} \exp\{-\frac{1}{2\pi z}(x-\mu)^2\}(x-\mu)^2\mathrm{d}x &= \frac{\pi}{(2\pi z)^{1/2}} \\
\int_{-\infty}^{\infty} \exp\{-\frac{1}{2\pi z}(x-\mu)^2\}(x-\mu)^2\mathrm{d}x &= \frac{2\pi z^2}{(2\pi z)^{1/2}} \\
\int_{-\infty}^{\infty} \exp\{-\frac{1}{2\pi z}(x-\mu)^2\}(x-\mu)^2\mathrm{d}x &= \frac{2\pi z^2}{2^{1/2}\pi^{1/2}z^{1/2}} \\
\int_{-\infty}^{\infty} \exp\{-\frac{1}{2\pi z}(x-\mu)^2\}(x-\mu)^2\mathrm{d}x &= 2^{1/2}\pi^{1/2}z^{3/2} \\
\int_{-\infty}^{\infty} \exp\{-\frac{1}{2\pi z}(x-\mu)^2\}(x-\mu)^2\mathrm{d}x &= (2\pi z)^{1/2}z \\
\frac{1}{(2\pi z)^{1/2}}\int_{-\infty}^{\infty} \exp\{-\frac{1}{2\pi z}(x-\mu)^2\}(x-\mu)^2\mathrm{d}x &= z \\
\overset{(iv)}{\Rightarrow} \quad \frac{1}{(2\pi\sigma^2)^{1/2}}\int_{-\infty}^{\infty} \exp\{-\frac{1}{2\pi\sigma^2}(x-\mu)^2\}(x-\mu)^2\mathrm{d}x &= \sigma^2 \qquad (2)
\end{aligned}$$

where in (i) we can push the derivative with respect to z inside the integral because the integral is being taken with respect to x and the integrand has continuous partial derivatives. In (ii) we take the derivative of the right hand side, using the power rule and chain rule. In (iii) we take the derivative on the left hand side and again use the power rule and chain rule. In (iv) we substitute from z back to σ^2. We now need to notice that the left hand side is an expectation of the expression $(x-\mu^2)$ with respect to the

distribution of x (Gaussian). From (2) we then have,

$$\begin{aligned} \mathbb{E}[(x-\mu)^2] &= \sigma^2 \\ \mathbb{E}[x^2 - 2\mu x + \mu^2] &= \sigma^2 \\ \overset{(i)}{\Rightarrow} \quad \mathbb{E}[x^2] - 2\mu\mathbb{E}[x] + \mu^2 &= \sigma^2 \\ \overset{(ii)}{\Rightarrow} \quad \mathbb{E}[x^2] - 2\mu^2 + \mu^2 &= \sigma^2 \\ \mathbb{E}[x^2] &= \sigma^2 + \mu^2 \end{aligned} \tag{1.50}$$

where in (i) we use the fact that the expectation is a linear operator and that the expectation of a constant is that constant. In (ii) we use the fact that the expectation with respect to x *of* x is μ for a Gaussian distributed random variable. ■

1.8.3 The Variance

Claim: The variance of a Gaussian random variable x is

$$\begin{aligned} \text{var}[x] &= \mathbb{E}[x^2] - \mathbb{E}[x]^2 && (1) \\ &= \sigma^2 && (1.51) \end{aligned}$$

Approach: Using the results from **1.8.1** and **1.8.2**, we can substitute $\mathbb{E}[x]$ and $\mathbb{E}[x^2]$ into the right hand side of (1) to show (1.51).

Proof: From **1.8.1** we have $\mathbb{E}[x] = \mu$ and from **1.8.2** we have $\mathbb{E}[x^2] = \mu^2 + \sigma^2$. Plugging this into (1),

$$\begin{aligned} \mathbb{E}[x^2] - \mathbb{E}[x]^2 &= (\mu^2 + \sigma^2) - \mu^2 \\ &= \sigma^2 \end{aligned}$$

■

1.9 The mode of the univariate Gaussian and multivariate Gaussian

1.9.1 The mode of the univariate Gaussian

Claim: The mode of the univariate Gaussian is μ.

Approach: Differentiate the Gaussian pdf with respect to x, set this equal to 0, and solve for x to get the critical point μ. Then take the second derivative of the Gaussian pdf with respect to x, and show that, when evaluated at the critical point μ, is positive. This shows that the critical point μ is a maximum.

Proof: We begin by differentiating the univariate Gaussian pdf with respect to x,

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}x} p(x) &= \frac{\mathrm{d}}{\mathrm{d}x} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\} \\ &= \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}(-\frac{1}{\sigma^2}(x-\mu)) \qquad (1) \end{aligned}$$

We then set (1) equal to 0 and solve for x,

$$\frac{1}{(2\pi\sigma^2)^{1/2}} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}(-\frac{1}{\sigma^2}(x-\mu)) = 0$$

When $x = \mu$ the equation holds, and so $x = \mu$ is a critical point. To show that it is a maximum, we can either note that the Gaussian distribution has no minimum with respect to x since the tails go off to infinity, or we can take the second derivative of $p(x)$, and show that when this second derivative is evaluated at $x = \mu$ it is negative. If it is negative, this means the function is concave down in the neighborhood of x, and thus x is a maximum. For the purpose of explicitness, we take the second route here. Taking the second derivative of $p(x)$ amounts to taking the derivative of (1) again,

