Contents

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

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

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

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

So X^TX is the A proposed in the problem.

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

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:

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

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

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

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

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

Setting it to zero:

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

Hence, $(X^TX + \lambda I)$ and X^Tt 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

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

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 $0 \le x \le 1$ and $x = \sin(y)$, where $0 \le y \le \pi/2$. Then according to the transformation formula, we have that

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

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 $\mathrm{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$ 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

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

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

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

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

Problem 1.7 - Gaussian integral via polar coordinate

First, we write

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

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

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

Hence, as a result

$$
\begin{aligned}
I^2 &= \int_0^{2\pi}\int_0^{\infty} \exp\left\{-\frac{r^2}{2\sigma^2}\right\} r dr d\theta \\
&=_1 \int_0^{2\pi}\int_0^{\infty} \exp(-u)\sigma^2 du d\theta \\
&= \int_0^{2\pi} \sigma^2 d\theta \int_0^{\infty} \exp(-u) du \\
&= 2\pi\sigma^2 \left[-\exp(-u)\right]_0^{\infty} = 2\pi\sigma^2.
\end{aligned}
$$

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

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

And as a result,

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

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 $h \in \mathbb{R}^D$

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

Note that and

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

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

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

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:

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

To rearrange, we get

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

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

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

Plug in $\mu = \mu_{ML}$ we get $\sigma^2_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu)^2$ 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: $\sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}x_ix_j = x^TWx$, 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,

$$x^TWx = x^T(W_S+W_A)x = x^TW_Sx + x^TW_Ax.$$

Notice that

$$x^TW_Ax = \frac{1}{2}(x^TW_Sx - x^TW^Tx) = \frac{1}{2}(x^TW_Ss - x^TWx) = 0,$$

where the last inequality follows from the fact that x^TW^Tx is a scalar and is equal to x^TWx. Since we have shown the sum, $\sum_{i,j} w_{ij}x_ix_j$, 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

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

introduces duplicate terms, e.g. if $w_{1,3,2}x_1x_3x_2$ and $w_{2,3,1}x_2x_3x_1$ are the same and can be combined into $(w_{1,3,2}+w_{2,3,1})x_1x_2x_3$, we can introduce an ordering that prevents such duplication from happening. Rewrite the sum in the newly introduced ordering yields

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

Thus, we have

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

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

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

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

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

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

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

2. We note that

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

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 $f : \mathbb{R}^D \to \mathbb{R}; x \mapsto e^{-\pi\|x\|^2}$. 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:

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

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

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

2. On the other hand, we have

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

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

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

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

2. First, we note

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

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

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

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

3. To show (1.149), first we note

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

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),

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

   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

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

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.

Problem 1.23 - Characterization for minimizing general expected loss

Note

$$\sum_k L_{kj}\mathbb{P}(\mathcal{C}_k|x) = \frac{1}{p(x)} \sum_k L_{kj}\mathbb{P}(x|\mathcal{C}_k)\mathbb{P}(\mathcal{C}_k).$$

Suppose $m = \min(\sum_k L_{kj}\mathbb{P}(x|\mathcal{C}_k)$, if we increase $\mathbb{P}(\mathcal{C}_k)$, we would have to decrease L_{kj} to keep the minimum. Hence, there is a direct trade-off between $\mathbb{P}(\mathcal{C}_k)$ and L_{kj}.

Problem 1.24 - Duality between decision and rejection criterion

1. According to Eq. (1.81) in the book, the decision of labels is found by computing $\arg\min_j \sum_k L_{kj}\mathbb{P}(\mathcal{C}_k|x)$. Since rejection option is also used, let $\hat{j}$ be the minimum, then the decision criterion can be modeled as a function $\varphi : \mathbb{N} \to \mathbb{N} \cup \{\emptyset\}$ by

$$j \mapsto \begin{cases} \arg\min_j \sum_k L_{kj}\mathbb{P}(\mathcal{C}_k|x) & \text{if } \min_j \sum_k L_{kj}\mathbb{P}(\mathcal{C}_k|x) \\ \emptyset & \text{otherwise} \end{cases}.$$

   Note the j defined in φ by default refers to the minimizer of $\sum_k L_{kj}\mathbb{P}(\mathcal{C}_k|x)$, and the mapping to empty set means rejection.

2. When $L = \mathbb{1}\mathbb{1}^T - I$, then we have by previous part that

$$\begin{aligned} \varphi(\hat{j}) = j &\iff \min_j \sum_k L_{kj}\mathbb{P}(\mathcal{C}_k|x) \le \lambda \\ &\iff \min_j 1 - \mathbb{P}(\mathcal{C}_j|x) \le \lambda \qquad \text{(by Problem 1.22)} \\ &\iff \max \mathbb{P}(\mathcal{C}_k|x) \ge 1 - \lambda. \end{aligned}$$

   Note that the last stipulation is equivalent to $\theta = 1 - \lambda$ in the reject option definition. Hence, the two criteria coincide when $\theta = 1 - \lambda$.

