1 Basic linear algebra background

1.1 Vectors and matrices

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 $\mathbb{R}^{m \times n}$. That's it: [1]

$$
x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad x^T = \begin{bmatrix} x_1 & x_2 & \dots & x_n \end{bmatrix} \quad x \in \mathbb{R}^n, x_i \in \mathbb{R} \tag{1}
$$

Similarly, if $A \in \mathbb{R}^{m \times n}$ we denote transposition as $A^T \in \mathbb{R}^{n \times m}$:

$$
A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \quad A^T = \begin{bmatrix} a_{11} & a_{21} & \dots & a_{m1} \\ a_{12} & a_{22} & \dots & a_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \dots & a_{mn} \end{bmatrix} \quad A \in \mathbb{R}^{m \times n}, a_{ij} \in \mathbb{R}
$$

We will write $x \geq 0$ and $x \neq 0$ to indicate componentwise relationships

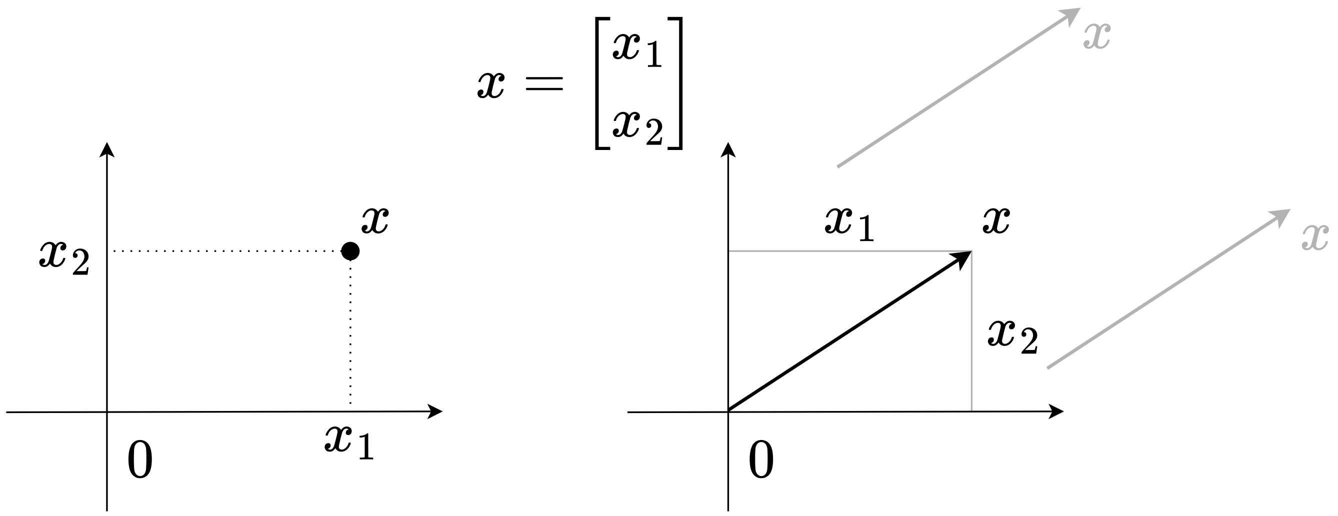


Figure 1: Equivivalent representations of a vector

A matrix is symmetric if $A = A^T$. It is denoted as $A \in \mathbb{S}^n$ (set of square symmetric matrices of dimension n). Note, that only a square matrix could be symmetric by definition.

A matrix $A \in \mathbb{S}^n$ is called **positive (negative) definite** if for all $x \neq 0 : x^T Ax > (<) 0$. We denote this as $A \succ (\prec) 0$. The set of such matrices is denoted as $\mathbb{S}^n_{++} (\mathbb{S}^n_{--})$

A matrix $A \in \mathbb{S}^n$ is called **positive (negative) semidefinite** if for all $x : x^T Ax \geq (\leq) 0$. We denote this as $A \succeq (\preceq) 0$. The set of such matrices is denoted as $\mathbb{S}^n_{+} (\mathbb{S}^n_{-})$

> 🤔 **Question**
>
> Is it correct, that a positive definite matrix has all positive entries?

1.2 Matrix and vector product

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:

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

This operation in a naive form requires $\mathcal{O}(n^3)$ arithmetical operations, where n is usually assumed as the largest dimension of matrices.

> 🤔 **Question**
>
> Is it possible to multiply two matrices faster, than $\mathcal{O}(n^3)$? How about $\mathcal{O}(n^2)$, $\mathcal{O}(n)$?

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:

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

Remember, that:

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

1.3 Norms and scalar products

Norm is a **qualitative measure of the smallness of a vector** and is typically denoted as $\|x\|$.

The norm should satisfy certain properties:

1. $\|\alpha x\| = |\alpha| \|x\|$, $\alpha \in \mathbb{R}$
2. $\|x + y\| \leq \|x\| + \|y\|$ (triangle inequality)
3. If $\|x\| = 0$ then $x = 0$

The distance between two vectors is then defined as

$$d(x, y) = \|x - y\|.$$

The most well-known and widely used norm is **Euclidean norm**:

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2},$$

which corresponds to the distance in our real life. If the vectors have complex elements, we use their modulus.