$$\begin{aligned}
\frac{\mathrm{d}^2}{\mathrm{d}x^2}p(x) &= \frac{\mathrm{d}}{\mathrm{d}x}\underbrace{\frac{1}{(2\pi\sigma^2)^{1/2}}\exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}}_{p(x)}\underbrace{(-\frac{1}{\sigma^2}(x-\mu))}_{f(x)} \\
&\overset{(i)}{=} (\frac{\mathrm{d}}{\mathrm{d}x}p(x))(f(x)) + (p(x))(\frac{\mathrm{d}}{\mathrm{d}x}f(x)) \\
&= \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}(-\frac{1}{\sigma^2}(x-\mu))(-\frac{1}{\sigma^2}(x-\mu)) \\
&+ \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}(-\frac{1}{\sigma^2}) \\
&= \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}(-\frac{1}{\sigma^2}(x-\mu))^2 \\
&+ \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}(-\frac{1}{\sigma^2}) \\
&\overset{(ii)}{=} \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}(\frac{1}{\sigma^4}(x-\mu)^2 - \frac{1}{\sigma^2}) \qquad (2)
\end{aligned}$$

where in (i) we use the product rule of derivatives, and in (ii) we factor out $p(x)$ and simplify terms. We now set $x = \mu$ in (2) to get,

$$\begin{aligned}
\frac{1}{(2\pi\sigma^2)^{1/2}}\exp\{0\}(0 - \frac{1}{\sigma^2}) &= 1(0 - \frac{1}{\sigma^2}) \\
&= -\frac{1}{\sigma^2}
\end{aligned}$$

And since σ^2 is always positive, the result is negative. This then shows that $x = \mu$ is the maximum of the Gaussian pdf.

1.9.2 The mode of the multivariate Gaussian

Claim: The mode of the multivariate Gaussian is $\boldsymbol{\mu}$.

Approach: Differentiate the multivariate Gaussian pdf with respect to $\mathbf{x}$, set it to zero, and then solve for $\mathbf{x}$. Then note that there is no minimum for a multivariate Gaussian as its tails (in all directions) tend to infinity and it is unimodal. Therefore, the value $\mathbf{x}$ takes on when the derivative is set to zero is the maximum (mode).

Proof: We begin by differentiating the univariate Gaussian pdf with respect to x,

$$\begin{aligned}
\nabla_{\mathbf{x}}\, p(\mathbf{x}) &= \nabla_{\mathbf{x}}\, \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\} \\
&\overset{(i)}{=} \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\}\nabla_{\mathbf{x}}\,[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})] \\
&\overset{(ii)}{=} \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\}(-\frac{1}{2}(\boldsymbol{\Sigma}^{-1} + (\boldsymbol{\Sigma}^{-1})^{\mathrm{T}})(\mathbf{x}-\boldsymbol{\mu})) \\
&\overset{(iii)}{=} \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\}(-\frac{1}{2}(\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-1})(\mathbf{x}-\boldsymbol{\mu})) \\
&= \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\}(-\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}))
\end{aligned}$$

Where in (i) we use the chain rule, in (ii) we use the matrix calculus identity $\frac{\partial \mathbf{z}^{\mathrm{T}}\mathbf{A}\mathbf{z}}{\partial \mathbf{z}} = (\mathbf{A}+\mathbf{A}^{\mathrm{T}})\mathbf{z}$ which holds when $\mathbf{A}$ is *not* a function of $\mathbf{z}$. In (iii) we first use the fact that if a matrix $\mathbf{A}$ is symmetric, then its inverse $\mathbf{A}^{-1}$ is also symmetric. We then use the fact that $\mathbf{\Sigma}$ is symmetric and so $(\mathbf{\Sigma}^{-1})^{\mathrm{T}} = \mathbf{\Sigma}^{-1}$. We now set the derivative equal to 0 and solve for $\mathbf{x}$.

$$\frac{1}{(2\pi)^{D/2}}\frac{1}{|\mathbf{\Sigma}|^{1/2}}\exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\}(-\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})) = 0$$

No result of an exponential can be 0, so we are left with looking at $(-\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})) = 0t$ which implies that $\mathbf{x} = \boldsymbol{\mu}$ yields zero, and so $\boldsymbol{\mu}$ is a critical point. We now note that there is no minimum for a multivariate Gaussian as its tail (in all directions) tend to infinity and it is unimodal. Therefore, $\boldsymbol{\mu}$ is the maximum (mode). ■

1.10 Linearity of expectation and variance

1.10.1 The expectation

Claim: Given two statistically independent random variables x and z, the expectation of their sum is the sum of their respective expectations,

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

Approach: Use the definition of expectation to expand the terms, use the property of independent variables that $p(x,z) = p(x)p(z)$, rearrange, and then use the definition of expectation again.

Proof:

$$\begin{aligned}
\mathbb{E}[x+z] &\overset{(i)}{=} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x+z)p(x,z)\mathrm{d}x\mathrm{d}z \\
&\overset{(ii)}{=} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x+z)p(x)p(z)\mathrm{d}x\mathrm{d}z \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xp(x)p(z)\mathrm{d}x\mathrm{d}z + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}zp(x)p(z)\mathrm{d}x\mathrm{d}z \\
&\overset{(iii)}{=} \int_{-\infty}^{\infty}xp(x)\mathrm{d}x + \int_{-\infty}^{\infty}zp(z)\mathrm{d}z \\
&\overset{(iv)}{=} \mathbb{E}[x] + \mathbb{E}[z]
\end{aligned}$$

where in (i) we use the definition of expectation, in (ii) we use the fact that $p(x,z) = p(x)p(z)$ if x and z are independent, in (iii) we use the normalization property of probability densities, and in (iv) we use the definition of expectation again. ■

In fact, we can prove $\mathbb{E}[x+z] = \mathbb{E}[x]+\mathbb{E}[z]$ even if x and z are *not* statistically independent. This is an important property of expectations which is often referred to as the *linearity of expectations*...

Claim: Given two random variables x and z, the expectation of their sum is the sum of their respective expectations,

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

Approach: Use the definition of expectation to expand the terms, rearrange and use the law of total probability (marginalization), then use the definition of expectation again.

