Matrix-vector differentiation

Mathematical Optimization

Georgiy Kormakov

CMC MSU

9 February 2024

Differentiation definition

Let U and V be *finite-dimensional linear spaces with norms*.
Examples: $\mathbb{R}$, $\mathbb{R}^n$, $\mathbb{R}^{n \times m}$, their Cartesian products.
Consider the function $f : X \to V, \quad X \subset U$.

Differentiation definition

Let U and V be *finite-dimensional linear spaces with norms*.
Examples: $\mathbb{R}$, $\mathbb{R}^n$, $\mathbb{R}^{n \times m}$, their Cartesian products.
Consider the function $f : X \to V, \quad X \subset U$.

Differentiation

Let $x \in X$ be the inner point of X, and $L : U \to V$ be a linear operator. We will say that the function f **is differentiable** at the point x with the derivative L if for all sufficiently small $h \in U$ it is true

$$f(x+h) = f(x) + L[h] + o(||h||) \quad \Longleftrightarrow \quad \lim_{h \to 0} \frac{||f(x+h) - f(x) - L[h]||}{||h||} = 0.$$

- Non-differentiable?

Differentiation definition

Let U and V be *finite-dimensional linear spaces with norms*.
Examples: $\mathbb{R}$, $\mathbb{R}^n$, $\mathbb{R}^{n\times m}$, their Cartesian products.
Consider the function $f : X \to V, \quad X \subset U$.

Differentiation

Let $x \in X$ be the inner point of X, and $L : U \to V$ be a linear operator. We will say that the function f **is differentiable** at the point x with the derivative L if for all sufficiently small $h \in U$ it is true

$$f(x+h) = f(x)+L[h]+o(||h||) \quad \Longleftrightarrow \quad \lim_{h\to 0} \frac{||f(x+h) - f(x) - L[h]||}{||h||} = 0.$$

- Non-differentiable? For any linear operator f does not satisfy the definition.
- What norm?

Differentiation definition

Let U and V be *finite-dimensional linear spaces with norms*.
Examples: $\mathbb{R}$, $\mathbb{R}^n$, $\mathbb{R}^{n\times m}$, their Cartesian products.
Consider the function $f : X \to V, \quad X \subset U$.

Differentiation

Let $x \in X$ be the inner point of X, and $L : U \to V$ be a linear operator. We will say that the function f **is differentiable** at the point x with the derivative L if for all sufficiently small $h \in U$ it is true

$$f(x+h) = f(x)+L[h]+o(||h||) \quad \Longleftrightarrow \quad \lim_{h\to 0} \frac{||f(x+h) - f(x) - L[h]||}{||h||} = 0.$$

- Non-differentiable? For any linear operator f does not satisfy the definition.
- What norm? Any!

Differential and directional derivative

Differential

The **differential** $df(x)[h] \in V$ at the point $x \in X$ differentiability of the function f and with an increment h is called the vector $f'(x)[h]$.

Notation: $df(x)[h] \equiv Df(x)[h] \equiv f'(x)dx$. In practice, h is removed, leaving $df(x)$, and x is removed, leaving df :)

Differential and directional derivative

Differential

The **differential** $df(x)[h] \in V$ at the point $x \in X$ differentiability of the function f and with an increment h is called the vector $f'(x)[h]$.

Notation: $df(x)[h] \equiv Df(x)[h] \equiv f'(x)dx$. In practice, h is removed, leaving $df(x)$, and x is removed, leaving df :)

Directional derivative

The derivative in the direction h of the function f at the point x is called

$$\frac{\partial f(x)}{\partial h} := \lim_{t \to +0} \frac{f(x+th) - f(x)}{t}.$$

- Partial derivative?

Differential and directional derivative

Differential

The **differential** $df(x)[h] \in V$ at the point $x \in X$ differentiability of the function f and with an increment h is called the vector $f'(x)[h]$.

Notation: $df(x)[h] \equiv Df(x)[h] \equiv f'(x)dx$. In practice, h is removed, leaving $df(x)$, and x is removed, leaving df :)

Directional derivative

The derivative in the direction h of the function f at the point x is called

$$\frac{\partial f(x)}{\partial h} := \lim_{t \to +0} \frac{f(x + th) - f(x)}{t}.$$

- Partial derivative? Simply take the unit element of space
- Connection with the differentiation definition?

Differential and directional derivative

Differential

The **differential** $df(x)[h] \in V$ at the point $x \in X$ differentiability of the function f and with an increment h is called the vector $f'(x)[h]$.

Notation: $df(x)[h] \equiv Df(x)[h] \equiv f'(x)dx$. In practice, h is removed, leaving $df(x)$, and x is removed, leaving df :)

Directional derivative

The derivative in the direction h of the function f at the point x is called

$$\frac{\partial f(x)}{\partial h} := \lim_{t \to +0} \frac{f(x+th) - f(x)}{t}.$$

- Partial derivative? Simply take the unit element of space
- Connection with the differentiation definition? Equal, if f is differentiable.

