... ONLY, IT ISN'T!

Introduction to Optimization

K. R. Sahasranand

Data Science

sahasranand@iitpkd.ac.in

Optimization so far...

For $\Omega = ?$

Least squares

$$\min_{x \in \Omega} \|Ax - b\|^2$$

Optimization so far...

For $\Omega = \{x \in \mathbb{R}^n\}$ Least squares

$$\min_{x \in \Omega} \|Ax - b\|^2$$

Optimization so far...

For $\Omega = \{x \in \mathbb{R}^n\}$ Least squares

$$\min_{x \in \Omega} \|Ax - b\|^2$$

For $\Omega = ?$ Minimum ℓ_2 norm

$$\min_{x \in \Omega} x_1^2 + \cdots + x_n^2.$$

Optimization so far...

For $\Omega = \{x \in \mathbb{R}^n\}$ Least squares

$$\min_{x \in \Omega} \|Ax - b\|^2$$

For $\Omega = \{x \in \mathbb{R}^n : Ax = b\}$ Minimum ℓ_2 norm

$$\min_{x \in \Omega} x_1^2 + \cdots + x_n^2.$$

Optimization so far...

For $\Omega = \{x \in \mathbb{R}^n\}$ Least squares

$$\min_{x \in \Omega} \|Ax - b\|^2$$

For $\Omega = \{x \in \mathbb{R}^n : Ax = b\}$ Minimum ℓ_2 norm

$$\min_{x \in \Omega} x_1^2 + \cdots + x_n^2.$$

For $\Omega = ?$ Linear Programming

$$\min_{x \in \Omega} c^\top x$$

Optimization so far...

For $\Omega = \{x \in \mathbb{R}^n\}$ Least squares

$$\min_{x \in \Omega} \|Ax - b\|^2$$

For $\Omega = \{x \in \mathbb{R}^n : Ax = b\}$ Minimum ℓ_2 norm

$$\min_{x \in \Omega} x_1^2 + \cdots + x_n^2.$$

For $\Omega = \{x \in \mathbb{R}^n : Ax \geq b\}$ Linear Programming

$$\min_{x \in \Omega} c^\top x$$

Optimization so far...

For $\Omega = \{x \in \mathbb{R}^n\}$ Least squares

$$\min_{x \in \Omega} \|Ax - b\|^2$$

For $\Omega = \{x \in \mathbb{R}^n : Ax = b\}$ Minimum ℓ_2 norm

$$\min_{x \in \Omega} x_1^2 + \cdots + x_n^2.$$

For $\Omega = \{x \in \mathbb{R}^n : Ax \geq b\}$ Linear Programming

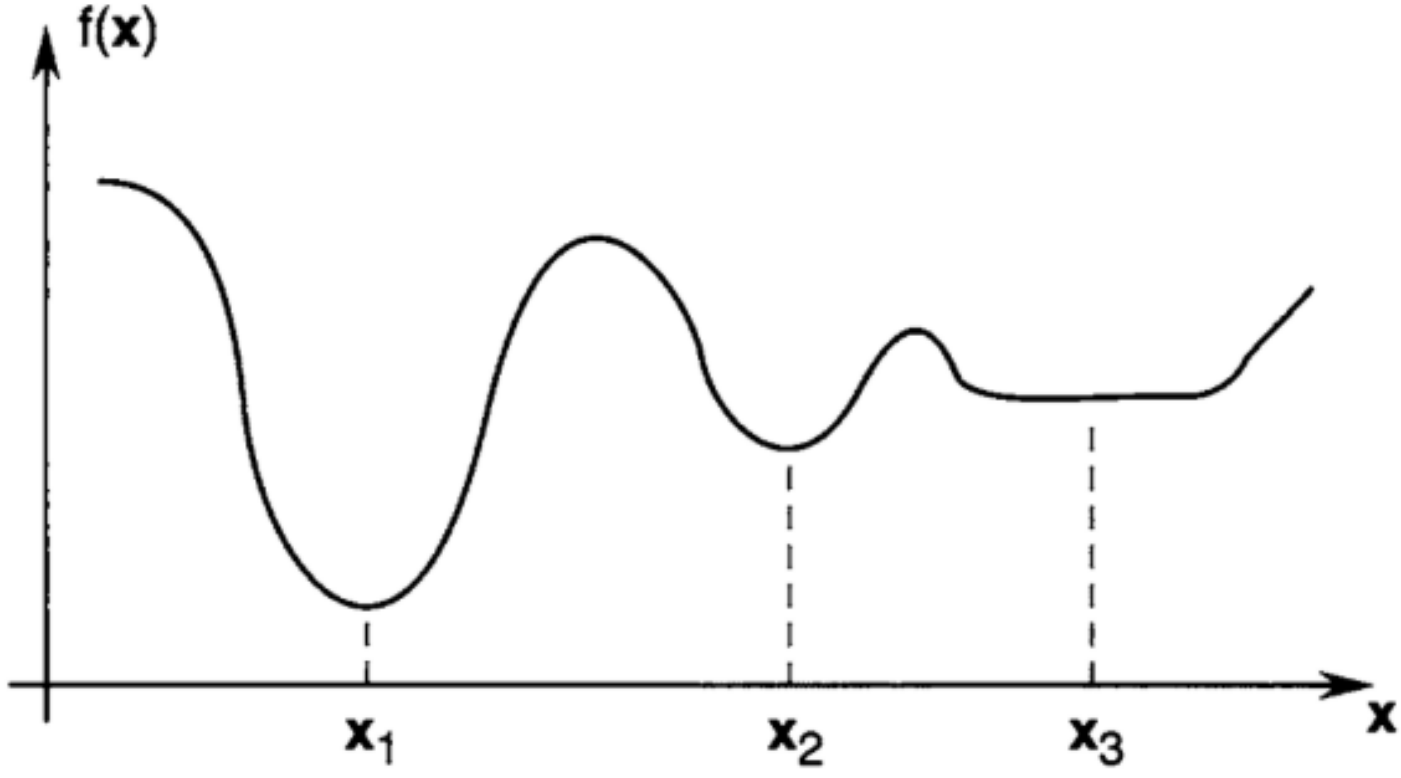
$$\min_{x \in \Omega} c^\top x$$

For $\Omega = \{x \in \mathbb{R}^n : x \text{ satisfies some constraint}\}$ General optimization

$$\min_{x \in \Omega} f(x)$$

Minimizers

global and local



What quantity is **zero** at the minimum?

