

Contents

1 Solutions for exercises to chapter 1	2
Problem 1.1 - Closed form solution to polynomial regression	2
Problem 1.2 - Closed form solution to regularized polynomial regression	3
Problem 1.3 - Bayes formula warm up	3
Problem 1.4 - Nonlinear transform of likelihood function doesn't preserve its extrema . . .	4
Problem 1.5 - Characterization of variance	4
Problem 1.6 - Covariance of two independent r.v. is zero	4
Problem 1.7 - Gaussian integral via polar coordinate	5
Problem 1.8 - Second moment of gaussian integral via Feymann's trick	5
Problem 1.9 - Gaussian density peaks at mean	6
Problem 1.10 - Linearity of expectation and variance	7
Problem 1.11 - MLE of gaussian	7
Problem 1.12 - Inconsistency gaussian MLE	8
Problem 1.14 - Independent terms of 2-nd order term in polynomial	8
Problem 1.16 - Independent terms of high order polynomial	10
Problem 1.17 - Gamma density warmup	11
Problem 1.18 - Volume of unit sphere in n-space	11
Problem 1.19 - High dimensional cubes concentrate on corners	13
Problem 1.20 - High dimensional gaussian concentrate on a thin strip	14
Problem 1.21 - Upper bound of bayesian classification error	16
Problem 1.22 - Uniform loss maximizes posterior probability	16

Chapter 1

Solutions for exercises to chapter 1

Problem 1.1 - Closed form solution to polynomial regression

We use a slightly better notation to write this problem. Let X be the matrix of the form

$$X = \begin{bmatrix} x_1^0 & x_1^1 & \cdots & x_1^M \\ x_2^0 & x_2^1 & \cdots & x_2^M \\ \vdots & \vdots & \ddots & \vdots \\ x_N^0 & x_N^1 & \cdots & x_N^M \end{bmatrix}, \quad t = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{bmatrix}$$

The the problem can be rewritten in the following form:

$$E(w) = \frac{1}{2} \left((Xw - t)^T (Xw - t) \right).$$

Now we differentiate w.r.t w , note that

$$\begin{aligned} E(w + h) &= \frac{1}{2} (X(w + h) - t)^T (X(w + h) - t) \\ &= \frac{1}{2} \left((Xw - t)^T + (Xh)^T \right) (Xw - t + Xh) \\ &= \frac{1}{2} \left[(Xw - t)^T (Xw - t) + (Xw - t)^T Xh + (Xh)^T (Xw - t) + (Xh)^T (Xh) \right] \\ &= E(w) + \left\langle (Xw - t)^T, Xh \right\rangle + \frac{1}{2} \langle Xh, Xh \rangle \\ &= E(w) + \left\langle X^T (Xw - t), h \right\rangle + \frac{1}{2} \langle Xh, Xh \rangle. \end{aligned}$$

Note that $\langle X^T (Xw - t), h \rangle \in \text{Hom}(\mathbb{R}^{M+1}, \mathbb{R})$ and

$$\frac{1}{2} \langle Xh, Xh \rangle \leq \frac{1}{2} \|Xh\| \|Xh\| \leq \frac{C}{2} \|X\|_\infty^2 \|h\| \xrightarrow{\|h\| \rightarrow 0} 0,$$

it follows that $\nabla E(w) = X^T (Xw - t)$. Set it to zero and we get

$$X^T (Xw - t) = 0 \iff X^T Xw = X^T t.$$

So $X^T X$ is the A proposed in the problem.

$$[X^T X]_{ij} = \sum_{n=1}^N (x_n^i x_n^j) = \sum_{n=1}^N x_n^{i+j}, \text{ and } [X^T t]_i = \sum_{n=1}^N x_n^i t_n,$$

as desired.

Problem 1.2 - Closed form solution to regularized polynomial regression

We use the same notation as in the previous problem and still rewrite the loss function in matrix form as follows:

$$\tilde{E}(w) = \frac{1}{2} \langle Xw - t, Xw - t \rangle + \frac{\lambda}{2} \langle w, w \rangle.$$

Still we differentiate the expression. Note that if we let $\varphi(w) = \frac{\lambda}{2} \langle w, w \rangle$, we have that

$$\begin{aligned} \varphi(w+h) &= \frac{\lambda}{2} (w+h)^T (w+h) \\ &= \frac{\lambda}{2} (w^T w + w^T h + h^T w + \|h\|^2) \\ &= \varphi(w) + \langle \lambda w, h \rangle + \underbrace{\frac{\lambda}{2} \|h\|^2}_{=o(\|h\|)}. \end{aligned}$$

Therefore, $\nabla \varphi(w) = \lambda w$, and as a result

$$\nabla \tilde{E}(w) = \nabla E(w) + \nabla \varphi(w) = X^T (Xw - t) + \lambda w.$$

Setting it to zero:

$$X^T (Xw - t) + \lambda w = 0 \iff (X^T X + \lambda I)w = X^T t.$$

Hence, $(X^T X + \lambda I)$ and $X^T t$ are the corresponding matrices.