Gradient

1. Differentiability at point $x \Rightarrow \exists \frac{\partial f(x)}{\partial h} \forall h$. The converse is not true. A sufficient condition for differentiability —

Gradient

1. Differentiability at point $x \Rightarrow \exists \frac{\partial f(x)}{\partial h} \forall h$. The converse is not true. A sufficient condition for differentiability — the continuity of all partial derivatives $\frac{\partial f(x)}{\partial x_i}$.
2. $f : \mathbb{R}^n \to \mathbb{R}$: $Df(x)[h] = \langle a_x, h \rangle$, where $a_x \in \mathbb{R}^n$ — the `gradient` of f $(\nabla f(x))$, depends on x.
   Taking $h = e_i$, we receive the standard form:
   $\nabla f(x) = \left(\frac{\partial f}{\partial x_1}(x), \dots, \frac{\partial f}{\partial x_n}(x) \right)^\top \in \mathbb{R}^n.$
3. $f : \mathbb{R}^{n \times m} \to \mathbb{R}$: $Df(X)[H] = \langle A_X, H \rangle$, with $A_X(X) \in \mathbb{R}^{n \times m}$ — the `gradient` of f $(\nabla f(X))$.
   We also receive the standard form by taking $h = e_{ij}$.

Jacobian and more...

1. $f : \mathbb{R}^n \to \mathbb{R}^m$

$$Df(x)[h] = J_f(x)h, \quad \text{where } J_f(x) \in \mathbb{R}^{n\times m}$$

Matrix $J_x(x)$ called `Jacobian` of $f(x)$ in point x.
Taking $h = e_i$, we receive the standard form:
$J_f(x) \equiv \frac{\partial f}{\partial x} := \left(\frac{\partial f_i}{\partial x_j}(x)\right)_{i,j} \in \mathbb{R}^{n\times m}$.

2. In all other cases, to construct a derivative, it is enough to find all partial derivatives in the form of a tensor

$$\frac{\partial f_{ij}}{\partial x_{kl}}(x).$$

What should we remember, when taking simply partial derivatives?

Summary

In \ Out	$\mathbb{R}$	$\mathbb{R}^n$	$\mathbb{R}^{n\times m}$
$\mathbb{R}$	$df(x) = f'(x)dx$ $f'(x)$ scalar, dx scalar.	-	-
$\mathbb{R}^m$	$df(x) = \langle \nabla f(x), dx \rangle$ $f(x)$ vector, dx vector	$df(x) = J_x dx$ J_x matrix, dx vector	-
$\mathbb{R}^{n'\times m'}$	$df(X) = \langle \nabla f(X), dX \rangle$ $\nabla f(X)$ matrix, dX matrix	-	-

Second derivative

Let $f : U \to V$ be differentiable at each point $x \in U$. Consider the differential of the function f with a fixed increment h_1 as a function of x:

$$g(x) = Df(x)[h_1].$$

Second derivative

If at some point x the function g has a derivative, then it is called the second derivative, and the second differential has the form

$$D^2 f(x)[h_1, h_2] := D(Df[h_1])(x)[h_2].$$

- Higher order?

Second derivative

Let $f : U \to V$ be differentiable at each point $x \in U$. Consider the differential of the function f with a fixed increment h_1 as a function of x:

$$g(x) = Df(x)[h_1].$$

Second derivative

If at some point x the function g has a derivative, then it is called the second derivative, and the second differential has the form

$$D^2 f(x)[h_1, h_2] := D(Df[h_1])(x)[h_2].$$

- Higher order? Yes, iteratively
- Continuously differentiable?

Second derivative

Let $f : U \to V$ be differentiable at each point $x \in U$. Consider the differential of the function f with a fixed increment h_1 as a function of x:

$$g(x) = Df(x)[h_1].$$

Second derivative

If at some point x the function g has a derivative, then it is called the second derivative, and the second differential has the form

$$D^2 f(x)[h_1, h_2] := D(Df[h_1])(x)[h_2].$$

- Higher order? Yes, iteratively
- Continuously differentiable? If $g(x)$ is continuous

Hessian and connection with Jacobian

What kind of a form is the second differential?

Hessian and connection with Jacobian

What kind of a form is the second differential? Bilinear.
So, in the case of $f : \mathbb{R}^n \to \mathbb{R}$

$$D^2 f(x)[h_1, h_2] = \langle H_x h_1, h_2 \rangle.$$

The matrix H_x is called the `Hessian` of the function f at the point x and is denoted by $\nabla^2 f(x)$.
What is the connection between Jacobian of ∇f and Hessian?

Hessian and connection with Jacobian

What kind of a form is the second differential? Bilinear.
So, in the case of $f : \mathbb{R}^n \to \mathbb{R}$

$$D^2 f(x)[h_1, h_2] = \langle H_x h_1, h_2 \rangle.$$

The matrix H_x is called the `Hessian` of the function f at the point x and is denoted by $\nabla^2 f(x)$.
What is the connection between Jacobian of ∇f and Hessian?

$$d(\nabla f(x)) = (\nabla^2 f)^\top dx \Leftrightarrow \nabla^2 f(x) = (J_{\nabla f})^\top.$$

In the standard basis, the Hessian has the form

$$\nabla^2 f(x) = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}(x) \right)_{ij}.$$

When is it symmetric?

