

Как вытащить гессиан?  $f: \mathbb{R}^n \rightarrow \mathbb{R}$

$$\nabla^2 f = \begin{pmatrix} \frac{\partial^2 f}{\partial x_i \partial x_j} \end{pmatrix}_{ij} \in \mathbb{R}^{n \times n}$$

1.

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

Теперь  $dx = dx_1 = \text{const}$

$$df = \langle \nabla f(x), dx_1 \rangle$$

$$d^2f = d(df) = \langle d(\nabla f(x)), dx_1 \rangle$$

Пример 1:  $f(x) = \frac{1}{2}x^T Ax - b^T x + C$  (гессиан  $dx = \text{const}$ )

$$1. df = \langle \frac{1}{2}(A+A^T)x - b, dx_1 \rangle$$

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

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

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

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

$$H_f = f'' =$$

Пример 2  $f(x) = -e^{-x^T x}$   $\nabla f, \nabla^2 f = ?$

Решение:

$$1. df = d(-e^{-\langle x, x \rangle}) = \underbrace{-e^{-\langle x, x \rangle}}_f \cdot d(-\langle x, x \rangle) = f: \mathbb{R}^n \rightarrow \mathbb{R}$$

$$= f \cdot (-\langle x, dx \rangle - \langle dx, x \rangle) = -f \cdot \langle 2x, dx \rangle$$

$$\Rightarrow \boxed{\nabla f = -2fx = 2 \cdot e^{-\langle x, x \rangle} \cdot x}$$

$$2. d^2 f = d(\langle -2fx, dx_1 \rangle) =$$

$$= \langle -2d(fx), dx_1 \rangle = -2 \langle \underbrace{df \cdot x}_{\text{inner product}}, \underbrace{dx_1}_{\text{outer product}} \rangle =$$

$$= -2 \langle (-f \langle 2x, dx \rangle)x + f dx, dx_1 \rangle =$$

$$= -2 \langle -2f \cdot x^T dx \cdot x + f \cdot I dx, dx_1 \rangle =$$

$$= -2f \langle -2 \cdot \underbrace{x \cdot x^T dx}_{\text{inner product}}, \underbrace{I dx, dx_1}_{\text{outer product}} \rangle =$$

$$= -2f \langle \underbrace{(-2x^T + I) dx}_{\text{inner product}}, \underbrace{dx_1}_{\text{outer product}} \rangle$$

$$\Rightarrow \boxed{\nabla^2 f = -2f \cdot (I - 2x^T)}$$

$$\boxed{\nabla^2 f = 2e^{-\langle x, x \rangle} (I - 2x^T)}$$

$$(I - 2x^T)^T = I - 2(x^T)^T = I - 2x^T$$

$$\begin{pmatrix} x_1 x_1 & x_1 x_2 \dots x_{k-1} \\ \vdots & \\ x_k x_1 & \dots x_{k-1} \end{pmatrix}$$

Задача:  $f(x) = \frac{1}{2} \|Ax - b\|_2^2$

$\nabla f = ?$

$\nabla^2 f = ?$

$\|z\|_2^2 = \langle z, z \rangle$

$d(\langle z, z \rangle) = \langle 2z, dz \rangle$

$\boxed{\nabla f = A^T(Ax - b)}$

$x \in \mathbb{R}^n$   
 $A \in \mathbb{R}^{m \times n}$   
 $b \in \mathbb{R}^m$

