



---

# **LINMA2472 - Algorithm in data science**

---

SIMON DESMIDT

Academic year 2025-2026 - Q1



UCLouvain

# Contents

<b>1</b>	<b>Automatic differentiation</b>	<b>2</b>
1.1	Chain rule . . . . .	2
1.2	Forward differentiation . . . . .	2
1.3	Backward differentiation . . . . .	3
1.4	Computational graph and multivariate differentiation . . . . .	4
1.5	Jacobian computation . . . . .	5
1.6	Memory usage . . . . .	5
1.7	Second order AD . . . . .	7
1.8	Implicit differentiation . . . . .	9
1.9	Sparse AD . . . . .	12
<b>2</b>	<b>Neural networks</b>	<b>17</b>
2.1	Autoregressive models . . . . .	18
2.2	Tokenization . . . . .	19
2.3	Embedding . . . . .	19
2.4	Recurrent neural networks (RNN) . . . . .	20
2.5	Attention head . . . . .	21
2.6	Decoder-only transformer . . . . .	22
2.7	Performances of transformers . . . . .	24
<b>3</b>	<b>Diffusion Models</b>	<b>25</b>
3.1	Tweedie's formula . . . . .	25
3.2	Langevin dynamics (sampling) . . . . .	25
3.3	Score matching . . . . .	26
<b>4</b>	<b>Kernels</b>	<b>27</b>
4.1	Reminders on scalar product . . . . .	27
4.2	Kernel methods for finite sets . . . . .	28
4.3	Kernels methods for continuous sets . . . . .	29
4.4	Polynomial kernels . . . . .	30

# Automatic differentiation

The Automatic differentiation is an algorithmic technique to compute automatically the derivative (gradient) of a function defined in a computer program. Unlike symbolic differentiation (done by hand) and numerical differentiation (finite difference approximation), automatic differentiation exploits the fact that every function can be decomposed into a sequence of elementary operations (addition, multiplication, sine, exponential, etc.) and so that we can apply the chain rule to compute the derivative of the whole function. Thus we can compute the gradient of a function exactly and efficiently.

Automatic differentiation is widely used in machine learning because for the neural networks, we need to compute the gradient of a loss function with respect to the parameters of the model (weights and biases) to update them during the training process and it would be difficult to compute this manually for each node.

## 1.1 Chain rule

There is two ways to apply the chain rule to compute the gradient of a function: forward differentiation and backward differentiation. Suppose that we have a composition of  $m$  functions. The chain rule gives us:

$$f'(x) = f'_m(f_{m-1}(f_{m-2}(\dots f_1(x)\dots))) \cdot \dots \cdot f'_2(f_1(x)) \cdot f'_1(x) \quad (1.1)$$

Let's define:

$$\begin{cases} s_0 &= x \\ s_k &= f_k(s_{k-1}) \end{cases} \quad (1.2)$$

We thus get:

$$f'(x) = f'_m(s_{m-1}) \cdot \dots \cdot f'_2(s_1) \cdot f'_1(s_0) \quad (1.3)$$

Based on this, we can define the forward and backward differentiation algorithms.

## 1.2 Forward differentiation

Also called forward mode, this algorithm consists in propagating forward the derivative and the values at the same time. The fact of propagating the values forward is called a **forward pass**. It can be represented by this graph where the blue part represents the values and the green part the derivatives:

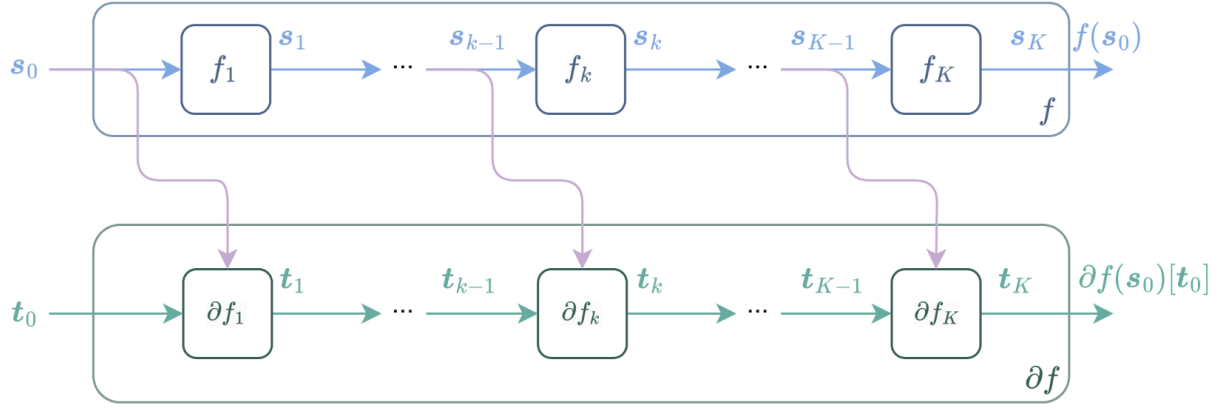


Figure 1.1: Forward differentiation

And it can be computed with the following recurrence relation:

$$\begin{cases} t_0 = 1 \\ t_k = f'_k(s_{k-1}) \cdot t_{k-1} \end{cases} \quad (1.4)$$

It is simple to implement and very efficient for functions with a small number of input variables. However, it becomes inefficient for functions with a large number of input variables because we need to compute the derivative for each input variable separately. So in practice for neural networks where we have a large number of input variables (weights and biases), we use the backward differentiation.

### 1.3 Backward differentiation

Also called backward mode, this algorithm consists in propagating the derivative backward and the values forward at the same time. The fact of propagating the derivative backward is called a **backward pass**. It can be represented by this graph where the blue part represents the values and the orange part the derivatives:

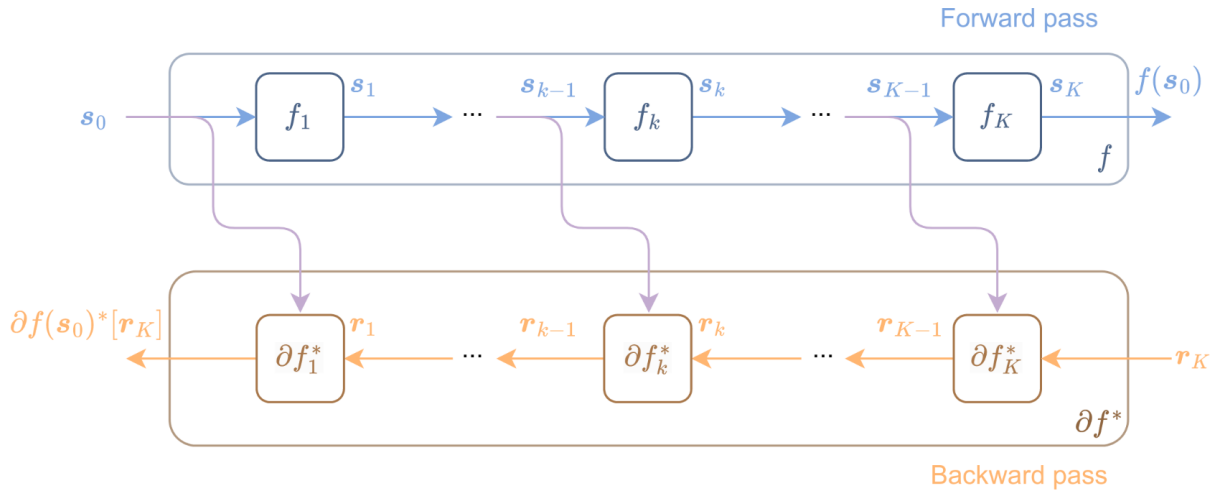


Figure 1.2: Backward differentiation

The idea is to compute all the intermediate values  $s_k$  in a forward pass and then compute the derivatives  $r_k$  based on the output in a backward pass. It can be computed with the following recurrence relation:

$$\begin{cases} r_m &= 1 \\ r_k &= r_{k+1} \cdot f'_{k+1}(s_k) \end{cases} \quad (1.5)$$

This method is more complex to implement but it is very efficient for functions with a large number of input variables and a small number of output variables typically 1, the loss function.

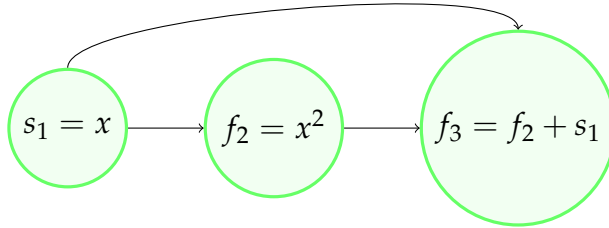
## 1.4 Computational graph and multivariate differentiation

### 1.4.1 Computational graph

To represent the computation of a function, we can use a computational graph. It is a directed acyclic graph where the nodes represent the operations and the edges represent the variables. For example, consider the function with  $f_1(x) = x = s_1$  and  $f_2(x) = x^2 = s_2$ :

$$f_3(s_1, s_2) = s_1 + s_2 = x + x^2 \quad (1.6)$$

The computational graph is:



### 1.4.2 Multivariate differentiation

Let's consider the function of the computational graph above:

$$f_3(f_1(x), f_2(x)) = s_3 = f_1(x) + f_2(x) = s_1 + s_2 = x + x^2 \quad (1.7)$$

following the chain rule, we have:

$$f'_3(x) = \frac{\partial f_3}{\partial s_1} \frac{\partial s_1}{\partial x} + \frac{\partial f_3}{\partial s_2} \frac{\partial s_2}{\partial x} \quad (1.8)$$

