Applied Linear Algebra in Data Analysis

Introduction to Optimization

Sivakumar Balasubramanian

Department of Bioengineering
Christian Medical College, Bagayam
Vellore 632002

Optimization

- Optimization is the process of finding the best solution to a problem from a set of possible solutions.

- Optimization problems come up in many applications in engineering, science, economics, biology, medicine, operations research, etc.

- Optimization problems can be classified in different ways, but one major classification gives us: **unconstrained** and **constrained** optimization problems.

A general optimization problem

- A general optimization problem can be fomulated as the following,

$$\begin{aligned} &\min_{\mathbf{x}\in\mathcal{X}} \; f(\mathbf{x}) \\ \text{subject to } &\mathbf{g}(\mathbf{x}) \leq \mathbf{0},\; \mathbf{g}(\mathbf{x}) = \begin{bmatrix} g_1(\mathbf{x}) & g_2(\mathbf{x}) & \cdots & g_p(\mathbf{x}) \end{bmatrix}^\top \\ &\mathbf{h}(\mathbf{x}) = \mathbf{0},\; \mathbf{h}(\mathbf{x}) = \begin{bmatrix} h_1(\mathbf{x}) & h_2(\mathbf{x}) & \cdots & h_q(\mathbf{x}) \end{bmatrix}^\top \end{aligned}$$

  where, $f(\mathbf{x})$ is the **objective function** and $\mathbf{g}(\mathbf{x})$ represents the set of **inquality constaints** and $\mathbf{h}(\mathbf{x})$ represents the set of **equality constraints**.

- In this course, we will only focus on optimization problems over $\mathbb{R}^n$, and mostly problems where the objective function and the constraints are differentiable.

A general optimization problem

- Most optimization problems of practical significance cannot be solved analytically, and we must resort to numerical iterative methods to find a solution.

- We can never solve these problems exactly through numerical means, and must content outselves with finding an approximate “good enough” solution.

Mathematical preliminaries: Sequences and Limits

We first review the notions of continuity and differentiability of functions of single and multiple variables, since we will be dealing with differentiable functions in optimization problems.

Sequences and Limits:

- A sequence of real numbers is a function whose domain is a set of natural numbers $1, 2, \ldots, k, \ldots$ and whose range is a set of real numbers. The sequence is denoted by $\{x_k\}_{k=1}^{\infty}$ or $\{x_k\}$.

- A number x^* is said to be the **limit** of the sequence $\{x_k\}$ if for every $\epsilon > 0$, there exists an integer K such that for all $k > K$, we have $|x_k - x^*| < \epsilon$.

$$\lim_{k \to \infty} x_k = x^* \quad \text{or} \quad x_k \to x^*$$

A sequence that has a limit is called a **convergent sequence**.

Sequences and Limits

We can extend these ideas to $\mathbb{R}^n$.

- A sequence in $\mathbb{R}^n$ is a function whose domain is a set of natural numbers $1, 2, \ldots, k, \ldots$ and whose range is $\mathbb{R}^n$. The sequence is denoted by $\{\mathbf{x}_k\}_{k=1}^{\infty}$ or $\{\mathbf{x}_k\}$.

- $\mathbf{x}^*$ is said to be the **limit** of the sequence $\{\mathbf{x}_k\}$ if for every $\epsilon > 0$, there exists an integer K such that for all $k > K$, we have $\|\mathbf{x}_k - \mathbf{x}^*\| < \epsilon$.

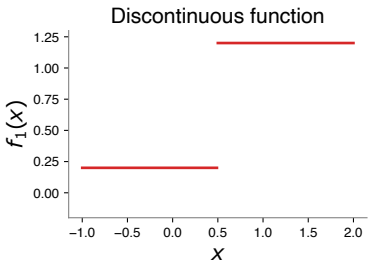
$$\lim_{k \to \infty} \mathbf{x}_k = \mathbf{x}^* \quad \text{or} \quad \mathbf{x}_k \to \mathbf{x}^*$$

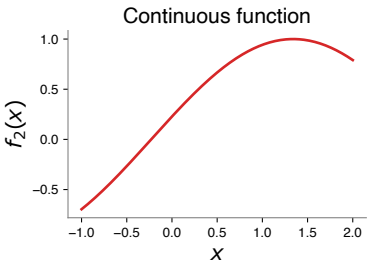
- The limit of a convergent sequence is unique.

