Calculus for ML

1 Gradients, Jacobians, and Hessians

Recall from lecture the following:

- The *gradient* of a scalar-valued function $f : \mathbb{R}^n \to \mathbb{R}$ is a column vector of length n, denoted as ∇f, containing the derivatives of components of f with respect to the input variables:

$$((\nabla f(x))_i = \frac{\partial f}{\partial x_i}(x), \quad i = 1, \ldots, n. \tag{1}$$

- The *Hessian* of a scalar-valued function $f : \mathbb{R}^n \to \mathbb{R}$ is an $n \times n$ matrix, denoted as $\nabla^2 f$, containing the second derivatives of components of f with respect to the input variables:

$$(\nabla^2 f(x))_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}, \quad i = 1, \ldots, n, \quad j = 1 \ldots, n. \tag{2}$$

- The *Jacobian* (or *Derivative*) of a vector-valued function $f : \mathbb{R}^n \to \mathbb{R}^m$ is an $m \times n$ matrix, denoted as $\frac{\partial f}{\partial x}$, containing the derivatives of components of f with respect to the input variables:

$$\left(\frac{\partial f}{\partial x}\right)_{ij} = \frac{\partial f_i}{\partial x_j}, \quad i = 1, \ldots, m, \quad j = 1 \ldots, n. \tag{3}$$

For the remainder of the class, we will repeatedly have to take gradients, Hessians, and derivatives of functions.

Furthermore, we outline two strategies one can use to answer the following parts. To illustrate this, we'll use the following example: $f(x) = w^\top x$

1. *Using computation via first principle.* In class we learned Taylor's Theorem which states: For a scalar function $f : \mathbb{R}^n \to \mathbb{R}$ around a point x_0, the first-order (linear) approximation is

$$\widehat{f}_1(x; x_0) = f(x_0) + \frac{\partial f}{\partial x}(x_0)(x - x_0). \tag{4}$$

   Let $x_0 = x + \Delta$, where Δ is tiny. Then, the first-order (linear) approximation of f around $x + \Delta$ is

$$f(x) = f(x + \Delta) + \frac{\partial f}{\partial x}(x + \Delta)(-\Delta). \tag{5}$$

   Rearranging gives us:

$$f(x + \Delta) = f(x) + \frac{\partial f}{\partial x}(x + \Delta)(\Delta). \tag{6}$$

   Via pattern matching, we see that

$$f(x + \Delta) = f(x) + (\textit{something})\Delta, \tag{7}$$

   where *something* is our derivative!

   Now, we use $f(x) = w^\top x$. Then we have

$$f(x + \Delta) = w^\top (x + \Delta) = w^\top x + w^\top \Delta = f(x) + w^\top \Delta. \tag{8}$$

   Comparing with equation (7), we conclude that

$$\frac{\partial f}{\partial x} = w^\top \quad \text{and, thus,} \quad \nabla f(x) = w. \tag{9}$$

2. *Using the formula.* The idea is to build up the gradient/derivative one component at a time. For $f(x) = w^\top x$, we have that $\sum_j w_j x_j$. Hence, we have

$$\frac{\partial f}{\partial x_i} = w_i. \tag{10}$$

Thus, we find that

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix} = w. \tag{11}$$

And $\frac{\partial f}{\partial x} = w^\top$.

For the following parts, before taking any gradients/Hessians/derivatives, identify what the gradient/Hessian/derivative looks like (is it a scalar, vector, or matrix?) and how we calculate each term in the gradient/Hessian/derivative. Then carefully solve for an arbitrary entry of the gradient/Hessian/derivative, then stack/arrange all of them to get a final result.

For the first two parts, suppose $X \in \mathbb{R}^{n \times n}$ is a square matrix whose entries are denoted X_{ij} and whose rows are denoted $X_1^\top, \ldots, X_n^\top$ and $b \in \mathbb{R}^n$ is a vector whose entries are denoted b_i.

(a) Compute the Jacobians for the following functions.

i. $f(w) = Xw$.

Solution

We compute each partial derivative and construct $\frac{\partial f}{\partial w}$ at the end. That is,

$$\begin{aligned} \left(\frac{\partial f}{\partial w}(w)\right)_{jk} &= \frac{\partial f_j}{\partial w_k}(w) && (12)\\ &= \frac{\partial (Xw)_j}{\partial w_k} && (13)\\ &= \frac{\partial (X_j^\top w)}{\partial w} && (14)\\ &= (X_j)_k && (15)\\ &= X_{jk}. && (16) \end{aligned}$$

Thus, $\frac{\partial f}{\partial w}(w) = X$.

ii. $g(w) = f(w)w$ where $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable.