For the forward automatic differentiation, we work the same way as before, we propagate the values and the derivatives forward. But when we have a node with multiple inputs, we need to use formula derived from the chain rule. For the function  $f_3$  that we want to evaluate in  $x = 3$ , we will have:

$$\begin{cases} t_0 &= 1 \\ t_1 &= f'_1(x)|_{x=3} \cdot t_0 = 1 \\ t_2 &= f'_2(x)|_{x=3} \cdot t_0 = 6 \\ t_3 &= \frac{\partial f_3}{\partial s_1}|_{x=3} \cdot t_1 + \frac{\partial f_3}{\partial s_2}|_{x=3} \cdot t_2 = 7 \end{cases} \quad (1.9)$$

For the backward automatic differentiation, first we need to initialize the gradient accumulator to 0.

$$\frac{\partial s_3}{\partial s_1} = \frac{\partial s_3}{\partial s_2} = \frac{\partial s_3}{\partial x} = 0 \quad (1.10)$$

Then we compute the intermediate values in a forward pass:

$$\begin{aligned} \frac{\partial s_3}{\partial s_1} + &= 1 \Rightarrow \frac{\partial s_3}{\partial x} + = 1 \cdot 1|_{x=3} \\ \frac{\partial s_3}{\partial s_2} + &= 1 \Rightarrow \frac{\partial s_3}{\partial x} + = 1 \cdot 2x|_{x=3} \end{aligned} \quad (1.11)$$

Finally we get:

$$\frac{\partial s_3}{\partial x} = 7 \quad (1.12)$$

## 1.5 Jacobian computation

When doing the forward and backward mode, we compute the Jacobian matrix of the function. Using this Jacobian we can do the forward mode like this:

$$J_f(x) \cdot v \quad (\text{JVP}) \quad (1.13)$$

where  $v$  is a vector of size  $n$  (number of input variables) and the backward mode like this:

$$v^T J_f(x) \quad (\text{VJP}) \quad (1.14)$$

Consider a function  $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$  then computing the full Jacobian requires  $n$  forward passes (JVP) or  $m$  backward passes (VJP). Therefore,

- If  $n \ll m$ , we use the forward mode because it's faster
- If  $n \gg m$ , we use the backward mode because it's faster
- If  $n \approx m$ , we can use either mode

## 1.6 Memory usage

The forward mode only needs to store the current value and the current derivative, so the memory usage is relatively constant.

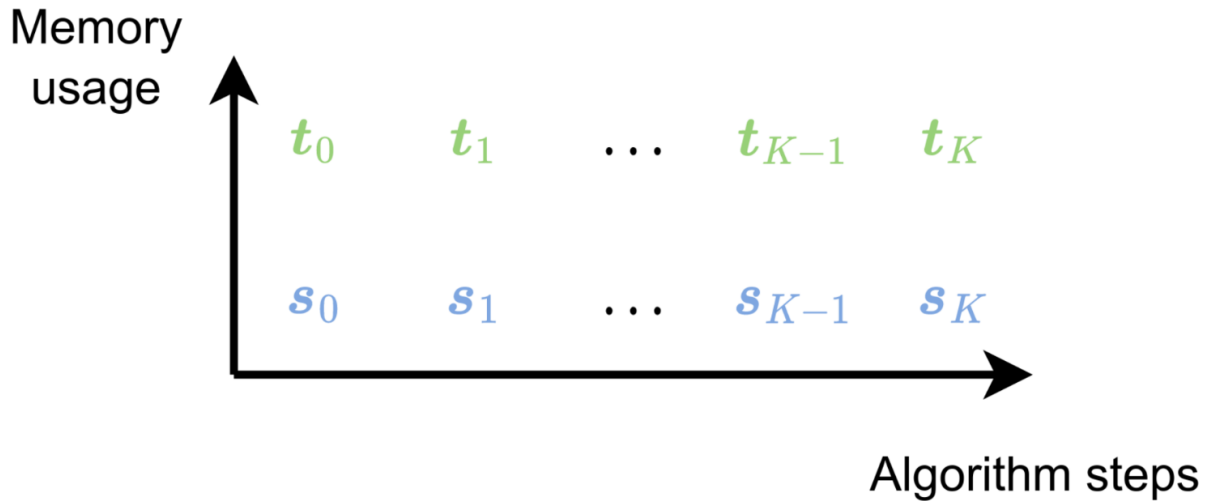


Figure 1.3: Forward mode memory usage

However, the backward mode needs to store all the intermediate values to compute the derivatives in the backward pass so the memory usage will first increase then reduce when we will start to use the derivatives previously computed.

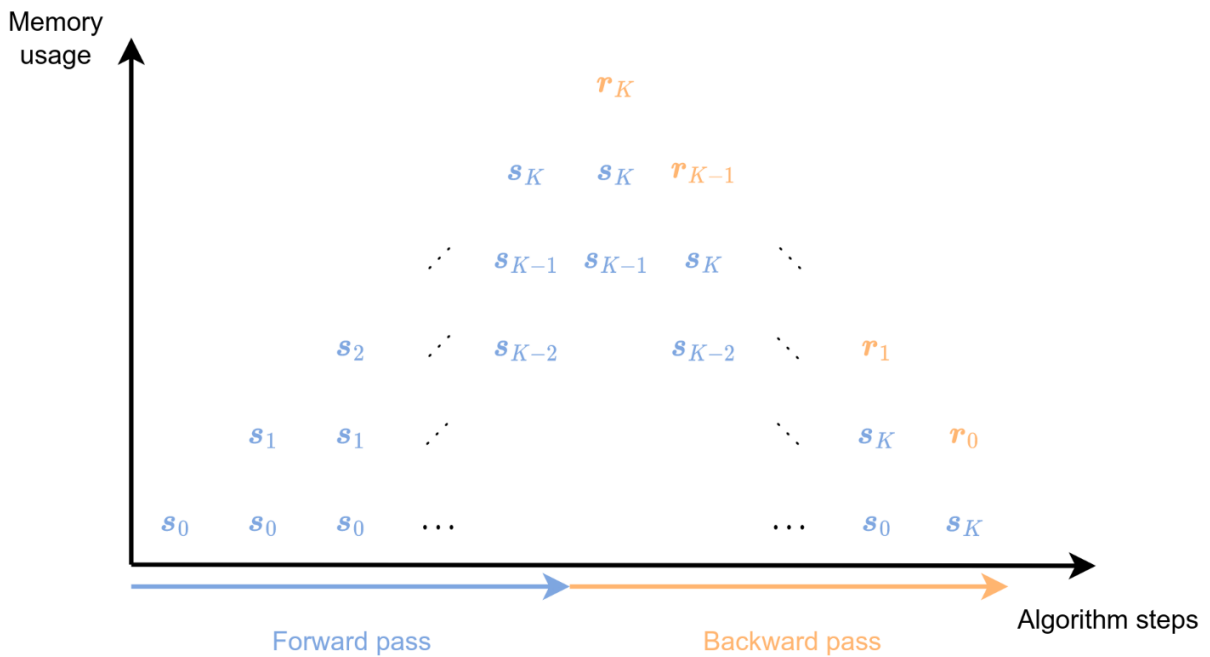


Figure 1.4: Backward mode memory usage

So the forward mode is more memory efficient than the backward mode. However, this factor may be less significant than the number of operations performed (JVP and VJP).

## 1.7 Second order AD

And what happens when we take a look at the second order derivative? Remember that we can compute the Hessian  $\nabla^2 f(x)$  like this:

$$(\nabla^2 f(x))_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j} \quad (1.15)$$

Or like this:

$$\nabla^2 f(x) = J_{\nabla f}(x) \quad (1.16)$$

We can easily see that with this definition, we can use an AD for the gradient and for the Jacobian to compute the Hessian. And because we can separate these two computations in distinct AD, it means that we are not forced to use the same mode for both. So we can do either forward or backward for both computations without taking care of the mode used for the other computation.

Based on the chain rule, let's rewrite  $\frac{\partial^2(f_2 \circ f_1)}{\partial x_i \partial x_j}$  into something more suitable for AD (consider  $\partial f$  as the gradient of  $f$ ):

$$\begin{aligned} \frac{\partial^2(f_2 \circ f_1)}{\partial x_i \partial x_j} &= \frac{\partial}{\partial x_j} \left( \frac{\partial(f_2 \circ f_1)}{\partial x_i} \right) \\ &= \frac{\partial}{\partial x_j} \left( \partial f_2 \frac{\partial f_1}{\partial x_i} \right) \\ &= \left( \partial^2 f_2 \frac{\partial f_1}{\partial x_j} \right) \frac{\partial f_1}{\partial x_i} + \partial f_2 \frac{\partial^2 f_1}{\partial x_i \partial x_j} \end{aligned} \quad (1.17)$$

Introducing the variables  $\mathbf{J}_k = \partial f_k$  and  $\mathbf{H}_{kj} = \frac{\partial}{\partial x_j} \mathbf{J}_k = \partial^2 f_k \frac{\partial f_{k-1}}{\partial x_j}$ , and so we get:

$$\frac{\partial^2(f_2 \circ f_1)}{\partial x_i \partial x_j} = \mathbf{H}_{2j} \frac{\partial f_1}{\partial x_i} + \mathbf{J}_2 \frac{\partial^2 f_1}{\partial x_i \partial x_j} \quad (1.18)$$

We can now define four different ways to compute the Hessian depending on the mode used for each computation.