«внешний  
нормы  
вектора»

$$f = \frac{1}{2} \langle Ax - b, Ax - b \rangle$$

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

$$\underbrace{A^T A}_{\substack{n \times m \\ m \times n}} \quad \nabla f \in \mathbb{R}^{n \times n}$$

$$d(Ax - b) =$$

$$= d(Ax) - db =$$

$$= Adx - 0$$

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

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

$$= \langle A^T A dx, dx_1 \rangle =$$

$$= \langle A^T A dx, dx \rangle$$

$$\Rightarrow \boxed{\nabla^2 f = A^T A}$$

$$\text{Zagora: } f(x) = \ln(1 + \exp(a, x))$$

$$\begin{matrix} x \in \mathbb{R}^n \\ a \in \mathbb{R}^n \end{matrix}$$

$$\nabla f, \nabla^2 f = ?$$

$$d(\ln f) = \frac{df}{f} \quad d(\exp(f)) = \exp(f) \cdot df$$

$$\text{Rechenve: } df = \frac{d(1 + \exp(a, x))}{1 + \exp(a, x)} = \frac{\exp(a, x) \cdot \langle a, dx \rangle}{1 + \exp(a, x)}$$

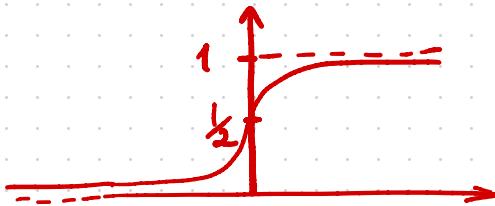
$$\Rightarrow \nabla f = \frac{\exp(a, x)}{1 + \exp(a, x)} \cdot a \Rightarrow \frac{1}{1 + \exp(-\langle a, x \rangle)} = \sigma(\langle a, x \rangle)$$

