FEDERAL STATE AUTONOMOUS
EDUCATIONAL INSTITUTION OF HIGHER EDUCATION
MOSCOW INSTITUTE OF PHYSICS AND TECHNOLOGY
(STATE UNIVERSITY)
PHYSTECH-SCHOOL OF APPLIED MATHEMATICS AND COMPUTER SCIENCE

Homework.

«Optimization»

3-rd year student, group B05-003
Kreinin Matvei

Moscow, 2022

I need to do: 2.3, 3.1, 3.2, 3.5, 4.3, 4.7, 5-th week. It's only 11 problems...

Содержание

1 Matrix calculus

1.1 Problem № 1

Find the gradient $\nabla f(x)$ and hessian $f''(x)$, if $f(x) = \frac{1}{2}\|Ax - b\|_2^2$

Solution:

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

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

$$\nabla f(x) = A^T(Ax - b)$$

$$df(x) = \langle A^T(Ax - b), dx\rangle$$

$$d^2f(x) = \langle d(A^T(Ax_2 - b), dx_1\rangle = \langle A^TAdx_2, dx_1\rangle = \langle dx_1, A^TAdx_2\rangle$$

$$d^2f(x) = \langle A^TAdx_1, dx_2\rangle$$

Answer: $\nabla f(x) = A^T(Ax - b),\ f''(x) = A^TA$

1.2 Problem № 2

Find gradient and hessian of $f : \mathbb{R}^n \to \mathbb{R}$, if:

$f(x) = \log\left(\sum\limits_{i=1}^{m} exp(a_i^Tx + b_i)\right), a_1, ..., am \in \mathbb{R}^n; b_1, ..., b_m \in \mathbb{R}$

Solution:

$$df(x) = \frac{d\left(\sum\limits_{i=1}^{m} exp(a_i^Tx + b_i)\right)}{\sum\limits_{i=1}^{m} exp(a_i^Tx + b_i)} = \frac{\sum\limits_{i=1}^{m} exp(a_i^Tx + b_i)a_i^Tdx}{\sum\limits_{i=1}^{m} exp(a_i^Tx + b_i)} = \frac{\langle\sum\limits_{i=1}^{m} exp(a_i^Tx + b_i)a_i^T, dx\rangle}{\sum\limits_{i=1}^{m} exp(a_i^Tx + b_i)}$$

$$\nabla f(x) = \frac{\sum\limits_{i=1}^{m} exp(a_i^Tx + b_i)a_i^T}{\sum\limits_{i=1}^{m} exp(a_i^Tx + b_i)}$$

$$d^2f(x) = \langle d\left(\frac{\sum\limits_{i=1}^{m} exp(a_i^T x_2 + b_i)a_i}{\sum\limits_{i=1}^{m} exp(a_i^T x_2 + b_i)}\right), dx_1\rangle$$

$$d^2f(x) = \langle\left(\frac{\sum\limits_{i=1}^{m} exp(a_i^T x_2 + b_i)a_i a_i^T}{\sum\limits_{i=1}^{m} exp(a_i^T x_2 + b_i)} + (\frac{\sum\limits_{i=1}^{m} exp(a_i^T x_2 + b_i)a_i a_i^T}{\left(\sum\limits_{i=1}^{m} exp(a_i^T x_2 + b_i)\right)^2}\right) dx_2, dx_1\rangle$$

$$d^2f(x) = \langle dx_1, \left(\frac{\sum\limits_{i=1}^{m} exp(a_i^T x_2 + b_i)a_i a_i^T}{\sum\limits_{i=1}^{m} exp(a_i^T x_2 + b_i)} + (\frac{\sum\limits_{i=1}^{m} exp(a_i^T x_2 + b_i)a_i a_i^T}{\left(\sum\limits_{i=1}^{m} exp(a_i^T x_2 + b_i)\right)^2}\right) dx_2\rangle$$

$$d^2f(x) = \langle\left(\frac{\sum\limits_{i=1}^{m} exp(a_i^T x_2 + b_i)a_i^T a_i}{\sum\limits_{i=1}^{m} exp(a_i^T x_2 + b_i)} + (\frac{\sum\limits_{i=1}^{m} exp(a_i^T x_2 + b_i)a_i^T a_i}{\left(\sum\limits_{i=1}^{m} exp(a_i^T x_2 + b_i)\right)^2}\right) dx_1, dx_2\rangle$$

Answer: $\nabla f(x) = \frac{\sum\limits_{i=1}^{m} exp(a_i^T x+b_i)a_i^T}{\sum\limits_{i=1}^{m} exp(a_i^T x+b_i)}$; $f''(x) = \left(\frac{\sum\limits_{i=1}^{m} exp(a_i^T x_2+b_i)a_i^T a_i}{\sum\limits_{i=1}^{m} exp(a_i^T x_2+b_i)} + (\frac{\sum\limits_{i=1}^{m} exp(a_i^T x_2+b_i)a_i^T a_i}{\left(\sum\limits_{i=1}^{m} exp(a_i^T x_2+b_i)\right)^2}\right)$

