Lecture 6

Math Foundations Team

BITS Pilani

Pilani | Dubai | Goa | Hyderabad

Introduction

Many algorithms in machine learning optimize an objective function with respect to a set of desired model parameters that control how well a model explains the data: Finding good parameters can be phrased as an optimization problem.

Examples include: linear regression, where we look at curve-fitting problems and optimize linear weight parameters to maximize the likelihood; neural-network auto-encoders for dimensionality reduction and data compression.

Power Method

Power method is an iterative method to compute dominant eigenvalue of a square matrix $\boldsymbol{A}$ where dominant eigenvalue λ is the eigenvalue such that $|\lambda|$ is greater than the absolute values of other eigenvalues.

1. Select an initial vector $\boldsymbol{x_0}$ whose largest entry is 1.
2. For $k = 0, 1, \cdots$
   a. Compute $\boldsymbol{A}\boldsymbol{x}_k$.
   b. Let μ_k be an entry in $\boldsymbol{A}\boldsymbol{x}_k$ whose absolute value is as large as possible.
   c. Compute $\boldsymbol{x}_{k+1} = (1/\mu_k)\boldsymbol{A}\boldsymbol{x}_k$.

For almost all choices of $\boldsymbol{x_0}$, the sequence μ_k approaches the dominant eigenvalue and the sequence $\boldsymbol{x}_k$ approaches a corresponding eigenvector. (Refer Excel sheet for an example)

- For a symmetric positive definite matrix $\boldsymbol{A}$ of order n, there exists an orthonormal basis consisting of eigenvectors $v_1, v_2, \cdots v_n$ of $\mathbb{R}^n$ corresponding to real eigenvalues $\lambda_1, \lambda_2, \cdots \lambda_n$ (Spectral theorem).
- Suppose there exists a dominant eigenvalue and WLOG we can assume

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \cdots |\lambda_n|$$

Let $\boldsymbol{y} \in \mathbb{R}^n$, then $\boldsymbol{y} = \sum_{i=1}^{n} c_i v_i$. This implies

$$\boldsymbol{A}^k \boldsymbol{y} = \sum_{i=1}^{n} c_i \boldsymbol{A}^k v_i = \sum_{i=1}^{n} c_i \lambda_i^k v_i = \lambda_1^k \{c_1 v_1 + \sum_{i=2}^{n} c_i \{\frac{\lambda_i}{\lambda_1}\}^k v_i\}$$

- Clearly $\{\frac{\lambda_i}{\lambda_1}\}^k \to 0$ as $k \to \infty$. But the convergence depend on the rate of convergence of $\{\frac{\lambda_2}{\lambda_1}\}^k$

Continuous Function

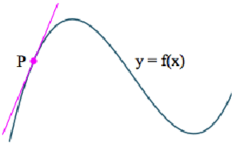
- A function $f : A \to \mathbb{R}$ is said to be continuous at $a \in A$ if for any $\epsilon > 0$, there exist a $\delta > 0$ such thatl $x \in A$ and $|x - a| < \delta$ implies $|f(x) - f(a)| < \epsilon$
- $f(x) = x^2 + 4$ is a continuous function whereas $h(x) = \begin{cases} 1, & \forall x > 0 \\ 2, & \forall x \leq 0 \end{cases}$ is not continuous at $x = 0$.
- A continuous function f on a closed and bounded interval $[a, b]$ is bounded and attains its bounds.
- A continuous function $f : A \to \mathbb{R}$ is increasing or decreasing at a point $a \in A$ implies that there exists an $\epsilon > 0$ such that f is increasing or decreasing in $N_\epsilon = \{x : |x - a| < \epsilon\} \cup A$

Differentiation of Univariate Functions

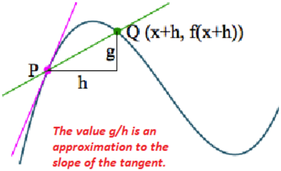
For $h > 0$, the derivative of f at x is defined as the limit

$$\frac{df}{dx} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} \tag{1}$$

The derivative of f points in the direction of steepest ascent of f.



Slope of the tangent at P.



Slope of the line PQ.

Taylor polynomial

The Taylor polynomial is a representation of a function f as an finite sum of terms. These terms are determined using derivatives of f evaluated at x_0.

