Background Mathematics[1]

(2013 - I)

Salvador Ruiz Correa

Centro de Investigación en Matemáticas (CIMAT)

[1] These slides are *adapted* from those that accompany the book *Bayesian Reasoning and Machine Learning*. The book and demos can be downloaded from `www.cs.ucl.ac.uk/staff/D.Barber/brml`. We acknowledge David Barber for providing the original slides.

Matrices

An $m \times n$ matrix $\mathbf{A}$ is a collection of scalar values arranged in a rectangle of m rows and n columns.

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$$

The i, j element of matrix $\mathbf{A}$ can be written A_{ij} or more conventionally a_{ij}. Where more clarity is required, one may write $[\mathbf{A}]_{ij}$ (for example $\left[\mathbf{A}^{-1}\right]_{ij}$).

Matrix addition

For two matrix $\mathbf{A}$ and $\mathbf{B}$ of the same size,

$$[\mathbf{A} + \mathbf{B}]_{ij} = [\mathbf{A}]_{ij} + [\mathbf{B}]_{ij}$$

Matrix multiplication

For an l by n matrix $\mathbf{A}$ and an n by m matrix B, the product $\mathbf{AB}$ is the l by m matrix with elements

$$[\mathbf{AB}]_{ik} = \sum_{j=1}^{n} [\mathbf{A}]_{ij} [\mathbf{B}]_{jk} \; ; \qquad i = 1, \ldots, l \quad k = 1, \ldots, m \, .$$

In general $\mathbf{BA} \neq \mathbf{AB}$. When $\mathbf{BA} = \mathbf{AB}$ we say they $\mathbf{A}$ and $\mathbf{B}$ commute.

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{pmatrix}$$

$$= \begin{pmatrix} a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} & a_{11}b_{12} + a_{12}b_{22} + a_{13}b_{32} \\ a_{21}b_{11} + a_{22}b_{21} + a_{23}b_{31} & a_{21}b_{12} + a_{22}b_{22} + a_{23}b_{32} \\ a_{31}b_{11} + a_{32}b_{21} + a_{33}b_{31} & a_{31}b_{12} + a_{32}b_{22} + a_{33}b_{32} \end{pmatrix}$$

Identity

The matrix $\mathbf{I}$ is the identity matrix, necessarily square, with 1's on the diagonal and 0's everywhere else. For clarity we may also write $\mathbf{I}_m$ for a square $m \times m$ identity matrix. Then for an $m \times n$ matrix $\mathbf{A}$, $\mathbf{I}_m\mathbf{A} = \mathbf{A}\mathbf{I}_n = \mathbf{A}$. The identity matrix has elements $[\mathbf{I}]_{ij} = \delta_{ij}$ given by the Kronecker delta:

$$\delta_{ij} \equiv \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

$$\mathbf{I} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

Transpose

The transpose $\mathbf{B}^{\mathsf{T}}$ of the n by m matrix $\mathbf{B}$ is the m by n matrix D with components

$$\left[\mathbf{B}^{\mathsf{T}}\right]_{kj} = \mathbf{B}_{jk}\,; \qquad k = 1, \ldots, m \quad j = 1, \ldots, n\,.$$

$\left(\mathbf{B}^{\mathsf{T}}\right)^{\mathsf{T}} = \mathbf{B}$ and $(\mathbf{AB})^{\mathsf{T}} = \mathbf{B}^{\mathsf{T}}\mathbf{A}^{\mathsf{T}}$. If the shapes of the matrices $\mathbf{A}$,$\mathbf{B}$ and $\mathbf{C}$ are such that it makes sense to calculate the product $\mathbf{ABC}$, then

$$(\mathbf{ABC})^{\mathsf{T}} = \mathbf{C}^{\mathsf{T}}\mathbf{B}^{\mathsf{T}}\mathbf{A}^{\mathsf{T}}$$

Vector algebra

Vectors

Let $\mathbf{x}$ denote the n-dimensional column vector with components

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

A vector can be considered a $n \times 1$ matrix.

