Mathematical Methods for Data Analysis
Seminar 3
Matrix-vector differentiation

As a rule, differentiable models are trained using gradient descent, and it is important for it to be able to calculate the gradient of the functional with respect to the model parameters. You can count the gradient coordinate-wise, and then look closely at the formulas and try to understand what it might look like in vector form. It is much easier to calculate the gradient directly – and for this, knowledge of gradients for basic functions and the basic rules of matrix-vector differentiation will help.

1 Derivation of the basic formulas

We introduce the following definitions:

- When function is a map from a vector to a scalar $f(x) : \mathbb{R}^n \to \mathbb{R}$

$$\nabla_x f(x) = \left[\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n}\right]^T$$

- When mapping a matrix to a number $f(A) : \mathbb{R}^{n\times m} \to \mathbb{R}$

$$\nabla_A f(A) = \left(\frac{\partial f}{\partial A_{ij}}\right)_{i,j=1}^{n,m}$$

Therefore, the gradient of the function with respect to the vector will be a vector, with respect to the matrix – matrix. Now let's practice differentiation:

Problem 1.1. *Let $a \in \mathbb{R}^n$ be a vector of parameters, and $x \in \mathbb{R}^n$ be a vector of variables. It is necessary to find the derivative of their scalar product with respect to the vector of variables $\nabla_x a^T x$.*

Solution.

$$\frac{\partial}{\partial x_i} a^T x = \frac{\partial}{\partial x_i} \sum_j a_j x_j = a_i,$$

so $\nabla_x a^T x = a$.

Note that $a^T x$ is a number, so $a^T x = x^T a$, hence,

$\nabla_x x^T a = a$.

■

Problem 1.2. *Now let's say $A \in \mathbb{R}^{n\times n}$. You need to find $\nabla_x x^T Ax$.*

Solution.

$$\frac{\partial}{\partial x_i} x^T Ax = \frac{\partial}{\partial x_i} \sum_j x_j (Ax)_j = \frac{\partial}{\partial x_i} \sum_j x_j \left(\sum_k a_{jk} x_k \right) = \frac{\partial}{\partial x_i} \sum_{j,k} a_{jk} x_j x_k =$$
$$= \sum_{j\neq i} a_{ji} x_j + \sum_{k\neq i} a_{ik} x_k + 2a_{ii} x_i = \sum_j a_{ji} x_j + \sum_k a_{ik} x_k = \sum_j (a_{ji} + a_{ij}) x_j.$$

So $\nabla_x x^T Ax = (A + A^T)x$.

■

Problem 1.3. *Let $A \in \mathbb{R}^{n\times n}$. You need to find $\nabla_A \det A$.*

Solution. Let's use Laplace's cofactor expansion:

$$\frac{\partial}{\partial A_{ij}} \det A = \frac{\partial}{\partial A_{ij}} \left[\sum_k (-1)^{i+k} A_{ik} M_{ik} \right] = (-1)^{i+j} M_{ij},$$

where M_{ik} is an additional minor of the matrix A. Also recall the formula for the elements of the inverse matrix

$$(A^{-1})_{ij} = \frac{1}{\det A} (-1)^{i+j} M_{ji}.$$

Substituting the expression for the additional minor, we get the answer $\nabla_A \det A = = (\det A) A^{-T}$.

■

Problem 1.4. *Let $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times n}$. You need to find $\nabla_A tr(AB)$.*

Solution.

$$\frac{\partial}{\partial A_{ij}}\text{tr}(AB) = \frac{\partial}{\partial A_{ij}}\sum_k (AB)_{kk} = \frac{\partial}{\partial A_{ij}}\sum_{k,l} A_{kl}B_{lk} = B_{ji}.$$

That is, the $\nabla_A \text{tr}(AB) = B^T$.