Solution

We again compute each partial derivative using the scalar product rule, obtaining

$$\begin{aligned} \left(\frac{\partial g}{\partial w}(w)\right)_{jk} &= \frac{\partial g_j}{\partial w_k}(w) && (17)\\ &= \frac{\partial (f(w)w_j)}{\partial w_k}(w) && (18)\\ &= f(w)\frac{\partial w_j}{\partial w_k} + w_j \frac{\partial f}{\partial w_k}(w) && (19)\\ &= w_j [\nabla f(w)]_k + \begin{cases} f(w) & \text{if } j = k, \\ 0 & \text{otherwise.} \end{cases} && (20) \end{aligned}$$

This gives a Jacobian whose entries are

$$\left(\frac{\partial g}{\partial w}(w)\right)_{jk} = w_j[\nabla f(w)]_k, \quad \forall j \neq k \tag{21}$$

$$\left(\frac{\partial g}{\partial w}(w)\right)_{jj} = w_j[\nabla f(w)]_k + f(x), \quad \forall j. \tag{22}$$

This was not necessary but one could also write the Jacobian as

$$\frac{\partial g}{\partial w}(w) = w[\nabla f(w)]^\top + f(w)I. \tag{23}$$

(b) Compute the gradient and Hessian for each of the following functions.

i. $h(w) = w^\top X w$

Solution

We discuss two approaches to solve this problem.
Using computation via first principle. Taking $h(w) = w^\top X w$ and expanding, we have

$$f(x+\Delta) = (w+\Delta)^\top X(w+\Delta) \tag{24}$$

$$= w^\top X w + \Delta^\top X w + w^\top X \Delta + \Delta^\top X \Delta \tag{25}$$

$$= h(w) + (w^\top X^\top + w^\top X)\Delta + \Delta^\top X \Delta \tag{26}$$

which yields (via pattern matching)

$$\frac{\partial h}{\partial w} = w^\top(X^\top + X) \quad \text{and, thus,} \quad \nabla h(w) = (X + X^\top)w. \tag{27}$$

Now, we expand $\nabla h(w) = (X + X^\top)w$ and find that

$$\nabla h(w+\Delta) = (X + X^\top)w + (X + X^\top)\Delta = \nabla h(w) + (X + X^\top)\Delta. \tag{28}$$

Relating with equation (7), we obtan that $\nabla^2 h(w) = X + X^\top$.
Using the formula. We have

$$h(w) = w^\top X w \tag{29}$$

$$= w^\top \begin{bmatrix} X_1^\top \\ \vdots \\ X_n^\top \end{bmatrix} w \tag{30}$$

$$= w^\top \begin{bmatrix} X_1^\top w \\ \vdots \\ X_n^\top w \end{bmatrix} \tag{31}$$

$$= \sum_{i=1}^{n} w_i (X_i^\top w). \tag{32}$$

Taking the derivative with respect to any w_j yields

$$\frac{\partial h}{\partial w_j}(x) = \frac{\partial}{\partial w_j}\sum_{i=1}^{n} w_i(X_i^\top w) \tag{33}$$

$$= \frac{\partial}{\partial w_j}\left(w_j(X_j^\top w) + \sum_{i\neq j} w_i(X_i^\top w)\right) \tag{34}$$

$$= w_j\frac{\partial}{\partial w_j}(X_j^\top w) + X_j^\top w + \sum_{i\neq j} w_i\frac{\partial}{\partial w_j}(X_i^\top w) \tag{35}$$

$$= w_j(X_j)_j + X_j^\top w + \sum_{i\neq j} w_i(X_i)_j \tag{36}$$

$$= X_j^\top w + \sum_{i=1}^{n} w_i(X_i)_j \tag{37}$$

$$= \sum_{i=1}^{n} X_{ji}w_i + \sum_{i=1}^{n} X_{ij}w_i \tag{38}$$

$$= (Xw)_j + (X^\top w)_j \tag{39}$$

$$= [(X + X^\top)w]_j. \tag{40}$$

Thus, we have

$$\nabla h(w) = (X + X^\top)w. \tag{41}$$

For the Hessian, notice that

$$[\nabla^2 h(w)]_{jk} = \frac{\partial^2 h}{\partial w_j \partial w_k}(w) \tag{42}$$