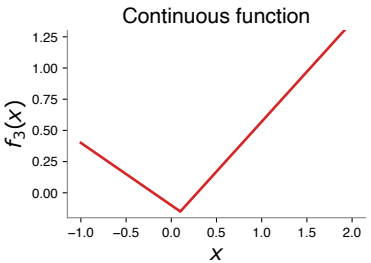
Continuity

Consider the function $f : \Omega \to \mathbb{R}$, where $\Omega \subseteq \mathbb{R}^n$. This function is continuous at the point $\mathbf{x}_0 \in \Omega$, if and only if,

$$\lim_{\mathbf{x} \to \mathbf{x}_0} f(\mathbf{x}) = f(\mathbf{x}_0)$$







Differentiability

Differentiability is a local property of a function, like continuity.
Consider a function $f : \Omega \to \mathbb{R}$, where $\Omega \subseteq \mathbb{R}$. Let $x_0 \in \Omega$,

$$\frac{\delta f\left(x_0\right)}{\delta x} = \frac{f\left(x_0 + \delta x\right) - f\left(x_0\right)}{\delta x}$$

The function f is said to be differentiable at the point $x_0 \in \Omega$, if and only if,

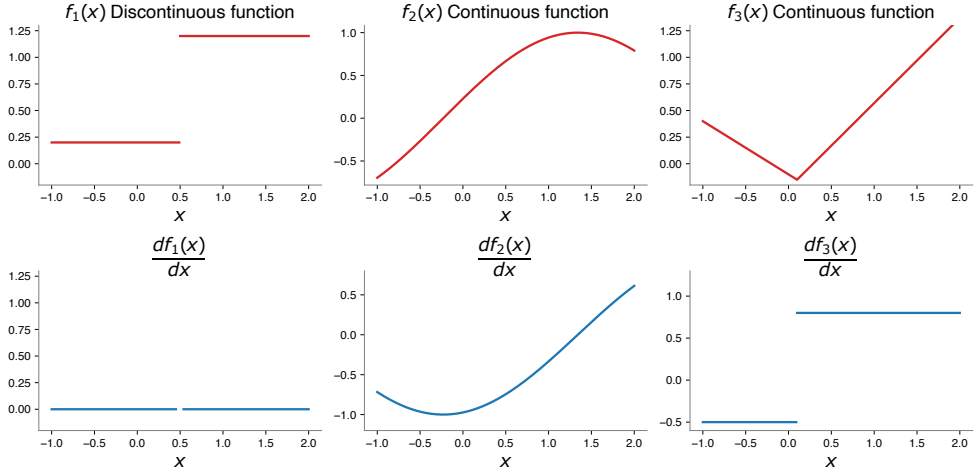
- $f\left(x\right)$ is continuous at x_0.
- $\lim_{\delta x \to 0} \frac{\delta f(x_0)}{\delta x} = \lim_{\delta x \to 0^-} \frac{\delta f(x_0)}{\delta x} = \lim_{\delta x \to 0^+} \frac{\delta f(x_0)}{\delta x}$
- $\lim_{\delta x \to 0} \frac{f(x_0)}{\delta x}$ is finite.

Then the derivative of the function f at the point x_0 is defined as,

$$\frac{f\left(x_0\right)}{dx} = \lim_{\delta x \to 0} \frac{f\left(x_0 + \delta x\right) - f\left(x_0\right)}{\delta x}$$

Differentiability

Three functions f_1, f_2, f_3 defined over the set $\Omega = [-1, 2] \subseteq \mathbb{R}$.



Differentiability in $\mathbb{R}^n$

Consider the function $f : \Omega \to \mathbb{R}$, where $\Omega \subseteq \mathbb{R}^n$.

$$f(\mathbf{x}) = f(x_1, x_2, \ldots, x_n)$$

f maps a column vector $\mathbf{x} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}^\top \in \mathbb{R}^n$ to a real number.

