2 Vector Differentiation

Gradient vectors and Hessian matrices play an important role in the development of iterative algorithms for inference and learning. In this chapter, we define the notions of first and second-order differentiation for functions of vector arguments, and introduce the notation for future chapters.

2.1 GRADIENT VECTORS

We describe two closely related differentiation operations that we will be using regularly in our treatment of inference and learning problems.

Let $z \in \mathbb{R}^M$ denote a real-valued *column* vector with M entries denoted by

$$z = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_M \end{bmatrix} \tag{2.1}$$

Let also $g(z) : \mathbb{R}^M \to \mathbb{R}$ denote a real-valued function of the vector argument, z. We denote the gradient vector of $g(z)$ with respect to z by the notation $\nabla_z\, g(z)$ and define it as the following *row* vector

$$\nabla_z\, g(z) \triangleq \begin{bmatrix} \frac{\partial g}{\partial z_1} & \frac{\partial g}{\partial z_2} & \dots & \frac{\partial g}{\partial z_M} \end{bmatrix}, \quad \begin{cases} z \text{ is a } \textbf{column} \\ \nabla_z\, g(z) \text{ is a } \textbf{row} \end{cases} \tag{2.2}$$

Note that the gradient is defined in terms of the partial derivatives of $g(z)$ relative to the individual entries of z.

Jacobian

Expression (2.2) for $\nabla_z\, g(z)$ is related to the concept of a *Jacobian* matrix for *vector-valued* functions.. Specifically, consider a second function $h(z) : \mathbb{R}^M \to \mathbb{R}^N$, which now maps z into a *vector*, assumed of dimension N and whose individual entries are denoted by

$$h(z) = \text{col}\Big\{h_1(z),\, h_2(z),\, h_3(z),\, \dots, h_N(z)\Big\} \tag{2.3}$$

The Jacobian of $h(z)$ relative to z is defined as the matrix:

$$\nabla_z\, h(z) \;\stackrel{\Delta}{=}\; \begin{bmatrix} \partial h_1/\partial z_1 & \partial h_1/\partial z_2 & \partial h_1/\partial z_3 & \ldots & \partial h_1/\partial z_M \\ \partial h_2/\partial z_1 & \partial h_2/\partial z_2 & \partial h_2/\partial z_3 & \ldots & \partial h_2/\partial z_M \\ \partial h_3/\partial z_1 & \partial h_3/\partial z_2 & \partial h_3/\partial z_3 & \ldots & \partial h_3/\partial z_M \\ \vdots & \vdots & \vdots & \vdots & \\ \partial h_N/\partial z_1 & \partial h_N/\partial z_2 & \partial h_N/\partial z_3 & \ldots & \partial h_N/\partial z_M \end{bmatrix} \tag{2.4}$$

Thus, note that if $h(z)$ were scalar-valued, with $N = 1$, then its Jacobian will reduce to the first row in the above matrix, which agrees with the definition for the gradient vector in (2.2).

In the same token, we will denote the gradient vector of $g(z)$ relative to the *row* vector z^{T} by $\nabla_{z^{\mathsf{T}}}\, g(z)$ and define it as the *column* vector:

$$\nabla_{z^{\mathsf{T}}}\, g(z) \;\stackrel{\Delta}{=}\; \begin{bmatrix} \partial g/\partial z_1 \\ \partial g/\partial z_2 \\ \vdots \\ \partial g/\partial z_M \end{bmatrix}, \qquad \begin{cases} z^{\mathsf{T}} \text{ is a } \textbf{row} \\ \nabla_{z^{\mathsf{T}}}\, g(z) \text{ is a } \textbf{column} \end{cases} \tag{2.5}$$

It is clear that

$$\nabla_{z^{\mathsf{T}}}\, g(z) \;=\; \Big(\nabla_z\, g(z)\Big)^{\mathsf{T}} \tag{2.6}$$

Notation

Observe that we are defining the gradient of a function with respect to a *column* vector to be a *row* vector, and the gradient with respect to a *row* vector to be a *column* vector:

gradient vector relative to a column is a row (2.7a)

gradient vector relative to a row is a column (2.7b)

Some references may reverse these conventions, such as defining the gradient vector of $g(z)$ relative to the column z to be the column vector (2.5). There is no standard convention in the literature. To avoid any ambiguity, we make a distinction between differentiating relative to z and z^{T}. Specifically, we adopt the convention that the gradient relative to a column (row) is a row (column). The main motivation for doing so is because the results of differentiation that follow from this convention will be consistent with what we would expect from traditional differentiation rules from the calculus of single variables. This is illustrated in the next examples.