Problem 1.25 - Generalized squared loss function

We follow the same procedure as in the 1 dimensional case. Note

$$\begin{aligned} \frac{\delta \mathbb{E}[L]}{\delta L} &= \frac{\delta}{\delta L}\left[\int\int \|y(x) - t\|^2 p(t,x)dxdt\right] \\ &= \int 2(y(x) - t)p(t,x)dt. \end{aligned}$$

Setting it to zero yields:

$$\begin{aligned} y(x)\int p(t,x)dt = \int tp(t,x)dt &\iff y(x) = \frac{\int tp(t,x)dt}{\int p(t,x)dt} \\ &\iff y(x) = \frac{\int tp(t,x)dt}{p(x)} = \int tp(t|x)dt \end{aligned}$$

as desired.

Problem 1.26 - Decomposition of expected squared loss

We use the similar argument as in deriving Eq. (1.90) in here. We write

$$\begin{aligned} \|y(x) - t\|^2 &= \|y(x) - \mathbb{E}[t|x] + \mathbb{E}[t|x] - t\|^2 \\ &= \|y(x) - \mathbb{E}[t|x]\|^2 - 2(y(x) - \mathbb{E}[y|x])^T(\mathbb{E}[t|x] - t) + \|\mathbb{E}[t|x] - t\|^2. \end{aligned}$$

Also note that we can rewrite $\mathbb{E}[t|x]-t = \mathbb{E}[t|x]-\mathbb{E}[\mathbb{E}[t|x]]$ and that $\mathbb{E}[y(x)-\mathbb{E}[y|x]] = \mathbb{E}[y]-\mathbb{E}[y] = 0$. Hence

$$\begin{aligned}\mathbb{E}\left[\|y(x)-t\|^2\right] &= \int \|y(x)-\mathbb{E}[t|x]\|^2 p(x)dx + \int \|\mathbb{E}[t|x-t\|^2 p(x)dx \\ &= \int \|y(x)-\mathbb{E}[t|x]\|^2 p(x)dx + \int \mathrm{Var}[t|x]p(x)dx.\end{aligned}$$

Hence, we see that $\mathbb{E}[\|y(x)-t\|^2]$ is minimized when $y(x) = \mathbb{E}[t|x]$, which is analogues to Eq. (1.90).

Problem 1.27 - Maximizer of L_1, L_{0^+} expected loss

According to Eq. (1.91), an application of Fubini's theorem we can rewrite the expected Minkowski loss in the following form:

$$\mathbb{E}[L] = \int \underbrace{\int |y(x)-t|^q \, p(x,t)dt}_{:=G(x,y(x))} \, dx$$

Here we need assume $G(x, y(x))$ converges uniformly so that we can differentiate under the improper (possibly) integral. As usual, we compute the first variation:

$$\begin{aligned}\frac{\delta\mathbb{E}[L]}{\delta y(x)} = \frac{\partial G(x,y(x))}{\partial y(x)} &= \int q\,|y(x)-t|^{q-1}\,\frac{(y(x)-t)}{|y(x)-t|}p(x,t)dt \\ &= p(x)\int q\,|y(x)-t|^{q-1}\,\mathrm{sgn}(y(x)-t)p(t|x)dt \\ &= p(x)\left(\int_{\{t\le y(x)\}} q\,|y(x)|^{q-1}\,p(t|x)dt - \int_{t>y(x)} q\,|y(x)-t|^{q-1}\,p(t|x)dt\right) \qquad (1)\end{aligned}$$

To find the stationary point when $q = 1$, we set Eq.(1) to zero:

$$\begin{aligned}& p(x)\left(\int_{\{t\le y(x)\}} q\,|y(x)|^{q-1}\,p(t|x)dt - \int_{\{t>y(x)\}} q\,|y(x)-t|^{q-1}\,p(t|x)dt\right) = 0 \\ \implies_1 & \int_{\{t\le y(x)\}} p(t|x)dt = \int_{\{t>y(x)\}} p(t|x), \qquad (2)\end{aligned}$$

where $\implies_1$ follows since we only need to care about $x \in \mathrm{supp}(p(x))$. Hence, the $y(x)$ that maximizes the expected loss function with $q = 1$ satisfies Eq.(2), which is the definition of the median.

