Differentiation and its applications

Levent Sagun

New York University

January 28, 2016

Example: Least Squares

Suppose we observe the input $x \in \mathbb{R}^n$, take action $A \in \mathbb{R}^{m \times n}$, and observe the output $b \in \mathbb{R}^m$, and evaluate through mean square.

- Loss function: $L(x) = \frac{1}{2}||Ax - b||_2^2$
- **GOAL:** Minimize $L(x)$ with a gradient-based method.

Example: Least Squares

Suppose we observe the input $x \in \mathbb{R}^n$, take action $A \in \mathbb{R}^{m \times n}$, and observe the output $b \in \mathbb{R}^m$, and evaluate through mean square.

- Loss function: $L(x) = \frac{1}{2}||Ax - b||_2^2$
- **GOAL:** Minimize $L(x)$ with a gradient-based method.
- Gradient: $\nabla_x L(x) = A^T(Ax - b)$
- Descent steps performed in the opposite direction of the gradient:

$$x \leftarrow x - \eta \nabla_x L(x)$$

Example: Least Squares

What are all the symbols in this update rule: $x \leftarrow x - \eta \nabla_x L(x)$?

Remark: Always be aware of what objects are there, and what operations are performed!

Example: Least Squares

What are all the symbols in this update rule: $x \leftarrow x - \eta \nabla_x L(x)$?

- x is a *vector*.
- The arrow replaces LHS with the RHS.
- Minus sign subtracts two *vectors*.
- η is a scalar *number*.
- $L(x)$ is also a scalar *number*.
- $\nabla_x L(x)$ is a *vector*.
- $\eta \nabla_x L(x)$ is a multiplication of a number with a vector.

Remark: Always be aware of what objects are there, and what operations are performed!

2-dimensional case

When $n = m = 2$, we have the following equation:

$$L(x) = \frac{1}{2}(a_{11}x_1 + a_{12}x_2 - b_1)^2 + \frac{1}{2}(a_{21}x_1 + a_{22}x_2 - b_2)^2$$

and its gradient can be computed by partial differentiation:

$$\begin{aligned}\nabla_x L(x) =& (\frac{\partial L(x)}{\partial x_1}, \frac{\partial L(x)}{\partial x_2}) \\ =& ((a_{11}x_1 + a_{12}x_2 - b_1)a_{11} + (a_{21}x_1 + a_{22}x_2 - b_2)a_{21}, \\ & (a_{11}x_1 + a_{12}x_2 - b_1)a_{12} + (a_{21}x_1 + a_{22}x_2 - b_2)a_{22})\end{aligned}$$

This is rather verbose and doesn't give us a hint on how to code derivatives efficiently. How can we get around this?

More examples with summation representation

- *Gradient vector*: For $x \in \mathbb{R}^n$, and A a square matrix, the function $f(x) = x^T Ax$ takes a vector and maps it to a number, $f(x) = \sum_{i,j=1}^{n} x_j a_{ij} x_i$. It's gradient is a vector. The kth component of this vector is:

$$\left(\frac{df}{dx}(x)\right)_k = \frac{df}{dx_k}(x) = \sum_{i=1,j=k}^{i=n} a_{ik}x_i + \sum_{i=k,j=1}^{j=n} x_j a_{kj}$$

- *Jacobian matrix:* $f(x) = Ax$ takes a vector and maps it to another vector, its ith component is given by $f_i(x) = \sum_{k=1}^{n} a_{ik}x_k$, which is a real valued function, hence its gradient can be calculated. Then, the total derivative evaluated at a point x is the matrix composed of component gradient vectors.

$$\left(\frac{df}{dx}(x)\right)_{ij} = \frac{df_i(x)}{dx_j} = a_{ij}$$

Converting back to matrix forms

The computation carried out in the previous slide can be best summarized in matrix form for ease of computation:

- kth component of the *Gradient vector*: $\sum_{i=1,j=k}^{i=n} a_{ik} x_i + \sum_{i=k,j=1}^{j=n} x_j a_{kj} = ((Ax)^T + x^T A)_k$
- An element of the *Jacobian matrix:* $a_{ij} = (A)_{ij}$

The following derivatives are useful to keep in mind:

- $\frac{d}{dx}(x^T Ax) = (Ax)^T + x^T A = x^T(A + A^T)$
- $\frac{d}{dx}(Ax) = A$
- $\frac{d}{dx}(y^T Ax) = y^T A$
- $\frac{d}{dy}(y^T Ax) = (Ax)^T = x^T A^T$

Exercises

- For a vector x and a matrix A identify the type of the following objects:
  1. $x^T x$
  2. $x^T A x$
  3. $x^T A^T + Ax$
  4. $(x^T(\frac{1}{2}A^T A)x)^T x$
- For $f : \mathbb{R}^n \to \mathbb{R}^m$, and $f = (f^1, \ldots, f^m)$ what are the types of the following expressions:
  1. $\frac{df(x)}{dx}$
  2. $\frac{\partial f(x)}{\partial x_i}$
  3. $\frac{\partial f^j(x)}{\partial x}$
  4. $\frac{\partial f^j(x)}{\partial x_i}$

Exercises

- Write the first and second derivatives of $H(x)$ evaluated at a point $x \in \mathbb{R}^n$ $H(x) = \sum_{i,j,k=1}^{N} J_{i,j,k} x_i x_j x_k$. If $J_{i,j,k} \sim \mathcal{N}(0,1)$ and iid, find the mean and variance of $H(x)$.
- Write the first and derivative of $\log L(x)$ where $L(x)$ is $\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\}$, and solve for zero.
- Given a real valued function f on $\mathbb{R}^n$, suppose the domain is constraint on the unit sphere: $S^{n-1}(1) \subset \mathbb{R}^n$. Write an expression for the appropriate gradient descent procedure.
- If a random variable U is uniformly distributed over $[0,1]$, find the distribution of $X = -\frac{1}{\lambda}\log(1-U)$.

Common mistakes: use of confusing indices, getting the wrong object (number instead of a vector), confusing operations (mistaking dot product with scalar multiplication), etc...

Back to gradient descent

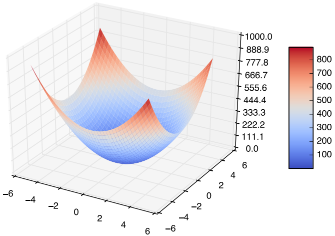
Calculation of the gradient of a scalar function leads to an optimization procedure. We need to be able to calculate the gradient to follow the direction it leads us. But where does this descent take us? If we keep following its lead where will we end up?

- GD takes us to a local minimum in a given landscape.
- There can be more than one such value.
- Not all critical points are local minima!
- Some points have higher index.

Hessian of a scalar-valued, differentiable function is the symmetric matrix formed by its second partial derivatives. It has real eigenvalues. The number of negative eigenvalues of the Hessian is called the index of the function at the evaluation point.

Critical points of scalar functions (with demo)

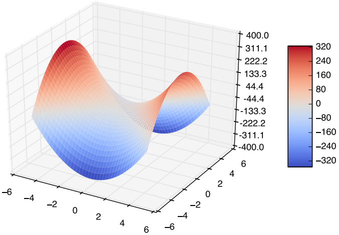
A quadratic function with a minimum (index $= 0$): $x_1^2 + x_2^2 = y$



If f is convex and finite near x, then either

- x minimizes f, or
- there is a descent direction for f at x

A quadratic function with a saddle point (index $= 0$): $x_1^2 - x_2^2 = y$



When does this theorem fail?

- Non-convex: saddles, valleys...

- Unbounded

Directional derivative

Let $f : \mathbb{R}^n \to \mathbb{R}$,

- The gradient at any point of $\mathbb{R}^n$, is the best linear approximation to f at that point.
- For $f(x) = f(x_1, \dots, x_n)$, say we are given a unit vector $v = (v_1, \dots, v_n)$, then the directional derivative in the direction of v is given by

$$\nabla_v f(x) = \lim_{h \to 0} \frac{f(x + hv) - f(x)}{h}$$

- This can be calculated using the gradient: $\nabla_v f(x) = \nabla f(x) \cdot v$
- It can be thought of the value of the rate of change in the direction v.
- Partial derivatives are special cases of this where the v vector are unit coordinate vectors.