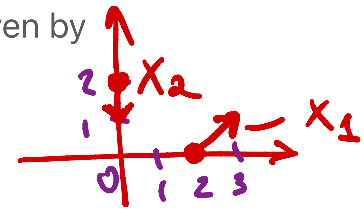
Useful definitions and notations

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

$\langle X_1, X_2 \rangle$

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.

$X \in R^{m \times n}$

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

$\langle X, X \rangle =$

$$
\boxed{\langle \underset{m\times n}{X}, \underset{m\times n}{Y} \rangle = \operatorname{tr}(\underbrace{X^\top Y}_{n\times m \ \ m\times n})} = \sum_{i=1}^m \sum_{j=1}^n X_{ij} Y_{ij} = \operatorname{tr}(Y^\top X) = \langle Y, X \rangle
$$

$= \|X\|_F^2$

след матрицы - сумма диаг. элементов

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

спектральное разл.

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

$A = U \Lambda U^*$

U - орт.

$UU^* = I$

$\operatorname{diag}(\lambda_1 \ldots \lambda_n)$

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

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

The largest and smallest eigenvalues satisfy

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

$Ax = \lambda x$

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

$x^T A x = x^T \lambda x$

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

$x^T A x = \lambda x^T x$

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

$\|A\|_p = \sup_{x \neq 0} \frac{\langle x, Ax \rangle}{\|x\|_p}$

" $\langle x, Ax \rangle$

$A \succ (\succeq) 0.$ $A = \begin{pmatrix} 1000 & 0 \\ 0 & 1 \end{pmatrix}$ Большое $\left(\begin{matrix} 10+ \\ 100+ \end{matrix}\right)$

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

$A^{-1} = \begin{pmatrix} \frac{1}{1000} & 0 \\ 0 & 1 \end{pmatrix}$

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

$\|A\|_2 = \sigma_{max}(A) = 1000$

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:

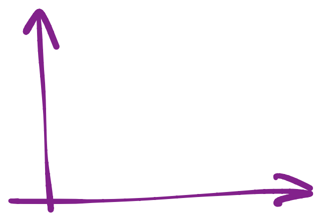
$O(n^3)$ - наивный алгоритм

$$C = AB$$

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

$m{\times}n \; n{\times}p$

$O(n)$



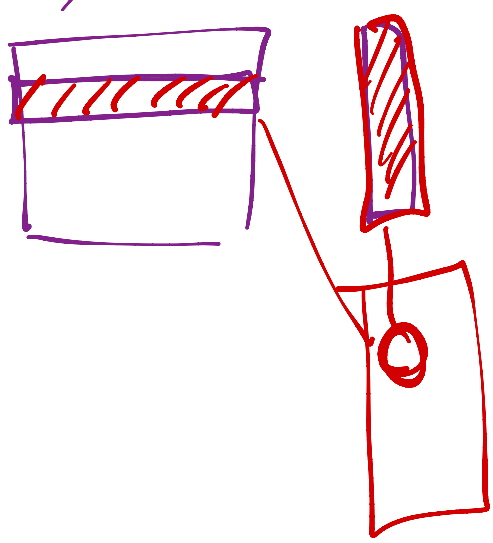
$$\boxed{\underbrace{c_{ij}}_{n^2} = \sum_{k=1}^{n} a_{ik} b_{kj}}$$

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:

$$z = Ax \quad O(n^2)$$

is given by:

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



Finally, just to remind:

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

$x^T A y \qquad (A^T x)^T y$

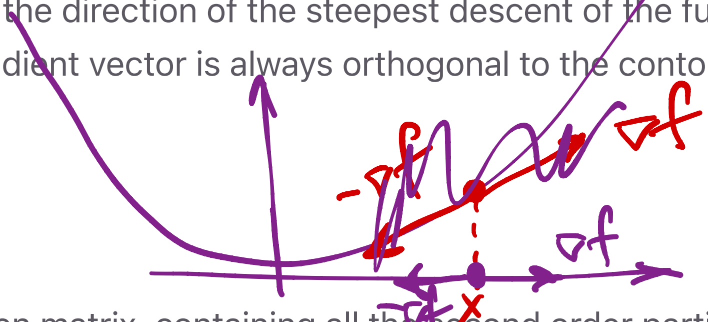
$\qquad\qquad\quad x^T A y$

Gradient

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

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

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.