Now we consider the case when $p \to 0$. Instead of taking the functional derivative, we will use a more delicate and analytical approach. First, we write

$$\mathbb{E}[L] = \lim_{q\to 0}\int |y(x)-t|^q \, d(F_x \times F_t).$$

Observe that $\lim_{q\to 0}|y(x)-t|^q = \mathbb{1}_{\{y(x)\neq t\}}$, which is in $L_2(\Omega)$ (we use Ω to denote the probability space). An application of DCT yields

$$\begin{aligned}\mathbb{E}[L] &= \int \mathbb{1}_{\{y(x)\neq t\}} d(F_x \times F_t) \\ &= \int d(F_x \times F_t) - \int \mathbb{1}_{\{y(x)=t\}} d(F_x \times F_t) \\ &= 1 - \underbrace{\int\int \mathbb{1}_{\{y(x)=t\}} p(x,t)dtdx}_{:=\mathcal{I}_1(y(x),x,t)}. \qquad \text{(by change of variable theorem)}\end{aligned}$$

In order to minimize $\mathbb{E}[L]$, it suffices to find $\arg\max_{y(x)} \mathcal{I}_1(y(x),x,t)$. First, we rewrite

$$\mathcal{I}_1(y(x),x,t) = \int p(x) \underbrace{\int \mathbb{1}_{\{t=y(x)\}} p(t|x)dt}_{:=\mathcal{I}_2(y(x),x,t)} dx.$$

Since $p(x) \geq 0$ for any $x \in \mathbb{R}^n$ and $\mathcal{I}_2(y(x),x,t) \geq 0$, it follows that $\arg\max_{y(x)} \mathcal{I}_1(y(x),x,t) = \arg\max_{y(x)} \mathcal{I}_2(y(x),x,t)$. Note that $\mathcal{I}_2(y(x),x,t) = 0$ since it is an integral w.r.t to a singleton point whose Lebesgue measure is zero. However, we can circumvent this in the following manner: note that

$$\begin{aligned}\mathcal{I}_2(y(x),x,t) &= \int \lim_{n\to\infty} \mathbb{1}_{\{t\in(y(x)-\frac{1}{2n},y(x)+\frac{1}{2n})\}} p(t|x)dt \\ &= \lim_{n\to\infty} \int \mathbb{1}_{\{t\in(y(x)-\frac{1}{2n},y(x)+\frac{1}{2n})\}} p(t|x)dt \\ &\leq \lim_{n\to\infty} \frac{1}{n} \sup_{t\in(y(x)-\frac{1}{2n},y(x)+\frac{1}{2n})} p(t|x) \qquad (3)\end{aligned}$$

If we define $F_n(y(x)) = \frac{1}{n}\sup_{t\in(y(x)-1/(2n),y(x)+1/(2n))} p(t|x)$, then it follows from Eq.(3) that $F_n(y(x)) \to 0$ as $n \to \infty$ for any $y(x) \in \mathbb{R}$. However, we would like to find a $\tilde{y}(x)$ s.t. $F_n(\tilde{y}(x)) \leq F_n(y(x))$ for any other choice of $y(x)$. We claim that $\tilde{y}(x) = \arg\max_{t\in\mathbb{R}} p(t|x)$. Indeed, if so, we have

$$\begin{aligned}F_n(\tilde{y}(x)) &= \frac{1}{n} \sup_{t\in(\arg\max_t p(t|x)-\frac{1}{2n},\arg\max_t p(t|x)+\frac{1}{2n})} p(t|x) = \frac{1}{n}\sup_{t\in\mathbb{R}^n} p(t|x) \\ &\geq \frac{1}{n} \sup_{t\in(y(x)-1/(2n),y(x)+1/(2n))} p(t|x) = F(y(x)).\end{aligned}$$

So to translate into heuristic terms, $y(x) = \arg\max_t p(t|x)$ minimizes the loss function in "each step" of the process of "approaching the limit of $q \to 0$".

Problem 1.28 - Derivation of information content

Assuming the randoms variables to be discrete does simplifies the argument but it also losses rigor. To achieve maximum amount of rigor possible, we use a measure theoretic language. For this reason, we will use a slightly different formulation, but the idea remains the same.

Instead of using x, y to denote random variable, we use X, Y. Note that for any $A \in \mathcal{B}(X)$,

$B \in \mathcal{B}(Y)$, where $\mathcal{B}$ denote the Borel sets, if X and Y are independent,

$$h(X \in A, Y \in B) = h\bigg(\int_{A\times B} d(F_X \times F_Y)\bigg) = h\bigg(\int_A dF_X \cdot \int_B dF_Y\bigg). \qquad \text{(by independence)}$$
$$= h(X \in A) + h(Y \in B) = h\bigg(\int_A dF_X\bigg) + h\bigg(\int_B dF_Y\bigg).$$

If we let $x = \int_A dF_X$ and $y = \int_B dF_Y$, then this problem reduces to the following form: find a representation of h such that $h(xy) = h(x) + h(y)$ for any $x, y \in [0, 1]$. This is variant of the Cauchy Functional Equation problem.