Observe further that the result of the gradient operation (2.5) relative to z^{T} is a vector that has the *same dimension* as z; we will also use the following alternative notation for this gradient vector when convenient:

$$\frac{\partial g(z)}{\partial z} \triangleq \begin{bmatrix} \partial g/\partial z_1 \\ \partial g/\partial z_2 \\ \vdots \\ \partial g/\partial z_M \end{bmatrix} \tag{2.8}$$

In this way, we end up with the following convention:

$$\begin{cases} \textbf{a)}\ z \text{ is } M \times 1 \text{ column vector;} \\ \textbf{b)}\ \partial g(z)/\partial z \text{ and } \nabla_{z^{\mathsf{T}}}\, g(z) \text{ coincide and have the same dimensions as } z; \\ \textbf{c)}\ \nabla_z\, g(z) \text{ has the same dimensions as } z^{\mathsf{T}} \end{cases} \tag{2.9}$$

The notation $\partial g(z)/\partial z$ organizes the partial derivatives in column form, while the notation $\nabla_z\, g(z)$ organizes the *same* partial derivatives in row form. The two operations of differentiation and gradient evaluation are the transpose of each other. We will be using these forms interchangeably.

Example 2.1 (**Calculations for vector arguments**) We consider a couple of examples.

(1) Let $g(z) = a^{\mathsf{T}} z$, where $\{a, z\}$ are column vectors in $\mathbb{R}^M$ with entries $\{a_m, z_m\}$. Then,

$$\begin{aligned} \nabla_z\, g(z) &\triangleq \begin{bmatrix} \partial g(z)/\partial z_1 & \partial g(z)/\partial z_2 & \ldots & \partial g(z)/\partial z_M \end{bmatrix} \\ &= \begin{bmatrix} a_1 & a_2 & \ldots & a_M \end{bmatrix} \\ &= a^{\mathsf{T}} \end{aligned} \tag{2.10}$$

Note further that since $g(z)$ is real-valued, we can also write $g(z) = z^{\mathsf{T}} a$ and a similar calculation gives

$$\nabla_{z^{\mathsf{T}}}\, g(z) = a \tag{2.11}$$

(2) Let $g(z) = \|z\|^2 = z^{\mathsf{T}} z$, where $z \in \mathbb{R}^M$. Then,

$$\begin{aligned} \nabla_z\, g(z) &\triangleq \begin{bmatrix} \partial g(z)/\partial z_1 & \partial g(z)/\partial z_2 & \ldots & \partial g(z)/\partial z_M \end{bmatrix} \\ &= \begin{bmatrix} 2z_1 & 2z_2 & \ldots & 2z_M \end{bmatrix} \\ &= 2z^{\mathsf{T}} \end{aligned} \tag{2.12}$$

Likewise, we get $\nabla_{z^{\mathsf{T}}}\, g(z) = 2z$.

(3) Let $g(z) = z^{\mathsf{T}} C z$, where C is a *symmetric* matrix in $\mathbb{R}^{M \times M}$ that does not depend on z. If we denote the individual entries of C by C_{mn}, then we can write

$$g(z) = \sum_{m=1}^{M} \sum_{n=1}^{M} C_{mn} z_m z_n \tag{2.13}$$

so that

$$\begin{aligned}\frac{\partial g(z)}{\partial z_m} &= 2C_{mm}z_m \;+\; \sum_{n\neq m}^{M}(C_{mn}+C_{nm})z_n \\ &= 2\sum_{n=1}^{M}C_{nm}z_n \\ &= 2z^{\mathsf{T}}C_{:,m} \end{aligned} \tag{2.14}$$

where in the second equality we used the fact that C is symmetric and, hence, $C_{mn} = C_{nm}$, and in the last equality we introduced the notation $C_{:,m}$ to refer to the $m-$th column of C. Collecting all partial derivatives $\{\partial g(z)/z_m\}$, for $m = 1, 2, \ldots, M$, into a row vector we conclude that