Euclidean norm, or 2-norm, is a subclass of an important class of p-norms:

$$\|x\|_p = \Big(\sum_{i=1}^{n} |x_i|^p\Big)^{1/p}.$$

There are two very important special cases. The infinity norm, or Chebyshev norm is defined as the element of the maximal absolute value:

$$\|x\|_\infty = \max_i |x_i|$$

L_1 norm (or **Manhattan distance**) which is defined as the sum of modules of the elements of x:

$$\|x\|_1 = \sum_i |x_i|$$

L_1 norm plays a very important role: it all relates to the **compressed sensing** methods that emerged in the mid-00s as one of the most popular research topics. The code for the picture below is available here: 🧑‍💻

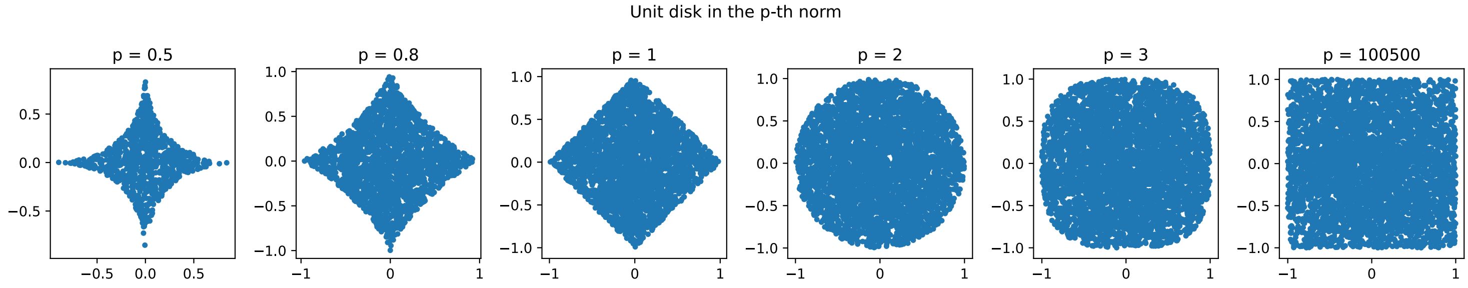


Figure 2: Balls in different norms on a plane

In some sense there is no big difference between matrices and vectors (you can vectorize the matrix), and here comes the simplest matrix norm **Frobenius** norm:

$$\|A\|_F = \left(\sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2\right)^{1/2}$$

Spectral norm, $\|A\|_2$ is one of the most used matrix norms (along with the Frobenius norm).

$$\|A\|_2 = \sup_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2},$$

It can not be computed directly from the entries using a simple formula, like the Frobenius norm, however, there are efficient algorithms to compute it. It is directly related to the **singular value decomposition** (SVD) of the matrix. It holds

$$\|A\|_2 = \sigma_1(A) = \sqrt{\lambda_{\max}(A^TA)}$$

where $\sigma_1(A)$ is the largest singular value of the matrix A.

> 🤔 **Question**
>
> Is it true, that all matrix norms satisfy the submultiplicativity property: $\|AB\| \leq \|A\|\|B\|$? Hint: consider Chebyshev matrix norm $\|A\|_C = \max\limits_{i,j} |a_{ij}|$.

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

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

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

> 🤔 **Question**
>
> Is there any connection between the norm $\|\cdot\|$ and scalar product $\langle \cdot, \cdot \rangle$?

> 🧑‍🏫 **Example**
>
> Prove, that you can switch the position of a matrix inside a scalar product with transposition: $\langle x, Ay \rangle = \langle A^Tx, y \rangle$ and $\langle x, yB \rangle = \langle xB^T, y \rangle$

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

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

Question

Is there any connection between the Frobenious norm $\|\cdot\|_F$ and scalar product between matrices $\langle\cdot,\cdot\rangle$?

Example

Simplify the following expression:

$$\sum_{i=1}^n \langle S^{-1}a_i, a_i\rangle,$$

where $S = \sum\limits_{i=1}^n a_i a_i^T, a_i \in \mathbb{R}^n, \det(S) \neq 0$

Solution

1. Let A be the matrix of columns vector a_i, therefore matrix A^T contains rows a_i^T
2. Note, that, $S = AA^T$ - it is the skeleton decomposition from vectors a_i. Also note, that A is not symmetric, while S, clearly, is.
3. The target sum is $\sum\limits_{i=1}^n a_i^T S^{-1} a_i$.
4. The most important part of this exercise lies here: we'll present this sum as the trace of some matrix M to use trace cyclic property.

$$\sum_{i=1}^n a_i^T S^{-1} a_i = \sum_{i=1}^n m_{ii},$$

   where m_{ii} - i-th diagonal element of some matrix M.
5. Note, that $M = A^T\left(S^{-1}A\right)$ is the product of 2 matrices, because i-th diagonal element of M is the scalar product of i-th row of the first matrix A^T and i-th column of the second matrix $S^{-1}A$. i-th row of matrix A^T, by definition, is a_i^T, while i-th column of the matrix $S^{-1}A$ is clearly $S^{>}-1a_i$.
   Indeed, $m_{ii} = a_i^T S^{-1} a_i$, then we can finish the exercise:

$$\begin{aligned}\sum_{i=1}^n a_i^T S^{-1} a_i &= \sum_{i=1}^n m_{ii} = \text{tr} M \\ &= \text{tr}\left(A^T S^{-1} A\right) = \text{tr}\left(AA^T S^{-1}\right) \\ &= \text{tr}\ \left(SS^{-1}\right) = \text{tr}\left(I\right) = n\end{aligned}$$

1.4 Eigenvalues, eigenvectors, and the singular-value decomposition

1.4.1 Eigenvalues

$\begin{pmatrix} 3 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix}$ нет с3

A scalar value λ is an eigenvalue of the $n \times n$ matrix A if there is a nonzero vector q such that

$$\boxed{Aq = \lambda q.}$$

Example

Consider a 2x2 matrix:

$$A = \begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix}$$

The eigenvalues of this matrix can be found by solving the characteristic equation:

$$\det(A - \lambda I) = 0$$

For this matrix, the eigenvalues are $\lambda_1 = 1$ and $\lambda_2 = 4$. These eigenvalues tell us about the scaling factors of the matrix along its principal axes.