### 1.7.1 Forward on forward

Define  $Dual(s_1, t_1)$  with  $s_1 = Dual(f_1(x), \frac{\partial f_1}{\partial x_j})$  and  $t_1 = Dual(\frac{\partial f_1}{\partial x_i}, \frac{\partial^2 f_1}{\partial x_i \partial x_j})$  then we can have this algorithm:

1. Compute  $s_2 = f_2(s_1) = (f_2(f_1(x)), \mathbf{J}_2 \frac{\partial f_1}{\partial x_j})$
2. Compute  $\mathbf{J}_{f_2}(s_1)$  which gives  $Dual(\mathbf{J}_2, \mathbf{H}_{2j})$
3. Compute

$$\begin{aligned} t_2 &= \mathbf{J}_{f_2}(s_1) t_1 \\ &= Dual(\mathbf{J}_2, \mathbf{H}_{2j}) Dual(\frac{\partial f_1}{\partial x_i}, \frac{\partial^2 f_1}{\partial x_i \partial x_j}) \\ &= Dual(\mathbf{J}_2 \frac{\partial f_1}{\partial x_i}, \mathbf{J}_2 \frac{\partial^2 f_1}{\partial x_i \partial x_j} + \mathbf{H}_{2j} \frac{\partial f_1}{\partial x_i}) \end{aligned} \quad (1.19)$$



#### 4. Repeat

All that can be resumed as these two equations with  $g_k(x) = f_k \circ \dots \circ f_1$ :

$$\begin{cases} s_k &= Dual(g_k(x), \frac{\partial g_k}{\partial x_j}) \\ t_k &= Dual(\frac{\partial g_k}{\partial x_i}, \frac{\partial^2 g_k}{\partial x_i \partial x_j}) \end{cases} \quad (1.20)$$

### 1.7.2 Forward on reverse

First the forward pass works the same as forward on forward, given  $s_1 = Dual(f_1(x), \frac{\partial f_1}{\partial x_j})$

1. Compute  $s_2 = f_2(s_1)$
2. Compute  $\mathbf{J}_{f_2}(s_1)$  which gives  $Dual(\mathbf{J}_2, \mathbf{H}_{2j})$

Then the backward pass, given  $r_2 = Dual((r_2)_1, (r_2)_2)$  compute:

$$\begin{aligned} r_2 \mathbf{J}_2 &= Dual((r_2)_1, (r_2)_2) Dual(\mathbf{J}_2, \mathbf{H}_{2j}) \\ &= Dual((r_2)_1 \mathbf{J}_2, (r_2)_1 \mathbf{H}_{2j} + (r_2)_2 \mathbf{J}_2) \end{aligned} \quad (1.21)$$

Given the last equation, we get the recurrence relation:

$$r_k = Dual(\frac{\partial f}{\partial s_k}, \frac{\partial^2 f}{\partial s_k \partial x_j}) \quad (1.22)$$

### 1.7.3 Reverse on forward

First we set up the forward pass, given  $s_1 = Dual(f_1(x), \frac{\partial f_1}{\partial x_i})$  (notice the difference with the previous modes ( $j \rightarrow i$ )). Then we compute:

1. Compute  $s_2 = f_2(s_1) = Dual(f_2(s_1), \mathbf{J}_2 \frac{\partial f_1}{\partial x_i})$
2. The reverse mode computes the local Jacobian

$$\frac{\partial s_2}{\partial s_1} = \begin{bmatrix} \mathbf{J}_2 & 0 \\ \mathbf{H}_{2i} & \mathbf{J}_2 \end{bmatrix} \quad (1.23)$$

Then the backward pass, which gives us:

$$\begin{cases} (r_1)_1 = (r_2)_1 \mathbf{J}_2 + (r_2)_2 \mathbf{H}_{2i} \\ (r_1)_2 = (r_2)_2 \mathbf{J}_2 \end{cases} \quad (1.24)$$

Which gives us the solution of recurrence equation:

$$r_k = Dual(\frac{\partial f}{\partial s_k}, \frac{\partial^2 f}{\partial s_k \partial x_j}) \quad (1.25)$$

### 1.7.4 Reverse on reverse

First we need to set up the forward pass, so we have  $s_2 = f_2(s_1)$ , again we need to compute its jacobian  $\mathbf{J}_2 = \frac{\partial s_2}{\partial s_1}$ . Then we need compute the backward pass with  $r_1 = r_2 \mathbf{J}_2$ . Then we need to compute again the backward pass, in order to get the second order derivative. Let  $\dot{r}_k$  be the second order reverse tangent for  $r_k$ , then we have:

$$\begin{aligned}\dot{r}_2 &= \mathbf{J}_2 \dot{r}_1 \\ \dot{s}_1 &= (r_2 \partial^2 f_2(s_1)) \dot{r}_1 + \dot{s}_2 \mathbf{J}_2\end{aligned}\tag{1.26}$$

So we can get the solution of the recurrence relation with  $\dot{s}_0 = e_i$ :

$$\begin{cases} r_k = \mathbf{J}_K \dots \mathbf{J}_{k+1} \\ \dot{r}_k = \mathbf{J}_k \dots \mathbf{J}_1 e_i \\ (r_k \partial^2 f_k) \dot{r}_{k-1} = r_k (\partial^2 f_k \dot{r}_{k-1}) \\ \quad \quad \quad = r_k \mathbf{H}_{ki} \\ \dot{s}_k = \sum_{k=1}^K r_k \mathbf{H}_{ki} \mathbf{J}_{k-1} \dots \mathbf{J}_1 \end{cases}\tag{1.27}$$

## 1.8 Implicit differentiation

Up to now, we have only considered functions that are explicitly defined, meaning that we know each step explicitly between the input and the output. But what about implicit function, when the output isn't given by an explicit formula. For example, iterative algorithms (Newton's method, Gradient Descent), or functions defined in other language and called from our main program ? As before we want to compute the derivative of these functions, but we do not know them. To solve some of those problems, we could use naive AD but it would be very inefficient because we would need to unroll the whole computation graph of the implicit function. Instead, we can use implicit differentiation, because we know that the derivatives depends on the solution and not on how you got there.

First let us consider a simple example, and show how we can get it's implicit equivalent. Consider the square root function  $x = \sqrt{a}$ , we want to rewrite it as a fixed point equation.

$$\begin{aligned}x &= \sqrt{a} \\ x^2 &= a \\ 2x^2 &= x^2 + a \\ g(x, a) &= x = \frac{1}{2} \left( x + \frac{a}{x} \right)\end{aligned}\tag{1.28}$$

With this fixed point equation, we can iterate (with  $x_{k+1} = g(x_k, a)$ ) starting from an initial guess  $x_0$  to get the solution  $x^*$  at convergence, which would be  $x = \sqrt{a}$ . The implicit equation is then  $F(x, a) = x - g(x, a)$  and it's solution satisfies  $F(x^*, a) = 0$ .

First we need to states the inverse function theorem:

**Theorem 1.1** (Inverse function theorem). If:

- $f : \mathcal{W} \rightarrow \mathcal{W}$  is  $C^2$

- $\partial f(w_0)$  is invertible.

Then  $f$  is bijective from a neighborhood of  $w_0$  to a neighborhood of  $f(w_0)$  and from this neighborhood  $w$ ,  $f^{-1}$  is  $C^2$  and:

$$\partial f^{-1}(f(w)) = \left( \partial f(f^{-1}(w)) \right)^{-1} \quad (1.29)$$

With the previous formula, that comes directly from the chain rule.

And the implicit function theorem:

**Theorem 1.2** (Implicit function theorem (IFT, univariate case)). If:

- $F : \mathbb{R}^2 \rightarrow \mathbb{R}$
- $\exists(w_0, \lambda_0)$  such that  $F(w_0, \lambda_0) = 0$  and  $\frac{\partial F}{\partial w}(w_0, \lambda_0) \neq 0$
- $F(w, \lambda)$  is  $C^2$  in a neighborhood  $\mathcal{U}$  of  $(w_0, \lambda_0)$

Then there exists a neighborhood  $\mathcal{V} \subseteq \mathcal{U}$  there exists  $w^*(\lambda)$  such that:

$$\begin{aligned} w^*(\lambda_0) &= w_0 \\ F(w^*(\lambda), \lambda) &= 0, \quad \forall (w^*(\lambda), \lambda) \in \mathcal{V} \\ \partial w^*(\lambda) &= - \frac{\frac{\partial F}{\partial \lambda}(w^*(\lambda), \lambda)}{\frac{\partial F}{\partial w}(w^*(\lambda), \lambda)} \end{aligned} \quad (1.30)$$

**Theorem 1.3** (Implicit function theorem (IFT, multivariate case)). If:

- $F : \mathcal{W} \times \Lambda \rightarrow \mathcal{W}$
- $\exists(w_0, \lambda_0)$  such that  $F(w_0, \lambda_0) = 0$  and  $\frac{\partial F}{\partial w}(w_0, \lambda_0) \neq 0$
- $F(w, \lambda)$  is  $C^2$  in a neighborhood  $\mathcal{U}$  of  $(w_0, \lambda_0)$

Then there exists a neighborhood  $\mathcal{V} \subseteq \mathcal{U}$  there exists  $w^*(\lambda)$  such that:

$$\begin{aligned} w^*(\lambda_0) &= w_0 \\ F(w^*(\lambda), \lambda) &= 0, \quad \forall (w^*(\lambda), \lambda) \in \mathcal{V} \\ \partial w^*(\lambda) &= - \frac{\frac{\partial F}{\partial \lambda}(w^*(\lambda), \lambda)}{\frac{\partial F}{\partial w}(w^*(\lambda), \lambda)} \end{aligned} \quad (1.31)$$

### 1.8.1 Implicit JVP and VJP

To propagate the derivative through the graph when we have an implicit function, we need to do it differently than before. Consider the implicit function defined by  $F(w^*, \lambda) = 0$ , we want to compute the JVP and VJP through this function.

