Matrix and Vector Differentiation

Rolf D. Svegstrup

August 7, 2019

1 Vector Differentiation

Definition 1.1. *Let $y = \psi(x) \in \mathbb{R}^m$ where x ranges over R^n. Then $\frac{\partial y}{\partial x}$ is an $m \times n$-matrix given by*

$$\frac{\partial y}{\partial x} = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial y_n}{\partial x_1} & \cdots & \frac{\partial y_n}{\partial x_m} \end{pmatrix} \tag{1}$$

The matrix is called the Jacobian *of the transformation ψ.*

Remark 1.2. *Note that if vectors $z = z(y)$ and $y = y(x)$, then $\frac{\partial z}{\partial x} = \frac{\partial z}{\partial y}\frac{\partial y}{\partial x}$. Also note that $\frac{\partial x}{\partial x} = I$.*

Notation 1.3. *Throughout this section it will be assumed that x is a vector variable, $y = y(x)$ and $z = y(x)$ are vectors depending on x and that A is a matrix not depending on x. Dimensions of x, y, z and A are such that the equations make sense.*

Proposition 1.4.

$$\frac{\partial Ay}{\partial x} = A\frac{\partial y}{\partial x} \tag{2}$$

Proof.

$$\left(\frac{\partial Ay}{\partial x}\right)_{ij} = \frac{\partial(\sum_k A_{ik} y_k)}{\partial x_j} = \sum_k A_{ik}\left(\frac{\partial y}{\partial x}\right)_{kj} = \left(A\frac{\partial y}{\partial x}\right)_{ij}.$$

□

Proposition 1.5.

$$\frac{\partial y^t A z}{\partial x} = z^t A^t \frac{\partial y}{\partial x} + y^t A \frac{\partial z}{\partial x}. \tag{3}$$

Proof.

$$\begin{aligned}
\left(\frac{\partial y^t A z}{\partial x}\right)_{1j} &= \frac{\partial \sum_{k,l} y_k A_{kl} z_l}{\partial x_j} \\
&= \sum_k \frac{\partial y_k}{\partial x_j}(Az)_k + \sum_l (y^t A)_{1l} \frac{\partial z_l}{\partial x_j} \\
&= \sum_k \left(\frac{\partial y}{\partial x}\right)_{kj} (Az)_k + \sum_l (y^t A)_{1l} \left(\frac{\partial z}{\partial x}\right)_{lj} \\
&= \left(\left(\left(\frac{\partial y}{\partial x}\right)^t Az\right)^t + y^t A \frac{\partial z}{\partial x}\right)_{1j}
\end{aligned}$$

□

Corollary 1.6. *Some simple corollaries follow:*

1. $\frac{\partial Ax}{\partial x} = A$.
2. $\frac{\partial x^t A x}{\partial x} = x^t(A + A^t)$.
3. $\frac{\partial x^t x}{\partial x} = 2x^t$.
4. $\frac{\partial y^t z}{\partial x} = z^t \frac{\partial y}{\partial x} + y^t \frac{\partial z}{\partial x}$.

2 Differentiating Matrices by Scalars

Definition 2.1. *Let A be an $m \times n$ matrix whose elements are functions of the scalar α. Then*

$$\frac{\partial A}{\partial \alpha} = \begin{pmatrix} \frac{\partial A_{11}}{\partial \alpha} & \cdots & \frac{\partial A_{1n}}{\partial \alpha} \\ \vdots & & \vdots \\ \frac{\partial A_{m1}}{\partial \alpha} & \cdots & \frac{\partial A_{mn}}{\partial \alpha} \end{pmatrix}. \tag{4}$$

Lemma 2.2.

$$\frac{\partial AB}{\partial \alpha} = \frac{\partial A}{\partial \alpha} B + A \frac{\partial B}{\partial \alpha}. \tag{5}$$

Proposition 2.3. *Let A be an invertible matrix. Then*

$$\frac{\partial A^{-1}}{\partial \alpha} = -A^{-1} \frac{\partial A}{\partial \alpha} A^{-1}. \tag{6}$$

Proof.

$$0 = \frac{\partial A^{-1} A}{\partial \alpha} = \frac{\partial A^{-1}}{\partial \alpha} A + A^{-1} \frac{\partial A}{\partial \alpha}.$$

Isolating $\frac{\partial A^{-1}}{\partial \alpha}$ completes the proof. □

3 Applications

3.1 Linear Regression Coefficients

Let there be given N observations (x_i, y_i) where $x_i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}$. Let $y = (y_i) \in \mathbb{R}^N$. Let $X \in \mathbb{R}^{N\times(n+1)}$ be the matrix in which the i-th row is 1 followed by x_i as a row vector. We seek a coefficient vector $\beta \in \mathbb{R}^{n+1}$ such that $||X\beta - y||_2$ is minimized.

We find the β minimizing the square of the norm using vector differentiation:

$$0 = \frac{\partial}{\partial\beta}(X\beta - y)^t(X\beta - y) = 2(X\beta - y)^t X. \tag{7}$$

Hence, $\beta^t X^t X = y^t X$. Assuming that $X^t X$ is invertible (columns of X being linearly independent), we find

$$\beta = (X^t X)^{-1} X^t y. \tag{8}$$

3.2 Regularised Linear Regression Coefficients

Using L^2-regularisation, we seek to minimize $||X\beta - y||_2^2 + \lambda||\beta||_2^2$.

$$0 = \frac{\partial}{\partial\beta}\big((X\beta - y)^t(X\beta - y) + \lambda\beta^t\beta\big) = 2(X\beta - y)^t X + 2\lambda\beta^t.$$

The matrix $(X^X + \lambda I)$ is invertible if $\lambda > 0$ and so we find:

$$\beta = (X^t X + \lambda I)^{-1} X^t y.$$