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

$$d\sigma(f) = \sigma(f) (1 - \sigma(f)) \cdot df$$

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

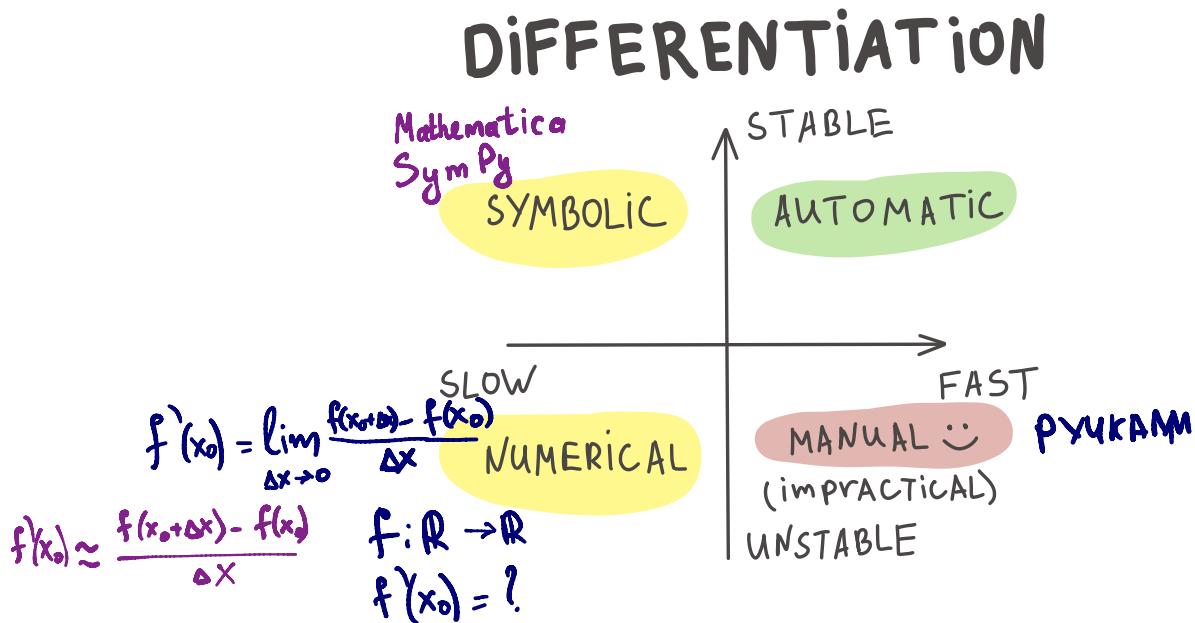
CKAMP nx1



$$\begin{aligned}
 d^2 f &= \langle d(\sigma(\langle a, x \rangle) \cdot a), dx_1 \rangle = \\
 &= \langle \underbrace{\sigma(\langle a, x \rangle)}_{\sigma} (1 - \sigma(\langle a, x \rangle)) \langle a, dx \rangle \cdot a, dx_1 \rangle \quad d^2 f = \langle H dx, dx \rangle \\
 &= \sigma(1 - \sigma) \langle a^T dx a, dx_1 \rangle = \quad H = H^T \\
 &= \sigma(1 - \sigma) \langle \underbrace{[a \ a^T] dx}_{\text{CKA resp}} \cdot dx, dx_1 \rangle = \quad H = \underbrace{\sigma(\langle a, x \rangle) (1 - \sigma(\langle a, x \rangle)) a a^T}_{n \times 1 \text{ resp } 1 \times n} \\
 &= \sigma(1 - \sigma) \langle a a^T dx_1, dx \rangle \Rightarrow \quad d^2 f = \langle \dots dx, dx \rangle
 \end{aligned}$$

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

# Idea



Automatic differentiation is a scheme, that allows you to compute a value of gradient of function with a cost of computing function itself only twice.

## Chain rule

We will illustrate some important matrix calculus facts for specific cases

### Univariate chain rule

Suppose, we have the following functions  $R: \mathbb{R} \rightarrow \mathbb{R}$ ,  $L: \mathbb{R} \rightarrow \mathbb{R}$  and  $W \in \mathbb{R}$ . Then

$$x \rightarrow W(x) \rightarrow L(W) \rightarrow R(L(W))$$

$$\frac{\partial R}{\partial W} = \frac{\partial R}{\partial L} \frac{\partial L}{\partial W}$$

нропзб.  
сюхноу  
оутикчум

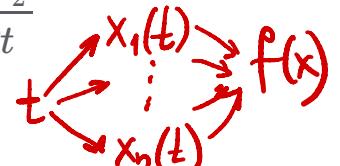
### Multivariate chain rule

The simplest example:

$$\frac{\partial R}{\partial x} = \frac{\partial R}{\partial L} \frac{\partial L}{\partial w} \frac{\partial w}{\partial x}$$

$$\frac{\partial}{\partial t} f(x_1(t), x_2(t)) = \frac{\partial f}{\partial x_1} \frac{\partial x_1}{\partial t} + \frac{\partial f}{\partial x_2} \frac{\partial x_2}{\partial t}$$

Now, we'll consider  $f: \mathbb{R}^n \rightarrow \mathbb{R}$ :



$$\frac{\partial}{\partial t} f(x_1(t), \dots, x_n(t)) = \frac{\partial f}{\partial x_1} \frac{\partial x_1}{\partial t} + \dots + \frac{\partial f}{\partial x_n} \frac{\partial x_n}{\partial t}$$

But if we will add another dimension  $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$ , than the  $j$ -th output of  $f$  will be:

$$\frac{\partial}{\partial t} f_j(x_1(t), \dots, x_n(t)) = \sum_{i=1}^n \frac{\partial f_j}{\partial x_i} \frac{\partial x_i}{\partial t} = \sum_{i=1}^n J_{ji} \frac{\partial x_i}{\partial t}, \quad \text{where } J \in \mathbb{R}^{m \times n}$$

where matrix  $J \in \mathbb{R}^{m \times n}$  is the jacobian of the  $f$ . Hence, we could write it in a vector way:

$$\begin{array}{c} f(x) \\ \left( \frac{\partial f}{\partial t} = J \frac{\partial x}{\partial t} \right) \\ \begin{matrix} m \times 1 & m \times n & n \times 1 \end{matrix} \end{array} \iff \left( \frac{\partial f}{\partial t} \right)^T = \left( \frac{\partial x}{\partial t} \right)^T J^T$$

