$$f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{y}\|^2 + \lambda \|\mathbf{x}\|_p^p. \tag{2.17}$$

Such a result will be needed when we get to the regression chapter.

Lastly, we mention a useful fact that the α-sublevel sets of a convex function must be convex sets.

Theorem 2.5 *Let $f : \Omega \subseteq \mathbb{R}^n \mapsto \mathbb{R}$ be a convex function. For any $\alpha \in \mathbb{R}$, $S_\alpha(f)$ is a convex set in $\mathbb{R}^n$.*

Proof Suppose $\mathbf{x}, \mathbf{y} \in S_\alpha(f)$, that is $f(\mathbf{x}) \le \alpha$ and $f(\mathbf{y}) \le \alpha$. For any $t \in (0, 1)$,

$$f(t\mathbf{x} + (1-t)\mathbf{y}) \le t f(\mathbf{x}) + (1-t) f(\mathbf{y}) \le t\alpha + (1-t)\alpha = \alpha.$$

Thus, $t\mathbf{x} + (1-t)\mathbf{y} \in S_\alpha(f)$. This shows that $S_\alpha(f)$ is indeed a convex set. □

Example 2.5 The following are all convex sets in $\mathbb{R}^n$ because they are sublevel sets of convex functions:

- **ℓ_p-balls**: $\{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_p \le r\}$ where $p \ge 1$ and $r > 0$;
- **Half spaces**: $\{\mathbf{x} \in \mathbb{R}^n : \mathbf{a}^T\mathbf{x} \le b\}$ where $\mathbf{a} \ne \mathbf{0} \in \mathbb{R}^n$ and $b \in \mathbb{R}$;
- **Ellipsoids** (centered at the origin): $\{\mathbf{x} \in \mathbb{R}^n : \mathbf{x}^T\mathbf{A}\mathbf{x} \le b\}$ where $\mathbf{A} \in \mathbb{R}^{n\times n}$ is a positive definite matrix and $b \in \mathbb{R}$.

See Figure 2.7.

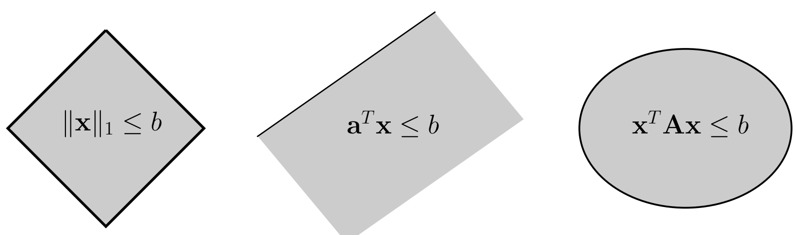

Fig. 2.7 Convex sets in $\mathbb{R}^n$: a solid ℓ_1-ball (left), a half-space (middle), and an ellipsoid (right).

2.3 Derivatives of a function of several variables

2.3.1 Basic concepts

Let $\Omega \subseteq \mathbb{R}^n$ be a set and $f : \Omega \mapsto \mathbb{R}$ a function over the set. The function f is said to be *differentiable* at an interior point $\mathbf{x}_0 \in \Omega$ if all the partial derivatives of f, $\{\frac{\partial f}{\partial x_i} : i = 1, \dots, n\}$, exist at $\mathbf{x}_0$. If additionally all the partial derivatives of f are continuous at $\mathbf{x}_0$, then f is said to be *continuously differentiable* at the point.

Similarly, if all the second-order partial derivatives of f, $\{\frac{\partial^2 f}{\partial x_i \partial x_j} : 1 \le i, j \le n\}$, exist at an interior point $\mathbf{x}_0 \in \Omega$, we say that f is *twice differentiable* at $\mathbf{x}_0$. If additionally all the second-order partial derivatives of f are continuous at $\mathbf{x}_0$, then f is said to be *twice continuously differentiable* at the point.