The vector q is called an eigenvector of A. The matrix A is nonsingular if none of its eigenvalues are zero. The eigenvalues of symmetric matrices are all real numbers, while nonsymmetric matrices may have imaginary eigenvalues. If the matrix is positive definite as well as symmetric, its eigenvalues are all positive real numbers.

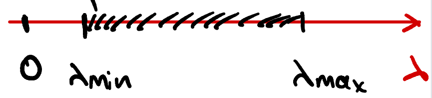
в обе стороны

набор сз матрицы – спектр матрицы

Theorem

$$\boxed{A \succeq 0 \Leftrightarrow \text{all eigenvalues of } A \text{ are } \geq 0}$$

$$\boxed{A \succ 0 \Leftrightarrow \text{all eigenvalues of } A \text{ are } > 0}$$



Proof

We will just prove the first point here. The second one can be proved analogously.

1. $\rightarrow$ Suppose some eigenvalue λ is negative and let x denote its corresponding eigenvector. Then

$$Ax = \lambda x \rightarrow x^T Ax = \lambda x^T x < 0$$

which contradicts the condition of $A \succeq 0$.

2. $\leftarrow$ For any symmetric matrix, we can pick a set of eigenvectors $v_1, \dots, v_n$ that form an orthogonal basis of $\mathbb{R}^n$. Pick any $x \in \mathbb{R}^n$.

$$\begin{aligned} x^T Ax &= (\alpha_1 v_1 + \ldots + \alpha_n v_n)^T A(\alpha_1 v_1 + \ldots + \alpha_n v_n) \\ &= \sum \alpha_i^2 v_i^T A v_i = \sum \alpha_i^2 \lambda_i v_i^T v_i \geq 0 \end{aligned}$$

here we have used the fact that $v_i^T v_j = 0$, for $i \neq j$.

Question

If a matrix has all positive eigenvalues, what can we infer about the matrix's definiteness?

$$Q^T \begin{pmatrix} \text{———} \\ \text{———} \\ \vdots \\ \text{———} \end{pmatrix}$$

Suppose $A \in S_n$, i.e., A is a real symmetric $n \times n$ matrix. Then A can be factorized as

$$\boxed{A = Q\Lambda Q^T,}$$

спектральное разложение матрицы A

$$Q \begin{pmatrix} q_1 & q_2 & \cdots & q_n \end{pmatrix}$$

where $Q \in \mathbb{R}^{n \times n}$ is orthogonal, i.e., satisfies $\boxed{Q^T Q = I,}$ and $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$. The (real) numbers λ_i are the eigenvalues of A and are the roots of the characteristic polynomial $\det(A - \lambda I)$. The columns of Q form an orthonormal set of eigenvectors of A. The factorization is called the spectral decomposition or (symmetric) eigenvalue decomposition of A. [2]

We usually order the eigenvalues as $\boxed{\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n}$. We use the notation $\lambda_i(A)$ to refer to the i-th largest eigenvalue of $A \in S$. We usually write the largest or maximum eigenvalue as $\lambda_1(A) = \lambda_{\max}(A)$, and the least or minimum eigenvalue as $\lambda_n(A) = \lambda_{\min}(A)$.

The largest and smallest eigenvalues satisfy

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

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

$$\boxed{\lambda_{\min}(A) x^T x \leq x^T Ax \leq \lambda_{\max}(A) x^T x} \quad \Big| : x^T x$$

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

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

число обусловленности матрицы A

Suppose $A \in \mathbb{R}^{m \times n}$ with rank $A = r$. Then A can be factored as

$$\boxed{A = U\Sigma V^T,} \qquad (A.12)$$

сингулярное разложение

$$\begin{pmatrix} \sigma_1 & & & 0 \\ & \sigma_2 & & \\ & & \ddots & \\ 0 & & & \sigma_r \end{pmatrix}$$

where $U \in \mathbb{R}^{m \times r}$ satisfies $\boxed{U^T U = I}$, $V \in \mathbb{R}^{n \times r}$ satisfies $\boxed{V^T V = I,}$ and Σ is a diagonal matrix with $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_r)$, such that

диаг. матрица

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0.$$

1.4.2 Singular value decomposition

This factorization is called the **singular value decomposition (SVD)** of A. The columns of U are called left singular vectors of A, the columns of V are right singular vectors, and the numbers σ_i are the singular values. The singular value decomposition can be written as сумма матриц ранга 1

$$\boxed{A = \sum_{i=1}^{r} \sigma_i \underbrace{u_i v_i^T}_{\text{outer product}},}$$

$A_{m\times n}$ Как сжать эту матрицу?

$\|A - A_r\|_2^2 \longrightarrow \min\limits_{A_r}$, $\operatorname{rg} A_r = r$

аппроксимация ранга r

where $u_i \in \mathbb{R}^m$ are the left singular vectors, and $v_i \in \mathbb{R}^n$ are the right singular vectors.

В матрице A m·n элементов

Аналитическое решение: truncated SVD ← теорема Эккарта-Юнга

Example

Consider a 2x2 matrix:

Для матрицы A_r нужно $m\cdot r + r\cdot r + r\cdot n = (m+n+r)\cdot r \approx (m+n) r \ll mn$

Обычно $r \ll \min(m,n)$

$$B = \begin{bmatrix} 4 & 0 \\ 0 & 2 \end{bmatrix}$$

$$\underset{m\times n}{A_r^*} = \underset{m\times r}{U_r} \cdot \underset{r\times r}{\Sigma_r} \cdot \underset{r\times n}{V_r^T} = U_r \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{pmatrix} V_r^T$$

The singular value decomposition of this matrix can be represented as:

$$B = U\Sigma V^T.$$

Where U and V are orthogonal matrices and Σ is a diagonal matrix with the singular values on its diagonal. For this matrix, the singular values are 4 and 2, which are also the eigenvalues of the matrix.

