MML — Exercises — Example Answers

October 27, 2021

1 Least-squares solution

Exercise: The loss function for linear regression is

$$L(\boldsymbol{\theta}) = \sum_{n=1}^{N} \left(y_n - \boldsymbol{\phi}(x_n)^T \boldsymbol{\theta}\right)^2 = ||\mathbf{y} - \boldsymbol{\Phi}(X)\boldsymbol{\theta}||^2$$

1. Find the gradient of the loss function with respect to $\boldsymbol{\theta}$.
2. Find $\boldsymbol{\theta}$ for which $L(\boldsymbol{\theta})$ is minimized.
3. Demonstrate $\boldsymbol{\theta}$ is in fact a minimum for $L(\boldsymbol{\theta})$.

Notice $\boldsymbol{\Phi}(X)$ contains all the samples evaluated using the different basis functions used. $\boldsymbol{\Phi}(X)$ is usually denoted as the *design matrix* and for N data points and M basis functions, it represents the following.

$$\boldsymbol{\Phi}(X) = \begin{pmatrix} \boldsymbol{\phi}(x_1)^T \\ \vdots \\ \boldsymbol{\phi}(x_N)^T \end{pmatrix} = \begin{pmatrix} \phi_1(x_1) & \dots & \phi_M(x_1) \\ \vdots & \ddots & \vdots \\ \phi_1(x_M) & \dots & \phi_M(x_M) \end{pmatrix}$$

Solution: Let us compute the gradient with respect to $\boldsymbol{\theta}$. For illustrative purposes, we will use the vector form of the loss function.

$$\begin{aligned} \frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \frac{\partial}{\partial \boldsymbol{\theta}} \left(||\mathbf{y} - \boldsymbol{\Phi}(X)\boldsymbol{\theta}||^2 \right) = \frac{\partial}{\partial \boldsymbol{\theta}} \left(\left(\mathbf{y} - \boldsymbol{\Phi}(X)\boldsymbol{\theta}\right)^T \left(\mathbf{y} - \boldsymbol{\Phi}(X)\boldsymbol{\theta}\right) \right) \\ &= \frac{\partial}{\partial \boldsymbol{\theta}} \left(\mathbf{y}^T\mathbf{y} - 2\boldsymbol{\theta}^T\boldsymbol{\Phi}(X)^T\mathbf{y} + \boldsymbol{\theta}^T\boldsymbol{\Phi}(X)^T\boldsymbol{\Phi}(X)\boldsymbol{\theta} \right) \\ &= -2\boldsymbol{\Phi}(X)^T\mathbf{y} + 2\boldsymbol{\Phi}(X)^T\boldsymbol{\Phi}(X)\boldsymbol{\theta} \end{aligned}$$

Now, we calculate $\boldsymbol{\theta}^*$ for which the gradient is zero, which should minimize the loss function.

$$\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} = 0$$

$$0 = -2\boldsymbol{\Phi}(X)^T\mathbf{y} + 2\boldsymbol{\Phi}(X)^T\boldsymbol{\Phi}(X)\boldsymbol{\theta}^*$$

$$\boldsymbol{\Phi}(X)^T\boldsymbol{\Phi}(X)\boldsymbol{\theta}^* = \boldsymbol{\Phi}(X)^T\mathbf{y}$$

$$\boldsymbol{\theta}^* = \left(\boldsymbol{\Phi}(X)^T\boldsymbol{\Phi}(X)\right)^{-1}\boldsymbol{\Phi}(X)^T\mathbf{y}$$

Notice that the following quantity

$$\boldsymbol{\Phi}(X)^\dagger = \left(\boldsymbol{\Phi}(X)^T\boldsymbol{\Phi}(X)\right)^{-1}\boldsymbol{\Phi}(X)^T$$

is commonly known as the *pseudo-inverse* of the matrix $\boldsymbol{\Phi}(X)$, which can be regarded as a generalization for the inverse of a non-square matrix.

Therefore, the $\boldsymbol{\theta}^*$ which minimizes the loss function should be

$$\boldsymbol{\theta}^* = \boldsymbol{\Phi}(X)^\dagger \mathbf{y}$$

We still need to confirm that $\boldsymbol{\theta}^*$ is actually a minimum of $L(\boldsymbol{\theta})$. We can do so by computing the Hessian. We know that if the resulting Hessian is positive definite, $\boldsymbol{\theta}^*$ is a minimum for the least squares problem.

$$\frac{\partial^2 L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} = \frac{\partial}{\partial \boldsymbol{\theta}} \Big(-2\boldsymbol{\Phi}(X)^T \mathbf{y} + 2\boldsymbol{\Phi}(X)^T \boldsymbol{\Phi}(X)\boldsymbol{\theta} \Big) = 2\boldsymbol{\Phi}(X)^T \boldsymbol{\Phi}(X)$$

We need to prove that $\boldsymbol{\Phi}(X)^T \boldsymbol{\Phi}(X)$ is positive definite. Recall $\boldsymbol{\Phi}(X) \in \mathbb{R}^{N \times M}$.

$$\boldsymbol{\Phi}(X)^T \boldsymbol{\Phi}(X) \text{ is positive definite } \iff \mathbf{z}^T \boldsymbol{\Phi}(X)^T \boldsymbol{\Phi}(X)\mathbf{z} > 0, \quad \forall \mathbf{z} \in \mathbb{R}^M \setminus \{\mathbf{0}\}$$

$$\mathbf{z}^T \boldsymbol{\Phi}(X)^T \boldsymbol{\Phi}(X)\mathbf{z} = \Big(\boldsymbol{\Phi}(X)\mathbf{z}\Big)^T \boldsymbol{\Phi}(X)\mathbf{z} = ||\boldsymbol{\Phi}(X)\mathbf{z}||^2 \geq 0$$

$\boldsymbol{\Phi}(X)^T \boldsymbol{\Phi}(X)$ is positive definite if we can show that $\boldsymbol{\Phi}(X)\mathbf{z} \neq 0, \quad \forall \mathbf{z} \in \mathbb{R}^M \setminus \{\mathbf{0}\}$. For that, we make two assumptions which usually hold in the least squares problem. The first one is that $N \geq M$, which means that we have more (or equal) data points than basis functions. The second one is that the design matrix $\boldsymbol{\Phi}(X)$ is full rank. We can re-arrange the matrix product $\boldsymbol{\Phi}(X)\mathbf{z}$ as a linear combination of each basis function $\boldsymbol{\phi}_i(\mathbf{x})$, where $\boldsymbol{\phi}_i(\mathbf{x})$ is the i-th basis function evaluated on each data point.

$$\boldsymbol{\Phi}(X)\mathbf{z} = \begin{pmatrix} \boldsymbol{\phi}_1(\mathbf{x}) & \boldsymbol{\phi}_2(\mathbf{x}) & \dots & \boldsymbol{\phi}_M(\mathbf{x}) \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_M \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

$$\sum_{i=1}^{M} \boldsymbol{\phi}_i(\mathbf{x}) z_i = \mathbf{0}$$

Since $\boldsymbol{\Phi}(X)$ is full rank and $M \leq N$, all the terms $\boldsymbol{\phi}_i(\mathbf{x})$ are linearly independent and thus, there exists no $\mathbf{z}$ for which we can obtain $\mathbf{0}$ in the previous expression. Consequently, we have that $||\boldsymbol{\Phi}(X)\mathbf{z}||^2 > 0$ and thus, $\boldsymbol{\Phi}(X)^T \boldsymbol{\Phi}(X)$ is positive definite, which means that $\boldsymbol{\theta}^*$ is in fact a minimum for the least squares problem.