Problem 1.3 - Bayes formula warm up

According to the Bayes formula, we get that

$$\begin{aligned} P(\text{apple}) &= P(\text{apple}|\text{r}) P(\text{r}) + P(\text{apple}|\text{g}) P(\text{g}) + P(\text{apple}|\text{b}) P(\text{b}) \\ &= \frac{3}{10} \cdot \frac{2}{10} + \frac{1}{2} \frac{2}{10} + \frac{3}{10} \frac{6}{10} = \frac{17}{50}. \end{aligned}$$

And again, we can use formula to get

$$\begin{aligned} P(\text{g}|\text{orange}) &= \frac{P(\text{orange}|\text{g}) P(\text{g})}{P(\text{orange}|\text{g}) P(\text{g}) + P(\text{orange}|\text{b}) P(\text{b}) + P(\text{orange}|\text{r}) P(\text{r})} \\ &= \frac{\frac{3}{10} \frac{6}{10}}{\frac{3}{10} \frac{6}{10} + \frac{2}{10} \frac{1}{2} + \frac{2}{10} \frac{4}{10}} \\ &= \frac{1}{2}. \end{aligned}$$

Problem 1.4 - Nonlinear transform of likelihood function doesn't preserve its extrema

We first observe that if x_* maximizes the likelihood function $p_x(x)$, then $p'_x(x_*) = 0$. By chain rule, we have that

$$\begin{aligned} \frac{dp_x(g(y))}{dy} |g'(y)| &= \frac{dp_x(g(y))}{dy} |g'(y)| + p_x(g(y)) \frac{d|g'(y)|}{dy} \\ &= \frac{dp_x(g(y))}{dg(y)} \frac{dg(y)}{dy} |g'(y)| + p_x(g(y)) \frac{d|g'(y)|}{dy}. \end{aligned} \quad (1)$$

Hence, if $x_* = g(y_*)$, the

$$\frac{dp_x(g(y_*))}{dg(y_*)} = \frac{dp_x(x_*)}{dx_*} = 0.$$

However, there is no guarantee that the second term of the RHS of Eq. 1 is zero. For example, if $p_x(x) = 2x$ for $0 \leq x \leq 1$ and $x = \sin(y)$, where $0 \leq y \leq \pi/2$. Then according to the transformation formula, we have that

$$p_y(y) = p_x(g(y))g'(y) = 2\sin(y)\cos(y) = \sin(2y) \text{ for } 0 \leq y \leq \frac{\pi}{2}.$$

Clearly, $p_y(y)$ reaches its peak at $y = \pi/4$ but $\sin(\pi/4) \neq x_* = 1$. Thus, we have found a counterexample.

On the other hand, if $g(y)$ is an affine map, then $g'(y)$ is a constant map and as a result

$$\frac{d|g'(y)|}{dy} = 0$$

Problem 1.5 - Characterization of variance

It suffices to show that $\text{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$ since any a measurable function of a random variable is again a random variable and in this case f although is not mentioned, it is safe to assume in this context that f is measurable. So note

$$\begin{aligned} \text{Var}[X] &= \mathbb{E}[X - \mathbb{E}[X]]^2 \\ &= \mathbb{E}[X^2] - 2\mathbb{E}[X]\mathbb{E}[X] + (\mathbb{E}[X])^2 \\ &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \end{aligned}$$

as desired.

Problem 1.6 - Covariance of two independent r.v. is zero

Since $X \perp Y$, then it follows that $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$. Then we have

$$\begin{aligned} \text{Cov}(X, Y) &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= \mathbb{E}[XY - X\mathbb{E}[Y] - \mathbb{E}[X]Y + \mathbb{E}[X]\mathbb{E}[Y]] \end{aligned}$$

$$\begin{aligned}
&= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y] \\
&= \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y] \\
&= 0.
\end{aligned}$$

Problem 1.7 - Gaussian integral via polar coordinate

First, we write

$$\begin{aligned}
I^2 &= \left(\int_{\mathbb{R}} \exp \left\{ -\frac{1}{2\sigma^2} x^2 \right\} dx \right) \left(\int_{\mathbb{R}} \exp \left\{ -\frac{1}{2\sigma^2} y^2 \right\} dy \right) \\
&= \int_{\mathbb{R} \times \mathbb{R}} \exp \left\{ -\frac{1}{2\sigma^2} (x^2 + y^2) \right\} dx dy.
\end{aligned}$$

Now using polar coordinate - let $x = r \cos \theta$ and $y = r \sin \theta$. Then we get the Jacobian matrix as

$$\frac{\partial(x, y)}{\partial(r, \theta)} = \begin{bmatrix} \cos \theta & -r \sin \theta \\ \sin \theta & r \cos \theta \end{bmatrix} \implies \left| \frac{\partial(x, y)}{\partial(r, \theta)} \right| = r(\cos^2 \theta + \sin^2 \theta) = r.$$

Hence, as a result

$$\begin{aligned}
I^2 &= \int_0^{2\pi} \int_0^\infty \exp \left\{ -\frac{r^2}{2\sigma^2} \right\} r dr d\theta \\
&= \int_0^{2\pi} \int_0^\infty \exp(-u) \sigma^2 du d\theta \\
&= \int_0^{2\pi} \sigma^2 d\theta \int_0^\infty \exp(-u) du \\
&= 2\pi \sigma^2 [-\exp(-u)]_0^\infty = 2\pi \sigma^2.
\end{aligned}$$