1.3 Problem № 3

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 $x_i(or, n = 1)$

Solution:

$$L = \frac{1}{n}\sum_{i=1}^{n} \|y_i - \tilde{y}\|^2 = \frac{1}{n}\sum_{i=1}^{n}\langle y_i - \tilde{y}, y_i - \tilde{y}\rangle = \frac{1}{n}\sum_{i=1}^{n}\langle y_i - Wx_i - b, y_i - Wx_i - b\rangle$$

$$dL(dW) = \frac{1}{n}\sum_{i=1}^{n}\langle y_i - Wx_i - b, -dWx_i,\rangle + \langle y_i - Wx_i - b, -dWx_i\rangle$$

$$dL(dW) = \frac{2}{n}\sum_{i=1}^{n}\langle -dWx_i, y_i - Wx_i - b\rangle = -\frac{2}{n}\sum_{i=1}^{n}\langle (y_i - Wx_i - b)x_i^T, dW\rangle$$

$$dL(db) = \frac{1}{n}\sum_{i=1}^{n}\langle -db, y_i - Wx_i - b\rangle + \langle y_i - Wx_i - b, -db\rangle = -\frac{2}{n}\sum_{i=1}^{n}\langle y_i - Wx_i - b, db\rangle$$

Answer: $\frac{\partial L}{\partial W} = -\frac{2}{n}\sum\limits_{i=1}^{n}(y_i - Wx_i - b)x_i^T$; $\frac{\partial L}{\partial b} = -\frac{2}{n}\sum\limits_{i=1}^{n} y_i - Wx_i - b$

1.4 Problem № 4

Calculate:

$$\frac{\partial}{\partial X}\sum \text{eig(X)}, \frac{\partial}{\partial X}\prod \text{eig(X)}, \frac{\partial}{\partial X}tr(X), \frac{\partial}{\partial X}\text{det(X)}$$

Solution:

$$\frac{\partial}{\partial X}\sum \text{eig(X)} = \frac{\partial}{\partial X}tr(X)$$

$$d(tr(X)) = tr(dX) = tr(I^T, dX) = \langle I, dX\rangle$$

$$\frac{\partial}{\partial X}\prod \text{eig(X)} = \frac{\partial}{\partial X}\text{det(X)}$$

$$det(X) = \sum_{i=1}^{n} x_{ij}M_{ij}; \frac{\partial X}{\partial x_{ij}} = \frac{\partial \sum\limits_{i=1}^{n} x_{ij}M_{ij}}{\partial x_{ij}} = M_{ij}$$

Т.к. $x_{ij}^{-1} = \frac{M_{ji}}{detX}$, тогда

$$\frac{\partial(det(X))}{\partial X} = det(X)X^{-T}$$

Answer: $\frac{\partial}{\partial X}\sum \text{eig(X)} = \frac{\partial}{\partial X}tr(X) = I$; $\frac{\partial}{\partial X}\prod \text{eig(X)} = \frac{\partial}{\partial X}\text{det(X)} = det(X)X^{-T}$

1.5 Problem № 5

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

$$df(t) = det(A - t \cdot I)\langle (A - t \cdot I)^{-T}, -Idt\rangle = -det(A - t \cdot I)\langle (A - t \cdot I)^{-T}, -Idt\rangle$$

$$df(t) = -f(t) \cdot tr\left((A - t \cdot I)^{-1}\right) dt$$

Okay, let's try to calculate second derivative of that nice function!

$$d^2 f(t) = -d\left(f(t) \cdot tr\left((A - t \cdot I)^{-1}\right) dt_1\right)$$

$$d^2 f(t) = -\nabla f(t) \cdot tr\left((A - t \cdot I)^{-1}\right) dt_2 \cdot dt_1 - f(t)\langle I, -(A - t \cdot I)^{-1}(-Idt_2)(A - t \cdot I)^{-1}\rangle dt_1$$

And then we get:

$$d^2 f(t) = -\left(\nabla f(t) \cdot tr\left((A - t \cdot I)^{-1}\right) + f(t) \cdot tr\left(((A - t \cdot I)^{-2})^T\right)\right) \cdot dt_1 \cdot dt_2$$

Answer: $\nabla f(t) = -f(t) \cdot tr\left((A - t \cdot I)^{-1}\right)$
$f''(t) = -\left(\nabla f(t) \cdot tr\left((A - t \cdot I)^{-1}\right) + f(t) \cdot tr\left(((A - t \cdot I)^{-2})^T\right)\right)$