Hessian and connection with Jacobian

What kind of a form is the second differential? Bilinear.
So, in the case of $f : \mathbb{R}^n \to \mathbb{R}$

$$D^2 f(x)[h_1, h_2] = \langle H_x h_1, h_2 \rangle.$$

The matrix H_x is called the `Hessian` of the function f at the point x and is denoted by $\nabla^2 f(x)$.
What is the connection between Jacobian of ∇f and Hessian?

$$d(\nabla f(x)) = (\nabla^2 f)^\top dx \Leftrightarrow \nabla^2 f(x) = (J_{\nabla f})^\top.$$

In the standard basis, the Hessian has the form

$$\nabla^2 f(x) = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}(x) \right)_{ij}.$$

When is it symmetric? For a doubly continuously differentiable function

The ways to compute derivatives?

Ok, we know what the derivatives look like. But how to calculate them in practice?

The ways to compute derivatives?

Ok, we know what the derivatives look like. But how to calculate them in practice?

1. Using definition, obviously... Take the partial derivatives as they are and evaluate coordinate by coordinate. Maybe, numerically...

The ways to compute derivatives?

Ok, we know what the derivatives look like. But how to calculate them in practice?

1. Using definition, obviously... Take the partial derivatives as they are and evaluate coordinate by coordinate. Maybe, numerically...
2. Or use simple rules as in simple calculus (sum of functions, multiplication, division, composition, and so on...)

Differentiation rules

1. (Linearity) Let $f : X \to V$ and $g : X \to V$.
   If f, g are differentiable at x, while $c_1, c_2 \in \mathbb{R}$ are numbers, then $c_1 f + c_2 g$ is differentiable at x and

$$d(c_1 f + c_2 g) = c_1 df + c_2 dg.$$

Differentiation rules

1. (Linearity) Let $f : X \to V$ and $g : X \to V$.
   If f, g are differentiable at x, while $c_1, c_2 \in \mathbb{R}$ are numbers, then $c_1 f + c_2 g$ is differentiable at x and
   $$d(c_1 f + c_2 g) = c_1 df + c_2 dg.$$

2. (Multiplication) Let $\alpha : X \to \mathbb{R}$ and $f : X \to V$ be functions.
   If α, f are differentiable at point x, then αf is differentiable at point x and
   $$D(\alpha f)(x)[h] = (D\alpha(x)[h])f(x) + \alpha(x)(Df(x)[h])$$
   for any increments of h.

Differentiation rules 2

3. (Composition) Let Y be a subset of V, $f : X \to Y$ be a function. Also let W be a linear space, $g : Y \to W$ be a function. If f is differentiable at x, g is differentiable at $f(x)$, then their composition is $(g \circ f)(x) \equiv g(f(x))$ is differentiable at the point x and

$$D(g \circ f)(x) = Dg(f(x))[df] \iff Dg(f(x))[Df(x)[h]].$$

Differentiation rules 2

3. (Composition) Let Y be a subset of V, $f : X \to Y$ be a function. Also let W be a linear space, $g : Y \to W$ be a function. If f is differentiable at x, g is differentiable at $f(x)$, then their composition is $(g \circ f)(x) \equiv g(f(x))$ is differentiable at the point x and

$$D(g \circ f)(x) = Dg(f(x))[df] \Longleftrightarrow Dg(f(x))[Df(x)[h]].$$

4. (Division) Let $\alpha : X \to \mathbb{R}$ and $f : X \to V$ be functions. If α, f are differentiable in x and does not converge to 0 by X, then $(1/\alpha)f$ is differentiable in x and

$$D\left(\frac{f}{\alpha}\right)(x)[h] = \frac{\alpha(x)(Df(x)[h]) - (D\alpha(x)[h])f(x)}{\alpha(x)^2}.$$

Differentiation rules 3

5. (Multiplication for matrix-valued functions) Let $f : X \to \mathbb{R}^{m \times n}$ and $g : X \to \mathbb{R}^{n \times k}$ be matrix-valued functions. If f, g are differentiable at x, then fg is differentiable at x and

$$D(fg)(x)[h] = (Df(x)[h])g(x) + f(x)(Dg(x)[h]).$$

Matrix multiplication is implied here.

The most frequent functions

- It follows from the product rule that for vector-valued functions $f : X \to \mathbb{R}^n$ and $g : X \to \mathbb{R}^n$ differentiable at x, the function $\langle f, g \rangle$ is differentiable in x and

$$d(\langle f, g \rangle) = \langle df, g \rangle + \langle f, dg \rangle.$$

The most frequent functions

- It follows from the product rule that for vector-valued functions $f : X \to \mathbb{R}^n$ and $g : X \to \mathbb{R}^n$ differentiable at x, the function $\langle f, g \rangle$ is differentiable in x and

$$d(\langle f, g \rangle) = \langle df, g \rangle + \langle f, dg \rangle.$$

- For a vector-valued function $f : X \to \mathbb{R}^n$ differentiable at a point x and a linear map $L : \mathbb{R}^n \to \mathbb{R}^m$ the differential and L are permutable:

$$D(L \circ f)(x)[h] = L[Df(x)[h]].$$

The most frequent functions

- It follows from the product rule that for vector-valued functions $f : X \to \mathbb{R}^n$ and $g : X \to \mathbb{R}^n$ differentiable at x, the function $\langle f, g\rangle$ is differentiable in x and

$$d(\langle f, g\rangle) = \langle df, g\rangle + \langle f, dg\rangle.$$

- For a vector-valued function $f : X \to \mathbb{R}^n$ differentiable at a point x and a linear map $L : \mathbb{R}^n \to \mathbb{R}^m$ the differential and L are permutable:

$$D(L \circ f)(x)[h] = L[Df(x)[h]].$$

- The Jacobi matrix of a composition $f(g(x))$ is equal to the product of Jacobi matrices of composites

$$J_{g(f(x))} = J_g J_f.$$

Tabular functions

1. For $f(x) = \langle c, x \rangle, x \in \mathbb{R}^n$ and increments of h we count
$$f(x+h) - f(x) = \langle c, x+h \rangle - \langle c, x \rangle = \langle c, h \rangle.$$
The mapping $h \to \langle c, h \rangle$ is linear, so it can be taken as a derivative by definition
$$Df(x)[h] = \langle c, h \rangle.$$

Tabular functions

1. For $f(x) = \langle c, x \rangle, x \in \mathbb{R}^n$ and increments of h we count
$$f(x+h) - f(x) = \langle c, x+h \rangle - \langle c, x \rangle = \langle c, h \rangle.$$
The mapping $h \to \langle c, h \rangle$ is linear, so it can be taken as a derivative by definition
$$Df(x)[h] = \langle c, h \rangle.$$
2. For $f(x) = \langle Ax, x \rangle, x \in \mathbb{R}^n, A \in \mathbb{R}^{n \times n}$ and increments of h we count
$$f(x+h) - f(x) = \langle Ax + Ah, x+h \rangle - \langle Ax, x \rangle =$$
$$= \langle (A + A^\top)x, h \rangle + \langle Ah, h \rangle.$$
Note that
$$\langle Ah, h \rangle \leq \|Ah\| \|h\| \leq \|A\| \|h\|^2 = o(\|h\|),$$
Again, by definition
$$Df(x)[h] = \langle (A + A^\top)x, h \rangle.$$

Tabular functions (some matrix example)

3. Let $S := \{X \in \mathbb{R}^{n\times n} : \det(X) \neq 0\}$ and the function $f : S \to S$ reverses the matrix $f(X) = X^{-1}$. For an arbitrary small increment of H, we calculate

$$f(X+H) - f(X) = (X+H)^{-1} - X^{-1} = (X(I_n + X^{-1}H))^{-1} - X^{-1} =$$
$$= ((I_n + X^{-1}H)^{-1} - I_n)X^{-1}.$$

Neumann series.

Let $A \in \mathbb{R}^{n\times n}$ be a matrix such that $\|A\| < 1$, then the matrix $(I_n - A)$ is invertible and

$$(I_n - A)^{-1} = \sum_{k=0}^{\infty} A^k.$$

Tabular functions (still some matrix example)

In our case, we can apply the Neumann series due to the smallness of H

$$(I_n + X^{-1}H)^{-1} = I_n - X^{-1}H + \sum_{k=2}^{\infty}(-X^{-1}H)^k.$$

Tabular functions (still some matrix example)

In our case, we can apply the Neumann series due to the smallness of H

$$(I_n + X^{-1}H)^{-1} = I_n - X^{-1}H + \sum_{k=2}^{\infty}(-X^{-1}H)^k.$$

Let's estimate the norm of the last term

$$\left\|\sum_{k=2}^{\infty}(-X^{-1}H)^k\right\| \leq \sum_{k=2}^{\infty}\|(-X^{-1}H)^k\| \leq \sum_{k=2}^{\infty}\|-X^{-1}\|^k\|H\|^k =$$

$$= \frac{\|X^{-1}\|^2\|H\|^2}{1-\|X^{-1}\|\|H\|} = o(\|H\|),$$

As a result, we get the difference

$$f(X+H) - f(X) = -X^{-1}HX^{-1} + o(\|H\|),$$

in this case, the mapping $H \to -X^{-1}HX^{-1}$ is linear. That is, by definition

$$Df(X)[H] = -X^{-1}HX^{-1}.$$

Tabular functions. The main table

Transformation rules
$d(\alpha X) = \alpha dX$
$d(AXB) = AdXB$
$d(X + Y) = dX + dY$
$d(X^T) = (dX)^T$
$d(XY) = (dX)Y + X(dY)$
$d\langle X, Y\rangle = \langle dX, Y\rangle + \langle X, dY\rangle$
$d\left(\frac{X}{\phi}\right) = \frac{\phi dX - (d\phi)X}{\phi^2}$
$d(g(f(x))) = g'(f)df(x)$
$J_{g(f)} = J_g J_f \Longleftrightarrow \frac{\partial g}{\partial x} = \frac{\partial g}{\partial f}\frac{\partial f}{\partial x}$

Standard derivatives table
$dA = 0$
$\langle A, X\rangle = \langle A, dX\rangle$
$d\langle Ax, x\rangle = \langle (A + A^\top)x, dx\rangle$
$d\mathrm{Tr}\,(X) = \mathrm{Tr}\,(dX)$
$d(\det(X)) = \det(X)\mathrm{Tr}\left(X^{-1}dX\right)$
$d(X^{-1}) = -X^{-1}(dX)X^{-1}$

Quadratic function. Direct method

Quadratic function. Find the first and second differential $df(x), d^2f(x)$, as well as the gradient $\nabla f(x)$ and the hessian $\nabla^2 f(x)$ functions

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

where $A \in \mathbb{R}^{n \times n}$, $b \in \mathbb{R}^n$, $c \in \mathbb{R}$.

