NONLINEAR OPTIMIZATION

JAYANT RAJGOPAL

Department of Industrial Engineering
University of Pittsburgh
Pittsburgh, PA 15261

August 2012

These notes draw significantly upon material developed by Professor **Dennis L. Bricker**, Department of Industrial Engineering, The University of Iowa and Professor **Ashok D. Belegundu**, Department of Mechanical Engineering, The Pennsylvania State University.

The contributions of these two individuals are gratefully acknowledged.

Mathematics Review

Vector $\boldsymbol{x} \in \mathbb{R}^n = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = [x_1 \;\; x_2 \;\; \ldots \;\; x_n]^{\mathrm{T}}$.

Special vectors: $\boldsymbol{1} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$; $\boldsymbol{0} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$; $\boldsymbol{e_i} = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \rightarrow i^{th}$ *position*

$\mathbb{R}^n_+ = \{ \boldsymbol{x} \in \mathbb{R}^n \mid x_i \geq 0,\ i=1,2,\ldots,n \}$; $\mathbb{R}^n_{++} = \{ \boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{x}_i > 0,\ i=1,2,\ldots,n \}$;

Norm of a vector ($\|\cdot\|$) is a real-valued function $\|\boldsymbol{x}\| = \mathbb{R}^n \rightarrow \mathbb{R}^n_+$ that is a measure of "length" or distance. It satisfies the following properties:

(i) $\|\boldsymbol{x}\| \geq 0 \forall \boldsymbol{x}$; (ii) $\|\boldsymbol{x}\| = 0 \Leftrightarrow \boldsymbol{x} = \boldsymbol{0}$; (iii) $\|\delta \boldsymbol{x}\| = |\delta| * \|\boldsymbol{x}\|$ for all real δ;

(iv) $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$ $\forall$ $\boldsymbol{x}$ and $\boldsymbol{y}$.

Various norms:

l_1-norm: $\|\boldsymbol{x}\|_1 = \sum_{i=1}^{n} |x_i|$

l_2-norm: $\|\boldsymbol{x}\|_2 = \left(\sum_{i=1}^{n} x_i^2 \right)^{1/2}$

l_p-norm: $\|\boldsymbol{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}$

l_∞-norm: $\|\boldsymbol{x}\|_\infty = Max_i \{ |x_i| \}$

Matrix Norms: Suppose $\boldsymbol{A}$ is some $m \times n$ matrix. Then the matrix norm $\|\boldsymbol{A}\|$ is said to be *induced* by the corresponding vector norm $\|\boldsymbol{x}\|$, and is defined as

$$\|\boldsymbol{A}\| = \underset{\|\boldsymbol{x}\|=1}{Max} \|\boldsymbol{A}\boldsymbol{x}\|.$$

Thus

- $\|A\|_1 = \underset{\|\boldsymbol{x}\|_1=1}{Max} \|\boldsymbol{A}\boldsymbol{x}\|_1 = \{Max \sum_i |(\sum_j a_{ij}x_j)|, \text{ st } \{\sum_j |x_j|\}=1\} \} = \underset{1 \le j \le n}{Max} \sum_{i=1}^{m} |a_{ij}|$

For example, if $\boldsymbol{A}=\begin{bmatrix} 1 & 3 & -2 & 4 \\ -5 & 7 & 9 & -3 \\ 2 & -1 & 6 & 8 \end{bmatrix}$ then $\boldsymbol{A}\boldsymbol{x}=\begin{bmatrix} x_1+3x_2-2x_3+4x_4 \\ -5x_1+7x_2+9x_3-3x_4 \\ 2x_1-x_2+6x_3+8x_4 \end{bmatrix}$ and so

$\|A\|_1 = \{Max \{|x_1+3x_2+2x_3+4x_4|+|-5x_1+7x_2+9x_3-3x_4|+|2x_1-x_2+6x_3+8x_4|\},$

st $\{|x_1|+|x_2|+|x_3|+|x_4|\}=1\} \}$ = Max(1+5+2, 3+7+1, 2+9+6, 4+3+8)

= **17** (…with $x_1=x_2=x_4=0$ and $x_3=1$ or -1 at the optimum)

Similarly,

- $\|A\|_2 = \underset{\|\boldsymbol{x}\|_2=1}{Max} \|\boldsymbol{A}\boldsymbol{x}\|_2 = \{Max \left\{ \sum_{i=1}^{m} \left(\sum_{j=1}^{n} a_{ij}x_j \right)^2 \right\}^{1/2}, \text{ st } \left\{ \sum_{j=1}^{n} x_j^2 \right\}^{1/2} = 1\} \} =$

  SQRT$\{\lambda_{max}(\boldsymbol{A}\text{T}\boldsymbol{A})\}$, where $\lambda_{max}(\boldsymbol{A}^{\text{T}}\boldsymbol{A})$ is the largest eigenvalue of the matrix $\boldsymbol{A}^{\text{T}}\boldsymbol{A}$ (may be shown…)

This matrix norm is also called the *spectral norm.* For our example, the eigenvalues of $\boldsymbol{A}^{\text{T}}\boldsymbol{A}$ are {0, 23.48, 108.05, 167.47} (***check using MATLAB or other package***….), so that $\|A\|_2 = \sqrt{167.47} =$ **12.94.**

Finally,

- $\|A\|_\infty = \underset{\|x\|_\infty=1}{Max} \|Ax\|_\infty = \left\{ Max\left(\underset{1\le i\le m}{Max} \left| \sum_{j=1}^{n} a_{ij} x_j \right| \right), \text{ st} \left(\underset{1\le j\le n}{max} | x_j | = 1 \right) \right\}$

$= \underset{1\le i\le m}{Max} \sum_{j=1}^{n} |a_{ij}|$

For our example,

$\|A\|_\infty = \{ Max [\{\text{Max } (|x_1+3x_2+2x_3+4x_4|, |-5x_1+7x_2+9x_3-3x_4|, |2x_1-x_2+6x_3+8x_4|\}],$