Addition

$$\mathbf{x} + \mathbf{y} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_n + y_n \end{pmatrix}$$

Scalar product

$$\mathbf{w} \cdot \mathbf{x} = \sum_{i=1}^{n} w_i x_i = \mathbf{w}^\mathsf{T} \mathbf{x}$$

The length of a vector is denoted $|\mathbf{x}|$, the squared length is given by

$$|\mathbf{x}|^2 = \mathbf{x}^\mathsf{T} \mathbf{x} = \mathbf{x}^2 = x_1^2 + x_2^2 + \cdots + x_n^2$$

A unit vector $\mathbf{x}$ has $\mathbf{x}^\mathsf{T} \mathbf{x} = 1$. The scalar product has a natural geometric interpretation as:

$$\mathbf{w} \cdot \mathbf{x} = |\mathbf{w}| \, |\mathbf{x}| \cos(\theta)$$

where θ is the angle between the two vectors. Thus if the lengths of two vectors are fixed their inner product is largest when $\theta = 0$, whereupon one vector is a constant multiple of the other. If the scalar product $\mathbf{x}^\mathsf{T} \mathbf{y} = 0$, then $\mathbf{x}$ and $\mathbf{y}$ are orthogonal.

Linear dependence

A set of vectors $\mathbf{x}^1, \ldots, \mathbf{x}^n$ is linearly dependent if there exists a vector $\mathbf{x}^j$ that can be expressed as a linear combination of the other vectors. If the only solution to

$$\sum_{i=1}^{n} \alpha_i \mathbf{x}^i = \mathbf{0}$$

is for all $\alpha_i = 0, i = 1, \ldots, n$, the vectors $\mathbf{x}^1, \ldots, \mathbf{x}^n$ are linearly independent.

Projections

Suppose we wish to resolve the vector $\mathbf{a}$ into its components along the orthogonal directions specified by the unit vectors $\mathbf{e}$ and $\mathbf{e}^*$. That is $|\boldsymbol{e}| = |\boldsymbol{e}|^* = 1$ and $\mathbf{e} \cdot \mathbf{e^*} = 0$. We are required to find the scalar values α and β such that

$$\mathbf{a} = \alpha\mathbf{e} + \beta\mathbf{e}^*$$

$$\mathbf{a} \cdot \mathbf{e} = \alpha\mathbf{e} \cdot \mathbf{e} + \beta\mathbf{e^*} \cdot \mathbf{e}, \qquad \mathbf{a} \cdot \mathbf{e^*} = \alpha\mathbf{e} \cdot \mathbf{e^*} + \beta\mathbf{e^*} \cdot \mathbf{e^*}$$

From orthogonality and unit lengths of the vectors $\mathbf{e}$ and $\mathbf{e}^*$, this becomes

$$\mathbf{a} \cdot \mathbf{e} = \alpha, \qquad \mathbf{a} \cdot \mathbf{e^*} = \beta$$

Hence

$$\mathbf{a} = (\mathbf{a} \cdot \mathbf{e})\,\mathbf{e} + (\mathbf{a} \cdot \mathbf{e}^*)\,\mathbf{e}^*$$

The projection of a vector $\mathbf{a}$ onto a direction specified by general $\mathbf{f}$ is $\frac{\mathbf{a}\cdot\mathbf{f}}{|\boldsymbol{f}|^2}\mathbf{f}$.