Reminder, we want to compute the forward tangent  $t_k$ , knowing the previous tangent  $t_{k-1}$  and the function at this step. Assuming the IFT holds, we have this equality:

$$t_k = - \left( \frac{\partial F}{\partial w}(w^*, \lambda) \right)^{-1} \cdot \frac{\partial F}{\partial \lambda}(w^*, \lambda) \cdot t_{k-1} \quad (1.32)$$

Knowing that  $t_k = \partial w^*(\lambda)/\partial s_0$  and  $t_{k-1} = \partial \lambda/\partial s_0$ . We thus have a linear system to solve for each JVP for an implicit function. Setting  $A = -\frac{\partial F}{\partial w}(w^*(\lambda), \lambda)$  and  $B = \frac{\partial F}{\partial \lambda}(w^*(\lambda), \lambda)$ , we need to solve:

$$t_k = A^{-1} B t_{k-1} \quad (1.33)$$

Once the forward pass is done,  $A$  is fixed and so we can do an LU decomposition of  $A$  to solve the linear system efficiently for each JVP.

$$\begin{aligned} LU t_k &= B t_{k-1} \\ \Downarrow \\ \begin{cases} Ly &= B t_{k-1} \\ Ut_k &= y \end{cases} \end{aligned} \quad (1.34)$$

And for the VJP, we want to compute the reverse tangent  $r_{k-1}$  knowing the next reverse tangent  $r_k$  and the function at this step. Assuming the IFT holds, we have this equality:

$$r_{k-1} = - \left( \frac{\partial F}{\partial \lambda}(w^*, \lambda) \right)^T \cdot \left( \frac{\partial F}{\partial w}(w^*, \lambda) \right)^{-T} \cdot r_k \quad (1.35)$$

Knowing that  $r_{k-1} = \partial s_0/\partial \lambda$  and  $r_k = \partial s_0/\partial w^*(\lambda)$ , and with  $A = \frac{\partial F}{\partial w}(w^*(\lambda), \lambda)$  and  $B = \frac{\partial F}{\partial \lambda}(w^*(\lambda), \lambda)$ . We thus have a linear system to solve for each VJP for an implicit function. And we can do it like this:

$$\begin{cases} A^T y = r_k \\ r_{k-1} = -B^T y \end{cases} \quad (1.36)$$

## 1.8.2 AD with optimization problem

Consider the optimization problem defined by:

$$\min c^T x \quad \text{s.t.} \quad Ax = b, \quad x \geq 0 \quad (1.37)$$

and its dual:

$$\max b^T y \quad \text{s.t.} \quad A^T y \leq c \quad (1.38)$$

The KKT conditions gives us the optimality conditions for this problem:

$$\begin{cases} Ax = b \\ (A^T y - c) \perp x \geq 0 \end{cases} \quad (1.39)$$

We can rewrite these conditions as an implicit function  $F((x, y), (A, b, c))$ :

$$F((x, y), (A, b, c)) = \begin{bmatrix} Ax - b \\ \text{Diag}(x)(A^T y - c) \end{bmatrix} \quad (1.40)$$

If the IFT holds for this function, then we can compute  $\frac{\partial F}{\partial w}(w^*, \lambda)$ , which gives us:

$$\frac{\partial F}{\partial (x, y)} = \begin{bmatrix} A & 0 \\ \text{Diag}(A^T y - c) & \text{Diag}(x)A^T \end{bmatrix} \quad (1.41)$$

## 1.9 Sparse AD

Often in practice, the Jacobian matrix is sparse, meaning that many of its entries are zero. In such cases, we can use sparse automatic differentiation techniques to exploit this sparsity and reduce the computational cost of computing the Hessian. For this section, we will suppose that we know the sparsity pattern of the Jacobian matrix. Consider for example the following Jacobian matrix:

$$J = \begin{bmatrix} 0 & * & 0 & * & 0 \\ 0 & 0 & 0 & * & * \\ 0 & * & * & 0 & 0 \\ * & 0 & * & 0 & 0 \end{bmatrix} \quad (1.42)$$

If we computed the Hessian using the standard AD techniques, we would need to compute 5 JVP or 4 VJP. But what if we could group some columns or rows together to reduce the number of passes needed? Yes we can combine the columns or rows that do not share any non-zero entry. For example, with the Jacobian matrix above, we can compute the Hessian with only 2 JVP or 2 VJP. the intuition behind this comes from the fact that when we compute a JVP or a VJP, the zero entries do not contribute to the result, so we can group together the columns or rows that do not share any non-zero entry. For example, let us compute the two firsts JVPs:

$$J \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ * \end{bmatrix} \quad \& \quad J \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} * \\ 0 \\ * \\ * \\ 0 \end{bmatrix} \quad (1.43)$$

Combining these two JVPs, we get:

$$J \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} * \\ 0 \\ * \\ * \\ * \end{bmatrix} \quad (1.44)$$

And because we know the sparsity pattern we can then put the results back in the right place, this is called the decompression. So to compute the Hessian, we combine the following columns and so use those vectors for the JVPs:

$$\begin{bmatrix} 0 & * & 0 & * & 0 \\ 0 & 0 & 0 & * & * \\ 0 & * & * & 0 & 0 \\ * & 0 & * & 0 & 0 \end{bmatrix} \Rightarrow \textcolor{red}{v}_1 = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix} \& \textcolor{blue}{v}_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} \quad (1.45)$$

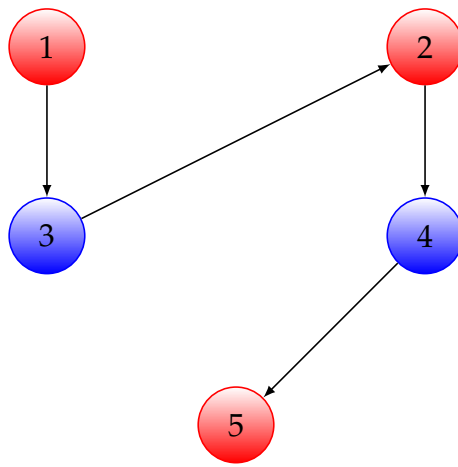
Or for the VJP:

$$\begin{bmatrix} 0 & * & 0 & * & 0 \\ 0 & 0 & 0 & * & * \\ 0 & * & * & 0 & 0 \\ * & 0 & * & 0 & 0 \end{bmatrix} \Rightarrow \textcolor{red}{v}_1 = [1 \ 0 \ 0 \ 1] \& \textcolor{blue}{v}_2 = [0 \ 1 \ 1 \ 0] \quad (1.46)$$

### 1.9.1 Sparsity detection

The problem of grouping the columns or rows of a sparse matrix can be seen as a graph coloring problem. There is multiple way to model this problem as a graph coloring problem. For example, we can create a graph where each column is a node and there is an edge between two nodes if they share a non-zero entry. Then we can use a graph coloring algorithm to color the graph such that no two adjacent nodes have the same color. Each color then represents a group of columns that can be combined together for the JVPs. The same can be done for the rows for the VJP. Here is a visual example:

$$J = \begin{bmatrix} 0 & * & 0 & * & 0 \\ 0 & 0 & 0 & * & * \\ 0 & * & * & 0 & 0 \\ * & 0 & * & 0 & 0 \end{bmatrix} \quad (1.47)$$



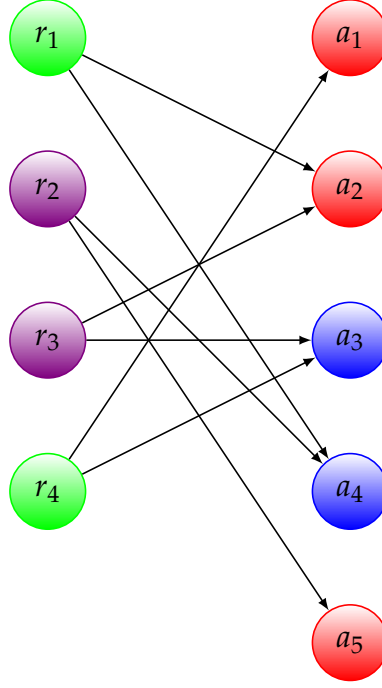
Be careful, with more complex sparsity patterns, we could have edges that link more than one times the same nodes. For example, consider the following Jacobian matrix:

$$J = \begin{bmatrix} 0 & 0 & * \\ * & * & 0 \\ * & * & 0 \end{bmatrix} \quad (1.48)$$

With that matrix, the node 1 and 2 should be linked by two edges because they share two non-zero entries. This means that when we color the graph, we need to make sure that the nodes linked by multiple edges do not share the same color.

Another way to model this problem is to create a bipartite graph where one set of nodes represents the columns and the other set represents the rows. There is an edge between a column node and a row node if the corresponding entry in the matrix is non-zero. Considering the same example as before:

$$J = \begin{bmatrix} 0 & * & 0 & * & 0 \\ 0 & 0 & 0 & * & * \\ 0 & * & * & 0 & 0 \\ * & 0 & * & 0 & 0 \end{bmatrix} \quad or \quad \begin{bmatrix} 0 & * & 0 & * & 0 \\ 0 & 0 & 0 & * & * \\ 0 & * & * & 0 & 0 \\ * & 0 & * & 0 & 0 \end{bmatrix} \quad (1.49)$$



With this modeling, the coloring problem is a distance-2 graph coloring problem, meaning that two nodes that are at distance 2 from each other cannot share the same color. This is because two columns that share a non-zero entry are connected to the same row node, so they are at distance 2 from each other. The same applies for the rows.