$$\nabla_z\, g(z) \;=\; 2z^{\mathsf{T}}C \tag{2.15}$$

(4) Let $g(z) = z^{\mathsf{T}}Cz$, where C is now an arbitrary (not necessarily symmetric) matrix in $\mathbb{R}^{M\times M}$. If we repeat the same argument as in part (3) we arrive at

$$\nabla_z\, g(z) \;=\; z^{\mathsf{T}}(C + C^{\mathsf{T}}) \tag{2.16}$$

(5) Let $g(z) = \kappa + a^{\mathsf{T}}z + z^{\mathsf{T}}b + z^{\mathsf{T}}Cz$, where κ is a scalar, $\{a, b\}$ are column vectors in $\mathbb{R}^M$, and C is a matrix in $\mathbb{R}^{M\times M}$. Then,

$$\nabla_z\, g(z) \;=\; a^{\mathsf{T}} + b^{\mathsf{T}} + z^{\mathsf{T}}(C + C^{\mathsf{T}}) \tag{2.17}$$

(6) Let $g(z) = Az$, where $A \in \mathbb{R}^{M\times M}$ does not depend on z. Then, the Jacobian matrix is given by $\nabla_z\, g(z) = A$.

2.2 HESSIAN MATRICES

Hessian matrices involve second-order partial derivatives. Consider again the real-valued function $g(z) : \mathbb{R}^M \to \mathbb{R}$. We continue to denote the individual entries of the column vector $z \in \mathbb{R}^M$ by $z = \text{col}\{z_1, z_2, \ldots, z_M\}$. The Hessian matrix of $g(z)$ is an $M \times M$ *symmetric* matrix function of z, denoted by $H(z)$, and whose $(m, n)-$th entry is constructed as follows:

$$[H(z)]_{m,n} \;\triangleq\; \frac{\partial^2 g(z)}{\partial z_m \partial z_n} \;=\; \frac{\partial}{\partial z_m}\left(\frac{\partial g(z)}{\partial z_n}\right) \;=\; \frac{\partial}{\partial z_n}\left(\frac{\partial g(z)}{\partial z_m}\right) \tag{2.18}$$

in terms of the partial derivatives of $g(z)$ with respect to the scalar arguments $\{z_m, z_n\}$. For example, for a two-dimensional argument z (i.e., $M = 2$), the four entries of the 2×2 Hessian matrix are:

$$H(z) \;=\; \begin{bmatrix} \dfrac{\partial^2 g(z)}{\partial z_1^2} & \dfrac{\partial^2 g(z)}{\partial z_1 \partial z_2} \\ \dfrac{\partial^2 g(z)}{\partial z_2 \partial z_1} & \dfrac{\partial^2 g(z)}{\partial z_2^2} \end{bmatrix} \tag{2.19}$$

It is straightforward to verify that $H(z)$ can also be obtained as the result of two successive gradient vector calculations with respect to z and z^{T} in the following manner (where the order of the differentiation does not matter):

$$H(z) \;\stackrel{\Delta}{=}\; \nabla_{z^{\mathsf{T}}}\Big(\nabla_z\, g(z)\Big) \;=\; \nabla_z\Big(\nabla_{z^{\mathsf{T}}}\, g(z)\Big) \tag{2.20}$$

For instance, using the first expression, the gradient operation $\nabla_z\, g(z)$ generates a $1 \times M$ (row) vector function and the subsequent differentiation with respect to z^{T} leads to the $M \times M$ Hessian matrix, $H(z)$. It is clear from (2.18) and (2.20) that the Hessian matrix is symmetric so that

$$H(z) \;=\; \Big(H(z)\Big)^{\mathsf{T}} \tag{2.21}$$

A useful property of Hessian matrices is that they help characterize the nature of stationary points for functions $g(z)$ that are twice differentiable. Specifically, if z^o is a stationary point of $g(z)$ (i.e., a point where $\nabla_z\, g(z) = 0$), then the following facts hold:

(a) z^o will correspond to a local minimum of $g(z)$ if $H(z^o) > 0$, i.e., if all eigenvalues of $H(z^o)$ are positive.

(b) z^o will correspond to a local maximum of $g(z)$ if $H(z^o) < 0$, i.e., if all eigenvalues of $H(z^o)$ are negative.