Determinant

For a square matrix $\mathbf{A}$, the determinant is the volume of the transformation of the matrix $\mathbf{A}$ (up to a sign change). That is, we take a hypercube of unit volume and map each vertex under the transformation. The volume of the resulting object is defined as the determinant. Writing $[\mathbf{A}]_{ij} = a_{ij}$,

$$\det \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = a_{11}a_{22} - a_{21}a_{12}$$

The determinant in the (3×3) case has the form

$$a_{11}\det \begin{pmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{pmatrix} - a_{12}\det \begin{pmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{pmatrix} + a_{13}\det \begin{pmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix}$$

More generally, the determinant can be computed recursively as an expansion along the top row of determinants of reduced matrices.
The absolute value of the determinant is the volume of the transformation.

$$\det\left(\mathbf{A}^{\mathsf{T}}\right) = \det\left(\mathbf{A}\right)$$

For square matrices $\mathbf{A}$ and $\mathbf{B}$ of equal dimensions,

$$\det\left(\mathbf{AB}\right) = \det\left(\mathbf{A}\right)\det\left(\mathbf{B}\right), \qquad \det\left(\mathbf{I}\right) = 1 \Rightarrow \det\left(\mathbf{A}^{-1}\right) = 1/\det\left(\mathbf{A}\right)$$

Matrix inversion

For a square matrix $\mathbf{A}$, its inverse satisfies

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{I} = \mathbf{A}\mathbf{A}^{-1}$$

It is not always possible to find a matrix $\mathbf{A}^{-1}$ such that $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$, in which case $\mathbf{A}$ singular. Geometrically, singular matrices correspond to projections: if we transform each of the vertices $\mathbf{v}$ of a binary hypercube using $\mathbf{A}\mathbf{v}$, the volume of the transformed hypercube is zero ($\mathbf{A}$ has determinant zero). Given a vector $\mathbf{y}$ and a singular transformation, $\mathbf{A}$, one cannot uniquely identify a vector $\mathbf{x}$ for which $\mathbf{y} = \mathbf{A}\mathbf{x}$. Provided the inverses exist

$$(\mathbf{A}\mathbf{B})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$$

Pseudo inverse

For a non-square matrix $\mathbf{A}$ such that $\mathbf{A}\mathbf{A}^{\mathsf{T}}$ is invertible,

$$\mathbf{A}^{\dagger} = \mathbf{A}^{\mathsf{T}}\left(\mathbf{A}\mathbf{A}^{\mathsf{T}}\right)^{-1}$$

satisfies $\mathbf{A}\mathbf{A}^{\dagger} = \mathbf{I}$.

Solving Linear Systems

Problem

Given square $N \times N$ matrix $\mathbf{A}$ and vector $\mathbf{b}$, find the vector $\mathbf{x}$ that satisfies

$$\mathbf{A}\mathbf{x} = \mathbf{b}$$

Solution

Algebraically, we have the inverse:

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$$

In practice, we solve solve for $\mathbf{x}$ numerically using Gaussian Elimination – more stable numerically.

Complexity

Solving a linear system is $O\left(N^3\right)$ – can be very expensive for large N.
Approximate methods include conjugate gradient and related approaches.

Matrix rank

For an $m \times n$ matrix $\mathbf{X}$ with n columns, each written as an m-vector:

$$\mathbf{X} = \left[\mathbf{x}^1, \ldots, \mathbf{x}^n\right]$$

the rank of $\mathbf{X}$ is the maximum number of linearly independent columns (or equivalently rows).

Full rank

An $n \times n$ square matrix is full rank if the rank is n, in which case the matrix is must be non-singular. Otherwise the matrix is reduced rank and is singular.

Trace and Det

$$\operatorname{trace}(\mathbf{A}) = \sum_i A_{ii} = \sum_i \lambda_i$$

where λ_i are the eigenvalues of $\mathbf{A}$.

$$\det(\mathbf{A}) = \prod_{i=1}^{n} \lambda_i$$

A matrix is singular if it has a zero eigenvalue.

Trace-Log formula

For a positive definite matrix $\mathbf{A}$,

$$\operatorname{trace}(\log \mathbf{A}) \equiv \log \det(\mathbf{A})$$

The above logarithm of a matrix is not the element-wise logarithm. In general for an analytic function $f(x)$, $f(\mathbf{M})$ is defined via the power-series expansion of the function. On the right, since $\det(\mathbf{A})$ is a scalar, the logarithm is the standard logarithm of a scalar. (A real matrix has a real logarithm if and only if it is invertible and each Jordan block belonging to a nonnegative eigenvalue occurs an even number of times. Also see pp. 558-561in the book *Matrix Computations* by Golub and Van Loan.)

