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,

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

  where, $f\left(\mathbf{x}\right)$ is the **objective function** and $\mathbf{g}\left(\mathbf{x}\right)$ represents the set of **inquality constaints** and $\mathbf{h}\left(\mathbf{x}\right)$ 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 $\|\mathbf{x}_k - \mathbf{x}^*\| < \epsilon$.

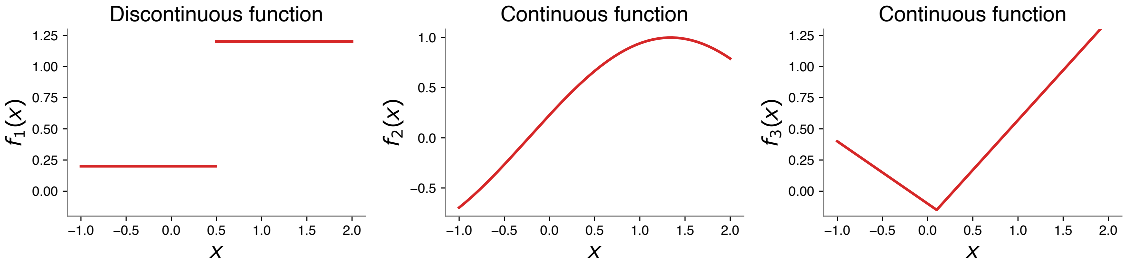
$$\lim_{k \to \infty} \mathbf{x}_k = \mathbf{x}^* \quad \text{or} \quad \mathbf{x}_k \to \mathbf{x}^*$$

- The limit of a convergent sequence is unique.

Continuity

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

$$\lim_{\mathbf{x} \rightarrow \mathbf{x}_0} f\left(\mathbf{x}\right) = f\left(\mathbf{x}_0\right)$$



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,

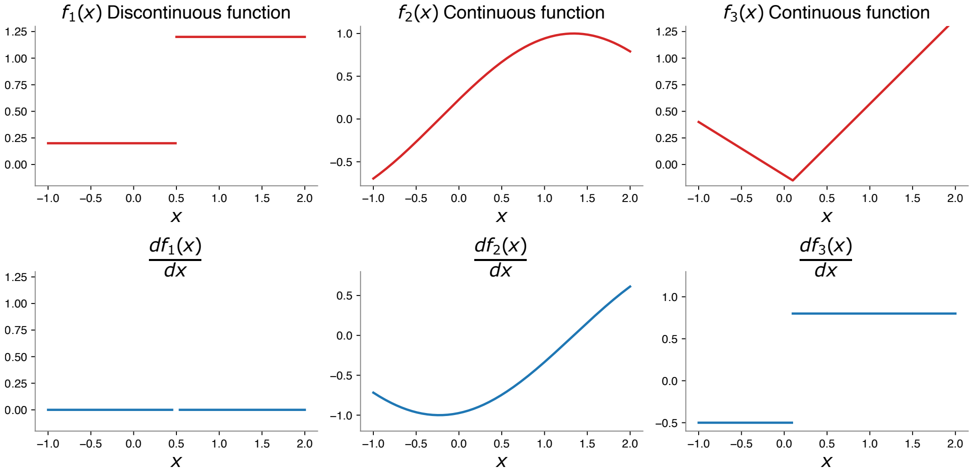
- $f\left(x\right)$ is continuous at x_0.
- $\lim_{\delta x \to 0} \frac{\delta f(x_0)}{\delta x} = \lim_{\delta x \to 0^-} \frac{\delta f(x_0)}{\delta x} = \lim_{\delta x \to 0^+} \frac{\delta f(x_0)}{\delta x}$
- $\lim_{\delta x \to 0} \frac{f(x_0)}{\delta x}$ is finite.

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,

$$\nabla f(\mathbf{x}_0) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(\mathbf{x}_0) & \frac{\partial f}{\partial x_2}(\mathbf{x}_0) & \cdots & \frac{\partial f}{\partial x_n}(\mathbf{x}_0) \end{bmatrix} \in \mathbb{R}^n$$

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 $\nabla_{\mathbf{x}} f(\mathbf{x})$.

$$\nabla_{\mathbf{x}} f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \cdots & \frac{\partial f}{\partial x_n} \end{bmatrix}$$

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

$$\nabla_{\mathbf{x}^\top} f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \cdots & \frac{\partial f}{\partial x_n} \end{bmatrix}^\top$$

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.

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

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

Steepest descent algorithm

- Consider the experiment tossing a dice, and we observe the count of the dots that turn on the top face of the dice.
  - Observed outcome is an even number. $A = \{2, 4, 6\} \subset S$
  - Observed outcome is a positive number. $A = S \implies$ **Sure event**
  - Observed outcome is 0. $A = \{\} \implies$ **Impossible event**
- For discrete sample spaces and **elementary event** is an event with just single sample point.
- We can combine events to produce other events that might be of interest to us. Set operations can be used to perform algebra on events.