Example 2.2 **(Quadratic cost functions)** Consider $g(z) = \kappa + 2a^{\mathsf{T}}z + z^{\mathsf{T}}Cz$, where κ is a real scalar, a is a real column vector of dimension $M \times 1$, and C is an $M \times M$ real *symmetric* matrix. We know from (2.17) that

$$\nabla_z\, g(z) = 2a^{\mathsf{T}} + 2z^{\mathsf{T}}C \tag{2.22}$$

Differentiating again gives:

$$H(z) \;\stackrel{\Delta}{=}\; \nabla_{z^{\mathsf{T}}}\left(\nabla_z\, g(z)\right) \;=\; \nabla_{z^{\mathsf{T}}}\Big(2a^{\mathsf{T}} + 2z^{\mathsf{T}}C\Big) \;=\; 2C \tag{2.23}$$

We find that for quadratic functions $g(z)$, the Hessian matrix is independent of z. Moreover, any stationary point z^o of $g(z)$ should satisfy

$$2a^{\mathsf{T}} + 2(z^o)^{\mathsf{T}}C = 0 \;\;\Longrightarrow\;\; Cz^o = -a \tag{2.24}$$

The stationary point will be unique if C is invertible. And the unique z^o will be a global minimum of $g(z)$ if, and only if, $C > 0$.

2.3 MATRIX DIFFERENTIATION

We end the chapter with a list of useful matrix differentiation results, collected in Table 2.1. We leave the derivations to Probs. 2.10–2.14. The notation used in the table refers to the following definitions.

Let $X \in \mathbb{R}^{M\times M}$ be a square matrix whose individual entries are functions of

Table 2.1 Some useful matrix differentiation results; inverses are assumed to exist whenever necessary. The last column provides the problem numbers where these properties are established.

property	result	problem
1.	$\partial(A(\alpha))^{-1}/\partial\alpha = -A^{-1}(\partial A/\partial\alpha)A^{-1}$	2.10
2.	$\partial \det A/\partial\alpha = \det(A)\,\mathrm{Tr}\{A^{-1}(\partial A/\partial\alpha)\}$	2.10
3.	$\partial \ln\lvert\det A\rvert/\partial\alpha = \mathrm{Tr}\Big\{(A^{-1}(\partial A/\partial\alpha)\Big\}$	2.10
4.	$\partial A(\alpha)B(\alpha)/\partial\alpha = A(\partial B/\partial\alpha) + (\partial A/\partial\alpha)B$	2.10
5.	$\nabla_{X^\mathsf{T}}\,\mathrm{Tr}(AXB) = A^\mathsf{T}B^\mathsf{T}$	2.11
6.	$\nabla_{X^\mathsf{T}}\,\mathrm{Tr}(AX^{-1}B) = -(X^{-1}BAX^{-1})^\mathsf{T}$	2.11
7.	$\nabla_{X^\mathsf{T}}\,\mathrm{Tr}(X^{-1}A) = -(X^{-1}A^\mathsf{T}X^{-1})^\mathsf{T}$	2.11
8.	$\nabla_{X^\mathsf{T}}\,\mathrm{Tr}(X^\mathsf{T}AX) = (A + A^\mathsf{T})X$	2.11
9.	$\nabla_{X^\mathsf{T}}\,\mathrm{Tr}(X^\mathsf{T}X) = 2X$	2.11
10.	$\nabla_{X^\mathsf{T}}\,f(X) = -X^{-\mathsf{T}}\left(\nabla_{X^{-1}}f(X)\right)X^{-\mathsf{T}}$	2.11
11.	$\nabla_{X^\mathsf{T}}\,\det(X) = \det(X)X^{-\mathsf{T}}$	2.12
12.	$\nabla_{X^\mathsf{T}}\,\det(X^{-1}) = -\det(X^{-1})X^{-\mathsf{T}}$	2.12
13.	$\nabla_{X^\mathsf{T}}\,\det(AXB) = \det(AXB)(X^{-1})^\mathsf{T}$	2.12
14.	$\nabla_{X^\mathsf{T}}\,\det(X^\mathsf{T}AX) = 2\det(X^\mathsf{T}X)(X^{-1})^\mathsf{T}$	2.12
15.	$\nabla_{X^\mathsf{T}}\,\ln\lvert\det(X)\rvert = (X^{-1})^\mathsf{T}$	2.12
16.	$\nabla_{X^\mathsf{T}}\,\lVert X\rVert_\mathrm{F}^2 = 2X$	2.13
17.	$\nabla_{X^\mathsf{T}}\,\mathrm{Tr}(X^p) = pX^{p-1},\ p \in \mathbb{R}$	2.14