Suppose that $f : \Omega \mapsto \mathbb{R}$ is differentiable in all or a nonempty subset of the interior of Ω. The *gradient* of f is a vector field (i.e., vector-valued function) whose components are the partial derivatives of f:

$$\nabla f : \Omega \subseteq \mathbb{R}^n \longmapsto \mathbb{R}^n, \qquad \text{with} \quad \nabla f = \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n}\right)^T. \tag{2.18}$$

When given a specific point $\mathbf{p} \in \Omega^o$ at which f is differentiable, we may evaluate the gradient ∇f at $\mathbf{p}$ to obtain a gradient vector:

$$\nabla f(\mathbf{p}) = \left(\frac{\partial f}{\partial x_1}(\mathbf{p}), \ldots, \frac{\partial f}{\partial x_n}(\mathbf{p})\right)^T \in \mathbb{R}^n. \tag{2.19}$$

On the other hand, if any of the partial derivatives is undefined at $\mathbf{p}$, we say that the gradient of the function f does not exist at the location $\mathbf{p}$.

In the cases where the expression of the function f contains several parameters besides the variable $\mathbf{x}$, e.g., $f(\mathbf{x}) = \mathbf{a}^T \mathbf{x}$ where $\mathbf{a}$ is a constant vector regarded as the parameter of the function, we denote the gradient of f by $\nabla_{\mathbf{x}} f$ or $\frac{\partial f}{\partial \mathbf{x}}$ instead (to indicate clearly that the partial derivatives are taken only with respect to $\mathbf{x}$).

If $\nabla f(\mathbf{p}) \neq \mathbf{0}$ for some interior point $\mathbf{p} \in \Omega$, then at that point the gradient vector is perpendicular to the level set $L_\alpha(f)$ in $\mathbb{R}^n$, where $\alpha = f(\mathbf{p})$. That is, $\nabla f(\mathbf{p})$ is perpendicular to the tangent plane $L_\alpha(f)$ at the point $\mathbf{p}$. Figure 2.8 displays a function of two variables and an α-level set (contour curve) where $\alpha = f(\mathbf{p})$. The gradient at $\mathbf{p}$ is perpendicular to the contour curve at location $\mathbf{p}$, which means that $\nabla f(\mathbf{p})$ is perpendicular to the tangent line of the curve at $\mathbf{p}$. Additionally, the direction of the gradient $\nabla f(\mathbf{p})$ is the direction in which the function f increases the most rapidly from the point $\mathbf{p}$, and the magnitude of $\nabla f(\mathbf{p})$ is the rate of increase in that direction. On the contrary, $-\nabla f(\mathbf{p})$ is the direction of the fastest decrease of f at location $\mathbf{p}$.

We say that a point in the domain of the given function f, $\mathbf{x}_0 \in \Omega$, is a *critical point* of f if $\nabla f(\mathbf{x}_0) = \mathbf{0}$ (such as the point $\mathbf{x}_0$ in Figure 2.8), or the gradient does not exist at $\mathbf{x}_0$. The *critical values* are the values of the function f at the critical points.

For any function $f : \Omega \longmapsto \mathbb{R}$ that is second-order differentiable with respect to each variable x_i in all or the same part of Ω^o, the *Hessian* of f, denoted as $\nabla^2 f$, is a matrix-valued function whose components are the second-order partial derivatives of f:

$$\nabla^2 f = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}\right)_{1 \le i, j \le n}. \tag{2.20}$$

Similarly, when the function f has several parameters besides the variables in $\mathbf{x}$, we denote the Hessian of f by $\nabla^2_{\mathbf{x}} f$ or $\frac{\partial^2 f}{\partial \mathbf{x}^2}$ instead.

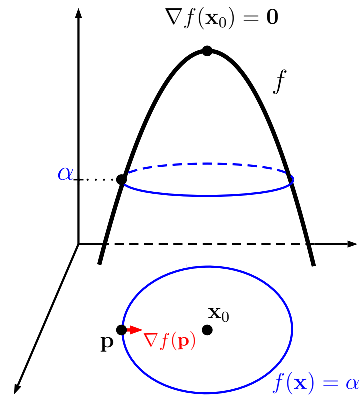

Fig. 2.8 The gradient vector at the point $\mathbf{p}$ is the direction of fastest increase of the function f from the point. The other labeled point, $\mathbf{x}_0$, is a critical point of f where the gradient vector vanishes.

