Chapter 17

Appendix

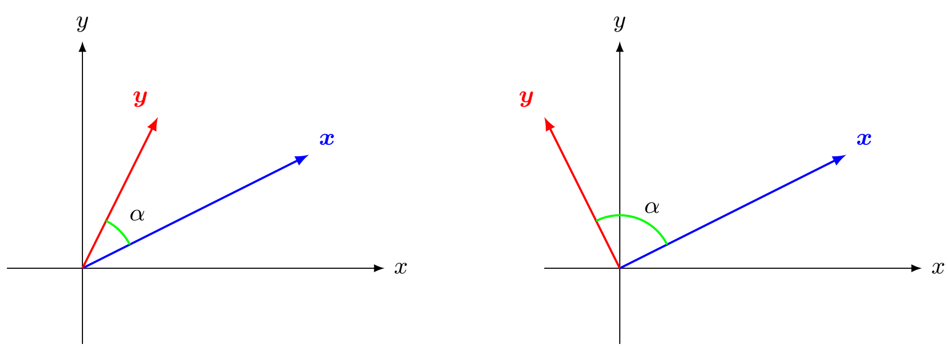
17.1 Review of Linear Algebra

For two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, the **inner product** $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ of $\boldsymbol{x}$ and $\boldsymbol{y}$ is

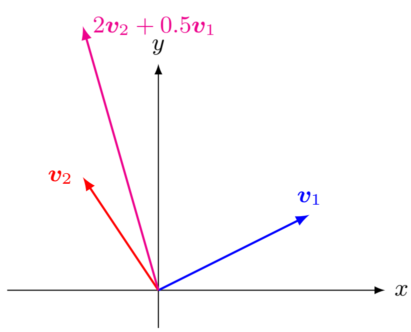
$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad \boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^T \boldsymbol{y} = \sum_{i=1}^{n} x_i y_i. \tag{17.1}$$

where $\boldsymbol{x}^T$ is the transpose of $\boldsymbol{x}$.

The **length** or the ℓ_2 norm of a vector $\boldsymbol{x}$ is $\|\boldsymbol{x}\| = \|\boldsymbol{x}\|_2 = \sqrt{\boldsymbol{x}^T \boldsymbol{x}}$ and we have $\|\boldsymbol{x}\|_2^2 = \boldsymbol{x}^T \boldsymbol{x}$. Let α be the angle between $\boldsymbol{x}$ and $\boldsymbol{y}$. Then $\boldsymbol{x}^T \boldsymbol{y} = \|\boldsymbol{x}\| \|\boldsymbol{y}\| \cos \alpha$. If $\boldsymbol{x}^T \boldsymbol{y} = 0$, then the two are called **orthogonal**.



For a collection of vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m$, a **linear combination** of these is any vector of the form $a_1 \boldsymbol{v}_1 + \cdots + a_m \boldsymbol{v}_m$, $a_i \in \mathbb{R}$. The set of all linear combinations of $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m$ is their **span** and denoted as $\text{Span}\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m\}$. This is a **subspace** (think line, plane, or the whole space). For a matrix A, the span of the columns of A is the **column space** of A.

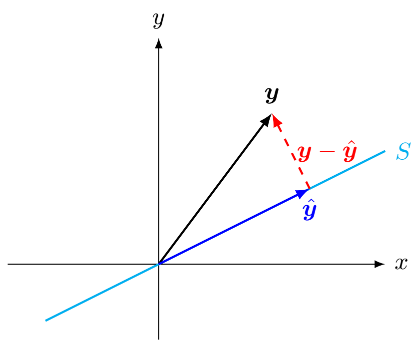


The vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m$ are **linearly independent** if there is no vector among them that can be written as a linear combination of the others, and linearly dependent otherwise. The vectors are linearly independent if and only if the only values for $a_1, \ldots, a_m$ satisfying $a_1\boldsymbol{v}_1 + \cdots + a_m\boldsymbol{v}_m = \mathbf{0}$ are $a_1, \ldots, a_m = 0$. In particular, the columns of a matrix A are linearly independent if and only if the only vector $\boldsymbol{a}$ satisfying $\mathsf{A}\boldsymbol{a} = \mathbf{0}$ is $\boldsymbol{a} = \mathbf{0}$.