## Backpropagation

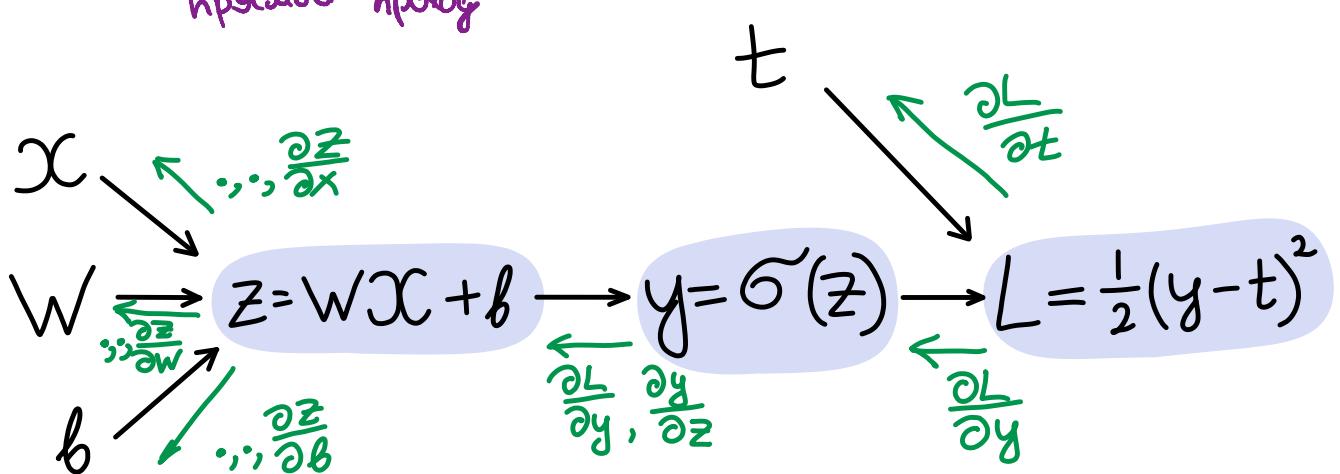
The whole idea came from the applying chain rule to the computation graph of primitive operations

$$W_{k+1} = W_k - \eta \frac{\partial L}{\partial W}(w_k)$$

$$L = L(y(z(w, x, b)), t)$$

ногорѣт ажыкым  
нотерп

### FORWARD PASS (COMPUTE LOSS)



### BACKWARD PASS (COMPUTE derivatives)

$$z = wx + b$$

$$\frac{\partial z}{\partial w} = x, \frac{\partial z}{\partial x} = w, \frac{\partial z}{\partial b} = 0$$

$$y = \sigma(z)$$

$$\frac{\partial y}{\partial z} = \sigma'(z)$$

$$L = \frac{1}{2}(y - t)^2$$

$$\frac{\partial L}{\partial y} = y - t, \frac{\partial L}{\partial t} = t - y$$

All frameworks for automatic differentiation construct (implicitly or explicitly) computation graph. In deep learning we typically want to compute the derivatives of

the loss function  $L$  w.r.t. each intermediate parameters in order to tune them via gradient descent. For this purpose it is convenient to use the following notation:

$$\overline{v_i} = \frac{\partial L}{\partial v_i}$$

Let  $v_1, \dots, v_N$  be a topological ordering of the computation graph (i.e. parents come before children).  $v_N$  denotes the variable we're trying to compute derivatives of (e.g. loss).