Proof:

$$\begin{aligned}
\mathbb{E}[x+z] &\overset{(i)}{=} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x+z)p(x,z)\mathrm{d}x\mathrm{d}z \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xp(x,z)\mathrm{d}z\mathrm{d}x + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}zp(x,z)\mathrm{d}x\mathrm{d}z \\
&\overset{(ii)}{=} \int_{-\infty}^{\infty}xp(x)\mathrm{d}x + \int_{-\infty}^{\infty}zp(z)\mathrm{d}z \\
&\overset{(iii)}{=} \mathbb{E}[x] + \mathbb{E}[z]
\end{aligned}$$

where in (i) we use the definition of expectation, in (ii) we use the fact that $\int p(x,z)\mathrm{d}z = p(x)$ and $\int p(x,z)\mathrm{d}x = p(z)$ which is referred to as the law of total probability or marginalization, and in (iii) we use the definition of expectation again. ■

1.10.2 The variance

Claim: Given two statistically independent random variables x and z, the variance of their sum is the sum of their respective variances,

$$\mathrm{var}[x+z] = \mathrm{var}[x] + \mathrm{var}[z]$$

Approach: Use the variance theorem (**1.5**) to expand the terms. Then use the expectation definition and linearity of expectation to expand the terms even more, then use the independence property $p(x,z) = p(x)p(z)$, then cancel out terms to result in $\mathrm{var}[x]$ and $\mathrm{var}[z]$.

Proof:

$$\mathrm{var}[x+z] = \underbrace{\mathbb{E}[(x+z)^2]}_{\text{term 1}} - \underbrace{\mathbb{E}[(x+z)]^2}_{\text{term 2}} \tag{1}$$

where we've used the variance theorem (**1.5**). Let's turn our attention to term 1:

$$\begin{aligned}
\mathbb{E}[(x+z)^2] &\overset{(i)}{=} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x+z)^2p(x,z)\mathrm{d}x\mathrm{d}z \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x^2+2xz+z^2)p(x,z)\mathrm{d}x\mathrm{d}z \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}x^2p(x,z)\mathrm{d}x\mathrm{d}z + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}2xzp(x,z)\mathrm{d}x\mathrm{d}z + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}z^2p(x,z)\mathrm{d}x\mathrm{d}z \\
&\overset{(ii)}{=} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}x^2p(x)p(z)\mathrm{d}x\mathrm{d}z + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}2xzp(x)p(z)\mathrm{d}x\mathrm{d}z + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}z^2p(x)p(z)\mathrm{d}x\mathrm{d}z \\
&\overset{(iii)}{=} \int_{-\infty}^{\infty}x^2p(x)\mathrm{d}x + 2\int_{-\infty}^{\infty}xp(x)\mathrm{d}x\int_{-\infty}^{\infty}zp(z)\mathrm{d}z + \int_{-\infty}^{\infty}z^2p(z)\mathrm{d}z \\
&\overset{(iv)}{=} \mathbb{E}[x^2] + 2\mathbb{E}[x]\mathbb{E}[z] + \mathbb{E}[z^2]
\end{aligned} \tag{2}$$

where in (i) we use the definition of expectation, in (ii) we use the property $p(x,z) = p(x)p(z)$ if $x \perp z$, in (iii) we evaluate the first term with respect to z and third term with respect to x which both yield 1 due to the normalization condition of probability densities, and we also use the fact $\int_A\int_B f(x)g(z)\mathrm{d}x\mathrm{d}z = \int_B f(x)\mathrm{d}x\int_A g(z)\mathrm{d}z$. Finally, in (iv) we use the definition of expectation again. This leaves us with (2) for the term 1 in

(1). Let's substitute it into (1),

$$\begin{aligned}
\text{var}[x+z] &= \underbrace{\mathbb{E}[(x+z)^2]}_{\text{term 1}} - \underbrace{\mathbb{E}[(x+z)]^2}_{\text{term 2}} \\
&= \underbrace{\mathbb{E}[x^2] + 2\mathbb{E}[x]\mathbb{E}[z] + \mathbb{E}[z^2]}_{\text{term 1}} - \underbrace{\mathbb{E}[(x+z)]^2}_{\text{term 2}} \\
&\overset{(i)}{=} \mathbb{E}[x^2] + 2\mathbb{E}[x]\mathbb{E}[z] + \mathbb{E}[z^2] - (\mathbb{E}[x] + \mathbb{E}[z])^2 \\
&= \mathbb{E}[x^2] + 2\mathbb{E}[x]\mathbb{E}[z] + \mathbb{E}[z^2] - (\mathbb{E}[x]^2 + 2\mathbb{E}[x]\mathbb{E}[z] + \mathbb{E}[z]^2) \\
&= \mathbb{E}[x^2] + 2\mathbb{E}[x]\mathbb{E}[z] + \mathbb{E}[z^2] - \mathbb{E}[x]^2 - 2\mathbb{E}[x]\mathbb{E}[z] - \mathbb{E}[z]^2 \\
&= \mathbb{E}[x^2] + \mathbb{E}[z^2] - \mathbb{E}[x]^2 - \mathbb{E}[z]^2 \\
&= (\mathbb{E}[x^2] - \mathbb{E}[x]^2) + (\mathbb{E}[z^2] - \mathbb{E}[z]^2) \\
&\overset{(ii)}{=} \text{var}[x] + \text{var}[z]
\end{aligned}$$

where in (i) we use the linearity of expectation and in (ii) we use the variance theorem.

■

1.11 Maximum likelihood for univariate Gaussian data

Problem setup: We have a set of N i.i.d variables $\mathbf{x} = \{x_1, ..., x_N\}$ that each follow the univariate Gaussian distribution $\mathcal{N}(x|\mu, \sigma^2)$. Note that unless otherwise specified, the data $\mathbf{x}$ is to be considered "fixed" or "realized".

1.11.1 Estimating μ

Claim: Using maximum likelihood estimation for the parameter μ results in

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

Approach: Take the derivative of the log likelihood (1.54) with respect to μ, set it to zero and solve for μ.

