Useful definitions and notations

We will treat all vectors as column vectors by default. The space of real vectors of length n is denoted by $\mathbb{R}^n$, while the space of real-valued $m \times n$ matrices is denoted by $\mathbb{R}^{m \times n}$.

Basic linear algebra background

The standard **inner product** between vectors x and y from $\mathbb{R}^n$ is given by

$$\langle x, y \rangle = x^\top y = \sum_{i=1}^n x_i y_i = y^\top x = \langle y, x \rangle$$

Here x_i and y_i are the scalar i-th components of corresponding vectors.

The standard **inner product** between matrices X and Y from $\mathbb{R}^{m \times n}$ is given by

$$\langle X, Y \rangle = \text{tr}(X^\top Y) = \sum_{i=1}^m \sum_{j=1}^n X_{ij} Y_{ij} = \text{tr}(Y^\top X) = \langle Y, X \rangle$$

The determinant and trace can be expressed in terms of the eigenvalues

$$\text{det} A = \prod_{i=1}^n \lambda_i, \qquad \text{tr} A = \sum_{i=1}^n \lambda_i$$

Don't forget about the cyclic property of a trace for a square matrices A, B, C, D:

$$\text{tr}(ABCD) = \text{tr}(DABC) = \text{tr}(CDAB) = \text{tr}(BCDA)$$

The largest and smallest eigenvalues satisfy

$$\lambda_{\text{min}}(A) = \inf_{x \neq 0} \frac{x^\top A x}{x^\top x}, \qquad \lambda_{\text{max}}(A) = \sup_{x \neq 0} \frac{x^\top A x}{x^\top x}$$

and consequently $\forall x \in \mathbb{R}^n$ (Rayleigh quotient):

$$\lambda_{\text{min}}(A) x^\top x \leq x^\top A x \leq \lambda_{\text{max}}(A) x^\top x$$

A matrix $A \in \mathbb{S}^n$ (set of square symmetric matrices of dimension n) is called **positive (semi)definite** if for all $x \neq 0 (\text{for all } x) : x^\top A x > (\geq) 0$. We denote this as

The **condition number** of a nonsingular matrix is defined as

$$\kappa(A) = \|A\|\|A^{-1}\|$$

Matrix and vector multiplication

Let A be a matrix of size $m \times n$, and B be a matrix of size $n \times p$, and let the product AB be:

$$C = AB$$

then C is a $m \times p$ matrix, with element (i, j) given by:

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$$

Let A be a matrix of shape $m \times n$, and x be $n \times 1$ vector, then the i-th component of the product:

$$z = Ax$$

is given by:

$$z_i = \sum_{k=1}^{n} a_{ik} x_k$$

Finally, just to remind:

- $C = AB \quad C^\top = B^\top A^\top$
- $AB \neq BA$
- $e^A = \sum\limits_{k=0}^{\infty} \frac{1}{k!} A^k$
- $e^{A+B} \neq e^A e^B$ (but if A and B are commuting matrices, which means that $AB = BA$, $e^{A+B} = e^A e^B$)
- $\langle x, Ay \rangle = \langle A^\top x, y \rangle$

Gradient

Let $f(x) : \mathbb{R}^n \to \mathbb{R}$, then vector, which contains all first order partial derivatives:

$$
\nabla f(x) = \frac{df}{dx} = \begin{pmatrix} \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}
$$

named gradient of $f(x)$. This vector indicates the direction of steepest ascent. Thus, vector $-\nabla f(x)$ means the direction of the steepest descent of the function in the point. Moreover, the gradient vector is always orthogonal to the contour line in the point.

Hessian

Let $f(x) : \mathbb{R}^n \to \mathbb{R}$, then matrix, containing all the second order partial derivatives:

$$
f''(x) = \frac{\partial^2 f}{\partial x_i \partial x_j} = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n} \end{pmatrix}
$$

In fact, Hessian could be a tensor in such a way: $(f(x) : \mathbb{R}^n \to \mathbb{R}^m)$ is just 3d tensor, every slice is just hessian of corresponding scalar function $(H\left(f_1(x)\right), H\left(f_2(x)\right), \ldots, H\left(f_m(x)\right))$.

Jacobian

The extension of the gradient of multidimensional $f(x) : \mathbb{R}^n \to \mathbb{R}^m$ is the following matrix:

$$
f'(x) = \frac{df}{dx^T} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}
$$

Summary

∂x

X	Y	G	Name
$\mathbb{R}$	$\mathbb{R}$	$\mathbb{R}$	$f'(x)$ (derivative)
$\mathbb{R}^n$	$\mathbb{R}$	$\mathbb{R}^n$	$\dfrac{\partial f}{\partial x_i}$ (gradient)
$\mathbb{R}^n$	$\mathbb{R}^m$	$\mathbb{R}^{m\times n}$	$\dfrac{\partial f_i}{\partial x_j}$ (jacobian)
$\mathbb{R}^{m\times n}$	$\mathbb{R}$	$\mathbb{R}^{m\times n}$	$\dfrac{\partial f}{\partial x_{ij}}$

General concept

Naive approach

The basic idea of naive approach is to reduce matrix/vector derivatives to the well-known scalar derivatives.

Matrix notation of a function

$$f(x) = c^\top x$$

Scalar notation of a function

$$f(x) = \sum_{i=1}^n c_i x_i$$

Matrix notation of a gradient

$$\nabla f(x) = c$$

$$\frac{\partial f(x)}{\partial x_k} = c_k$$

Simple derivative

$$\frac{\partial f(x)}{\partial x_k} = \frac{\partial \left(\sum_{i=1}^n c_i x_i\right)}{\partial x_k}$$

One of the most important practical tricks here is to separate indices of sum (i) and

Differential approach

The guru approach implies formulating a set of simple rules, which allows you to calculate derivatives just like in a scalar case. It might be convenient to use the differential notation here.

Differentials

After obtaining the differential notation of df we can retrieve the gradient using following formula:

$$df(x) = \langle \nabla f(x), dx \rangle$$

Then, if we have differential of the above form and we need to calculate the second derivative of the matrix/vector function, we treat "old" dx as the constant dx_1, then calculate $d(df) = d^2 f(x)$

$$d^2 f(x) = \langle \nabla^2 f(x) dx_1, dx_2 \rangle = \langle H_f(x) dx_1, dx_2 \rangle$$

Properties

Let A and B be the constant matrices, while X and Y are the variables (or matrix functions).

- $dA = 0$
- $d(\alpha X) = \alpha (dX)$
- $d(AXB) = A(dX)B$
- $d(X + Y) = dX + dY$
- $d(X^\top) = (dX)^\top$
- $d(XY) = (dX)Y + X(dY)$
- $d\langle X, Y \rangle = \langle dX, Y \rangle + \langle X, dY \rangle$
- $d\left(\dfrac{X}{\phi}\right) = \dfrac{\phi dX - (d\phi)X}{\phi^2}$
- $d\left(\det X\right) = \det X \langle X^{-\top}, dX \rangle$
- $d\left(\text{tr } X\right) = \langle I, dX \rangle$
- $df(g(x)) = \dfrac{df}{dg} \cdot dg(x)$
- $H = (J(\nabla f))^T$

References

- Convex Optimization book by S. Boyd and L. Vandenberghe - Appendix A. Mathematical background.
- Numerical Optimization by J. Nocedal and S. J. Wright. - Background Material.
- Matrix decompositions Cheat Sheet.
- Good introduction
- The Matrix Cookbook
- MSU seminars (Rus.)
- Online tool for analytic expression of a derivative.
- Determinant derivative