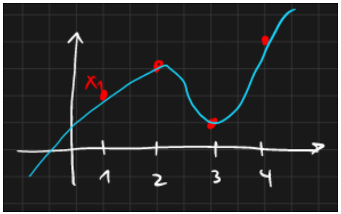
Calculus

8 Nonlinear Aspects

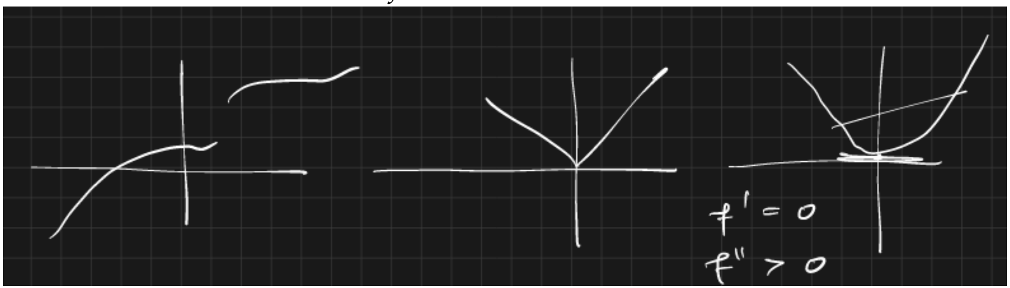
We will touch upon the following topics:

- continuous and differentiable functions
- partial derivatives, gradient, Jacobian
- (chain rule)
- In exercise: Taylor approximation and Newton's method

Until now we have worked with "discrete objects", say $x \in \mathbb{R}^n$, $\{1, \ldots, n\} \to \mathbb{R}$, $i \mapsto x_i$



Now, vectors become functions $f : \mathbb{R} \to \mathbb{R}$



8.1 Motivation

Let us first recall the definition of a linear function. Consider a function $f : \mathbb{R}^n \to \mathbb{R}^m$, then

$$f \text{ linear } :\overset{\text{Def}}{\Leftrightarrow} \forall\, x, y \in \mathbb{R}^n,\ \lambda \in \mathbb{R}:\ f(\lambda \cdot x + y) = \lambda \cdot f(x) + f(y).$$

The prototype of a linear function between finite dimensional spaces is the matrix–vector product, more precisely,

$$A \in \mathbb{R}^{m\times n},\ f_A(x) := Ax.$$

We say f is **nonlinear**, if it is not linear.

Nonlinear function may extend our modeling choice significantly and may help to explain complicated relations, such as

$$\underset{[\text{image}]}{\underset{\in \mathbb{R}^p}{z_i}} \quad \to \quad \underset{[\text{feature}]}{\underset{\in \mathbb{R}^q}{y_i}}, \ i = 1, \ldots, m.$$

Until now, we have consider models with *linear* dependency of the parameters:

$$f_x(z) = \sum_{k=1}^{n} x_k \cdot f_k(z) \approx y.$$

We determined the parameters $x = (x_k)_k$ by solving a (potentially regularized) least squares problem of the form

$$\min_x L(x; (z_i, y_i)) + R(x), \quad \left(\text{e.g., Ridge Regression } R(x) := \frac{\delta}{2}\|x\|_2^2\right),$$

where the cost function has the form

$$\sum_{i=1}^{m} \|f_x(z_i) - y_i\|_2^2 = \|A_z x - y\|_2^2 =: L(x, (z_i, y_i)).$$

The specialty of this kind of minimization problem is that we can solve it via the normal equation, which is a *linear* equation.

Now let us consider a **nonlinear model** (e.g. Neural Network); more precisely, nonlinear with respect to the sought-after parameters. More specifically, let us for example consider a model of the form

$$f_x(z) = (f_M \circ \cdots \circ f_1)(z) = f_M(f_{M-1}(f \ldots (f_1(z)) \ldots))$$

where the building blocks f_k, also called **layers**, are given by

$$f_k \colon \mathbb{R}^p \to [0, +\infty)^q, \quad f_k(z) := (A_k z + b_k)_+ \quad \textit{(applied element-wise)},$$

with

$$\mathbb{R} \to [0, +\infty), \quad w_+ := \begin{cases} 0 : w < 0 \\ w : \text{ else} \end{cases}$$

being the so-called **ReLU function** (Rectified Linear Unit), an example of a so-called **activation function**.