1.6 Problem № 6

Find the gradient $\nabla f(x)$, if $f(x) = tr(AX^2BX^{-T})$.
Solution:

$$df(X) = d(tr(AX^2BX^{-T}) = \langle I, d(AX^2BX^{-T})\rangle$$

$$df(X) = \langle I, A(XdX + dXX)BX^{-T} - AX^2BX^{-T}dX^TX^{-T}\rangle$$

$$df(X) = \langle (BX^{-T}AX)^T, dX\rangle + \langle (XBX^{-T}A)^T, dX\rangle + \langle (X^{-T}AX^2BX^{-T})^T, dX^T\rangle$$

$$df(X) = \langle X^TA^TX^{-1}B^T, dX\rangle + \langle A^TX^{-1}B^TX^T, dX\rangle - \langle X^{-1}B^TX^TX^TA^TX^{-1}, dX^T\rangle$$

$$df(X) = \langle 2X^TA^TX^{-1}B^T, dX\rangle - \langle X^{-1}B^TX^TX^TA^TX^{-1}, dX^T\rangle$$

$$df(X) = \langle X^TA^TX^{-1}B^T + A^TX^{-1}B^TX^T, dX\rangle - \langle (X^{-1}B^TX^TX^TA^TX^{-1})^T, dX\rangle$$

$$df(X) = \langle X^TA^TX^{-1}B^T + A^TX^{-1}B^TX^T - X^{-T}AXXBX^{-T}, dX\rangle$$

Answer: $\nabla f(x) = X^TA^TX^{-1}B^T + A^TX^{-1}B^TX^T - X^{-T}AXXBX^{-T}$

2 Automatic differentiation

```
[ ]: import jax
     import numpy

     from numpy.linalg import inv
     from jax import numpy as jnp
     from jax import grad
```

1 Prolem №1

You will work with the following function for exercise, $f(x,y) = e^{-(sin(x)+cos(y))^2}$

Draw the computational graph for the function. Note, that it should contain only primitive operations - you need to do it automatically.

```
[ ]: #Function of first problem
     def func_p1(x, y):
       return jnp.exp(- jnp.power((jnp.sin(x[0]) + jnp.cos(y[0])), 2))


     def dfunc_p1(x, y):
       return grad(func_p1, argnums=(0, 1))(x, y)
```

```
[ ]: z=jax.xla_computation(dfunc_p1)(numpy.random.rand(1), numpy.random.rand(1))

     with open("t1.txt", "w") as f:
       f.write(z.as_hlo_text())

     with open("t1.dot", "w") as f:
       f.write(z.as_hlo_dot_graph())
```

2 Problem №2

Compare analytic and autograd approach for the hessian of: $f(x) = \frac{1}{2}x^T Ax + b^T x + c$

```
[ ]: from jax import jacfwd, jacrev
```

```
[ ]: A = numpy.random.rand(100, 100)
     b = numpy.random.rand(100)
     c = 1

     def func_p2(x):
      return 0.5 * x.T @ A @ x + b @ x + c

     def hessian(f):
       return jax.jacfwd(jax.grad(f))
```

1

```
def d2func_p2(x):
  return hessian(func_p2)(x)

hessian_auto2 = d2func_p2(numpy.random.rand(100))
hessian_anal2 = (A + A.T) / 2
```

Difference between autograde and analytical solution:

[ ]:
```
numpy.linalg.norm(hessian_anal2 - hessian_auto2)
```

[ ]: 2.1407686e-06

Cringe moment for visualising it

[ ]:
```
z = jax.xla_computation(d2func_p2)(numpy.random.rand(100))

with open("t2.txt", "w") as f:
   f.write(z.as_hlo_text())

with open("t2.dot", "w") as f:
  f.write(z.as_hlo_dot_graph())
```

3 Problem №3

Suppose we have the following function $f(x) = \frac{1}{2}||x||^2$, select a random point $x_0 \in \mathbb{B}^{1000} = \{0x_i1|i\}$. Consider 10 steps of the gradient descent starting from the point x_0:

$x_{k+1} = x_k - \alpha_k \nabla f(x_k)$.

Your goal in this problem to write the function, that takes 10 scalar values α_i and return the result of the gradient descent on function $L = f(x_{10})$. And optimize this function using gradient descent on $\alpha \in \mathbb{R}^{10}$. Suppose, $\alpha_0 = 1$.

$\alpha_{k+1} = \alpha_k - \beta \cdot \frac{\partial L}{\partial \alpha}$

Choose any β and the number of steps your need. Describe obtained results.

[ ]:
```
def func_p3(x):
  return 0.5 * x.T @ x

def dfunc_p3(x):
  return grad(func_p3)(x)
```

[ ]:
```
# Do it later...
def gradient(x0, alpha0, num_steps=10):
  x = x0
  alpha = alpha0
  for i in range(0, num_steps):
    x = x - alpha * dfunc_p3(x)
```

2

4 Problem №4

Compare analytic and autograd approach for the gradient of: $f(X) = -log(det(X))$

Analytical gradient: $df = -\frac{1}{det(X)} \cdot det(X)\langle X^{-T}, dX\rangle$

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

$\nabla f = -X^{-T}$

```
[ ]: X = numpy.random.rand(100, 100)

func_p4 = lambda X: -jnp.log(jnp.linalg.det(X))
dfunc_p4 = lambda X: grad(func_p4)(X)

grad_auto4 = dfunc_p4(X)
grad_anal4 = -(inv(X)).T

print("Difference between analytical and autograde methods:", numpy.linalg.
↪norm(grad_auto4 - grad_anal4))
```

```
Difference between analytical and autograde methods: 0.00039553354
```