Orthogonal matrix

A square matrix $\mathbf{A}$ is orthogonal if

$$\mathbf{A}\mathbf{A}^{\mathsf{T}} = \mathbf{I} = \mathbf{A}^{\mathsf{T}}\mathbf{A}$$

From the properties of the determinant, we see therefore that an orthogonal matrix has determinant ± 1 and hence corresponds to a volume preserving transformation.

Linear transformations

Cartesian coordinate system

Define $\mathbf{u}_i$ to be the vector with zeros everywhere expect for the i^{th} entry, then a vector can be expressed as $\mathbf{x} = \sum_i x_i \mathbf{u}_i$.

Linear transformation

A linear transformation of $\mathbf{x}$ is given by matrix multiplication by some matrix $\mathbf{A}$

$$\mathbf{A}\mathbf{x} = \sum_i x_i \mathbf{A}\mathbf{u}_i = \sum_i x_i \mathbf{a}_i$$

where $\mathbf{a}_i$ is the i^{th} column of $\mathbf{A}$.

Eigenvalues and eigenvectors

For an $n \times n$ square matrix $\mathbf{A}$, $\mathbf{e}$ is an eigenvector of $\mathbf{A}$ with eigenvalue λ if

$$\mathbf{A}\mathbf{e} = \lambda\mathbf{e}$$

For an $(n \times n)$ dimensional matrix, there are (including repetitions) n eigenvalues, each with a corresponding eigenvector. We can write

$$\underbrace{(\mathbf{A} - \lambda\mathbf{I})}_{\mathbf{B}}\,\mathbf{e} = \mathbf{0}$$

If $\mathbf{B}$ has an inverse, then the only solution is $\mathbf{e} = \mathbf{B}^{-1}\mathbf{0} = \mathbf{0}$, which trivially satisfies the eigen-equation. For any non-trivial solution we therefore need $\mathbf{B}$ to be non-invertible. Hence λ is an eigenvalue of $\mathbf{A}$ if

$$\det\left(\mathbf{A} - \lambda\mathbf{I}\right) = 0$$

It may be that for an eigenvalue λ the eigenvector is not unique and there is a space of corresponding vectors.

Spectral decomposition

A real symmetric matrix $N \times N$ $\mathbf{A}$ has an eigen-decomposition

$$\mathbf{A} = \sum_{i=1}^{n} \lambda_i \mathbf{e}_i \mathbf{e}_i^{\mathsf{T}}$$

where λ_i is the eigenvalue of eigenvector $\mathbf{e}_i$ and the eigenvectors form an orthogonal set,

$$\left(\mathbf{e}^i\right)^{\mathsf{T}} \mathbf{e}^j = \delta_{ij} \left(\mathbf{e}^i\right)^{\mathsf{T}} \mathbf{e}^i$$

In matrix notation

$$\mathbf{A} = \mathbf{E}\mathbf{\Lambda}\mathbf{E}^{\mathsf{T}}$$

where $\mathbf{E}$ is the orthogonal matrix of eigenvectors and $\mathbf{\Lambda}$ the corresponding diagonal eigenvalue matrix. More generally, for a square non-symmetric 'diagonalisable' $\mathbf{A}$ we can write

$$\mathbf{A} = \mathbf{E}\mathbf{\Lambda}\mathbf{E}^{-1}$$

Computational Complexity

It generally takes $O\left(N^3\right)$ time to compute the eigen-decomposition.