The matrices $A_k \in \mathbb{R}^{q \times p}$ and vectors $b_k \in \mathbb{R}^q$ are the parameters (also called **weights**) that need to be determined. If A_k is dense, the function f_k is called fully connected layer and if, e.g., A_k is Toeplitz, then f_k is called **convolutional layer**.

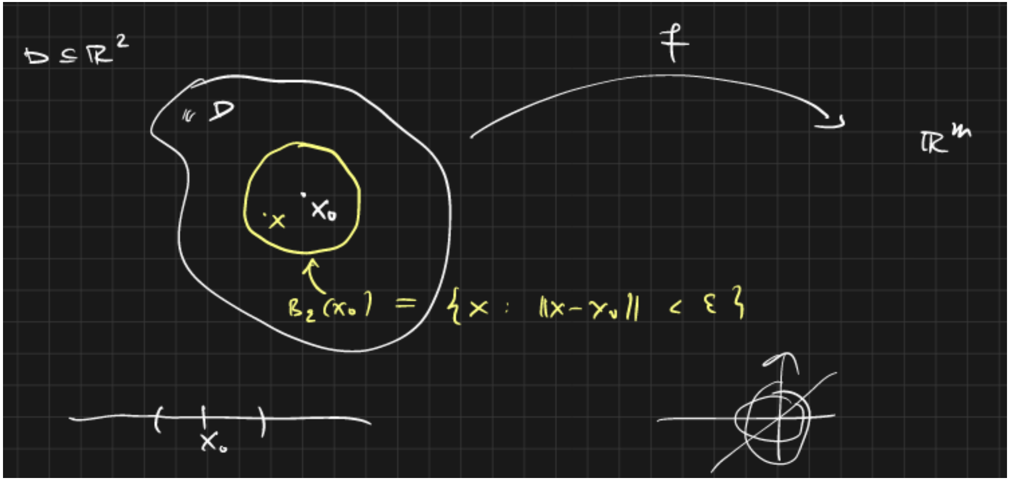
Due to the ReLU function $(\cdot)_+$ the concatenated model f_x is highly nonlinear.

Similarly to the linear case, we aim to find suitable parameters/weights $x := (A_k, b_k)_k$ that best describe the model with respect to a certain cost function:

$$\min_{x := (A_k, b_k)_k} L(x; (z_i, y_i)) + R(x) =: F(x) \qquad (\leftarrow \ F \text{ highly nonlinear})$$

Before we continue with some standard definitions from calculus, a preliminary remark:

The concepts of continuity and differentiability in the context of a function $f : \mathbb{R}^n \to \mathbb{R}^m$ are "local" concepts, i.e., they are required to hold in a small neighborhood of a point $x_0 \in \mathbb{R}^n$.



8.2 Continuity and Differentiability

In the following we consider neighborhoods of the form $B_\varepsilon(x_0) := \{x \in \mathbb{R}^n \colon \|x - x_0\| < \varepsilon\}$.

Definition 8.1 (*Continuous and differentiable function*)
Let $D \subseteq \mathbb{R}^n$, $f : D \to \mathbb{R}^m$ and $x_0 \in D$ with $B_\varepsilon(x_0) \subseteq D$ for some $\varepsilon > 0$. Then

i) *f is called **continuous** at x_0, if*

$$\lim_{n\to 0} \|f(x_n) - f(x_0)\|_2 = 0$$

for all sequences $(x_n)_{n\in\mathbb{N}} \subseteq B_\varepsilon(x_0)$ for which $x_n \to x_0$.

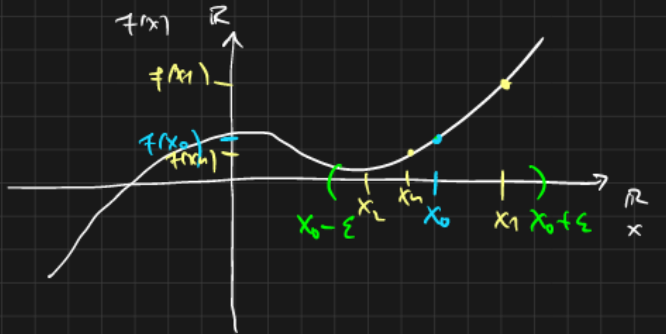
ii) *f is called **differentiable** at x_0, if there is a linear mapping $A : \mathbb{R}^n \to \mathbb{R}^m$ such that*