$$= \frac{\partial [\nabla h]_j}{\partial w_k}(w) \tag{43}$$

$$= \frac{\partial}{\partial w_k}\left[\sum_{i=1}^{n} X_{ji}w_i + \sum_{i=1}^{n} X_{ij}w_i\right] \tag{44}$$

$$= \left[\sum_{i=1}^{n} \frac{\partial}{\partial w_k} X_{ji}w_i + \sum_{i=1}^{n} \frac{\partial}{\partial w_k} X_{ij}w_i\right] \tag{45}$$

$$= X_{jk} + X_{kj} \tag{46}$$

$$= (X + X^\top)_{jk}. \tag{47}$$

Thus we have

$$\nabla^2 h(w) = X + X^\top. \tag{48}$$

Notice the important special case that if X is symmetric then $\nabla h(w) = 2Xw$ and $\nabla^2 h(w) = 2X$.

ii. $l(w) = \|w\|_2^2$.

Solution

If we take $X = I$ in h from part i. we get

$$\nabla l(w) = 2w \quad \text{and, thus,} \quad \nabla^2 l(w) = 2I. \tag{49}$$

(c) Given $x \in \mathbb{R}^n, z \in \mathbb{R}^m$, $z = q(x)$, and $p(z) = z^\top z$, find the gradient $\nabla_z p(z)$, then find the gradient $\nabla_x p(z)$ in terms of $\frac{\partial z}{\partial x}$ and z.

Solution

Note that $z^\top z$ is a scalar and can be expressed element-wise as:

$$z^\top z = \sum_{i=1}^{n} z_j^2. \tag{50}$$

Therefore, the partial derivatives of $p(z)$ can be expressed as:

$$\frac{\partial p}{\partial z_i} = 2z_i. \tag{51}$$

The gradient is therefore:

$$\nabla_z p(z) = 2z. \tag{52}$$

Now, applying the Chain Rule for Multivariate Functions, we get:

$$\frac{\partial p}{\partial x_i} = \sum_{j=1}^{m} \frac{\partial p}{\partial z_j} \frac{\partial z_j}{\partial x_i} \tag{53}$$

$$= \sum_{j=1}^{m} (2z_j) \frac{\partial z_j}{\partial x_i} \tag{54}$$

Notice that this summation looks like a dot product! In fact, it is the dot product between the vector $2z^\top$ and the vector $\frac{\partial z}{\partial x_i}$. In matrix notation, this becomes the i-th component of the row vector $2z^\top \frac{\partial z}{\partial x}$, where $\frac{\partial z}{\partial x}$ is the Jacobian of $z = g(x)$. Therefore,

$$\frac{\partial p}{\partial x} = 2z^\top \left(\frac{\partial z}{\partial x}\right) \quad \text{and, thus,} \quad \nabla_x p(z) = 2\left(\frac{\partial z}{\partial x}\right)^\top z. \tag{55}$$

2 Taylor's Theorem

In class, we learned that we can approximate any function $f(x)$ near a point x_0 using polynomials. This concept is incredibly useful in Deep Learning when we interpret how the parameters (input variables) of our model (function f) change after one step of gradient descent. Don't worry if you don't understand what this means–you will soon!

For the following problem, plot/hand draw the level sets of the function. Also, point out the gradient directions in the level-set diagram. Additionally, compute the first and second order Taylor series approximation around the point $x_0 = (1, 1)$ for the function and comment on how accurately they approximate the true function.

(a) $g(x_1, x_2) = \frac{x_1^2}{4} + \frac{x_2^2}{9}$

Solution

We first compute the first and second order partial derivatives of g as follows:

$$\frac{\partial g}{\partial x_1}(x_1, x_2) = x_2, \quad \frac{\partial g}{\partial x_2}(x_1, x_2) = x_1, \tag{56}$$

$$\frac{\partial^2 g}{\partial x_1^2}(x_1, x_2) = 0, \quad \frac{\partial^2 g}{\partial x_2 x_1}(x_1, x_2) = 1, \tag{57}$$

$$\frac{\partial^2 g}{\partial x_2^2}(x_1, x_2) = 0, \quad \frac{\partial^2 g}{\partial x_1 x_2}(x_1, x_2) = 1. \tag{58}$$