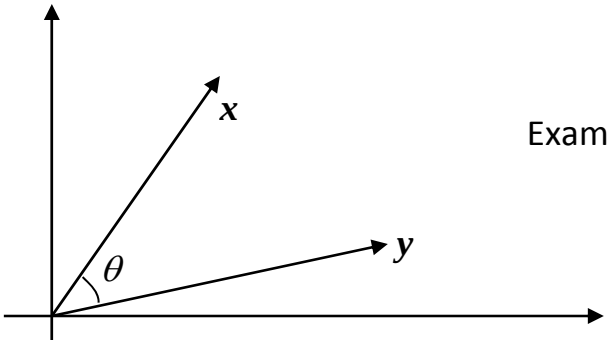
$\text{st Max}(|x_1|,|x_2|,|x_3|,|x_4|)=1 \}$

= Max(1+3+2+4, 5+7+9+3, 2+1+6+8) = Max (10, 24, 17)

= **24** (…with $x_1=x_4 = -1$ and $x_2= x_3 = 1$ at the optimum…)

Suppose $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{y} \in \mathbb{R}^n$; and $\boldsymbol{x, y \neq 0}$. Then $\boldsymbol{x^T y} = \langle \boldsymbol{x}, \boldsymbol{y} \rangle = \sum_i x_i y_i = \|\boldsymbol{x}\| \|\boldsymbol{y}\| \cos\theta$,

where θ is the angle between the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$.



Example in $\mathbb{R}^2$

Note that if

$\boldsymbol{x^T y} = 0$, then $\cos\theta = 0 \Rightarrow \theta = 90°$ (orthogonal vectors)

$\boldsymbol{x^T y} > 0$ then $\cos\theta > 0 \Rightarrow \theta < 90°$ (acute angle between vectors)

$\boldsymbol{x^T y} < 0$ then $\cos\theta < 0 \Rightarrow \theta > 90°$ (obtuse angle between vectors)

Also, $\boldsymbol{x}^T\boldsymbol{x} = \sum_i x_i^2 = \|x\|\|x\|\cos 0^\circ = \|x\|^2$ (as defined by normal vector multiplication)

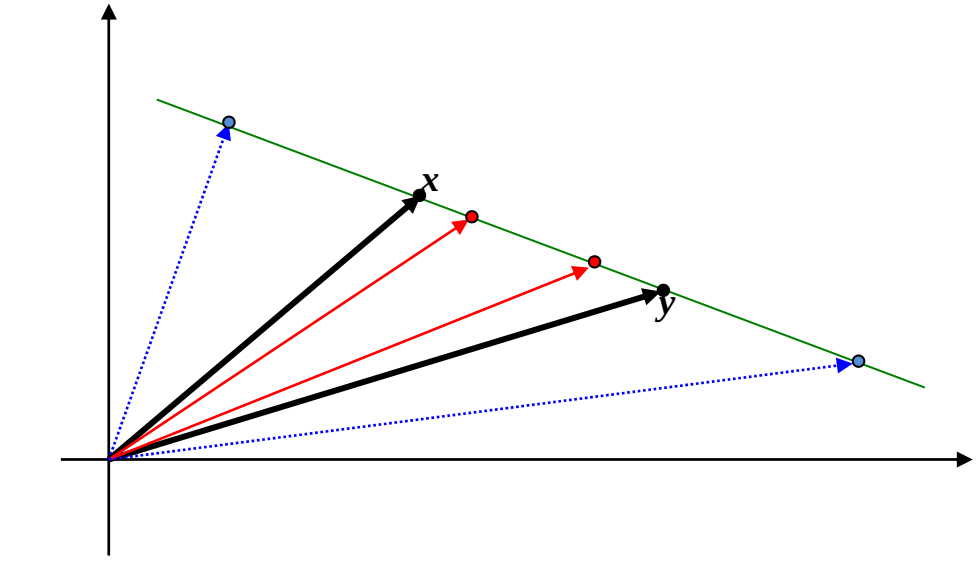
Note that a vector also has associated with it a direction (i.e., the vector $[6\ \ 6]^T$ and the vector $[1\ \ 1]^T$ have the same direction). This is easily seen by looking at their product as $[6\quad 6]\cdot[1\quad 1]^T = 12 = \|6\quad 6\|\|1\quad 1\|\cos\theta$ (by the definition above), so that $\cos\theta = 12/(\sqrt{72}\sqrt{2}) = 1$, i.e, $\theta = 0°$.

<u>Combinations of vectors</u>: Given $\boldsymbol{x}^1, \boldsymbol{x}^2 \ldots, \boldsymbol{x}^n$

Linear combination = $\sum_j \lambda_j \boldsymbol{x}^j$

Affine combination = $\sum_j \lambda_j \boldsymbol{x}^j$, where $\sum_j \lambda_j = 1$

Convex combination = $\sum_j \lambda_j \boldsymbol{x}^j$, where $\sum_j \lambda_j = 1$ and all $\lambda_j \geq 0$



Affine combination of $\boldsymbol{x}$ and $\boldsymbol{y}$

Convex (as well as affine) combination of $\boldsymbol{x}$ and $\boldsymbol{y}$

Note that every vector in $\mathbb{R}^2$ is a linear combination of $\boldsymbol{x}$ and $\boldsymbol{y}$.

Linear Independence: The vectors $\boldsymbol{x^1}, \boldsymbol{x^2}\ldots, \boldsymbol{x^n}$ are said to be linearly independent if $\sum_j \lambda_j \boldsymbol{x^j} = 0$ implies that $\lambda_j=0$ for all j.

Span: The vectors $\boldsymbol{x^1}, \boldsymbol{x^2}\ldots, \boldsymbol{x^k}$ are said to span the space $L\subset \mathbb{R}^n$ if *any* $\boldsymbol{x}\in L$ can be written as a linear combination of $\boldsymbol{x^1}, \boldsymbol{x^2}\ldots, \boldsymbol{x^k}$.