2 Exercises chapter 5

In some of the solutions, earlier solutions are re-used. In an exam, you need to ensure to state what identity is used, or you may need to prove sub-results if requested. In an exam, you may refer to earlier derivation in your exam transcript, **but you must do so clearly and unambiguously**, e.g. with an equation number.

5.1.

$$f(x) = \log(x^4)\sin(x^3)$$

$$f'(x) = \frac{\partial \log(x^4)}{\partial x}\sin(x^3) + \log(x^4)\frac{\partial \sin(x^3)}{\partial x}$$

$$f'(x) = \frac{1}{x^4} 4x^3 \sin(x^3) + \log(x^4)\cos(x^3) 3x^2$$

$$f'(x) = \frac{4}{x}\sin(x^3) + 3x^2 \log(x^4)\cos(x^3)$$

5.2.

$$f(x) = \frac{1}{1+\exp(-x)}$$

$$f'(x) = \frac{-1}{(1+\exp(-x))^2}\exp(-x)(-1)$$

$$f'(x) = \frac{\exp(-x)}{(1+\exp(-x))^2}$$

5.3.

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

$$f'(x) = \exp(-\frac{1}{2\sigma^2}(x-\mu)^2)\big(\frac{-1}{2\sigma^2}2(x-\mu)\big)$$

$$f'(x) = \frac{(\mu-x)}{\sigma^2}\exp(-\frac{1}{2\sigma^2}(x-\mu)^2)$$

5.5.

- $f_1(\mathbf{x}) = \sin(x_1)\cos(x_2), \quad \mathbf{x} \in \mathbb{R}^2$

  $$\frac{\partial f_1}{\partial \mathbf{x}} \in \mathbb{R}^{1\times 2}$$

  $$\begin{aligned}\frac{\partial f_1}{\partial \mathbf{x}} &= \left[\frac{\partial f_1}{\partial x_1}, \frac{\partial f_1}{\partial x_2}\right] \\ &= \left[\cos(x_1)\cos(x_2), -\sin(x_1)\sin(x_2)\right]\end{aligned}$$

- $f_2(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T\mathbf{y}, \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$

  $$\frac{\partial f_2}{\partial \mathbf{x}} \in \mathbb{R}^{1\times n}$$

  We can solve this directly using basic rules of vector calculus

  $$\frac{\partial f_2}{\partial \mathbf{x}} = \frac{\partial(\mathbf{x}^T\mathbf{y})}{\partial \mathbf{x}} = \mathbf{y}^T$$

  We can confirm this result holds by confirming the notation used in the lectures. First, let us calculate the value $f_2(\mathbf{x}, \mathbf{y})$

  $$f_2(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T\mathbf{y} = \sum_{i=1}^{n} x_i y_i$$

  $$\frac{\partial f_2}{\partial x_j} = \frac{\partial}{\partial x_j}\sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n}\frac{\partial x_j}{\partial x_i}y_i = \sum_{i=1}^{n}\delta_{ij}y_i = y_j$$

  $$\frac{\partial f_2}{\partial \mathbf{x}} = \left[\frac{\partial f_2}{\partial x_1}, \ldots, \frac{\partial f_2}{\partial x_n}\right] = [y_1, \ldots, y_n] = \mathbf{y}^T$$

- $\mathbf{f}_3(x) = \mathbf{x}\mathbf{x}^T, \quad \mathbf{x} \in \mathbb{R}^n$

  $$\frac{\partial \mathbf{f}_3}{\partial \mathbf{x}} \in \mathbb{R}^{(n\times n)\times n}$$

  $$\frac{\partial \mathbf{f}_3}{\partial \mathbf{x}} = \mathbf{C} \quad \text{where } C \text{ is a 3D tensor.}$$

  $$C_{ijk} = \frac{\partial f_3(\mathbf{x})_{ij}}{\partial x_k}$$

  $$\mathbf{x}\mathbf{x}^T = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \begin{pmatrix} x_1 & \dots & x_n \end{pmatrix} = \begin{pmatrix} x_1^2 & x_1x_2 & \dots & x_1x_n \\ x_2x_1 & x_2^2 & & \vdots \\ \vdots & & \ddots & \vdots \\ x_nx_1 & \dots & \dots & x_n^2 \end{pmatrix}$$

  $$C_{ijk} = \frac{\partial(x_ix_j)}{\partial x_k} = \frac{\partial x_i}{\partial x_k}x_j + \frac{\partial x_j}{\partial x_k}x_i = \delta_{ik}x_j + \delta_{jk}x_i = \begin{cases} 0 & \text{if } k \neq i \text{ and } k \neq j \\ x_i & \text{if } k = j \text{ and } i \neq j \\ x_j & \text{if } k = i \text{ and } i \neq j \\ 2x_i & \text{if } k = i = j \end{cases}$$

5.6.

- $f(\mathbf{t}) = \sin\big(\log(\mathbf{t}^T\mathbf{t})\big) \quad \mathbf{t} \in \mathbb{R}^D$

  We directly apply the chain rule

  $$\frac{\partial f}{\partial \mathbf{t}} = \frac{\partial \sin\big(\log(\mathbf{t}^T\mathbf{t})\big)}{\partial \log(\mathbf{t}^T\mathbf{t})} \cdot \frac{\partial \log(\mathbf{t}^T\mathbf{t})}{\partial(\mathbf{t}^T\mathbf{t})} \cdot \frac{\partial(\mathbf{t}^T\mathbf{t})}{\partial \mathbf{t}}$$

  All of the terms are one dimensional except for $\frac{\partial(\mathbf{t}^T\mathbf{t})}{\partial \mathbf{t}} \in \mathbb{R}^{1\times D}$. Let us calculate the value using the notation for vector calculus in the lectures. As in 5.5, we first calculate the value of $\mathbf{t}^T\mathbf{t}$ and its derivative w.r.t. t_i.