5 Problem №5

Compare analytic and autograd approach for the gradient and hessian of: $f(x) = x^T x x^T x$

```
[ ]: def func_p5(x):
  return jnp.dot(x.T, x)* jnp.dot(x.T, x)

def dfunc_p5(x):
  return grad(func_p5)(x)

def d2func_p5(x):
  return hessian(func_p5)(x)
```

Analytical gradient: \$df $= 4\ \langle x, x\rangle \cdot \langle x, dx\rangle$\$

$\nabla f = 4\langle x, x\rangle \cdot x$

Analytical hessian: $d^2 f = 4 \cdot \big(\langle dx_2, x\rangle\langle x, dx_1\rangle + \langle x, dx_2\rangle\langle x, dx_1\rangle + \langle x, x\rangle\langle dx_2, dx_1\rangle\big)$

$d^2 f = 4 \cdot x^T \big(3x \cdot dx_2^T\big) dx_1 = 12x^T \cdot x dx_2^T \cdot dx_1$

$hessian(f) = 12x \cdot x^T$ - it will be matrix...

```
[ ]: x = numpy.random.rand(2)
#x = numpy.ones(2)
grad_auto5 = grad(func_p5)(x)
grad_anal5 = 4 * jnp.dot(x, x) * x
```

3

```
print("Difference between analytic and auto gradient",numpy.linalg.
 ↪norm(grad_auto5 - grad_anal5))
```

```
Difference between analytic and auto gradient 0.0
```

[ ]:
```
hessian_auto5 = d2func_p5(x)
hessian_anal5 = 12 * jnp.outer(x, x.T)

print("Difference between analytic and auto hessian:",numpy.linalg.
 ↪norm(hessian_auto5 - hessian_anal5))

#print("Hessian auto: ", hessian_auto5, hessian_auto5.shape)
#print("Hessian anal: ", hessian_anal5, hessian_anal5.shape)
```

```
Difference between analytic and auto hessian: 5.4479795
```

4

3 Convex sets

3.1 Problem №3

Prove, that if S is convex, then $S + S = 2S$. Give an counterexample in case, when S – is not convex.

Solution:

$\forall \alpha \in [0, 1]\ \forall (x, y) \in 2S \hookrightarrow \alpha \cdot (x, y) + (1 - \alpha) \cdot (x, y) \in S$

Let's rewrite this expression: $(\alpha \cdot x, \alpha \cdot y) + ((1 - \alpha) \cdot x, (1 - \alpha) \cdot y) \in 2S$ And it's right because $x \in S$ and S is convex set, the same for y. Due to that $2S = S + S$, and $\alpha \cdot x + (1 - \alpha) \cdot x \in S$, $\alpha \cdot y + (1 - \alpha) \cdot y \in S$ follows that $2S$ - convex set.

3.2 Problem №4

Let $x \in \mathbb{R}$ is a random variable with a given probability distribution of $\mathbb{P}(x = a_i) = p_i$, where $i = 1, ..., n$, and $a_1 < ... < a_n$. It is said the probability vector of outcomes of $p \in \mathbb{R}^n$ belongs to the probabilistic simpex, i.e. $P = \{p | \mathbf{1}^T p = 1, p \succcurlyeq 0\} = \{p | p_1 + ... + p_n = 1, p_i \geq 0\}$.

Determine if the following sets of p are convex:

- $\alpha < \mathbb{E}f(x) < \beta$, where $\mathbb{E}f(x)$ stands for expected value of $f(x) : \mathbb{R} \to \mathbb{R}$, i.e. $\mathbb{E}f(x) = \sum\limits_{i=1}^{n} p_i \cdot f(a_i)$
- $\mathbb{E}x^2 \leq \alpha$
- $\mathbb{V}x \leq \alpha$

Solution:

A. It's right because we reduce constraints on p as constraints on the half-space, it will follow from this that the set is convex.

$$\alpha < \mathbb{E}f(x) = \sum_{i=1}^{n} p_i f(a_i) < \beta$$

From the geometry it is half-space and convex, and it means that our set is also convex.

B. Here, we have the same idea like in **A.**. We reduce constraints on p as constraints on the half-space, it will follow this that set is convex.

$$\mathbb{E}x^2 = \sum_{i=1}^{n} p_i a_i^2 \leq \alpha$$

<u>**C.**</u>

$$0 \leq \mathbb{V}x = \mathbb{E}x^2 - (\mathbb{E}x)^2 = \sum_{i=1}^{n} p_i a_i^2 - \left(\sum_{i=1}^{n} p_i a_i\right)^2 = -p^T X p + d^T p \leq \alpha$$

where $d_i = a_i$ and $X = aa^T$, $X \succ 0$. This is a parabola with branches down in multidimensional space. Under the graph of a parabola is a convex set. And we also get that a convex set cut off by a hyperplane is a convex set.