Basis: The vectors $\boldsymbol{x^1}, \boldsymbol{x^2}\ldots, \boldsymbol{x^k}$ constitute a basis for $\mathbb{R}^n$ if they span $\mathbb{R}^n$, but if one of them is dropped then the remaining vectors do not span span $\mathbb{R}^n$.

MATRICES

$\boldsymbol{A}_{m\times n} = \{a_{ij}\}$ where i=1,2,…,m is the row index and j=1,2,…,n is the column index. We also denote $\boldsymbol{A}_{m\times n} = [\boldsymbol{a^1}\ \ \boldsymbol{a^2}\ \ \ldots\ \boldsymbol{a^n}]$ where each $\boldsymbol{a^j}\in \mathbb{R}^m$.

Special matrices: $\boldsymbol{I_n} \equiv n\times n$ Identity Matrix $[\mathbf{0}]\equiv$ the zero matrix

$\boldsymbol{A^T}$ =transpose of $\boldsymbol{A}$ where $(\boldsymbol{A^T})_{ij} = \boldsymbol{A_{ji}}$. If $\boldsymbol{A^T}=\boldsymbol{A}$ then $\boldsymbol{A}$ is said to be symmetric.

Inverse of a matrix: Given $\boldsymbol{A_{n\times n}}$, if $\exists$ $n\times n$ matrix $\boldsymbol{A^{-1}}$ ∋ $\boldsymbol{AA^{-1}= A^{-1}A = I}$, then $\boldsymbol{A^{-1}}$ is called the inverse of $\boldsymbol{A}$; otherwise $\boldsymbol{A}$ is said to be **singular**.
$\boldsymbol{A}$ is said to have full rank if the maximum number of linearly independent columns or rows (= rank of $\boldsymbol{A}$) = Min$\{n,m\}$.

SETS

A set **S** is a collection of elements. e.g.,

S = {1,2,3,4} finite, countable **S** = {1,2,3,…} infinite but countable

S = $\{x\in\mathbb{R}|\ 0\le x\le 1\}$ infinite and not countable

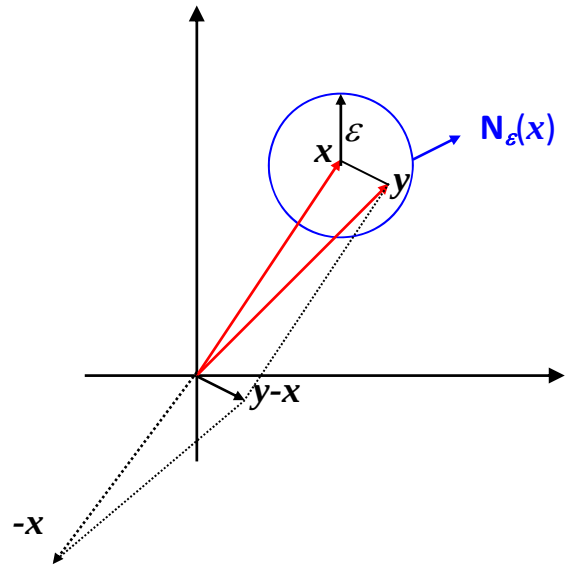
In nonlinear optimization problems, typically we deal with sets of the form

$\mathbf{S} = \{\boldsymbol{x} \in \mathbb{R}^n \mid$ "some set of conditions"$\}$;

e.g., $\mathbf{S} = \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{a}^T\boldsymbol{x} \geq 0\}$, i.e. the set of vectors in $\mathbb{R}^n$ making an angle of 90° or less with the vector $\boldsymbol{a}$. The **empty set** Φ is the set with no elements in it.

<u>Neighborhood</u>: Given a point $\boldsymbol{x} \in \mathbb{R}^n$**,** its ε-neighborhood (where $\varepsilon \in \mathbb{R}_{++}$) is defined as the set $\mathbf{N}_\varepsilon(\boldsymbol{x}) = \{\boldsymbol{y} \in \mathbb{R}^n \mid \|\boldsymbol{y} - \boldsymbol{x}\| < \varepsilon\}$. The set $\mathbf{N}_\varepsilon(\boldsymbol{x})$ is also called a **ball** (or more precisely, an open *ball*).

Note that for $\boldsymbol{y} \in \mathbf{N}_\varepsilon(\boldsymbol{x})$, $\|\boldsymbol{y} - \boldsymbol{x}\| < \varepsilon$



We may also define an ε-neighborhood for an entire set **S** in an analogous fashion: $\mathbf{N}_\varepsilon(\mathbf{S}) = \left\{\boldsymbol{y} \in \mathbb{R}^n \mid \underset{\boldsymbol{x} \in \mathrm{S}}{Min} \|\boldsymbol{y} - \boldsymbol{x}\| < \varepsilon\right\}$.

"LOCAL" ≈ Points belonging to some ε-neighborhood

"GLOBAL" ≈ If ε is "extremely large," i.e., "everywhere."

Sets could be (i) open, (ii) closed, (iii) neither open nor closed or, (iv) both open and closed.

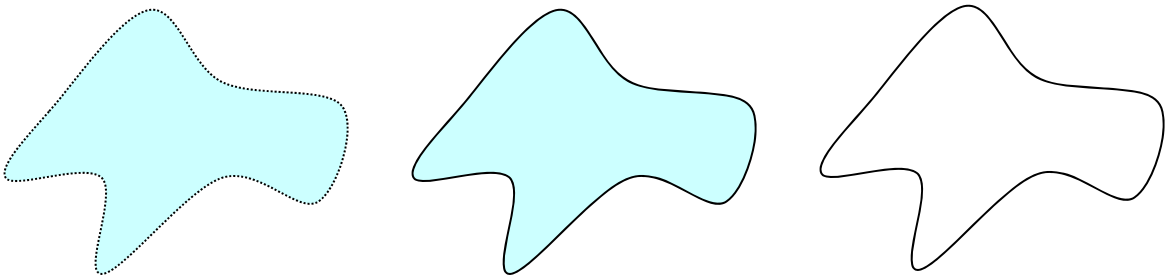
Suppose **S** is a subset of $\mathbb{R}^n$.