  $$\mathbf{t}^T\mathbf{t} = \sum_{i=1}^{D} t_i^2, \quad \frac{\partial(\mathbf{t}^T\mathbf{t})}{\partial t_i} = 2t_i$$

  $$\frac{\partial(\mathbf{t}^T\mathbf{t})}{\partial \mathbf{t}} = \left[\frac{\partial(\mathbf{t}^T\mathbf{t})}{\partial t_1}, \dots, \frac{\partial(\mathbf{t}^T\mathbf{t})}{\partial t_D}\right] = [2t_1 \dots, 2t_D] = 2\mathbf{t}^T$$

  We can now use this result to proceed with the derivative of $f(\mathbf{t})$.

  $$\frac{\partial f}{\partial \mathbf{t}} = \cos\big(\log(\mathbf{t}^T\mathbf{t})\big) \cdot \frac{1}{\mathbf{t}^T\mathbf{t}} \cdot 2\mathbf{t}^T$$

  $$\frac{\partial f}{\partial \mathbf{t}} = 2\mathbf{t}^T \frac{\cos\big(\log(\mathbf{t}^T\mathbf{t})\big)}{\mathbf{t}^T\mathbf{t}}$$

- $g(\mathbf{X}) = tr(\mathbf{A}\mathbf{X}\mathbf{B}), \quad \mathbf{A} \in \mathbb{R}^{D\times E}, \mathbf{X} \in \mathbb{R}^{E\times F}, \mathbf{B} \in \mathbb{R}^{F\times D}$

  Eq. 101 from Matrix cookbook gives us the direct result.

  $$\frac{\partial g}{\partial \mathbf{X}} = \mathbf{A}^T\mathbf{B}^T$$

  Alternative proof: Use index notation introduced in the course.

  $$g(\mathbf{X}) = tr(\mathbf{A}\mathbf{X}\mathbf{B}) = \sum_{i=1}^{D} (\mathbf{A}\mathbf{X}\mathbf{B})_{ii}$$

In order to fully compute $g(\mathbf{X})$, we need to calculate $(\mathbf{AXB})_{ii}$

$$(\mathbf{AXB})_{ii} = \sum_{k=1}^{F} (\mathbf{AX})_{ik} b_{ki} = \sum_{k=1}^{F} \left(\sum_{l=1}^{E} a_{il} x_{lk} \right) b_{ki}$$

Thus

$$g(\mathbf{X}) = \sum_{i=1}^{D} \sum_{k=1}^{F} \sum_{l=1}^{E} a_{il} x_{lk} b_{ki}$$

Now we can just calculate the derivative using index notation

$$\frac{\partial g}{\partial x_{nm}} = \frac{\partial}{\partial x_{nm}} \sum_{i=1}^{D} \sum_{k=1}^{F} \sum_{l=1}^{E} a_{il} x_{lk} b_{ki} = \sum_{i=1}^{D} \sum_{k=1}^{F} \sum_{l=1}^{E} a_{il} \frac{\partial x_{lk}}{\partial x_{nm}} b_{ki} = \sum_{i=1}^{D} \sum_{k=1}^{F} \sum_{l=1}^{E} a_{il} \delta_{ln} \delta_{km} b_{ki}$$

Notice that in the last expression, all the terms in the summation cancel except when $k = m$ and $l = n$. Therefore

$$\frac{\partial g}{\partial x_{nm}} = \sum_{i=1}^{D} a_{in} b_{mi} = \sum_{i=1}^{D} b_{mi} a_{in} = (\mathbf{BA})_{mn}$$

Using this last result, we can calculate the derivative w.r.t. $\mathbf{X}$.

$$\frac{\partial g}{\partial \mathbf{X}} = (\mathbf{BA})^T = \mathbf{A}^T \mathbf{B}^T$$

Alternative proof 2: Use properties 4.19 and 5.100 from the course book. From 4.19

$$g(\mathbf{X}) = tr(\mathbf{AXB}) = tr(\mathbf{XBA}) = tr(\mathbf{XC}), \quad \mathbf{C} = \mathbf{BA}$$

and from 5.100

$$\frac{\partial g}{\partial \mathbf{X}} = \frac{\partial tr(\mathbf{XC})}{\partial \mathbf{X}} = tr\left(\frac{\partial (\mathbf{XC})}{\partial \mathbf{X}} \right), \quad \text{where } \frac{\partial (\mathbf{XC})}{\partial \mathbf{X}} \in \mathbb{R}^{(E \times E) \times (E \times F)}$$

We need to calculate $\frac{\partial (\mathbf{XC})_{ij}}{\partial X_{kl}}$, and we find convenient to write the pairs i, j of the product $\mathbb{I}\mathbf{XC}$, where $\mathbb{I} \in \mathbb{R}^{E \times E}$ is the identity matrix.

$$(\mathbb{I}\mathbf{XC})_{ij} = \sum_{e=1}^{E} \sum_{f=1}^{F} \delta_{ie} x_{ef} c_{fj}$$

$$\frac{\partial (\mathbf{XC})_{ij}}{\partial X_{kl}} = \frac{\partial (\mathbb{I}\mathbf{XC})_{ij}}{\partial X_{kl}} = \delta_{ik} c_{lj}$$

in the prevous expression, all the terms in the sum vanish except the ones that contain x_{kl} in it.

Now, we take into account the definition of the trace for any 4D tensor $\mathbf{A} \in \mathbb{R}^{(N \times N) \times (P \times Q)}$ given in the course book:

$$tr(\mathbf{A})_{ij} = \sum_{k=1}^{N} a_{kkij}, \quad \text{where } tr(\mathbf{A}) \in \mathbb{R}^{P \times Q}$$

We use this definition to calculate our result.

$$tr\left(\frac{\partial (\mathbf{XC})}{\partial \mathbf{X}} \right)_{ij} = \sum_{k=1}^{E} \frac{\partial (\mathbf{XC})_{kk}}{\partial X_{ij}} = \sum_{k=1}^{E} \delta_{ki} c_{jk} = c_{ji}$$

all the terms will be 0 except when $k = i$.

$$tr\left(\frac{\partial (\mathbf{XC})}{\partial \mathbf{X}} \right) = \mathbf{C}^T = (\mathbf{BA})^T = \mathbf{A}^T \mathbf{B}^T$$

5.7.

a. $f(z) = \log(1+z), \quad z = \mathbf{x}^T\mathbf{x}, \quad \mathbf{x} \in \mathbb{R}^D$

$$\frac{\partial f}{\partial \mathbf{x}} = \frac{\partial f}{\partial z}\frac{\partial z}{\partial \mathbf{x}} = \frac{\partial \log(1+z)}{\partial z}\frac{\partial(\mathbf{x}^T\mathbf{x})}{\partial \mathbf{x}} = \frac{2\mathbf{x}^T}{1+z} = \frac{2\mathbf{x}^T}{1+\mathbf{x}^T\mathbf{x}}$$