Singular Value Decomposition

The SVD decomposition of a $n \times p$ matrix $\mathbf{X}$ is

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathsf{T}}$$

where $\dim \mathbf{U} = n \times n$ with $\mathbf{U}^{\mathsf{T}}\mathbf{U} = \mathbf{I}_n$. Also $\dim \mathbf{V} = p \times p$ with $\mathbf{V}^{\mathsf{T}}\mathbf{V} = \mathbf{I}_p$. The matrix $\mathbf{S}$ has $\dim \mathbf{S} = n \times p$ with zeros everywhere except on the diagonal entries. The singular values are the diagonal entries $[\mathbf{S}]_{ii}$ and are positive. The singular values are ordered so that the upper left diagonal element of $\mathbf{S}$ contains the largest singular value.

Computational Complexity

It generally takes $O\left(\max(n,p)\left(\min(n,p)\right)^2\right)$ time to compute the SVD-decomposition.

Positive definite matrix

A symmetric matrix $\mathbf{A}$ with the property that $\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} \geq 0$ for any vector $\mathbf{x}$ is called positive semidefinite. A symmetric matrix $\mathbf{A}$, with the property that $\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} > 0$ for any vector $\mathbf{x} \neq 0$ is called positive definite. A positive definite matrix has full rank and is thus invertible.

Eigen-decomposition

Using the eigen-decomposition of $\mathbf{A}$,

$$\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} = \sum_i \lambda_i \mathbf{x}^\mathsf{T}\mathbf{e}^i(\mathbf{e}^i)^\mathsf{T}\mathbf{x} = \sum_i \lambda_i \left(\mathbf{x}^\mathsf{T}\mathbf{e}^i\right)^2$$

which is greater than zero if and only if all the eigenvalues are positive. Hence $\mathbf{A}$ is positive definite if and only if all its eigenvalues are positive.

Multivariate Calculus

Partial derivative

Consider a function of n variables, $f(x_1, x_2, \ldots, x_n)$ or $f(\mathbf{x})$. The partial derivative of f *w.r.t.* x_i is defined as the following limit (when it exists)

$$\frac{\partial f}{\partial x_i} = \lim_{h \to 0} \frac{f(x_1, \ldots, x_{i-1}, x_i + h, x_{i+1}, \ldots, x_n) - f(\mathbf{x})}{h}$$

Gradient vector

For function f the gradient is denoted ∇f or $\mathbf{g}$:

$$\nabla f(\mathbf{x}) \equiv \mathbf{g}(\mathbf{x}) \equiv \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}$$

Interpreting the gradient vector

Consider a function $f(\mathbf{x})$ that depends on a vector $\mathbf{x}$. We are interested in how the function changes when the vector $\mathbf{x}$ changes by a small amount : $\mathbf{x} \rightarrow \mathbf{x} + \boldsymbol{\delta}$, where $\boldsymbol{\delta}$ is a vector whose length is very small:

$$f\left(\mathbf{x} + \boldsymbol{\delta}\right) = f(\mathbf{x}) + \sum_i \delta_i \frac{\partial f}{\partial x_i} + O\left(\boldsymbol{\delta}^2\right)$$

We can interpret the summation above as the scalar product between the vector ∇f with components $[\nabla f]_i = \frac{\partial f}{\partial x_i}$ and $\boldsymbol{\delta}$.

$$f\left(\mathbf{x} + \boldsymbol{\delta}\right) = f(\mathbf{x}) + (\nabla f)^{\mathsf{T}} \boldsymbol{\delta} + O\left(\boldsymbol{\delta}^2\right)$$

Interpreting the Gradient

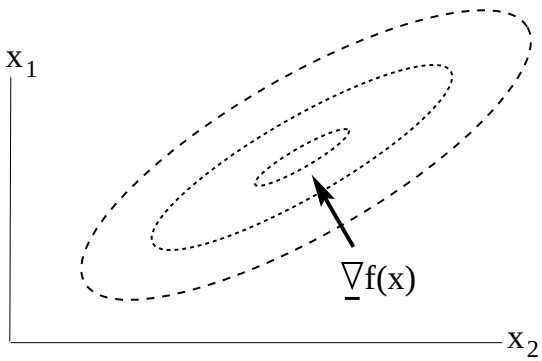

Figure: Interpreting the gradient. The ellipses are contours of constant function value, $f = \text{const}$. At any point $\mathbf{x}$, the gradient vector $\nabla f(\mathbf{x})$ points along the direction of maximal increase of the function.