■

Problem 1.5. *Let $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{n\times m}$, $y \in \mathbb{R}^m$. You need to find $\nabla_A x^T Ay$.*

Solution. Taking advantage of the cyclic property of the matrix trace (for matrices of suitable size):

$$\text{tr}(ABC) = \text{tr}(BCA) = \text{tr}(CAB)$$

and the result of the previous task, we get

$$\nabla_A x^T Ay = \nabla_A \text{tr}(x^T Ay) = \nabla_A \text{tr}(Ayx^T) = xy^T.$$

■

Finally, we will learn how to count gradients for composition of functions. Let's say the functions $f : \mathbb{R}^n \to \mathbb{R}^m$ and $g : \mathbb{R}^m \to \mathbb{R}$ are given. Then the gradient of their composition can be calculated as

$$\nabla_x g\left(f(x)\right) = J_f^T(x)\nabla_z\, g(z)|_{z=f(x)}\,,$$

where $J_f(x) = \left(\frac{\partial f_i(x)}{\partial x_j}\right)_{i,j=1}^{m,n}$ — Jacobian matrix of the function f. If $m = 1$ and the function $g(z)$ has only one argument, then the formula is simplified:

$$\nabla_x g\left(f(x)\right) = g'(f(x))\nabla_x f(x).$$

Problem 1.6. *Calculate the gradient of the logistic loss function for a linear model from the parameters of this model:*

$$\nabla_w \log\left(1 + \exp(-y\langle w, x\rangle)\right).$$

Solution.

$$\begin{aligned}
&\nabla_w \log\left(1 + \exp(-y\langle w, x\rangle)\right) = \\
&\quad = \frac{1}{1 + \exp(-y\langle w, x\rangle)}\nabla_w\left(1 + \exp(-y\langle w, x\rangle)\right) = \\
&\quad = \frac{1}{1 + \exp(-y\langle w, x\rangle)}\exp(-y\langle w, x\rangle)\nabla_w\left(-y\langle w, x\rangle\right) = \\
&\quad = -\frac{1}{1 + \exp(-y\langle w, x\rangle)}\exp(-y\langle w, x\rangle)yx = \\
&\quad = \left\{\sigma(z) = \frac{1}{1 + \exp(-z)}\right\} = \\
&\quad = -\sigma(-y\langle w, x\rangle)yx
\end{aligned}$$

■

2 Solving the regression problem for the multidimensional case

Let's remember why we wanted to learn how to differentiate. In general, we have a sample of $\{(x_i, y_i)\}_{i=1}^{\ell}$, $x_i \in \mathbb{R}^d, y_i \in \mathbb{R}$ $i = \overline{1,\ell}$, and we want to find the best model parameters $a(x) = \langle w, x\rangle$ in terms of minimizing the loss function

$$Q(w) = (y - Xw)^T(y - Xw).$$

Here $X \in \mathbb{R}^{\ell \times d}$ is the matrix «objects-feature» for the training sample (also assume that examples are linearly independent), $y \in \mathbb{R}^{\ell}$ is the vector of values of the target variable on the training sample, $w \in \mathbb{R}^d$ is the vector of parameters. Let's write out the gradient of the loss function by w:

$$\begin{aligned}\nabla_w Q(w) &= \nabla_w[y^T y - y^T Xw - w^T X^T y + w^T X^T Xw] = \\ &= 0 - X^T y - X^T y + (X^T X + X^T X)w = 0.\end{aligned}$$

Thus, the desired parameter vector is expressed as

$$w = (X^T X)^{-1} X^T y.$$

Note that this is a general formula, and there is no need to output a formula for regression of the form $a(x) = Xw + w_0$, since we can always add a feature (column of the matrix X), which will always be equal to 1, and from the already derived formula we will find the parameter w_0.

Let's show why the found point is the minimum point if the matrix $X^T X$ is invertible. From the course of calculus, we know that if the Hesse matrix of a function is positively semi-defined at a point whose gradient is zero, then this point is the local minimum.

$$\nabla^2 Q(w) = 2X^T X.$$

It is necessary to understand whether the matrix $X^T X$ is positive semi-definite. We write down the definition of the positive semi-definiteness of the matrix $X^T X$:

$$z^T X^T Xz > 0,\ \forall z \in \mathbb{R}^d, z \neq 0.$$

We see that the square of the norm of the vector Xz is written here, that is, this expression will not be less than zero. If the matrix X has a «portrait» orientation (rows are not less than columns) and has a full rank (there are no linearly dependent columns), then the vector Xz cannot be zero, which means that

$$z^T X^T Xz = ||Xz||^2 > 0,\ \forall z \in \mathbb{R}^d, z \neq 0.$$

That is, $X^T X$ is a positive semi-definite matrix and invertible, and the solution exists. If there are fewer rows than columns, or X is not full-rank, then $X^T X$ is irreversible and the solution to w is ambiguous.