### 1.9.2 Symmetric matrix

When the Hessian matrix is symmetric, we can use a more efficient algorithm to compute the HVP. For example, consider the following symmetric Hessian matrix:

$$J = \begin{bmatrix} a_1 & a_2 & a_3 & 0 \\ a_2 & a_4 & 0 & a_5 \\ a_3 & 0 & a_6 & 0 \\ 0 & a_5 & 0 & a_7 \end{bmatrix} \quad (1.50)$$

With the previous sparse AD techniques, we would need to compute 3 HVP or 3 VHP, instead of 4. But if we take into account the symmetry of the matrix we could even do it in 2 HVP or 2 VHP. Indeed, we can group the columns like this:

$$J = \begin{bmatrix} a_1 & a_2 & a_3 & 0 \\ a_2 & a_4 & 0 & a_5 \\ a_3 & 0 & a_6 & 0 \\ 0 & a_5 & 0 & a_7 \end{bmatrix} \quad (1.51)$$

At first sight, it seems that the columns 2 and 3 cannot be grouped together because they share a non-zero entry at position (1,2) and (1,3). But because of the symmetry, we already have the information of those entries in the column 1 and thus we will be able to retrieve the full matrix. Let us compute the two HVPs and see how we can retrieve

the full matrix:

$$J \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} a_1 & a_2 + a_3 \\ a_2 + a_5 & a_4 \\ a_3 & a_6 \\ a_7 & a_5 \end{bmatrix} \quad (1.52)$$

And as we can see, we can solve a little linear equation system to retrieve the value of  $a_2$ , and with all the values know and knowing the sparsity pattern, we can reform the full matrix. The same can be done for the VHP by grouping the rows instead of the columns.

### 1.9.3 Star coloring

We have a graph  $G = (V, E)$ , a coloring function  $\phi : V \rightarrow \{1, \dots, p\}$ . A star coloring is distance-1 coloring with the additional constraint that every path 4 vertices uses at least 3 colors. In other words, no path of length 3 is bi-chromatic. This prevents thus length-3 dependency chains.

Let us see an example considering the same symmetric Hessian matrix as before:

$$J = \begin{bmatrix} a_1 & a_2 & a_3 & 0 \\ a_2 & a_4 & 0 & a_5 \\ a_3 & 0 & a_6 & 0 \\ 0 & a_5 & 0 & a_7 \end{bmatrix} \quad (1.53)$$

If we color its associated graph with the classic distance-1 coloring, we would get the same matrix as (1.51) and thus if we do the HVPs we would get the matrix (1.52) and we would need to solve a linear equation system to retrieve the value  $a_2$ . But if we use a star coloring, we could retrieve explicitly all the element. Let us show how:

$$J = \begin{bmatrix} a_1 & a_2 & a_3 & 0 \\ a_2 & a_4 & 0 & a_5 \\ a_3 & 0 & a_6 & 0 \\ 0 & a_5 & 0 & a_7 \end{bmatrix} \quad (1.54)$$

The use of the green color ensure the star coloring condition. Now if we compute the 3 HVPs:

$$J \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} a_1 & a_2 + a_3 & 0 \\ a_2 & a_4 & a_5 \\ a_3 & a_6 & 0 \\ 0 & a_5 & a_7 \end{bmatrix} \quad (1.55)$$

And we see that we can directly retrieve all the values of the matrix without solving any linear equation system. It is thus more efficient than the classic distance-1 coloring (4 HVPs), and more efficient than the symmetric distance-1 coloring (2 HVPs + linear system).

### 1.9.4 Acyclic coloring

We have a graph  $G = (V, E)$ , a coloring function  $\phi : V \rightarrow \{1, \dots, p\}$ . An acyclic coloring is distance-1 coloring with the additional constraint that every cycle uses at least 3 colors. This allows us to forms forests of disjoints trees with the subgraph of each color pair.

**TODO better explain HVP computation with acyclic coloring and give an example**



### 1.9.5 Chromatic number ( $\tilde{\zeta}$ )

For every graph  $G$ ,

$$\tilde{\zeta}_1(G) \leq \tilde{\zeta}_{acyclic}(G) \leq \tilde{\zeta}_{star}(G) \leq \tilde{\zeta}_2(G) \quad (1.56)$$

$\tilde{\zeta}_1$  is the classic distance-1 chromatic number,  $\tilde{\zeta}_2$  the distance-2 chromatic number.

# Neural networks

Neural networks are a class of machine learning models inspired by the structure and function of the human brain. They are composed of layers of interconnected nodes (neurons) that process and transmit information.

First let's define some variables:

- $X$ : input data (matrix)
- $y$ : target data
- $W_k$ : weights matrix at layer  $k$
- $b_k$ : bias vector at layer  $k$
- $\sigma$ : activation function (ReLU, sigmoid, etc)
- $\ell(\cdot)$ : loss function
- $H$ : number of hidden layers
- $S_i$ : intermediate state

To propagate and update the information through the network we use forward pass and backward pass and so automatic differentiation.

We can describe the forward pass of a neural network in two equivalent ways:

Right to left:	Left to right:	
$S_0 = X$	$S_0 = x$	
$S_{2k-1} = W_k S_{2k-2} + b_k$	$S_{2k-1} = S_{2k-2} W_k + b_k$	(2.1)
$S_{2k} = \sigma(S_{2k-1})$	$S_{2k} = \sigma(S_{2k-1})$	
$S_{2H+1} = W_{k+1} S_{2H}$	$S_{2H+1} = S_{2H} W_{k+1}$	
$S_{2H+2} = \ell(S_{2H+1}, Y)$	$S_{2H+2} = \ell(S_{2H+1}, Y)$	

The principal differences is that the weights acts on the left or on the right of the data, it can be useful depending whether you represents inputs as rowvectors or column-vectors.

It can be represented by this computational graph:

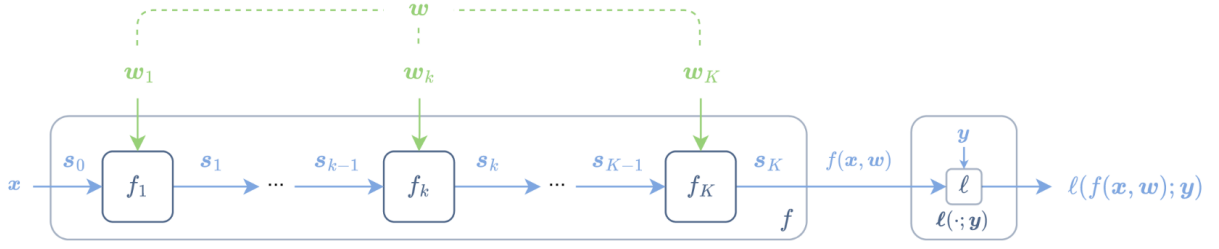


Figure 2.1: Neural network forward pass

And the backward pass can be represented like this:

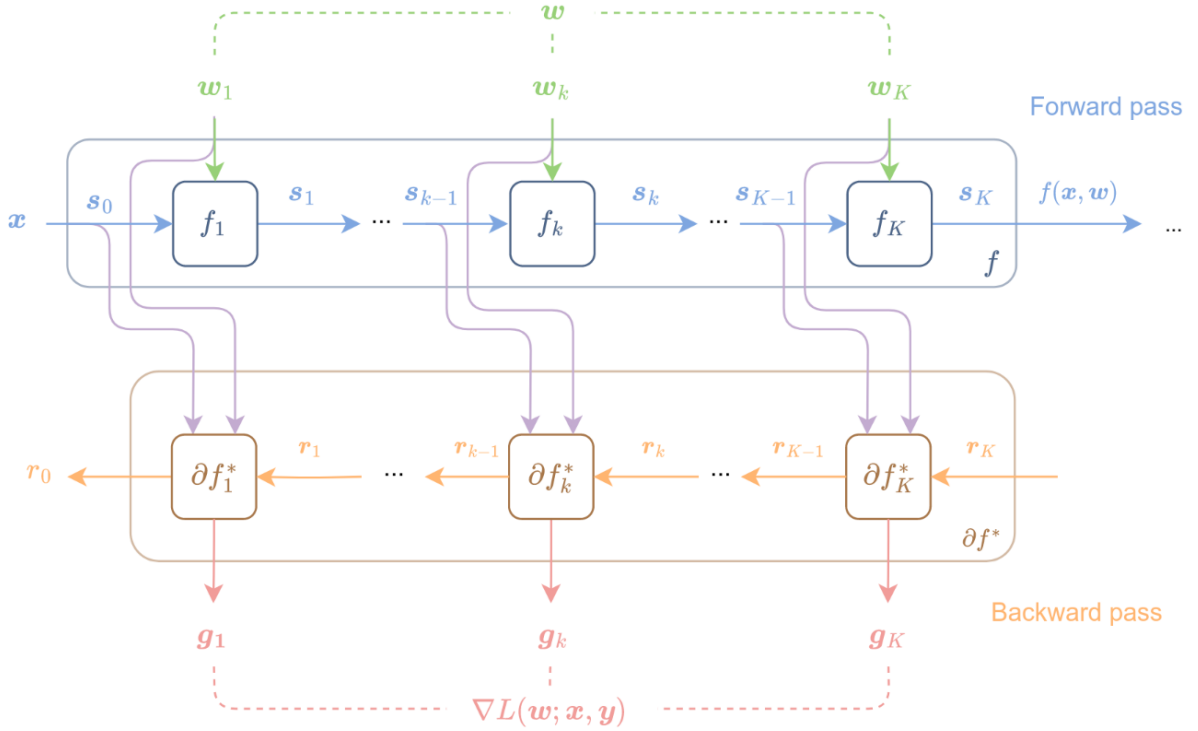


Figure 2.2: Neural network backward pass

## 2.1 Autoregressive models

An Autoregressive model wants to predict the next values in a sequence based on its previous values. Given a sequence of  $n_{cxt}$  (context) past vectors  $x_{-1}, x_{-2}, \dots, x_{-n_{cxt}}$ , the model aims to predict the next vector  $x_0$  and maybe more  $x_1$ . Example, we want to predict  $x_0$  and  $x_1$  based on the past:

$$\begin{aligned} p(x_0, x_1 | x_{-1}, \dots, x_{-n_{cxt}}) &= p(x_0 | x_{-1}, \dots, x_{-n_{cxt}}) \cdot p(x_1 | x_0, x_{-1}, \dots, x_{-n_{cxt}+1}, x_{-n_{cxt}}) \\ &\approx p(x_0 | x_{-1}, \dots, x_{-n_{cxt}}) \cdot p(x_1 | x_0, x_{-1}, \dots, x_{-n_{cxt}+1}) \end{aligned} \quad (2.2)$$

So the models aims to predict the probability distribution of the next vector  $x_0$  with  $\hat{p}(x_0 | X)$  with  $X$  that concatenates all the context vectors. And then we use the cross-entropy to measure how well the predicted distribution  $\hat{p}$  matches the true distribution

$p$ :

$$\mathcal{L}(X) = - \sum_{x_0} p(x_0|X) \log(\hat{p}(x_0|X)) \quad (2.3)$$

## 2.2 Tokenization

How can we use this autoregressive model for LLM (Large Language Models) for example ? Consider that we have  $n_{cxt}$  characters of context and we want to predict the next characters, this means that we would have a one-hot encoding in  $\mathbb{R}^{26}$ , which means  $n_{voc} = 26$ . This means that  $n_{cxt}$  should be a very large number, but this is annoying because transformers have a quadratic complexity in  $n_{cxt}$ .