Problem 1.8 - Second moment of gaussian integral via Feymann's trick

The differentiation under the integral needs a bit more theoretical justification. We won't reproduce the related theorems here. But they could be found in e.g. Theorem 3.2, Theorem 3.3 in Chapter XIII of [Lan97] or in [Con00]. With this in mind, we get

$$\begin{aligned}
\frac{d}{d\sigma^2} \int_{\mathbb{R}} \exp \left\{ -\frac{1}{2\sigma^2} (x - \mu)^2 \right\} dx &= \int_{\mathbb{R}} \frac{d}{d\sigma^2} \exp \left\{ -\frac{1}{2\sigma^2} (x - \mu)^2 \right\} dx \\
&= \int_{\mathbb{R}} \exp \left\{ -\frac{1}{2\sigma^2} (x - \mu)^2 \right\} (x - \mu)^2 \left(-\frac{1}{2} \right) (\sigma^{-2})^2 dx
\end{aligned}$$

On the the other hand, we have

$$\frac{d}{d\sigma^2} (2\pi \sigma^2)^{1/2} = -\frac{1}{2} (2\pi) (\sigma^2)^{-1/2}.$$

So combined together, we get

$$\int_{\mathbb{R}} \exp \left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\} (x-\mu)^2 \left(-\frac{1}{2} \right) (\sigma^{-2})^2 dx = \left(-\frac{1}{2} \right) (2\pi)^{1/2} (\sigma^2)^{-1/2}.$$

One step of reduction, we get

$$\begin{aligned} \mathbb{E}[(x - \mathbb{E}[x])^2] &= \text{Var}[x] \\ &= \frac{1}{(2\pi\sigma^2)^{1/2}} \int_{\mathbb{R}} \exp \left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\} (x-\mu)^2 dx \\ &= \sigma^2. \end{aligned}$$

And as a result,

$$\mathbb{E}[x^2] = \text{Var}[x] + (\mathbb{E}[x])^2 = \sigma^2 + \mu^2.$$

Problem 1.9 - Gaussian density peaks at mean

It suffices to show the result holds in the multidimensional case since 1-dim is just a special case. Recall that the density of the Gaussian distribution in D dimension is

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

Differentiate w.r.t. x and we get:

$$\nabla_x N(x|u, \Sigma) = \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu) \right\} \nabla_x \left(\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu) \right).$$

Now note that $\varphi(x) = (x-\mu)^T \Sigma^{-1}(x-\mu)$ for $x \in \mathbb{R}^d$, then note for any $h \in \mathbb{R}^D$

$$\begin{aligned} \varphi(x+h) &= (x-u+h)^T \Sigma^{-1}(x-\mu+h) \\ &= (x-\mu)^T \Sigma^{-1}(x-\mu+h) + h^T \Sigma^{-1}(x-\mu+h) \\ &= (x-\mu)^T \Sigma^{-1}(x-\mu) + (x-\mu)^T \Sigma^{-1}h + h^T \Sigma^{-1}(x-\mu) + h^T \Sigma^{-1}h \\ &= (x-\mu)^T \Sigma^{-1}(x-\mu) + \langle 2\Sigma^{-1}(x-\mu), h \rangle + h^T \Sigma^{-1}h \end{aligned}$$

Note that and

$$h^T \Sigma^{-1}h = \left\langle h\Sigma^{-1/2}, h\Sigma^{-1/2} \right\rangle \leq \left\| h\Sigma^{-1/2} \right\|^2 \leq C \|h\|^2 \|\Sigma\|_{\infty}^2 = o(\|h\|),$$

and that $\langle 2\Sigma^{-1}(x-\mu), h \rangle \in \text{Hom}(\mathbb{R}^d, \mathbb{R})$. It follows that

$$\nabla_x \varphi(x) = 2\Sigma^{-1}(x-\mu),$$

whence

$$\nabla_x \varphi(x) = 0 \iff 2\Sigma^{-1}(x-\mu) = 0 \iff x = \mu.$$

Problem 1.10 - Linearity of expectation and variance

1. Note

$$\begin{aligned}
\mathbb{E}[x + y] &= \int_{\text{supp}(x)} \int_{\text{supp}(y)} (x + y) f_{(x,y)}(x, y) dx dy \\
&= \int_{\text{supp}(x)} \int_{\text{supp}(y)} (x + y) f_x(x) f_y(y) dx dy \\
&= \int_{\text{supp}(x)} \int_{\text{supp}(y)} x f_x(x) f_y(y) dx dy + \int_{\text{supp}(x)} \int_{\text{supp}(y)} y f_x(x) f_y(y) dx dy \\
&= \int_{\text{supp}(x)} x f_x(x) dx \int_{\text{supp}(y)} f_y(y) dy + \int_{\text{supp}(x)} f_x(x) dx \int_{\text{supp}(y)} y f_y(y) dy \\
&= \mathbb{E}[x] + \mathbb{E}[y].
\end{aligned}$$

2. Note

$$\begin{aligned}
\text{Var}[x + y] &= \mathbb{E}[x + y]^2 - (\mathbb{E}[x + y])^2 \\
&= \mathbb{E}[x^2] + \mathbb{E}[y^2] + \underbrace{2\mathbb{E}[xy]}_{\mathbb{E}[x]\mathbb{E}[y]} - (\mathbb{E}[x])^2 - (\mathbb{E}[y])^2 - 2\mathbb{E}[x]\mathbb{E}[y] \\
&= \mathbb{E}[x^2] - (\mathbb{E}[x])^2 + \mathbb{E}[y^2] - (\mathbb{E}[y])^2 \\
&= \text{Var}[x] + \text{Var}[y].
\end{aligned}$$

Problem 1.11 - MLE of gaussian

Recall that the log-likelihood function for Gaussian distribution is

$$\ln p(x|\mu, \sigma^2) = -\frac{1}{2\sigma^2} \sum_{i=1}^N (x_n - \mu)^2 - \frac{N}{2} \ln \sigma^2 - \frac{N}{2} \ln(2\pi).$$