What is the derivative of f at x_0?

The subject matter of this Chapter is nothing else but the elementary theorems of Calculus, which however are presented in a way which will probably be new to most students. That presentation, which throughout adheres strictly to our general "geometric" outlook on Analysis, aims at keeping as close as possible to the fundamental idea of Calculus, namely the "local" approximation of functions by linear functions. In the classical teaching of Calculus, the idea is immediately obscured by the accidental fact that, on a one-dimensional vector space, there is a one-to-one correspondence between linear forms and numbers, and therefore the derivative at a point is defined as a number instead of a linear form.

This *slavish subservience to the shibboleth of numerical interpretation at any cost* becomes much worse when dealing with functions of several variables...

- J.Dieudonné, Foundations of Modern Analysis, Chapter VIII: Differential Calculus.

What is the derivative of f at x_0?

Look for a good affine approximation of f near x_0

We wish to find an affine function $\mathcal{A}$ that approximates f near x_0.

$$\mathcal{A}(x) = mx + c$$

What is the derivative of f at x_0?

Look for a good affine approximation of f near x_0

We wish to find an affine function $\mathcal{A}$ that approximates f near x_0.

$$\mathcal{A}(x) = mx + c$$

It is natural to impose the condition:

$$\mathcal{A}(x_0) = f(x_0);$$

What is the derivative of f at x_0?

Look for a good affine approximation of f near x_0

We wish to find an affine function $\mathcal{A}$ that approximates f near x_0.

$$\mathcal{A}(x) = mx + c$$

It is natural to impose the condition:

$\mathcal{A}(x_0) = f(x_0)$; That is, $f(x_0) = mx_0 + c$.

What is the derivative of f at x_0?

Look for a good affine approximation of f near x_0

We wish to find an affine function $\mathcal{A}$ that approximates f near x_0.

$$\mathcal{A}(x) = mx + c$$

It is natural to impose the condition:

$$\mathcal{A}(x_0) = f(x_0); \qquad \text{That is, } f(x_0) = mx_0 + c.$$

Then,

$$\mathcal{A}(x) = mx + c = mx + f(x_0) - mx_0$$

What is the derivative of f at x_0?

Look for a good affine approximation of f near x_0

We wish to find an affine function $\mathcal{A}$ that approximates f near x_0.

$$\mathcal{A}(x) = mx + c$$

It is natural to impose the condition:

$$\mathcal{A}(x_0) = f(x_0); \qquad \text{That is, } f(x_0) = mx_0 + c.$$

Then,

$$\begin{aligned}\mathcal{A}(x) = mx + c &= mx + f(x_0) - mx_0 \\ &= m(x - x_0) + f(x_0).\end{aligned}$$

What is the derivative of f at x_0?

Look for a good affine approximation of f near x_0

We wish to find an affine function $\mathcal{A}$ that approximates f near x_0.

$$\mathcal{A}(x) = mx + c$$

Then,

$$\begin{aligned}\mathcal{A}(x) = mx + c &= mx + f(x_0) - mx_0 \\ &= m(x - x_0) + f(x_0).\end{aligned}$$

We need: approximation error $\to 0$ faster than $x \to x_0$.

$$\lim_{x \to x_0} \frac{|f(x) - \mathcal{A}(x)|}{|x - x_0|} = 0$$

What is the derivative of f at x_0?

Look for a good affine approximation of f near x_0

We wish to find an affine function $\mathcal{A}$ that approximates f near x_0.

$$\mathcal{A}(x) = mx + c$$

Then,

$$\mathcal{A}(x) = f(x_0) + m(x - x_0).$$

We need: approximation error $\to 0$ faster than $x \to x_0$.

$$\lim_{x \to x_0} \frac{|f(x) - \mathcal{A}(x)|}{|x - x_0|} = 0$$

Equivalently,

$$\lim_{x \to x_0} \frac{|f(x) - f(x_0)|}{|x - x_0|} = m$$

What is the derivative of f at x_0?

Look for a good affine approximation of f near x_0

We wish to find an affine function $\mathcal{A}$ that approximates f near x_0.

$$\mathcal{A}(x) = mx + c$$

Then,

$$\mathcal{A}(x) = f(x_0) + m(x - x_0).$$

We need: approximation error $\to 0$ faster than $x \to x_0$.

$$\lim_{x \to x_0} \frac{|f(x) - \mathcal{A}(x)|}{|x - x_0|} = 0$$

Equivalently,

$$\lim_{x \to x_0} \frac{|f(x) - f(x_0)|}{|x - x_0|} = m$$

m is denoted $f'(x_0)$ The derivative of f at x_0!

What is the derivative of f at x_0?

Look for a good affine approximation of f near x_0

We wish to find an affine function $\mathcal{A}$ that approximates f near x_0.

$$\mathcal{A}(x) = mx + c$$

Then,

$$\boxed{\mathcal{A}(x) = f(x_0) + f'(x_0)(x - x_0)}$$

We need: approximation error $\to 0$ faster than $x \to x_0$.

$$\lim_{x \to x_0} \frac{|f(x) - \mathcal{A}(x)|}{|x - x_0|} = 0$$

Equivalently,

$$\lim_{x \to x_0} \frac{|f(x) - f(x_0)|}{|x - x_0|} = m$$

m is denoted $f'(x_0)$ The derivative of f at x_0!