Quadratic function. Direct method

Quadratic function. Find the first and second differential $df(x), d^2f(x)$, as well as the gradient $\nabla f(x)$ and the hessian $\nabla^2 f(x)$ functions

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

where $A \in \mathbb{R}^{n\times n}, b \in \mathbb{R}^n, c \in \mathbb{R}$.

Solution. Let's try to apply both approaches to solve this problem.

- First we use the direct method and write out an explicit scalar dependency $f(x_1, \dots, x_n)$

$$f(x_1, \dots, x_n) = \frac{1}{2}\sum_{i=1}^{n} x_i \sum_{j=1}^{n} A_{ij}x_j + \sum_{i=1}^{n} x_i b_i + c =$$

$$= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}x_i x_j + \sum_{i=1}^{n} x_i b_i + c.$$

Quadratic function. Direct method

Find the partial derivative by x_k

$$f(x_1,\dots,x_n) = \frac{1}{2}A_{kk}x_k^2 + \frac{1}{2}\sum_{i\neq k}A_{ik}x_ix_k + \frac{1}{2}\sum_{j\neq k}A_{kj}x_kx_j + x_kb_k + \left(\frac{1}{2}\sum_{i\neq k}\sum_{j\neq k}A_{ij}x_ix_j + \sum_{i\neq k}x_ib_i + c\right).$$

Taking the partial derivative, we get

$$\frac{\partial f}{\partial x_k} = \frac{1}{2}\cdot 2A_{kk}x_k + \frac{1}{2}\sum_{i\neq k}A_{ik}x_i + \frac{1}{2}\sum_{j\neq k}A_{kj}x_j + b_k = \frac{1}{2}(Ax)_k + \frac{1}{2}(A^\top x)_k + b_k.$$

Substituting coordinate, we calculate the gradient

$$\nabla f(x) = \frac{1}{2}(A + A^\top)x + b.$$

Quadratic function. Direct method

To calculate the Hessian, we find the double partial derivative of x_k, x_l

$$\frac{\partial^2 f}{\partial x_l \partial x_k} = \frac{\partial \frac{1}{2} \sum\limits_{i=1}^{n} A_{ik} x_i + \frac{1}{2} \sum\limits_{j=1}^{n} A_{kj} x_j + b_k}{\partial x_l} = \frac{1}{2} A_{lk} + \frac{1}{2} A_{kl} = \frac{1}{2}(A + A^\top)_{kl}.$$

Therefore, the Hessian is

$$\nabla^2 f(x) = \frac{1}{2}(A + A^\top).$$

Quadratic function. Differential approach

Now we use differential calculus

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

Therefore, by reducing to the standard form $df = \langle \nabla f(x), dx\rangle$, we get the gradient

$$\nabla f(x) = \frac{1}{2}(A + A^\top)x + b.$$

Next, for the hessian, we fix the first increment of dx_1 at the first differential and take another differential from it

Quadratic function. Differential approach

$$d^2f = d(df) = d\left\langle \frac{1}{2}(A + A^\top)x + b, dx_1 \right\rangle = \left\langle d\left(\frac{1}{2}(A + A^\top)x + b\right), dx_1 \right\rangle$$
$$+ \left\langle \frac{1}{2}(A + A^\top)x + b, \underbrace{d(dx_1)}_{=0} \right\rangle = \left\langle \frac{1}{2}(A + A^\top)dx, dx_1 \right\rangle.$$

We transfer and transpose the matrix in the scalar product, but since $A + A^\top$ is symmetric, it does not change.

$$d^2f = \left\langle dx, \frac{1}{2}(A + A^\top)^\top dx_1 \right\rangle = \left\langle \frac{1}{2}(A + A^\top)dx_1, dx \right\rangle.$$

Leading to the standard form $d^2f = \langle \nabla^2 f(x) \cdot dx_1, dx \rangle$, we get the hessian

$$\nabla^2 f(x) = \frac{1}{2}(A + A^\top).$$

Note that if A is symmetric, then $\nabla f(x) = Ax + b, \quad \nabla^2 f(x) = A.$

Log quadratic

Find the first and second differential $df(x), d^2f(x)$, as well as the gradient $\nabla f(x)$ and the hessian $\nabla^2 f(x)$ functions $f(x) = \ln\langle Ax, x\rangle$ where $x \in \mathbb{R}^n, A \in \mathbb{S}^n_{++}$.