Now we differentiate it w.r.t. μ and setting it to zero:

$$\frac{\partial \ln p(x|\mu, \sigma^2)}{\partial \mu} = -\frac{1}{2\sigma^2} \cdot 2 \cdot \sum_{i=1}^N (x_n - \mu) = 0 \iff \sum_{i=1}^N (x_n - \mu) = 0 \iff \mu_{ML} = \frac{1}{n} \sum_{i=1}^N x_n.$$

Now we differentiate it w.r.t. σ^2 and setting it to zero:

$$\frac{\partial \ln(p|\mu, \sigma^2)}{\partial \sigma^2} = \underbrace{\sum_{n=1}^N (x_n - \mu)^2 \left(-\frac{1}{2}\right) (-1)(\sigma^2)^{-2} - \frac{N}{2\sigma^2}}_{(\star)} = 0.$$

To rearrange, we get

$$(\star) \iff \sum_{n=1}^N (x_n - \mu)^2 \sigma^{-4} = \frac{N}{\sigma^2}$$

$$\begin{aligned} &\iff \sum_{n=1}^N (x_n - \mu)^2 = \sigma^2 N \\ &\iff \sigma_{ML}^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \mu)^2. \end{aligned}$$

Plug in $\mu = \mu_{ML}$ we get $\sigma_{ML}^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \mu)^2$ as desired.

Problem 1.12 - Inconsistency gaussian MLE

Problem 1.14 - Independent terms of 2-nd order term in polynomial

We rewrite the sum in matrix form: $\sum_{i=1}^D \sum_{j=1}^D w_{ij} x_i x_j = x^T W x$, where $[W]_{ij} = w_{ij}$. Define

$$W_S = \frac{1}{2}(W + W^T) \text{ and } W_A = \frac{1}{2}(W - W^T).$$

Clearly, W_S is symmetric and $W_A^T = \frac{1}{2}(W^T - W) = -W_A$ is anti-symmetric and $W_S + W_A = W$. Therefore,

$$x^T W x = x^T (W_S + W_A) x = x^T W_S x + x^T W_A x.$$

Notice that

$$x^T W_A x = \frac{1}{2}(x^T W_S x - x^T W^T x) = \frac{1}{2}(x^T W_S x - x^T W x) = 0,$$

where the last inequality follows from the fact that $x^T W^T x$ is a scalar and is equal to $x^T W x$. Since we have shown the sum, $\sum_{i,j} w_{ij} x_i x_j$, only depends on a symmetric matrix, W_S , whose independent items is of the cardinality of $\sum_{i=1}^D i = D(D+1)/2$ if we assume its of dimension $D \times D$, we have established our claim.

Problem 1.15 - Independent terms of M -th order term in polynomial

1. Since by writing the M -th order in the form of

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

introduces duplicate terms, e.g. if $w_{1,3,2} x_1 x_3 x_2$ and $w_{2,3,1} x_2 x_3 x_1$ are the same and can be combined into $(w_{1,3,2} + w_{2,3,1}) x_1 x_2 x_3$, we can introduce an ordering that prevents such duplication from happening. Rewrite the sum in the newly introduced ordering yields

$$\sum_{i_1=1}^D \sum_{i_2=1}^{i_1} \cdots \sum_{i_M=1}^{i_{M-1}} w_{i_1, i_2, \dots, i_M} x_{i_1} x_{i_2} \cdots x_{i_M}.$$

Thus, we have

$$\begin{aligned}
 n(D, M) &= \sum_{i_1=1}^D \sum_{i_2=1}^{i_1} \cdots \sum_{i_M=1}^{i_{M-1}} w_{i_1, i_2, \dots, i_M} x_{i_1} x_{i_2} \cdots x_{i_M} \\
 &= \sum_{i_1=1}^D \left(\sum_{i_2=1}^{i_1} \cdots \sum_{i_M=1}^{i_{M-1}} w_{i_1, i_2, \dots, i_M} x_{i_1} x_{i_2} \cdots x_{i_M} \right) \\
 &= \sum_{i_1=1}^D n(i_1, M-1).
 \end{aligned}$$

2. To show the equality holds using induction, we note for the base case of $D = 1$,

$$\text{LHS} = \frac{(1 + M - 2)!}{0!(M-1)!} = \frac{(M-1)!}{(M-1)!} = 1.$$

And

$$\text{RHS} = \frac{(1 + M - 1)!}{(D-1)!M!} = \frac{M!}{M!} = 1.$$

Now suppose $D = k$ and the equality holds. Then

$$\begin{aligned}
 \sum_{i=1}^{k+1} \frac{(i + M - 2)!}{(i-1)!(M-1)!} &= \sum_{i=1}^k \frac{(i + M - 2)!}{(i-1)!(M-1)!} + \frac{(k+1 + M - 2)!}{k!(M-1)!} \\
 &= \frac{(k + M - 1)!}{(k-1)!M!} + \frac{(k + M - 1)!}{k!(M-1)!} \\
 &= \frac{(k + M - 1)!(k + M)}{k!(M-1)!} \\
 &= \frac{(k + M)!}{k!M!} \\
 &= \frac{((k+1) + M - 1)!}{(k+1-1)!M!},
 \end{aligned} \tag{1}$$

where Eq. (1) follows from induction hypothesis.