Extend to $f : \mathbb{R}^n \to \mathbb{R}$

Look for a good affine approximation of f near x_0

Extend to $f : \mathbb{R}^n \to \mathbb{R}$

Look for a good affine approximation of f near x_0

We wish to find an affine function $\mathcal{A}$ that approximates f near x_0.

$$\mathcal{A}(x) = \mathcal{L}(x) + c$$

Extend to $f : \mathbb{R}^n \to \mathbb{R}$

Look for a good affine approximation of f near x_0

We wish to find an affine function $\mathcal{A}$ that approximates f near x_0.

$$\mathcal{A}(x) = \mathcal{L}(x) + c$$

It is natural to impose the condition:

$$\mathcal{A}(x_0) = f(x_0);$$

Extend to $f : \mathbb{R}^n \to \mathbb{R}$

Look for a good affine approximation of f near x_0

We wish to find an affine function $\mathcal{A}$ that approximates f near x_0.

$$\mathcal{A}(x) = \mathcal{L}(x) + c$$

It is natural to impose the condition:

$\mathcal{A}(x_0) = f(x_0)$; That is, $f(x_0) = \mathcal{L}(x_0) + c$.

Extend to $f : \mathbb{R}^n \to \mathbb{R}$

Look for a good affine approximation of f near x_0

We wish to find an affine function $\mathcal{A}$ that approximates f near x_0.

$$\mathcal{A}(x) = \mathcal{L}(x) + c$$

It is natural to impose the condition:

$$\mathcal{A}(x_0) = f(x_0); \qquad \text{That is, } f(x_0) = \mathcal{L}(x_0) + c.$$

Then,

$$\mathcal{A}(x) = \mathcal{L}(x) + c = \mathcal{L}(x) + f(x_0) - \mathcal{L}(x_0)$$

Extend to $f : \mathbb{R}^n \to \mathbb{R}$

Look for a good affine approximation of f near x_0

We wish to find an affine function $\mathcal{A}$ that approximates f near x_0.

$$\mathcal{A}(x) = \mathcal{L}(x) + c$$

It is natural to impose the condition:

$$\mathcal{A}(x_0) = f(x_0); \qquad \text{That is, } f(x_0) = \mathcal{L}(x_0) + c.$$

Then,

$$\begin{aligned}\mathcal{A}(x) = \mathcal{L}(x) + c &= \mathcal{L}(x) + f(x_0) - \mathcal{L}(x_0) \\ &= \mathcal{L}(x - x_0) + f(x_0).\end{aligned}$$

Extend to $f : \mathbb{R}^n \to \mathbb{R}$

Look for a good affine approximation of f near x_0

We wish to find an affine function $\mathcal{A}$ that approximates f near x_0.

$$\mathcal{A}(x) = \mathcal{L}(x) + c$$

Then,

$$\begin{aligned}\mathcal{A}(x) = \mathcal{L}(x) + c &= \mathcal{L}(x) + f(x_0) - \mathcal{L}(x_0) \\ &= \mathcal{L}(x - x_0) + f(x_0).\end{aligned}$$

We need: approximation error $\to 0$ *faster than* $x \to x_0$.

$$\lim_{x \to x_0} \frac{|f(x) - \mathcal{A}(x)|}{\|x - x_0\|} = 0$$

Extend to $f : \mathbb{R}^n \to \mathbb{R}$

Look for a good affine approximation of f near x_0

We wish to find an affine function $\mathcal{A}$ that approximates f near x_0.

$$\mathcal{A}(x) = \mathcal{L}(x) + c$$

Then,

$$\mathcal{A}(x) = f(x_0) + \mathcal{L}(x - x_0).$$

We need: approximation error $\to 0$ *faster than* $x \to x_0$.

$$\lim_{x \to x_0} \frac{|f(x) - \mathcal{A}(x)|}{\|x - x_0\|} = 0$$

That is, f is said to be differentiable at x_0 if $\exists$ $\mathcal{L}$ linear, such that

$$\lim_{x \to x_0} \frac{|f(x) - \{\mathcal{L}(x - x_0) + f(x_0)\}|}{\|x - x_0\|} = 0.$$

Extend to $f : \mathbb{R}^n \to \mathbb{R}$

Look for a good affine approximation of f near x_0

We wish to find an affine function $\mathcal{A}$ that approximates f near x_0.

$$\mathcal{A}(x) = \mathcal{L}(x) + c$$

Then,

$$\mathcal{A}(x) = f(x_0) + \mathcal{L}(x - x_0).$$

We need: approximation error $\to 0$ *faster than* $x \to x_0$.

$$\lim_{x \to x_0} \frac{|f(x) - \mathcal{A}(x)|}{\|x - x_0\|} = 0$$

That is, f is said to be differentiable at x_0 if $\exists$ $\mathcal{L}$ linear, such that

$$\lim_{x \to x_0} \frac{|f(x) - \{\mathcal{L}(x - x_0) + f(x_0)\}|}{\|x - x_0\|} = 0.$$

Such an $\mathcal{L}$ is called the `Frechét derivative` of f at x_0.

What does $\mathcal{L}$ look like?

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

- Recall that $\mathcal{L}$ is a linear function from $\mathbb{R}^n \to \mathbb{R}$.

What does $\mathcal{L}$ look like?

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

- Recall that $\mathcal{L}$ is a linear function from $\mathbb{R}^n \to \mathbb{R}$.
- We can write $\mathcal{L}(x) = Lx$ where L is a $1 \times n$ matrix

What does $\mathcal{L}$ look like?

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

- Recall that $\mathcal{L}$ is a linear function from $\mathbb{R}^n \to \mathbb{R}$.
- We can write $\mathcal{L}(x) = Lx$ where L is a $1 \times n$ matrix
- Consider the vectors

$$x_j = x_0 + t\mathbf{e_j}; \qquad j = 1, \cdots, n.$$

What does $\mathcal{L}$ look like?