Log quadratic

Find the first and second differential $df(x), d^2f(x)$, as well as the gradient $\nabla f(x)$ and the hessian $\nabla^2 f(x)$ functions $f(x) = \ln\langle Ax, x\rangle$ where $x \in \mathbb{R}^n, A \in \mathbb{S}^n_{++}$. **Solution.** Find the first differential

$$df = d\ln\langle Ax, x\rangle = \frac{1}{\langle Ax, x\rangle} d\langle Ax, x\rangle = \frac{2\langle Ax, dx\rangle}{\langle Ax, x\rangle} = \left\langle \frac{2Ax}{\langle Ax, x\rangle}, dx \right\rangle.$$

Now let's find the gradient differential

$$d\left(\frac{2Ax}{\langle Ax, x\rangle}\right) = \frac{d(2Ax)\langle Ax, x\rangle - (2Ax)d\langle Ax, x\rangle}{\langle Ax, x\rangle^2} =$$

$$= \frac{2\langle Ax, x\rangle Adx - 4Ax\langle Ax, dx\rangle}{\langle Ax, x\rangle^2} = \left(\frac{2A}{\langle Ax, x\rangle} - \frac{4Axx^\top A}{\langle Ax, x\rangle^2}\right) dx = J_{\nabla f} dx.$$

Since $\nabla^2 f = (J_{\nabla f})^\top$, and the hessian is symmetric due to continuity, then

$$\nabla^2 f = \frac{2A}{\langle Ax, x\rangle} - \frac{4Axx^\top A}{\langle Ax, x\rangle^2}.$$

Euclidean norm

Find the first and second differential $df(x), d^2f(x)$, as well as the gradient $\nabla f(x)$ and the hessian $\nabla^2 f(x)$ functions $f(x) = \|x\|_2, \quad x \in \mathbb{R}^n \setminus \{0\}$.

Euclidean norm

Find the first and second differential $df(x), d^2f(x)$, as well as the gradient $\nabla f(x)$ and the hessian $\nabla^2 f(x)$ functions $f(x) = \|x\|_2, \quad x \in \mathbb{R}^n \setminus \{0\}$.

Solution. Find the first differential

$$df(x) = d(\langle x, x\rangle^{\frac{1}{2}}) = \left\{ dy^{\frac{1}{2}} = \frac{1}{2y^{\frac{1}{2}}} dy \right\} = \frac{d(\langle x, x\rangle)}{2\langle x, x\rangle^{\frac{1}{2}}} =$$

$$= \left\langle \frac{2x}{2\langle x, x\rangle^{\frac{1}{2}}}, dx \right\rangle = \left\langle \frac{x}{\|x\|}, dx \right\rangle.$$

After that, we bring df to the standard form $df = \langle \nabla f, dx\rangle$ and get the gradient

$$\nabla f(x) = \frac{x}{\|x\|}.$$

Euclidean norm. Second differential

Now let's calculate the second differential by fixing the increment dx_1 of the first one

$$df^2(x) = d\left(\left\langle \frac{x}{\|x\|}, dx_1 \right\rangle\right) = \left\langle d\left(\frac{x}{\|x\|}\right), dx_1 \right\rangle = \textit{Division rule}$$

$$= \left\langle \frac{dx\|x\| - xd(\|x\|)}{\|x\|^2}, dx_1 \right\rangle = \left\langle \frac{dx\|x\| - x\left\langle \frac{x}{\|x\|}, dx \right\rangle}{\|x\|^2}, dx_1 \right\rangle$$

$$= \left\langle \frac{I_n\|x\| - \frac{xx^\top}{\|x\|}}{\|x\|^2} dx, dx_1 \right\rangle = \left\langle \left(\frac{I_n\|x\| - \frac{xx^\top}{\|x\|}}{\|x\|^2}\right)^\top dx_1, dx \right\rangle.$$

By representing $d^2 f$ in the standard form $\langle \nabla^2 f(x) \cdot dx_1, dx \rangle$, we get

$$\nabla^2 f(x) = \frac{I_n}{\|x\|} - \frac{xx^\top}{\|x\|^3}.$$

Euclidean norm. Important note

Note that at the point $x = 0$ the function is not differentiable. BUT at the same time we can calculate the derivative in any direction h:

$$\frac{\partial f}{\partial h}(0) = \lim_{t \to 0} \frac{f(0 + th) - f(0)}{t} = \lim_{t \to +0} \frac{||th||}{t} = ||h||.$$

If the function would be differentiable, then

$$df(x)[h] = ||h||,$$

and this is a nonlinear function of h.