Recall that the Cauchy functional equation in its standard form is as follows: find a function f that satisfies $f(x + y) = f(x) + f(y)$. The obvious solution to f is the linear one: $x \mapsto cx$ for $x \in \mathbb{R}^n$. However, without additional assumptions, one can obtain other complicated solutions as well. But generally these solutions serves as pedagogical example. A classical result is that if we assume f to be either continuous or monotone, then $f(x) = cx$ for arbitrary $c \in \mathbb{R}$ is then the only solution. (cf. [Kuc09]).

With this in mind, to solve for h, we define $g(x) = h(e^x)$. Then we see that

$$g(x + y) = h(e^x e^y) = h(e^x) + h(e^y) = g(x) + g(y).$$

Then if we require g to be continuous, then $g(x)$ is uniquely represented as cx for any $c \in \mathbb{R}$. Then note that for any $x \in \mathbb{R}^+$,

$$h(x) = h(e^{\ln x}) = g(\ln x) = c \ln x, \text{ for any } c \in \mathbb{R}.$$

Therefore, we have $h(x) \propto \ln(x)$ as desired.

Problem 1.29 - Upper bound for entropy of discrete variables

We directly apply Jensen's inequality:

$$H(x) = -\sum_{i=1}^{M} p(x_i) \ln(x_i) = \sum_{i=1}^{M} p(x_i) \ln \frac{1}{p(x_i)} \leq \ln\bigg(\sum_{i=1}^{M} p(x_i) \frac{1}{p(x_i)}\bigg) = \ln M,$$

where the $\leq$ follows since $\ln(x)$ is concave.

Problem 1.30 - KL-divergence for Gaussian

We use the original definition of KL-divergence:

$$KL(p||q) = \underbrace{-\int p(x) \ln q(x) dx}_{(1)} \underbrace{-\bigg(-\int p(x) \ln p(x) dx\bigg)}_{(2)}.$$

We compute it term by term, first note that

$$\begin{aligned}
(1) &= -\int \varphi(x|\mu,\sigma^2)\ln\left[\varphi(x|m,s^2)\right]dx \\
&= -\int \varphi(x|\mu,\sigma^2)\left\{\ln\frac{1}{(2\pi s^2)^{1/2}} - \frac{(x-m)^2}{2s^2}\right\} \\
&= \int \varphi(x|\mu,\sigma^2)\left[\frac{1}{2}\ln(2\pi s^2) + \frac{(x-m)^2}{2s^2}\right]dx \\
&= \frac{1}{2}\int \varphi(x|\mu,\sigma^2)\ln(2\pi s^2)dx + \frac{1}{2s^2}\left[\int \varphi(x|\mu,\sigma^2)x^2dx - 2m\int \varphi(x|\mu,\sigma^2)xdx + \int \varphi(x|\mu,\sigma^2)m^2dx\right] \\
&= \frac{1}{2}\ln(2\pi s^2) + \frac{1}{2s^2}\left[\sigma^2+\mu^2-2m\mu+m^2\right].
\end{aligned}$$

And similarly

$$\begin{aligned}
(2) &= -\int \varphi(x|\mu,\sigma^2)\ln\left[\varphi(x|\mu,\sigma^2)\right]dx \\
&= \frac{1}{2}\int \varphi(x|\mu,\sigma^2)\ln(2\pi\sigma^2)dx + \frac{1}{2\sigma^2}\left[\int \varphi(x|\mu,\sigma^2)x^2dx - 2\mu\int \varphi(x|\mu,\sigma^2)xdx + \mu^2\int \varphi(x|\mu,\sigma^2)dx\right] \\
&= \frac{1}{2}\ln(2\pi\sigma^2) + \frac{1}{2\sigma^2}\left[\sigma^2+\mu^2-2\mu^2+\mu^2\right] \\
&= \frac{1}{2}\ln(2\pi\sigma^2) + \frac{1}{2}.
\end{aligned}$$

Hence, it follows that

$$\begin{aligned}
KL(p||q) &= \frac{1}{2}\ln(2\pi s^2) + \frac{1}{2s^2}\left[\sigma^2+\mu^2-2m\mu+m^2\right] - \frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2} \\
&= \frac{1}{2s^2}\left[(m-\mu)^2 + (\sigma^2-s^2) + s^2\log\frac{s^2}{\sigma^2}\right]
\end{aligned}$$

Problem 1.31 - Differential entropy and independence

In this problem, we extend the definition to of KL-divergence to a more general setting as follows:

Definition 1.1. If P and Q are probability measures over a set Ω, if P is absolutely continuous w.r.t. Q, then the KL divergence is defined as

$$KL(P||Q) = \int_\Omega \log\frac{dP}{dQ}dP,$$

where dP/dQ is the Radon-Nikodym derivative, whose existence is guaranteed by the fact that P is absolutely continuous w.r.t Q.