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

- Recall that $\mathcal{L}$ is a linear function from $\mathbb{R}^n \to \mathbb{R}$.
- We can write $\mathcal{L}(x) = Lx$ where L is a $1 \times n$ matrix
- Consider the vectors

$$x_j = x_0 + t\mathbf{e_j}; \qquad j = 1, \cdots, n.$$

- By the definition of Frechét derivative, we have

$$\lim_{t \to 0} \frac{f(x_j) - \{tL\mathbf{e_j} + f(x_0)\}}{t} = 0; \qquad j = 1, \cdots, n.$$

What does $\mathcal{L}$ look like?

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

- Recall that $\mathcal{L}$ is a linear function from $\mathbb{R}^n \to \mathbb{R}$.
- We can write $\mathcal{L}(x) = Lx$ where L is a $1 \times n$ matrix
- Consider the vectors

$$x_j = x_0 + t\mathbf{e_j}; \qquad\qquad j = 1, \cdots, n.$$

- By the definition of Frechét derivative, we have

$$\lim_{t \to 0} \frac{f(x_j) - \{tL\mathbf{e_j} + f(x_0)\}}{t} = 0; \qquad j = 1, \cdots, n.$$

- That is,

$$L\mathbf{e_j} = \lim_{t \to 0} \frac{f(x_j) - f(x_0)}{t}$$

What does $\mathcal{L}$ look like?

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

- Recall that $\mathcal{L}$ is a linear function from $\mathbb{R}^n \to \mathbb{R}$.
- We can write $\mathcal{L}(x) = Lx$ where L is a $1 \times n$ matrix
- Consider the vectors

$$x_j = x_0 + t\mathbf{e_j}; \qquad j = 1, \cdots, n.$$

- By the definition of Frechét derivative, we have

$$\lim_{t \to 0} \frac{f(x_j) - \{tL\mathbf{e_j} + f(x_0)\}}{t} = 0; \qquad j = 1, \cdots, n.$$

- That is,

$$L\mathbf{e_j} = \lim_{t \to 0} \frac{f(x_j) - f(x_0)}{t}$$

j^{th} entry of L

$\frac{\partial f}{\partial x_j}(x_0)$

Gradient & Hessian

The "matrix" L is often denoted $Df(x_0)$ "Jacobian matrix"

$$Df(x_0) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x_0) & \cdots & \frac{\partial f}{\partial x_n}(x_0) \end{bmatrix}$$

Gradient & Hessian

The "matrix" L is often denoted $Df(x_0)$ "Jacobian matrix"

$$Df(x_0) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x_0) & \cdots & \frac{\partial f}{\partial x_n}(x_0) \end{bmatrix}$$

Gradient: $\nabla f(x_0) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x_0) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x_0) \end{bmatrix} = Df(x_0)^\top$

Gradient & Hessian

The "matrix" L is often denoted $Df(x_0)$ "Jacobian matrix"

$$Df(x_0) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x_0) & \cdots & \frac{\partial f}{\partial x_n}(x_0) \end{bmatrix}$$

Gradient: $\nabla f(x_0) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x_0) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x_0) \end{bmatrix} = Df(x_0)^\top$

Hessian: $F = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_2 \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_1} \\ \frac{\partial^2 f}{\partial x_1 \partial x_2} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n} & \frac{\partial^2 f}{\partial x_2 \partial x_n} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix} = D^2 f$

Chain rule

Recall: for $f, g : \mathbb{R} \to \mathbb{R}$, we have

$$\frac{d}{dt} f(g(t)) = f'(g(t))g'(t)$$

Chain rule

Recall: for $f, g : \mathbb{R} \to \mathbb{R}$, we have

$$\frac{d}{dt} f(g(t)) = f'(g(t))g'(t)$$

Similarly, for $g : \mathbb{R} \to \mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}$, let $h(t) = f(g(t))$.

Chain rule

Recall: for $f, g : \mathbb{R} \to \mathbb{R}$, we have

$$\frac{d}{dt} f(g(t)) = f'(g(t))g'(t)$$

Similarly, for $g : \mathbb{R} \to \mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}$, let $h(t) = f(g(t))$. Then,

$$h'(t) = Df(g(t))Dg(t).$$

(refer textbook for proof)

Big-O and little-o

See Page 74 (Sec 5.6) in textbook

$$f = o(g) \quad \text{means} \quad \lim_{x \to 0} \frac{\|f(x)\|}{|g(x)|} = 0.$$

Big-O and little-o

See Page 74 (Sec 5.6) in textbook

$$f = o(g) \quad \text{means} \quad \lim_{x \to 0} \frac{\|f(x)\|}{|g(x)|} = 0.$$

Examples:

$$x^2 = o(x) \qquad \begin{bmatrix} x^3 \\ 2x^2 + 3x^4 \end{bmatrix} = o(x)$$

$$x^3 = o(x^2) \qquad x = o(1).$$

Big-O and little-o

See Page 74 (Sec 5.6) in textbook

$$f = o(g) \quad \text{means} \quad \lim_{x \to 0} \frac{\|f(x)\|}{|g(x)|} = 0.$$

Examples:

$$x^2 = o(x) \qquad \begin{bmatrix} x^3 \\ 2x^2 + 3x^4 \end{bmatrix} = o(x)$$

$$x^3 = o(x^2) \qquad x = o(1).$$

$f = O(g)$ means $\exists\ K > 0, \delta > 0$ such that if $\|x\| < \delta$, then $\dfrac{\|f(x)\|}{|g(x)|} \leq K$.