<u>**Answer:**</u> **A.-C.** convex

4 Convex functions

4.1 Problem №1

Is $f(x) = -x \cdot lnx - (1-x)ln(1-x)$ convex?

Solution:

$$\nabla^2 f(x) = \frac{\partial}{\partial x}(-lnx - 1 + ln(1-x) + 1) = \frac{-1}{x} - \frac{-1}{1-x} = \frac{-1}{x(1-x)} < 0$$

because $x \in (0,1)$ it is right. From this we get that f(x) is concave function, but not convex.

Answer: No, it's not convex function, it's concave function.

4.2 Problem №2

Let x be a real variable with the values $a_1 < a_2 < ... < a_n$ with probabilities $\mathbb{P}(x = a_i) = p_i$. Derive the convexity or concavity of the following functions from p on the set of $\left\{ p \mid \sum\limits_{i=1}^{n} p_i = 1,\ p_i \geq 0 \right)$

Solution: We know that linear function: $a^T x + b$ convex and concave function at the same time.

1. $\mathbb{E}x = \sum\limits_{i=1}^{n} a_i \cdot p_i$ – that's linear function, therefore math expectation is convex and concave function.

2. $\mathbb{P}\{x \geq \alpha\} = \sum\limits_{i: a_i \geq \alpha} p_i$ – it's linear function, therefore it's convex and concave function.

3. $\mathbb{P}\{\alpha \leq x \leq \beta\} = \sum\limits_{i: \alpha \leq a_i \leq b_i}^{n} p_i$, we get that is convex and concave function.

4. $\sum\limits_{i=1}^{n} p_i \log p_i$.

Okay, let's check is $f(x) = x \log x$ - convex

$\nabla^2 f(x) = (x \log x)'' = (\log x + 1)' = \frac{1}{x} > 0$, because $x > 0$ - yep, it's convex function. And our function is non-negative sum of convex function, then our function is convex function.

5. $\mathbb{V}x = \mathbb{E}(x - \mathbb{E}x)^2 = \mathbb{E}x^2 - (\mathbb{E}x)^2$

Okay, let's take counterexample for this function:

1. $p_a = (1,0),\ x = (0,1),\ \mathbb{V}x = 1 \cdot 1^2 - (1 \cdot 1)^2 = 0$

2. $p_b = (0,1),\ x = (0,1),\ \mathbb{V}x = 0 \cdot 1^2 + 1 \cdot 0^2 - (1 \cdot 0 + 0 \cdot 1)^2 = 0$

3. $p_c = 0.5 \cdot p_a + 0.5 \cdot p_b = (\frac{1}{2}, \frac{1}{2})$, $x = (0, 1)$, $\mathbb{V}x = 0.5 \cdot 1^2 - (0.5 \cdot 1)^2 = 0.25$

By definition of convex function the following equality is right:

$$f(\theta x + (1-\theta)y) \le \theta f(x) + (1-\theta) f(y)$$

But:

$$f(0.5p_a + 0.5p_b) = \frac{1}{4} \le \frac{1}{2} f(p_a) + \frac{1}{2} f(p_b) = 0 + 0 = 0$$

And we get that it's not convex function, it's concave function.

6. quatile(x) $= \inf\{\beta \mid \mathbb{P}\{x \le \beta\} \ge 0.25\}$

quartile is not continuous function, because x can take discrete values, then it is defined on a discrete set of points, that is not convex. It means that function is not convex and concave.

Answer: **a-c** convex and concave function, **d.** convex function, **e.** concave function. **f.** not concave, not convex function.

4.3 Problem №4

Prove that, $f(X) = -\log\det X$ is convex on $X \in \mathbb{S}^n_{++}$

Solution:

$$df(x) = -\frac{1}{detX} detX \langle X^{-T}, dX \rangle = -\langle X^{-T}, dX \rangle$$

$$d^2 f(x) = -d(tr(X^{-1}, dX_1) = -tr(d(X^{-1})dX_1) = tr(X^{-1}dX_2X^{-1}dX_2)$$

For a reason that $X^{-1}, dX_1, dX_2 \in \mathbb{S}^n_{++}$ trace is positive, the hessian of f(x) is positive, then we get that f(x) is convex function on $\mathbb{S}^n_{++}$

4.4 Problem №5

Prove, that adding $\lambda||x||_2^2$ to any convex function $g(x)$ ensures strong convexity function of a resulting function $f(x) = g(x) + \lambda||x||_2^2$. Find the constant of the strong convexity μ.

Solution:

$$\frac{\partial^2}{\partial x_l \partial x_k} \sum_{i=1}^{n} x_i = \frac{\partial}{\partial x_l} 2x_k = 0$$

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

The hessian of $||X||_2^2$ is $2I$. Then $\nabla^2 f(x) = \nabla^2 g(x) + \lambda \cdot 2I \succcurlyeq \lambda I$. It's right because g(x) is convex function.