Example

Let $A \in \mathbb{R}^{m\times n}$, and let $q := \min\{m, n\}$. Show that

докажем

$$\boxed{\|A\|_F^2 = \sum_{i=1}^{q} \sigma_i^2(A),}$$

1) $\|A\|_F^2 = \langle A, A\rangle = tr(A^TA) \ominus$

2) $A = U\Sigma V^T \Rightarrow A^T = V\Sigma U^T$

$\Rightarrow \ominus\ tr(V\Sigma \underbrace{U^TU}_{I}\Sigma V^T) = tr(V\Sigma^2 V^T) \boxminus$

where $\sigma_1(A) \geq \ldots \geq \sigma_q(A) \geq 0$ are the singular values of matrix A. Hint: use the connection between Frobenius norm and scalar product and SVD.

$\boxminus\ tr(V^TV\Sigma^2) = tr(\Sigma^2) = \sum_{i=1} \sigma_i^2$.

Solution

Question

Suppose, matrix $A \in \mathbb{S}^n_{++}$. What can we say about the connection between its eigenvalues and singular values?

Question

How do the singular values of a matrix relate to its eigenvalues, especially for a symmetric matrix?

? ? ?

$tr(ABC) = tr(CAB) = tr(BCA)$

$\forall$ матриц A, B, C если согласованы размерности

1.4.3 Skeleton decomposition

Simple, yet very interesting decomposition is Skeleton decomposition, which can be written in two forms:

$$A = UV^T \quad A = \hat{C}\hat{A}^{-1}\hat{R}$$

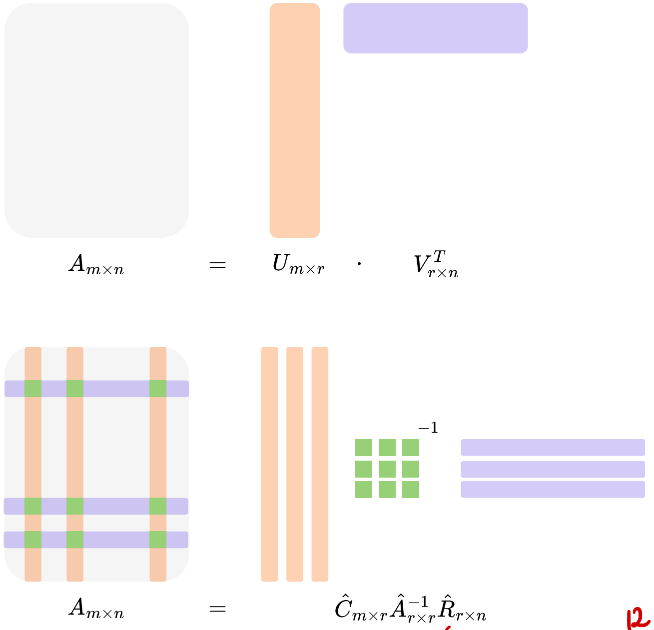
The latter expression refers to the fun fact: you can randomly choose r linearly independent columns of a matrix and any r linearly independent rows of a matrix and store only them with the ability to reconstruct the whole matrix exactly.

$A^TA = V\Sigma^2V^T$ ← спектр. разложение матрицы A^TA

$\Rightarrow \sigma^2(A)$ это $\lambda(A^TA)$ $\forall A$

если $A \in S^n_+$, то $\sigma(A) = \lambda(A)$

NLA.SKOLTECH.RU



$m \cdot n = 10^{15}$

$r \leq \min(m, n)$

$m = 10^9$

$n = 10^6$

$r = 10^3$

$(m+n)r \approx 10^{12}$

Figure 3: Illustration of Skeleton decomposition

> 🤔 **Question**
>
> How does the choice of columns and rows in the Skeleton decomposition affect the accuracy of the matrix reconstruction?

Use cases for Skeleton decomposition are:

- Model reduction, data compression, and speedup of computations in numerical analysis: given rank-r matrix with $r \ll n, m$ one needs to store $\mathcal{O}((n+m)r) \ll nm$ elements.
- Feature extraction in machine learning, where it is also known as matrix factorization
- All applications where SVD applies, since Skeleton decomposition can be transformed into truncated SVD form.

1.5 Canonical tensor decomposition

One can consider the generalization of Skeleton decomposition to the higher order data structure, like tensors, which implies representing the tensor as a sum of r primitive tensors.