Lemma 1.1. *$KL(P||Q) \geq 0$ for any pair of probability measures P and Q such that $P \ll Q$, the equality if P and Q are equal.*

Proof. This is a directly application of Jensen's inequality. Note that

$$KL(P||Q) = -\int_\Omega \log\frac{dQ}{dP}dP \geq -\log\left(\int_\Omega \frac{dQ}{dP}dP\right) = -\log\int_\Omega dQ = 0.$$

Recall that Jensen's inequality attains the equality if and only if when the function is affine or its argument is constant. In this case, $\log(t)$ is not constant, and thus $KL(P||Q) = 0$ iff $dP/dQ = C$ for some constant $C \in \mathbb{R}$. We claim that $C = 1$, since otherwise we would have

$$\int_\Omega dP = \int_\Omega \frac{dP}{dQ} dQ = C \int_\Omega dQ = C \neq 1,$$

which is a contradiction since P is a probability measure. Then we claim that P and Q are equal. For any set A in the (predefined) sigma algebra, we have

$$P(A) = \int_A dP = \int_A \frac{dP}{dQ} dQ = \int_A 1 dQ = Q(A).$$

Hence, $P = Q$. □

Now we come back to the problem. We instead use X and Y two denote the random variables and $f_X, f_Y, f_{X,Y}$, to denote their (marginal) densities. Suppose X and Y are independent. Then it follows that

$$\begin{aligned}
H(X,Y) &= \int\int f_{X,Y}(x,y) \log f_{X,Y}(x,y) dxdy \\
&= \int\int f_X(x) f_Y(y) (\log f_X(x) + \log f_Y(y)) dxdy \\
&= \int\int f_X(x) f_Y(y) \log f_X(x) dxdy + \int\int f_X(x) f_Y(y) \log f_Y(y) dxdy \\
&= \int f_X(x) \log f_X(x) dx + \int f_Y(y) \log f_Y(y) dy \\
&= H(X) + H(Y).
\end{aligned}$$

Now on the other hand, suppose $H(X,Y) = H(X) + H(Y)$. Then since $H(X,Y) = H(Y|X) + H(X)$ according to Eq. (1.112), it follows that $H(Y|X) = H(Y)$. Note that

$$\begin{aligned}
H(Y|X) - H(Y) &= -\int\int f(x,y) \log f(y|x) dxdy + \int\int f(x,y) \log f(y) dxdy \\
&= \int\int f(x,y) \log \frac{f(y)}{f(y|x)} dxdy \\
&= KL(f_Y(y)||f_{Y|X}(y|x)).
\end{aligned}$$

Then according to Lem. 1.1, $f_Y(y) = f_{Y|X}(y|x)$ almost surely, and as a result $f_{X,Y}(x,y) = f_X(x) f_Y(y)$ which implies that X and Y are independent.

Problem 1.32 - Entropy under linear transformation

Since this problem uses transformation theorem, we first recall this classical result:

Theorem 1.1 ([Bil12, Thm. 17.2]). *Let T be a continuously differentiable map of the open set U onto V. Suppose that T is injective an that $J(x) \neq 0$ for all x. If f is non-negative, then*

$$\int_U f(Tx) |J(x)| dx = \int_{V=TU} f(y) dy.$$

Remark 1.2. We can use this theorem to the get change of variable formula for random variables in $\mathbb{R}^d$ in the following way. Suppose X is a random variable in $\mathbb{R}^d$ with density f_X and $g(\cdot)$ is a C^1 diffeomorphism in $\mathbb{R}^d$, whose inverse is denoted as T and $J_T(x) \neq 0$, then it follows that

$$\mathbb{P}[g(X) \in A] = \mathbb{P}[X \in g^{-1}(A)] = \mathbb{P}[X \in TA] = \int_{TA} f_X(y)dy.$$

Now apply Thm. 1.1, and we get

$$\int_{TA} f_X(y)dy = \int_A f_X(Tx)|J_T(x)|dx = \int_A f_X(g^{-1}(x))|J_{g^{-1}}(x)|dx.$$

Hence, from

$$\begin{aligned}\mathbb{P}(g(X) \in A) &= \int \mathbb{1}_A dF_{g(X)} = \int \mathbb{1}_A \frac{dF_{g(X)}}{dx}dx \qquad (dx \text{ refers to Lebesgue measure})\\ &= \int_A f_X(g^{-1}(x)) \left|J_{g^{-1}}(x)\right| dx\end{aligned}$$

and the fact that Radon-Nikodym derivative is unique it follows that $g(X)$ has density of the form $f_X(g^{-1}(x)) \left|J_{g^{-1}}(x)\right|$.