The **inverse** of a square matrix A is a matrix A^{-1} such that $\mathsf{A}\mathsf{A}^{-1} = \mathsf{A}^{-1}\mathsf{A} = I$, where I is the **identity matrix**, which has 1s on the diagonal and 0s elsewhere. A matrix that has an inverse is called **invertible**. For a square matrix A, the following are equivalent:

- It is invertible.
- For all distinct vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, we have $\mathsf{A}\boldsymbol{a} \neq \mathsf{A}\boldsymbol{b}$.
- The only solution to $\mathsf{A}\boldsymbol{x} = \mathbf{0}$ is $\boldsymbol{x} = \mathbf{0}$.
- Its columns are linearly independent.
- Its determinant $|\mathsf{A}|$ is nonzero. (We also have $|\mathsf{A}^{-1}| = \frac{1}{|\mathsf{A}|}$.)

Given a subspace S (e.g., a plane or the column space of a matrix) and a vector $\boldsymbol{y}$, let $\hat{\boldsymbol{y}}$ be the vector in the subspace that is closest to $\boldsymbol{y}$. That is, we find $\hat{\boldsymbol{y}} \in S$ such that $\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|$ is minimized. Then $\hat{\boldsymbol{y}}$ is called the **projection** of $\boldsymbol{y}$ onto the subspace S.



Lemma 17.1 (Projection Lemma). *Let $\hat{\boldsymbol{y}}$ be the projection of a vector $\boldsymbol{y}$ onto a subspace S. Then $\boldsymbol{y} - \hat{\boldsymbol{y}}$ is orthogonal to every vector in S.*

Proof. Suppose that this is not the case. Then there is a nonzero vector $\boldsymbol{v} \in S$ such that $(\boldsymbol{y} - \hat{\boldsymbol{y}})^T\boldsymbol{v} \neq 0$. We will show that this contradicts the minimality of $\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|$. For any $a \in \mathbb{R}$,

$$\|\boldsymbol{y} - \hat{\boldsymbol{y}} - a\boldsymbol{v}\|_2^2 = (\boldsymbol{y} - \hat{\boldsymbol{y}} - a\boldsymbol{v})^T(\boldsymbol{y} - \hat{\boldsymbol{y}} - a\boldsymbol{v}) \tag{17.2}$$

$$= \|\boldsymbol{y} - \hat{\boldsymbol{y}}\|_2^2 - 2a\boldsymbol{v}^T(\boldsymbol{y} - \hat{\boldsymbol{y}}) + a^2\|\boldsymbol{v}\|_2^2. \tag{17.3}$$

This is a convex function in a. So setting the derivative to 0 gives the value of a that minimizes the error:

$$\frac{\partial}{\partial a}\|\boldsymbol{y} - \hat{\boldsymbol{y}} - a\boldsymbol{v}\|_2^2 = -2\boldsymbol{v}^T(\boldsymbol{y} - \hat{\boldsymbol{y}}) + 2a\|\boldsymbol{v}\|_2^2 = 0 \Rightarrow a = \frac{\boldsymbol{v}^T(\boldsymbol{y} - \hat{\boldsymbol{y}})}{\|\boldsymbol{v}\|_2^2} \neq 0. \tag{17.4}$$

Let

$$\hat{\boldsymbol{y}}' = \hat{\boldsymbol{y}} + \frac{\boldsymbol{v}^T(\boldsymbol{y} - \hat{\boldsymbol{y}})}{\boldsymbol{v}^T\boldsymbol{v}}\boldsymbol{v}, \tag{17.5}$$

and note that $\hat{\boldsymbol{y}}'$ is also in S but it is closer to $\boldsymbol{y}$, contradicting the optimality of $\hat{\boldsymbol{y}}$. □