The partial derivative of the function $f(\mathbf{x})$ at $\mathbf{x}_0$ is defined as,

$$\frac{\partial f(\mathbf{x}_0)}{\partial x_i} = \lim_{\delta x \to 0} \frac{f(\mathbf{x}_0 + \delta x\, \mathbf{e}_i) - f(\mathbf{x}_0)}{\delta x}$$

$\frac{\partial f(\mathbf{x})}{\partial x_i}$ is the rate of change of the function f when move along the i-th coordinate direction at the point $\mathbf{x}_0$.

The function f is said to be differentiable at the point $\mathbf{x}_0 \in \Omega$, if and only if, the partial derivatives of the function f w.r.t. all x_i.

Differentiability in $\mathbb{R}^n$

The derivative of the function $f : \Omega \to \mathbb{R}$, where $\Omega \subseteq \mathbb{R}^n$ with respect to the column vector $\mathbf{x}$ at the point $\mathbf{x}_0 \in \Omega$ is defined as the following,

$$\nabla f(\mathbf{x}_0) = \left[\frac{\partial f}{\partial x_1}(\mathbf{x}_0) \quad \frac{\partial f}{\partial x_2}(\mathbf{x}_0) \quad \cdots \quad \frac{\partial f}{\partial x_n}(\mathbf{x}_0) \right] \in \mathbb{R}^n$$

Notice that $\nabla f(\mathbf{x}_0)$ is a row vector, and it is called the *gradient* of the function f at the point $\mathbf{x}_0$.

We follow the following convention when dealing with derivative of functions of multiple variables $f : \Omega \to \mathbb{R}$:

- The gradient with respect to a column vector $\mathbf{x}$ is a row vector $\nabla_{\mathbf{x}} f(\mathbf{x})$.

$$\nabla_{\mathbf{x}} f(\mathbf{x}) = \left[\frac{\partial f}{\partial x_1} \quad \frac{\partial f}{\partial x_2} \quad \cdots \quad \frac{\partial f}{\partial x_n} \right]$$

- The gradient with respect to a row vector $\mathbf{x}^\top$ is a column vector $\nabla_{\mathbf{x}^\top} f(\mathbf{x})$.

$$\nabla_{\mathbf{x}^\top} f(\mathbf{x}) = \left[\frac{\partial f}{\partial x_1} \quad \frac{\partial f}{\partial x_2} \quad \cdots \quad \frac{\partial f}{\partial x_n} \right]^\top$$

Differentiability in $\mathbb{R}^n$: Jacobian of a Vector-valued function

Consider the function $\mathbf{h} : \mathbb{R}^q \to \mathbb{R}^p$, where

$$\mathbf{h}(\mathbf{x}) = \begin{bmatrix} h_1(\mathbf{x}) & h_2(\mathbf{x}) & \cdots & h_p(\mathbf{x}) \end{bmatrix}^\top \quad \mathbf{x} \in \mathbb{R}^q$$

The *Jacobian* of the function $\mathbf{h}(\mathbf{x})$ with respect to $\mathbf{x} \in \mathbb{R}^q$ is defined as the following matrix,

$$\nabla_{\mathbf{x}} \mathbf{h}(\mathbf{x}) \triangleq \begin{bmatrix} \nabla_{\mathbf{x}} h_1(\mathbf{x}) \\ \nabla_{\mathbf{x}} h_2(\mathbf{x}) \\ \vdots \\ \nabla_{\mathbf{x}} h_q(\mathbf{x}) \end{bmatrix}^\top \in \mathbb{R}^{p \times q}$$

Differentiability in $\mathbb{R}^n$: Hessian Matrices

Consider the function $f : \mathbb{R}^n \to \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^n$.

The Hessian matrix $\mathbf{H}_f(\mathbf{x})$ of the function $f(\mathbf{x})$ is defined as the symmetric matrix $n \times n$ matrix of the second order partial derivatives of f with respect to the components of $\mathbf{x}$, assuming all the second order partial derivatives exists.