$f(x)\ \mathbb{R} \to \mathbb{R}$

$\nabla f \in \mathbb{R}$

Hessian

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

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

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 $\left(H\left(f_1(x)\right), H\left(f_2(x)\right), \ldots, H\left(f_m(x)\right)\right)$.

Jacobian

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

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

$m \times n$

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)
$\mathbb{R}^n$	$\mathbb{R}^m$	$\mathbb{R}^{m\times n}$	$\frac{\partial f_i}{\partial x_j}$ (jacobian)
$\mathbb{R}^{m\times n}$	$\mathbb{R}$	$\mathbb{R}^{m\times n}$	$\frac{\partial f}{\partial x_{ij}}$

General concept

Naive approach

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

$\nabla f = ?$

Matrix notation of a function

$$\boxed{f(x) = c^\top x}$$

Matrix notation of a gradient

$$\boxed{\nabla f(x) = c}$$

Scalar notation of a function

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

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

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.

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.

$f: \mathbb{R}^n \to \mathbb{R}$

$f(x+dx) - f(x) \approx df$

Differentials

After obtaining the differential notation of df we can retrieve the gradient using following formula: $f = c^T x = \langle c, x \rangle$

$\nabla f = c$

$dx = dx_1$

$df = d(\langle c, x \rangle) =$
$= \langle c, dx \rangle$ ∇f

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

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 $d(df) = d^2 f(x)$

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

Properties

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

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

X, dX одной природы

$\langle X, Y \rangle = \text{tr}(X^T Y)$

$\text{tr} X = \text{tr}\, I^T X = \langle I, X \rangle$

$d(\langle I, X \rangle) = \langle I, dX \rangle \Rightarrow \nabla f = I$

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

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

Matrix calculus

1. Find the derivatives of $f(x) = Ax, \quad \nabla_x f(x) = ?, \nabla_A f(x) = ?$
2. Find $\nabla f(x)$, if $f(x) = c^T x$.
3. Find $\nabla f(x)$, if $f(x) = \frac{1}{2} x^T Ax + b^T x + c$.
4. Find $\nabla f(x), f''(x)$, if $f(x) = -e^{-x^T x}$.
5. Find the gradient $\nabla f(x)$ and hessian $f''(x)$, if $f(x) = \frac{1}{2} \|Ax - b\|_2^2$.
6. Find $\nabla f(x)$, if $f(x) = \|x\|_2, x \in \mathbb{R}^p \setminus \{0\}$.
7. Find $\nabla f(x)$, if $f(x) = \|Ax\|_2, x \in \mathbb{R}^p \setminus \{0\}$.
8. Find $\nabla f(x), f''(x)$, if $f(x) = \frac{-1}{1 + x^\top x}$.
9. Calculate $df(x)$ and $\nabla f(x)$ for the function $f(x) = \log(x^\top \mathrm{A} x)$.
10. Find $f'(X)$, if $f(X) = \det X$

    Note: here under $f'(X)$ assumes first order approximation of $f(X)$ using Taylor series: $f(X + \Delta X) \approx f(X) + \mathbf{tr}(f'(X)^\top \Delta X)$
11. Find $f''(X)$, if $f(X) = \log \det X$

    Note: here under $f''(X)$ assumes second order approximation of $f(X)$ using Taylor series: $f(X + \Delta X) \approx f(X) + \mathbf{tr}(f'(X)^\top \Delta X) + \frac{1}{2}\mathbf{tr}(\Delta X^\top f''(X) \Delta X)$
12. Find gradient and hessian of $f : \mathbb{R}^n \to \mathbb{R}$, if:

$$f(x) = \log \sum_{i=1}^{m} \exp(a_i^\top x + b_i), \quad a_1, \ldots, a_m \in \mathbb{R}^n; \quad b_1, \ldots, b_m \in \mathbb{R}$$

