Chapter 16

Appendix

16.1 Vector and matrix differentiation

Definition 16.1 (The three derivatives)**.** For a matrix A, scalar z, and two vectors $\boldsymbol{x}, \boldsymbol{y}$ (possibly one-dimensional), let

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

Lemma 16.2. *For a scalar a, vectors $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{v}$, and constant matrices A and S,*

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

Lemma 16.3. *For matrix A and constant vector $\boldsymbol{x}$,*

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

Definition 16.4. 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 16.2,

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

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 16.5. 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 $f(\boldsymbol{x}) = \log \mathbf{1}^T \boldsymbol{z}$.

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

△

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

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

16.2 Properties of Expectation, Correlation, and Covariance for Vectors

Elementary properties of expectation, correlation, and covariance for vectors follow immediately from similar properties for ordinary scalar random variables. These properties include the following (here A and C are nonrandom matrices and b and d are nonrandom vectors).

1. $E[AX + b] = AE[X] + b$
2. $\text{Cov}(X, Y) = E[X(Y - E[Y])^T] = E[(X - E[X])Y^T] = E[XY^T] - (E[X])(E[Y])^T$
3. $E[(AX)(CY)^T] = AE[XY^T]C^T$
4. $\text{Cov}(AX + b, CY + d) = A\text{Cov}(X, Y)C^T$
5. $\text{Cov}(AX + b) = A\text{Cov}(X)A^T$
6. $\text{Cov}(W + X, Y + Z) = \text{Cov}(W, Y) + \text{Cov}(W, Z) + \text{Cov}(X, Y) + \text{Cov}(X, Z)$.