Now we return to the problem. We instead use $f_Y(y)$ and $f_X(x)$ to denote the density function for X and Y. First, by previous remark, we see that $f_Y(y) = f_X(A^{-1}y) \left|J_{A^{-1}}\right| = f_X(A^{-1}y) \left|\det(A)^{-1}\right|$. So,

$$\begin{aligned}H(Y) &= -\int \ln f_Y(y) dF_Y = \int \ln f_X(A^{-1}y) \left|\det(A)^{-1}\right| dF_Y\\ &= -\int \ln \left[f_X(A^{-1}y) \left|\det(A)^{-1}\right|\right] f_X(A^{-1}y) \left|\det(A)^{-1}\right| dy\\ &= -\int \ln \left[f_X(A^{-1}Ax) \left|\det(A)^{-1}\right|\right] f_X(A^{-1}Ax) \left|\det(A)^{-1}\right| |\det(A)|\, dx \qquad (1)\\ &= -\int \ln \left[f_X(x) \left|\det(A)^{-1}\right|\right] f_X(x)dx\\ &= -\int f_X(x) \ln f_X(x)dx + \int (\ln \det A) f_X(x)dx\\ &= H(X) + \ln(\det A)\end{aligned}$$

as desired. Note that the justification for Eq. (1) is as follows: we abbreviate

$$\varphi(x) = f_X(x) \left|\det(A)^{-1}\right| f_X(x) \left|\det(A)^{-1}\right|,$$

then again by an application of Thm. 1.1

$$(1) = -\int \varphi \circ L d\mu = -\det(L^{-1}) \int \varphi \circ L \circ L^{-1} d\mu = -\det(L^{-1}) \int \varphi d\mu.$$

where μ is the Lebesgue measure. Here since L is represented by A^{-1}, L^{-1} is thus represented by A.

Problem 1.33 - Zero conditional entropy implies singleton concentration

Instead of x, y, we use X, Y to denote random variables. First, we reformulate $H(Y|X)$ as follows:

$$\begin{aligned} H(Y|X) &= -\sum_i \sum_j \mathbb{P}(X = x_i, Y = y_j) \log \mathbb{P}(Y_j = y_j | X = x_i) \\ &= -\sum_i \sum_j \mathbb{P}(Y = y_j | X = x_i) \mathbb{P}(X = x_i) \log \mathbb{P}(Y_j = y_j | X = x_i) \\ &= \sum_i \mathbb{P}(X = x_i) \sum_j f(x_{ij}), \end{aligned}$$

where $f : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}$ is defined as $x \mapsto -x \log x$. We now observe that f is strictly positive for $x \in (0, 1)$ and zero for $x = 1$or 0. The latter is straightforward by direct substitution. To see the former, note

$$f(x) = x \log \frac{1}{x} > x \log 1 = 0 \text{ for } x \in (0, 1).$$

Without loss of generality, we assume $\mathbb{P}(X = x_i) > 0$ since otherwise we get remove these zeros terms without affect the sum. Note that

$$\begin{aligned} H(Y|X) = 0 &\implies \sum_i x \mathbb{P}(X = x_i) \sum_j f(x_{ij}) = 0 \\ &\implies \sum_j f(x_{ij}) = 0 \quad \text{for any given } i. \qquad (\text{since } \mathbb{P}(X = x_i) > 0 \text{ for any } i) \end{aligned}$$

Since $f(x) = 0$ iff $x_{ij} = 0$ or 1, it follows that for any given i, $\mathbb{P}(Y = y_j | X = x_i) = 0$ or 1 for any j. Clearly, there must be only j such that $\mathbb{P}(Y = y_j | X = x_i) = 1$ and $\mathbb{P}(Y = y_j | X = x_i) = 0$ for all other j's since otherwise $\sum_j \mathbb{P}(Y = y_j | X = x_i) \neq 0$, causing a contradiction.

Problem 1.34 - Gaussian distribution maximizes entropy under constraints

To facilitate the notation, we define

$$F(p(x) = -\int_{\mathbb{R}} p(x) \ln p(x) dx + \lambda_1 \left(\int_{\mathbb{R}} p(x) dx - 1 \right) + \lambda_2 \left(\int_{\mathbb{R}} x p(x) dx - \mu \right) + \lambda_3 \left(\int_{\mathbb{R}} (x - \mu)^2 p(x) dx - \sigma^2 \right).$$

First, we rearrange to get

$$F(p(x)) = \int_{\mathbb{R}} \underbrace{-p(x) \ln p(x) dx + \lambda_1 p(x) + \lambda_2 x p(x) + \lambda_3 (x - \mu)^2 p(x)}_{:= G(p(x), x)} dx - (\lambda_1 + \lambda_2 \mu + \lambda_3 \sigma^2).$$

To get the stationary point, we take the functional derivative:

$$\frac{\delta F(p(x))}{\delta p(x)} = \frac{\partial G(p(x), x)}{\partial p(x)} = -\ln p(x) - 1 + \lambda_1 + \lambda_2 x + \lambda_3 (x - \mu)^2.$$

Setting it to zero yields,

$$\ln(p(x)) = \lambda_1 + \lambda_2 x + \lambda_3 (x - \mu)^2 - 1 \implies p(x) = \exp\{\lambda_1 + \lambda_2 x + \lambda_3 (x - \mu)^2 - 1\}.$$

Now we need to eliminate the λ's by substituting back to the constraints

1. $\int p(x)dx = 1$
2. $\int xp(x)dx = \mu$
3. $\int (x-\mu)^2 p(x)dx = \sigma^2$.