Another idea could be to turn each word into a one-hot encoding, but the vocabulary size is often very large (+200k words), with this we could have a relative low  $n_{cxt}$ , but  $n_{voc}$  would be too big.

An intermediate solution is to use byte pair encoding algorithm which greedily merges the most frequent pairs of characters into new tokens.

Example: consider the word "abracadabra", we see that we have two frequent pairs "ab" and "ra", so we can merge them into new tokens  $X$  and  $Y$  respectively:

$$\text{abracadabra} \rightarrow \text{XYcadXY} \quad (2.4)$$

then we can repeat the process until we reach the desired vocabulary size.

The next iteration could give:

$$\text{XYcadXY} \rightarrow \text{ZcadZ} \quad (2.5)$$

This tokenization method allows us to have good trade-off between the vocabulary size  $n_{voc}$  and the context size  $n_{cxt}$ . But if the model isn't trained enough we could have an issue which is that the model doesn't output fully constructed words, for example it could output half of a word because that what makes more sense even though the word is half complete.

## 2.3 Embedding

Consider a vocabulary of size  $n_{voc}$ , a bigram model and a network with  $d$  layers. The model would be:

$$\hat{p}(x_0|x_{-1}) = \text{softmax}(W_d \tanh(\dots \tanh(W_1 x_{-1}) \dots)) \quad (2.6)$$

The matrix  $W_1$  has  $n_{voc}$  columns and  $W_d$  has  $n_{voc}$  rows, so when  $n_{voc}$  is large, the model becomes very big.

So an idea would be to use an encoder and a decoder to reduce the sizes, it is called an embedding size  $d_{emb} \ll n_{voc}$ . The encoder ( $C \in \mathbb{R}^{d_{emb} \times n_{voc}}$ ) maps the one-hot encoding of size  $n_{voc}$  to a dense vector of size  $d_{emb}$  and the decoder ( $D \in \mathbb{R}^{n_{voc} \times d_{emb}}$ ) maps back the dense vector to a vector of size  $n_{voc}$ . So the model becomes:

$$\hat{p}(x_0|x_{-1}) = \text{softmax}(DW_d \tanh(\dots \tanh(W_1 C x_{-1}) \dots)) \quad (2.7)$$

If we choose wisely  $d_{emb}$ , which means much smaller than  $n_{voc}$ , then it is faster to compute  $W_1(Cx_{-1})$  than in the previous model. Moreover we are forcing  $W_1C$  to be low-rank which can help to reduce overfitting but reduce the expressiveness (capacity to capture a range of possible relation between input and output) of the model. It is useful to share the embedding between different input and output when  $n_{ctx} > 1$ , we also found that forcing  $D = C^T$  appears to work well in practice.

### 2.3.1 Shared embedding

Let's investigate the case where we have  $n_{ctx} > 1$ . When  $n_{ctx} > 1$ , the encoder  $C$  is shared by all tokens, so we get this model:

$$\hat{p}(x_0|x_{-1}, \dots, x_{-n_{ctx}}) = \text{softmax}(DW_d \tanh(\dots \tanh(W_1 \begin{bmatrix} Cx_{-1} \\ \vdots \\ Cx_{-n_{ctx}} \end{bmatrix}) \dots)) \quad (2.8)$$

We can note that now  $W_1$  has  $n_{ctx}d_{emb}$  columns. Assuming  $d_{emb} \ll n_{voc}$  and  $n_{ctx} \gg 1$ , this is much smaller than the  $n_{ctx}n_{voc}$  that we had before the embedding. The number of rows of  $W_2$  is not affected by the embedding so it remains  $n_{voc}$ .

We could represent this model like this:

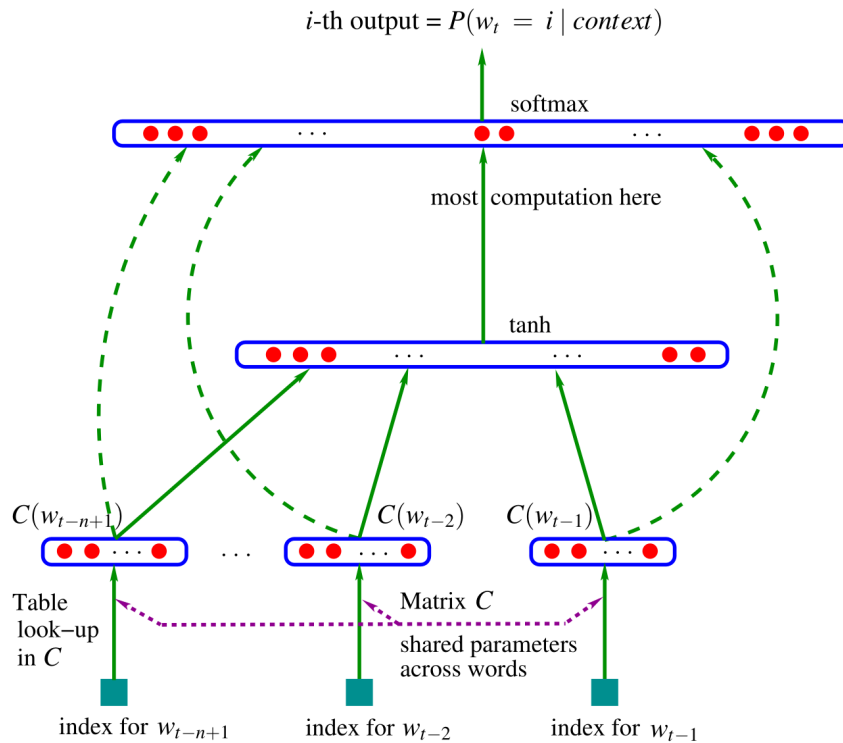


Figure 2.3: Neural network with shared embedding

## 2.4 Recurrent neural networks (RNN)

A RNN is a type of neural network that is designed to process sequential data by taking advantages of the memory of previous inputs. The idea is to have a hidden state  $h_t$  that

depends on the current input  $x_t$  and the previous hidden state  $h_{t-1}$ . With this idea, the network reuses the same weights at every time step and thus gives meaning to the sequence. The equations of a simple RNN are:

$$\begin{cases} h_{t+1} = \tanh(Wh_t + Ux_{t+1} + b) \\ \hat{y}_t = \text{softmax}(Vh_t + c) \end{cases} \quad (2.9)$$

Where  $W, U, V$  are weight matrices and  $b, c$  are bias vectors.

This model presents some limitations, for example it is difficult to learn long-term dependencies because of the vanishing/exploding gradient problem. The fact that it processes a sequence step by step makes it difficult to parallelize the computations. The fact that it need to store the hidden state at each time step can be also memory-intensive for long sequences.

## 2.5 Attention head

First we need to define a numerical dictionary. Consider keys  $k_i \in \mathbb{R}^{d_k}$ , values  $v_i \in \mathbb{R}^{d_v}$ . Given a query  $q \in \mathbb{R}^{d_k}$ , we want to retrieve the value  $v_i$  corresponding to the key  $k_i$  that is the most similar to the query  $q$ . To do this, we can use the dot-product. With that definition, we can define the attention head:

$$\text{Attention}(Q, K, V) = \sum_{i=1}^{n_{ctx}} \alpha_i v_i \quad (2.10)$$

where  $\alpha = \text{softmax}(\langle q, k_1 \rangle, \dots, \langle q, k_{n_{ctx}} \rangle)$  is the attention weight for key  $k_i$ . But in practice, we have vectors of queries, keys and values, each contained in matrices  $Q, K, V$ . So we can rewrite the attention head as:

$$\text{Attention}(Q, K, V) = V \text{softmax} \left( \frac{K^T Q}{\sqrt{d_k}} \right) \quad (2.11)$$

Where the division by  $\sqrt{d_k}$  is used to prevent the dot-product from growing too large. For a transformers, we could explain  $Q, K, V$  like this, for each token we have a vector  $q, k, v$  that represent:

- $q$ : What am I looking for?
- $k$ : What do I contain?
- $v$ : What information do I pass if I'm selected?