17.2 Vector and matrix differentiation

Definition 17.2 (The three derivatives)**.** For a matrix A, scalar z, and two vectors $\boldsymbol{x}, \boldsymbol{y}$ (possibly one-dimensional), let

$$\frac{d\mathsf{A}}{dz} = \begin{pmatrix} \frac{\partial \mathsf{A}_{11}}{\partial z} & \cdots & \frac{\partial \mathsf{A}_{1n}}{\partial z} \\ \vdots & \ddots & \vdots \\ \frac{\partial \mathsf{A}_{m1}}{\partial z} & \cdots & \frac{\partial \mathsf{A}_{mn}}{\partial z} \end{pmatrix}, \qquad \frac{dz}{d\mathsf{A}} = \begin{pmatrix} \frac{\partial z}{\partial \mathsf{A}_{11}} & \cdots & \frac{\partial z}{\partial \mathsf{A}_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial z}{\partial \mathsf{A}_{m1}} & \cdots & \frac{\partial z}{\partial \mathsf{A}_{mn}} \end{pmatrix}, \qquad \frac{d\boldsymbol{y}}{d\boldsymbol{x}} = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_n}{\partial x_1} & \cdots & \frac{\partial y_n}{\partial x_m} \end{pmatrix}$$

Lemma 17.3. *For a scalar a, vectors $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{v}$, and constant matrices A and S,*

$$\begin{aligned} \frac{d\boldsymbol{y}}{d\boldsymbol{v}} &= \frac{d\boldsymbol{y}}{d\boldsymbol{x}}\frac{d\boldsymbol{x}}{d\boldsymbol{v}}, \\ \frac{d}{d\boldsymbol{v}}(a\boldsymbol{x}) &= a\frac{d\boldsymbol{x}}{d\boldsymbol{v}} + \boldsymbol{x}\frac{da}{d\boldsymbol{v}}, \\ \frac{d}{d\boldsymbol{v}}(\boldsymbol{y}^T\mathsf{A}\boldsymbol{x}) &= \boldsymbol{y}^T\mathsf{A}\frac{d\boldsymbol{x}}{d\boldsymbol{v}} + \boldsymbol{x}^T\mathsf{A}^T\frac{d\boldsymbol{y}}{d\boldsymbol{v}}, \\ \frac{d}{d\boldsymbol{v}}(\boldsymbol{y}^T\mathsf{S}\boldsymbol{y}) &= 2\boldsymbol{y}^T\mathsf{S}\frac{d\boldsymbol{y}}{d\boldsymbol{v}}, \qquad (\mathsf{S} \textit{ is symmetric}) \\ \frac{d}{d\boldsymbol{v}}(\mathsf{A}\boldsymbol{x}) &= \mathsf{A}\frac{d\boldsymbol{x}}{d\boldsymbol{v}}. \end{aligned}$$

Lemma 17.4. *For matrix A and constant vector $\boldsymbol{x}$,*

$$\begin{aligned} \frac{d}{d\mathsf{A}}(\boldsymbol{x}^T\mathsf{A}\boldsymbol{x}) &= \boldsymbol{x}\boldsymbol{x}^T \\ \frac{d}{d\mathsf{A}}\ln|\mathsf{A}| &= \mathsf{A}^{-T} \end{aligned}$$

Definition 17.5. Let $f : \mathbb{R}^m \to \mathbb{R}$. The gradient of $f(\boldsymbol{x})$ with respect to $\boldsymbol{x}$ is defined as

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = \left(\frac{df(\boldsymbol{x})}{d\boldsymbol{x}}\right)^T = \begin{pmatrix} \frac{\partial f(\boldsymbol{x})}{\partial x_1} \\ \vdots \\ \frac{\partial f(\boldsymbol{x})}{\partial x_m} \end{pmatrix}$$

and the Hessian of $f(\boldsymbol{x})$ with respect to $\boldsymbol{x}$ is defined as

$$\mathsf{H}_{\boldsymbol{x}}(f(\boldsymbol{x})) = \frac{d\nabla_{\boldsymbol{x}} f(\boldsymbol{x})}{d\boldsymbol{x}} = \begin{pmatrix} \frac{\partial f(\boldsymbol{x})}{\partial x_1 \partial x_1} & \cdots & \frac{\partial f(\boldsymbol{x})}{\partial x_m \partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial f(\boldsymbol{x})}{\partial x_1 \partial x_m} & \cdots & \frac{\partial f(\boldsymbol{x})}{\partial x_m \partial x_m} \end{pmatrix}$$