Proof: We begin with taking the derivative of the log likelihood function (1.54) with respect to μ,

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

$$\begin{aligned}
\frac{\partial}{\partial\mu}\ln p(\mathbf{x}|\mu, \sigma^2) &= \frac{\partial}{\partial\mu}\left[-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)\right] \\
&\overset{(i)}{=} \frac{\partial}{\partial\mu}\left[-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2\right] \\
&\overset{(ii)}{=} \left[-\frac{1}{2\sigma^2}\sum_{n=1}^{N}\frac{\partial}{\partial\mu}(x_n - \mu)^2\right] \\
&= -\frac{1}{2\sigma^2}\sum_{n=1}^{N} 2(x_n - \mu)(-1) \\
&= \frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n - \mu) \qquad (1)
\end{aligned}$$

where in (i) we ignore all terms that do not depend on μ, and in (ii) we use the fact that $(f+g)' = (f'+g')$ but extended to finite sums. We now set the derivative (1) to zero and solve for μ,

$$\begin{aligned}
\frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n-\mu) &= 0 \\
\sum_{n=1}^{N}(x_n-\mu) &= 0 \\
\sum_{n=1}^{N}x_n - \sum_{n=1}^{N}\mu &\overset{(i)}{=} 0 \\
\sum_{n=1}^{N}x_n - N\mu &= 0 \\
\frac{1}{N}\sum_{n=1}^{N}x_n &= \mu
\end{aligned}$$

where in (i) we use the distributive property of finite sums. ■

1.11.2 Estimating σ^2

Claim: Using maximum likelihood estimation for the parameter σ^2 results in

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

Approach: Take the derivative of the log likelihood (1.54) with respect to σ^2, set it to zero and solve for σ^2.

Proof: We begin with taking the derivative of the log likelihood function (1.54) with respect to σ^2,

$$\begin{aligned}
\ln p(\mathbf{x}|\mu,\sigma^2) &= -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi) && (1.54)\\
\frac{\partial}{\partial\sigma^2}\ln p(\mathbf{x}|\mu,\sigma^2) &= \frac{\partial}{\partial\sigma^2}\left[-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)\right] \\
&= \frac{\partial}{\partial\sigma^2}\left[-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2\right] \\
&= \sum_{n=1}^{N}(x_n-\mu)^2\frac{\partial}{\partial\sigma^2}\left[-\frac{1}{2\sigma^2}\right] - \frac{\partial}{\partial\sigma^2}\left[\frac{N}{2}\ln\sigma^2\right] \\
&= \sum_{n=1}^{N}(x_n-\mu)^2\frac{\partial}{\partial\sigma^2}\left[-\frac{1}{2}(\sigma^2)^{-1}\right] - \frac{\partial}{\partial\sigma^2}\left[\frac{N}{2}\ln\sigma^2\right] \\
&= \sum_{n=1}^{N}(x_n-\mu)^2\frac{1}{2}(\sigma^2)^{-2} - \frac{N}{2}(\sigma^2)^{-1} && (1)
\end{aligned}$$

We now set (1) to zero and solve for σ^2,

$$\begin{aligned}
\sum_{n=1}^{N}(x_n-\mu)^2\frac{1}{2}(\sigma^2)^{-2}-\frac{N}{2}(\sigma^2)^{-1}&=0\\
\sum_{n=1}^{N}(x_n-\mu)^2\frac{1}{2}(\sigma^2)^{-2}&=\frac{N}{2}(\sigma^2)^{-1}\\
\sum_{n=1}^{N}(x_n-\mu)^2&=N\frac{(\sigma^2)^{-1}}{(\sigma^2)^{-2}}\\
\frac{1}{N}\sum_{n=1}^{N}(x_n-\mu)^2&=\sigma^2\\
\overset{(i)}{\Rightarrow}\quad \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{\text{ML}})^2&=\sigma^2
\end{aligned}$$

where in (i) we are using the fact that the maximum likelihood estimate for μ decouples from that for σ^2 in the case of the Gaussian. Therefore, the joint maximization is equivalent to maximizing with respect to μ to get μ_{ML} and then using that solution in the maximization with respect to σ^2. ■

1.12 Expectation of maximum likelihood estimates for univariate Gaussian

Claim: There are two claims, where the first helps to prove the second. First, given two univariate Gaussian data points x_n and x_m (which are independent if $n \neq m$),

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

where $I_{nm} = 1$ if $n = m$ and 0 otherwise. That is, if $x_n = x_m$, then the expectation evaluates to $\mu^2 + \sigma^2$. But if $x_n \neq x_m$, the expectation evaluates to μ^2. The second claim is,

$$\mathbb{E}[\mu_{\text{ML}}] = \mu \tag{1.57}$$

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

where μ_{ML} and σ^2_{ML} are the maximum likelihood estimates.

Approach: As directed by the text, we will use $\mathbb{E}[x] = \mu$ (1.49) and $\mathbb{E}[x^2] = \mu^2 + \sigma^2$ (1.50) to prove the first claim. Then to prove (1.57) and (1.58) we will expand out the maximum likelihood estimates and then simplify, eventually having to use the first claim (1.130).

Proof: We first prove the first claim (1.130). Consider two univariate Gaussian data points x_n and x_m (which are independent if $n \neq m$).