13. What is the gradient, Jacobian, Hessian? Is there any connection between those three definitions?
14. Calculate: $\frac{\partial}{\partial X} \sum \text{eig}(X), \quad \frac{\partial}{\partial X} \prod \text{eig}(X), \quad \frac{\partial}{\partial X} \text{tr}(X), \quad \frac{\partial}{\partial X} \det(X)$
15. Calculate the Frobenious norm derivative: $\frac{\partial}{\partial X} \|X\|_F^2$
16. Calculate the gradient of the softmax regression $\nabla_\theta L$ in binary case ($K = 2$) n - dimensional objects:

Пример:

$x \in \mathbb{R}^n$

$$f(x) = \ln\langle x, Ax\rangle$$

1. $df = ?$
2. $\nabla f = ?$

Решение:

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

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

$$= \frac{\langle Ax, dx\rangle + \langle x, Adx\rangle}{\langle x, Ax\rangle} =$$

$$\boxed{df = \langle \ldots, dx\rangle}$$

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

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

$$df = \left\langle \frac{(A+A^T)x}{\langle x, Ax\rangle}, dx_1\right\rangle$$

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

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

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

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

$$x\,((A+A^T)x)^T dx =$$

$$= x x^T (A+A^T) dx$$

$$x^T A x \overset{I}{\vee} dx - x \cdot x^T A dx - x \cdot x^T A^T dx$$

$$\left\langle \frac{(A+A^T)\left(\underline{x^T A x} I - \underset{n\times 1}{x}\underset{1\times n}{x^T}\underset{nn}{A} - x x^T A^T\right)dx}{(\langle x, Ax\rangle)^2}, dx_1 \right\rangle$$

$$(x x^T A + x x^T A^T) = (x x^T (A + A^T))$$

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

$$f(x) = \frac{1}{2}\underbrace{x^T}_{1\times n}\underbrace{Ax}_{n\times 1} + b^T x + c \qquad df = ? \quad \nabla f = ?$$

$x \in \mathbb{R}^n$ $\quad \langle x, Ax\rangle$

$$df = d\left(\frac{1}{2}\langle x, Ax\rangle + \langle b, x\rangle + c\right) \qquad A \in \mathbb{R}^{m\times n}$$

$$= \frac{1}{2}\langle dx, Ax\rangle + \frac{1}{2}\langle x, A\,dx\rangle + \langle b, dx\rangle \qquad b \in \mathbb{R}^n, \; c \in \mathbb{R}$$

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

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

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

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

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

$$d^2 f = \langle \quad dx_1, dx\rangle$$

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

$$= \left\langle \boxed{\frac{1}{2}(A+A^T)} dx_1, dx\right\rangle$$

$$(A+A^T)^T = A^T + A$$

$$\boxed{\nabla^2 f = \frac{1}{2}(A+A^T)}$$

$$
h_\theta(x) = \begin{bmatrix} P(y=1|x;\theta) \\ P(y=2|x;\theta) \\ \vdots \\ P(y=K|x;\theta) \end{bmatrix} = \frac{1}{\sum_{j=1}^K \exp(\theta^{(j)\top} x)} \begin{bmatrix} \exp(\theta^{(1)\top} x) \\ \exp(\theta^{(2)\top} x) \\ \vdots \\ \exp(\theta^{(K)\top} x) \end{bmatrix}
$$

$$
L(\theta) = -\left[\sum_{i=1}^n (1-y^{(i)})\log(1-h_\theta(x^{(i)})) + y^{(i)}\log h_\theta(x^{(i)})\right]
$$