The gradient points along the direction in which the function increases most rapidly. Why? Consider a direction $\hat{\mathbf{p}}$ (a unit length vector). Then a displacement, δ units along this direction changes the function value to

$$f(\mathbf{x} + \delta\hat{\mathbf{p}}) \approx f(\mathbf{x}) + \delta\nabla f(\mathbf{x}) \cdot \hat{\mathbf{p}}$$

The direction $\hat{\mathbf{p}}$ for which the function has the largest change is that which maximises the overlap

$$\nabla f(\mathbf{x}) \cdot \hat{\mathbf{p}} = |\nabla f(\mathbf{x})||\hat{\mathbf{p}}| \cos\theta = |\nabla f(\mathbf{x})| \cos\theta$$

where θ is the angle between $\hat{\mathbf{p}}$ and $\nabla f(\mathbf{x})$. The overlap is maximised when $\theta = 0$, giving $\hat{\mathbf{p}} = \nabla f(\mathbf{x})/|\nabla f(\mathbf{x})|$. Hence, the direction along which the function changes the most rapidly is along $\nabla f(\mathbf{x})$.

Higher derivatives

The 'second-derivative' of an n-variable function is defined by

$$\frac{\partial}{\partial x_i}\left(\frac{\partial f}{\partial x_j}\right) \qquad i = 1, \ldots, n; \quad j = 1, \ldots, n$$

which is usually written

$$\frac{\partial^2 f}{\partial x_i \partial x_j}, \quad i \neq j \qquad \frac{\partial^2 f}{\partial {x_i}^2}, \quad i = j$$

If the partial derivatives $\partial f/\partial x_i$, $\partial f/\partial x_j$ and $\partial^2 f/\partial x_i \partial x_j$ are continuous, then $\partial^2 f/\partial x_i \partial x_j$ exists and

$$\partial^2 f/\partial x_i \partial x_j = \partial^2 f/\partial x_j \partial x_i \, .$$

This is also denoted by $\nabla\nabla f$. These n^2 second partial derivatives are represented by a square, symmetric matrix called the Hessian matrix of $f(\mathbf{x})$.

$$\mathbf{H}_f(\mathbf{x}) = \begin{pmatrix} \frac{\partial^2 f}{\partial {x_1}^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n} & \cdots & \frac{\partial^2 f}{\partial {x_n}^2} \end{pmatrix}$$

Chain rule

Let each x_j be parameterized by $u_1, \ldots, u_m$, *i.e.* $x_j = x_j(u_1, \ldots, u_m)$.

$$\frac{\partial f}{\partial u_\alpha} = \sum_{j=1}^{n} \frac{\partial f}{\partial x_j} \frac{\partial x_j}{\partial u_\alpha}$$

or in vector notation

$$\frac{\partial}{\partial u_\alpha} f(\mathbf{x}(\mathbf{u})) = \nabla f^{\mathsf{T}}(\mathbf{x}(\mathbf{u})) \frac{\partial \mathbf{x}(\mathbf{u})}{\partial u_\alpha}$$

Derivatives with vectors

The derivative of a vector $\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$, by a scalar x is written as:

$$\frac{\partial \mathbf{y}}{\partial x} = \begin{pmatrix} \frac{\partial y_1}{\partial x} \\ \vdots \\ \frac{\partial y_n}{\partial x} \end{pmatrix}.$$

The derivative of a scalar x by a vector $\mathbf{y}$ is $\nabla x(\mathbf{y})$ and $\nabla x(\mathbf{y}) \cdot \mathbf{n}$ is the directional derivative of x in the $\mathbf{n}$ direction, where $\|\mathbf{n}\| = 1$.