Definition: The Taylor polynomial of degree n of $f : \mathbb{R} \to \mathbb{R}$ at x_0 is defined as

$$T_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k \tag{2}$$

where $f^{(k)}(x_0)$ is the kth derivative of f at x_0 which we assume exists.

Taylor series

Definition: The Taylor series of smooth (continuously differentiable infinite many times) function $f : \mathbb{R} \to \mathbb{R}$ at x_0 is defined as

$$T_\infty(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k \tag{3}$$

For $x_0 = 0$, we obtain the Maclaurin series as a a special instance of the Taylor series.

Remark: In general, a Taylor polynomial of degree n is an approximation of a function, which does not need to be a polynomial. The Taylor polynomial is similar to f in a neighborhood around x_0. However, a Taylor polynomial of degree n is an exact representation of a polynomial f of degree $k \leq n$ since all derivatives $f^{(i)} = 0$, for $i > k$.

Taylor Polynomial example

Consider the polynomial $f(x) = x^4$. Find the Taylor polynomial T_6 evaluated at $x_0 = 1$.
We compute $f^{(k)}(1)$ for $k = 0, 1, 2..., 6$
$f(1) = 1$, $f'(1) = 4$, $f''(1) = 12$, $f^{(3)}(1) = 24$, $f^{(4)}(1) = 24$, $f^{(5)}(1) = 0$, $f^{(6)}(1) = 0$. The desired Taylor polynomial is

$$\begin{aligned} T_6(x) &= \sum_{k=0}^{6} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k \\ &= 1 + 4(x-1) + 12(x-1)^2 + 24(x-1)^3 + 24(x-1)^4 \\ &= x^4 = f(x) \end{aligned} \tag{4}$$

we obtain an exact representation of the original function.

Taylor Series example

Consider the smooth function $f(x) = sin(x) + cos(x)$. We compute Taylor series expansion of f at $x_0 = 0$, which is the Maclaurin series expansion of f. We obtain the following derivatives:
$f(0) = sin(0) + cos(0) = 1$
$f^{'}(0) = cos(0) - sin(0) = 1$
$f^{''}(0) = -sin(0) - cos(0) = -1$
$f^{(3)}(0) = -cos(0) + sin(0) = -1$
$f^{(4)}(0) = sin(0) + cos(0) = f(0) = 1$
The coefficients in our Taylor series are only ± 1 (since $sin(0) = 0$), each of which occurs twice before switching to the other one.
Furthermore, $f^{(k+4)}(0) = f^{k}(0)$

Taylor Series example

Therefore, the full Taylor series expansion of f at $x_0 = 0$ is given by

$$\begin{aligned}
T_\infty(x) &= \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k \\
&= 1 + x - \frac{1}{2!}x^2 - \frac{1}{3!}x^3 + \frac{1}{4!}x^4 + \frac{1}{5!}x^5 - \dots \\
&= 1 - \frac{1}{2!}x^2 + \frac{1}{4!}x^4 \mp \dots x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5 \mp \dots \\
&= \sum_{k=0}^{\infty}(-1)^k \frac{1}{(2k)!}x^{2k} + \sum_{k=0}^{\infty}(-1)^k \frac{1}{(2k+1)!}x^{2k+1} \\
&= cos(x) + sin(x)
\end{aligned} \tag{5}$$

Differentiation Rules

We denote the derivative of f by $f^{'}$

- Product Rule: $(f(x)g(x))^{'} = f^{'}(x)g(x) + f(x)g^{'}(x)$
- Sum Rule: $(f(x) + g(x))^{'} = f^{'}(x) + g^{'}(x)$
- Quotient Rule: $(\frac{f(x)}{g(x)})^{'} = \frac{f^{'}(x)g(x) - f(x)g^{'}(x)}{(g(x))^2}$
- Chain Rule: $(g(f(x))^{'} = (g \circ f)^{'}(x) = g^{'}(f(x))f^{'}(x)$