Given a point $\mathbf{p} \in \Omega^o$ at which f is twice differentiable, we may evaluate the components of the Hessian at $\mathbf{p}$ to obtain the *Hessian matrix* of f at $\mathbf{p}$:

$$\nabla^2 f(\mathbf{p}) = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{p}) \right) \in \mathbb{R}^{n \times n}. \tag{2.21}$$

If all the second-order partial derivatives of the function f are continuous at $\mathbf{p}$, then

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{p}) = \frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{p}), \quad \text{for all } i \neq j \tag{2.22}$$

In this case, $\nabla^2 f(\mathbf{p})$ is a symmetric matrix.

Example 2.6 Let $f(x_1, x_2) = x_1^2 + 2x_2^2 + 2x_1 x_2$ which is a function on $\mathbb{R}^2$. The gradient of f is

$$\nabla f = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{pmatrix} = \begin{pmatrix} 2x_1 + 2x_2 \\ 4x_2 + 2x_1 \end{pmatrix}.$$

The gradient vector at $\mathbf{x} = \mathbf{1}$ is

$$\nabla f(\mathbf{1}) = \begin{pmatrix} 2x_1 + 2x_2 \\ 4x_2 + 2x_1 \end{pmatrix}\Bigg|_{x_1 = x_2 = 1} = \begin{pmatrix} 4 \\ 6 \end{pmatrix}.$$

The Hessian of f is constant everywhere in $\mathbb{R}^2$

$$\nabla^2 f = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} \end{pmatrix} = \begin{pmatrix} 2 & 2 \\ 2 & 4 \end{pmatrix}.$$

Lastly, we mention the notion of the Jacobian of a vector-valued function (such as the gradient function). Let $\mathbf{f} : \Omega \subseteq \mathbb{R}^n \mapsto \mathbb{R}^m$ be a vector-valued function with differentiable component functions f_i, i.e.,

$$\mathbf{f}(\mathbf{x}) = \begin{pmatrix} f_1(\mathbf{x}) \\ \vdots \\ f_m(\mathbf{x}) \end{pmatrix}, \quad \text{for all } \mathbf{x} \in \Omega. \tag{2.23}$$

The Jacobian of **f** is a matrix-valued function

$$\nabla\mathbf{f} : \Omega \subseteq \mathbb{R}^n \mapsto \mathbb{R}^{n\times m}, \qquad \text{with} \quad (\nabla\mathbf{f})_{ij} = \frac{\partial f_j}{\partial x_i} \tag{2.24}$$

Using this terminology, the Hessian of a scalar-valued function $f : \Omega \subseteq \mathbb{R}^n \mapsto \mathbb{R}$ is just the Jacobian of the gradient of f, namely,

$$\nabla^2 f = \nabla(\nabla f). \tag{2.25}$$

Given a point $\mathbf{p} \in \Omega$ at which all f_i's are differentiable, we can evaluate every component of the Jacobian $\nabla\mathbf{f}$ at $\mathbf{p}$ to obtain the Jacobian matrix of $\mathbf{f}$ at $\mathbf{p}$:

$$\nabla\mathbf{f}(\mathbf{p}) = \left(\frac{\partial f_j}{\partial x_i}(\mathbf{p})\right)_{1\le i\le n,\, 1\le j\le m}. \tag{2.26}$$

2.3.2 Useful formulas

Next, we derive a few formulas concerning the gradients and Hessians of functions of $\mathbf{x}$ like $\mathbf{x}^T\mathbf{A}\mathbf{x}$ and $\|\mathbf{x}\|^2$. These formulas are frequently needed in the derivation of statistics and machine learning algorithms.

