CQF Module 4.3: Mathematics Toolbox for Machine Learning Answers

Panos Parpas

Exercise 1(Dot products)
$\boldsymbol{x}^\top \boldsymbol{y} = 1 \times 0 + (-2) \times 4 + 5 \times (-3) + (-1) \times 7 = 0 + (-8) + (-15) + (-7) = -30.$

Exercise 2(Matrix product)
$\boldsymbol{y} = (24, -14, -12)^\top$, $||\boldsymbol{x}||_2 = \sqrt{23}$, $||\boldsymbol{y}||_2 = \sqrt{916}$.
Note that by definition the ℓ_2 norm of a vector is $||\boldsymbol{x}||_2 = \sqrt{\boldsymbol{x}^\top \boldsymbol{x}}$.

Exercise 3(Basis) 1, 2.
A set of vectors $\{\mathbf{b}_1, ..., \mathbf{b}_K\}$ with $\mathbf{b}_k \in \mathbb{R}^d$ can form a basis of $\mathbb{R}^d$ iff $K = d$ the vectors are linearly independent to each other. Apart from 1,2 all others are linearly dependent.

Exercise 4(Span of vectors) 2, 5.
A point $\boldsymbol{x} \in \mathbb{R}^d$ is in $span(\{\mathbf{b}_1, ..., \mathbf{b}_K\})$ with $\mathbf{b}_k \in \mathbb{R}^d$ iff we can find $a_1, ..., a_K \in \mathbb{R}$ such that $\boldsymbol{x} = \sum_{k=1}^K a_k \mathbf{b}_k$.

Exercise 5(Eigen decomposition). a) When A is symmetric, then $A = Q\Lambda Q^\top$, and $\boldsymbol{x}^\top A \boldsymbol{x} = \boldsymbol{x}^\top Q \Lambda Q^\top \boldsymbol{x} = (Q^\top \boldsymbol{x})^\top \Lambda (Q^\top \boldsymbol{x})$. As Q is an orthonormal matrix, we have $\boldsymbol{x} \to Q^\top \boldsymbol{x}$ a one-to-one mapping. Therefore we have

$$\boldsymbol{x}^\top A \boldsymbol{x} = \boldsymbol{z}^\top \Lambda \boldsymbol{z} = \sum_{i=1}^d \lambda_i z_i^2, \quad \boldsymbol{z} = (z_1, ..., z_d)^\top = Q^\top \boldsymbol{x}.$$

Therefore $\boldsymbol{x}^\top A \boldsymbol{x} \geq 0 \Leftrightarrow \sum_{i=1}^d \lambda_i z_i^2 \geq 0$. This is true for any $\boldsymbol{x} \in \mathbb{R}^{d \times 1}$ if and only if $\lambda_i \geq 0$ for all $i = 1, ..., d$.

b) We use the permutation invariance property of matrix trace to show the result:

$$\mathrm{Tr}(A) = \mathrm{Tr}(Q\Lambda Q^{-1}) = \mathrm{Tr}(Q^{-1}Q\Lambda) = \mathrm{Tr}(\Lambda) = \sum_{i=1}^{d} \lambda_i.$$

c) We use the product rule of matrix determinant to show the result:

$$\begin{aligned}\det(A) = \det(Q\Lambda Q^{-1}) = \det(Q)\det(\Lambda)\det(Q^{-1}) &= \det(Q)\det(\Lambda)\det(Q)^{-1} \\ &= \det(\Lambda) = \prod_{i=1}^{d} \lambda_i.\end{aligned}$$

Exercise 6(Vector notation) Given $p(\mathbf{x}) = \frac{1}{C}(x_1^2 + x_1x_2x_2^2 + 2x_2x_3)$, we need to rearrange the terms to find an expression as follows:

$$p(\mathbf{x}) = \frac{1}{C}(\mathbf{x}^T A\mathbf{x}), \quad A \in \mathbb{R}^{3\times 3}. \tag{1}$$

Inspection of the terms in $p(\mathbf{x})$ gives the following solution

$$\begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix}^T \begin{pmatrix} 1 & \frac{1}{2} & 0 \\ \frac{1}{2} & 1 & 1 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix} = \begin{pmatrix} x_1 + \frac{x_2}{2} \\ \frac{x_1}{2} + x_2 + x_3 \\ x_2 \end{pmatrix} \begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix} = Cp(\mathbf{x}).$$

Thus,

$$p(\mathbf{x}) = \frac{1}{C}(\mathbf{x}^T A\mathbf{x}), \quad A = \begin{pmatrix} 1 & \frac{1}{2} & 0 \\ \frac{1}{2} & 1 & 1 \\ 0 & 1 & 0 \end{pmatrix}. \tag{2}$$

Exercise 7(Index notation).