A set **S** is said to be **<u>open</u>** if $\forall\ \boldsymbol{x}\in\mathbf{S}, \exists\ \varepsilon>0\ \ni\ \mathbf{N}_{\varepsilon}(\boldsymbol{x})\in\mathbf{S}$.

The **<u>interior</u>** of a set **S** is defined as $int(\mathbf{S}) = \{\boldsymbol{x}\in\mathbf{S} | \exists\ \varepsilon>0\ \ni\ \mathbf{N}_{\varepsilon}(\boldsymbol{x})\in\mathbf{S}\}$, i.e., it is the collection of points in **S** that have some sufficiently small ε-neighborhood that is contained entirely within **S**. It should be clear that a set **S** is open if $\mathbf{S} = int(\mathbf{S})$.

The **<u>closure</u>** of a set **S** is defined as $cl(\mathbf{S}) = \{\boldsymbol{x}\in \mathbb{R}^n \mid \forall\ \varepsilon>0, \mathbf{S}\cap\mathbf{N}_{\varepsilon}(\boldsymbol{x})\neq\Phi\}$. In other words it is the set of all points in $\mathbb{R}^n$ for which an arbitrarily small ε-neighborhood can still not be separated from **S** (obviously the closure also includes the interior of **S**), i.e., it is the set of all points that are in, or can be made "arbitrarily close to" the set **S**. It should be clear that a set **S** is closed if $\mathbf{S} = cl(\mathbf{S})$.

The set of **<u>boundary points</u>** of a set **S** (denoted by $\partial\mathbf{S}$) is defined as the set of all points $\boldsymbol{x}\in \mathbb{R}^n$ such that for every $\varepsilon>0$, $\mathbf{N}_{\varepsilon}(\boldsymbol{x})$ has at least one point that belongs to **S** and one that does not belong to **S**. Note that $\partial\mathbf{S} = \{\mathbf{S}\backslash int(\mathbf{S})\}$. (check)



$int(\mathbf{S})$ $cl(\mathbf{S})$ $\partial\mathbf{S}$

Examples:

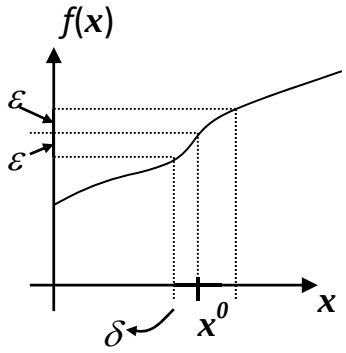
- $\mathbf{S} = \{x \in \mathbb{R} \mid a \le x \le b\}$ is the closed set $[a,b]$. Note that $int(\mathbf{S}) = \{x \in \mathbb{R} \mid a < x < b\}$, $cl(\mathbf{S}) = \mathbf{S}$ and $\partial\mathbf{S} = \{a,b\}$.
- $\mathbf{S} = \{x \in \mathbb{R} \mid a < x < b\}$ is the open set (a,b). Note that $int(\mathbf{S}) = \mathbf{S}$, while $cl(\mathbf{S}) = \{x \in \mathbb{R} \mid a \le x \le b\}$ and $\partial\mathbf{S} = \Phi$.
- $\mathbf{S} = \{x \in \mathbb{R} \mid a < x \le b\}$ is the set (a,b] that is neither open nor closed. Note that $int(\mathbf{S}) = \{x \in \mathbb{R} \mid a < x < b\}$, $cl(\mathbf{S}) = \{x \in \mathbb{R} \mid a \le x \le b\}$ and $\partial\mathbf{S} = \{b\}$. (check)

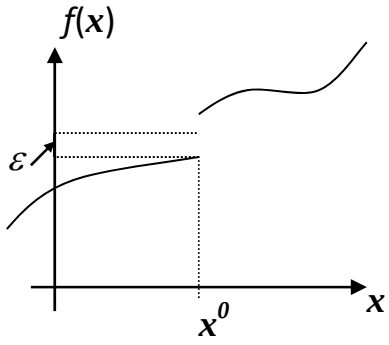
A set **S** is said to be **<u>bounded</u>** if it can be contained within a ball of finite radius, i.e., for all $\boldsymbol{x} \in \mathbf{S}$, $\|x\| < M$ for some $M > 0$. Note that all three sets above are bounded, whereas the set $\mathbf{S} = \{x \in \mathbb{R} \mid x \ge 0\}$ is closed but unbounded.

A set **S** is said to be **<u>compact</u>** if it is both closed as well as bounded.

FUNCTIONS

Given a domain $\mathbf{D} \subseteq \mathbb{R}^n$, a function f is a mapping $f: \mathbf{D} \to \mathbb{R}$ that maps each point in the domain on to a real number. A function f is said to be **<u>continuous</u>** at a point $\boldsymbol{x}^0 \in \mathbf{D}$, if $\forall \varepsilon > 0, \exists\ \delta > 0\ \ni \boldsymbol{y} \in \mathbf{D}$, $\left\|\boldsymbol{y} - \boldsymbol{x}^0\right\| < \delta \Rightarrow |f(\boldsymbol{y}) - f(\boldsymbol{x}^0)| < \varepsilon$.





Fact: Continuous functions attain their maxima or minima over compact sets. For example, Max $f(x) = x$; st $0 \le x \le 1$ yields a maximum at $x=1$ since the maximum is over a compact set. On the other hand, for the problem Max $f(x) = x$; st $0 \le x < 1$, the maximization is over a non-compact set, and thus even though the *supremum* of $f(x)$ is equal to 1, this is never attained at any x, i.e., there is no maximizing point. Note that sometimes the maxima or minima over non-compact sets may still be attained (e.g., the minimum for the latter problem is attained at $x=0$).