- If $n = m$ then x_n and x_m are the same variable and $x_n x_m = x^2$. And so

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

- If $n \neq m$, then $x_n \perp x_m$ and so

$$\mathbb{E}[x_n x_m] = \mathbb{E}[x_n]\mathbb{E}[x_m] = \mu^2$$

This proves the first claim. We now move onto the second claim. First we have,

$$\begin{aligned}\mathbb{E}[\mu_{\text{ML}}] &\overset{(i)}{=} \mathbb{E}[\frac{1}{N}\sum_{n=1}^{N} x_n] \\ &= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[x_n] \\ &\overset{(ii)}{=} \frac{1}{N}N\mu \\ &= \mu\end{aligned}$$

where in (i) we substitute the expression for the maximum likelihood estimate for μ, and in (ii) we use the fact $\mathbb{E}[x] = \mu$ when x is a univariate Gaussian variable. Next we have,

$$\begin{aligned}\mathbb{E}[\sigma^2_{\text{ML}}] &\overset{(i)}{=} \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{\text{ML}})^2\right] \\ &\overset{(ii)}{=} \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}(x_n - \frac{1}{N}\sum_{m=1}^{M} x_m)^2\right] \\ &= \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}(x_n^2 - \frac{2}{N}x_n\sum_{m=1}^{M} x_m + (\frac{1}{N}\sum_{m=1}^{N} x_m)^2)\right] \\ &\overset{(iii)}{=} \frac{1}{N}\mathbb{E}\left[\sum_{n=1}^{N} x_n^2 - \frac{2}{N}\sum_{n=1}^{N}(x_n\sum_{m=1}^{M} x_m) + \sum_{n=1}^{N}(\frac{1}{N}\sum_{m=1}^{N} x_m)^2\right] \\ &\overset{(iv)}{=} \frac{1}{N}\left[\underbrace{\sum_{n=1}^{N}\mathbb{E}[x_n^2]}_{(1)} - \underbrace{\frac{2}{N}\mathbb{E}\left[\sum_{n=1}^{N}(x_n\sum_{m=1}^{M} x_m)\right]}_{(2)} + \underbrace{\mathbb{E}\left[\sum_{n=1}^{N}(\frac{1}{N}\sum_{m=1}^{N} x_m)^2\right]}_{(3)}\right]\end{aligned}$$

Where in (i) we substitute the maximum likelihood expression for σ^2_{ML}, in (ii) we substitute the maximum likelihood expression for μ_{ML}. In (iii) we distribute the sum $\sum_{n=1}^{N}$, and in (iv) we use the linearity of expectation. Let's evaluate each of the specified terms (1), (2), and (3) separately. Beginning with (1),

$$\sum_{n=1}^{N}\mathbb{E}[x_n^2] \overset{(i)}{=} N(\mu^2 + \sigma^2)$$

where in (i) we use the fact $\mathbb{E}[x^2] = \mu^2 + \sigma^2$ when x is a univariate Gaussian variable. Now with (2),

$$\begin{aligned}\frac{2}{N}\mathbb{E}\left[\sum_{n=1}^{N}(x_n\sum_{m=1}^{M} x_m)\right] &\overset{(i)}{=} \frac{2}{N}\mathbb{E}\left[\underbrace{(x_1x_1 + \cdots + x_1x_N) + \cdots + (x_Nx_1 + \cdots + x_Nx_N)}_{N \text{ times}}\right] \\ &\overset{(ii)}{=} \frac{2}{N}\left(\underbrace{\mathbb{E}\left[(x_1x_1 + \cdots + x_1x_N)\right] + \cdots + \mathbb{E}\left[(x_Nx_1 + \cdots + x_Nx_N)\right]}_{N \text{ times}}\right) \quad (4) \\ &= \overset{(iii)}{=} \frac{2}{N}\left(\underbrace{(N\mu^2 + \sigma^2) + \cdots + (N\mu^2 + \sigma^2)}_{N \text{ times}}\right) \quad (5) \\ &= \frac{2}{N}(N(N\mu^2 + \sigma^2)) \\ &= 2(N\mu^2 + \sigma^2)\end{aligned}$$

where in (i) we expand out the sums, in (ii) we use the linearity of expectation, and in (iii) we use the fact that each expectation in (4) is itself the sum of expectations $\mathbb{E}[x_n x_m]$ where only once does $n = m$, leading to $N\mu^2 + \sigma^2$ in (5). Finally, in (3) we have,

$$\begin{aligned}
\mathbb{E}\left[\sum_{n=1}^{N}(\frac{1}{N}\sum_{m=1}^{N}x_m)^2\right] &\overset{(i)}{=} \mathbb{E}\left[\sum_{n=1}^{N}\frac{1}{N^2}\sum_{m=1}^{N}\sum_{k=1}^{N}x_m x_k\right] \\
&= \frac{1}{N^2}\sum_{n=1}^{N}\mathbb{E}\left[\sum_{m=1}^{N}x_m\sum_{k=1}^{N}x_k\right] \\
&= \frac{1}{N^2}\sum_{n=1}^{N}\mathbb{E}\left[\underbrace{(x_1x_1+\cdots+x_1x_N)+\cdots+(x_Nx_1+\cdots+x_Nx_N)}_{N \text{ times}}\right] \\
&= \frac{1}{N^2}\sum_{n=1}^{N}\left(\underbrace{\mathbb{E}\left[(x_1x_1+\cdots+x_1x_N)\right]+\cdots+\mathbb{E}\left[(x_Nx_1+\cdots+x_Nx_N)\right]}_{N \text{ times}}\right) \qquad (6) \\
&\overset{(ii)}{=} \frac{1}{N^2}\sum_{n=1}^{N}\left(\underbrace{(N\mu^2+\sigma^2)+\cdots+(N\mu^2+\sigma^2)}_{N \text{ times}}\right) \qquad (7) \\
&= \frac{N}{N^2}N(N\mu^2+\sigma^2) \\
&= N\mu^2+\sigma^2
\end{aligned}$$

where in (i) we use the fact $(\sum_{i=1}^{K}a_i)^2 = \sum_{i=1}^{K}a_i\sum_{j=1}^{K}a_j$, and in (ii) we use the fact that each expectation in (6) is itself the sum of expectations $\mathbb{E}[x_n x_m]$ where only once does $n = m$, leading to $N\mu^2 + \sigma^2$ in (7). Substituting the expressions we found for (1), (2), and (3) back in,