### 2.5.1 Masked attention

To ensure that the model only attends to previous tokens and not future tokens, we can use a masked attention mechanism. This is done by applying a mask to the attention weights before the softmax operation. The mask is typically a lower triangular matrix that sets the weights corresponding to future tokens to negative infinity, effectively

preventing them from contributing to the attention output. So with masked attention, the equation becomes:

$$\left\{ \begin{array}{l} M = \begin{bmatrix} 0 & 0 & \dots & 0 \\ -\infty & 0 & \ddots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ -\infty & -\infty & \dots & 0 \end{bmatrix} \\ \text{Attention}(Q, K, V) = V \text{softmax} \left( M + \frac{K^T Q}{\sqrt{d_k}} \right) \end{array} \right. \quad (2.12)$$

## 2.5.2 Multi-head attention

Instead of having a single attention head, we can have multiple attention heads that can understand different aspects of the input data. Each head has its own set of weights and computes its own attention output, all in parallel. After computing all the heads, their outputs are concatenated and linearly transformed to be in the right dimension with  $W^O$ , this is thus gathering all of the meanings of the attention different heads to a single representation. So the multi-head attention can be defined as:

$$\left\{ \begin{array}{l} \text{head}_j = \text{Attention}(W_j^Q Q, W_j^K K, W_j^V V) \\ \text{MultiHead}(Q, K, V) = W^O \begin{bmatrix} \text{head}_1 \\ \text{head}_2 \\ \vdots \\ \text{head}_h \end{bmatrix} \end{array} \right. \quad (2.13)$$

## 2.6 Decoder-only transformer

A decoder-only transformer is a type of transformer architecture that is used for autoregressive model. It consists of reading the tokens and predicting the next token in the sequence. In this type of transformer, the matrices  $Q, K, V$  are the same, and we define them as the input tokens  $CX$ ,  $C$  being the embedding matrix. So the decoder-only transformer can be computed with  $\text{MultiHead}(CX, CX, CX)$ . Let's seek the different mechanisms used in a decoder-only transformer.

### 2.6.1 Positional encoding

The positional encoding is essential in a transformer model because it provides information about the order of the tokens in the input sequence. Since transformers do not have a built-in notion of sequence order like RNNs, positional encodings are added to the input embeddings to give the model a sense of position.

We cannot use one-hot encoding for the position because the dimension would be  $n_{ctx}$  and not  $d_{emb}$  as expected. The classic approach is to use sines and cosines with an angle that depends on the position of the token and the embedding dimension. The positional encoding for a token at position  $pos$  and with the embedding dimension

index  $i$  and embedding dimension  $d_{emb}$  could be defined as:

$$\begin{cases} angle = \frac{pos}{10000^{2i/d_{emb}}} \\ PE[pos, 2i] = \sin(angle) \\ PE[pos, 2i + 1] = \cos(angle) \end{cases} \quad (2.14)$$

And then we add this positional encoding to the input embeddings before doing the attention mechanism:

$$\text{MultiHead}(CX + PE, CX + PE, CX + PE) \quad (2.15)$$

### 2.6.2 Residual connection

The principle of residual connection is to add the input of a layer to its output before applying the layer normalization (next subsection). This helps to mitigate the vanishing gradient problem and help the network to learn because its learning is more controlled, because we remind it of the original input.

### 2.6.3 Layer normalization

The norm of the gradient increases exponentially with the number of layers, so to prevent this we can use layer normalization. This helps stabilize the training process and improve convergence, by rescaling the data before the activation function, which means that we control the mean and the variance of the data. There is two way of doing this normalization, the batch normalization and the layer normalization. The batch normalization is difficult and not very efficient to use in transformers, so we use layer normalization. It is done like this:

$$\text{LayerNorm}(y_i) = \gamma \frac{y_i - \mu_i}{\sigma_i} + \beta \quad (2.16)$$

With  $\gamma$  and  $\beta$  two learnable parameters,  $\mu_i$  the mean of the elements of  $y_i$  and  $\sigma_i$  the standard deviation of the elements of  $y_i$ . Then we can add this in the transformer architecture like this:

$$\text{LayerNorm}(\text{MultiHead}(CX + PE, CX + PE, CX + PE) + (CX + PE)) \quad (2.17)$$

### 2.6.4 Feed-forward network (FFN)

The feed-forward network consists in doing this operation:

$$\text{FFN}(x) = W_2 \text{ReLU}(W_1 x + b_1) + b_2 \quad (2.18)$$

With  $W_1 \in \mathbb{R}^{d_{ff} \times d_{emb}}$  and  $W_2 \in \mathbb{R}^{d_{emb} \times d_{ff}}$ ,  $d_{ff}$  being the dimension of the feed-forward network, generally  $d_{ff} = 4d_{emb}$ . This feed-forward network is applied to each token independently and identically. The feed-forward network is used to introduce non-linearity and increase the model's capacity to learn complex patterns in the data. It is used the memorize more complex information in the weights, which helps the model to better understand the complex relationships between tokens in the sequence.



## 2.6.5 Transformers variation

To investigate ?

## 2.7 Performances of transformers

Before analysing the complexity of a transformer, let us remind the main dimension of the variables:

- $n_{ctx}$ : context size (number of tokens in the input sequence)
- $n_{voc}$ : vocabulary size (number of unique tokens)
- $d_{emb}$ : embedding size (dimension of embeddings)
- $d_k, d_v$ : key and value dimension (per head)
- $d_{ff}$ : feed-forward network dimension
- $N$ : number of layers

We can thus analyse the time complexity of a transformer step by step (ignoring the embedding step):

1. Linear projections for  $Q, K, V$  (with  $W_j^Q, W_j^K, W_j^V$ ):  $\mathcal{O}(d_k d_{emb} n_{ctx})$
2. Attention score computation ( $K^T Q$ ):  $\mathcal{O}((d_k + d_v) n_{ctx}^2)$
3. Output projection ( $W^O$ ):  $\mathcal{O}(d_v d_{emb} n_{ctx})$
4. FFN:  $\mathcal{O}(d_{emb} d_{ff} n_{ctx})$

This brings the total cost for a single layer to:

$$\mathcal{O}((d_k + d_v) n_{ctx}^2 + (d_k + d_v + d_{ff}) d_{emb} n_{ctx}) \quad (2.19)$$

To get the total cost for  $N$  layers, we just need to multiply by  $N$ .

$$\mathcal{O}(N(d_k + d_v) n_{ctx}^2 + N(d_k + d_v + d_{ff}) d_{emb} n_{ctx}) \quad (2.20)$$

Knowing that generally  $d_k \approx d_v \approx d_{emb} \approx d_{ff}$ , we can simplify the complexity to:

$$\mathcal{O}(N d_{emb} n_{ctx}^2 + N d_{emb}^2 n_{ctx}) = \mathcal{O}(N d_{emb} n_{ctx} (n_{ctx} + d_{emb})) \quad (2.21)$$

Here we can see the two dominant terms of the complexity:

- $N d_{emb} n_{ctx}^2$ : comes from the attention score computation, it is the most critical term when  $n_{ctx}$  is large.
- $N d_{emb}^2 n_{ctx}$ : comes from the FFN and the projection, it is the most critical term when  $d_{emb}$  and thus for wide models.

We can also observe that the dimension of the vocabulary  $n_{voc}$  does not appear in the complexity. It is because it is only involved in the embedding complexity.

# Diffusion Models

The objective of diffusion models is to generate data by learning how to reverse a gradual noising process, turning random noise into structured samples through iterative denoising.

## 3.1 Tweedie's formula

First, we need to understand why diffusion models can be trained by denoising and why predicting the noise is equivalent to learning a score. For that, we need to introduce Tweedie's formula. Consider that we observe a noisy version of a random variable  $X$ . The noisy observation is given by  $Y = X + \sigma\epsilon$ , where  $\epsilon \sim \mathcal{N}(0,1)$ . We want, given this noisy observation  $Y = y$ , to estimate the original variable  $X$ . The MMSE estimator is given by  $\mathbb{E}[X|Y = y]$ . Tweedie's formula states that:

$$\mathbb{E}[X|Y = y] = y + \sigma^2 \nabla_y \log f_Y(y) \quad (3.1)$$

where  $f_Y(y)$  is the probability density function of the noisy observation  $Y$ . And thus  $\nabla_y \log f_Y(y)$  is a score function that tells us in which direction the probability mass increases and thus where cleaner data is. We can also estimate the noise  $\epsilon$ :

$$\mathbb{E}[\epsilon|Y = y] = -\sigma \nabla_y \log f_Y(y) \quad (3.2)$$

And this means that if we can predict the noise, we can also estimate the score function and thus the denoised data. This is the principle used in diffusion models.

## 3.2 Langevin dynamics (sampling)

Langevin dynamics is a way to sample from a probability distribution ( $p(y)$ ) when you only know its log-density gradient (the score), not the density itself. The idea is to iteratively update a sample by taking small steps in the direction of the score, while also adding some random noise to ensure exploration of the space. The update rule for Langevin dynamics is given by:

$$y_{k+1} = y_k + \delta_k \nabla_y \log p(y_k) + \sqrt{2\delta_k} w_k \quad (3.3)$$

with  $w_k \sim \mathcal{N}(0,1)$  and  $\delta_k$  is a small step size. By repeating this process many times, the samples  $y_k$  will converge to the target distribution  $p(y)$ . Adding this noise is important because it helps the samples to explore the space more effectively and avoid getting stuck in local modes of the distribution. The Langevin dynamics is seen in diffusion models at each denoising step.

### 3.3 Score matching

This is the part "learning the score" of diffusion models. Here the idea is to fix a noise level  $\sigma$  and to corrupt the data  $X$  with Gaussian noise to get  $Y = X + \sigma\epsilon$ , with  $\epsilon \sim \mathcal{N}(0, 1)$ . Then we want to minimize the error to learn the noisy data distribution ( $p_{X+\sigma\epsilon}$ ):

$$\mathbb{E} \left[ \|\epsilon_\theta(X + \sigma\epsilon) - \epsilon\|^2 \right] \quad (3.4)$$

But with this,  $\sigma$  needs to be scaled properly, otherwise the model will not learn well. If  $\sigma$  is too small, then the support of  $X + \sigma\epsilon$  may not cover the whole space, and thus  $\epsilon_\theta(y)$  may be inaccurate. If  $\sigma$  is too large, then the noise dominates the signal, and the model may struggle to learn meaningful patterns.

