1. Calculating ∇f
2. Calculating $\nabla^2 f, H_f$

Brief recap of matrix calculus

Useful definitions and notations

We will treat all vectors as column vectors by default.

Matrix and vector multiplication

Let A be $m \times n$, and B be $n \times p$, and let the product AB be: $O(n^3)$

$$
\underset{m\times p}{C} = \underset{m\times n}{A}\,\underset{n\times p}{B}
$$

$O(n^2)$

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

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

Let A be $m \times n$, and x be $n \times 1$, then the typical element of the product:

matvec

$$
z = Ax
$$

$O(n^2)$

is given by:

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

Finally, just to remind:

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

$\langle z^T, p^T B \rangle = \langle z^T B^T, p^T \rangle$

Gradient

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

$f(x) = c^T x$

$$\nabla f(x) = \frac{df}{dx} = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix} \in \mathbb{R}^n$$

$f(X) = \det X$

$\operatorname{tr} X$

$\frac{\partial f}{\partial X} = \left(\frac{\partial f}{\partial x_{ij}} \right) \forall i,j$ ← MATRIX

Hessian

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

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

But actually, 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 $(H(f_1(x)), H(f_2(x)), \ldots, H(f_m(x)))$.

Jacobian

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

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

Summary

$$f(x) : X \to Y; \qquad \frac{\partial f(x)}{\partial x} \in G$$

INPUT OUTPUT gradient

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

$\left(\frac{\partial f}{\partial x_{ij}} \right)_{i,j}$ ← matrix

∇f

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.

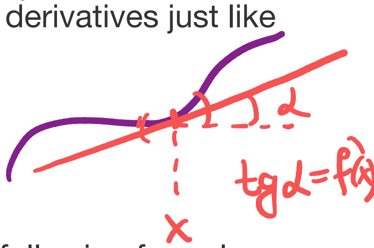
$X \in \mathbb{R}^n \quad X + dX \quad \|dX\|_2 \to 0$

General concept

$df = f(X + dX) - f(X) \to 0 \qquad \lim_{dX \to 0} \frac{df}{dX} = f'(x)$

The idea 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.

$f(x) = c^T x$



Differentials

$df = c^T(x + dx) - c^T x = c^T dx = \langle c; dx \rangle \Rightarrow \boxed{\nabla f = c}$

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

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

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 $d(df)$

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

Properties

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

$\langle \underset{m\times n}{A}, \underset{m\times n}{B} \rangle = tr(\underset{n\times m}{A^T} \underset{m\times n}{B}) = tr(B^T A)$

$\langle A, A \rangle = tr(A^T A) = \|A\|_F^2$

$\|A\|_F = \sqrt{\sum_{ij} a_{ij}^2}$

- $dA = 0$
- $d(\alpha X) = \alpha(dX)$
- $d(AXB) = A(dX)B$
- $d(X + Y) = dX + dY$
- $d(X^\top) = (dX)^\top$
- $d(XY) = (dX)Y + X(dY)$
- $d\langle X, Y \rangle = \langle dX, Y \rangle + \langle X, dY \rangle$
- $d\left(\frac{X}{\phi}\right) = \frac{\phi dX - (d\phi)X}{\phi^2}$
- $\boxed{d(\det X) = \det X \langle X^{-\top}, dX \rangle}$ ← we'll derive it manually
- $d(\text{tr } X) = \langle I, dX \rangle$
- $df(g(x)) = \frac{df}{dg} \cdot dg(x)$
- $H = (J(\nabla f))^T$
- $d(X^{-1}) = -X^{-1}(dX)X^{-1}$

References

- Good introduction
- The Matrix Cookbook
- MSU seminars (Rus.)
- Online tool for analytic expression of a derivative.
- Determinant derivative

$\nabla f = ?$ Gradient

$\nabla^2 f = ?$ Hessians

1. $df = \dots\dots$
2. $df = \langle \dots , dx \rangle$, where the dotted part is ∇f

Examples

Example 1

Find $\nabla f(x)$, if $f(x) = \dfrac{1}{2}x^T Ax + b^T x + c$.

1) $f = \frac{1}{2}\langle x, Ax\rangle + \langle b, x\rangle + c$

2) $df = d\left(\frac{1}{2}\langle x, Ax\rangle + \langle b,x\rangle + c\right) = d\left(\frac{1}{2}\langle x, Ax\rangle\right) + d(\langle b, x\rangle) + \cancel{dc}^{\,0} =$

$= \frac{1}{2} d(\langle x, Ax\rangle) + \langle b, dx\rangle = \frac{1}{2}\left(\langle dx, Ax\rangle + \langle x, d(Ax)\rangle\right) + \langle b, dx\rangle =$

$= \frac{1}{2}\left(\langle Ax, dx\rangle + \langle x, Adx\rangle\right) + \langle b, dx\rangle = \frac{1}{2}\left(\langle Ax, dx\rangle + \langle A^T x, dx\rangle\right) + \langle b, dx\rangle =$

$= \langle \frac{1}{2}(A + A^T)x + b, dx\rangle \Rightarrow \boxed{\nabla f = \frac{1}{2}(A+A^T)x + b}$

Example 2

Find $\nabla f(x)$, $f''(x)$, if $f(x) = -e^{-x^T x}$.