Matrix-vector expressions to index notation:

1. $\boldsymbol{ABC}\mathbf{x} = \sum_{jkl} A_{ij}B_{jk}C_{kl}x_l$
2. $\mathrm{Tr}(\boldsymbol{A}) = \sum_i A_{ii}$
3. $\mathrm{Tr}(\boldsymbol{AB}) = \sum_{ij} A_{ij}B_{ji}$
4. $\mathbf{y}^\intercal \boldsymbol{A}^\intercal \mathbf{x} = \sum_{ij} y_i A_{ji} x_j$

Exercise 8 (Jacobian).

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

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

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

2. $f(\mathbf{x}) = \mathbf{x}^T\mathbf{y}, \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$

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

We can solve this directly using basic rules of vector calculus

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

We can confirm this result holds with index notation. First, let us calculate the value $f(\mathbf{x})$

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

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

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

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

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

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

$$c_{ijk} = \frac{\partial f(\mathbf{x})_{ij}}{\partial x_k}$$

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

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

4. $f(\mathbf{t}) = \sin\left(\log(\mathbf{t}^T\mathbf{t})\right) \quad \mathbf{t} \in \mathbb{R}^D$

   We directly apply the chain rule

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

   All of the terms are one dimensional except for $\frac{\partial(\mathbf{t}^T\mathbf{t})}{\partial \mathbf{t}} \in \mathbb{R}^{1\times D}$. 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 $f(\mathbf{t})$.

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

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

5. $f(X) = \mathrm{Tr}(AXB), \quad A \in \mathbb{R}^{D\times E}, X \in \mathbb{R}^{E\times F}, B \in \mathbb{R}^{F\times D}$

   Use index notation:

$$f(X) = \mathrm{Tr}(AXB) = \sum_{i=1}^{D}(AXB)_{ii}$$

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

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

Thus

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

Now we can just calculate the derivative using index notation

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

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

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

Using this last result, we can calculate the derivative w.r.t. X.

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

Exercise 9[Chain rule]

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

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

Dimensions are

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

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

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

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

Dimensions are

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

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

   Let us expand $f(x)$.

$$f(x) = \sum_{i=1}^{D}\Big((\mathbf{x}\mathbf{x}^T)_{ii} + \sigma^2\Big)$$
$$= \sum_{i=1}^{D}(\mathbf{x}\mathbf{x}^T)_{ii} + D\sigma^2 = \sum_{i=1}^{D} x_i^2 + D\sigma^2$$

   We already know that $(\mathbf{x}\mathbf{x}^T)_{ij} = x_i x_j$. Therefore

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

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

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

   where we used $\frac{d\tanh(v)}{dv} = 1 - \tanh^2(v)$.

   Dimensions are

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

Exercise 10(Hessian of Linear Regression).

The objective function and gradient w.r.t. $\boldsymbol{\theta}$ (see lectures) for Linear Regression is

$$L(\boldsymbol{\theta}) = |\mathbf{y} - \Phi(X)\boldsymbol{\theta}|^2\,, \qquad \frac{\mathrm{d}L}{\mathrm{d}\boldsymbol{\theta}} = 2(\Phi(X)\boldsymbol{\theta} - \mathbf{y})^\intercal \Phi(X)\,.$$