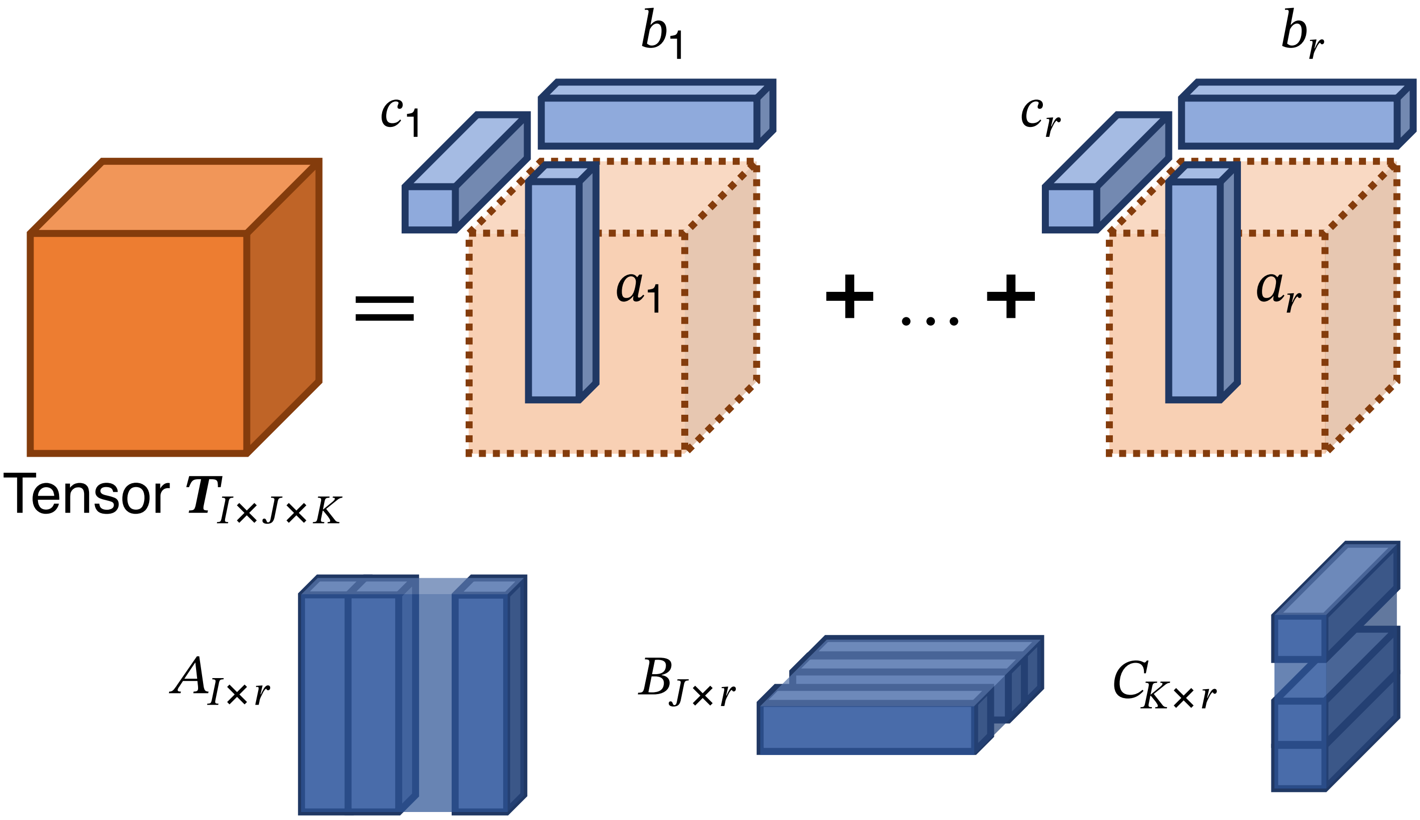


Figure 4: Illustration of Canonical Polyadic decomposition

Example

Note, that there are many tensor decompositions: Canonical, Tucker, Tensor Train (TT), Tensor Ring (TR), and others. In the tensor case, we do not have a straightforward definition of *rank* for all types of decompositions. For example, for TT decomposition rank is not a scalar, but a vector.

Question

How does the choice of rank in the Canonical tensor decomposition affect the accuracy and interpretability of the decomposed tensor?

1.6 Determinant and trace

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

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

The determinant has several appealing (and revealing) properties. For instance,

- $\det A = 0$ if and only if A is singular;
- $\det AB = (\det A)(\det B)$;
- $\det A^{-1} = \frac{1}{\det A}$.

Don't forget about the cyclic property of a trace for arbitrary matrices A, B, C, D (assuming, that all dimensions are consistent):

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

Example

For the matrix:

$$C = \begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix}$$

The determinant is $\det(C) = 6 - 1 = 5$, and the trace is $\operatorname{tr}(C) = 2 + 3 = 5$. The determinant gives us a measure of the volume scaling factor of the matrix, while the trace provides the sum of the eigenvalues.

🤔 **Question**

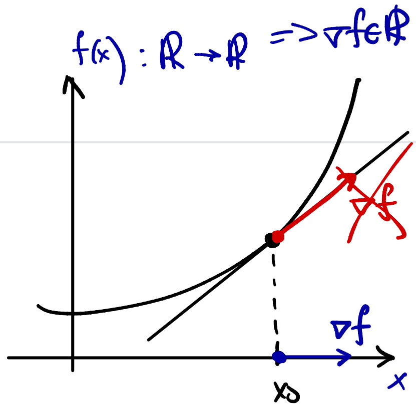
How does the determinant of a matrix relate to its invertibility?

🤔 **Question**

What can you say about the determinant value of a positive definite matrix? > 0

2 Optimization bingo

2.1 Gradient



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

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

named gradient of $f(x)$. This vector indicates the direction of the 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.

🧑‍🏫 **Example**

For the function $f(x, y) = x^2 + y^2$, the gradient is:

$$\nabla f(x, y) = \begin{bmatrix} 2x \\ 2y \end{bmatrix}$$

This gradient points in the direction of the steepest ascent of the function.

🤔 **Question**

How does the magnitude of the gradient relate to the steepness of the function?

2.2 Hessian

Теорема Шварца: если f - дважды непрерывно дифф-ма, то H_f - симм. матрица

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

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

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 $(H(f_1(x)), H(f_2(x)), \dots, H(f_m(x)))$.

🧑‍🏫 **Example**

For the function $f(x, y) = x^2 + y^2$, the Hessian is:

$$H_f(x, y) = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

This matrix provides information about the curvature of the function in different directions.

🤔 **Question**

How can the Hessian matrix be used to determine the concavity or convexity of a function?

2.3 Jacobian

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

$$
\underset{n \times m}{J_f} = f'(x) = \frac{df}{dx^T} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_2}{\partial x_1} & \dots & \frac{\partial f_m}{\partial x_1} \\ \frac{\partial f_1}{\partial x_2} & \frac{\partial f_2}{\partial x_2} & \dots & \frac{\partial f_m}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_1}{\partial x_n} & \frac{\partial f_2}{\partial x_n} & \dots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}_{n \times m}
$$

Example

For the function

$$
f(x, y) = \begin{bmatrix} x + y \\ x - y \end{bmatrix},
$$

the Jacobian is:

$$
J_f(x, y) = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}
$$

This matrix provides information about the rate of change of the function with respect to its inputs.