Theorem 2.6 *For any fixed symmetric matrix* $\mathbf{A} \in S^n(\mathbb{R})$*, fixed matrix* $\mathbf{B} \in \mathbb{R}^{m\times n}$*, and fixed vector* $\mathbf{a} \in \mathbb{R}^n$*, we have*

$$\begin{aligned} \frac{\partial}{\partial\mathbf{x}}(\mathbf{a}^T\mathbf{x}) &= \mathbf{a} \\ \frac{\partial}{\partial\mathbf{x}}(\|\mathbf{x}\|^2) &= 2\mathbf{x} \\ \frac{\partial}{\partial\mathbf{x}}(\mathbf{x}^T\mathbf{A}\mathbf{x}) &= 2\mathbf{A}\mathbf{x} \\ \frac{\partial}{\partial\mathbf{x}}(\|\mathbf{B}\mathbf{x}\|^2) &= 2\mathbf{B}^T\mathbf{B}\mathbf{x} \end{aligned}$$

Proof The top two identities can be verified by direct calculation of the kth partial derivative: For each $1 \le k \le n$:

$$\frac{\partial}{\partial x_k}\left(\mathbf{a}^T\mathbf{x}\right) = \frac{\partial}{\partial x_k}\left(\sum_{i=1}^{n} a_i x_i\right) = a_k$$
$$\frac{\partial}{\partial x_k}\left(\|\mathbf{x}\|^2\right) = \frac{\partial}{\partial x_k}\left(\sum_{i=1}^{n} x_i^2\right) = 2x_k.$$

For the third identity involving $\mathbf{x}^T\mathbf{A}\mathbf{x}$,

$$\begin{aligned}
\frac{\partial}{\partial x_k}(\mathbf{x}^T\mathbf{A}\mathbf{x}) &= \frac{\partial}{\partial x_k}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i x_j\right) \\
&= \frac{\partial}{\partial x_k}\left(\sum_{j\neq k} a_{kj}x_k x_j + \sum_{i\neq k} a_{ik}x_i x_k + a_{kk}x_k^2\right) \\
&= \sum_{j\neq k} a_{kj}x_j + \sum_{i\neq k} a_{ik}x_i + 2a_{kk}x_k \\
&= \sum_{j=1}^{n} a_{kj}x_j + \sum_{i=1}^{n} x_i a_{ik} \\
&= A_k\mathbf{x} + \mathbf{x}^T\mathbf{a}_k \\
&= 2A_k\mathbf{x} \qquad \text{(since } \mathbf{A} \text{ is symmetric)}
\end{aligned}$$

Collectively, we have

$$\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^T\mathbf{A}\mathbf{x}) = \begin{bmatrix} \frac{\partial}{\partial x_1}(\mathbf{x}^T\mathbf{A}\mathbf{x}) \\ \vdots \\ \frac{\partial}{\partial x_n}(\mathbf{x}^T\mathbf{A}\mathbf{x}) \end{bmatrix} = \begin{bmatrix} 2A_1\mathbf{x} \\ \vdots \\ 2A_n\mathbf{x} \end{bmatrix} = 2\mathbf{A}\mathbf{x}$$

The last identity can then be verified by writing

$$\|\mathbf{B}\mathbf{x}\|^2 = (\mathbf{B}\mathbf{x})^T(\mathbf{B}\mathbf{x}) = \mathbf{x}^T(\mathbf{B}^T\mathbf{B})\mathbf{x}$$

and applying the third identity. □

The next theorem concerns the Hessians of functions like $\mathbf{x}^T\mathbf{A}\mathbf{x}$ and $\|\mathbf{x}\|^2$.