Differentiability

Let $\mathbf{S}$ be a nonempty set in $\mathbb{R}^n$. A function f: $\mathbf{S} \to \mathbb{R}$ is differentiable at a point $\boldsymbol{x^0} \in int(\mathbf{S})$ if for all $\boldsymbol{h}$ in $\mathbb{R}^n$ that satisfy $(\boldsymbol{x^0}+\boldsymbol{h}) \in \mathbf{S}$, there exist

($\boldsymbol{a}$) a vector $\nabla \boldsymbol{f}(\boldsymbol{x^0})$ and ($\boldsymbol{b}$) a function $\beta(\boldsymbol{x^0},\boldsymbol{h})$

such that

$$f(\boldsymbol{x^0}+\boldsymbol{h}) = f(\boldsymbol{x^0}) + [\nabla \boldsymbol{f}(\boldsymbol{x^0})]^{\mathrm{T}}\boldsymbol{h} + \|\boldsymbol{h}\|\,\beta(\boldsymbol{x^0},\boldsymbol{h}) \qquad (\otimes)$$

where $\beta(\boldsymbol{x^0},\boldsymbol{h})$ has the property that $\lim_{\|\boldsymbol{h}\| \to 0} \beta(\boldsymbol{x^0},\boldsymbol{h}) = 0$.

Here $\nabla \boldsymbol{f}(\boldsymbol{x^0}) = \left[\frac{\partial f}{\partial x_1}(\boldsymbol{x^0}),\ \frac{\partial f}{\partial x_2}(\boldsymbol{x^0}),\ \ldots,\ \frac{\partial f}{\partial x_n}(\boldsymbol{x^0}) \right]^T$ is called the gradient vector.

Note that the above representation of f as given in ($\otimes$) is called a *first-order Taylor series expansion* of f at the point $\boldsymbol{x^0}$ and without the remainder term $\|\boldsymbol{h}\|\,\beta(\boldsymbol{x^0},\boldsymbol{h})$ it is called a *first-order Taylor series approximation* of f at $\boldsymbol{x^0}$. An equivalent form is $f(\boldsymbol{x}) \approx f(\boldsymbol{x^0}) + [\boldsymbol{x}-\boldsymbol{x^0}]^{\mathrm{T}} \nabla \boldsymbol{f}(\boldsymbol{x^0})$

Another important concept: Using the same definitions for f and $\boldsymbol{h}$ as above, f is said to be **twice differentiable** at $\boldsymbol{x}^0 \in int(\mathbf{S})$ if in addition to $\nabla \boldsymbol{f}(\boldsymbol{x}^0)$ and $\beta(\boldsymbol{x}^0,\boldsymbol{h})$ as defined above, there exists an $n \times n$ symmetric matrix $\boldsymbol{H}(\boldsymbol{x}^0)$ called the Hessian ∋

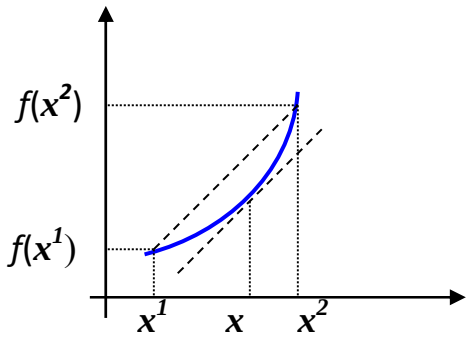
$$f(\boldsymbol{x}^0+\boldsymbol{h}) = f(\boldsymbol{x}^0) + [\nabla \boldsymbol{f}(\boldsymbol{x}^0)]^{\mathrm{T}}\boldsymbol{h} + (\tfrac{1}{2})\boldsymbol{h}^{\boldsymbol{T}}\boldsymbol{H}(\boldsymbol{x}^0)\,\boldsymbol{h} + \|\boldsymbol{h}\|^2 \beta(\boldsymbol{x}^0,\boldsymbol{h}) \qquad (\oplus)$$

Here $\boldsymbol{H}(\boldsymbol{x}^0)$ has as its $(i\text{-}j)^{\text{th}}$ element the second partial $\dfrac{\partial^2 f}{\partial x_i \partial x_j}(\boldsymbol{x}^0)$.

Again, the above representation of f as given in $(\oplus)$ is called a *second-order Taylor series expansion* of f at the point $\boldsymbol{x}^0$ and without the remainder term $\|\boldsymbol{h}\|\beta(\boldsymbol{x}^0,\boldsymbol{h})$ it is called a *second-order Taylor series approximation* of f at $\boldsymbol{x}^0$**.** An equivalent form is

$$f(\boldsymbol{x}) \approx f(\boldsymbol{x}^0) + [\boldsymbol{x}\text{-}\boldsymbol{x}^0]^{\mathrm{T}}\,\nabla f(\boldsymbol{x}^0) + (\tfrac{1}{2})[\boldsymbol{x}\text{-}\boldsymbol{x}^0]^{\mathrm{T}}\boldsymbol{H}(\boldsymbol{x}^0)\,[\boldsymbol{x}\text{-}\boldsymbol{x}^0]$$

Mean Value Theorem: Given **S,** a nonempty convex open set in $\mathbb{R}^n$ and a function f:$\mathbf{S}\rightarrow\mathbb{R}$ that is differentiable. Then $\forall\ \boldsymbol{x}^1, \boldsymbol{x}^2 \in \mathbf{S}$, $\exists\ \lambda \in (0,1)$ ∋ $f(\boldsymbol{x}^2)=f(\boldsymbol{x}^1)+[\boldsymbol{x}^2\text{-}\boldsymbol{x}^1]^{\mathrm{T}}\nabla \boldsymbol{f}(\boldsymbol{x})$, i.e., $f(\boldsymbol{x}^2)\text{-}f(\boldsymbol{x}^1)= [\boldsymbol{x}^2\text{-}\boldsymbol{x}^1]^{\mathrm{T}}\nabla \boldsymbol{f}(\boldsymbol{x})$, where $\boldsymbol{x}=\lambda\boldsymbol{x}^1+(1\text{-}\lambda)\boldsymbol{x}^2$



An illustration for a univariate function is shown alongside…

Taylor's Theorem: Given **S,** a nonempty convex open set in $\mathbb{R}^n$ and a function f:**S**$\rightarrow\mathbb{R}$ that is twice differentiable.