### Forward pass:

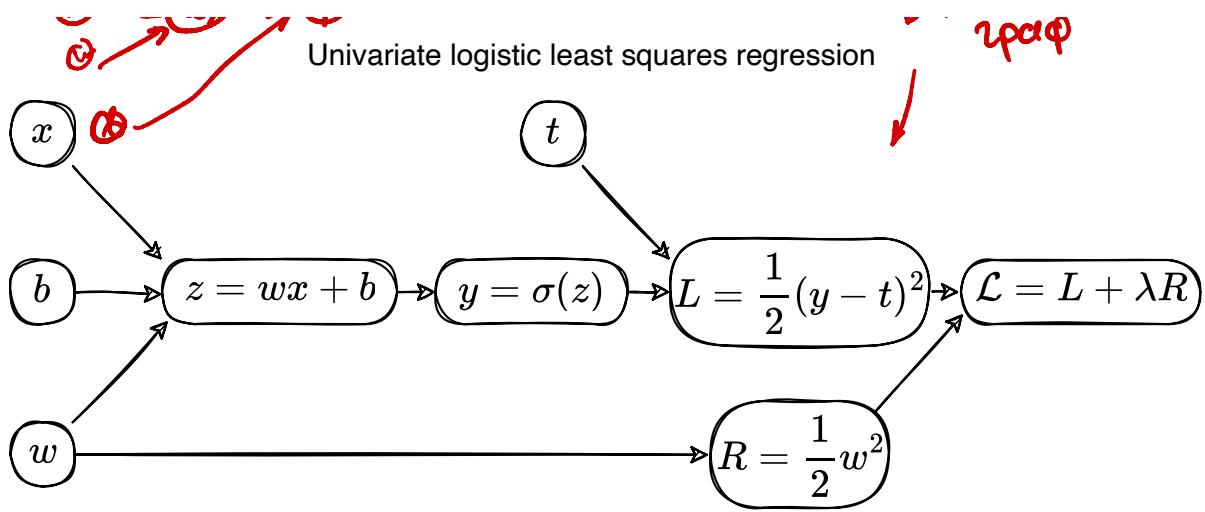
- For  $i = 1, \dots, N$ :
  - Compute  $v_i$  as a function of its parents.

### Backward pass:

- $\overline{v_N} = 1$
- For  $i = N - 1, \dots, 1$ :

- Compute derivatives  $\overline{v_i} = \sum_{j \in \text{Children}(v_i)} \overline{v_j} \frac{\partial v_j}{\partial v_i}$

Note, that  $\overline{v_j}$  term is coming from the children of  $\overline{v_i}$ , while  $\frac{\partial v_j}{\partial v_i}$  is already precomputed effectively.



Forward pass

$$\begin{aligned} z &= wx + b \\ y &= \sigma(z) \\ L &= \frac{1}{2}(y - t)^2 \\ R &= \frac{1}{2}w^2 \\ \mathcal{L} &= L + \lambda R \end{aligned}$$

Backward pass

$$\begin{aligned} \bar{L} &= 1 \\ \bar{R} &= \bar{L} \frac{d\mathcal{L}}{dR} = \bar{L}\lambda \\ \bar{L} &= \bar{L} \frac{d\mathcal{L}}{dL} = \bar{L} \\ \bar{y} &= \bar{L} \frac{dL}{dy} = \bar{L}(y - t) \end{aligned}$$

$$\begin{aligned} \bar{z} &= \bar{y} \frac{dy}{dz} = \bar{y}\sigma'(z) \\ \bar{w} &= \bar{z} \frac{dz}{dw} + \bar{R} \frac{dR}{dw} = \bar{z}x + \bar{R}w \\ \bar{b} &= \bar{z} \frac{dz}{db} = \bar{z} \\ \bar{x} &= \bar{z} \frac{dz}{dx} = \bar{z}w \end{aligned}$$

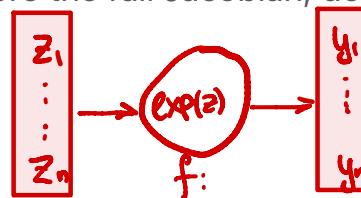
## Jacobian vector product

The reason why it works so fast in practice is that the Jacobian of the operations are already developed in effective manner in automatic differentiation frameworks.