Chain rule. Consider $h : \mathbb{R}^m \to \mathbb{R}$, $g : \mathbb{R} \to \mathbb{R}$, and $f(\boldsymbol{x}) = g(h(\boldsymbol{x}))$. From Lemma 17.3,

$$\begin{aligned} \nabla f(\boldsymbol{x}) &= g'(h(\boldsymbol{x}))\nabla h(\boldsymbol{x}), \\ \mathsf{H} f(\boldsymbol{x}) &= g'(h(\boldsymbol{x}))\mathsf{H} h(\boldsymbol{x}) + g''(h(\boldsymbol{x}))\nabla h(\boldsymbol{x})\nabla^T h(\boldsymbol{x}) \end{aligned}$$

since

$$\begin{aligned} \mathsf{H} f(\boldsymbol{x}) &= \frac{d\nabla f}{d\boldsymbol{x}} \\ &= \frac{d(g'(h(\boldsymbol{x}))\nabla h(\boldsymbol{x}))}{d\boldsymbol{x}} \\ &= g'(h(\boldsymbol{x}))\frac{d\nabla h(\boldsymbol{x})}{d\boldsymbol{x}} + \nabla h(\boldsymbol{x})\frac{d(g'(h(\boldsymbol{x})))}{d\boldsymbol{x}} \\ &= g'(h(\boldsymbol{x}))\mathsf{H} h(\boldsymbol{x}) + \nabla h(\boldsymbol{x})\nabla^T h(\boldsymbol{x}) g''(h(\boldsymbol{x})) \end{aligned}$$

Example 17.6. Let us find the derivatives of $f(\boldsymbol{x}) = \log\sum_{i=1}^m e^{x_i}$. Let $\boldsymbol{z} = (\exp(x_i))_{i=1}^m$ so that $f(\boldsymbol{x}) = \log \mathbf{1}^T\boldsymbol{z}$.

$$\nabla f(\boldsymbol{x}) = \frac{\boldsymbol{z}}{\mathbf{1}^T\boldsymbol{z}},$$
$$\mathsf{H}f(\boldsymbol{x}) = \frac{\operatorname{diag}(\boldsymbol{z})}{\mathbf{1}^T\boldsymbol{z}} - \frac{\boldsymbol{z}\boldsymbol{z}^T}{(\mathbf{1}^T\boldsymbol{z})^2}.$$

△

Chain rule. Let $\boldsymbol{h} = (h_1, \dots, h_n) : \mathbb{R}^m \to \mathbb{R}^n$, $g : \mathbb{R}^n \to \mathbb{R}$, and $f(\boldsymbol{x}) = g(\boldsymbol{h}(\boldsymbol{x}))$. Then

$$\frac{\partial f}{\partial x_i} = \sum_{j=1}^n \frac{\partial g}{\partial h_j}\frac{\partial h_j}{\partial x_i} = \frac{dg}{d\boldsymbol{h}} \cdot \frac{d\boldsymbol{h}}{dx_i} = \nabla^T g \cdot \frac{d\boldsymbol{h}}{dx_i},$$
$$\frac{df}{d\boldsymbol{x}} = \frac{dg}{d\boldsymbol{h}}\frac{d\boldsymbol{h}}{d\boldsymbol{x}} = \nabla^T g \frac{d\boldsymbol{h}}{d\boldsymbol{x}}, \qquad \nabla_{\boldsymbol{x}} f = \left(\frac{df}{d\boldsymbol{x}}\right)^T = \left(\frac{d\boldsymbol{h}}{d\boldsymbol{x}}\right)^T \nabla g$$