This is system of integral equations. To solve it using first principles would require a lot more work (plus I don't know if Gaussian density is the unique solution). But since we are only required to show that Gaussian density is indeed one solution, we are relieved from the burden of proving uniqueness. And we can just directly compare the coefficients. Note that

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

Hence, if we let

$$\exp\left\{\lambda_2 x + \lambda_3 (x-\mu)^2\right\} = \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \implies \lambda_3 = -\frac{1}{2\sigma^2}, \lambda_2 = 0 \text{ is a solution}$$

and

$$\exp\left\{\lambda_1 - 1\right\} = \frac{1}{(2\pi\sigma^2)^{1/2}} \implies \lambda_1 = 1 - \frac{1}{2}\ln 2\pi\sigma^2 \text{ is a solution.}$$

Hence, we have shown that we can find admissible $\lambda_1, \lambda_2, \lambda_3$ such that $p(x)$ satisfies the constraint, and the resulting distribution with this set of λ's is Gaussian. Therefore, Gaussian distribution is a minimizer.

Remark 1.3. One can potentially ask is Gaussian a unique minimizer for this optimization problem? I don't know on the top of my head. This is equivalent to showing that the solution to the integral constraints with $p(x) = \exp\{\lambda_1 + \lambda_2 x + \lambda_3 (x-\mu)^2 - 1\}$, has unique solution. I would guess some deep theorems are needed to prove this result, assuming it is true.

Problem 1.35 - Entropy of Gaussian

We let $\varphi(x|\mu,\sigma^2)$ denote the density of Gaussian distribution. Let X be a Gaussian random variable, then

$$\begin{aligned}
H(X) &= -\int \varphi(x|\mu,\sigma^2) \ln\left[\varphi(x|\mu,\sigma^2)\right] dx \\
&= \frac{1}{2}\int \varphi(x|\mu,\sigma^2)\ln(2\pi\sigma^2)dx + \frac{1}{2\sigma^2}\left[\int \varphi(x|\mu,\sigma^2)x^2 dx - 2\mu\int \varphi(x|\mu,\sigma^2)x dx + \mu^2\int \varphi(x|\mu,\sigma^2)dx\right] \\
&= \frac{1}{2}\ln(2\pi\sigma^2) + \frac{1}{2\sigma^2}\left[\sigma^2 + \mu^2 - 2\mu^2 + \mu^2\right] \\
&= \frac{1}{2}\ln(2\pi\sigma^2) + \frac{1}{2} \\
&= \frac{1}{2}(1 + \ln(2\pi\sigma^2))
\end{aligned}$$

as desired.

Problem 1.36 - Second order characterization of convexity

We prove a slightly more generalized version. First, we recall the definition of the convexity.

Definition 1.2. A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if its domain $\mathcal{D}_f$ is a convex set and for any $x, y \in \mathcal{D}_f$ and $\lambda \in [0, 1]$,

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y). \tag{1}$$

The result of this problem is an direct consequence of the following proposition.

Proposition 1.1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function. Then the following statements are equivalent.*

1. *f is convex.*
2. *$f(y) \geq f(x) + \nabla f(x)^T (y - x)$ assuming f is differentiable.*
3. *The Hessian matrix $H_f(x)$ is positive semidefinite, assuming f is twice differentiable and $\mathcal{D}_f$ is open.*

Proof. (1) $\Rightarrow$ (2). Suppose f is convex. Then by definition for any $y, x \in \mathcal{D}_f$,

$$\begin{aligned} f(\lambda y + (1-\lambda)x) &= f(x + \lambda(y - x)) \\ &\leq \lambda f(y) + (1-\lambda) f(x) \\ &= f(x) + \lambda(f(y) - f(x)). \end{aligned}$$

Rearranging the expression yields

$$\frac{f(x + \lambda(y-x)) - f(x)}{\lambda} \leq f(y) - f(x).$$

Now we take the limit:

$$\lim_{\lambda \to 0} \frac{f(x + \lambda(y-x)) - f(x)}{\lambda} = \nabla f(x)^T (y - x).$$

This equality can be derived from the following argument: note the Taylor expansion of f at $x + h$ is

$$f(x + th) = f(x) + t \langle \nabla f(x), h \rangle + o(\|th\|).$$

Then by rearranging we get

$$\frac{f(x+th) - f(x)}{t} = \langle \nabla f(x), h \rangle + \frac{o(\|th\|)}{t \|h\|} \|h\| \xrightarrow{t \to 0} \langle \nabla f(x), h \rangle = \nabla f(x)^T h.$$

Hence, we have

$$\nabla f(x)^T (y - x) \leq f(y) - f(x) \iff f(y) \geq f(x) + f(x)^T (y - x)$$

as desired.