3. We establish the identity by inducting on M . By Problem 1.14, it follows that

$$n(D, 2) = \frac{1}{2}D(D+1) = \frac{(D+2-1)!}{(D-1)!2!} = \frac{(D+1)!}{(D-1)!2!},$$

which proves the base case. Now suppose the statement holds for $M = k$. Then for $M = k+1$, we have

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

using part-2.

Problem 1.16 - Independent terms of high order polynomial

1. The first equality just follows from that summing up all the independent terms:

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

2. We prove this inequality by inducting on M . Now for the base case, $M = 0$, we note that

$$\text{LHS} = n(D, 0) = \frac{(D+0-1)!}{(D-1)!0!} = 1 = \frac{(D+0)!}{D!0!} = \text{RHS}.$$

Now assume that the claim holds for $M = k$. Then for $M = k + 1$, we have

$$\begin{aligned} N(D, k+1) &= \sum_{i=0}^k n(D, i) + n(D, k+1) \\ &= \frac{(D+k)!}{D!k!} + \frac{(D+k+1-1)!}{(D-1)!(k+1)!} \\ &= \frac{(D+k)!(D+k+1)}{D!(k+1)!} \\ &= \frac{(D+k+1)!}{D!(k+1)!}, \end{aligned}$$

proving the inducting step.

3. Now we show that $N(D, M)$ grows in polynomial fashion like D^M . Assume $D \ll M$. First, we write

$$\begin{aligned} N(D, M) &= \frac{(D+M)!}{D!M!} \\ &\simeq \frac{(D+M)^{D+M} e^{-(D+M)}}{D!M^M e^{-M}} \quad (\text{by Stirling's approximation}) \\ &= \frac{1}{D!M^M} \left(1 + \frac{D}{M}\right)^{D+M} M^{D+M} \frac{e^{-(D+M)}}{e^{-M}} \\ &= \frac{e^{-D}}{D!} \left(1 + \frac{D}{M}\right)^{D+M} M^D. \end{aligned} \tag{1}$$

Now we take a more delicate look at the term $(1 + \frac{D}{M})^{D+M}$. Note that

$$\begin{aligned} \left(1 + \frac{D}{M}\right)^{D+M} &= \left(1 + \frac{D}{M}\right)^M \left(1 + \frac{D}{M}\right)^D \\ &= \left(\left(1 + \frac{1}{M/D}\right)^{M/D}\right)^D \left(1 + \frac{D}{M}\right)^D \\ &\leq e^D 2^D, \end{aligned}$$

where the inequality comes from the fact that $(1 + 1/x)^x$ is an increasing function and $D < M \Rightarrow$

$D/M \leq 1$. Substitution back into Eq (1), we get

$$N(D, M) \leq \frac{e^{-D}}{D!} e^D 2^D M^D = \frac{2^D}{D!} M^D.$$

The case for $M \ll D$ follows by symmetry.

Problem 1.17 - Gamma density warmup

1. Note

$$\begin{aligned} \Gamma(x+1) &= \int_0^\infty u^x e^{-u} du \\ &= [-u^x e^{-u}]_{u=0}^\infty + \int_0^\infty x u^{x-1} e^{-u} du \\ &= x \Gamma(x). \end{aligned}$$

2. We note that

$$\Gamma(1) = \int_0^\infty e^{-u} du = [e^{-u}]_0^\infty = 1.$$

And as a result, by recursion

$$\Gamma(x+1) = x \Gamma(x) = \cdots = x! \text{ for } x \in \mathbb{N}.$$

Problem 1.18 - Volume of unit sphere in n-space

To state the problem statement in a clearer manner, we solve this problem in several steps. In this problem, we let $d\mu$ denote the Lebesgue measure.

1. First we derive Eq (1.142) in the book. We first rewrite the LHS in the following way. Let $x \in \mathbb{R}^D$ be arbitrary, then

$$\begin{aligned} \int_{\mathbb{R}^d} e^{-\|x\|^2} dx &= \int_{\mathbb{R}} \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} e^{-(x_1^2 + x_2^2 + \cdots + x_n^2)} dx_1 dx_2 \cdots dx_n \\ &= \prod_{i=1}^D \int_{\mathbb{R}} e^{-x_i^2} dx_i. \end{aligned}$$

Next, we evaluate this integral. In order to make the computation easier, we choose to let the integrand be $e^{-\pi\|x\|^2}$ instead (it doesn't effect the final result, and one could always get the original integral by scaling). Note that using the same argument as above, we have

$$\int_{\mathbb{R}^D} e^{-\pi\|x\|^2} dx = \left(\int_{\mathbb{R}} e^{-\pi x^2} dx \right)^D.$$

Next, we have

$$\left(\int_{\mathbb{R}} e^{-\pi x^2} dx \right)^2 = \left(\int_{\mathbb{R}} e^{-\pi x_1^2} dx_1 \right) \left(\int_{\mathbb{R}} e^{-\pi x_2^2} dx_2 \right)$$

$$\begin{aligned}
&= \int_{\mathbb{R} \times \mathbb{R}} e^{-\pi(x_1^2 + x_2^2)} d(x_1 \times x_2) && \text{(by Fubini's theorem)} \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} e^{-\pi(x_1^2 + x_2^2)} dx_1 dx_2 && \text{(by Fubini's theorem)} \\
&= \int_{[0, 2\pi]} \int_{\mathbb{R}} e^{-\pi r^2} r dr d\theta && \text{(switch to polar coordinates)} \\
&= \int_{[0, 2\pi]} d\theta \int_{\mathbb{R}} e^{-\pi r^2} r dr \\
&= 2\pi \left[-\frac{1}{2\pi} e^{-\pi r^2} \right]_0^\infty \\
&= 1.
\end{aligned}$$

Since $\int_{\mathbb{R}} e^{\pi x^2} dx > 0$, it follows that $\int_{\mathbb{R}^D} e^{-\pi \|x\|^2} dx = 1$.