#### 3.3.1 Variance dependent score

To address the issue of choosing the right noise level  $\sigma$ , we can introduce a variance-dependent score function. Instead of using a fixed  $\sigma$ , we can define a score function that depends on the noise level. The idea is to train a model  $\epsilon_\theta(y, \sigma)$  that takes both the noisy observation  $y$  and the noise level  $\sigma$  as inputs. The training objective becomes:

$$\mathbb{E} \left[ \|\epsilon_\theta(X + \sigma\epsilon, \sigma) - \epsilon\|^2 \right] \quad (3.5)$$

# Kernels

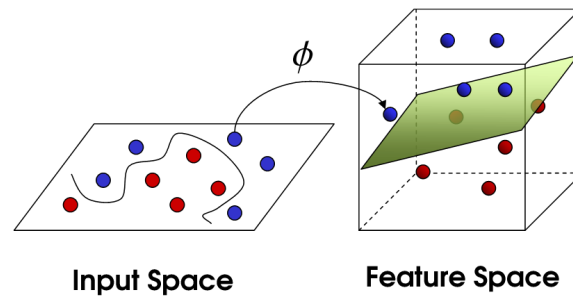


Figure 4.1: Illustration of kernels

A kernel is a function that transforms a dataset into another, typically of higher dimension. This helps separate the nonlinear feature, to use the usual linear tools. For example, the canonical kernel is

$$r(x, y) = (x^T y)^2 \quad (4.1)$$

where the kernel function then is

$$\phi(x) = \begin{pmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}x_1x_2 \end{pmatrix} \quad (4.2)$$

## 4.1 Reminders on scalar product

→ Reminder: a Euclidean space is a finite-dimensional vector space endowed with a scalar product.

A scalar product  $\langle \cdot, \cdot \rangle_V$  verifies

- Symmetry:  $\forall x, y \in V, \langle x, y \rangle_V = \langle y, x \rangle_V$ ;
- Definite positive: if  $x \in V$  and  $x \neq 0$ , then  $\langle x, x \rangle > 0$ , and if  $x = 0$ , then  $\langle x, x \rangle = 0$ ;
- Bilinearity:  $\forall x, y, z \in V, \forall \alpha, \beta \in \mathbb{F}, \langle (\alpha x + \beta y), z \rangle_V = \alpha \langle x, z \rangle_V + \beta \langle y, z \rangle_V$ .

### 4.1.1 Equivalence

Consider a positive definite symmetric matrix  $M \succ 0 \in \mathbb{R}^{n \times n}$  with the scalar product  $\langle x, y \rangle_M = x^T M y$ . Any such matrix can be factored as  $M = L^T L$ , where  $L$  is invertible. We can do a change of coordinates to re-express under the cartesian scalar product:

$$\begin{cases} w = Lx \\ z = Ly \end{cases} \implies \langle x, y \rangle_M = \langle w, z \rangle_{\mathbb{R}^n} \quad (4.3)$$

Then, all  $n$ -dimensional Euclidean spaces are equivalent up to a change of coordinates.

### 4.1.2 Hilbert space

For functions integrable on an interval  $[a, b]$ , we define the scalar product as

$$\langle f, g \rangle = \int_a^b f(t)g(t)dt \quad (4.4)$$

## 4.2 Kernel methods for finite sets

### 4.2.1 Example – word embedding

Text embedding has as objective to represent a text with a vector. The general idea is one-hot encoding, i.e. for a data set  $X$  of  $N$  words, we create one-hot vectors of dimension  $N$ , using the feature map (or embedding)  $\phi : X \rightarrow H$ . A sentence is simply the sum of those vectors, weighted by the number of occurrences. From this, we can create the kernel matrix  $K$ , where each element  $K_{ij}$  is the comparison between the basis vectors  $e_i, e_j$ . We suppose that the  $\phi$  function has a unitary norm.

$$K_{ij} = \langle \phi(x_i), \phi(x_j) \rangle_K \in [-1, 1] \quad (4.5)$$

A value of 1 means that the words  $x_i$  and  $x_j$  are identical, but if the value is  $-1$ , their meaning are opposite. In the case of a 0, there is no connection between the words. Now, we can calculate the Cholesky decomposition  $K = L^T L$  for a new basis, and use the usual scalar product.

→ Note: if the matrix  $K$  is not invertible, we just hit a degenerate case where the  $\phi(x_i)$  are not linearly independent. This is not a problem.

### 4.2.2 Kernel trick

The kernel trick is to never explicitly define the function  $\phi$ : we can work with the matrix  $K$  only, as any vector can be expressed as a linear combination of  $\phi(x_i)$  and the main ingredient of linear methods is generally the scalar product, here defined only using  $K$ .

$$\langle \phi(y), \phi(z) \rangle = \left\langle \sum_{i=1}^N \alpha_i \phi(x_i), \sum_{i=1}^N \beta_i \phi(x_i) \right\rangle_K = \sum_{i=1}^N \sum_{j=1}^N \alpha_i \beta_j \langle \phi(x_i), \phi(x_j) \rangle_K = \alpha^T K \beta \quad (4.6)$$

### 4.2.3 Reconstruction

Given the matrix  $K$ , we can reconstruction the embedding ( $\phi$  and  $\langle \cdot, \cdot \rangle_K$ ) up to a rotation. For  $X$  the set of words and  $N$  the dimension of the space, we have  $\phi : X \rightarrow \mathbb{R}^N$  and we define the scalar product:

$$\langle x, y \rangle_M = \langle x, y \rangle_{K^{-1}} \quad (4.7)$$

Then, remembering  $\phi(x_i) = Ke_i$ ,

$$\langle \phi(x_i), \phi(x_j) \rangle_M = (Ke_i)^T K^{-1} (Ke_j) = e_i^T K^T K^{-1} Ke_j = e_i^T Ke_j \quad (4.8)$$

Let us prove the "up to a rotation" part. Define

$$\begin{cases} \phi' : X \rightarrow \mathbb{R}^N \\ \phi'(x) = Q\phi(x) \\ \langle x, y \rangle_{M'} = \langle x, y \rangle_{QK^{-1}Q^T} \end{cases} \quad (4.9)$$

where  $Q$  is a rotation matrix, i.e.  $Q^T Q = I$ . Then,

$$\begin{aligned} \langle \phi'(x), \phi'(y) \rangle_{M'} &= (Q\phi(x))^T QK^{-1}Q^T (Q\phi(y)) \\ &= \phi(x)^T K^{-1} \phi(y) = \langle \phi(x), \phi(y) \rangle_{K^{-1}} \end{aligned} \quad (4.10)$$

The result is true for any rotation matrix  $Q$ .

→ Note: the complexity of computing  $K$  is  $\mathcal{O}(N^2)$ , while the Cholesky decomposition to find the new basis has a complexity of  $\mathcal{O}(N^3)$ . This decomposition is to be avoided if not really necessary.

### 4.2.4 Advantages

Let us consider proteins. The alphabet is of size 20 (number of existing amino-acids), and a protein is made of 4 of these. This means that our space is initially  $H = \mathbb{R}^{20^4} = \mathbb{R}^{160.000}$ . Consider  $N = 100$  proteins to classify.

The embedding consists in storing  $N$  vectors of size  $h = 160.000$ , meaning 1.6M scalars. However, we only need the  $K$  matrix of size  $N^2 = 10.000$ , and the entries are easy to compute. This means that, using the kernel trick, i.e. working with  $K$  directly, we are much more efficient when  $N \ll h$ .

## 4.3 Kernels methods for continuous sets

### 4.3.1 Reproducibility Kernel Hilbert Space

Consider richer sets than finite  $X$ , e.g. infinite or uncountable sets, with distances defined but not scalar product. Then, let  $H$  be a vector space of continuous functions  $X \rightarrow \mathbb{R}$ .  $H$  is a RKHS if it is a Hilbert space and

$$\forall x \in X, \exists v \in H \text{ such that } \forall f \in H : \langle v, f \rangle_H = f(x) \quad (4.11)$$

Mathematically, the kernel function is the equivalent of the kernel matrix, but for infinite spaces:

$$\phi : X \rightarrow \mathbb{R}^N \text{ such that } k(x, y) = \langle \phi(x), \phi(y) \rangle_K = \alpha_x^T K \alpha_y \quad (4.12)$$

For a RKHS  $H$ , we verify

$$\begin{cases} \forall x \in X : k(x, \cdot) \in H \\ \forall x \in X, \forall f \in H, f(x) = \langle f, k(x, \cdot) \rangle_K \end{cases} \quad (4.13)$$

The reproducing kernel property is

$$k(x, y) = \langle k(x, \cdot), k(y, \cdot) \rangle_H \quad (4.14)$$

### 4.3.2 Properties

**Theorem 4.1.** A continuous map  $k : X \times X \rightarrow \mathbb{R}$  is the kernel of some RKHS  $H \subset \mathcal{C}(X, \mathbb{R})$

- iff  $k(x, y) = \langle \phi(x), \phi(y) \rangle_H$  for some feature map  $\phi : X \rightarrow H$ ;
- iff, for all finite subsets  $X_0 \subset X$ , the kernel matrix is symmetric positive definite.

## 4.4 Polynomial kernels

Polynomial kernels are a particular type of continuous kernels, where the feature map is a function

$$\phi : \mathbb{R}^n \rightarrow \mathbb{R}^\ell \quad (4.15)$$

Let us take our initial example 4.2, and generalize it to

$$k(x, y) = (x^T y)^d \quad X = \mathbb{R}^n \quad (4.16)$$

Then, the dimension of the out space  $H = \mathbb{R}^\ell$  is  $\ell = \binom{n+d-1}{d}$ .