1. $f = -e^{-\langle x, x\rangle}$

2. $df = d\left(-e^{-\langle x,x\rangle}\right) = -d\left(e^{-\langle x,x\rangle}\right) = -\underbrace{e^{-\langle x,x\rangle}}_{f} \cdot d(-\langle x,x\rangle) =$

$= f \cdot -(\langle dx, x\rangle + \langle x, dx\rangle) = -f \cdot \langle 2x, dx\rangle = \langle \underbrace{-2fx}_{\nabla f}, dx\rangle$

$\Rightarrow \boxed{\nabla f = 2\, e^{-\langle x,x\rangle} \cdot x} \in \mathbb{R}^n$

$1\times1 \quad 1\times1 \quad n\times1$

$$df = -f \cdot \langle 2x, dx_1 \rangle$$

$$d^2 f = d(df) = d\left(-f \cdot \langle 2x, dx_1 \rangle\right) = -d\left(f \cdot \langle 2x, dx_1 \rangle\right) =$$

$$= -\underline{df} \cdot \langle 2x, dx_1 \rangle - f \cdot d\left(\langle 2x, dx_1 \rangle\right) =$$

$$= +f \langle 2x, dx \rangle \cdot \langle 2x, dx_1 \rangle - f \langle 2dx, dx_1 \rangle =$$

$$dx$$
$$\|$$
$$d^2 f = \langle H\, dx_1, dx_2 \rangle$$

$$= 2f\left[2\, x^T dx \cdot x^T dx_1 - dx^T \cdot dx_1\right] =$$

$$\langle (x^T dx_1 \cdot x^T)^T, dx \rangle$$

$$= 2f\left[2\, x^T dx_1 \cdot x^T dx - dx_1^T\, dx\right] =$$

$$\langle dx_1, dx \rangle$$

$$= 2f \cdot \langle 2\, x^T \underline{dx_1}\, x - \underline{dx_1}, dx \rangle =$$

$$= 2f \langle 2\, x x^T dx_1 - dx_1, dx \rangle =$$

$$= 2f \langle (2\, \underset{n\times 1}{x} \underset{1\times n}{x^T} - I_n)\, dx_1, dx \rangle$$

$$I_n \cdot dx_1 = dx_1$$

$$d^2 f = \langle 2f \cdot (2xx^T - I)\, dx_1, dx_2 \rangle$$

$$\Rightarrow \boxed{\nabla^2 f = 2f \cdot (2xx^T - I)}$$

$df = \langle \frac{1}{2}(A+A^T)x + b, dx_1 \rangle$

$d^2f = d(df) = d\left(\langle \frac{1}{2}(A+A^T)x+b, \underline{dx_1} \rangle\right) = \langle \frac{1}{2} d\left[(A+A^T)x+b\right], dx_1 \rangle =$

$= \langle \frac{1}{2}(A+A^T)\underline{dx_2}, dx_1 \rangle = \langle dx_2, \frac{1}{2}(A+A^T)dx_1 \rangle =$

$= \langle \frac{1}{2}(A+A^T)dx_1, dx_2 \rangle \Rightarrow \boxed{\nabla^2 f = \frac{1}{2}(A+A^T)}$

$d(\ln \varphi) = \frac{d\varphi}{\varphi}$

Example 3

$\ln x = \log_e x$

Find $\nabla f(X)$, if $f(X) = \langle S, X \rangle - \log \det X$.

1. $df = d\left(\langle S, X \rangle - \ln \det X\right) = d(\langle S, X \rangle) - d(\ln \det X) =$

$= \langle S, dX \rangle - \frac{d(\det X)}{\det X} = \langle S, dX \rangle - \frac{\cancel{\det X} \cdot \langle X^{-T}, dX \rangle}{\cancel{\det X}} =$ $\det X \neq 0$

$= \langle S - X^{-T}, dX \rangle \Rightarrow \boxed{\nabla f = S - X^{-T}}$

Example 4

Find $\nabla f(X)$, if $f(X) = \ln \langle Ax, x \rangle$, $A \in \mathbb{S}^n_{++}$

$$f = \ln\langle Ax, x\rangle \qquad A \in S^n_{++} \qquad A = A^T$$

$$df = d(\ln\langle Ax,x\rangle) = \frac{d(\langle Ax,x\rangle)}{\langle Ax,x\rangle} = \frac{\langle A\,dx, x\rangle + \langle Ax, dx\rangle}{\langle Ax,x\rangle} =$$

$$= \frac{\langle A^T x, dx\rangle + \langle Ax, dx\rangle}{\langle Ax, x\rangle} = \left\langle \frac{2Ax}{\langle Ax,x\rangle}, dx\right\rangle$$

$$\Rightarrow \boxed{\nabla f = \frac{2Ax}{\langle Ax, x\rangle}}$$

$f(X) = \det X$. What is $df = ?$

$\det X \neq 0$

$\nabla f = ?$

$df = f(X + dX) - f(X)$

$df = \langle \nabla f, dX \rangle$

$\|dX\| \to 0 \quad I = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}$

$df = \det(X + dX) - \det X = \det X(I + X^{-1}dX) - \det X =$

$= \det X \cdot \det(I + X^{-1}dX) - \det X = \det X\left[\det(I + X^{-1}dX) - 1\right] =$

$\det(I + X^{-1}dX) = \prod_{i=1}^{n} \lambda_i (I + X^{-1}dX) = \prod_{i=1}^{n} \left(1 + \lambda_i(X^{-1}dX)\right) =$

$= 1 + \sum_{i=1}^{n} \lambda_i (X^{-1}dX) + O(\|dX\|) \approx 1 + \underbrace{\sum_{i=1}^{n} \lambda_i (X^{-1}dX)}_{= tr(X^{-1}dX)}$ when $\|dX\| \to 0$

$= \det X[\cancel{1} + tr(X^{-1}dX) - \cancel{1}] = \det X \cdot tr(X^{-1}dX) =$

tr

$\langle X^{-T}, dX \rangle =$

$= tr(X^{-1}dX)$

$= \langle \det X \cdot X^{-T}, dX \rangle$

$$\boxed{\nabla f = \det X \cdot X^{-T}}$$