Question ГЕССИАН – ЭТО ЯКОБИАН ГРАДИЕНТА

How does the Jacobian matrix relate to the gradient for scalar-valued functions?

Question

Can we somehow connect those three definitions above (gradient, jacobian, and hessian) using a single correct statement?

2.4 Summary

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

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

2.5 Taylor approximations

Taylor approximations provide a way to approximate functions locally by polynomials. The idea is that for a smooth function, we can approximate it by its tangent (for the first order) or by its parabola (for the second order) at a point.

2.5.1 First-order Taylor approximation

The first-order Taylor approximation, also known as the linear approximation, is centered around some point x_0. If $f : \mathbb{R}^n \to \mathbb{R}$ is a differentiable function, then its first-order Taylor approximation is given by:

$$
\boxed{f_{x_0}^I(x) = f(x_0) + \nabla f(x_0)^T (x - x_0)}
$$

Where:

- $f(x_0)$ is the value of the function at the point x_0.
- $\nabla f(x_0)$ is the gradient of the function at the point x_0.

It is very usual to replace the $f(x)$ with $f_{x_0}^I(x)$ near the point x_0 for simple analysis of some approaches.

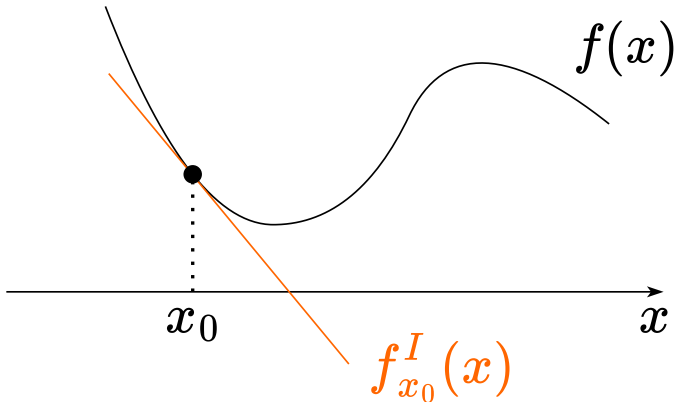


Figure 5: First order Taylor approximation near the point x_0

Example

For the function $f(x) = e^x$ around the point $x_0 = 0$, the first order Taylor approximation is:

$$
f_{x_0}^I(x) = 1 + x
$$

The second-order Taylor approximation is:

$$
f_{x_0}^{II}(x) = 1 + x + \frac{x^2}{2}
$$

These approximations provide polynomial representations of the function near the point x_0.

2.5.2 Second-order Taylor approximation

The second-order Taylor approximation, also known as the quadratic approximation, includes the curvature of the function. For a twice-differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, its second-order Taylor approximation centered at some point x_0 is:

$$
\boxed{f_{x_0}^{II}(x) = f(x_0) + \nabla f(x_0)^T (x - x_0) + \frac{1}{2}(x - x_0)^T \nabla^2 f(x_0)(x - x_0)}
$$

Where:

- $\nabla^2 f(x_0)$ is the Hessian matrix of f at the point x_0.

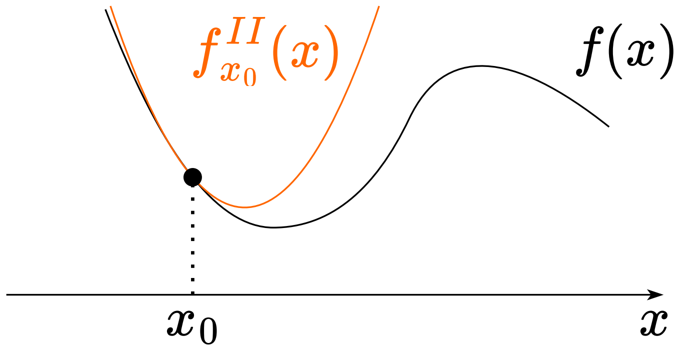


Figure 6: Second order Taylor approximation near the point x_0

When using the linear approximation of the function is not sufficient one can consider replacing the $f(x)$ with $f_{x_0}^{II}(x)$ near the point x_0. In general, Taylor approximations give us a way to locally approximate functions. The first-order approximation is a plane tangent to the function at the point x_0, while the second-order approximation includes the curvature and is represented by a parabola. These approximations are especially useful in optimization and numerical methods because they provide a tractable way to work with complex functions.

Example

Calculate first and second order Taylor approximation of the function $f(x) = \frac{1}{2}x^TAx - b^Tx + c$ вблизи т. x_0

Solution

$\cancel{\nabla f(x_0) = Ax_0 - b}$?

$\cancel{\nabla^2 f(x_0) = A}$?