Dimensions are

$$\frac{\partial f}{\partial \mathbf{x}} \in \mathbb{R}^D, \quad \frac{\partial f}{\partial z} \in \mathbb{R}, \quad \frac{\partial z}{\partial \mathbf{x}} \in \mathbb{R}^D$$

b. $f(\mathbf{z}) = \sin(\mathbf{z}), \quad \mathbf{z} = \mathbf{A}\mathbf{x} + \mathbf{b}, \quad \mathbf{A} \in \mathbb{R}^{E\times D}, \mathbf{x} \in \mathbb{R}^D, \mathbf{b} \in \mathbb{R}^E$

$$\frac{\partial f}{\partial \mathbf{x}} = \frac{\partial f}{\partial \mathbf{z}}\frac{\partial \mathbf{z}}{\partial \mathbf{x}} = \frac{\partial \sin(\mathbf{z})}{\partial \mathbf{z}}\frac{\partial(\mathbf{A}\mathbf{x}+\mathbf{b})}{\partial \mathbf{x}}$$

Notice that $\frac{\partial f}{\partial \mathbf{z}} \in \mathbb{R}^{E\times E}$. We already know that $\sin(\cdot)$ is applied to each element independently, thus

$$\frac{\partial f_i}{\partial z_j} = \begin{cases} 0 & \text{if } i \neq j \\ \cos(z_i) & \text{if } i = j \end{cases}$$

We also have $\frac{\partial \mathbf{z}}{\partial \mathbf{x}} \in \mathbb{R}^{E\times D}$. Although this has already shown in the lectures, let us review the result $\frac{\partial \mathbf{z}}{\partial \mathbf{x}}$ using the notation of the course.

$$z_i = \sum_{j=1} A_{ij}x_j + b_i$$

We can now easily compute $\frac{\partial z_i}{\partial x_j}$

$$\frac{\partial z_i}{\partial x_j} = A_{ij}, \quad \frac{\partial \mathbf{z}}{\partial \mathbf{x}} = \mathbf{A}$$

Let us use all the previous results to compute the derivative of $f(\mathbf{z})$ w.r.t. $\mathbf{x}$.

$$\frac{\partial f}{\partial \mathbf{x}} = diag(\cos(\mathbf{z}))\mathbf{A}, \quad \text{where } diag(\mathbf{a}) = \begin{pmatrix} a_1 & 0 & \dots & 0 \\ 0 & a_2 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \dots & \dots & a_N \end{pmatrix}, \quad \mathbf{a} \in \mathbb{R}^N$$

$$\frac{\partial f}{\partial \mathbf{x}} = diag(\cos(\mathbf{A}\mathbf{x}+\mathbf{b}))\mathbf{A}$$

Dimensions are

$$\frac{\partial f}{\partial \mathbf{x}} \in \mathbb{R}^{E\times D}, \quad \frac{\partial f}{\partial \mathbf{z}} \in \mathbb{R}^{E\times E}, \quad \frac{\partial \mathbf{z}}{\partial \mathbf{x}} \in \mathbb{R}^{E\times D}$$

5.8.

a. $f(z) = \exp(-\frac{1}{2}z), \quad z = g(\mathbf{y}) = \mathbf{y}^T\mathbf{S}^{-1}\mathbf{y}, \quad \mathbf{y} = h(\mathbf{x}) = \mathbf{x} - \boldsymbol{\mu}, \quad \mathbf{x}, \boldsymbol{\mu} \in \mathbb{R}^D, \mathbf{S} \in \mathbb{R}^{D\times D}$

$$\begin{aligned}\frac{\partial f}{\partial \mathbf{x}} &= \frac{\partial f}{\partial z}\frac{\partial z}{\partial \mathbf{y}}\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \frac{\partial \exp(-\frac{1}{2}z)}{\partial z}\frac{\partial(\mathbf{y}^T\mathbf{S}^{-1}\mathbf{y})}{\partial \mathbf{y}}\frac{\partial(\mathbf{x}-\boldsymbol{\mu})}{\partial \mathbf{x}} = \exp\left(-\frac{1}{2}z\right)\left(-\frac{1}{2}\right)\mathbf{y}^T(\mathbf{S}^T + \mathbf{S}^{-T})\mathbb{I} \\ &= -\frac{1}{2}\exp\left(-\frac{1}{2}\Big((\mathbf{x}-\boldsymbol{\mu})^T\mathbf{S}^{-1}(\mathbf{x}-\boldsymbol{\mu})\Big)\right)(\mathbf{x}-\boldsymbol{\mu})^T(\mathbf{S}^T + \mathbf{S}^{-T})\end{aligned}$$

where $\mathbf{S}^{-T} = \left(\mathbf{S}^{-1}\right)^T$, and we use (5.107) to calculate $\frac{\partial(\mathbf{y}^T\mathbf{S}^{-1}\mathbf{y})}{\partial \mathbf{y}}$.
Dimensions are

$$\frac{\partial f}{\partial \mathbf{x}} \in \mathbb{R}^D, \quad \frac{\partial f}{\partial z} \in \mathbb{R}, \quad \frac{\partial z}{\partial \mathbf{y}} \in \mathbb{R}^D, \quad \frac{\partial \mathbf{y}}{\partial \mathbf{x}} \in \mathbb{R}^{D\times D}$$

b. $f(\mathbf{x}) = tr(\mathbf{x}\mathbf{x}^T + \sigma^2\mathbb{I}), \quad \mathbf{x} \in \mathbb{R}^D$
Let us expand $f(x)$.

$$\begin{aligned} f(x) &= \sum_{i=1}^{D}\left((\mathbf{x}\mathbf{x}^T)_{ii} + \sigma^2\right) \\ &= \sum_{i=1}^{D}(\mathbf{x}\mathbf{x}^T)_{ii} + D\sigma^2 = \sum_{i=1}^{D} x_i^2 + D\sigma^2 \end{aligned}$$