Theorem 2.7 *For any fixed symmetric matrix* $\mathbf{A} \in S^n(\mathbb{R})$*, fixed matrix* $\mathbf{B} \in \mathbb{R}^{m\times n}$ *and fixed vector* $\mathbf{a} \in \mathbb{R}^n$*, we have*

$$\frac{\partial^2}{\partial \mathbf{x}^2}(\mathbf{a}^T\mathbf{x}) = \mathbf{O}$$
$$\frac{\partial^2}{\partial \mathbf{x}^2}(\|\mathbf{x}\|^2) = 2\mathbf{I}$$
$$\frac{\partial^2}{\partial \mathbf{x}^2}(\mathbf{x}^T\mathbf{A}\mathbf{x}) = 2\mathbf{A}$$
$$\frac{\partial^2}{\partial \mathbf{x}^2}(\|\mathbf{B}\mathbf{x}\|^2) = 2\mathbf{B}^T\mathbf{B}$$

Proof The first three results can be proved by using the corresponding partial derivatives derived in the proof of the preceding theorem: For any $1 \le k, \ell \le n$,

$$\frac{\partial}{\partial x_\ell}\left[\frac{\partial}{\partial x_k}\left(\mathbf{a}^T\mathbf{x}\right)\right] = \frac{\partial}{\partial x_\ell}(a_k) = 0 \qquad \Longrightarrow \qquad \frac{\partial^2}{\partial \mathbf{x}^2}(\mathbf{a}^T\mathbf{x}) = \mathbf{O}$$
$$\frac{\partial}{\partial x_\ell}\left[\frac{\partial}{\partial x_k}\left(\|\mathbf{x}\|^2\right)\right] = \frac{\partial}{\partial x_\ell}(2x_k) = 2\delta_{k\ell} \qquad \Longrightarrow \qquad \frac{\partial^2}{\partial \mathbf{x}^2}(\|\mathbf{x}\|^2) = 2\mathbf{I}$$
$$\frac{\partial}{\partial x_\ell}\left[\frac{\partial}{\partial x_k}(\mathbf{x}^T\mathbf{A}\mathbf{x})\right] = \frac{\partial}{\partial x_\ell}(2A_k\mathbf{x}) = 2a_{k\ell} \qquad \Longrightarrow \qquad \frac{\partial^2}{\partial \mathbf{x}^2}(\mathbf{x}^T\mathbf{A}\mathbf{x}) = 2\mathbf{A}$$

The last result can be proved similarly by writing $\|\mathbf{B}\mathbf{x}\|^2 = \mathbf{x}^T(\mathbf{B}^T\mathbf{B})\mathbf{x}$ and applying the third result. □

Lastly, we introduce the gradient of a function $f : \mathbb{R}^{m\times n} \mapsto \mathbb{R}$ whose inputs are matrices (regarded as vectors). That is, for any $\mathbf{X} \in \mathbb{R}^{m\times n}$, we think of $f(\mathbf{X})$ as $f(\text{vec}(\mathbf{X}))$ and compute the gradient of f with respect to each component of vec($\mathbf{X}$), but we will represent the gradient in a matrix form consistent in size and order with $\mathbf{X}$.

Formally, the gradient of a function $f : \mathbf{X} \in \mathbb{R}^{m\times n} \mapsto f(\mathbf{X}) \in \mathbb{R}$ is a matrix-valued function $\nabla f : \mathbb{R}^{m\times n} \mapsto \mathbb{R}^{m\times n}$, whose components are the partial derivatives of f with respect to each input variable x_{ij}:

$$\nabla f = \left(\frac{\partial f}{\partial x_{ij}}\right)_{1\le i\le m,\, 1\le j\le n}. \tag{2.27}$$

Other kinds of notation for the gradient are $\nabla_{\mathbf{X}} f$ and $\frac{\partial f}{\partial \mathbf{X}}$.

For example, let

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix}, \qquad \text{and} \quad f(\mathbf{X}) = x_{11}^2 + x_{22} - x_{12}x_{21}.$$

Then

$$\frac{\partial f}{\partial \mathbf{X}} = \begin{pmatrix} \frac{\partial f}{\partial x_{11}} & \frac{\partial f}{\partial x_{12}} \\ \frac{\partial f}{\partial x_{21}} & \frac{\partial f}{\partial x_{22}} \end{pmatrix} = \begin{pmatrix} 2x_{11} & x_{21} \\ x_{12} & 1 \end{pmatrix}.$$

We are now ready to present the following results.