Big-O and little-o

See Page 74 (Sec 5.6) in textbook

$$f = o(g) \quad \text{means} \quad \lim_{x \to 0} \frac{\|f(x)\|}{|g(x)|} = 0.$$

Examples:

$$x^2 = o(x) \qquad \begin{bmatrix} x^3 \\ 2x^2 + 3x^4 \end{bmatrix} = o(x)$$

$$x^3 = o(x^2) \qquad x = o(1).$$

$$f = O(g) \quad \text{means} \quad \exists\ K > 0, \delta > 0 \text{ such that if } \|x\| < \delta, \text{ then } \frac{\|f(x)\|}{|g(x)|} \leq K.$$

Examples:

$$x = O(x) \qquad \begin{bmatrix} x^3 \\ 2x^2 + 3x^4 \end{bmatrix} = O(x^2)$$

$$\cos x = O(1) \qquad \sin x = O(x).$$

Big-O and little-o

See Page 74 (Sec 5.6) in textbook

$$f = o(g) \quad \text{means} \quad \lim_{x \to 0} \frac{\|f(x)\|}{|g(x)|} = 0.$$

$$f = O(g) \quad \text{means} \quad \exists\ K > 0, \delta > 0 \text{ such that if } \|x\| < \delta, \text{ then } \frac{\|f(x)\|}{|g(x)|} \leq K.$$

- If $f = o(g(x))$, then $f = O(g(x))$ (but the converse is not true)

(proof left as exercise)

Big-O and little-o

See Page 74 (Sec 5.6) in textbook

$$f = o(g) \quad \text{means} \quad \lim_{x \to 0} \frac{\|f(x)\|}{|g(x)|} = 0.$$

$$f = O(g) \quad \text{means} \quad \exists\ K > 0, \delta > 0 \text{ such that if } \|x\| < \delta, \text{then } \frac{\|f(x)\|}{|g(x)|} \leq K.$$

- If $f = o(g(x))$, then $f = O(g(x))$ (but the converse is not true)
- If $f = O(\|x\|^p)$, then $f = o(\|x\|^{p-\varepsilon})$ for any $\varepsilon > 0$.

(proof left as exercise)

Taylor's theorem

The single most important theorem in non-linear optimization

"smooth functions $\approx$ polynomials"

Taylor's theorem

The single most important theorem in non-linear optimization

"smooth functions $\approx$ polynomials"

- For $f : \mathbb{R} \to \mathbb{R}$, $f \in \mathcal{C}^m$ on $[a, b]$, $h = b - a$,

$$f(b) = f(a) + \frac{h}{1!} f'(a) + \frac{h^2}{2!} f''(a) + \cdots + \frac{h^m}{m!} f^{(m)'}(a + \theta h)$$

where $\theta \in (0, 1)$.

Taylor's theorem

The single most important theorem in non-linear optimization

"smooth functions $\approx$ polynomials"

- For $f : \mathbb{R} \to \mathbb{R}$, $f \in \mathcal{C}^m$ on $[a, b]$, $h = b - a$,

$$f(b) = f(a) + \frac{h}{1!} f'(a) + \frac{h^2}{2!} f''(a) + \cdots + \frac{h^m}{m!} f^{(m)'}(a + \theta h)$$

where $\theta \in (0, 1)$.

- For $f : \mathbb{R}^n \to \mathbb{R}$, f is differentiable,

$$f(x) = f(x_0) + Df(x_0)(x - x_0) + \texttt{error term}.$$

$$\boxed{f = o(g) \quad \text{means} \quad \lim_{x \to 0} \frac{\|f(x)\|}{|g(x)|} = 0}$$

Taylor's theorem

The single most important theorem in non-linear optimization

"smooth functions $\approx$ polynomials"

- For $f : \mathbb{R} \to \mathbb{R}$, $f \in \mathcal{C}^m$ on $[a, b]$, $h = b - a$,

$$f(b) = f(a) + \frac{h}{1!} f'(a) + \frac{h^2}{2!} f''(a) + \cdots + \frac{h^m}{m!} f^{(m)'}(a + \theta h)$$

where $\theta \in (0, 1)$.

- For $f : \mathbb{R}^n \to \mathbb{R}$, f is differentiable,

$$f(x) = f(x_0) + Df(x_0)(x - x_0) + o(\|x - x_0\|).$$

$$f = o(g) \quad \text{means} \quad \lim_{x \to 0} \frac{\|f(x)\|}{|g(x)|} = 0$$

Taylor's theorem

The single most important theorem in non-linear optimization

"smooth functions $\approx$ polynomials"

- For $f : \mathbb{R} \to \mathbb{R}$, $f \in \mathcal{C}^m$ on $[a, b]$, $h = b - a$,

$$f(b) = f(a) + \frac{h}{1!} f'(a) + \frac{h^2}{2!} f''(a) + \cdots + \frac{h^m}{m!} f^{(m)'}(a + \theta h)$$

where $\theta \in (0, 1)$.

- For $f : \mathbb{R}^n \to \mathbb{R}$, $f \in \mathcal{C}^2$,

$$f(x) = f(x_0) + \frac{1}{1!} Df(x_0)(x - x_0) + \frac{1}{2!} (x - x_0)^\top D^2 f(x_0)(x - x_0) + \texttt{error term}.$$

$$\boxed{f = o(g) \quad \text{means} \quad \lim_{x \to 0} \frac{\|f(x)\|}{|g(x)|} = 0}$$

Taylor's theorem

The single most important theorem in non-linear optimization

"smooth functions $\approx$ polynomials"

- For $f : \mathbb{R} \to \mathbb{R}$, $f \in \mathcal{C}^m$ on $[a, b]$, $h = b - a$,

$$f(b) = f(a) + \frac{h}{1!} f'(a) + \frac{h^2}{2!} f''(a) + \cdots + \frac{h^m}{m!} f^{(m)'}(a + \theta h)$$

where $\theta \in (0, 1)$.