$$\begin{aligned}
&\frac{1}{N}\left[\underbrace{\sum_{n=1}^{N}\mathbb{E}[x_n^2]}_{(1)}-\underbrace{\frac{2}{N}\mathbb{E}\left[\sum_{n=1}^{N}(x_n\sum_{m=1}^{M}x_m)\right]}_{(2)}+\underbrace{\mathbb{E}\left[\sum_{n=1}^{N}(\frac{1}{N}\sum_{m=1}^{N}x_m)^2\right]}_{(3)}\right] \\
=&\frac{1}{N}\left[N(\mu^2+\sigma^2)-2(N\mu^2+\sigma^2)+(N\mu^2+\sigma^2)\right] \\
=&\frac{1}{N}\left[N\mu^2+N\sigma^2-2N\mu^2-2\sigma^2+N\mu^2+\sigma^2\right] \\
=&\mu^2+\sigma^2-2\mu^2-\frac{2}{N}\sigma^2+\mu^2+\frac{1}{N}\sigma^2 \\
=&(\mu^2-2\mu^2+\mu^2)+(\sigma^2-\frac{2}{N}\sigma^2+\frac{1}{N}\sigma^2) \\
=&(\frac{N-1}{N})\sigma^2
\end{aligned}$$

■

1.13 Estimator of the true variance

Claim: If we take the the maximum likelihood estimator of the variance of a Gaussian (1.56),

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

and switch the the maximum likelihood estimate μ_{ML} for the true value μ,

$$\begin{aligned}\sigma^2_{\text{ML}} =& \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{\text{ML}})^2 \\ \Rightarrow \quad & \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^2, \qquad \text{(new estimator)}\end{aligned}$$

then this new estimator has the property that its expectation is given by the true variance σ^2.

Approach: We take the expectation of the new estimator and this gives us (straightforwardly) the true variance σ^2.

Proof:

$$\begin{aligned}\mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^2\right] &\overset{(i)}{=} \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[x_n^2 - 2x_n\mu + \mu^2\right] \\ &\overset{(ii)}{=} \frac{1}{N}\sum_{n=1}^{N}((\mu^2 + \sigma^2) - 2\mu^2 + \mu^2) \\ &= \frac{1}{N}\sum_{n=1}^{N}(\sigma^2) \\ &= \frac{1}{N}N\sigma^2 \\ &= \sigma^2\end{aligned}$$

where in (i) we use the linearity of expectation and in (ii) we use the fact $E[x^2] = \mu^2 + \sigma^2$ for a Gaussian variable x. ■

1.14 NOT IMPLEMENTED

1.15 NOT IMPLEMENTED

1.16 NOT IMPLEMENTED

1.17 The gamma function

Claim: The gamma function is defined by

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

has the following properties:

- $\Gamma(x+1) = x\Gamma(x)$.
- $\Gamma(1) = 1$ and hence $\Gamma(x+1) = x!$ when x is an integer.

Approach: To prove the first property, as directed by the text we will use integration by parts. To prove the second property, we first evaluate $\Gamma(1)$ directly, and then illustrate how knowing that $\Gamma(x+1) = x\Gamma(x)$ and $\Gamma(1) = 1$ leads to $\Gamma(x) = x!$ when x is an integer.

Proof: We start with expressing $\Gamma(x+1)$ and $x\Gamma(x)$ by their respective definitions,

$$\begin{aligned}&\Gamma(x+1) = x\Gamma(x) \\ \overset{(i)}{\Rightarrow} \quad &\int_0^\infty u^x e^{-u}\mathrm{d}u = x\int_0^\infty u^{x-1}e^{-u}\mathrm{d}u\end{aligned}$$

where in (i) we use the definition of the Gamma function (1.141). We now evaluate the left handside of the equation, an integral, using integration by parts,

$$\int_0^\infty u^x e^{-u} \mathrm{d}u$$

Let,

$$\begin{aligned} z = u^x \quad &\Rightarrow \quad \mathrm{d}z = xu^{x-1}\mathrm{d}u \\ y = -e^{-u} \quad &\Rightarrow \quad \mathrm{d}y = e^{-u}\mathrm{d}u \end{aligned}$$

Applying integration by parts,

$$\begin{aligned} \int_0^\infty u^x e^{-u}\mathrm{d}u &= \left[-u^x e^{-u}\right]\Big|_0^\infty + \int_0^\infty xu^{x-1}e^{-u}\mathrm{d}u \\ &= \lim_{u\to\infty}\left[\frac{-u^x}{e^u}\right] + \int_0^\infty xu^{x-1}e^{-u}\mathrm{d}u \\ &= 0 + \int_0^\infty xu^{x-1}e^{-u}\mathrm{d}u \\ &= x\int_0^\infty u^{x-1}e^{-u}\mathrm{d}u \\ &\overset{(i)}{=} x\Gamma(x) \end{aligned}$$

where in (i) we use the definition of the Gamma function (1.141). Therefore $\Gamma(x+1) = x\Gamma(x)$. We now show that $\Gamma(1) = 1$ by evaluating the integral that defines it,

$$\begin{aligned} \Gamma(1) &= \int_0^\infty u^0 e^{-u}\mathrm{d}u \\ &= \int_0^\infty e^{-u}\mathrm{d}u \\ &= -e^{-u}\Big|_0^\infty \\ &= \lim_{u\to\infty}\left[-\frac{1}{e^u}\right] - (-e^0) \\ &= 0 + 1 \\ &= 1 \end{aligned}$$

And finally we show how the properties we just proved can be used to show $\Gamma(x+1) = x!$ when x is an integer,

$$\begin{aligned} \Gamma(x+1) &= x\Gamma(x) \\ &= x(x-1)\Gamma(x-1) \\ &= x(x-1)(x-2)\Gamma(x-2) \\ &\dots \\ &= x(x-1)(x-2)\dots(x-(x-2))(x-(x-1))\Gamma(x-(x-1)) \\ &= x(x-1)(x-2)\dots(2)(1)\Gamma(1) \\ &= x(x-1)(x-2)\dots(2)(1)(1) \\ &= x! \end{aligned}$$

■

1.18 Surface area and volume of a D-dimensional unit sphere

Claim: The claim has 3 parts.