2. Consider the function $f : \mathbb{R}^D \rightarrow \mathbb{R}; x \mapsto e^{-\pi \|x\|^2}$. We just showed in part-1 that $f \in L^1(\mathbb{R}^D)$. Therefore, using generalized spherical coordinate (e.g. Theorem 6.3.4 in [Ste05]), we have that

$$\begin{aligned}
1 &= \int_{\mathbb{R}^D} f(x) dx = \int_{S^{D-1}} \left(\int_{\mathbb{R}^+} f(r\gamma) r^{D-1} dr \right) d\sigma(\gamma) \\
&= \int_{S^{D-1}} \left(\int_{\mathbb{R}^+} e^{-\pi \|r\gamma\|^2} r^{D-1} dr \right) d\sigma(\gamma) \\
&= \int_{S^{D-1}} \left(\int_{\mathbb{R}^+} e^{-\pi r^2} r^{D-1} dr \right) d\sigma(\gamma) \\
&= \int_{S^{D-1}} d\sigma(r) \int_{\mathbb{R}^+} e^{-\pi r^2} r^{D-1} dr \\
&= \sigma(S^{D-1}) \int_{\mathbb{R}^+} e^{-\pi r^2} r^{D-1} dr.
\end{aligned}$$

Now we evaluate the integral on the RHS:

$$\begin{aligned}
\int_{\mathbb{R}^+} e^{-\pi r^2} r^{D-1} dr &= \int_0^\infty e^{-u} \left(\frac{u}{\pi} \right)^{\frac{D-1}{2}} \frac{1}{2\pi(u/\pi)^{1/2}} du \\
&= \frac{1}{2\pi} \int_0^\infty e^{-u} \left(\frac{u}{\pi} \right)^{\frac{D}{2}-1} du \\
&= \frac{1}{2\pi} \pi^{1-\frac{D}{2}} \int_0^\infty e^{-u} u^{\frac{D}{2}-1} du \\
&= \frac{1}{2} \pi^{-\frac{D}{2}} \Gamma\left(\frac{D}{2}\right).
\end{aligned}$$

Therefore, substituting back we get

$$\sigma(S^{D-1}) = \frac{1}{\int_{\mathbb{R}^+} e^{-\pi r^2} r^{D-1} dr} = \frac{2\pi^{D/2}}{\Gamma(D/2)}.$$

This $\sigma(S^{D-1})$ is the S_D in the problem.

3. Now we calculate the volume of the ball. Let B_1 denote the unit ball in \mathbb{R}^D . Note that again by

generalized spherical coordinate,

$$\begin{aligned}
 V_D &= \int_{\mathbb{R}^D} \mathbb{1}_{B_1}(x) d\mu \\
 &= \int_{S^{D-1}} \int_{\mathbb{R}^+} \mathbb{1}_{B_1}(r\gamma) r^{D-1} d\sigma(\gamma) \\
 &= \int_{S^{D-1}} \left(\int_{[0,1]} r^{D-1} dr \right) d\sigma(\gamma) \\
 &= \left(\int_{S^{D-1}} d\sigma(\gamma) \right) \left(\int_{[0,1]} r^{D-1} dr \right) \\
 &= \sigma(S^{D-1}) \left[\frac{1}{D} r^D \right]_0^1 \\
 &= \frac{\pi^{D/2}}{\Gamma(D/2)(D/2)} \\
 &= \frac{\pi^{D/2}}{\Gamma(D/2 + 1)}.
 \end{aligned}$$

as desired.

4. When $D = 2$, we get

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

When $D = 2$, we get

$$S_D = \frac{2\pi^{3/2}}{\Gamma(3/2)} = \frac{2\pi^{3/2}}{\pi^{1/2}/2} = 4\pi \text{ and } V_D = \frac{4}{3}\pi.$$

Remark 1.1. This problem could have been solved heuristically. But it loses rigor. What was showed was a rigorous mathematical way to treat this problem.

Problem 1.19 - High dimensional cubes concentrate on corners

1. Using the result of the previous problem, and the fact that $m_d(rB) = r^d m(B)$, where m_d is the Lebesgue measure in d -dimensional Euclidean space (e.g. Exercise 1.6 in [Ste05]), we have that

$$\begin{aligned}
 \frac{V_{\text{sphere}}}{V_{\text{cube}}} &= \frac{\pi^{D/2} a^D}{\Gamma(D/2 + 1) 2^D a^D} = \frac{\pi^{D/2}}{\Gamma(D/2 + 1) 2^D} \\
 &\simeq \frac{\pi^{D/2}}{(2\pi)^{1/2} e^{-D/2} (D/2)^{D/2+1/2} 2^D} && \text{(by Stirling formula)} \\
 &= C \frac{\pi^{D/2} e^{D/2}}{(D/2)^{D/2}} \frac{1}{D^{1/2}} 2^{-D} && (C \text{ is some constant}) \\
 &= C \left(\frac{2\pi e}{D} \right)^{D/2} \frac{1}{D^{1/2} 2^D} \xrightarrow{D \rightarrow \infty} 0.
 \end{aligned}$$

2. On the other hand, we have

$$\begin{aligned}\text{dist}(\text{center to corner}) &= \sqrt{Da^2} = a\sqrt{D} \\ \text{dist}(\text{center to top}) &= a.\end{aligned}$$

And thus the ratio is \sqrt{D} .

