Foundations of Machine Learning

Brett Bernstein

August 22, 2018

Lab 1: Gradients and Directional Derivatives

Multivariate Differentiation

Learning Objectives

1. Define the directional derivative, and use it to find a linear approximation to $f(\mathbf{x}+h\mathbf{u})$.
2. Define partial derivative and the gradient. Show how to compute an arbitrary directional derivative using the gradient.
3. For a differentiable function, give a linear approximation near a point $\mathbf{x}$ using the gradient.
4. Show that the gradient gives the direction of steepest ascent, and the negative gradient gives the direction of steepest descent.

Concept Check Questions

1. If $f'(x;u) < 0$ show that $f(x+hu) < f(x)$ for sufficiently small $h > 0$.

   Solution. The directional derivative is given by

   $$f'(x;u) = \lim_{h\to 0} \frac{f(x+hu) - f(x)}{h} < 0.$$

   By the definition of a limit, there must be a $\delta > 0$ such that

   $$\frac{f(x+hu) - f(x)}{h} < 0$$

   whenever $|h| < \delta$. If we restrict $0 < h < \delta$ then we have

   $$f(x+hu) - f(x) < 0 \implies f(x+hu) < f(x)$$

   as required.

2. Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable, and assume that $\nabla f(x) \neq 0$. Prove

$$\arg\max_{\|u\|_2=1} f'(x;u) = \frac{\nabla f(x)}{\|\nabla f(x)\|_2} \quad \text{and} \quad \arg\min_{\|u\|_2=1} f'(x;u) = -\frac{\nabla f(x)}{\|\nabla f(x)\|_2}.$$

Solution. By Cauchy-Schwarz we have, for $\|u\|_2 = 1$,

$$|f'(x;u)| = |\nabla f(x)^T u| \leq \|\nabla f(x)\|_2 \|u\|_2 = \|\nabla f(x)\|_2.$$

Note that

$$\nabla f(x)^T \frac{\nabla f(x)}{\|\nabla f(x)\|_2} = \|\nabla f(x)\|_2 \quad \text{and} \quad \nabla f(x)^T \frac{-\nabla f(x)}{\|\nabla f(x)\|_2} = -\|\nabla f(x)\|_2,$$

so these achieve the maximum and minimum bounds given by Cauchy-Schwarz.

One way to understand the Cauchy-Schwarz inequality is to recall that the dot-product between two vectors $v, w \in \mathbb{R}^d$ can be written as

$$v^T w = \|v\|_2 \|w\|_2 \cos(\theta),$$

where θ is the angle between v and w. This value is maximized at $\cos(0) = 1$ and minimized at $\cos(\pi) = -1$.

Computing Gradients

Learning Objectives

1. Find the gradient of a function by computing each partial derivative separately.
2. Use the chain rule to perform gradient computations.
3. Compute the gradient of a differentiable function by determining the form of a general directional derivative.

Concept Check Questions

1. Let $f : \mathbb{R}^2 \to \mathbb{R}$ be given by $f(x,y) = x^2 + 4xy + 3y^2$. Compute the gradient $\nabla f(x,y)$.

   Solution. Computing the partial derivatives gives

   $$\partial_1 f(x,y) = 2x + 4y \quad \text{and} \quad \partial_2 f(x,y) = 4x + 6y.$$

   Thus the gradient is given by

   $$\nabla f(x,y) = \begin{pmatrix} 2x + 4y \\ 4x + 6y \end{pmatrix}.$$

2. Compute the gradient of $f : \mathbb{R}^n \to \mathbb{R}$ where $f(x) = x^T Ax$ and $A \in \mathbb{R}^{n\times n}$ is any matrix.

   Solution. Here we show two methods. In either case we can obtain differentiability by noticing the partial derivatives are continuous.