The ij^{th} element of the Hessian matrix of $f(\mathbf{x})$ is given by.

$$[\mathbf{H}_f(\mathbf{x})]_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x}) = \frac{\partial}{\partial x_i}\left(\frac{\partial f}{\partial x_j}(\mathbf{x})\right) = \frac{\partial}{\partial x_j}\left(\frac{\partial f}{\partial x_i}(\mathbf{x})\right)$$

$$\mathbf{H}_f(\mathbf{x}) \triangleq \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix} \qquad \mathbf{H}_f(\mathbf{x}) = \nabla_{\mathbf{x}^\top}(\nabla_{\mathbf{x}} f(\mathbf{x})) = \nabla_{\mathbf{x}}(\nabla_{\mathbf{x}^\top} f(\mathbf{x}))$$

Gradient of a function $f : \mathbb{R}^n \to \mathbb{R}$

The levels set of the function $f : \mathbb{R}^n \to \mathbb{R}$ at the level $c \in \mathbb{R}$ is defined as,

$$S = \{\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^n . f(\mathbf{x}) = c\}$$

A level set is a curve for functions $f : \mathbb{R}^2 \to \mathbb{R}$, and is a surface when $f : \mathbb{R}^3 \to \mathbb{R}$.

The different level sets are also called the contours of the function f.

The gradient of the function f at a point $\mathbf{x}_0$ is orthogonal to the level set of the function f at the value $f(\mathbf{x}_0)$.

The gradient is also the direction in $\mathbb{R}^n$ of maximal increase of the value of the function f. This is also called the direction of *steepest ascent*.

Taylor's Theorem

Many results from analysis are used in optimization problems – one of them is the "Taylor's" theorem.
The Taylor's theorem gives an polynomial approximation of a k time differentiaable function $f : \mathbb{R} \to \mathbb{R}$ around at a given point x_0 by a k^{th} order **Taylor polynomial**.
For a smooth function (infinitely differentiable), the k^{th} order Taylor polynomial is a truncation at the order k of the Taylor series expansion of the function f around the point x_0.
Taylor's Theoerem: Suppose a function $f : \mathbb{R} \to \mathbb{R}$ is k times differentiable at a point x_0, then the function f can be approximated by the following polynomial with $\epsilon = x - x_0$,

$$f(x) = f(x_0) + Df(x_0)\frac{\epsilon}{1!} + D^2 f(x_0)\frac{\epsilon^2}{2!} + \cdots + D^k f(x_0)\frac{\epsilon^k}{(k)!} + o\left(\epsilon^k\right)$$

where, $D^l f(x_0)$ is the l^{th} order derivative of the function f at the point x_0, and $o\left(\epsilon^k\right)$ is the remainder term which trends to zero faster than the function ϵ^k as $\epsilon \to 0$.

Taylor's Theorem

Now consider a function $f : \mathbb{R}^n \to \mathbb{R}$ and $\mathbf{x}_0 \in \mathbb{R}^n$, and let's assume that f is differentiable twice with respect to $\mathbf{x}$, and let $\boldsymbol{\epsilon} = \mathbf{x} - \mathbf{x}_0$. The polynomial approximation of f is given by,

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \frac{1}{1!}\mathbf{g}(\mathbf{x}_0)^\top \boldsymbol{\epsilon} + \frac{1}{2!}\boldsymbol{\epsilon}^\top \mathbf{H}(\mathbf{x}_0)\boldsymbol{\epsilon} + o\left(\|\boldsymbol{\epsilon}\|_2^2\right)$$

where, $\mathbf{g}(\mathbf{x}_0) = \nabla_{\mathbf{x}^\top} f(\mathbf{x}_0)$ is the gradient of the function f with respect to the $\mathbf{x}^\top$ computed at $\mathbf{x}_0$, and $\mathbf{H}(\mathbf{x}_0)$ is the Hessian of the functon f computed at $\mathbf{x}_0$.

Local and Global Minimizers

We distinguih between two types of minimizers of a function $f : \mathbb{R}^n \to \mathbb{R}$: Global and local minimizers.

Global minimizer: A point $\mathbf{x}^* \in \mathbb{R}^n$ is said to be a *global minimizer* of the function $f(\mathbf{x})$ if and only if, $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^n - \{\mathbf{x}^*\}$.

A global minimizer is a *strict global minimizer* if $f(\mathbf{x}^*) < f(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^n - \{\mathbf{x}^*\}$.

Local minimizer: A point $\mathbf{x}^* \in \mathbb{R}^n$ is said to be a *local minimizer* of the function $f(\mathbf{x})$ over the set $\Omega \subset \mathbb{R}^n$, if there exists $\epsilon > 0$ such that $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^n - \{\mathbf{x}^*\}$ and $\|\mathbf{x} - \mathbf{x}^*\| < \epsilon$.