17 Find $\nabla f(X)$, if $f(X) = \operatorname{tr} AX$

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

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

20 Find the gradient $\nabla f(x)$ and hessian $f''(x)$, if

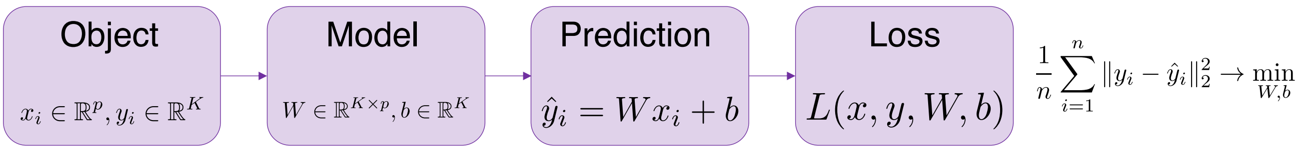
$$
f(x) = \ln\left(1 + \exp\langle a, x\rangle\right)
$$

21 Find the gradient $\nabla f(x)$ and hessian $f''(x)$, if $f(x) = \frac{1}{3}\|x\|_2^3$

22 Calculate $\nabla f(X)$, if $f(X) = \|AX - B\|_F, X \in \mathbb{R}^{k\times n}, A \in \mathbb{R}^{m\times k}, B \in \mathbb{R}^{m\times n}$

23 Calculate the derivatives of the loss function with respect to parameters $\frac{\partial L}{\partial W}, \frac{\partial L}{\partial b}$ for the single object x_i (or, $n = 1$)

Learning



24 Find the gradient $\nabla f(x)$ and hessian $f''(x)$, if $f(x) = \langle x, x\rangle^{\langle x, x\rangle}, x \in \mathbb{R}^p \setminus \{0\}$

25 Find the gradient $\nabla f(x)$ and hessian $f''(x)$, if $f(x) = \frac{\langle Ax, x\rangle}{\|x\|_2^2}, x \in \mathbb{R}^p \setminus \{0\}, A \in \mathbb{S}^n$

26 Find the gradient $\nabla f(x)$ and hessian $f''(x)$, if $f(x) = \frac{1}{2}\|A - xx^\top\|_F^2, A \in \mathbb{S}^n$

27 Find the gradient $\nabla f(x)$ and hessian $f''(x)$, if $f(x) = \|xx^\top\|_2$

28 Find the gradient $\nabla f(x)$ and hessian $f''(x)$, if

$$
f(x) = \frac{1}{n}\sum_{i=1}^n \log\left(1 + \exp(a_i^\top x)\right) + \frac{\mu}{2}\|x\|_2^2,\ a_i \in \mathbb{R}^n,\ \mu > 0.
$$

29 Match functions with their gradients:

- [ ] $f(\mathrm{X}) = \mathrm{Tr}\mathrm{X}$

- [ ] $f(\mathrm{X}) = \mathrm{Tr}\mathrm{X}^{-1}$

- [ ] $f(\mathrm{X}) = \det \mathrm{X}$

- [ ] $f(\mathrm{X}) = \ln \det \mathrm{X}$

a $\nabla f(\mathrm{X}) = \mathrm{X}^{-1}$
b $\nabla f(\mathrm{X}) = \mathrm{I}$
c $\nabla f(\mathrm{X}) = \det(\mathrm{X}) \cdot (\mathrm{X}^{-1})^\top$
d $\nabla f(\mathrm{X}) = -\left(\mathrm{X}^{-2}\right)^\top$

30 Calculate the first and the second derivative of the following function $f : S \to \mathbb{R}$ $f(t) = \det(A - tI_n)$, where $A \in \mathbb{R}^{n \times n}$, $S := \{t \in \mathbb{R} : \det(A - tI_n) \neq 0\}$.

31 Find the gradient $\nabla f(x)$, if $f(x) = \mathrm{tr}\left(AX^2BX^{-\top}\right)$.

$$f(X) = \langle S, X \rangle - \ln\det X \qquad \nabla f = ?$$

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

$$\det X > 0$$

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

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

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

$$\boxed{\nabla f = S - X^{-T}} = \left(\frac{\partial f}{\partial x_{ij}} \right)_{ij = \overline{1,n}}$$