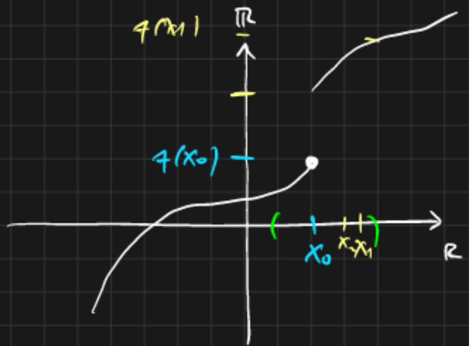
$$\lim_{n\to\infty} \frac{\|(f(x_0) + Ah_n) - f(x_0 + h_n)\|}{\|h_n\|} = 0$$

for all sequences $(h_n)_n$ with $x_0 + h_n \subseteq B_\varepsilon(x_0)$, $\lim_{n\to\infty} \|h_n\| \to 0$.
Since the linear function A depends on f and x_0, we denote it as $Df(x_0) := A$ and call it (Fréchet) derivative.

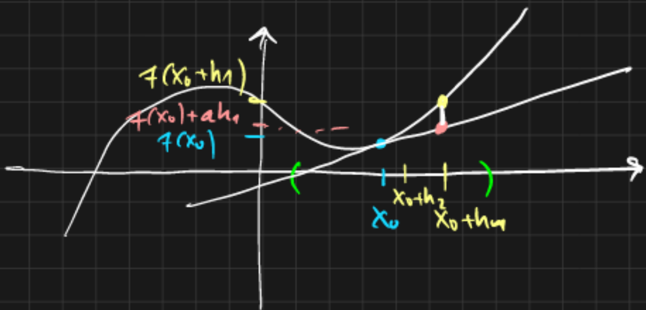
If f is continuous/differentiable at any point $x_0 \in D$, we call f simply continuous/differentiable.

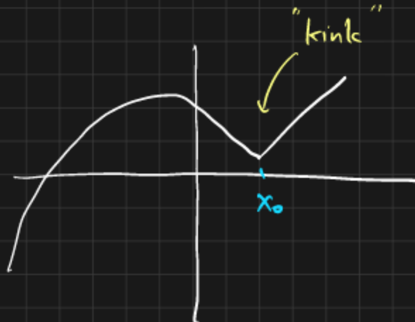
Continuity: $f: \mathbb{R} \to \mathbb{R}$





Differentiable: $f: \mathbb{R} \to \mathbb{R}$, $Df(x_0)$, $\frac{1}{|h|} \left| f(x_0) + a \cdot h - f(x_0+h) \right|$





Examples: Continuity

i) $f : \mathbb{R} \to \mathbb{R},\ x \mapsto |x| = \begin{cases} x : & x \geq 0 \\ -x : & x < 0 \end{cases}$

Let $x_0 \in \mathbb{R},\ (x_n)_{n\in\mathbb{N}},\ x_n \xrightarrow{n\to\infty} x_0$, then

$$0 \leq |f(x_n) - f(x_0)| = ||x_n| - |x_0|| \leq |x_n - x_0| \xrightarrow{n\to 0} 0$$

$\Rightarrow \quad f$ is continuous.

ii) $f : \mathbb{R} \to \mathbb{R},\ x \mapsto x^2$

Let $x_0 \in \mathbb{R},\ x_n \to x_0$, then

$$|f(x_n) - f(x_0)| = |x_n^2 - x_0^2| = |(x_n - x_0)(x_n + x_0)| = \underbrace{|x_n - x_0|}_{\to 0}\,\underbrace{|x_n + x_0|}_{\to 2x_0} \xrightarrow{n\to\infty} 0$$

$\Rightarrow \quad f$ is continuous.

iii) $f : \mathbb{R} \to \mathbb{R}, f(x) := \begin{cases} 1 : & x > 0 \\ -1 : & x \leq 0 \end{cases}$

Let $x_0 = 0,\ x_n \to 0^+$, then

$$|f(x_n) - f(x_0)| = |1 - (-1)| = 2 \not\to 0$$

$\Rightarrow \quad f$ is not continuous.

