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. $f\colon \mathbb{R}^m \to \mathbb{R}^n$, then 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, be 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 $n = 1$?

Answer

1. a) With $f\colon \mathbb{R}^d \to \mathbb{R}^n$, $g\colon \mathbb{R}^d \to \mathbb{R}^n$, and for $x \in \mathbb{R}^d$ where f and g are differentiable, $f(x)^\top g(x) = \sum_{k=1}^n f_k(x) g_k(x)$. Therefore, for an index $i \in [d]$,

$$\begin{aligned}
\partial_{x_i}(f^\top g)(x) = \partial_{x_i}(f(x)^\top g(x)) &= \partial_{x_i} \sum_{l=1}^n f_k(x) g_k(x) \\
&= \sum_{l=1}^n \partial_{x_i} f_k(x) g_k(x) + f_k(x) \partial_{x_i} g_k(x) \\
&= \sum_{l=1}^n J_f(x)_{ki} g_k(x) + J_g(x)_{ki} f_k(x) \\
&= \sum_{l=1}^n J_f(x)_{ik}^\top g_k(x) + J_g(x)_{ik}^\top f_k(x) \\
&= J_f^\top(x)_{i,:} g(x) + J_g^\top(x)_{i,:} f(x),
\end{aligned}$$

where $A_{i,:}$ stand for the row i of a matrix A. Since this is true for all $i \in [d]$, and i is a row index on both sides, we can write

$$\implies \qquad \nabla_x(f^\top g)(x) = J_f^\top(x) g(x) + J_g^\top(x) f(x).$$

Always check the dimensions!

b) In the case $n = 1$, then $J_f^\top = \nabla f$ and $J_g^\top = \nabla g$, therefore

$$\nabla(fg)(x) = \nabla f(x) g(x) + f(x) \nabla g(x)$$

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 $\nabla_w \mathcal{L}(x(w), y(w), z(w)) = \frac{\partial \mathcal{L}(x,y,z)}{\partial x} \nabla x(w) + \frac{\partial \mathcal{L}(x,y,z)}{\partial y} \nabla y(w) + \frac{\partial \mathcal{L}(x,y,z)}{\partial z} \nabla z(w)$.

Answer

2. a) Since $g \circ f\colon \mathbb{R}^m \to \mathbb{R}$ is real-valued, we have

$$\nabla(g \circ f)(x) = (J_{g\circ f}(x))^\top = J_f(x)^\top J_g(f(x)) = J_f(x)^\top \nabla g(f(x)),$$

since $g\colon \mathbb{R}^k \to \mathbb{R}$ is real valued.

b) Applying the previous rule with $g = h_2 \circ h_1$, we get

$$\begin{aligned}\nabla(h_2 \circ h_1 \circ f)(x) &= J_f(x)^\top \nabla(h_2 \circ h_1)(f(x)) \\ &= \nabla f(x) \cdot (h_2 \circ h_1)'(f(x)) \\ &= \nabla f(x) \cdot h_2'(h_1(f(x))) \cdot h_1'(f(x)),\end{aligned}$$

since h_1, h_2 are 1D functions and f is real valued.

c) We write the parametrization of x, y, z as $\phi\colon \mathbb{R}^p \to \mathbb{R}^3,\ w \mapsto \phi(w) = \begin{pmatrix} x(w) \\ y(w) \\ z(w) \end{pmatrix}$.

The function on w is then $(\mathcal{L} \circ \phi)(w)$, of which we can take the gradient

$$\begin{aligned}\nabla_w(\mathcal{L} \circ \phi)(w) &= J_\phi^\top(w)\nabla\mathcal{L}(\phi(w)) \\ &= \begin{pmatrix} \partial_w x(w) \\ \partial_w y(w) \\ \partial_w z(w) \end{pmatrix}^\top \begin{pmatrix} \partial_x \mathcal{L}(\phi(w)) \\ \partial_y \mathcal{L}(\phi(w)) \\ \partial_z \mathcal{L}(\phi(w)) \end{pmatrix} \qquad \text{(Recall the notation } \partial_w x(w) = (\nabla_w x(w))^\top\text{)} \\ &= \begin{pmatrix} \nabla_w x(w) & \nabla_w y(w) & \nabla_w z(w) \end{pmatrix} \begin{pmatrix} \partial_x \mathcal{L}(x,y,z) \\ \partial_y \mathcal{L}(x,y,z) \\ \partial_z \mathcal{L}(x,y,z) \end{pmatrix} \\ &= \partial_x \mathcal{L}(x,y,z)\nabla_w x(w) + \partial_y \mathcal{L}(x,y,z)\nabla_w y(w) + \partial_z \mathcal{L}(x,y,z)\nabla_w z(w),\end{aligned}$$

where the dependency on w for x, y, z has been dropped where it was not relevant.