- For $f : \mathbb{R}^n \to \mathbb{R}$, $f \in \mathcal{C}^2$,

$$f(x) = f(x_0) + \frac{1}{1!} Df(x_0)(x - x_0) + \frac{1}{2!} (x - x_0)^\top D^2 f(x_0)(x - x_0) + o(\|x - x_0\|^2).$$

$$f = o(g) \quad \text{means} \quad \lim_{x \to 0} \frac{\|f(x)\|}{|g(x)|} = 0$$

Taylor's theorem

The single most important theorem in non-linear optimization

"smooth functions $\approx$ polynomials"

- For $f : \mathbb{R} \to \mathbb{R}$, $f \in \mathcal{C}^m$ on $[a, b]$, $h = b - a$,

$$f(b) = f(a) + \frac{h}{1!} f'(a) + \frac{h^2}{2!} f''(a) + \cdots + \frac{h^m}{m!} f^{(m)'}(a + \theta h)$$

where $\theta \in (0, 1)$.

- For $f : \mathbb{R}^n \to \mathbb{R}$, $f \in \mathcal{C}^3$,

$$f(x) = f(x_0) + \frac{1}{1!} Df(x_0)(x - x_0) + \frac{1}{2!}(x - x_0)^\top D^2 f(x_0)(x - x_0) + O(\|x - x_0\|^3).$$

$f = O(g)$ means $\exists\ K > 0, \delta > 0$ such that if $\|x\| < \delta$, then $\dfrac{\|f(x)\|}{|g(x)|} \leq K$

Feasible direction

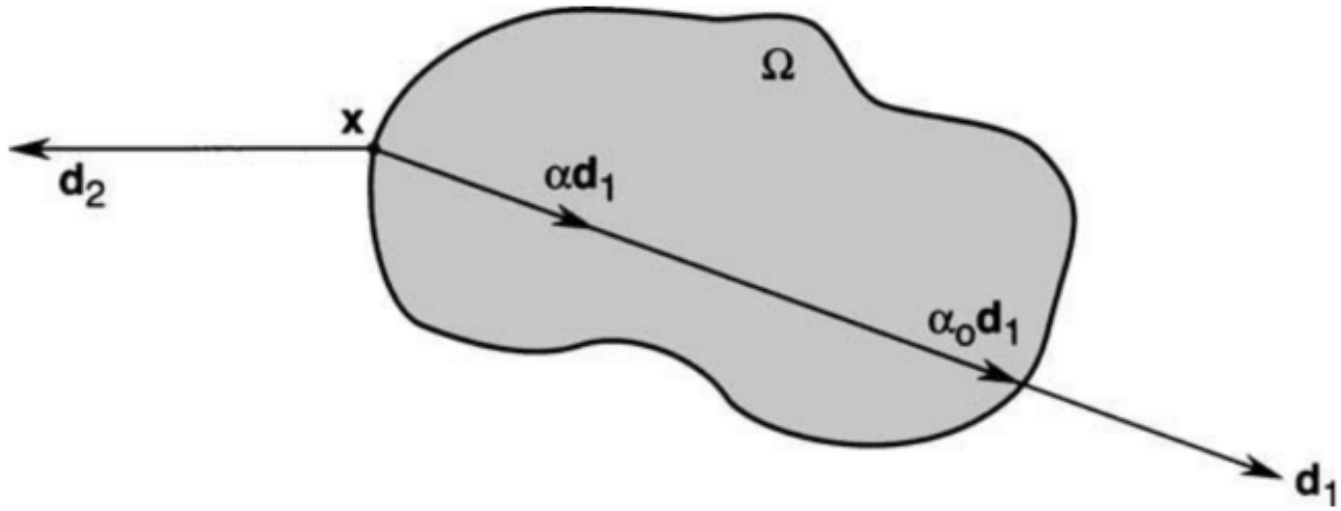
Constraint set Ω

– A vector $d \in \mathbb{R}^n$, $d \neq 0$, is a feasible direction at $x \in \Omega$ if $\exists\ \alpha_0 > 0$ such that $x + \alpha d \in \Omega$ for all $\alpha \in [0, \alpha_0]$.

Feasible direction

Constraint set Ω

– A vector $d \in \mathbb{R}^n$, $d \neq 0$, is a feasible direction at $x \in \Omega$ if $\exists\ \alpha_0 > 0$ such that $x + \alpha d \in \Omega$ for all $\alpha \in [0, \alpha_0]$.



d_1 is feasible, d_2 is not.

Directional derivative

More generally, Gateaux derivative

· Let $f : \mathbb{R}^n \to \mathbb{R}$, d a feasible direction at $x \in \Omega$. The directional derivative of f in the direction d is given by

Directional derivative

More generally, Gateaux derivative

· Let $f : \mathbb{R}^n \to \mathbb{R}$, d a feasible direction at $x \in \Omega$. The directional derivative of f in the direction d is given by

$$\frac{\partial f}{\partial d}(x) = \lim_{\alpha \to 0} \frac{f(x + \alpha d) - f(x)}{\alpha}.$$

Directional derivative

More generally, Gateaux derivative

· Let $f : \mathbb{R}^n \to \mathbb{R}$, d a feasible direction at $x \in \Omega$. The directional derivative of f in the direction d is given by

$$\frac{\partial f}{\partial d}(x) = \lim_{\alpha \to 0} \frac{f(x + \alpha d) - f(x)}{\alpha}.$$

· For a given x and d,

$$\frac{\partial f}{\partial d}(x) = \frac{d}{d\alpha} f(x + \alpha d)\Big|_{\alpha=0}$$

Directional derivative

More generally, Gateaux derivative

· Let $f : \mathbb{R}^n \to \mathbb{R}$, d a feasible direction at $x \in \Omega$. The directional derivative of f in the direction d is given by

$$\frac{\partial f}{\partial d}(x) = \lim_{\alpha \to 0} \frac{f(x + \alpha d) - f(x)}{\alpha}.$$

· For a given x and d,

$$\begin{aligned} \frac{\partial f}{\partial d}(x) &= \frac{d}{d\alpha} f(x + \alpha d)\Big|_{\alpha=0} \\ &= \nabla f(x)^\top d \qquad \text{(chain rule)} \end{aligned}$$