Examples: Differentiability

i) $f : \mathbb{R} \to \mathbb{R},\ x \mapsto ax,\ a \in \mathbb{R}$

Let us consider the surrogate $Df(x_0)(h) := ah$ and a sequence $h_n \to 0$. Then

$$\frac{1}{|h_n|}|f(x_0) + Df(x_0)h_n - f(x_0 + h_n)| = \frac{1}{|h_n|}|ax_0 + ah_n - a(x_0 + h_n)| = 0 \xrightarrow{n\to\infty} 0$$

$\Rightarrow$ f is differentiable.

ii) $f : \mathbb{R}^n \to \mathbb{R}^m,\ x \mapsto Ax,\ A \in \mathbb{R}^{m\times n}$

Let us consider the surrogate $Df(x_0)(h) := Ah$ and a sequence $h_n \to 0$. Then

$$\frac{1}{|h_n|}|f(x_0) + Df(x_0)h_n - f(x_0 + h_n)| = \frac{1}{|h_n|}|Ax_0 + Ah_n - A(x_0 + h_n)| = 0 \xrightarrow{n\to\infty} 0$$

$\Rightarrow$ f is differentiable.

iii) $f : \mathbb{R} \to \mathbb{R},\ x \mapsto |x|$

f is **not** differentiable at $x_0 = 0$.

Remark: How can we identify continuous/differentiable functions?

- Many elementary functions (polynomials, trigonometric functions, exponential function,...) and operations ("+","·",...) to combine such elementary functions are continuous/differentiable.

- The concatenation of such functions is also continuous/differentiable!

- Examples:

  – monomial x^k and polynomial (=linear combination) $p(x) = \sum_{j=0}^{m} a_j x^j$

  – expotential function e^x and sine function $\sin(x) = \frac{1}{2i}(e^{ix} - e^{-ix})$

We will show in the exercise that differentiability is a stronger requirement than continuity:

Theorem 8.2 *Every differentiable function is also continuous.*

Next, we introduce the directional derivative which often serves as a good starting point to find the (Fréchet) derivative of a function (especially in complex and confusing situations):

Definition 8.3 (*Directional derivative*) *We assume that $f : \mathbb{R}^n \to \mathbb{R}^m$ is (Fréchet-) differentiable at $x_0 \in \mathbb{R}^n$ with derivative $Df(x_0)$. For a $v \in \mathbb{R}^n$, the limit*

$$Gf(x_0)(v) := \lim_{t \to 0^+} \frac{f(x_0 + tv) - f(x_0)}{t}$$

exists and it concides with the Fréchet derivative, i.e., $Gf(x_0)(v) = Df(x_0)(v)$.
*We call $Gf(x_0)(v)$ the **directional derivative** at x_0 in the direction v. (Gâteaux derivative)*

Remark:
The Gâteaux derivative may exist, even if f is not Fréchet differentiable (e.g. $x \mapsto |x|,\ x_0 = 0$).

Examples:

i) $f : \mathbb{R} \to \mathbb{R},\ x \mapsto |x|,\ x_0 = 0$ (not Fréchet-differentiable)

a) $v \geq 0 :\ Gf(x_0)(v) = \lim_{t\to 0^+} \frac{1}{t}(f(x_0+tv) - f(x_0)) = \lim_{t\to 0^+} \frac{1}{t}(tv) = 1 \cdot v$
b) $v < 0 :\ Gf(x_0)(v) = \lim_{t\to 0^+} \frac{1}{t}(\underbrace{f(x_0+tv)}_{=-tv} - f(x_0)) = (-1) \cdot v$

ii) $f : \mathbb{R}^n \to \mathbb{R},\ x \mapsto \|x\|_2^2 = x^T x,\ v \in \mathbb{R}^n,\ x_0 \in \mathbb{R}^n$

$$
\begin{aligned}
Gf(x_0)(v) &= \lim_{t\to 0^+} \frac{f(x_0+tv) - f(x_0)}{t} \\
&= \lim_{t\to 0^+} \big(\underline{(x_0+tv)^T(x_0+tv)}_{x_0^T x_0 + 2t x_0^T v + t^2 v^T v} - x_0^T x_0\big)\frac{1}{t} \\
&= (2x_0)^T v
\end{aligned}
$$

Consider $v = \sum_{j=1}^n v_j e_j$, where $e_1, \dots, e_n$ denote the standard basis in $\mathbb{R}^n$, then

$$
Df(x_0)(v) = \sum_{j=1}^n v_j \underbrace{\underbrace{Df(x_0)}_{\mathbb{R}^n \to \mathbb{R}^m}(e_j)}_{\in \mathbb{R}^m}
$$

Definition 8.4 (*Partial derivative*) *Let $f : D \to \mathbb{R}^m$, $D \subseteq \mathbb{R}^n$ be (Fréchet-)differentiable in $x_0 \in D$. We define the so-called **partial derivatives** of f at x_0 with respect to the j-th variable by:*

$$\frac{\partial}{\partial x_j} f(x_0) := Df(x_0)(e_j),$$

where e_j is the j-th standard basis vector.