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

Answer

3. a) With $a \in \mathbb{R}^n$, $x \in \mathbb{R}^n$, we have $\langle a, x\rangle = \sum_{j=1}^n a_j x_j$, and therefore, for $i \in [n]$,

$$\partial_{x_i}\langle a, x\rangle = \partial_{x_i} \sum_{j=1}^n a_j x_j = \sum_{j=1}^n a_i \partial_{x_i} x_j = \sum_{j=1}^n a_j \delta_{ij},$$

where $\delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise} \end{cases}$. Therefore, the only non-zero term in the sum is when $j = i$, and

$$\partial_{x_i}\langle a, x\rangle = a_i \implies \nabla_x\langle a, x\rangle = a,$$

since i is a row index on both sides of the equation.

b) With $A = (a_{ij})_{i=1,j=1}^{n,m} \in \mathbb{R}^{n\times m}$, $x \in \mathbb{R}^m$, and $i \in [n]$, we have that $(Ax)_i = \sum_{k=1}^m a_{ik}x_k$, and therefore, for $(i, j) \in [n] \times [m]$, one gets

$$\begin{aligned} \partial_{x_j}(f(x))_i = \partial_{x_j}(Ax)_i &= \partial_{x_j}\sum_{k=1}^m a_{ik}x_k \\ &= \sum_{k=1}^m a_{ik}\partial_{x_j}x_k \\ &= \sum_{k=1}^m a_{ik}\delta_{jk} \\ &= a_{ij}. \end{aligned}$$

Since i (resp. j) is a row (resp. column) index on both sides of the equation, we can conclude that

$$J_f(x) = A$$

c) We can use the product formula (Exercise 1a) with $f(x) = x$ and $g(x) = Ax$: $\nabla_x\langle x, Ax\rangle = J_f(x)^\top Ax + J_g(x)^\top x = I_nAx + A^\top x = (A + A^\top)x$.

d) Here, similar to the gradient of a real-valued vector function, the definition of the gradient of a real-valued **matrix** function $f\colon \mathbb{R}^{n\times m} \to \mathbb{R}$ with respect to its input is implied to be a matrix (of same size as the input), such that

$$\nabla_A f(A) = (\partial_{a_{ij}} f(A))_{i=1,j=1}^{n,m} = \begin{pmatrix} \partial_{a_{11}} f(A) & \cdots & \\ \vdots & \ddots & \\ & & \partial_{a_{nm}} f(A) \end{pmatrix}.$$

It comes up in some situations in machine learning, typically when the weights with respect to which we are computing the gradient are organized in matrices.

Here, $A \in \mathbb{R}^{n\times n}$, so we compute, for $(i, j) \in [n] \times [n]$,

$$\begin{aligned} \partial_{a_{ij}} f(A) = \partial_{a_{ij}}\langle x, Ax\rangle &= \partial_{a_{ij}}\sum_{k,l} x_kx_la_{kl} \\ &= \sum_{k,l} x_kx_l\partial_{a_{ij}}a_{kl} \\ &= \sum_{k,l} x_kx_l\delta_{ik}\delta_{jl} \end{aligned}$$

again with $\delta_{ij} = \mathbb{1}(i = j)$, since $\partial_{a_{ij}} a_{kl} = \begin{cases} 1 & \text{if } k = i \text{ and } l = j \\ 0 & \text{otherwise} \end{cases}$. Therefore

$$\partial_{a_{ij}} f(A) = x_i x_j.$$

The matrix that have entries $x_i x_j$ for i a row index and j a column index is the matrix $xx^\top$. Therefore,

$$\nabla_A \langle x, Ax \rangle = xx^\top.$$

General functions

4. Compute the gradients of the functions:
   a) $f \colon \mathbb{R}^3 \to \mathbb{R}$, $(x, y, z) \mapsto \frac{1}{2}x^2 + yz - \ln(1 + \exp(x^2 y^3 z))$
   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) $h \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, $(x, w) \mapsto \ln\big(1 + \exp(-w^\top x)\big)$. Compute the gradient with respect to w: $\nabla_w h(x, w)$.

Answer

4. a) One computes

$$\nabla f(x, y, z) = \begin{pmatrix} x - \frac{2xy^3 z}{1+\exp(x^2 y^3 z)} \\ z - \frac{3y^2 x^2 z}{1+\exp(x^2 y^3 z)} \\ y - \frac{x^2 y^3}{1+\exp(x^2 y^3 z)} \end{pmatrix}$$

   b) i. One way to compute is to use the product rule 1a with $f = g = \mathrm{id}$:

$$\nabla_x(\frac{1}{2}x^\top x) = \frac{1}{2}(I_d x + I_d x) = x$$

   ii. Another way is to go back to the definition $\frac{1}{2}\|x\|^2 = \sum_i x_i^2$, and evaluate at each coordinate $j \in [d]$: $\partial_{x_j} \frac{1}{2}\|x\|^2 = \frac{1}{2}\sum_i \partial_{x_j} x_i^2 = x_j$.

   c) For this h, the formula found in Exercise 2b is adapted:

$$\nabla_w h(x, w) = \frac{\nabla_w \exp(-x^\top w)}{1 + \exp(-x^\top w)} = -\frac{\exp(-x^\top w)}{1 + \exp(-x^\top w)} x.$$