This is a *strict local minimizer* if $f(\mathbf{x}^*) < f(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^n - \{\mathbf{x}^*\}$ and $\|\mathbf{x} - \mathbf{x}^*\| < \epsilon$.

Conditions for Local Minimizers

Consider the twice differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, and let with gradient vector $\nabla_{\mathbf{x}^\top} f(\mathbf{x})$ and Hessian matrix $\mathbf{H}(\mathbf{x})$.

First order necessary condition (FONC) for local minimizers: If $\mathbf{x}^*$ is a local minimizer of f, then

$$\nabla_{\mathbf{x}^\top} f(\mathbf{x}^*) = \mathbf{0}$$

Second order necessary condition (SONC) for local minimizers: If $\mathbf{x}^*$ is a local minimizer of f, then

$$\nabla_{\mathbf{x}^\top} f(\mathbf{x}^*) = \mathbf{0} \quad \text{and} \quad \mathbf{d}^\top \mathbf{H}(\mathbf{x}^*)\,\mathbf{d} \geq 0,\ \mathbf{d} \in \mathbb{R}^n$$

Second order sufficient condition (SONC) for local minimizers: If $\mathbf{x}^*$ is a local minimizer of f, then

$$\nabla_{\mathbf{x}^\top} f(\mathbf{x}^*) = \mathbf{0} \quad \text{and} \quad \mathbf{d}^\top \mathbf{H}(\mathbf{x}^*)\,\mathbf{d} > 0,\ \mathbf{d} \in \mathbb{R}^n$$

Unconstrained Optimization: Single variable case

Consider a function $f : \mathbb{R} \to \mathbb{R}$, and we are interested in finding the minimizer x^*.

The SOSC for this case is: $\frac{df(x)}{dx} = 0$ and $\frac{d^2 f(x)}{dx^2} > 0$.

We might not be able to solve things analytically even for the single variable case, and will need to resort to iterative approaches. Such methods are called *line search* methods.

Iterative search methods: We start with an initial guess x_0, and then update the guess using a rule,

$$x_{k+1} = x_k + \alpha_k h\left(f\left(x_k\right)\right), \ k = 0, 1, 2, \ldots$$

where, α_k is the step size, and $h\left(f\left(x_k\right)\right)$ is the search direction.

The iteration is continued until some stopping criteria are satisfied.

Line Search Algorithm: Newton's method

Line search algorithms may use the value of the function f at different points, the first derivative f' or even the second derivative f''.

One of the most common line search methods is the Newton's methods, which uses the first and second derivatives of the function f to iteratively compute a local minimizer for a function.
At any given iteration k, the Newton's method uses $f(x_k)$, $f'(k)$, and $f''(x_k)$ to fit a quadratic approximation of the function as the following,

$$q(x) = f(x_k) + f'(x_k)(x - x_k) + \frac{1}{2} f''(x_k)(x - x_k)^2$$

We mininmize quadratic this approximiation $q(x)$ to find the next guess x_{k+1}. By settinng, $q'(x) = f1(x_k) + f''(x - x_k) = 0$, we get the next guess as,

$$x_{k+1} = x_k - \frac{f`(x_k)'}{f''(x_k)}$$

Line Search Algorithm: Secant method

What if we did not have access to the f''? We can use an approximation for f'' instead, which gives us the Secant method for line search.

f' **is unknown**: We can use f' to approximate f'' as the following,

$$\hat{f}''(x_k) = \frac{f'(x_k) - f'(x_{k-1})}{x_k - x_k - 1}$$

Using this approximation in the Newton's method and simplifying the expression, we get

$$x_{k+1} = \frac{f'(x_k)\, x_{k-1} - f'(x_{k-1})\, x_k}{f'(x_k) - f'(x_{k-1})}$$

It left as an exercise to shown that x_{k+1} is the minimizer of the quadratic approximation of the function f at the point x_k.

Line Search Algorithm: Secant method

Both the Newton's and Secant methods are examples of *quadratic fit* methods.

A third possible method foregoes the requirement of the first derivative f' and use only the value of the function at three points to fit a quadratic approximation.

The derivation of the iteration rule for this method is left as an exercise.