Now again with $v = \sum_{j=1}^n v_j e_j$ we find

$$\begin{aligned}
Df(x_0)(v) &= \sum_{j=1}^n v_j Df(x_0)(e_j) \\
&= \begin{pmatrix} | & & | \\ Df(x_0)(e_1) & \cdots & Df(x_0)(e_n) \\ | & & | \end{pmatrix} \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} \\
&= \begin{pmatrix} | & & | \\ \frac{\partial}{\partial x_1} f(x_0) & \cdots & \frac{\partial}{\partial x_n} f(x_0) \\ | & & | \end{pmatrix} v \\
&= \underbrace{J_f(x_0)}_{\in \mathbb{R}^{m \times n}} \cdot v
\end{aligned}$$

$$f : \mathbb{R}^n \to \mathbb{R}^m,\ f(x) = \begin{pmatrix} f_1(x) \\ \vdots \\ f_m(x) \end{pmatrix},\ f_i : \mathbb{R}^n \to \mathbb{R}$$

Since $Df(x_0) : \mathbb{R}^n \to \mathbb{R}^m$ is linear it can be represented by a matrix:

Lemma 8.5 (*Jacobian*) *Let $f : \mathbb{R}^n \supset D \to \mathbb{R}^m$ be differentiable at $x_0 \in D$ with derivative $Df(x_0) : \mathbb{R}^n \to \mathbb{R}^m$. Then the so-called* ***Jacobian matrix***

$$J_f(x_0) := \mathcal{M}_I^I(Df(x_0)) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(x_0) & \cdots & \frac{\partial f_1}{\partial x_n}(x_0) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(x_0) & \cdots & \frac{\partial f_m}{\partial x_n}(x_0) \end{pmatrix} \in \mathbb{R}^{m \times n}$$

is the matrix representation of $Df(x_0)$ with respect to the standard bases in $\mathbb{R}^n$ and $\mathbb{R}^m$.

In the special case, that the Jacobian matrix is just one row we give it a special name:

Definition 8.6 (*Gradient*) *Let $f : \mathbb{R}^n \supset D \to \mathbb{R}$ be differentiable at $x_0 \in D$, then*

$$J_f(x_0)^T = \begin{pmatrix} \frac{\partial f}{\partial x_1}(x_0) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x_0) \end{pmatrix} =: \nabla f(x_0)$$

is called the ***gradient of** f at $x_0 \in D$.*

Example

8.3 Solving Nonlinear Equations: Taylor Approximation and Newton's Method

The next result is on the approximation quality of the derivative:

Lemma 8.7 (*Taylor approximation*) *Let $f : \mathbb{R}^n \supset B_\varepsilon(\hat{x}) \to \mathbb{R}^n$ be differentiable at $\hat{x}$ with some $\varepsilon > 0$. Assume further that there is a (Lipschitz) constant $L \geq 0$ such that the Jacobian J_f satisfies*

$$\|J_f(y) - J_f(x)\| \leq L\|y - x\|, \quad \forall x, y \in B_\varepsilon(\hat{x}). \tag{18}$$

Then, there holds

$$\|f(y) - [f(x) + J_f(x)(y - x)]\| \leq \frac{L}{2}\|y - x\|^2, \quad \forall x, y \in B_\varepsilon(\hat{x})$$

which we rephrase with the notation:

$$f(y) = f(x) + J_f(x)(y - x) + \mathcal{O}(\|y - x\|^2).$$

Let us apply Taylor approximation to solve nonlinear systems: The idea is to locally approximate the nonlinear function by its linear derivative and then solve many linear systems.