- **Claim 1**: The surface area of a sphere with unit radius in D dimensions can be written as

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

- **Claim 2**: The volume of a sphere with unit radius in D dimensions can be written as

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

- **Claim 3**: When $D = 2$, (1.143) reduces to $2\pi r = 2\pi$ and (1.144) reduces to $\pi r^2 = \pi$. When $D = 3$, they reduce to $4\pi r^2 = 4\pi$ and $\frac{4}{3}\pi r^3 = \frac{4}{3}\pi$, respectively.

Approach of 1: As directed by the text we consider the following relationship involving the surface area of a sphere with unit radius in D dimensions S_D,

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

We evaluate both sides of (1) and use the definition of the Gamma function (1.141) and an intermediate result of exercise **1.7**,

$$\int_{-\infty}^{\infty} \exp\{-\frac{1}{2\sigma^2}x^2\}\mathrm{d}x = (2\pi\sigma^2)^{1/2} \tag{2}$$

to show (1.143).

Proof of 1: We begin with evaluating the left hand side integral of (1),

$$\begin{aligned}\prod_{i=1}^{D} \int_{-\infty}^{\infty} e^{-x_i^2} \mathrm{d}x_i &\overset{(i)}{=} \prod_{i=1}^{D} \pi^{1/2} \\ &= \pi^{D/2}\end{aligned}$$

where in (i) we use the fact that $\int_{-\infty}^{\infty} \exp\{-x^2\}\mathrm{d}x = (\pi)^{1/2}$ which follows from (2). Now evaluating the right hand side integral of (1),

$$S_D \int_0^{\infty} e^{-r^2} r^{D-1} \mathrm{d}r$$

Let us use a change of variables defined by the relation $u = r^2$ so that,

$$\begin{aligned} u = r^2 \quad &\Rightarrow \quad \mathrm{d}u = 2r\mathrm{d}r \\ r = u^{1/2} \quad &\Rightarrow \quad \mathrm{d}r = \frac{1}{2}u^{-(1/2)}\mathrm{d}u \end{aligned}$$

It follows then,

$$\begin{aligned} S_D \int_0^{\infty} e^{-r^2} r^{D-1} \mathrm{d}r &\overset{(i)}{=} S_D \frac{1}{2} \int_0^{\infty} e^{-u} (u^{\frac{1}{2}})^{D-1} u^{-\frac{1}{2}} \mathrm{d}u \\ &= \frac{S_D}{2} \int_0^{\infty} e^{-u} u^{\frac{D-2}{2}} \mathrm{d}u \\ &= \frac{S_D}{2} \int_0^{\infty} e^{-u} u^{\frac{D}{2}-1} \mathrm{d}u \\ &\overset{(ii)}{=} S_D \Gamma(D/2)/2 \end{aligned}$$

where in (i) we express r and dr in terms of u and $\mathrm{d}u$ and in (ii) we use the definition of the Gamma function (1.141). We now have,

$$\begin{aligned}\prod_{i=1}^{D}\int_{-\infty}^{\infty} e^{-x_i^2}\mathrm{d}x_i &= S_D\int_0^{\infty} e^{-r^2}r^{D-1}\mathrm{d}r \\ \pi^{D/2} &= S_D\Gamma(D/2)/2 \\ \frac{2\pi^{D/2}}{\Gamma(D/2)} &= S_D \end{aligned} \tag{1.143}$$

■

Approach of 2: To prove the second claim, we need to know a few basic facts about the volume and surface area of a general sphere. The first is how the surface area of a D-dimensional sphere with radius r is related to the surface area of its unit radius counterpart. The second is how volume is classically related to surface area. We use these two relationships to derive an expression with V_D and S_D. Solving the integral in that expression leads to the proof of the claim.

Proof of 2: We first note that the surface area of a D-dimensional sphere with radius r can be written as

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

where S_D is the surface area of the unit sphere, a constant. Secondly, we note that a D-dimensional sphere can be thought of as a union of spherical shells, so the volume is the integration of a bunch of surface areas of each of the shells:

$$\begin{aligned} V_D(r) &= \int_0^r S_D(q)\mathrm{d}q \\ &\overset{(i)}{=} \int_0^r S_D q^{D-1}\mathrm{d}q \\ &= S_D\int_0^r q^{D-1}\mathrm{d}q \end{aligned}$$

where (i) follows directly from (1). Because we want the volume of a sphere with unit radius, we set $r = 1$:

$$\begin{aligned} V_D &= S_D\int_0^1 q^{D-1}\mathrm{d}q \\ &= S_D\left[\frac{1}{D}r^D\right]\Big|_0^1 \\ &= S_D(\frac{1}{D}) \\ &= \frac{S_D}{D} \end{aligned}$$

■

Approach of 3: As provided by the text (and from what we've proved before), we use $\Gamma(1) = 1$ and $\Gamma(3/2) = \sqrt{\pi}/2$ to show the results of claim 3.

Proof of 3: When $D = 2$,

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

$$V_D = \frac{S_D}{D} = \frac{2\pi}{2} = \pi$$

When $D = 3$,

$$S_D = \frac{2\pi^{D/2}}{\Gamma(D/2)} = \frac{2\pi^{3/2}}{\Gamma(3/2)} = \frac{2\pi^{3/2}}{\pi^{1/2}/2} = 4\pi$$

$$V_D = \frac{S_D}{D} = \frac{4\pi}{3} = \frac{4}{3}\pi$$

■

1.19 Volume of cube as its dimension grows

1.19.1 Relationship between volume of sphere and volume of cube

Claim: Given a D-dimensional sphere with radius a and a concentric hypercube with side $2a$, so that the sphere touches the hypercube at the centres of each of its sides, it is true that

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

Approach: We use the results of the last question **1.18** as well as the relationship between volume $V_D(r)$ and surface area $S_D(r)$ to show (1.145).