Then $\forall\ \boldsymbol{x^1}, \boldsymbol{x^2}\in\mathbf{S}$, $\exists\ \lambda\in(0,1)$ and $\boldsymbol{x}=\lambda\boldsymbol{x^1}+(1-\lambda)\boldsymbol{x^2}$ ϶

$$f(\boldsymbol{x^2})=f(\boldsymbol{x^1})+[\boldsymbol{x^2}-\boldsymbol{x^1}]^{\mathrm{T}}\nabla\boldsymbol{f}(\boldsymbol{x^1})+\left(\tfrac{1}{2}\right)[\boldsymbol{x^2}-\boldsymbol{x^1}]^{\mathrm{T}}\boldsymbol{H}(\boldsymbol{x})\,[\boldsymbol{x^2}-\boldsymbol{x^1}]$$

Functions with continuous first derivatives are said to be continuously differentiable, functions with continuous first and second derivatives are said to be twice-continuously differentiable, etc. In general, the class of functions with continuous derivatives of order 1 through k is denoted by C^k. Functions with high degrees of differentiability are said to be "smooth" functions.

SEQUENCES

A sequence of vectors $\{\boldsymbol{x^k}\}$ (= $\boldsymbol{x^1}, \boldsymbol{x^2},\ldots$) converges to a limit point $\boldsymbol{x^*}$ if $\left\|\boldsymbol{x}^k-\boldsymbol{x}^*\right\|$ $\rightarrow 0$ as $k\rightarrow\infty$, i.e., given $\varepsilon>0$, $\exists$ a positive integer N_ε ϶ $\left\|\boldsymbol{x}^k-\boldsymbol{x}^*\right\|<\varepsilon$ for $k\geq N_\varepsilon$. We define $\lim_{k\rightarrow\infty}\left\{\boldsymbol{x^k}\right\}=\boldsymbol{x^*}$.

A subsequence of $\{\boldsymbol{x^k}\}$ is a subset of $\boldsymbol{x^1}, \boldsymbol{x^2},\ldots$, and is denoted by $\{\boldsymbol{x^k}\}_\Psi$ where $\Psi\subset\{1,2,3,\ldots\}$

Bolzano-Weierstrass Theorem: Every sequence $\{\boldsymbol{x^k}\}$ in a compact set **S** has a convergent subsequence with limit $\boldsymbol{x^*}$ in **S**.

If for a sequence, given any $\varepsilon>0$, $\exists\ N_\varepsilon>0$ ϶ $\left\|\boldsymbol{x}^k-\boldsymbol{x}^m\right\|<\varepsilon\ \forall\ k, m\geq N_\varepsilon$, then the sequence is called a ***Cauchy*** sequence. A sequence in $\mathbb{R}^n$ has a limit if, and only if, it is Cauchy.

NUMERICAL SOLUTION OF EQUATIONS

Used when analytical solutions are not possible; e.g.,

$2x - \sin x = 0.5,$

$x^2 + x \log x = 1$

$x^4 + 5 - 2x = 0$

Methods are iterative and provide successively better approximations at each iteration.

NEWTON RAPHSON METHOD

To solve $f(x) = 0$.

1) Start with x^0 - an initial 'guess.'

2) If $|f(x^0)| > 0$, then we try to find a 'correction', say Δx, so that

$$|f(x^0 + \Delta x)| < |f(x^0)|.$$

3) Let $x^0 + \Delta x = x^1$, our new (and better) approximation. Repeat the procedure until $|f(x^n)| = |f(x^*)| \cong 0$ ('sufficiently' close to zero)

Q. How do we get Δx?

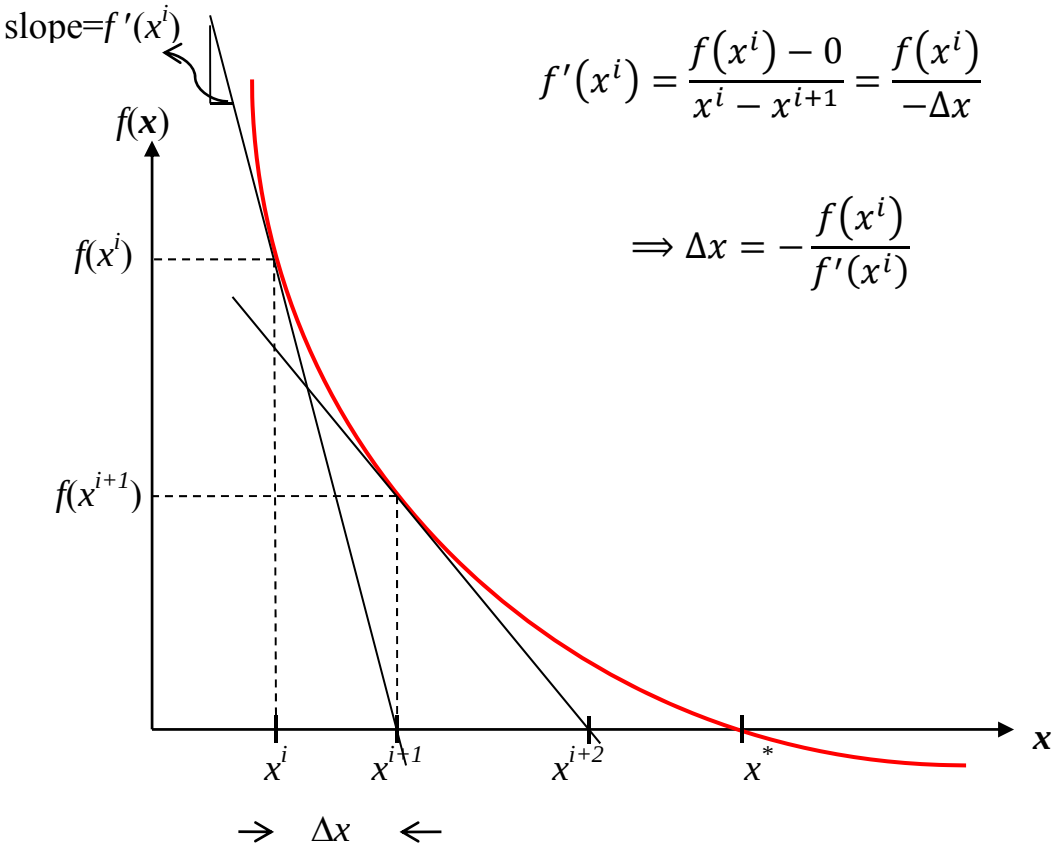
A. Use a TAYLOR SERIES expansion about x^0:

$$\underbrace{f(x^0 + \Delta x)}_{\text{want this to be zero}} = f(x^0) + \Delta x\, f'(x^0) + \underbrace{\tfrac{1}{2} (\Delta x)^2 f''(x)}_{\text{small...negligible,}} + \ldots$$

Thus, at iteration $\underline{i}$ we set the first order Taylor series approximation to zero and solve: $f(x^i) + \Delta x\, f'(x^i) \cong 0 \quad \Rightarrow \quad \boxed{\Delta x = - [f(x^i) / f'(x^i)]}$

Then $x^{i+1} = x^i + \Delta x$

<u>GEOMETRIC INTERPRETATION</u>: We look for x^* such that $f(x^*)=0$, i.e., where f intersects the X-axis,



$$f'(x^i) = \frac{f(x^i) - 0}{x^i - x^{i+1}} = \frac{f(x^i)}{-\Delta x}$$

$$\Rightarrow \Delta x = -\frac{f(x^i)}{f'(x^i)}$$

Note that $x^{i+1} = x^i + \Delta x$ is the intersection of the tangent to the curve at $[x^i, f(x^i)]$ with the X-axis.

<u>Example</u>: Given $a>0$, find $\sqrt{a}$ using only the operations (+, –, ÷, ×),

Let $\sqrt{a} = x \quad \Rightarrow \quad x^2 - a = 0,$

To solve the equation $f(x) = x^2 - a = 0$, let $x^i \equiv$ the guess at iteration i,

$$\Longrightarrow \quad \Delta x = -\frac{f(x^i)}{f'(x^i)} = -\frac{(x^i)^2 - a}{2x^i}$$

and the improved guess $x^{i+1} = x^i + \Delta x = x^i - \frac{(x^i)^2 - a}{2x^i}$

$$= \frac{(x^i)^2 + a}{2x^i} = 0.5\left(x^i + {}^{a}/_{x^i}\right)$$

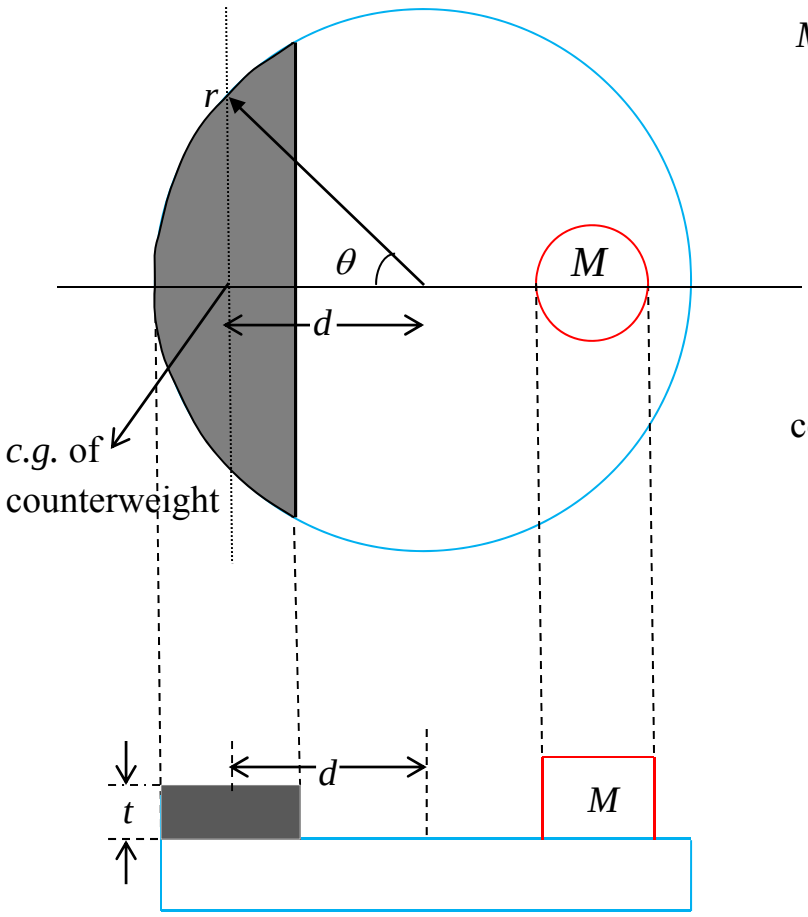
e.g., **Suppose a = 5, and the initial guess x^0 =2**, and we use a tolerance of 0.001

i	x^i	$f(x^i)$	x^{i+1}
0	2	-1	0.5(2+5/2)=2.25
1	2.25	0.0625	0.5(2.25+5/2.25)= 2.2361
2	2.2361	0.00014	*STOP*

Suppose a = 2, and the initial guess x^0 =1

i	x^i	$f(x^i)$	x^{i+1}
0	1	-1	0.5(1+2/1)=1.5
1	1.5	0.25	0.5(1.5+2/1.5)= 1.4167
2	1.4167	0.00704	0.5(1.4167+2/1.4167)= 1.4142
3	1.4142	-0.00004	*STOP*