- Situation: Consider for a potentially nonlinear function $f : \mathbb{R}^n \to \mathbb{R}^n$ and the nonlinear system $f(\hat{x}) = 0$
- Aim: Determine the solution $\hat{x}$ (iteratively/numerically)
- Idea: Define an iterative scheme $x^{k+1} := x^k + \Delta x^k$ where the increment is derived as follows:

$$0 \stackrel{!}{=} f(x^{k+1}) = f(x^k + \Delta x^k) \approx f(x^k) + J_f(x^k)\Delta x^k \quad (\rightsquigarrow \text{ solve for } \Delta x^k)$$

$$\Leftrightarrow \quad J_f(x^k) \cdot \Delta x^k = -f(x^k) \qquad \text{(linear equation)}$$

$$\Leftrightarrow \quad \Delta x^k = -J_f(x^k)^{-1} f(x^k) \qquad \text{(invertibility of the derivative at each } x_k \text{ assumed!)}$$

$$x^k \to \hat{x}$$

One can show the following convergence result of this approach:

Theorem 8.8 (*simplified Newton-Kantorovich*) *Let $f : \mathbb{R}^n \supset B_\varepsilon(\hat{x}) \to \mathbb{R}^n$ be differentiable with invertible derivative for some $\varepsilon > 0$ and $f(\hat{x}) = 0$. Assume the Lipschitz condition* (18) *and the existence of an upper bound $\|J_f(x)^{-1}\| < M$ for some $M < \infty$ and for all $x \in B_\varepsilon(\hat{x})$. Then, the Newton iteration*

$$\boxed{x^{k+1} := x^k + \Delta x^k\,, \quad \textit{where} \;\; \Delta x^k \;\; \textit{solves} \;\; f(x^k) + J_f(x^k)\Delta x^k = 0}$$

converges quadratically to $\hat{x}$, provided x^1 is chosen sufficiently close to $\hat{x}$, i.e.

$$\|x^{k+1} - \hat{x}\| \leq c\|x^k - \hat{x}\|^2\,, \quad c < \infty.$$

Remark

In many cases, Newton's method does not work right out of the box, because the starting vector x^1 is too far away from the solution. Then, techniques for adaptive step-length reduction (damping, relaxation, line-search) have to be used in order to enforce convergence. Details of these approaches fill multiple books. When Newton's method works, i.e., after an initial damped phase, it gets super fast.

Take-away messages:

- Derivatives $\rightarrow$ local linear approximation to the function
- Newton's method $\rightarrow$ solves nonlinear systems by solving many linear problems in each step

8.4 The Chain Rule and Back Propagation

The chain rule lies at the heart of back propagation. It tells us how to compute the derivative of concatenated functions:

Theorem 8.9 (*Chain rule*) *Consider mappings $g : \mathbb{R}^\ell \supset D_g \to D_f \subset \mathbb{R}^m$ differentiable in $x_0 \in D_g$ with Jacobian $J_g(x_0)$ and $f : \mathbb{R}^m \supset D_f \to \mathbb{R}^n$, differentiable in $g(x_0) \in D_f$ with Jacobian $J_f(g(x_0))$. Then, the concatenation is differentiable with Jacobian $J_{f\circ g}(x_0)$ and*

$$D(f \circ g)(x_0) = Df(g(x_0)) \circ Dg(x_0) \quad \text{and} \quad \boxed{J_{f\circ g}(x_0) = J_f(g(x_0)) \cdot J_g(x_0)}.$$

Example 8.10 Let us revisit our regularizer from the imaging example:
Consider $D \in \mathbb{R}^{p\times n}$ and the linear function $g \colon \mathbb{R}^n \to \mathbb{R}^p$, $g(x) := Dx$. Then for all $x \in \mathbb{R}^n$ we easily find

$$J_g(x) = D.$$

Also, let $f \colon \mathbb{R}^p \to \mathbb{R}$, $f(y) := \frac{1}{2}y^\top y = \frac{1}{2}\|y\|_2^2$, then we have seen above that, for all $y \in \mathbb{R}^p$,

$$J_f(y)^\top = \nabla f(y) = \tfrac{1}{2}2y = y.$$

Then the concatenation $h := (f \circ g) : \mathbb{R}^n \to \mathbb{R}$ is given by

$$h(x) = \tfrac{1}{2}\|Dx\|_2^2$$

with gradient, at $x \in \mathbb{R}^n$, obtained from the chain rule

$$\nabla h(x) = J_h(x)^\top = \left(J_f(g(x)) \cdot J_g(x)\right)^\top = D^\top \nabla f(g(x)) = D^\top g(x) = D^\top Dx.$$