Proof: Consider a D-dimensional sphere with radius a and a concentric hypercube with side $2a$ so that the sphere touches the cube at the centres of each of the cubes' sides. Let's first focus on the sphere: as outlined in the proof of **Claim 2** of question **1.18**, a D-dimensional sphere with radius a has volume $V_D(a)$ which is related to its surface area $S_D(a)$ in the following way,

$$V_D(a) = \int_0^a S_D(r)\mathrm{d}r \tag{2}$$

and further, the surface area $S_D(r)$ is related to the surface area of a D-dimensional unit sphere by

$$S_D(r) = S_D r^{D-1}. \tag{3}$$

Using (2) and (3) we have,

$$\begin{aligned}
\text{volume of sphere} &= V_D(a) \\
&\overset{(i)}{=} \int_0^a S_D(r)\mathrm{d}r \\
&\overset{(ii)}{=} \int_0^a S_D r^{D-1}\mathrm{d}r \\
&= S_D \int_0^a r^{D-1}\mathrm{d}r \\
&= S_D \left[\frac{1}{D}r^D\right]\Big|_0^a \\
&= S_D(\frac{1}{D}a^D) \\
&\overset{(iii)}{=} \frac{2\pi^{D/2}}{\Gamma(D/2)}(\frac{1}{D}a^D) \\
&= \frac{2\pi^{D/2}a^D}{D\Gamma(D/2)}
\end{aligned}$$

where in (i) we use (2), in (ii) we use (3), and in (iii) we use the result from last question (1.143). Moving onto the cube,

$$\text{volume of cube} = (2a)^D.$$

And so together we have,

$$\begin{aligned}\frac{\text{volume of sphere}}{\text{volume of cube}} &= \frac{2\pi^{D/2}a^D}{2^D D\Gamma(D/2)a^D} \\ &= \frac{\pi^{D/2}}{D2^{D-1}\Gamma(D/2)}\end{aligned}$$

■

1.19.2 Volume of cube as its dimension tends to infinity

Claim: Given a D-dimensional sphere with radius a and a concentric hypercube with side $2a$, so that the sphere touches the hypercube at the centres of each of its side, it is true that

$$\lim_{D\to\infty} \frac{\text{volume of sphere}}{\text{volume of cube}} = 0. \tag{1}$$

Furthermore, 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 $\sqrt{D}$. Thus, this ratio tends 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".

Approach: As directed by the text, we make use of Stirling's formula in the form

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

which is valid for $x \gg 1$. Making use of both (2) and (1.145), we show (1). To show the second claim about the ratio of the distance between the center and corner of the cube and the distance between the center and side of the cube, we use the general distance formula for Euclidean spaces to get an expression for both the distances, and then show their ratio is $\sqrt{D}$.

Proof: First let us write $\Gamma(D/2)$ using (2),

$$\Gamma(D/2) \approx (2\pi)^{1/2} \left[e^{-D/2} e\right] \left[(D/2)^{D/2}(D/2)^{1/2}\right] \tag{3}$$

Now consider

$$\begin{aligned}\frac{\text{volume of sphere}}{\text{volume of cube}} &= \frac{\pi^{D/2}}{D2^{D-1}\Gamma(D/2)} && (1.145) \\ &\overset{(i)}{=} \frac{(\pi e)^{D/2}}{D2^{D-1}(2\pi)^{1/2}(D/2)^{D/2}(D/2)^{1/2}e} && (4)\end{aligned}$$

Where in (i) we used (3). $(D/2)^{D/2}$ will dominate as the limit $D \to \infty$ is taken of (4), and so

$$\lim_{D\to\infty} \frac{\text{volume of sphere}}{\text{volume of cube}} = 0.$$

This demonstrates that for high dimension D, much of the volume contained in the cube is in its corners...where the volume is not shared with the sphere.

We now show 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 $\sqrt{D}$. First let's denote the center of the hypercube as a vector $(0, 0, \ldots, 0) \in \mathbb{R}^D$. Then one of the corners

of the hypercube can be represented as the vector $(a, a, \ldots, a) \in \mathbb{R}^D$. One of the sides of the hypercube is the one in moving in the 1st dimension by a from the center, so we can represent that point as a vector $(a, 0, \ldots, 0) \in \mathbb{R}^D$. We now note the general formula for the distance in Euclidean space between two D-dimensional vectors (an extension of Pythagorean's theorem),

$$d(\mathbf{v}, \mathbf{u}) = \sqrt{(v_1 - u_1)^2 + (v_2 - u_2)^2 + \cdots + (v_D - u_D)^2}. \tag{5}$$

Applying (5) to the center of the hypercube and one of its corners,

$$\begin{aligned} d(\text{center, a corner}) &= d((0, 0, \ldots, 0), (a, a, \ldots, a)) \\ &= \sqrt{\underbrace{(0-a)^2 + (0-a)^2 + \cdots + (0-a)^2}_{D \text{ times}}} \\ &= \sqrt{Da^2} \\ &= \sqrt{D}a \end{aligned} \tag{6}$$

Now applying (5) to the center of the hypercube and one of its sides,

$$\begin{aligned} d(\text{center, a corner}) &= d((0, 0, \ldots, 0), (a, 0, \ldots, 0)) \\ &= \sqrt{\underbrace{(0-a)^2 + (0-0)^2 + \cdots + (0-0)^2}_{D \text{ times}}} \qquad = \sqrt{a^2} \\ &= a \end{aligned} \tag{7}$$

So the ratio of the distance between the center and one of the hypercubes' corners (6) and the distance between the center and one of the hypercubes' sides (7) is

$$\frac{\sqrt{D}a}{a} = \sqrt{D}.$$

And so we see that as the dimensionality D tends to ∞, the hypercube's corners become very long "spikes". ■