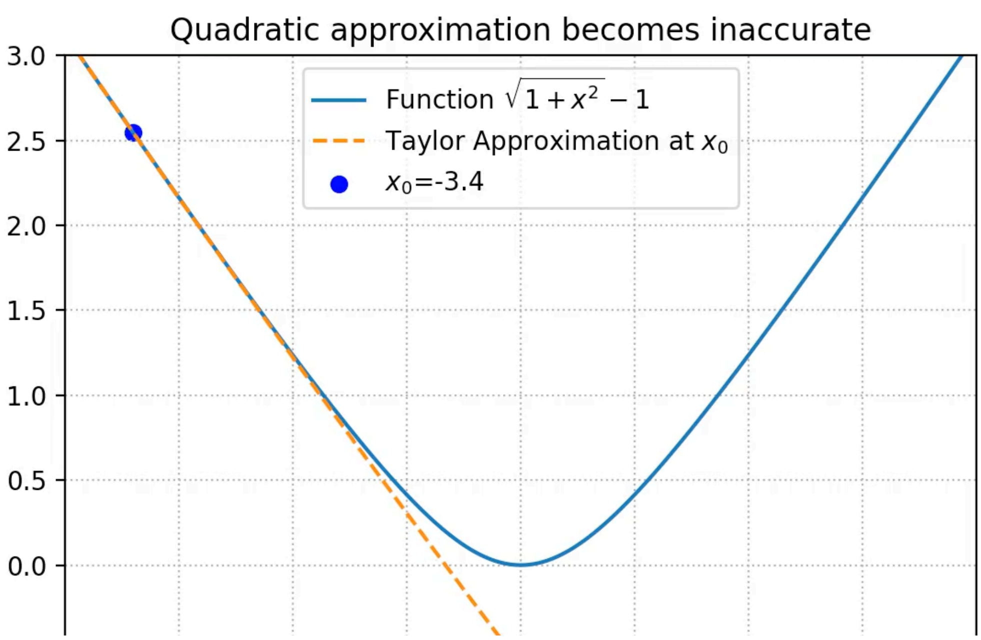
$\nabla f(x_0) = \frac{A+A^T}{2} x_0 - b$

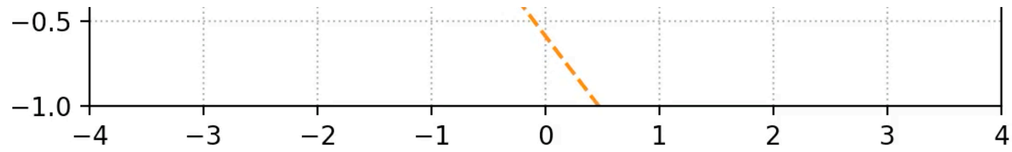
$\nabla^2 f(x_0) = \frac{A+A^T}{2}$

Question

Why might one choose to use a Taylor approximation instead of the original function in certain applications?

Note, that even the second-order approximation could become inaccurate very quickly. The code for the picture below is available here:





3 Derivatives

3.1 Naive approach

The basic idea of the 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 partial derivatives (k). Ignoring this simple rule tends to produce mistakes.

3.2 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. [3]

🤓 **Theorem**

Let $x \in S$ be an interior point of the set S, and let $D : U \to V$ be a linear operator. We say that the function f is differentiable at the point x with derivative D if for all sufficiently small $h \in U$ the following decomposition holds:

$$f(x + h) = f(x) + D[h] + o(\|h\|)$$

If for any linear operator $D : U \to V$ the function f is not differentiable at the point x with derivative D, then we say that f is not differentiable at the point x.

3.2.1 Differentials

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

$$df = \lim_{\|dx\| \to 0} \left(f(x+dx) - f(x) \right)$$

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

Then, if we have a 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 $d(df) = d^2 f(x)$

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

КАК ОЦЕНИТЬ ГЕССИАН?

1. Посчитать $df = \langle \dots, dx \rangle$
2. Посчитать $d^2 f = d(df)$, считая $dx = dx_1 = \text{const}$
3. Привести к виду

$d^2 f = \langle \nabla^2 f \, dx_1, dx \rangle$ ← матрица

3.2.2 Properties

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

- $dA = 0$
- $d(\alpha X) = \alpha(dX)$
- $d(AXB) = A(dX)B$
- $d(X+Y) = dX + dY$
- $d(X^T) = (dX)^T$
- $d(XY) = (dX)Y + X(dY)$
- $d\langle X, Y \rangle = \langle dX, Y \rangle + \langle X, dY \rangle$
- $d\left(\dfrac{X}{\phi}\right) = \dfrac{\phi dX - (d\phi)X}{\phi^2}$
- $d(\det X) = \det X \langle X^{-T}, dX \rangle$
- $d(\text{tr } X) = \langle I, dX \rangle$
- $df(g(x)) = \dfrac{df}{dg} \cdot dg(x)$
- $H = (J(\nabla f))^T$
- $d(X^{-1}) = -X^{-1}(dX)X^{-1}$

$df = \langle \dots, dx \rangle$

размерность такая же, как и у f ↑ ↑ градиент

$f: \mathbb{R}^m \to \mathbb{R} \Rightarrow df \in \mathbb{R}$

$d(\langle b, x \rangle) = \langle \overset{=0}{db}, x \rangle + \langle b, dx \rangle = \langle b, dx \rangle$

Example

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

Solution

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

$= \frac{1}{2} d(\langle Ax, x \rangle) - \langle b, dx \rangle + 0 = \frac{1}{2}\langle d(Ax), x \rangle + \frac{1}{2}\langle Ax, dx \rangle - \langle b, dx \rangle =$

$= \frac{1}{2}\underbrace{\langle A dx, x \rangle} + \frac{1}{2}\langle Ax, dx \rangle - \langle b, dx \rangle = \frac{1}{2}\underbrace{\langle A^T x, dx \rangle} + \frac{1}{2}\langle Ax, dx \rangle - \langle b, dx \rangle =$

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

если $A = A^T$, то $\nabla f = Ax - b$

2) Посчитаем $\nabla^2 f = ?$

$d(df) = d\left(\langle \frac{A+A^T}{2} x - b, dx_1 \rangle\right) = \langle \frac{A+A^T}{2} dx, dx_1 \rangle =$

$= \langle \frac{A+A^T}{2} dx_1, dx \rangle \Rightarrow \boxed{\nabla^2 f = \frac{A+A^T}{2}}$

Example

Find df, $\nabla f(x)$, if $f(x) = \ln\langle x, Ax \rangle$.