Directional derivative

More generally, Gateaux derivative

· Let $f : \mathbb{R}^n \to \mathbb{R}$, d a feasible direction at $x \in \Omega$. The directional derivative of f in the direction d is given by

$$\frac{\partial f}{\partial d}(x) = \lim_{\alpha \to 0} \frac{f(x + \alpha d) - f(x)}{\alpha}.$$

· For a given x and d,

$$\begin{aligned} \frac{\partial f}{\partial d}(x) &= \frac{d}{d\alpha} f(x + \alpha d)\Big|_{\alpha=0} \\ &= \nabla f(x)^\top d \qquad \text{(chain rule)} \end{aligned}$$

· If $\|d\| = 1$ (d is a unit vector), then

$\frac{\partial f}{\partial d}(x)$ is the rate of increase of f at the point x in the direction d.

Directional derivative

More generally, Gateaux derivative

· Let $f : \mathbb{R}^n \to \mathbb{R}$, d a feasible direction at $x \in \Omega$. The directional derivative of f in the direction d is given by

$$\frac{\partial f}{\partial d}(x) = \lim_{\alpha \to 0} \frac{f(x + \alpha d) - f(x)}{\alpha}.$$

· For a given x and d,

$$\boxed{\frac{\partial f}{\partial d}(x) = \nabla f(x)^\top d}$$

$$\begin{aligned} \frac{\partial f}{\partial d}(x) &= \frac{d}{d\alpha} f(x + \alpha d)\Big|_{\alpha=0} \\ &= \nabla f(x)^\top d \qquad \text{(chain rule)} \end{aligned}$$

· If $\|d\| = 1$ (d is a unit vector), then

$\frac{\partial f}{\partial d}(x)$ is the rate of increase of f at the point x in the direction d.

Directional derivative

More generally, Gateaux derivative

· Let $f : \mathbb{R}^n \to \mathbb{R}$, d a feasible direction at $x \in \Omega$. The directional derivative of f in the direction d is given by

$$\frac{\partial f}{\partial d}(x) = \lim_{\alpha \to 0} \frac{f(x + \alpha d) - f(x)}{\alpha}.$$

· For a given x and d,

$$\boxed{\frac{\partial f}{\partial d}(x) = \nabla f(x)^\top d}$$

$$\begin{aligned} \frac{\partial f}{\partial d}(x) &= \frac{d}{d\alpha} f(x + \alpha d)\Big|_{\alpha=0} \\ &= \nabla f(x)^\top d \qquad \text{(chain rule)} \end{aligned}$$

· Dir. der. of f at x in the direction $d = \Big\langle$Gradient at x, Direction $d\Big\rangle$

An example

$\frac{\partial f}{\partial d}(x) = \nabla f(x)^\top d$

- Let $f : \mathbb{R}^3 \to \mathbb{R}$ be defined as

$$f(x) = x_1 x_2 x_3.$$

Calculate the directional derivative of f at

$$\begin{bmatrix} 1 & 1 & 1 \end{bmatrix}^\top$$

in the direction

$$d = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{\sqrt{2}} \end{bmatrix}^\top.$$

(solution on board)

First Order Necessary Condition (FONC)

If x^* is a local minimizer, then the condition holds

Theorem – Let $\Omega \subseteq \mathbb{R}^n$ and $f : \Omega \to \mathbb{R}$, $f \in \mathcal{C}^1$.

If x^* is a local minimizer of f over Ω, then

for any feasible direction d at x^*, we have

$$d^\top \nabla f(x^*) \geq 0.$$

(proof on board/see Theorem 6.1 in textbook)

First Order Necessary Condition (FONC)

Interior case

$x \in S$ is an *interior point* if S contains a "neighborhood" of x.

First Order Necessary Condition (FONC)

Interior case

$x \in S$ is an *interior point* if S contains a "neighborhood" of x.

Corollary – Let $\Omega \subseteq \mathbb{R}^n$ and $f : \Omega \to \mathbb{R}$, $f \in \mathcal{C}^1$.

If x^* is a local minimizer of f over Ω and

if x^* is an interior point of Ω, then

$$\nabla f(x^*) = 0.$$

(Corollary 6.1 in textbook)

Second Order Necessary Condition (SONC)

Theorem 6.2 in textbook

Theorem – Let $\Omega \subseteq \mathbb{R}^n$ and $f : \Omega \to \mathbb{R}$, $f \in \mathcal{C}^2$.

Let x^* be a local minimizer of f over Ω, and

and d a feasible direction at x^*. If $d^\top \nabla f(x^*) = 0$, then

$$d^\top F(x^*) d \geq 0; \qquad F \text{ Hessian of } f.$$

Second Order Necessary Condition (SONC)

Theorem 6.2 in textbook

Theorem – Let $\Omega \subseteq \mathbb{R}^n$ and $f : \Omega \to \mathbb{R}$, $f \in \mathcal{C}^2$.

Let x^* be a local minimizer of f over Ω, and

and d a feasible direction at x^*. If $d^\top \nabla f(x^*) = 0$, then

$$d^\top F(x^*) d \geq 0; \qquad F \text{ Hessian of } f.$$

Proof sketch – Suppose for a feasible d at x^*, $d^\top \nabla f(x^*) = 0$ and $d^\top F(x^*) d < 0$.

Second Order Necessary Condition (SONC)

Theorem 6.2 in textbook

Theorem – Let $\Omega \subseteq \mathbb{R}^n$ and $f : \Omega \to \mathbb{R}$, $f \in \mathcal{C}^2$.