4.5 Problem №6

Study the following function of two variables $f(x, y) = e^{xy}$.

a. Is this function convex?

b. Prove, that this function will be convex on the line $x = y$.

c. Find another set in $\mathbb{R}^2, on which this function will be convex$

<u>**Solution:**</u>

a. No this function is not convex on $\mathbb{R}^2$, because the hessian equals $-(1 + 2xy)e^{2xy}$ and it is less than zero for $x = 0, y = 1$.

b. $f(x, x) = e^{x^2}$, $\nabla^2 f(x) = 2(1 + 2x^2)e^{x^2} > 0$, $\forall x \in \mathbb{R}$, and from that we get that $f(x)$ is strong convex function.

c. Let's take $y = x^3$, $f(x, x) = e^{x^4}$, $\nabla^2 f(x) = (12x^2 + 16x^6)e^{x^4} \geq 0$, $\forall x \in \mathbb{R}$

<u>**Answer:**</u> **a.** No, this function is not convex on $\mathbb{R}^2$, **b.** proved, **c.** $y = x^3$

5 Conjugate sets

6 Conjugate functions

6.1 Problem №1

Find $f^*(y)$, if $f(x) = p \cdot x - q$

Solution:

$$f^*(y) = \sup_{x \in dom(f)} (\langle y, x \rangle - f(x)) = \sup_{x \in \mathbf{R}} (xy - px + q)$$

$g(x) = xy - px + q$, $\nabla g(x) = y - p$ And we can easy get a answer:

Answer: $f^*(y) = \begin{cases} q, y = p \\ +\inf, y \neq p. \end{cases}$

6.2 Problem №2

Find $f^*(y), if f(x) = \frac{1}{2} x^T A x,\ A \in \mathbf{S}^n_{++}$

Solution:

$$f^*(y) = \sup_{x \in dom(f)} (\langle y, x \rangle - f(x)) = \sup_{x \in \mathbf{R}^n} \left(y^T x - \frac{1}{2} x^T A x \right)$$

$g(x) = y^T x - \frac{1}{2} x^T A x$

$\nabla g(x) = y - Ax$, because $A \in \mathbf{S}^n_{++}$, $\nabla g(x) = 0 \hookrightarrow y = Ax$

$$f^*(y) = \sup_{x \in \mathbf{R}^n} \left((Ax)^T x - \frac{1}{2} x^T A X \right) = \sup_{x \in \mathbf{R}^n} \left(\frac{1}{2} x^T A X \right) = +\infty$$

Answer: $f^*(y) = +\infty$

6.3 Problem №3

Find $f^*(y), if f(x) = log \left(\sum_{i=1}^{n} e^{x_i} \right)$

Solution:

$$f^*(y) = \sup_{x \in dom(f)} (\langle y, x \rangle - f(x)) = \sup_{x \in \mathbf{R}^n} \left(y^T x - log \left(\sum_{i=1}^{n} e^{x_i} \right) \right)$$

$d(y^t x - f(x)) = \langle y, dx \rangle - \frac{\langle e^x, dx \rangle}{\sum_{i=1}^{n} e^{x_i}} = 0$, e^x in meaning that it equals $(e^{x_1}, e^{x_2}, ..., e^{x_n})$.

And we get that: $y = \frac{e^x}{\sum_{i=1}^{n} e^{x_i}}$

$$f^*(y) = \sup_{x\in\mathbf{R}^n}\left(\sum_{i=1}^{n} y_i x_i - log\left(\sum_{i=1}^{n} e^{x_i}\right)\right) = \sup_{x\in\mathbf{R}^n}\left(\sum_{i=1}^{n} log(e^{y_i x_i}) - log\left(\sum_{i=1}^{n} e^{x_i}\right)\right)$$

$$f^*(y) = \sum_{i=1}^{n} log\left(\frac{e^{y_i}e^{x_i}}{\sum\limits_{k=1}^{n} e^{x_k}}\right) = \sum_{i=1}^{n} log\left(e^{y_i}y_i\right) = \sum_{i=1}^{n} y_i log(y_i)$$

Oh wow, it's crossentropyloss ($y \succ 0$, $y^T\mathbb{1} = 1$, so y - probability vector)
Now we need to find dom f^*.

1. if $\exists y_i : y_i < 0$, let's take $x_j = -\alpha$, and x_l for all l : $l \neq j$, then:

$$y^t x - f(x) = \alpha - log\alpha \xrightarrow{\alpha \to +\infty} +\infty$$

2. if $y \succ 0$, but $y^T\mathbb{1} \neq 1$, let's take $x = \alpha \cdot \mathbb{1}$

$$y^T x - f(x) = y^T\alpha\mathbb{1} - \alpha - \log(n)$$

if $y^T\mathbb{1} > 1$, then we take $\alpha \to +\infty$ and get that $f^*(y) = +\infty$
if $y^T\mathbb{1} < 1$, then we take $\alpha \to -\infty$ and get that $f^*(y) = -\infty$

Answer:

$$f^*(y) = \begin{cases} \sum\limits_{i=1}^{n} y_i \cdot log(y_i), \text{if } \succ 0 \text{ and } y^T\mathbb{1} = 1 \\ +\infty, otherwise \end{cases}$$