We begin by finding the Hessian, i.e. the matrix containing all second partial derivatives. It is easier to do this using index notation. So we first write the derivative in index notation, and then we take the derivative again, after which we return to vector notation:

$$\frac{\partial}{\partial\theta_j}\left(\frac{\partial L}{\partial\theta_i}\right) = \frac{\partial}{\partial\theta_j}\left(2\sum_k\left(\sum_m \Phi_{km}\theta_m - y_k\right)\Phi_{ki}\right) = \frac{\partial}{\partial\theta_j}\left(2\sum_k\left(\sum_m \Phi_{km}\theta_m - y_k\right)\Phi_{ki}\right) \tag{3}$$

$$= 2\sum_{km}\Phi_{km}\delta_{mj}\Phi_{ki} = 2\sum_k \Phi_{kj}\Phi_{ki}\,, \tag{4}$$

$$\implies \boldsymbol{H}_{\boldsymbol{\theta}}(L) = 2\Phi(X)^\intercal\Phi(X)\,. \tag{5}$$

The Hessian doesn't depend on the parameter $\boldsymbol{\theta}$, so if we prove that the matrix is positive definite, then if $\frac{\mathrm{d}L}{\mathrm{d}\boldsymbol{\theta}} = 0$ we will have a local minimum. For a matrix to be PD, we need $\mathbf{v}^\intercal\boldsymbol{H}\mathbf{v} > 0$ for all $\mathbf{v}$. We substitute our Hessian into $\boldsymbol{H}$ to prove this

$$\mathbf{v}^\intercal\boldsymbol{H}\mathbf{v} = 2\mathbf{v}^\intercal\Phi(X)^\intercal\Phi(X)\mathbf{v} \tag{6}$$

$$= \mathbf{w}^\intercal\mathbf{w} = \sum_i w_i^2\,, \qquad \text{with } \mathbf{v} = \Phi(X)\mathbf{v}\,. \tag{7}$$

This already shows that $\mathbf{v}^\intercal\boldsymbol{H}\mathbf{v} \geq 0$, with equality if there exists a $\mathbf{v}$ such that $\Phi(X)\mathbf{v} = 0$. So now we need to prove or assume that that *there cannot be* a $\mathbf{v}$ for which $\Phi(X)\mathbf{v} = 0$.

If we are coding up a linear regression problem, and we want to check numerically for a *specific* regression problem whether there is a unique solution, we can compute the eigenvalues of $\Phi(X)^\intercal\Phi(X)$, and see if they are all positive. This implies a PD Hessian because

$$\mathbf{v}^\intercal\boldsymbol{H}\mathbf{v} = \mathbf{v}^\intercal\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{-1}\mathbf{v} \quad \text{(eigenvalue decomposition)} \tag{8}$$

$$= \mathbf{v}^\intercal\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\intercal\mathbf{v} \quad (\boldsymbol{H} = \boldsymbol{H}^\intercal\text{, so } \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{-1} = (\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{-1})^\intercal\text{, so } \boldsymbol{Q}^{-1} = Q^\intercal) \tag{9}$$

$$= \mathbf{z}^\intercal\Lambda\mathbf{z}\,, \tag{10}$$

which is only > 0 if all the elements in the diagonal matrix Λ are positive.

Exercise 11(SVD and PCA) Let us write an SVD of $\mathbf{X}$ as $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^\top$ with $\mathbf{U} \in \mathbb{R}^{N\times N}, \Sigma \in \mathbb{R}^{N\times D}$ and $\mathbf{V} \in \mathbb{R}^{D\times D}$. Note that the covariance on $\mathcal{D}$ can be computed as $\mathbf{S}\ \mathbf{S} = \mathbf{X}^\top\mathbf{X}$. Plugging in the SVD of $\mathbf{X}$:

$$\mathbf{S} = \mathbf{X}^\top\mathbf{X} = \mathbf{V}\Sigma^\top\mathbf{U}^\top\mathbf{U}\Sigma\mathbf{V}^\top = \mathbf{V}\Sigma^\top\Sigma\mathbf{V}^\top.$$

Note that in an SVD, Σ is a rectangular diagonal matrix, i.e., only the leading diagonal terms have non-zero values. This also means $\Sigma^\top\Sigma \in \mathbb{R}^{D\times D}$ is a diagonal matrix with non-negative diagonal values. Therefore $\mathbf{V}\Sigma^\top\Sigma\mathbf{V}^\top$ is an eigendecomposition of $\mathbf{S}$, therefore by sorting the diagonal values in $\Sigma^\top\Sigma$ in descending order, we can retrieve the corresponding columns in $\mathbf{V}$ as the principal components.