Softmax

Find the Jacobi matrix of the function $s(x) = \text{softmax}(x)$

$$\text{softmax}(x) := \left(\frac{\exp(x_1)}{\sum_{i=1}^{n} \exp(x_i)}, \ldots, \frac{\exp(x_n)}{\sum_{i=1}^{n} \exp(x_i)} \right)^\top.$$

Softmax

Find the Jacobi matrix of the function $s(x) = \text{softmax}(x)$ $\text{softmax}(x) := \left(\frac{\exp(x_1)}{\sum_{i=1}^n \exp(x_i)}, \dots, \frac{\exp(x_n)}{\sum_{i=1}^n \exp(x_i)}\right)^\top$. **Solution.** We consider partial derivatives by definition

1. at $k \neq j$

$$
\begin{aligned}
\frac{\partial s_k}{\partial x_j} &= \frac{\partial}{\partial x_j} \frac{\exp(x_k)}{\sum_{i=1}^n \exp(x_i)} = \exp(x_k) \frac{\partial}{\partial x_j} \frac{1}{\sum_{i=1}^n \exp(x_i)} \\
&= \exp(x_k) \frac{-1}{(\sum_{i=1}^n \exp(x_i))^2} \frac{\partial}{\partial x_j} \left(\sum_{i=1}^n \exp(x_i) \right) = \\
&= -\frac{\exp(x_k)\exp(x_j)}{(\sum_{i=1}^n \exp(x_i))^2} = -s_k \cdot s_j,
\end{aligned}
$$

Softmax

2. when $k = j$

$$\frac{\partial s_j}{\partial x_j} = \frac{\partial}{\partial x_j} \frac{\exp(x_j)}{\sum_{i=1}^n \exp(x_i)} =$$

$$= \frac{\exp(x_j)(\sum_{i=1}^n \exp(x_i)) - \exp(x_j)\frac{\partial}{\partial x_j}(\sum_{i=1}^n \exp(x_i))}{(\sum_{i=1}^n \exp(x_i))^2} =$$

$$= \frac{\exp(x_j)}{\sum_{i=1}^n \exp(x_i)} - \frac{\exp(x_j)\exp(x_j)}{(\sum_{i=1}^n \exp(x_i))^2} = s_j(1 - s_j).$$

Total,

$$J_{k,j} = \begin{cases} -s_k \cdot s_j, & k \neq j \\ s_j(1 - s_j), & k = j. \end{cases}$$

Coordinate-wise operations

Find the gradient and Hessian of the function $f(x) = h(g(x))$, where $g(x) = \sin(x)$ element by element, $h(u) = \sum_{i=1}^{n} u_i$.

Coordinate-wise operations

Find the gradient and Hessian of the function $f(x) = h(g(x))$, where $g(x) = \sin(x)$ element by element, $h(u) = \sum_{i=1}^{n} u_i$. **Solution.** The incoming functions are not standard, but are enough easy to evaluate partial derivatives directly.
It is also useful to recall the rule of the Jacobi matrix of a complex function

$$J_f = J_{h(g)} J_g, \text{ the form with gradients: } \nabla f = J_g^\top \nabla h.$$

Coordinate-wise operations

Find the gradient and Hessian of the function $f(x) = h(g(x))$, where $g(x) = \sin(x)$ element by element, $h(u) = \sum_{i=1}^{n} u_i$. **Solution.** The incoming functions are not standard, but are enough easy to evaluate partial derivatives directly.

It is also useful to recall the rule of the Jacobi matrix of a complex function

$$J_f = J_{h(g)} J_g, \text{ the form with gradients: } \nabla f = J_g^\top \nabla h.$$

Next, we calculate the Jacobi matrix of the coordinate function of the form

$$g(x) = \begin{pmatrix} g(x_1) \\ \vdots \\ g(x_n) \end{pmatrix}, \quad J_g = \mathrm{diag}(g'(x_1), \ldots, g'(x_n)) = \mathrm{diag}(g'(x)) = J_g^\top.$$

When multiplying J_g by a vector, it is convenient to use element-wise matrix multiplication, denoted by $\odot$

$$(A \odot B)_{ij} = A_{ij} * B_{ij}.$$

Coordinate-wise operations

The result of multiplying J_g by the vector y is

$$J_g y = \begin{pmatrix} g'(x_1) \\ \vdots \\ g'(x_n) \end{pmatrix} \odot y = g'(x) \odot y.$$

Note that this operation is fairly quickly computable and easily amenable to parallelization.

Now let's proceed to our example

$$J_g = \text{diag}(\cos(x_1), \ldots, \cos(x_n)) = \text{diag}(\cos(x)),$$

$$\{\nabla h(u)\}_j = \frac{\partial(\sum_{i=1}^n u_i)}{\partial u_j} = 1 \quad \rightarrow \quad \nabla h(u) = \mathbf{1},$$

$$\nabla f = J_g^\top \nabla h = \cos(x) \odot \mathbf{1} = \cos(x).$$

And Hessian: $\nabla^2 f(x) = J_{\nabla f}^\top = \text{diag}(-\sin(x))$.

Logistic Regression

Find the first and second differential $df(x), d^2f(x)$, as well as the gradient $\nabla f(x)$ and the hessian $\nabla^2 f(x)$ functions

$$f(x) = \ln(1 + \exp(\langle a, x \rangle)),$$

where $a \in \mathbb{R}^n$.

Logistic Regression

Find the first and second differential $df(x), d^2f(x)$, as well as the gradient $\nabla f(x)$ and the hessian $\nabla^2 f(x)$ functions

$$f(x) = \ln(1 + \exp(\langle a, x\rangle)),$$

where $a \in \mathbb{R}^n$.