some real scalar α. We denote the individual entries of X by X_{mn} and define its derivative relative to α as the $M \times M$ matrix whose individual entries are the partial derivatives:

$$\frac{\partial X(\alpha)}{\partial\alpha} = \left[\frac{\partial X_{mn}}{\partial\alpha}\right]_{m,n}, \qquad (M \times M) \tag{2.25}$$

Likewise, let $f(X)$ denote some scalar real-valued function of a real-valued $M \times N$ matrix argument X. We employ two closely related operations to refer to differentiation operations applied to $f(X)$. The derivative of $f(\cdot)$ relative to X is defined as the $M \times N$ matrix whose individual entries are the partial derivatives of $f(X)$:

$$\frac{\partial f(X)}{\partial X} \;\triangleq\; \left[\frac{\partial f(X)}{\partial X_{mn}}\right]_{m,n} \;=\; \nabla_{X^\mathsf{T}}\,f(X), \qquad (M \times N) \tag{2.26}$$

Observe that the result has the *same* dimensions as X. Likewise, we define the gradient matrix of $f(\cdot)$ relative to X as the $N \times M$ matrix

$$\nabla_X f(X) = \left(\left[\frac{\partial f(X)}{\partial X_{mn}} \right]_{m,n} \right)^{\mathsf{T}}, \qquad (N \times M) \tag{2.27}$$

The result has the same dimensions as the transpose matrix, X^{T}. This construction is consistent with our earlier convention for the differentiation and gradient operations for vector arguments. In particular note that

$$\nabla_X f(X) \;=\; \left(\frac{\partial f(X)}{\partial X} \right)^{\mathsf{T}} \;=\; \Big(\nabla_{X^{\mathsf{T}}} f(X) \Big)^{\mathsf{T}} \tag{2.28}$$

2.4 COMMENTARIES AND DISCUSSION

Gradients and Hessians. Gradient vectors help identify the location of stationary points of a function and play an important role in the development of iterative algorithms for seeking these locations. Hessian matrices, on the other hand, help clarify the nature of a stationary point such as deciding whether it corresponds to a local minimum, a local maximum, or a saddle point. Hessian matrices are named after the German mathematician **Ludwig Hesse (1811–1874)** who introduced them in his study of quadratic and cubic curves — see the work collection by Dyck *et al.* (1972). A useful listing of gradient vector calculations for functions of real arguments is given by Petersen and Pedersen (2012). Some of these results appear in Table 2.1. For more discussion on first and second-order differentiation for functions of several variables, the reader may refer to Fleming (1987), Edwards (1995), Zorich (2004), Moskowitz and Paliogiannis (2011), Hubbard and Hubbard (2015), and Bernstein (2018).

PROBLEMS

2.1 Let $g(x,z) = x^{\mathsf{T}} C z$, where $x, z \in \mathbb{R}^M$ and C is a matrix. Verify that

$$\nabla_z\, g(x,z) = x^{\mathsf{T}} C, \quad \nabla_x\, g(x,z) = z^{\mathsf{T}} C^{\mathsf{T}}$$

2.2 Let $g(z) = Ah(z)$, where $A \in \mathbb{R}^{M\times M}$ and both $g(z)$ and $h(z)$ are vector-valued functions. Show that the Jacobian matrices of $g(z)$ and $h(z)$ are related as follows:

$$\nabla_z\, g(z) \;=\; A\, \nabla_z\, h(z)$$

2.3 Let $g(z) = (h(z))^{\mathsf{T}}\, h(z) = \|h(z)\|^2$, where $g(z)$ is scalar-valued while $h(z)$ is vector-valued with $z \in \mathbb{R}^M$. Show that

$$\nabla_z\, g(z) \;=\; 2\,(h(z))^{\mathsf{T}}\, \nabla_z\, h(z)$$

in terms of the Jacobian matrix of $h(z)$.