Example: Chain Rule

Compute the derivative of function $h(x) = (2x+1)^4$

$h(x) = (2x+1)^4 = g(f(x))$

$f(x) = 2x+1,$

$g(f) = f^4$

Derivatives of f and g are

$f^{'}(x) = 2$

$g^{'}(f) = 4f^3$

$h^{'}(x) = g^{'}(f)f^{'}(x) = (4f^3).2 = 8(2x+1)^3$

Maxima and Minima

- Let f be a real valued function which is differentiable in an open interval (a, b). Then at a point of local maxima and local minima, $f'(x) = 0$.
- Now the error between the function and the Taylor's n degree polynomial is equal to $R_n(x) = \frac{f^{(k+1)}(c)}{(k+1)!}(x - x_0)^{k+1}$ where c lies between x_0 and x. This implies $f(x+h) = f(x) + hf'(x) + \frac{h^2}{2!}f''(x+\theta h)$, for some $\theta \in (0, 1)$.
- Then at a point of local maxima and local minima, $f(x+h) = f(x) + \frac{h^2}{2!}f''(x+\theta h)$. If $f''(x+\theta h) > 0$, x is a point of minima and if $f''(x+\theta h) < 0$, x is a point of maxima. If the second derivative is equal to zero, x is a point of inflection.

An Example

Let

$$\begin{aligned} f(x) &= \cos(x) \\ \text{Then } f^{'}(x) &= 0 \Rightarrow -\sin(x) = 0 \Rightarrow x = 0, \pm n\pi \\ f^{''}(x) &= -\cos(x) \end{aligned}$$

$$\text{Then } f^{''}(x) = \begin{cases} 1, \ \forall x = \pm(2n+1)\pi \\ -1, \ \forall x = \pm(2n)\pi \end{cases} \quad \forall n = 0, 1, \cdots$$

Therefore , $x = \pm(2n+1)\pi$ are points of minima and $x = \pm(2n)\pi$ are points of maxima.

Partial Differentiation and Gradients

Differentiation applies to functions f of a scalar variable $x \in R$. In the following, we consider the general case where the function f depends on one or more variables $x \in R^n$, e.g.,$f(x) = f(x_1, x_2)$. The generalization of the derivative to functions of several variables is the gradient. We find the gradient of the function f with respect to x by varying one variable at a time and keeping the others constant. The gradient is then the collection of these partial derivatives.

Partial derivatives and Gradients

Definition: For a function $f : \mathbb{R}^n \to \mathbb{R}$, $x \to f(x)$, $x \in R^n$ of n variables $x_1, \ldots, x_n$ we define the *partial derivatives* as

$$
\begin{aligned}
\frac{\partial f}{\partial x_1} &= \lim_{h \to 0} \frac{f(x_1 + h, x_2, \ldots, x_n) - f(x_1, x_2, \ldots, x_n)}{h} \\
\frac{\partial f}{\partial x_2} &= \lim_{h \to 0} \frac{f(x_1, x_2 + h, \ldots, x_n) - f(x_1, x_2, \ldots, x_n)}{h} \\
&\vdots \\
\frac{\partial f}{\partial x_n} &= \lim_{h \to 0} \frac{f(x_1, x_2, \ldots, x_n + h) - f(x_1, x_2, \ldots, x_n)}{h}
\end{aligned}
$$

We collect them in the row vector called the gradient of f or Jacobian

$$\Delta_x f = gradf = \frac{df}{dx} = \left[\frac{\partial f(x)}{\partial x_1}, \frac{\partial f(x)}{\partial x_2}, \ldots, \frac{\partial f(x)}{\partial x_n}\right] \tag{6}$$

Example 1: Find the partial derivatives of $f(x, y) = (x + 2y^3)^2$

$$\frac{\partial f(x,y)}{\partial x} = 2(x + 2y^3)\frac{\partial(x + 2y^3)}{\partial x} = 2(x + 2y^3) \tag{7}$$

$$\frac{\partial f(x,y)}{\partial y} = 2(x + 2y^3)\frac{\partial(x + 2y^3)}{\partial y} = 12y^2(x + 2y^3) \tag{8}$$

here we used the chain rule to compute the partial derivatives.