From previous exercices, we know that $(\mathbf{x}\mathbf{x}^T)_{ij} = x_i x_j$. Therefore

$$\frac{\partial f}{\partial \mathbf{x}} = \frac{\partial\left(\sum_{i=1}^{D} x_i^2 + D\sigma^2\right)}{\partial \mathbf{x}} = 2\mathbf{x}^T$$

c. $f(\mathbf{z}) = \tanh(\mathbf{z}) \in \mathbb{R}^M, \quad \mathbf{z} = \mathbf{A}\mathbf{x} + \mathbf{b}, \quad \mathbf{x} \in \mathbb{R}^N, \mathbf{A} \in \mathbb{R}^{M\times N}, \mathbf{b} \in \mathbb{R}^M$

$$\begin{aligned} \frac{\partial f}{\partial \mathbf{x}} &= \frac{\partial f}{\partial \mathbf{z}}\frac{\partial \mathbf{z}}{\partial \mathbf{x}} = \frac{\partial \tanh(\mathbf{z})}{\partial \mathbf{z}}\frac{\partial(\mathbf{A}\mathbf{x}+\mathbf{b})}{\partial \mathbf{x}} = diag\left(1 - \tanh^2(\mathbf{z})\right)\mathbf{A} \\ &= diag\left(1 - \tanh^2(\mathbf{A}\mathbf{x}+\mathbf{b})\right)\mathbf{A} \end{aligned}$$

where we used $\frac{d\tanh(a)}{da} = 1 - \tanh^2(a)$.
Dimensions are

$$\frac{\partial f}{\partial \mathbf{x}} \in \mathbb{R}^{M\times N}, \quad \frac{\partial f}{\partial \mathbf{z}} \in \mathbb{R}^{M\times M}, \quad \frac{\partial z}{\partial \mathbf{x}} \in \mathbb{R}^{M\times N}$$

5.9.

$$g(\mathbf{z}, \boldsymbol{\nu}) := \log p(\mathbf{x}, \mathbf{z}) - \log q(\mathbf{z}, \boldsymbol{\nu}), \quad \mathbf{z} = t(\boldsymbol{\epsilon}, \boldsymbol{\nu})$$

We apply the chain rule straightforwardly.

$$\begin{aligned} \frac{d}{d\boldsymbol{\nu}} g(\mathbf{z}, \boldsymbol{\nu}) &= \frac{\partial g(\mathbf{z}, \boldsymbol{\nu})}{\partial \mathbf{z}}\frac{\partial \mathbf{z}}{\partial \boldsymbol{\nu}} + \frac{\partial g(\mathbf{z}, \boldsymbol{\nu})}{\partial \boldsymbol{\nu}} \\ &= \frac{\partial\left(\log p(\mathbf{x}, \mathbf{z}) - \log q(\mathbf{z}, \boldsymbol{\nu})\right)}{\partial \mathbf{z}}\frac{\partial t(\boldsymbol{\epsilon}, \boldsymbol{\nu})}{\partial \boldsymbol{\nu}} + \frac{\partial\left(\log p(\mathbf{x}, \mathbf{z}) - \log q(\mathbf{z}, \boldsymbol{\nu})\right)}{\partial \boldsymbol{\nu}} \\ &= \left(\frac{1}{p(\mathbf{x}, \mathbf{z})}\frac{\partial p(\mathbf{x}, \mathbf{z})}{\partial \mathbf{z}} - \frac{1}{q(\mathbf{z}, \boldsymbol{\nu})}\frac{\partial q(\mathbf{z}, \boldsymbol{\nu})}{\partial \mathbf{z}}\right)\frac{\partial t(\boldsymbol{\epsilon}, \boldsymbol{\nu})}{\partial \boldsymbol{\nu}} - \frac{1}{q(\mathbf{z}, \boldsymbol{\nu})}\frac{\partial q(\mathbf{z}, \boldsymbol{\nu})}{\partial \boldsymbol{\nu}} \\ &= \frac{1}{p(\mathbf{x}, \mathbf{z})}\frac{\partial p(\mathbf{x}, \mathbf{z})}{\partial \mathbf{z}}\frac{\partial t(\boldsymbol{\epsilon}, \boldsymbol{\nu})}{\partial \boldsymbol{\nu}} - \frac{1}{q(\mathbf{z}, \boldsymbol{\nu})}\left(\frac{\partial q(\mathbf{z}, \boldsymbol{\nu})}{\partial \mathbf{z}}\frac{\partial t(\boldsymbol{\epsilon}, \boldsymbol{\nu})}{\partial \boldsymbol{\nu}} + \frac{\partial q(\mathbf{z}, \boldsymbol{\nu})}{\partial \boldsymbol{\nu}}\right) \end{aligned}$$

3 Sets and probabilities

Let us denote the three axioms of probability theory:

1. The probability of an event E is a non-negative number.

$$E \in \mathbb{R}, \quad P(E) \geq 0$$

2. The probability that at least one of the events will occur is 1.

$$P(\Omega) = 1$$

3. Any countable sequence of disjoint sets $E_1, E_2, \ldots$ satisfies

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i) \quad \text{when } E_i \cap E_j = \varnothing, \quad \forall i, j, \quad i \neq j$$

Demonstrate the following properties.

a. $P(\neg A) = 1 - P(A)$

b. $P(\varnothing) = 0$, where $\varnothing$ is the empty set

c. $0 \leq P(A) \leq 1$

d. $A \subseteq B \implies P(A) \leq P(B)$

Hint: Consider the following definition. $B \backslash A = \{x \in B : x \notin A\}$

e. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

f. (*) if $\{A_i\}_{i=1}^{\infty} \subseteq \Omega$ and $A_i \leq A_{i+1} \forall i$ then:

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \lim_{i \to \infty} P(A_i)$$

Hint: Use axiom 3.

*: This question is beyond the course content.

Solution:

a. $P(\neg A) = 1 - P(A)$

Let us consider a collection of events contained in the sample space $\{A_1, \ldots, A_N\} \subseteq \Omega$. Let us select the first i events (where $i \leq N$) and denote them as A. The rest of them will be the complementary set, denoted as $\neg A$.

$$A = \{A_1, \ldots, A_i\} \qquad \neg A = \{A_{i+1}, \ldots, A_N\} \qquad A \cap \neg A = \varnothing$$

Using axiom (2), we have

$$P(A_1 \cup A_2 \cup \cdots \cup A_N) = P(\Omega) = 1$$

And using axiom (3), we have

$$P(A_1, \ldots, A_i) + P(A_{i+1}, \ldots, A_N) = P(A_1 \cup A_2 \cup \cdots \cup A_N)$$

$$P(A) + P(\neg A) = 1$$

Thus

$$P(\neg A) = 1 - P(A)$$

b. $P(\varnothing) = 0$, where $\varnothing$ is the empty set

We can just consider the sample space, Ω, where its complementary is the empty set $\varnothing$. Using the previous property and axiom 2, we have.

$$P(\Omega) = 1$$

$$P(\varnothing) = P(\neg\Omega) = 1 - P(\Omega) = 1 - 1 = 0$$

$$P(\varnothing) = 0$$

c. $0 \leq P(A) \leq 1$

Here we can also use property (a) and the first axiom. Consider any event A.

$$P(A) \geq 0$$

$$P(\neg A) = 1 - P(A) \geq 0$$

$$1 \geq P(A)$$

We can join the previous inequalities and obtain the following.

$$0 \leq P(A) \leq 1$$

d. $A \subseteq B \implies P(A) \leq P(B)$

Hint: Consider the following definition. $B \backslash A = \{x \in B : x \notin A\}$