2.4 Let $g(z) = \frac{1}{2}\,(h(z))^{\mathsf{T}}\, A^{-1} h(z)$, where $A > 0$ and $g(z)$ is scalar-valued while $h(z)$ is vector-valued with $z \in \mathbb{R}^M$. Show that

$$\nabla_z\, g(z) \;=\; (h(z))^{\mathsf{T}}\, A^{-1} \nabla_z\, h(z)$$

in terms of the Jacobian matrix of $h(z)$.

2.5 Let $g(z) = x^{\mathsf{T}}Cw$, where $x, w \in \mathbb{R}^P$ and both are functions of a vector $z \in \mathbb{R}^M$, i.e., $x = x(z)$ and $w = w(z)$, while C is a matrix that is independent of z. Establish the chain rule

$$\nabla_z\, g(z) = x^{\mathsf{T}} C\, \nabla_z\, w(z) \;+\; w^{\mathsf{T}} C^{\mathsf{T}}\, \nabla_z\, x(z)$$

in terms of the Jacobian matrices of $x(z)$ and $w(z)$ relative to z.

2.6 Let $g(z)$ be a real-valued differentiable function with $z \in \mathbb{R}^M$. Assume the entries of z are functions of a scalar parameter t, i.e., $z = \text{col}\{z_1(t),\, z_2(t), \ldots, z_M(t)\}$. Introduce the column vector $dz/dt = \text{col}\{dz_1/dt,\, dz_2/dt, \ldots, dz_M/dt\}$. Show that

$$\frac{dg}{dt} \;=\; \nabla_z\, g(z)\, \frac{dz}{dt}$$

2.7 Let $g(z)$ be a real-valued function with $z \in \mathbb{R}^M$. Let $f(t)$ be a real-valued function with $t \in \mathbb{R}$. Both functions are differentiable in their arguments. Show that

$$\nabla_z\, f(g(z)) \;=\; \left(\left.\frac{df(t)}{dt}\right|_{t=g(z)} \right) \times \nabla_z\, g(z)$$

2.8 Let $g(z)$ be a real-valued twice-differentiable function with $z \in \mathbb{R}^M$. Define $f(t) = g(z + t\Delta z)$ for $t \in [0, 1]$. Show from first principles that

$$\begin{aligned} \frac{df(t)}{dt} &= (\nabla_z\, g(z + t\Delta z))\, \Delta z \\ \frac{d^2 f(t)}{dt^2} &= (\Delta z)^{\mathsf{T}} \left(\nabla_z^2\, g(z + t\Delta z)\right) \Delta z \end{aligned}$$

2.9 Let $\|x\|_p$ denote the $p-$th norm of vector $x \in \mathbb{R}^M$. Verify that

$$\nabla_{x^{\mathsf{T}}}\, \|x\|_p = \frac{x \odot |x|^{p-2}}{\|x\|_p^{p-1}}$$

where $\odot$ denotes the Hadamard (elementwise) product of two vectors, and the notation $|x|^{p-2}$ refers to a vector whose individual entries are the absolute values of the entries of x raised to the power $p - 2$.

2.10 Let $A \in \mathbb{R}^{M\times M}$ be a matrix whose individual entries are functions of some real scalar α. We denote the individual entries of A by A_{mn} and define its derivative relative to α as the $M \times M$ matrix whose individual entries are given by the partial derivatives:

$$\left[\frac{\partial A}{\partial \alpha}\right]_{m,n} = \frac{\partial A_{mn}}{\partial \alpha}$$

Establish the following relations:
(a) $\partial A^{-1}/\partial\alpha = -A^{-1}(\partial A/\partial\alpha)A^{-1}$.
(b) $\partial AB/\partial\alpha = A(\partial B/\partial\alpha) + (\partial A/\partial\alpha)B$, for matrices A and B that depend on α.
(c) $\partial \det A/\partial\alpha = \det(A)\, \text{Tr}\Big\{A^{-1}(\partial A/\partial\alpha)\Big\}$.
(d) $\partial \ln|\det A|/\partial\alpha = \text{Tr}\Big\{A^{-1}(\partial A/\partial\alpha)\Big\}$.