(2) $\Rightarrow$ (1). Now assume $f(y) \geq f(x) + \nabla f(x)^T (y - x)$ for any $x, y \in \mathcal{D}_f$. Fix $x, y \in \mathcal{D}_f$. Then note that since $\mathcal{D}_f$ is convex, $\lambda x + (1-\lambda)y \in \mathcal{D}_f$. We first apply it to the pair $(\lambda x + (1-\lambda)y, y)$:

$$f(y) \geq f(\lambda x + (1-\lambda)y) + \nabla f(\lambda x + (1-\lambda)y)^T (y - \lambda x - (1-\lambda)y)$$

$$= f(\lambda x + (1-\lambda)y) + \nabla f(\lambda x + (1-\lambda)y)^T \lambda(y-x). \tag{2}$$

Similarly, we apply it to the pair $(\lambda x + (1-\lambda)y, x)$:

$$f(x) \geq f(\lambda x + (1-\lambda)y) + \nabla f(\lambda x + (1-\lambda)y)(1-\lambda)(x-y). \tag{3}$$

Now, we note that for $\lambda \in (0,1)$,

$$\begin{aligned}(1-\lambda) \times \text{Eq.}(2) + \lambda \times \text{Eq.}(3) =& (1-\lambda f(y)) + \lambda f(x) \\ \geq& (1-\lambda+\lambda) f(\lambda x + (1-\lambda)y) \\ =& f(\lambda x + (1-\lambda)y),\end{aligned}$$

which is the definition of convexity in defined in Eq. (1).

(2) $\Rightarrow$ (3). Pick arbitrary $x, h \in \mathcal{D}_f$. Since $\mathcal{D}_f$ is open, we can find a sufficiently small λ such that $x + \lambda h \in \mathcal{D}_f$. We first write out the second order Taylor expansion of f at $x + \lambda h$,

$$f(x+\lambda h) = f(x) + \lambda \langle \nabla f(x), h \rangle + \frac{\lambda^2}{2} H_f(x)(h,h) + o(\|\lambda h\|^2). \tag{4}$$

Since f is convex, it follows that $f(x+\lambda h) \geq f(x) + \lambda \langle \nabla f(x), h \rangle$. Substituting back to Eq.(4) yields

$$\begin{aligned}\lambda^2 H_f(x)(h,h) + o(\|\lambda h\|^2) \geq 0 &\implies H_f(x)(h,h) + \frac{o(\|\lambda h\|^2)}{\|\lambda h\|^2} \|h\|^2 \geq 0 \qquad (\text{any } \lambda \in (0,1)) \\ &\implies \lim_{\lambda \to 0^+} \left[H_f(x)(h,h) + \frac{o(\|\lambda h\|^2)}{\|\lambda h\|^2} \|h\|^2 \right] \geq 0 \\ &\implies H_f(x)(h,h) \geq 0.\end{aligned}$$

Since h is arbitrary, it follows that $H_f(x)$ is positive semidefinite.

(3) $\Rightarrow$ (2). Suppose H_f is positive semidefinite. Then for any $x, y \in \mathcal{D}_f$, since $\mathcal{D}_f$ is convex, $\lambda x + (1-\lambda)y \in \mathcal{D}_f$ for any $\lambda \in (0,1)$. Then by a second order Taylor formula, we can write

$$f(y) = f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2} H_f(z)(y-x, y-x)$$

for some z in the segment $[x,y] := \{\text{all points of form } \lambda x + (1-\lambda)y \text{ for } \lambda \in (0,1)\}$. Since H_f is positive semidefinite, it follows that $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$. □

Bibliography

B

[Bil12] Patrick Billingsley. *Probability and measure*. Wiley, Hoboken, N.J, 2012. 23

C

[Con00] Keith Conrad. Differentiating under the integral sign. 2000. 6

K

[Kuc09] Marek Kuczma. *An introduction to the theory of functional equations and inequalities : Cauchy's equation and Jensen's inequality*. Birkhauser, Basel Boston, 2009. 21

L

[Lan97] Serge Lang. *Undergraduate Analysis*. Springer-Verlag New York, 2 edition, 1997. 6

S

[Ste05] Elias Stein. *Real analysis : measure theory, integration, and Hilbert spaces*. Princeton University Press, Princeton, N.J. Oxford, 2005. 13, 14