We can construct B as the union of two disjoint sets.

$$B = B \backslash A \cup A$$

where $B \backslash A \cap A = \varnothing$ by definition of $B \backslash A$. Let us use axiom 3 and 1.

$$P(B) = P(B \backslash A) + P(A) \geq P(A)$$

where by axiom 1, we have $P(B \backslash A) \geq 0$. Thus

$$P(A) \leq P(B)$$

e. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Let us define the union $(A \cup B)$ in terms of two disjoint sets.

$$(A \cup B) = A \cup B \backslash A$$

where $A \cap B \backslash A = \varnothing$ by definition. Using axiom 3 we have

$$P(A \cup B) = P(A) + P(B \backslash A)$$

To calculate $P(B \backslash A)$, let us define B in terms of A, and the union of two disjoint sets.

$$B = (B \cap A) \cup (B \backslash A)$$

where $(B \cap A) \cap (B \backslash A) = \varnothing$ by definition. Using also axiom 3, we have.

$$P(B) = P(B \cap A) + P(B \backslash A)$$

$$P(B \backslash A) = P(B) - P(B \cap A)$$

Therefore, the probability of $(A \cup B)$ is the following

$$P(A \cup B) = P(A) + P(B \backslash A) = P(A) + P(B) - P(B \cap A)$$

f. (*****) if $\{A_i\}_{i=1}^{\infty} \subseteq \Omega$ and $A_{i-1} \subseteq A_i \quad \forall i > 0$ then:

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \lim_{i \to \infty} P(A_i)$$

Hint: Use axiom 3.

Let us define the following

$$A := \bigcup_{i=1}^{\infty} A_i$$

We would like to write A in terms of disjoint sets so as to use axiom 3.

$$A_{i-1} \subseteq A_i \quad \forall i > 0 \implies A = \bigcup_{i=1}^{\infty} A_i \backslash A_{i-1}$$

The previous expression holds if we have $A_0 = \varnothing$. Notice this new expression can be regarded as starting with A_1 and adding the information form $A_2, A_3, \dots$ which is not previously considered (e.g $A_2 \backslash A_1, A_3 \backslash A_2, \dots$). Since this construction is a union of disjoint sets, we now can use axiom 3.

$$P(A) = P\left(\bigcup_{i=1}^{\infty} A_i \backslash A_{i-1}\right) = \sum_{i=1}^{\infty} P(A_i \backslash A_{i-1})$$

The infinite summation is in fact defined as a limit.

$$P(A) = \sum_{i=1}^{\infty} P(A_i \backslash A_{i-1}) = \lim_{n \to \infty} \sum_{i=1}^{n} P(A_i \backslash A_{i-1})$$

Notice the result in exercise (d), where we obtained $P(B) = P(B \backslash A) + P(A)$ for $A \subseteq B$. Therefore

$$P(A_i) = P(A_i \backslash A_{i-1}) + P(A_{i-1})$$
$$P(A_i \backslash A_{i-1}) = P(A_i) - P(A_{i-1})$$
$$P(A) = \lim_{n \to \infty} \sum_{i=1}^{n} P(A_i) - P(A_{i-1}) = \lim_{n \to \infty} \left(\sum_{i=1}^{n} P(A_i) - \sum_{i=1}^{n-1} P(A_i) \right) = \lim_{n \to \infty} P(A_n)$$

where we used $P(A_0) = P(\varnothing) = 0$. In conclusion,

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = P(A) = \lim_{i \to \infty} P(A_i)$$

4 Lecture 'Moments' exercise solutions

- $\mathbb{E}_{X,Y}[X + Y] = \mathbb{E}_X[X] + \mathbb{E}_Y[Y]$

$$\begin{aligned}
\mathbb{E}_{X,Y}[X + Y] &= \iint_{x,y \in \mathbb{R}} (x + y) p(x, y) dx dy \\
&= \iint_{x,y \in \mathbb{R}} x p(x, y) dx dy + \iint_{x,y \in \mathbb{R}} y p(x, y) dx dy \\
&= \int_{x \in \mathbb{R}} x p(x) dx + \int_{y \in \mathbb{R}} y p(y) dy \\
&= \mathbb{E}_X[X] + \mathbb{E}_Y[Y]
\end{aligned}$$

- $\mathbb{V}_X[X] = \mathbb{E}_X[X^2] - (\mathbb{E}_X[X])^2$

$$\begin{aligned}
\mathbb{V}_X[X] &= \mathbb{E}_X\left[\left(X - \mathbb{E}_X[X]\right)^2\right] \\
&= \mathbb{E}_X\left[X^2 - 2\mathbb{E}_X[X]X + \left(\mathbb{E}_X[X]\right)^2\right] \\
&= \mathbb{E}_X[X^2] - 2\mathbb{E}_X[X]\mathbb{E}_X[X] + \left(\mathbb{E}_X[X]\right)^2 = \\
&= \mathbb{E}_X[X^2] - \left(\mathbb{E}_X[X]\right)^2
\end{aligned}$$

- $\mathbb{V}_{X,Y}[X+Y] = \mathbb{V}_X[X] + \mathbb{V}_Y[Y]$, if Y and X are independent.

$$\begin{aligned}
\mathbb{V}_{X,Y}[X+Y] &= \mathbb{E}_{X,Y}\Big[\big(X+Y-\mathbb{E}_{X,Y}[X+Y]\big)^2\Big] = \mathbb{E}_{X,Y}\Big[\Big((X+Y)-\big(\mathbb{E}_X[X]+\mathbb{E}_Y[Y]\big)\Big)^2\Big] \\
&= \mathbb{E}_{X,Y}\Big[\Big(\big(X-\mathbb{E}_X[X]\big)+\big(Y-\mathbb{E}_Y[Y]\big)\Big)^2\Big] \\
&= \mathbb{E}_{X,Y}\Big[\big(X-\mathbb{E}_X[X]\big)^2+\big(Y-\mathbb{E}_Y[Y]\big)^2\Big] + 2\mathbb{E}_{X,Y}\Big[\big(X-\mathbb{E}_X[X]\big)\big(X-\mathbb{E}_Y[Y]\big)\Big] \\
&= \mathbb{E}_X\Big[\big(X-\mathbb{E}_X[Y]\big)^2\Big] + \mathbb{E}_Y\Big[\big(Y-\mathbb{E}_Y[Y]\big)^2\Big] + 2\mathbb{E}_{X,Y}\Big[\big(X-\mathbb{E}_X[X]\big)\big(Y-\mathbb{E}_Y[Y]\big)\Big] \\
&= \mathbb{V}_X[X] + \mathbb{V}_Y[Y] + 2\mathbb{E}_{X,Y}\Big[\big(X-\mathbb{E}_X[X]\big)\big(Y-\mathbb{E}_Y[Y]\big)\Big]
\end{aligned}$$