Problem 1.20 - High dimensional gaussian concentrate on a thin strip

First, note that the density given in the problem is that of a Gaussian in D dimensional Euclidean space with $\Sigma = \text{diag}(\sigma^2)$.

1. To show that the density is of the form exhibited in (1.148), we note that again by generalized spherical coordinate we have

$$\begin{aligned}\int_{\mathbb{R}^D} p(x)dx &= \int_{S^{D-1}} \int_{\mathbb{R}^+} p(\gamma r) dr d\sigma(\gamma) \\ &= \int_{S^{D-1}} \int_{\mathbb{R}^+} \frac{1}{(2\pi\sigma^2)^{D/2}} \exp\left\{-\frac{\|\gamma r\|^2}{2\sigma^2}\right\} r^{D-1} dr d\sigma(\gamma) \\ &= \int_{S^{D-1}} \int_{\mathbb{R}^+} \frac{1}{(2\pi\sigma^2)^{D/2}} \exp\left\{-\frac{\|\gamma\|^2 r^2}{2\sigma^2}\right\} r^{D-1} dr d\sigma(\gamma) \\ &= \int_{S^{D-1}} d\sigma(\gamma) \int_{\mathbb{R}^+} \frac{1}{(2\pi\sigma^2)^{D/2}} \exp\left\{-\frac{r^2}{2\sigma^2}\right\} r^{D-1} dr \\ &= \sigma(S^{D-1}) \int_{\mathbb{R}^+} \frac{1}{(2\pi\sigma^2)^{D/2}} \exp\left\{-\frac{r^2}{2\sigma^2}\right\} r^{D-1} dr.\end{aligned}$$

This is the formula in (1.148) if we relabel $\sigma(S^{D-1}) = S_D$.

2. First, we note

$$\begin{aligned}\frac{d}{dr}p(r) &= C \cdot \frac{d}{dr} \left[r^{D-1} \exp\left\{-\frac{r^2}{2\sigma^2}\right\} \right] \\ &= C \cdot \left[(D-1)r^{D-2} \exp\left\{-\frac{r^2}{2\sigma^2}\right\} + r^{D-1} \exp\left\{-\frac{r^2}{2\sigma^2}\right\} \left(-\frac{1}{\sigma^2}\right) 2r \right] \\ &= C \left[(D-1)r^{D-2} - \frac{r^D}{\sigma^2} \right] \exp\left\{-\frac{r^2}{2\sigma^2}\right\}.\end{aligned}$$

To find the stationary point, we set it to zero:

$$\begin{aligned}\frac{d}{dr}p(r) = 0 &\iff C \left[(D-1)r^{D-2} - \frac{r^D}{\sigma^2} \right] \exp\left\{-\frac{r^2}{2\sigma^2}\right\} = 0 \\ &\iff (D-1)r^{D-2} - \frac{r^D}{\sigma^2} = 0 \\ &\iff \hat{r} = \sqrt{(D-1)\sigma^2} \simeq \sqrt{D}\sigma,\end{aligned}$$

where the approximation follows since $\sqrt{D+1} = \sqrt{D}$ for large D .

3. To show (1.149), first we note

$$\begin{aligned}
p(\hat{r} + \varepsilon) &= \frac{S_D(\hat{r} + \varepsilon)^{D-1}}{(2\pi\sigma^2)^{D/2}} \exp\left\{-\frac{(\hat{r} + \varepsilon)^2}{2\sigma^2}\right\} \\
&= \frac{S_D}{(2\pi\sigma^2)^{D/2}} \exp\left\{-\frac{(\hat{r} + \varepsilon)^2}{2\sigma^2} + (D-1)\log(\hat{r} + \varepsilon)\right\} \\
&= \frac{S_D}{(2\pi\sigma^2)^{D/2}} \exp\left\{-\frac{(\hat{r} + \varepsilon)^2}{2\sigma^2} + (D-1)\left[\log\left(1 + \frac{\varepsilon}{\hat{r}}\right) + \log \hat{r}\right]\right\} \\
&= \frac{S_D \hat{r}^{D-1}}{(2\pi\sigma^2)^{D/2}} \exp\left\{-\frac{\hat{r}^2}{2\sigma^2} - \frac{\hat{r}\varepsilon}{\sigma^2} - \frac{\varepsilon^2}{2\sigma^2} + (D-1)\left(\frac{\varepsilon}{\hat{r}} - \frac{\varepsilon^2}{2\hat{\gamma}^2} + o\left(\frac{\varepsilon^2}{\hat{\gamma}^2}\right)\right)\right\} \\
&= \underbrace{\frac{S_D \hat{r}^{D-1}}{(2\pi\sigma^2)^{D/2}} \exp\left\{-\frac{\hat{r}^2}{2\sigma^2}\right\}}_{=p(r)} \underbrace{\exp\left\{-\frac{\hat{r}\varepsilon}{\sigma^2} - \frac{\varepsilon^2}{2\sigma^2} + (D-1)\left(\frac{\varepsilon}{\hat{r}} - \frac{\varepsilon^2}{2\hat{\gamma}^2} + o\left(\frac{\varepsilon^2}{\hat{\gamma}^2}\right)\right)\right\}}_{:=\mathcal{E}(\varepsilon, \sigma, \hat{r})}. \quad (1)
\end{aligned}$$