   (a) Since
   $$f(x) = x^T Ax = \sum_{i,j=1}^{n} a_{ij}x_i x_j$$
   we have
   $$\partial_k f(x) = \sum_{j=1}^{n}(a_{kj} + a_{jk})x_j$$
   so
   $$\nabla f(x) = (A + A^T)x.$$

   (b) Note that
   $$\begin{aligned} f(x+tv) &= (x+tv)^T A(x+tv) \\ &= x^T Ax + tx^T Av + tv^T Ax + t^2 v^T Av \\ &= f(x) + t(x^T A + x^T A^T)v + t^2(v^T Av). \end{aligned}$$
   Thus
   $$f'(x;v) = \lim_{t\to 0}\frac{f(x+tv) - f(x)}{t} = \lim_{t\to 0}(x^T A + x^T A^T)v + t(v^T Av) = (x^T A + x^T A^T)v.$$
   This shows
   $$\nabla f(x) = (A + A^T)x.$$

3. Compute the gradient of the quadratic function $f : \mathbb{R}^n \to \mathbb{R}$ given by
   $$f(x) = b + c^T x + x^T Ax,$$
   where $b \in \mathbb{R}$, $c \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n\times n}$.

   Solution. First consider the linear function $g(x) = c^T x$. Note that
   $$g(x+tv) = c^T(x+tv) = c^T x + tc^T v \implies \nabla f(x) = c.$$
   As the derivative is linear we can combine this with the previous problem to obtain
   $$\nabla f(x) = c + (A + A^T)x.$$

4. Fix $s \in \mathbb{R}^n$ and consider $f(x) = (x-s)^T A(x-s)$ where $A \in \mathbb{R}^{n\times n}$. Compute the gradient of f.

   Solution. We give two methods.

(a) Let $g(x) = x^T Ax$ and $h(x) = x - s$ so that $f(x) = g(h(x))$. By the vector-valued form of the chain rule we have

$$\nabla f(x) = \nabla g(h(x))^T Dh(x) = (A + A^T)(x - s),$$

where $Dh(x) = \mathbf{I}_{n\times n}$ is the Jacobian matrix of h.

(b) We have

$$(x - s)^T A(x - s) = x^T Ax - s^T(A + A^T)x + s^T As.$$

Computing the gradient gives

$$\nabla f(x) = (A + A^T)x - (A + A^T)s = (A + A^T)(x - s).$$

5. Consider the ridge regression objective function

$$f(w) = \|Aw - y\|_2^2 + \lambda\|w\|_2^2,$$

where $w \in \mathbb{R}^n$, $A \in \mathbb{R}^{m\times n}$, $y \in \mathbb{R}^m$, and $\lambda \in \mathbb{R}_{\geq 0}$.

(a) Compute the gradient of f.

(b) Express f in the form $f(w) = \|Bw - z\|_2^2$ for some choice of B, z. What do you notice about B?

(c) Using either of the parts above, compute

$$\arg\min_{w\in\mathbb{R}^n} f(w).$$

Solution.

(a) We can express $f(w)$ as

$$f(w) = (Aw - y)^T(Aw - y) + \lambda w^T w = w^T A^T Aw - 2y^T Aw + y^T y + \lambda w^T w.$$

Applying our previous results gives (noting $w^T w = w^T \mathbf{I}_{n\times n} w$)

$$\nabla f(w) = 2A^T Aw - 2A^T y + 2\lambda w = 2(A^T A + \lambda\mathbf{I}_{n\times n})w - 2A^T y.$$

(b) Let

$$B = \begin{pmatrix} A \\ \sqrt{\lambda}\mathbf{I}_{n\times n} \end{pmatrix} \quad \text{and} \quad z = \begin{pmatrix} y \\ \mathbf{0}_{n\times 1} \end{pmatrix}$$

written in block-matrix form. Note B is full rank.

(c) The argmin is $w = (A^T A + \lambda\mathbf{I}_{n\times n})^{-1}A^T y$. To see why the inverse is valid, see the linear algebra questions below.

6. Compute the gradient of

$$f(\theta) = \lambda\|\theta\|_2^2 + \sum_{i=1}^{n} \log(1 + \exp(-y_i\theta^T x_i)),$$

where $y_i \in \mathbb{R}$ and $\theta \in \mathbb{R}^m$ and $x_i \in \mathbb{R}^m$ for $i = 1, \ldots, n$.

Solution. As the derivative is linear, we can compute the gradient of each term separately and obtain

$$\nabla f(\theta) = 2\lambda\theta - \sum_{i=1}^{n} \frac{\exp(-y_i\theta^T x_i)}{1 + \exp(-y_i\theta^T x_i)} y_i x_i,$$

where we used the techniques from Recitation 1 to differentiate the log terms.