Let x^* be a local minimizer of f over Ω, and

and d a feasible direction at x^*. If $d^\top \nabla f(x^*) = 0$, then

$$d^\top F(x^*) d \geq 0; \qquad F \text{ Hessian of } f.$$

Proof sketch – Suppose for a feasible d at x^*, $d^\top \nabla f(x^*) = 0$ and $d^\top F(x^*) d < 0$. Let $\phi(\alpha) := f(x^* + \alpha d)$.

Second Order Necessary Condition (SONC)

Theorem 6.2 in textbook

Theorem – Let $\Omega \subseteq \mathbb{R}^n$ and $f : \Omega \to \mathbb{R}$, $f \in \mathcal{C}^2$.

Let x^* be a local minimizer of f over Ω, and

and d a feasible direction at x^*. If $d^\top \nabla f(x^*) = 0$, then

$$d^\top F(x^*) d \geq 0; \qquad F \text{ Hessian of } f.$$

Proof sketch – Suppose for a feasible d at x^*, $d^\top \nabla f(x^*) = 0$ and $d^\top F(x^*) d < 0$. Let $\phi(\alpha) := f(x^* + \alpha d)$. We know that

$$\phi''(\alpha) = D\left[d^\top \nabla f(x^* + \alpha d)\right] = d^\top F(x^* + \alpha d) d.$$

Second Order Necessary Condition (SONC)

Theorem 6.2 in textbook

Theorem – Let $\Omega \subseteq \mathbb{R}^n$ and $f : \Omega \to \mathbb{R}$, $f \in \mathcal{C}^2$.

Let x^* be a local minimizer of f over Ω, and

and d a feasible direction at x^*. If $d^\top \nabla f(x^*) = 0$, then

$$d^\top F(x^*) d \geq 0; \qquad F \text{ Hessian of } f.$$

Proof sketch – Suppose for a feasible d at x^*, $d^\top \nabla f(x^*) = 0$ and $d^\top F(x^*) d < 0$. Let $\phi(\alpha) := f(x^* + \alpha d)$. We know that

$$\phi''(\alpha) = D\left[d^\top \nabla f(x^* + \alpha d)\right] = d^\top F(x^* + \alpha d) d.$$

By Taylor's theorem,

$$\phi(\alpha) - \phi(0) = \phi''(0)\frac{\alpha^2}{2} + o(\alpha^2) < 0$$

for α small enough (since $\phi''(0) = d^\top F(x^*) d < 0$).

Second Order Necessary Condition (SONC)

Theorem 6.2 in textbook

Theorem – Let $\Omega \subseteq \mathbb{R}^n$ and $f : \Omega \to \mathbb{R}$, $f \in \mathcal{C}^2$.

Let x^* be a local minimizer of f over Ω, and

and d a feasible direction at x^*. If $d^\top \nabla f(x^*) = 0$, then

$$d^\top F(x^*) d \geq 0; \qquad F \text{ Hessian of } f.$$

Proof sketch – Suppose for a feasible d at x^*, $d^\top \nabla f(x^*) = 0$ and $d^\top F(x^*) d < 0$. Let $\phi(\alpha) := f(x^* + \alpha d)$. We know that

$$\phi''(\alpha) = D\left[d^\top \nabla f(x^* + \alpha d)\right] = d^\top F(x^* + \alpha d) d.$$

By Taylor's theorem,

$$\phi(\alpha) - \phi(0) = \phi''(0)\frac{\alpha^2}{2} + o(\alpha^2) < 0$$

for α small enough (since $\phi''(0) = d^\top F(x^*) d < 0$).That is,

$$f(x^* + \alpha d) < f(x^*),$$

contradicting the assumption that x^* is a minimizer. (see proof in textbook) □

Second Order Necessary Condition (SONC)

Interior case: Corollary 6.2 in textbook

Corollary – Let x^* be an interior point of $\Omega \subseteq \mathbb{R}^n$. If x^* is a local minimizer of $f : \Omega \to \mathbb{R}$, $f \in \mathcal{C}^2$, then

$$\nabla f(x^*) = 0$$

and $F(x^*)$ is positive semidefinite.

(follows from the SONC Theorem and the corollary to FONC)

Second Order Sufficient Condition (SOSC)

If the condition holds, then x^* is a local minimizer

Let x^* be an interior point of Ω and $f : \Omega \to \mathbb{R}$, $f \in \mathcal{C}^2$. Suppose that

1. $\nabla f(x^*) = 0$
2. $F(x^*)$ is positive definite.

Then, x^* is a strict local minimizer of f.

Second Order Sufficient Condition (SOSC)

If the condition holds, then x^* is a local minimizer

Let x^* be an interior point of Ω and $f : \Omega \to \mathbb{R}$, $f \in \mathcal{C}^2$. Suppose that

1. $\nabla f(x^*) = 0$
2. $F(x^*)$ is positive definite.

Then, x^* is a strict local minimizer of f.

Clairaut's theorem – If $f \in \mathcal{C}^2$, then F is symmetric.

Second Order Sufficient Condition (SOSC)

If the condition holds, then x^* is a local minimizer

Let x^* be an interior point of Ω and $f : \Omega \to \mathbb{R}$, $f \in \mathcal{C}^2$. Suppose that

1. $\nabla f(x^*) = 0$
2. $F(x^*)$ is positive definite.

Then, x^* is a strict local minimizer of f.

Clairaut's theorem – If $f \in \mathcal{C}^2$, then F is symmetric.

Rayleigh's inequality – see lecture slides on Linear Algebra review

Second Order Sufficient Condition (SOSC)

If the condition holds, then x^* is a local minimizer

Let x^* be an interior point of Ω and $f : \Omega \to \mathbb{R}$, $f \in \mathcal{C}^2$. Suppose that

1. $\nabla f(x^*) = 0$
2. $F(x^*)$ is positive definite.

Then, x^* is a strict local minimizer of f.

Clairaut's theorem – If $f \in \mathcal{C}^2$, then F is symmetric.

Rayleigh's inequality – see lecture slides on Linear Algebra review

(see Theorem 6.3 in textbook for proof)