Derivatives with vectors (contd.)

The derivative of a vector function $\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix}$, wrt. an independent vector $\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ is written as:

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial y_m}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_n} \end{pmatrix}.$$

Matrix calculus

The derivative of a matrix function $\mathbf{Y}$ by a scalar x is known as the tangent matrix and is given (in numerator layout notation) by

$$\frac{\partial \mathbf{Y}}{\partial x} = \begin{pmatrix} \frac{\partial y_{11}}{\partial x} & \cdots & \frac{\partial y_{1n}}{\partial x} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_{m1}}{\partial x} & \cdots & \frac{\partial y_{mn}}{\partial x} \end{pmatrix}.$$

The derivative of a scalar y function of a matrix $\mathbf{X}$ of independent variables, with respect to the matrix $\mathbf{X}$, is given by

$$\frac{\partial y}{\partial \mathbf{X}} = \begin{pmatrix} \frac{\partial y}{\partial x_{11}} & \cdots & \frac{\partial y}{\partial x_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial y}{\partial x_{m1}} & \cdots & \frac{\partial y}{\partial x_{mn}} \end{pmatrix}.$$

Matrix calculus (contd.)

For matrices $\mathbf{A}$ and $\mathbf{B}$

$$\frac{\partial}{\partial \mathbf{A}} \text{trace}\left(\mathbf{A}\mathbf{B}\right) \equiv \mathbf{B}^{\mathsf{T}}$$

$$\partial \log\det\left(\mathbf{A}\right) = \partial \text{trace}\left(\log \mathbf{A}\right) = \text{trace}\left(\mathbf{A}^{-1}\partial\mathbf{A}\right)$$

So that

$$\frac{\partial}{\partial \mathbf{A}} \log\det\left(\mathbf{A}\right) = \mathbf{A}^{-\mathsf{T}}$$

For an invertible matrix $\mathbf{A}$,

$$\partial \mathbf{A}^{-1} \equiv -\mathbf{A}^{-\mathsf{T}}\partial\mathbf{A}\mathbf{A}^{-1}$$

Example:

Let $\mathbf{A}$ and $\mathbf{B}$ be matrices of $m \times n$ and $n \times m$ elements, respectively. Let's compute $\frac{\partial}{\partial \mathbf{A}} \text{trace}\,(\mathbf{AB})$.

$$y = \text{trace}\,(\mathbf{AB}) = \sum_{i=1}^{m} \sum_{k=1}^{n} A_{ik} B_{ki}$$

$$\frac{\partial}{\partial A_{rs}} \text{trace}\,(\mathbf{AB}) = \sum_{i=1}^{m} \sum_{k=1}^{n} \frac{\partial}{\partial A_{rs}} A_{ik} B_{ki} = \sum_{i=1}^{m} \sum_{k=1}^{n} \delta_{ir} \delta_{ks} B_{ki} = B_{sr}$$

$$\frac{\partial y}{\partial \mathbf{A}} = \begin{pmatrix} \frac{\partial y}{\partial A_{11}} & \cdots & \frac{\partial y}{\partial A_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial y}{\partial A_{m1}} & \cdots & \frac{\partial y}{\partial A_{mn}} \end{pmatrix} = \begin{pmatrix} B_{11} & \cdots & B_{n1} \\ \vdots & \ddots & \vdots \\ B_{1m} & \cdots & B_{nm} \end{pmatrix} = \mathbf{B}^{\mathsf{T}}$$

Einstenin's notation

According to this convention, when an index variable appears twice in a single term it implies summation of that term over all the values of the index. Fore example:

$$C_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj},$$

can be written as

$$C_{ij} = A_{ik} B_{ki}.$$

Homework

Show that

$$\frac{\partial}{\partial \mathbf{A}} \log \det (\mathbf{A}) = \mathbf{A}^{-\mathsf{T}}.$$