  $\mathbb{E}_{X,Y}\Big[\big(X-\mathbb{E}_X[X]\big)\big(Y-\mathbb{E}_Y[Y]\big)\Big]$ is the covariance between x and y. Let us show that it is 0 for independent variables, that is, for $p(x,y) = p(x)p(y)$.

$$\begin{aligned}
Cov[x,y] &= \mathbb{E}_{X,Y}\Big[XY + \mathbb{E}_X[X]\mathbb{E}_Y[Y] - \mathbb{E}_X[X]Y - \mathbb{E}_Y[Y]X\Big] \\
&= \mathbb{E}_{X,Y}[XY] + \mathbb{E}_X[X]\mathbb{E}_Y[Y] - \mathbb{E}_X[X]\mathbb{E}_Y[Y] - \mathbb{E}_Y[Y]\mathbb{E}_X[X] \\
&= \mathbb{E}_{X,Y}[XY] - \mathbb{E}_X[X]\mathbb{E}_Y[Y] \\
&= \iint_{x,y\mathbb{R}} xyp(x,y)dxdy - \mathbb{E}_X[X]\mathbb{E}_Y[Y] \\
&= \int_{x\in\mathbb{R}} xp(x)dx \int_{y\in\mathbb{R}} yp(y)dy - \mathbb{E}_X[X]\mathbb{E}_Y[Y] \\
&= \mathbb{E}_X[X]\mathbb{E}_Y[Y] - \mathbb{E}_X[X]\mathbb{E}_Y[Y] = 0
\end{aligned}$$

  Notice this is true since we used $p(x,y) = p(x)p(y)$. Therefore, the previous identity holds.

$$\mathbb{V}_{X,Y}[X+Y] = \mathbb{V}_X[X] + \mathbb{V}_Y[Y]$$

- $\mathbb{V}_{X,Y}[X-Y] = \mathbb{V}_X[X] + \mathbb{V}_Y[Y]$, if Y and X are independent.

$$\begin{aligned}
\mathbb{V}_{X,Y}[X-Y] &= \mathbb{E}_{X,Y}\Big[\big(X-Y-\mathbb{E}_{X,Y}[X-Y]\big)^2\Big] = \mathbb{E}_{X,Y}\Big[\Big((X-Y)-\big(\mathbb{E}_X[X]-\mathbb{E}_Y[Y]\big)\Big)^2\Big] \\
&= \mathbb{E}_{X,Y}\Big[\Big(\big(X-\mathbb{E}_X[X]\big)-\big(Y-\mathbb{E}_Y[Y]\big)\Big)^2\Big] \\
&= \mathbb{E}_{X,Y}\Big[\big(X-\mathbb{E}_X[X]\big)^2+\big(Y-\mathbb{E}_Y[Y]\big)^2\Big] - 2\mathbb{E}_{X,Y}\Big[\big(X-\mathbb{E}_X[X]\big)\big(Y-\mathbb{E}_Y[Y]\big)\Big] \\
&= \mathbb{E}_X\Big[\big(X-\mathbb{E}_X[X]\big)^2\Big] + \mathbb{E}_Y\Big[\big(Y-\mathbb{E}_Y[Y]\big)^2\Big] \\
&= \mathbb{V}_X[X] + \mathbb{V}_Y[Y]
\end{aligned}$$

  We already showed that $\mathbb{E}_{X,Y}\Big[\big(X-\mathbb{E}_X[X]\big)\big(Y-\mathbb{E}_Y[Y]\big)\Big] = 0$ if x and y are independent.

- $\mathbb{V}_X[cX] = c^2\mathbb{V}_X[X]$

$$\begin{aligned}
\mathbb{V}_X[cX] &= \mathbb{E}_X\Big[\big(cX - \mathbb{E}_X[cX]\big)^2\Big] \\
&= \mathbb{E}_X\Big[c^2X^2 - 2cX\mathbb{E}_X[cX] + \big(\mathbb{E}_X[cX]\big)^2\Big] \\
&= \mathbb{E}_X\Big[c^2X^2 - 2c^2X\mathbb{E}_X[X] + c^2\big(\mathbb{E}_X[X]\big)^2\Big] \\
&= c^2\mathbb{E}_X\Big[X^2 - 2X\mathbb{E}_X[X] + \big(\mathbb{E}_X[X]\big)^2\Big] \\
&= c^2\mathbb{E}_X\Big[\big(X - \mathbb{E}_X[X]\big)^2\Big] \\
&= c^2\mathbb{V}_X[X]
\end{aligned}$$

- $\mathbb{V}_{X,Y}[X+Y] = \mathbb{V}_X[X] + \mathbb{V}_Y[Y] + 2Cov[X,Y]$

  We have already proven that this expression holds.

$$\mathbb{V}_{X,Y}[X+Y] = \mathbb{V}_X[X]+\mathbb{V}_Y[Y]+2\mathbb{E}_{X,Y}\Big[\big(X-\mathbb{E}_X[X]\big)\big(Y-\mathbb{E}_Y[Y]\big)\Big] = \mathbb{V}_X[X]+\mathbb{V}_Y[Y]+2Cov[X,Y]$$

- Show that sampling with replacement gives an unbiased gradient estimator.

  Let us consider $f(x)$, where x is a random variable, where $x \sim X$. We would like to calculate the gradient of the following quantity.

$$\nabla\mathbb{E}_x[f(x)]$$

  Since this might yield complex analytical calculation, we define an estimator which calculates the previous value using sampling with replacement. In other words, we consider $\{x_1, \ldots, x_N\}$ which are i.i.d. and sampled according to x, $x_i \sim X$. The estimator is defined as follows.

$$\nabla\mathbb{E}_x[f(x)] \approx \nabla\frac{1}{N}\sum_{i=1}^{N} f(x_i)$$

  Let us prove that this estimator is unbiased.

$$\begin{aligned}
&\mathbb{E}_{x_1,\ldots,x_N}\Big[\nabla\frac{1}{N}\sum_{i=1}^{N} f(x_i)\Big] - \nabla\mathbb{E}_x[f(x)], \qquad \mathbb{E}_X[\nabla X] = \nabla\mathbb{E}_X[X] \\
=&\nabla\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{x_i}[f(x_i)] - \nabla\mathbb{E}_x[f(x)], \qquad \mathbb{E}_{X,Y}[X+Y] = \mathbb{E}_X[X] + \mathbb{E}_Y[Y] \\
=&\nabla\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_x[f(x)] - \nabla\mathbb{E}_x[f(x)], \qquad x_i \sim X, x \sim X \implies \mathbb{E}_{x_i}[f(x_i)] = \mathbb{E}_x[f(x)] \\
=&\nabla\frac{N}{N}\mathbb{E}_x[f(x)] - \nabla\mathbb{E}_x[f(x)] = 0
\end{aligned}$$

  Thus, the estimator is unbiased.