Example 2

Find the partial derivatives of $f(x_1, x_2) = x_1^2 x_2 + x_1 x_2^3$

$$\frac{\partial f(x_1, x_2)}{\partial x_1} = 2x_1 x_2 + x_2^3 \tag{9}$$

$$\frac{\partial f(x_1, x_2)}{\partial x_2} = x_1^2 + 3x_1 x_2^2 \tag{10}$$

So the gradient is then

$$\frac{df}{dx} = [\frac{\partial f(x_1, x_2)}{\partial x_1}, \frac{\partial f(x_1, x_2)}{\partial x_2}] = [2x_1 x_2 + x_2^3, x_1^2 + 3x_1 x_2^2] \in \mathbb{R}^{1\times 2} \tag{11}$$

Basic rules of partial differentiation

When we compute derivatives with respect to vectors $x \in \mathbb{R}^n$ we need to pay attention: Our gradients now involve vectors and matrices, and matrix multiplication is not commutative i.e., the order matters.

$$\text{Product rule: } \frac{\partial}{\partial x}(f(x)g(x)) = \frac{\partial f}{\partial x}g(x) + f(x)\frac{\partial g}{\partial x} \tag{12}$$

$$\text{Sum rule: } \frac{\partial}{\partial x}(f(x)+g(x)) = \frac{\partial f}{\partial x} + \frac{\partial g}{\partial x} \tag{13}$$

$$\text{chain rule: } \frac{\partial}{\partial x}(g \circ f)(x) = \frac{\partial}{\partial x}(g(f(x)) = \frac{\partial g}{\partial f}\frac{\partial f}{\partial x} \tag{14}$$

Chain Rule

Consider a function $f : \mathbb{R} \to \mathbb{R}$ of two variables x_1, x_2. Furthermore, $x_1(t)$ and $x_2(t)$ are themselves functions of t.

To compute the gradient of f with respect to t, we need to apply the chain rule for multivariate functions as

$$\frac{df}{dt} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix} \begin{bmatrix} \frac{\partial x_1(t)}{\partial t} \\ \frac{\partial x_2(t)}{\partial t} \end{bmatrix} = \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial t} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial t} \tag{15}$$

where d denotes the gradient and ∂ partial derivatives.

Example

Consider $f(x_1, x_2) = x_1^2 + 2x_2$, where $x_1 = \sin t$ and $x_2 = \cos t$ then

$$\begin{aligned}
\frac{df}{dt} &= \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial t} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial t} \\
&= 2\sin t\frac{\partial \sin t}{\partial t} + 2\frac{\partial \cos t}{\partial t} \\
&= 2\sin t\cos t - 2\sin t = 2\sin t(\cos t - 1)
\end{aligned}$$

is the corresponding derivative of f with respect to t.

If $f(x_1, x_2)$ is a function of x_1 and x_2, where $x_1(s,t)$ and $x_2(s,t)$ are themselves functions of two variables s and t, the chain rule yields the partial derivatives:

$$\frac{\partial f}{\partial s} = \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial s} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial s} \tag{16}$$

$$\frac{\partial f}{\partial t} = \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial t} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial t} \tag{17}$$

and the gradient is obtained by the matrix multiplication

$$\begin{aligned}\frac{df}{d(s,t)} &= \frac{\partial f}{\partial x}\frac{\partial x}{\partial (s,t)} \\ &= \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix} \begin{bmatrix} \frac{\partial x_1}{\partial s} & \frac{\partial x_1}{\partial t} \\ \frac{\partial x_2}{\partial s} & \frac{\partial x_2}{\partial t} \end{bmatrix}\end{aligned}$$

Gradients of Vector-Valued Functions

We have discussed partial derivatives and gradients of functions $f : \mathbb{R}^n \to \mathbb{R}$ mapping to the real numbers. Now we will generalize the concept of the gradient to vector-valued functions $f : \mathbb{R}^n \to \mathbb{R}^m$, where $n \geq 1$ and $m > 1$.
For a function $f : \mathbb{R}^n \to \mathbb{R}^m$ and a vector $x = [x_1, \ldots, x_n]^T$ corresponding vector of function values is given as

$$f(x) = \begin{bmatrix} f_1(x) \\ \vdots \\ f_m(x) \end{bmatrix} \in \mathbb{R}^m \tag{18}$$

where each $f_i : \mathbb{R}^n \to \mathbb{R}$

Gradients of Vector-Valued Functions

Therefore, the partial derivative of a vector-valued function $f : \mathbb{R}^n \to \mathbb{R}^m$ w.r.t. $x_i \in R$, $i = 1, \dots n$ is given as the vector