Solution. Find the first differential

$$d(\ln(1 + \exp(\langle a, x\rangle))) = \left\{d \ln y = \frac{1}{y} dy\right\} = \frac{1}{1 + \exp(\langle a, x\rangle)} d(1 + \exp(\langle a, x\rangle))$$

$$= \{d \exp(y) = \exp(y) dy\} = \frac{1}{1 + \exp(\langle a, x\rangle)} \exp(\langle a, x\rangle) d(\langle a, x\rangle) =$$

$$= \left\langle \frac{\exp(\langle a, x\rangle)}{1 + \exp(\langle a, x\rangle)} a, dx \right\rangle.$$

Logistic Regression

For convenience, we introduce the sigmoid function $\sigma(x) := \frac{1}{1+\exp(-x)}$. Note that $\sigma(-x) = 1 - \sigma(x)$ and $\sigma'(x) = \sigma(x)(1 - \sigma(x))$. After that, we bring df to the standard form $df = \langle \nabla f, dx \rangle$ and get the gradient

$$\nabla f(x) = \sigma(\langle a, x \rangle) a.$$

Thus, the gradient $\nabla f(x)$ is a vector collinear to the vector a with the coefficient $\sigma(\langle a, x \rangle) \in (0, 1)$. Depending on the point, x changes only the length of the gradient, but not the direction.

Logistic Regression. Hessian

Now let's calculate the second differential by fixing the increment dx_1 of the first one

$$
\begin{aligned}
d(df) = d(\langle \sigma(\langle a, x\rangle)a, dx_1\rangle) = \langle d(\sigma(\langle a, x\rangle))a, dx_1\rangle = \\
= \langle \sigma'(\langle a, x\rangle)d(\langle a, x\rangle)a, dx_1\rangle = \langle \sigma(\langle a, x\rangle)(1-\sigma(\langle a, x\rangle))\langle a, dx\rangle a, dx_1\rangle \\
= \sigma(\langle a, x\rangle)(1-\sigma(\langle a, x\rangle))\langle\langle dx, a\rangle a, dx_1\rangle = \\
= \sigma(\langle a, x\rangle)(1-\sigma(\langle a, x\rangle))(dx^\top aa^\top dx_1) \\
= \sigma(\langle a, x\rangle)(1-\sigma(\langle a, x\rangle))\langle aa^\top dx_1, dx\rangle.
\end{aligned}
$$

By representing $d^2 f$ in the standard form $\langle \nabla^2 f(x) \cdot dx_1, dx\rangle$, we get

$$
\nabla^2 f(x) = \sigma(\langle a, x\rangle)(1-\sigma(\langle a, x\rangle))aa^\top.
$$

Note that $\nabla^2 f$ is a peer matrix proportional to $aa^\top$ with the coefficient $\sigma(\langle a, x\rangle)(1-\sigma(\langle a, x\rangle)) \in (0, 0.25)$. The point x only affects the coefficient.

Derivative on scalar

Consider the function of the scalar argument α

$$\phi(\alpha) := f(x + \alpha p), \quad \alpha \in \mathbb{R},$$

$x, p \in \mathbb{R}^n$, $f : \mathbb{R}^n \to \mathbb{R}$ is a twice continuously differentiable function. Find the first and second derivatives of $\phi'(\alpha), \phi''(\alpha)$ and express them in terms of $\nabla f, \nabla^2 f$.

Derivative on scalar

Consider the function of the scalar argument α

$$\phi(\alpha) := f(x + \alpha p), \quad \alpha \in \mathbb{R},$$

$x, p \in \mathbb{R}^n$, $f : \mathbb{R}^n \to \mathbb{R}$ is a twice continuously differentiable function. Find the first and second derivatives of $\phi'(\alpha), \phi''(\alpha)$ and express them in terms of $\nabla f, \nabla^2 f$. **Solution.** It is important to remember that differentiation does not occur according to the standard vector x, but according to the scalar α with all the following properties

$$\begin{aligned} d\phi &= \{df = \langle \nabla f(y), dy \rangle\} = \langle \nabla f(x + \alpha p), d(x + \alpha p) \rangle = \\ &= \langle \nabla f(x + \alpha p), d(\alpha) p \rangle = \langle \nabla f(x + \alpha p), p \rangle d(\alpha). \end{aligned}$$

Note that we grafted the differential to the standard form $d\phi = \phi'(\alpha) \cdot d\alpha$, that is, the multiplier before $d\alpha$ is the derivative

$$\phi'(\alpha) = \langle \nabla f(x + \alpha p), p \rangle.$$

Derivative on scalar. Second

Now find the second derivative

$$
\begin{aligned}
d(\phi'(\alpha)) &= d\langle \nabla f(x+\alpha p), p\rangle = \{d(\nabla f(y)) = (\nabla^2 f(y))^\top dy\} = \\
&= \langle (\nabla^2 f(x+\alpha p))^\top d(x+p\alpha), p\rangle = \langle (\nabla^2 f(x+\alpha p))^\top p d\alpha, p\rangle = \\
&= \{\nabla^2 f(y) = (\nabla^2 f(y))^\top\} = \langle \nabla^2 f(x+\alpha p)p, p\rangle d\alpha.
\end{aligned}
$$

It turns out that

$$
\phi''(\alpha) = \langle \nabla^2 f(x+\alpha p)p, p\rangle.
$$