The first order Taylor series approximation around $(1, 1)$ can be computed as:

$$g(x_1, x_2) \approx g(1, 1) + (\nabla g(1, 1))^\top \begin{bmatrix} x_1 - 1 \\ x_2 - 1 \end{bmatrix} \tag{59}$$

$$= 1 + x_1 - 1 + x_2 - 1 \tag{60}$$

$$= x_1 + x_2 - 1. \tag{61}$$

The second order Taylor series approximation around $(1, 1)$ can be computed as:

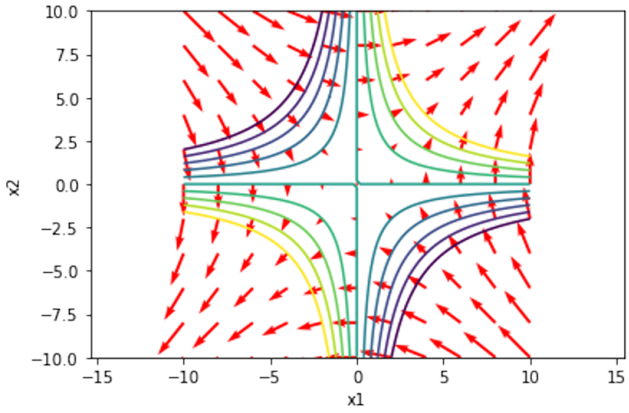
$$g(x_1, x_2) \approx g(1, 1) + (\nabla g(1, 1))^\top \begin{bmatrix} x_1 - 1 \\ x_2 - 1 \end{bmatrix} + \frac{1}{2} \begin{bmatrix} x_1 - 1 & x_2 - 1 \end{bmatrix} \nabla^2 g(1, 1) \begin{bmatrix} x_1 - 1 \\ x_2 - 1 \end{bmatrix} \tag{62}$$

$$= x_1 + x_2 - 1 + \frac{1}{2}(2(x_1 - 1)(x_2 - 1)) \tag{63}$$

$$= (x_1 - 1)(x_2 - 1) + x_1 + x_2 - 1 \tag{64}$$

$$= x_1 x_2 \tag{65}$$

The original function evaluated at $(1.1, 1.1)$ is 1.21. The first order approximation around $(1.1, 1.1)$ is 1.2, but the second order approximation again exactly represents the function! See below for the level sets and gradient directions for the function $g(x_1, x_2) = x_1 x_2$.



3 Least Squares

In our Linear Algebra for ML lecture, you learned about the following optimization problem:

$$w^* = \underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \|y - Xw\|_2^2. \tag{66}$$

where $X \in \mathbb{R}^{n \times d}$ is a full-rank data matrix, $y \in \mathbb{R}^n$ is the target vector of measurement values, and $w \in \mathbb{R}^d$ is a weight vector.

In this problem, we will learn how to solve this using Calculus.

(a) Let $\mathcal{L}(w) = \|y - Xw\|_2^2$ be our objective function. Expand $\mathcal{L}$.

Solution

$$\begin{aligned} \mathcal{L}(w) &= \|y - Xw\|_2^2 && (67)\\ &= (y - Xw)^\top (y - Xw) && (68)\\ &= y^\top y - y^\top Xw - w^\top X^\top y + w^\top X^\top Xw && (69)\\ &= w^\top X^\top Xw - 2w^\top X^\top y + y^\top y. && (70) \end{aligned}$$

(b) Find the gradient of $\mathcal{L}$ *with respect to* w, i.e., find $\nabla_w \mathcal{L}(w)$.

Solution

$$\begin{aligned} \nabla_w \mathcal{L}(w) &= \nabla_w \left(w^\top X^\top Xw - 2w^\top X^\top y + y^\top y \right) && (71)\\ &= \nabla_w (w^\top X^\top Xw) - \nabla_w 2w^\top X^\top y + \nabla_w y^\top y && (72)\\ &= 2X^\top Xw - 2X^\top y. && (73) \end{aligned}$$

(c) Use the Main Theorem to find the optimal w^*. It is important to note that finding w^* in an arbitrary optimization problem by this method may not guarantee a minimum. However, the Least Squares problem has some nice properties that allows us to use this theorem directly.

Solution

By invoking the Main Theorem, we take the gradient and set it to zero.

$$\nabla_w L(w^*) = 2X^\top Xw - 2X^\top y = 0, \tag{74}$$

and solve for w^*:

$$\begin{aligned} (X^\top X)w^* &= X^\top y && (75)\\ w^* &= (X^\top X)^{-1} X^\top y. && (76) \end{aligned}$$

Contributors

- Patrick Mendoza