Gradients manipulations

Optimization for Machine Learning — Exercise 01

Monday 17th April, 2023

Recall that the gradient of a differentiable function $f\colon \mathbb{R}^m \to \mathbb{R}$ at $x \in \mathbb{R}^m$ is a vector in $\mathbb{R}^m$, usually denoted by $\nabla f(x)$, such that

$$\nabla f(x) = \begin{pmatrix} \partial_{x_1} f(x) \\ \vdots \\ \partial_{x_n} f(x) \end{pmatrix},$$

where $\partial_{x_i} f(x) := \frac{\partial f(x)}{\partial x_i}$ is the partial derivative of f at x with respect to x_i for $i \in [n] := \{1, \ldots, n\}$.

When f is multivalued, i.e. $f\colon \mathbb{R}^m \to \mathbb{R}^n$, its *Jacobian* at x, denoted by $J_f(x)$, is the $n \times m$ matrix such that, if $y = f(x) \in \mathbb{R}^n$,

$$J_f(x) = \left(\frac{\partial y_i}{\partial x_j}\right)_{(i,j)\in[n]\times[m]} = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_m} \\ \frac{\partial y_2}{\partial x_1} & \ddots & & \vdots \\ \vdots & & & \\ & & & \frac{\partial y_n}{\partial x_m} \end{pmatrix} = \begin{pmatrix} \partial_x y_1 \\ \partial_x y_2 \\ \vdots \\ \partial_x y_n \end{pmatrix} = \partial_x f(x),$$

where $\partial_x y$ is understood as having as **column indices** the indices from x, and **row indices** the indices of y. This notation might be confusing, but it is sometimes useful, see e.g. Exercise 2c.

When $n = 1$, note that we have $\nabla f(x) = (J_f(x))^\top = (\partial_x f(x))^\top$.

Product rule Similar to the 1D case, a product rule exists for multivariate functions.

1. a) If $f\colon \mathbb{R}^d \to \mathbb{R}^n$ and $g\colon \mathbb{R}^d \to \mathbb{R}^n$, show that $\nabla(f^\top g)(x) = J_f(x)^\top g(x) + J_g(x)^\top f(x)$.

   b) What happens with $n = 1$?

Chain rule The chain rule for Jacobian is, for $f\colon \mathbb{R}^m \to \mathbb{R}^k$, and $g\colon \mathbb{R}^k \to \mathbb{R}^n$, $J_{g\circ f}(x) = J_g(f(x))J_f(x)$ (when the composition makes sense, and everything is differentiable). Compare with the chain rule in the 1D case: $(g \circ f)'(x) = g'(f(x)) \cdot f'(x)$.

2. a) Compute the gradient of $g \circ f$ at $x \in \mathbb{R}^m$, when $f\colon \mathbb{R}^m \to \mathbb{R}^k$ and $g \circ \mathbb{R}^k \to \mathbb{R}$, as a function of $J_f(x)$ and $\nabla g(f(x))$ (note the difference between $\nabla(g \circ f)(x)$ and $\nabla g(f(x))$).

   b) Compute the gradient of $h_2 \circ h_1 \circ f$, when $h_1\colon \mathbb{R} \to \mathbb{R}$ and $h_2\colon \mathbb{R} \to \mathbb{R}$ are single valued functions, and $f\colon \mathbb{R}^d \to \mathbb{R}$ is a vector function.

   c) Assume that $x(w), y(w), z(w)$ are function of $w \in \mathbb{R}^p$, and that $\mathcal{L}\colon \mathbb{R}^3 \to \mathbb{R}$ is a function of x, y, z. Show that $\nabla_w \mathcal{L}(x(w), y(w), z(w)) = \frac{\partial \mathcal{L}(x,y,z)}{\partial x}\nabla x(w) + \frac{\partial \mathcal{L}(x,y,z)}{\partial y}\nabla y(w) + \frac{\partial \mathcal{L}(x,y,z)}{\partial z}\nabla z(w)$.

Classical vector functions Often, the functions that will appear are build from simpler ones, such as the linear product $\langle a, b\rangle = a^\top b$, or a matrix-vector multiplication $A \cdot b$, etc. The easiest to find the gradient of such function is usually to go back to the expression with the indices, e.g. $\langle a, b\rangle = \sum_i a_i b_i$. Another method, especially for more complex cases, is to look up if the formula is in the Matrix Cookbook.

3. a) Let $a \in \mathbb{R}^n$. Show that $\nabla_x \langle a, x\rangle = a$,

   b) Let $A \in \mathbb{R}^{n\times m}$, and $f\colon \mathbb{R}^m \to \mathbb{R}^n$, $x \mapsto Ax$. Show that $J_f(x) = A$.

   c) What is $\nabla_x \langle x, Ax\rangle$? ($A \in \mathbb{R}^{n\times n}$)

   d) What is $\nabla_A \langle x, Ax\rangle$? ($A \in \mathbb{R}^{n\times n}$)

General functions

4. Compute the gradients of the functions:

   a) $f\colon \mathbb{R}^3 \to \mathbb{R}$, $(x, y, z) \mapsto \frac{1}{2}x^2 + yz - \ln(1 + \exp(x^2y^3z))$

   b) $g\colon \mathbb{R}^d \to \mathbb{R}$, $x \mapsto \frac{1}{2}\|x\|^2 = \frac{1}{2}x^\top x = \frac{1}{2}\sum_{i=1}^d x_i^2$

   c) $h\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, $(x, w) \mapsto \ln\left(1 + \exp(-w^\top x)\right)$. Compute the gradient with respect to w: $\nabla_w h(x, w)$.