Theorem 2.8 *Let* $\mathbf{X} \in \mathbb{R}^{m\times n}$ *be a matrix of mn variables. For any fixed matrix* $\mathbf{A} \in \mathbb{R}^{m\times n}$*, and fixed symmetric matrix* $\mathbf{B} \in \mathbb{R}^{m\times m}$*, we have*

$$\frac{\partial}{\partial \mathbf{X}} \operatorname{tr}(\mathbf{A}^T\mathbf{X}) = \mathbf{A}, \qquad \text{and} \qquad \frac{\partial}{\partial \mathbf{X}} \operatorname{tr}(\mathbf{X}^T\mathbf{B}\mathbf{X}) = 2\mathbf{B}\mathbf{X} \tag{2.28}$$

In particular, if $\mathbf{B} = \mathbf{I}$*, then the second identity reduces to*

$$\frac{\partial}{\partial \mathbf{X}} \operatorname{tr}(\mathbf{X}^T\mathbf{X}) = 2\mathbf{X} \tag{2.29}$$

Proof We just need to verify the entrywise partial derivatives. For any integers $1 \le i \le m$ and $1 \le j \le n$,

$$\frac{\partial}{\partial x_{ij}} \operatorname{tr}(\mathbf{A}^T\mathbf{X}) = \frac{\partial}{\partial x_{ij}} \sum_{k=1}^{m} (\mathbf{A}^T\mathbf{X})_{kk} = \frac{\partial}{\partial x_{ij}} \sum_{k=1}^{m} \left(\sum_{s=1}^{n} a_{sk} x_{sk} \right) = \frac{\partial}{\partial x_{ij}} (a_{ij} x_{ij}) = a_{ij}.$$

The completes the proof of the first identity.

For any integers $1 \le i \le m$ and $1 \le j \le n$,

$$\frac{\partial}{\partial x_{ij}} \operatorname{tr}(\mathbf{X}^T\mathbf{B}\mathbf{X}) = \frac{\partial}{\partial x_{ij}} \sum_{k=1}^{n} (\mathbf{X}^T\mathbf{B}\mathbf{X})_{kk} = \frac{\partial}{\partial x_{ij}} \sum_{k=1}^{n} \left(\sum_{s=1}^{m} \sum_{t=1}^{m} x_{sk} b_{st} x_{tk} \right).$$

Observe that the term-by-term partial derivatives are zero unless $k = j$. Thus,

$$\frac{\partial}{\partial x_{ij}} \operatorname{tr}(\mathbf{X}^T\mathbf{B}\mathbf{X}) = \frac{\partial}{\partial x_{ij}} \sum_{s=1}^{m} \sum_{t=1}^{m} x_{sj} b_{st} x_{tj}.$$

To proceed further, we divide the double sum into four terms depending on whether each of s, t is equal to i:

$$\begin{aligned}
\frac{\partial}{\partial x_{ij}} \operatorname{tr}(\mathbf{X}^T\mathbf{B}\mathbf{X}) &= \frac{\partial}{\partial x_{ij}} \left(b_{ii} x_{ij}^2 + \sum_{t=1, t\neq i}^{m} x_{ij} b_{it} x_{tj} + \sum_{s=1, s\neq i}^{m} x_{sj} b_{si} x_{ij} + \sum_{s\neq i} \sum_{t\neq i} x_{sj} b_{st} x_{tj} \right) \\
&= 2 b_{ii} x_{ij} + \sum_{t=1, t\neq i}^{m} b_{it} x_{tj} + \sum_{s=1, s\neq i}^{m} x_{sj} b_{si} + 0 \\
&= \sum_{t=1}^{m} b_{it} x_{tj} + \sum_{s=1}^{m} x_{sj} b_{si} \\
&= (\mathbf{B}\mathbf{X})_{ij} + (\mathbf{B}^T\mathbf{X})_{ij} \\
&= 2(\mathbf{B}\mathbf{X})_{ij} \quad \text{(by using the symmetry of } \mathbf{B}\text{)}.
\end{aligned}$$

This thus completes the proof of the second identity. □