$$\frac{\partial f}{\partial x_i} = \begin{bmatrix} \frac{\partial f_1}{\partial x_i} \\ \vdots \\ \frac{\partial f_m}{\partial x_i} \end{bmatrix}$$

$$= \begin{bmatrix} \lim_{h \to 0} \frac{f_1(x_1, \dots, x_{i-1}, x_i + h, x_{i+1}, \dots, x_n) - f_1(x)}{h} \\ \vdots \\ \lim_{h \to 0} \frac{f_m(x_1, \dots, x_{i-1}, x_i + h, x_{i+1}, \dots, x_n) - f_m(x)}{h} \end{bmatrix} \in \mathbb{R}^m$$

Gradients of Vector-Valued Functions

We know that the gradient of f with respect to a vector is the row vector of the partial derivatives. Every partial derivative $\frac{\partial f}{\partial x_i}$ is itself a column vector. Therefore, we obtain the gradient of $f : \mathbb{R}^n \to \mathbb{R}^m$ with respect to $x \in \mathbb{R}^n$ by collecting these partial derivatives:

$$\frac{df(x)}{dx} = \left[\frac{\partial f(x)}{\partial x_1} \cdots \frac{\partial f(x)}{\partial x_n} \right]$$

$$= \begin{bmatrix} \frac{\partial f_1(x)}{\partial x_1} & \cdots & \frac{\partial f_1(x)}{\partial x_n} \\ & \vdots & \\ \frac{\partial f_m(x)}{\partial x_1} & \cdots & \frac{\partial f_m(x)}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{m \times n}$$

Example 1: Gradients of Vector-Valued Functions

Given $f(x) = Ax, \quad f(x) \in \mathbb{R}^M, \quad A \in \mathbb{R}^{M\times N}, \quad x \in \mathbb{R}^N$

Since $f : \mathbb{R}^N \to \mathbb{R}^M$, it follows that $df/dx \in \mathbb{R}^{M\times N}$. To compute the gradient we determine the partial derivatives of f w.r.t x_j:

$$f_i(x) = \sum_{i=1}^{N} A_{ij}x_j \implies \frac{\partial f_i}{\partial x_j} = A_{ij} \tag{19}$$

We obtain the gradient using Jacobian

$$\frac{df}{dx} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} \cdots \frac{\partial f_1}{\partial x_N} \\ \vdots \\ \frac{\partial f_M}{\partial x_1} \cdots \frac{\partial f_M}{\partial x_N} \end{bmatrix} = \begin{bmatrix} A_{11} \dots A_{1N} \\ \vdots \\ A_{M1} \dots A_{MN} \end{bmatrix} = A \in \mathbb{R}^{M\times N} \tag{20}$$

Example 2: Gradients of Vector-Valued Functions

Consider the function $h : \mathbb{R} \to \mathbb{R}, \quad h(t) = (f \circ g)(t)$ with $f(x) = exp(x_1 x_2^2)$

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = g(t) = \begin{bmatrix} t \cos t \\ t \sin t \end{bmatrix} \tag{21}$$

and compute the gradient of h w.r.t. t. Since $f : \mathbb{R}^2 \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}^2$ we note that

$$\frac{\partial f}{\partial x} \in \mathbb{R}^{1\times 2} \text{ and } \frac{\partial g}{\partial t} \in \mathbb{R}^{2\times 1} \tag{22}$$

The desired gradient is computed by applying the chain rule:

$$
\begin{aligned}
\frac{dh}{dt} &= \frac{\partial f}{\partial x}\frac{\partial x}{\partial t} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix} \begin{bmatrix} \frac{\partial x_1}{\partial t} \\ \frac{\partial x_2}{\partial t} \end{bmatrix} \\
&= \begin{bmatrix} exp(x_1 x_2^2) x_2^2 & 2exp(x_1 x_2^2) x_1 x_2 \end{bmatrix} \begin{bmatrix} \cos t - t \sin t \\ \sin t + t \cos t \end{bmatrix} \\
&= exp(x_1 x_2^2)(x_2^2(\cos t - t \sin t) + 2 x_1 x_2 (\sin t + t \cos t))
\end{aligned}
$$

where $x_1 = t \cos t$ and $x_2 = t \sin t$;