Typically, we even do not construct or store the full Jacobian, doing matvec directly instead. ~~Handbook~~  $O(n^2)$

Example: element-wise exponent

$$y = \exp(z) \quad J = \text{diag}(\exp(z)) \quad \bar{z} = \bar{y}J$$



$$y_i = e^{z_i}$$

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

$$\frac{\partial y_p}{\partial z_q}$$

$$\frac{\partial y_1}{\partial z_1} = e^{z_1}$$

$$\frac{\partial y_1}{\partial z_2} = 0$$

$$\begin{pmatrix} e^{z_1} & 0 & 0 & 0 & 0 \\ 0 & e^{z_2} & 0 & 0 & 0 \\ 0 & 0 & e^{z_3} & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & 0 & e^{z_n} \end{pmatrix}$$

See the examples of Vector-Jacobian Products from autodidact library:

```
defvjp(anp.add,
       lambda g, ans, x, y : unbroadcast(x, g),
       lambda g, ans, x, y : unbroadcast(y, g))
defvjp(anp.multiply,
       lambda g, ans, x, y : unbroadcast(x, y * g),
       lambda g, ans, x, y : unbroadcast(y, x * g))
defvjp(anp.subtract,
       lambda g, ans, x, y : unbroadcast(x, g),
       lambda g, ans, x, y : unbroadcast(y, -g))
defvjp(anp.divide,
       lambda g, ans, x, y : unbroadcast(x, g / y),
       lambda g, ans, x, y : unbroadcast(y, -g * x / y**2))
```

$$\begin{pmatrix} e^{z_1} & 0 & 0 & 0 & 0 \\ 0 & e^{z_2} & 0 & 0 & 0 \\ 0 & 0 & e^{z_3} & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & 0 & e^{z_n} \end{pmatrix}$$

$$= \begin{pmatrix} e^{z_1} \cdot x_1 \\ \vdots \\ e^{z_n} \cdot x_n \end{pmatrix}$$

```
defvjp(anp.true_divide, lambda g, ans, x, y : unbroadcast(x, g / y),
       lambda g, ans, x, y : unbroadcast(y, -g * x / y**2))
```

## Hessian vector product

Interesting, that the similar idea could be used to compute Hessian-vector products, which is essential for second order optimization or conjugate gradient methods. For a scalar-valued function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  with continuous second derivatives (so that the Hessian matrix is symmetric), the Hessian at a point  $x \in \mathbb{R}^n$  is written as  $\partial^2 f(x)$ . A Hessian-vector product function is then able to evaluate

$$v \mapsto \partial^2 f(x) \cdot v$$

for any vector  $v \in \mathbb{R}^n$ .

The trick is not to instantiate the full Hessian matrix: if  $n$  is large, perhaps in the millions or billions in the context of neural networks, then that might be impossible to store. Luckily, `grad` (in the jax/autograd/pytorch/tensorflow) already gives us a way to write an efficient Hessian-vector product function. We just have to use the identity

$$\partial^2 f(x)v = \partial[x \mapsto \partial f(x) \cdot v] = \partial g(x),$$

where  $g(x) = \partial f(x) \cdot v$  is a new vector-valued function that dots the gradient of  $f$  at  $x$  with the vector  $v$ . Notice that we're only ever differentiating scalar-valued functions of vector-valued arguments, which is exactly where we know `grad` is efficient.

```
import jax.numpy as jnp
```

```
def hvp(f, x, v):
    return grad(lambda x: jnp.vdot(grad(f)(x), v))(x)
```

1. Строится вложенный график при помощи векторных операций.  $\oplus \otimes \odot \odiv e^x$

2. Каждая из операций складывается в матрицу  $J$  (произведение вектора на вектор) якобианом

3. Во время backward pass собираются нужные производные.  
4. Важно, что внутри реализовано

## Code

Open in Colab

## Materials

- Autodidact - a pedagogical implementation of Autograd
- CSC321 Lecture 6
- CSC321 Lecture 10
- Why you should understand backpropagation :)
- JAX autodiff cookbook

Вычисление  $f(x) \sim t_0$   
Вычисление всех  
произв.  $\frac{\partial f}{\partial \dots}$  в坝.  $\sim 2t$   
графе