2.11 Let $f(X)$ denote a scalar real-valued function of a real-valued $M \times N$ matrix argument X. Let X_{mn} denote the $(m, n)-$th entry of X. The gradient of $f(\cdot)$ relative to X^{T} is defined as the $M \times N$ matrix whose individual entries are given by the partial derivatives:

$$\Big[\nabla_{X^{\mathsf{T}}}\, f(X)\Big]_{m,n} = \frac{\partial f(X)}{\partial X_{mn}}$$

Establish the following differentiation results (assume X is square and/or invertible when necessary):
(a) $\nabla_{X^{\mathsf{T}}}\, \text{Tr}(AXB) = A^{\mathsf{T}}B^{\mathsf{T}}$.

(b) $\nabla_{X^\mathsf{T}} \operatorname{Tr}(AX^{-1}B) = -(X^{-1}BAX^{-1})^\mathsf{T}$.
(c) $\nabla_{X^\mathsf{T}} \operatorname{Tr}(X^{-1}A) = -(X^{-1}A^\mathsf{T}X^{-1})^\mathsf{T}$.
(d) $\nabla_{X^\mathsf{T}} \operatorname{Tr}(X^\mathsf{T}AX) = (A + A^\mathsf{T})X$.
(e) $\nabla_{X^\mathsf{T}} \operatorname{Tr}(X^\mathsf{T}X) = 2X$.
(f) $\nabla_{X^\mathsf{T}} f(X) = -X^{-\mathsf{T}} \left(\nabla_{X^{-1}} f(X)\right) X^{-\mathsf{T}}$.

2.12 Consider the same setting of Prob. 2.11. Show that
(a) $\nabla_{X^\mathsf{T}} \det(X) = \det(X)X^{-\mathsf{T}}$.
(b) $\nabla_{X^\mathsf{T}} \det(X^{-1}) = -\det(X^{-1})X^{-\mathsf{T}}$.
(c) $\nabla_{X^\mathsf{T}} \det(AXB) = \det(AXB)(X^{-1})^\mathsf{T}$.
(d) $\nabla_{X^\mathsf{T}} \det(X^\mathsf{T}AX) = 2\det(X^\mathsf{T}X)(X^{-1})^\mathsf{T}$, for square invertible X.
(e) $\nabla_{X^\mathsf{T}} \ln|\det(X)| = (X^{-1})^\mathsf{T}$.

2.13 Let $\|X\|_\mathrm{F}$ denote the Frobenius norm of matrix X, as was defined in Table 1.4. Show that $\nabla_{X^\mathsf{T}} \|X\|_\mathrm{F}^2 = 2X$.

2.14 Let $p \in \mathbb{R}$ and consider a positive-definite matrix, X. Show that $\nabla_{X^\mathsf{T}} \operatorname{Tr}(X^p) = pX^{p-1}$.

2.15 For the purposes of this problem, let the notation $\|X\|_1$ denote the sum of the absolute values of all entries of X. Find $\nabla_{X^\mathsf{T}} \|XX^\mathsf{T}\|_1$.

REFERENCES

Bernstein, D. S. (2018), *Scalar, Vector, and Matrix Mathematics: Theory, Facts, and Formulas*, revised edition, Princeton University Press, NJ.

Dyck, W., S. Gundelfinger, J. Luroth, and M. Noether (1972), *Ludwig Otto Hesse's Gesammelte Werke*, published originally in 1897 by the Bavarian Academy of Sciences and Humanities. Republished in 1972 by Chelsea Publishing Co., NY.

Edwards, Jr., C. H. (1995), *Advanced Calculus of Several Variables*, Dover, NY.

Fleming, W. (1987), *Functions of Several Variables*, 2nd edition, Springer, NY.

Hubbard, B. B. and J. Hubbard (2015), *Vector Calculus, Linear Algebra, and Differential Forms: A Unified Approach*, 5th edition, Matrix Editions, NY.

Moskowitz, M. and F. Paliogiannis (2011), *Functions of Several Real Variables*, World Scientific, Singapore.

Petersen, K. B. and M. S. Pedersen (2012), "The matrix cookbook," *available online at http://matrixcookbook.com*.

Zorich, V. A. (2004), *Mathematical Analysis*, vols. I and II, Springer, Berlin Heidelberg.