Now, we just need to massage last term in the RHS of (1): since $\hat{r} = \sqrt{D-1}\sigma$, we get

$$\begin{aligned}
\mathcal{E}(\varepsilon, \sigma, \hat{r}) &= \exp\left\{-\frac{\sqrt{D-1}\varepsilon}{\sigma} - \frac{\varepsilon^2}{2\sigma^2} + \frac{\sqrt{D-1}\varepsilon}{\sigma} - \frac{\varepsilon^2}{2\sigma^2} + o\left(\frac{\varepsilon^2}{\hat{\gamma}^2}\right)\right\} \\
&= \exp\left\{-\frac{\varepsilon^2}{\sigma^2}\right\} \exp\left\{o\left(\frac{\varepsilon^2}{\hat{\gamma}^2}\right)\right\}.
\end{aligned}$$

Since by assumption $\varepsilon \ll \hat{r}$, it follows that $\mathcal{E}(\varepsilon, \sigma, \hat{r}) \simeq \exp\{-\varepsilon^2/\sigma^2\}$. Substituting back we get

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

as desired.

4. Note that we have

$$p(x=0) = \frac{1}{(2\pi\sigma^2)^{D/2}},$$

and

$$\begin{aligned}
p(x \in \Gamma | \Gamma = \{\gamma \in \mathbb{R}^d | \|\gamma\| = \sqrt{D-1}\sigma\}) &= \frac{1}{(2\pi\sigma^2)^{D/2}} \exp\left\{-\frac{(D-1)\sigma^2}{2\sigma^2}\right\} \\
&= \frac{1}{(2\pi\sigma^2)^{D/2}} \exp\left\{-\frac{D-1}{2}\right\},
\end{aligned}$$

whence

$$\begin{aligned}
\frac{p(x \in \Gamma | \Gamma = \{\gamma \in \mathbb{R}^d | \|\gamma\| = \sqrt{D-1}\sigma\})}{p(x=0)} &= \exp\left\{-\frac{D-1}{2}\right\} \\
&\simeq \exp\left\{-\frac{D}{2}\right\} \text{ when } D \text{ is large}
\end{aligned}$$

Problem 1.21 - Upper bound of bayesian classification error

1. Since $x \mapsto \sqrt{x}$ is monotonically increasing and $a \leq b$, it follows that $0 \leq a^{1/2} \leq b^{1/2}$, which then implies $a \leq a^{1/2}b^{1/2}$ after multiplying both sides with $a^{1/2}$.
2. To show the desired inequality, we note (for notation, we let \mathcal{X} be the ambient input space),

$$\begin{aligned}
\mathbb{P}(\text{mistake}) &= \int_{\mathcal{R}_1} \mathbb{P}(x, \mathcal{C}_2) dx + \int_{\mathcal{R}_2} \mathbb{P}(x, \mathcal{C}_1) dx \\
&\leq \int_{\mathcal{R}_1} \mathbb{P}(x, \mathcal{C}_1) \mathbb{P}(x, \mathcal{C}_2) dx + \int_{\mathcal{R}_2} \mathbb{P}(x, \mathcal{C}_1) \mathbb{P}(x, \mathcal{C}_2) dx && \text{(by part-1)} \\
&= \int_{\mathcal{R}_1 \cup \mathcal{R}_2} \mathbb{P}(x, \mathcal{C}_1) \mathbb{P}(x, \mathcal{C}_2) dx \\
&= \int_{\mathcal{X}} \mathbb{P}(x, \mathcal{C}_1) \mathbb{P}(x, \mathcal{C}_2) dx,
\end{aligned}$$

where the last inequality follows since we are working in a two-class setting and the fact that decision regions partition the input space.

Problem 1.22 - Uniform loss maximizes posterior probability

For concise notation, we write the loss matrix as $L = \mathbb{1}\mathbb{1}^T - I$, where here $\mathbb{1}$ stands for vector of 1's and $\vec{\mathbb{P}}(\mathcal{C}|x)$ as a vector of $\mathbb{P}(\mathcal{C}_k, x)$'s. Then we can rewrite Eq. (1.81) in the book as

$$\begin{aligned}
\min_k \sum_k L_{kj} \mathbb{P}(\mathcal{C}_k|x) &= \min \vec{\mathbb{P}}(\mathcal{C}|x)^T (\mathbb{1}\mathbb{1}^T - I) e_j \\
&= \min \vec{\mathbb{P}}(\mathcal{C}|x)^T \mathbb{1} - \mathbb{P}(\mathcal{C}_j|x) \\
&= \min 1 - \mathbb{P}(\mathcal{C}_j|x) \\
&= \max_j \mathbb{P}(\mathcal{C}_j|x).
\end{aligned}$$

where the second equality follows from the fact the conditional distribution sums to 1.

We can interpret this loss in the following way: this loss assigns unit weight to each misclassified labels and zero weight to correctly classified labels and therefore minimizing the expectation represents minimizing the misclassification rate.

Bibliography

C

[Con00] Keith Conrad. Differentiating under the integral sign. 2000. [5](#)

L

[Lan97] Serge Lang. *Undergraduate Analysis*. Springer-Verlag New York, 2 edition, 1997. [5](#)

S

[Ste05] Elias Stein. *Real analysis : measure theory, integration, and Hilbert spaces*. Princeton University Press, Princeton, N.J. Oxford, 2005. [12](#), [13](#)