$\nabla^2 f = ?$

$A \in \mathbb{R}^{n \times n}$

более того, $A \in \mathbb{S}^n_{++}$

Solution

1. It is essential for A to be positive definite, because it is a logarithm argument. So, $A \in \mathbb{S}^n_{++}$ Let's find the differential first:

самостоятельно

$f = \ln x$

$df = \dfrac{dx}{x}$

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

$$= \frac{\langle Ax, dx \rangle + \langle x, Adx \rangle}{\langle x, Ax \rangle} = \frac{\langle Ax, dx \rangle + \langle A^T x, dx \rangle}{\langle x, Ax \rangle} = \frac{\langle (A + A^T)x, dx \rangle}{\langle x, Ax \rangle}$$

2. Note, that our main goal is to derive the form $df = \langle \cdot, dx \rangle$

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

Hence, the gradient is $\nabla f(x) = \dfrac{2Ax}{\langle x, Ax \rangle}$

$d^2 f = \langle d\left(\frac{2Ax}{\langle x, Ax \rangle}\right), dx_1 \rangle =$

$= \langle \dfrac{d(2Ax)\langle x, Ax \rangle - 2Ax \, d(\langle x, Ax \rangle)}{(\langle x, Ax \rangle)^2}, dx_1 \rangle =$

$= \langle \dfrac{\langle x, Ax \rangle 2Adx - 2Ax \langle 2Ax, dx \rangle}{(\langle x, Ax \rangle)^2}, dx_1 \rangle =$

$$= \left\langle \frac{\langle x, Ax\rangle 2A\cdot dx - 2Ax\cdot 2x^TA^T\cdot dx}{(\langle x, Ax\rangle)^2}, dx_1 \right\rangle = \left\langle \frac{(\langle x,Ax\rangle \cdot 2A - 4Ax\cdot (Ax)^T)dx}{(\langle x, Ax\rangle)^2}, dx_1 \right\rangle$$

$$\Rightarrow \boxed{\nabla^2 f = \frac{2\langle x,Ax\rangle\cdot A - 4(Ax)(Ax)^T}{(\langle x, Ax\rangle)^2}}$$

Example

Find $df, \nabla f(X)$, if $f(X) = \|AX - B\|_F = \sqrt{\langle AX-B, AX-B\rangle}$.

Solution

$$1.\ df = \frac{1}{2}(AX-B)^{-\frac{1}{2}} \cdot d(\langle AX-B, AX-B\rangle) =$$

$$= \frac{1}{\cancel{2}\|AX-B\|_F}\cdot \cancel{2}\langle AX-B, d(AX-B)\rangle =$$

$$= \frac{\langle AX-B, AdX\rangle}{\|AX-B\|_F} \Rightarrow \boxed{\nabla f = \frac{A^T(AX-B)}{\|AX-B\|_F}}$$

Example

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

САМОСТОЯТЕЛЬНО

Solution

$$d(\det X) = \det X \cdot \langle X^{-T}, dX\rangle = \det X \cdot \mathrm{tr}(X^{-1}dX)$$

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

НЕ ЗАБЫВАЕМ $\Rightarrow \boxed{\nabla f = S - X^{-T}}$

Example

Find the gradient $\nabla f(x)$ and hessian $\nabla^2 f(x)$, if $f(x) = \ln(1 + \exp\langle a, x\rangle)$

ЛОГ. РЕГРЕССИЯ

САМОСТОЯТЕЛЬНО

Solution

$$1.\ df = \frac{\exp\langle a,x\rangle \cdot \langle a, dx\rangle}{1+\exp\langle a,x\rangle} = \sigma(\langle a,x\rangle)\cdot\langle a, dx\rangle \Rightarrow \boxed{\nabla f = \sigma(\langle a,x\rangle)\cdot a}$$

$$\sigma(x) = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}$$

$$\sigma'(x) = \sigma(x)\cdot(1-\sigma(x))$$

$$2.\ d^2f = d(\sigma(\langle a,x\rangle)\langle a, dx_1\rangle) =$$

$$= \langle d\sigma\cdot a + \cancel{\sigma da}^{\,0}, dx_1\rangle \ [=]$$

$$d\sigma = d(\sigma(\langle a,x\rangle)) = \sigma(\langle a,x\rangle)(1-\sigma(\langle a,x\rangle))\cdot d(\langle a,x\rangle) = \sigma(1-\sigma)\langle a, dx\rangle$$

$$[=]\ \langle \sigma(1-\sigma)a^Tdx\cdot a, dx_1\rangle = \langle \sigma(1-\sigma)\overbrace{a\cdot a^T}^{\text{МАТРИЦА}}\underbrace{a^Tdx}_{\text{ЧИСЛО}}, dx_1\rangle =$$

$$= \langle \sigma(1-\sigma)aa^T dx_1, dx\rangle$$

$$\Rightarrow \boxed{\nabla^2 f = \sigma(\langle a,x\rangle)(1-\sigma(\langle a,x\rangle))\cdot aa^T}$$

4 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
- Introduction to Applied Linear Algebra – Vectors, Matrices, and Least Squares - book by Stephen Boyd & Lieven Vandenberghe.
- Numerical Linear Algebra course at Skoltech

Footnotes

1. A full introduction to applied linear algebra can be found in Introduction to Applied Linear Algebra – Vectors, Matrices, and Least Squares - book by Stephen Boyd & Lieven Vandenberghe, which is indicated in the source. Also, a useful refresher for linear algebra is in Appendix A of the book Numerical Optimization by Jorge Nocedal Stephen J. Wright.↩
2. A good cheat sheet with matrix decomposition is available at the NLA course website.↩
3. The most comprehensive and intuitive guide about the theory of taking matrix derivatives is presented in these notes by Dmitry Kropotov team.↩

Пример: $f(X) = \mathrm{tr}\,X \qquad \nabla f = ?$

1) $\mathrm{tr}\,X = \mathrm{tr}(I^T X) = \langle I, X\rangle$

2) $df = d(\langle I, X\rangle) = \langle I, dX\rangle \Rightarrow \boxed{\nabla f = I}$