Solution.:

$$\frac{\partial}{\partial A_{rs}} \log \det (\mathbf{A}) = \frac{1}{\det (\mathbf{A})} \frac{\partial}{\partial A_{rs}} \sum_{i=1}^{n} (-1)^{i+j} A_{ij} M_{ij} = \frac{1}{\det (\mathbf{A})} (-1)^{r+s} M_{rs}.$$

where M_{rs} is the cofactor of A_{rs}. But $C_{sr} = (-1)^{r+s} M_{rs}$, where $\mathbf{C} = \operatorname{adj}\left(\mathbf{A}^{\mathsf{T}}\right)$. Namely, $(\operatorname{adj}(\mathbf{A}))_{sr} = (-1)^{r+s} M_{rs}$. Therefore we have:

$$\frac{\partial}{\partial A_{rs}} \log \det (\mathbf{A}) = \frac{\operatorname{adj}\left(\mathbf{A}^{\mathsf{T}}\right)}{\det (\mathbf{A})} = \mathbf{A}^{-\mathsf{T}}.$$

Numerical issues: rounding error

- Often in machine learning we have a large number of terms to sum, for example when computing the log likelihood in for a large number of datapoints.
- It's good to be aware of potential numerical limitations and ways to improve accuracy, should this be a problem. Double floats have a relative error of around 1×10^{-16}.
- Operations that are mathematical identities may not remain so. For example

$$\sum_n x_i^n x_j^n$$

should give rise to a symmetric matrix. However, this symmetry can be lost due to roundoff.

- In general, it's worth checking key points in your code for such issues.

Numerical issues: rounding error

- Consider

$$S = \sum_{i=1}^{N} x_i$$

If x_i cannot be represented exactly by your machine, round-off error will potentially accumulate in computing S.

- Let y be an 'approximation' to each x_i, then

$$S = \sum_{i=1}^{N} (x_i - y + y) = Ny + \sum_{i=1}^{N} (x_i - y)$$

If each x_i is close to y, then the term $\sum_{i=1}^{N} (x_i - y)$ is small and the sum is dominated by the numerically more accurate term Ny.
See `testacc.m` for an example.

logsumexp

- It's common in ML to come across expression such as

$$S = \exp(a) + \exp(b)$$

for large (in absolute value) a or b. If $a = 1000$, Matlab will return ∞ (0 for $a = -1000$).

- It's not sufficient to simply compute the log:

$$\log S = \log\left(\exp(a) + \exp(b)\right)$$

since this requires the exponentiation still of each term.

- Let $m = \max(a, b)$.

$$\log S = m + \log\left(\exp(a - m) + \exp(b - m)\right)$$

Let's say that $m = a$, then

$$\log S = a + \log\left(1 + \exp(b - a)\right)$$

Since $a > b$ then $\exp(b - a) < 1$ and $\log\left(1 + \exp(b - a)\right) < \log 2$. We can compute $\log S$ more accurately this way.

- More generally, we define the `logsumexp` function

$$\texttt{logsumexp}(\mathbf{x}) = m + \sum_i \log\left(\sum_{i=1}^{N} \exp(x_i - m)\right), \quad m = \max(x_1, \ldots, x_N)$$

logsumexp: example

In a classification problem of a 100 dimensional vector $\mathbf{x}$,

$$p(c = i|\mathbf{x}) = \frac{e^{-(\mathbf{x}-\mathbf{m}_i)^2}}{\sum_j e^{-(\mathbf{x}-\mathbf{m}_j)^2}}$$

A naive implementation of this is likely to lead to $\frac{0}{0}$ and a numerical error.

$$\log p(c = i|\mathbf{x}) = -(\mathbf{x} - \mathbf{m}_i)^2 - \texttt{logsumexp}(\mathbf{y})$$

where

$$y_j = -(\mathbf{x} - \mathbf{m}_j)^2$$