6.4 Problem №4

Prove, that if $f(x) = g(Ax)$, then $f^*(y) = g^*(A^{-T}y)$

Solution:

$$f^*(y) = \sup_{x\in\mathbf{R}^n}\left(y^T x - f(x)\right)$$

$y^T dx - \nabla f(x)^T dx = 0$, $y = \nabla f(x)$

$$f^*(y) \sup_{x\in\mathbf{R}^n}\left((\nabla f(x))^T x - f(x)\right)$$

The same way we can get:

$$g^*(y) = \sup_{x \in \mathbf{R}^n} \big((\nabla g(x))^T x - g(x)\big)$$

$$g^*(A^{-T}y) = \sup_{x \in \mathbf{R}^n} \big((A^{-T}\nabla(g(x))^T x - g(x)\big)$$

$x = Ax_1$

$$g^*(A^{-T}y) = \sup_{x_1 \in \mathbf{R}^n} \big(\nabla g(Ax_1)^T A^{-1}Ax_1 - g(Ax_1)\big) = \sup_{x_1 \in \mathbf{R}^n} \big(\nabla g(Ax_1)^T x_1 - g(Ax_1)\big)$$

Because $g(Ax_1) = f(x_1)$, we can get:

$$g^*(A^{-T}y) = \sup_{x_1 \in \mathbf{R}^n} \big(\nabla f(x_1)^T x_1 - f(x_1)\big) = f^*(y)$$

WOHOO!

6.5 Problem №5

Find $f^*(Y)$, if $f(X) = -\log(detX), X \in \mathbf{S}^n_{++}$

Solution:

$$f^*(Y) = \sup_{x \in \mathbf{R}^{n \times n}} (\langle Y, X\rangle + \log(detX))$$

$\langle Y, dX\rangle + \frac{1}{detX} d(detX) = \langle Y, dX\rangle + \frac{1}{detX} detX \langle X^{-T}, dX\rangle = 0$
And we get: $Y = -X^{-T}$, $X = -Y^{-T}$

$$f^*(Y) = \langle Y, -Y^{-T}\rangle + \log(det(-Y^{-1})) = tr(-E) + \log\big(det(-Y^{-1})\big)$$

$$f^*(Y) = -n + \log\big(det(-Y^{-1})\big)$$

Answer:

$$f^*(Y) = -n + log(det(-Y^-1)), \text{ where } Y \in -\mathbb{S}^n_{++}$$

7 Subgradient and subdifferential

7.1 Problem № 1

Find $\partial f(x)$, if $f(x) =$ **Leaky ReLU**$(x) = \begin{cases} x, if x > 0 \\ 0.01x, otherwise \end{cases}$

Solution: By Dubovitsky-Milutin theorem we can get:

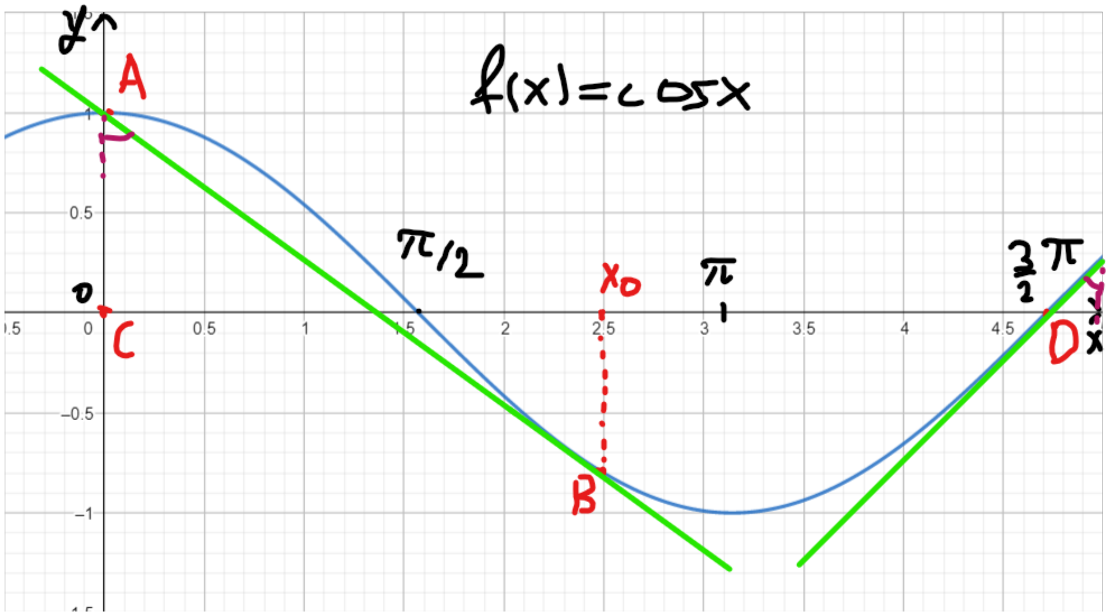
$$\partial f(x) = \begin{cases} 1, x > 0 \\ [0.01; 1], x = 0 \\ 0.01, x < 0 \end{cases}$$

Answer: $\partial f(x) = \begin{cases} 1, x > 0 \\ [0.01; 1], x = 0 \\ 0.01, x < 0 \end{cases}$

7.2 Problem №2

Find subdifferential of a function $f(x) = cosx$ on the set $X = [0, \frac{3}{2}\pi]$.