- In terms of v, the variance of a gradient estimator using a single element in the minibatch, compute the variance for a minibatch of size M.

  Let us consider a single element in the minibatch $\{x_1\}$.

$$v = \mathbb{V}_{x_1}\big[\nabla f(x_1)\big]$$

Let us now compute the variance for a minibatch of size M, with $\{x_1, \ldots, x_M\}$ being i.i.d. and sampled from X, as before $x_i \sim X$.

$$\begin{aligned}
\mathbb{V}_{x_1,\ldots,x_M}\left[\nabla \frac{1}{M}\sum_{i=1}^{M} f(x_i)\right] &= \frac{1}{M^2}\mathbb{V}_{x_1,\ldots,x_M}\left[\nabla \sum_{i=1}^{M} f(x_i)\right], \qquad \mathbb{V}_X[cX] = c^2\mathbb{V}_X[X] \\
&= \frac{1}{M^2}\mathbb{V}_{x_1,\ldots,x_M}\left[\sum_{i=1}^{M} \nabla f(x_i)\right] \\
&= \frac{1}{M^2}\sum_{i=1}^{M} \mathbb{V}_{x_i}\big[\nabla f(x_i)\big], \quad X, Y \text{ ind.} \implies \mathbb{V}_{X,Y}[X+Y] = \mathbb{V}_X[X] + \mathbb{V}_Y[Y] \\
&= \frac{1}{M^2}\sum_{i=1}^{M} v = \frac{M}{M^2}v = \frac{v}{M}
\end{aligned}$$

Thus

$$\mathbb{V}_{x_1,\ldots,x_M}\left[\nabla \frac{1}{M}\sum_{i=1}^{M} f(x_i)\right] = \frac{v}{M}$$

The variance of the estimator decreases with a factor of M for increasing sizes of minibatch.

5 Exercises chapter 6

6.1

$$p(x) = \sum_{y \in Y} p(x,y) = \begin{bmatrix} 0.01 + 0.05 + 0.1 \\ 0.02 + 0.1 + 0.05 \\ 0.03 + 0.05 + 0.03 \\ 0.1 + 0.07 + 0.05 \\ 0.1 + 0.2 + 0.04 \end{bmatrix} = \begin{bmatrix} 0.16 \\ 0.17 \\ 0.11 \\ 0.22 \\ 0.34 \end{bmatrix}$$

$$p(y) = \sum_{x \in X} p(x,y) = \begin{bmatrix} 0.01 + 0.02 + 0.03 + 0.1 + 0.1 \\ 0.05 + 0.1 + 0.05 + 0.07 + 0.2 \\ 0.1 + 0.05 + 0.03 + 0.05 + 0.04 \end{bmatrix} = \begin{bmatrix} 0.26 \\ 0.47 \\ 0.27 \end{bmatrix}$$

$$p(x|Y = y_1) = \frac{p(x, Y = y_1)}{p(Y = y_1)} = \frac{1}{0.26}\big[0.01,\ \ 0.02,\ \ 0.03,\ \ 0.1,\ \ 0.1\big]$$

$$p(x|Y = y_1) \approx \big[0.038,\ \ 0.077,\ \ 0.115,\ \ 0.385,\ \ 0.385\big]$$

$$p(y|X = x_3) = \frac{p(y, X = x_3)}{p(X = x_3)} = \frac{1}{0.11}\big[0.03,\ \ 0.05,\ \ 0.03\big] \approx \big[0.273,\ \ 0.273,\ \ 0.454\big]$$

6.4
Bag 1: 4 mangos, 2 apples.

$$p(mango|heads) = \frac{2}{3} \quad p(apple|heads) = \frac{1}{3}$$

Bag 2: 4 mangos, 4 apples.

$$p(mango|tails) = \frac{1}{2} \quad p(apple|tails) = \frac{1}{2}$$

Heads and tails distrib. is:

$$p(heads) = 0.6 \quad p(tails) = 0.4$$

Mango is presented as evidence. Therefore, we apply Bayes' rule to infer the probability that we picked the mango from bag 2. This is equivalent to computing the posterior distribution of obtaining *tails* given that a mango is taken.

$$p(tails|mango) = \frac{p(tails, mango)}{p(mango)}$$

$$p(tails, mango) = p(mango|tails)p(tails) = 0.5 \cdot 0.4 = 0.2$$

$$p(heads, mango) = p(mango|heads)p(heads) = \frac{2}{3} \cdot 0.6 = 0.4$$

$$p(mango) = p(tails, mango) + p(heads, mango) = 0.2 + 0.4$$

Finally

$$p(tails|mango) = \frac{p(tails, mango)}{p(mango)} = \frac{0.2}{0.2 + 0.4} = \frac{1}{3}$$

The probability of taking a mango from bag 2 is 0.333.

Note on why we need $p(tails|mango)$ and not $p(tails, mango)$: The answer is simple. Since we know that the outcome of the experiment is a *mango*, we consider it as evidence. Therefore, the posterior is the answer to this question.

6.6

Let us prove (6.44): $\mathbb{V}_X[x] = \mathbb{E}_X[x^2] - (\mathbb{E}_X[x])^2$. This is already solved above in this document.

$$\begin{aligned}
\mathbb{V}_X[x] &= \mathbb{E}_X\left[\left(x - \mathbb{E}_X[x]\right)^2\right] \\
&= \mathbb{E}_X\left[x^2 - 2\mathbb{E}_X[x]x + \left(\mathbb{E}_X[x]\right)^2\right] \\
&= \mathbb{E}_X[x^2] - 2\mathbb{E}_X[x]\mathbb{E}_X[x] + \left(\mathbb{E}_X[x]\right)^2 = \\
&= \mathbb{E}_X[x^2] - \left(\mathbb{E}_X[x]\right)^2
\end{aligned}$$

6.11

Let us consider the random variables x, y with joint distribution $p(x, y)$.

$$\begin{aligned}
\mathbb{E}_X[x] &= \mathbb{E}_Y\left[\mathbb{E}_X[x|y]\right] \\
&= \int_{y\in\mathbb{R}} \mathbb{E}_X[x|y]p(y)dy \\
&= \int_{y\in\mathbb{R}} \int_{x\in\mathbb{R}} xp(x|y)p(y)dxdy \\
&= \int_{x\in\mathbb{R}} x\left(\int_{y\in\mathbb{R}} p(x, y)dy\right)dx \\
&= \int_{x\in\mathbb{R}} xp(x)dx = \mathbb{E}_X[x]
\end{aligned}$$

Notice that we just swapped the integrals and marginalized y over the joint distribution to obtain $p(x)$.