Solution:



Answer: $\partial f(x) = \begin{cases} [-\infty, -sinx], x = 0 \\ \varnothing, x \in (0, x_0) \\ -sinx, x \in [x_0, \frac{3}{2}\pi) \\ [1, +\infty], x = \frac{3}{2}\pi \end{cases}$

7.3 Problem №3

Find $\partial f(x)$, if $f(x) = ||Ax - b||_1^2$

Solution: By property of subdifferential we can get that:

$$\partial(||Ax - b||_1^2) = ||Ax - b||_1 \partial(||Ax - b||_1)$$

From the seminar we know that:

$$\partial||y||_1 = \{\alpha : ||\alpha||_\infty \leq 1, \alpha^T y = ||y||_1\}$$

And now we can get that (by another property of subdifferential:

$$\partial\left(||Ax - b||_1^2\right) = ||Ax - b||_1 \partial\left(||Ax - b||_1\right)(x) = ||Ax - b||_1 A^T \partial||Ax + b||_1$$

$$||Ax - b||_1 A^T \partial||Ax + b||_1 = ||Ax - b||_1 A^T \cdot \{\alpha : ||\alpha||_\infty \leq 1, \alpha^T (Ax + b) = ||Ax + b||_1\}$$

Answer: $\partial f(x) = ||Ax - b||_1 A^T \cdot \{\alpha : ||\alpha||_\infty \leq 1,\ \alpha^T (Ax + b) = ||Ax + b||_1\}$

7.4 Problem №4

Suppose, that if $f(x) = ||x||_\infty$. Prove that $\partial f(0) = \textbf{conv}\ \{\pm e_1, ..., \pm e_n\}$, where e_i is i-th canonical basis vector (column of identity matrix).

Solution: By the definition: $f(x) = ||x||_\infty = \max\limits_i |x_i|$

We know that subbdifferential for module is equal:

$$\partial|x_i| = \begin{cases} x_i, x_i > 0 \\ [-1, 1], x_i = 0 \\ -x_i, x_i < 0 \end{cases}$$

Because $|x_i|$ – convex functions, by Dubovitsky - Milutin theorem we can get:

$$\partial f(0) = conv \left\{ \bigcup_{i \in \overline{1,n}} \partial|x_i|_{x_i = 0} \right\} = conv\left\{\pm e_1, ..., \pm e_n\right\}, \ e_i \text{ is i-th canonical basis vector}$$

WOHOO, we proved that!

7.5 Problem №5

Find $\partial f(x)$, if $f(x) = e^{||x||}$.
Try do the task for an arbitrary norm. At least, try $||\cdot|| = ||\cdot||_{\{2,1,\infty\}}$

Solution:

By the property of subdifferential we: $\partial f(x) = \partial(e^{||x||}) = e^{||x||}\partial(||x||)$
And now we need to find $\partial||x||$ for $||x||_1, ||x||_2$ and $||x||_\infty$
1. In the seminar we find that and it equals:

$$\partial||x||_1 = \{\alpha : ||\alpha||_\infty \leq 1, \alpha^T x = ||x||_1\}$$

2. By definition $||x||_2 = \sqrt{\sum\limits_{i=1}^{n} x_i^2}$, this function is differentiable everywhere except zero.
For $x \neq 0$, $\partial f(x) = \nabla||x||_2 = \frac{x}{||x||_2}$
Now we need to consider $x = 0$: let's find such interesting limit: $\lim\limits_{\beta \longrightarrow 0+} \frac{||\beta e||_2}{\beta} = ||e||_2$, where e is unit vector on unit sphere.

But by definition $\frac{x}{||x||_2} = e$, and we get that

$$\partial||x||_2 = \{e \mid ||e||_2 \leq 1\}$$

3. In problem №4 I find that (x_i - maximum element by modules):

$$||x||_\infty = \begin{cases} [-1, 1], \text{ if } ix_i = 0 \\ sign(x_i), x_i \neq 0 \end{cases}$$

Answer:

- for $||\cdot||_1$: $\partial f(x) = e^{||x||_1} \cdot \{\alpha : ||\alpha||_\infty \leq 1, \alpha^T x = ||x||_1\}$
- for $||\cdot||_2$: $\partial f(x) = e^{||x||_2} \cdot \{e \mid ||e||_2 \leq 1\}$
- for $||\cdot||_\infty$: $\partial f(x) = e^{||x||_\infty} \cdot \begin{cases} [-1, 1], \text{ if } x_i = 0 \\ sign(x_i), x_i \neq 0 \end{cases}$, where is x_i – maximum element by modules