WHEEL-COUNTERWEIGHT PROBLEM



r=radius of wheel,

M = unbalanced moment,

t = thickness of the counterweight

ρ = density of the counterweight

d = lever arm; wheel center to center of gravity

θ = arc; half angle

A = area of counterweight

PROBLEM: Design a sector shaped counterweight for a locomotive driver wheel to balance a moment M,

e.g. M = 600 Newton-m,

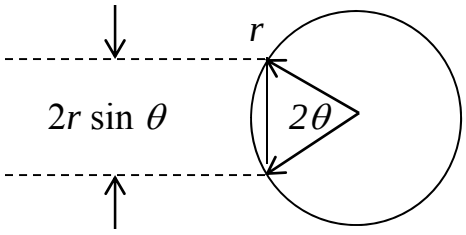
t = 0.1 m,

ρ= 700 kg/m^3,

r = 1.0 m,

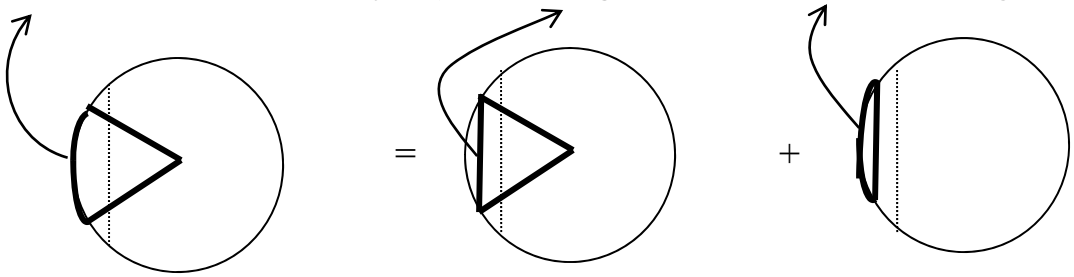
Moment (M) = (Force)*(Distance)= (Mass)*(g)*(Distance)

= (Volume)*(Density)*(g)*(Distance); i.e., $\boldsymbol{M = (At) * (\rho) * (g) * (d)}$

Note that $2r \sin\theta$ 

and $\cos\theta = d/r$

$r^2\theta$ (area of sector subtended by 2θ) *c.g.* ½ (counterweight area)



i.e., the counterweight area A is given by

$A = 2\{r^2\theta - \frac{1}{2}(2r\sin\theta)d\}$ [Note that the *c.g.* is "inside" the counterweight]

$= 2\{r^2\theta - r^2\sin\theta\cos\theta\} = r^2(2\theta - 2\sin\theta\cos\theta) = r^2(2\theta - \sin2\theta)$

Thus we have: $\boldsymbol{M = r^2(2\theta - \sin2\theta)\, t\, \rho\, g\, r\, (\cos\theta)}$

SOLVE FOR θ...,

ANALYTICAL APPROACH

$[2\theta - \sin2\theta]\cos\theta = M/(t\rho g r^3)$ CLEARLY IMPOSSIBLE TO SOLVE !

NEWTON RAPHSON APPROACH

Let $M/(t\rho g r^3) = K$ (from given values)

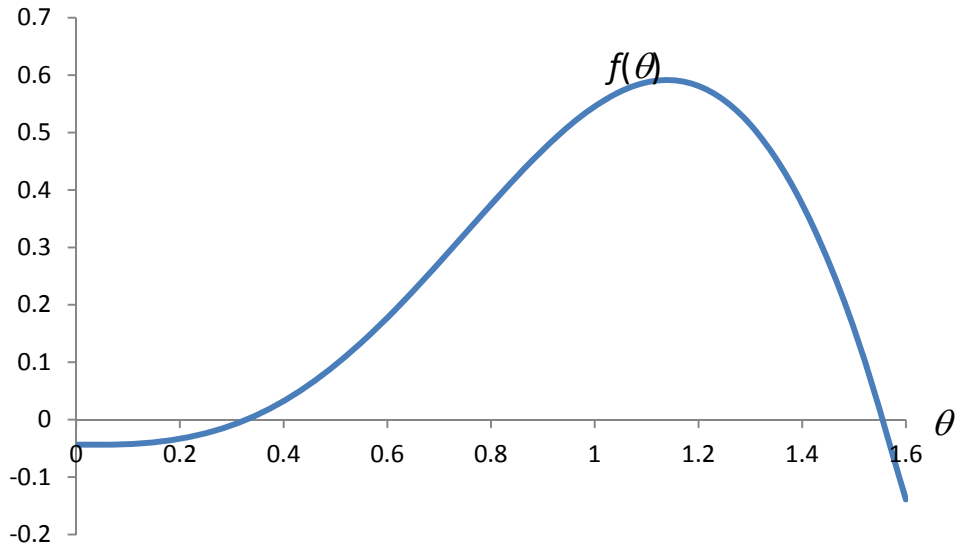
Then we solve

$$f(\theta) = [2\theta - \sin 2\theta]\cos\theta - K = 0$$

$$\Rightarrow \quad f'(\theta) = \cos\theta\,[2-2\cos 2\theta] + (2\theta - \sin 2\theta)^* (-\sin\theta)$$

If θ^i = approximation at iteration i, then $\Delta\theta = -f(\theta^i) \div f'(\theta^i)$, and the improved approximation at the next iteration is

$$\theta^{i+1} = \theta^i - \{f(\theta^i) \div f'(\theta^i)\}$$



```
C*********************************************************
C   DESIGN OF WHEEL COUNTER WEIGHT BY NEWTON-RAPHSON METHOD*
C*********************************************************
        REAL  M,T,TOL,R,RHO,F,FF,C,CORR,X
C
        WRITE (*,*), 'ENTER UNBALANCED MOMENT (M) IN NEWTON-METERS'
        READ *, M
        WRITE (*,*), 'ENTER THICKNESS OF COUNTERWEIGHT (T) IN METERS'
        READ *, T
        WRITE (*,*), 'ENTER DENSITY (RHO) OF COUNTERWEIGHT IN
KG/CU.M'
        READ *, RHO
        WRITE (*,*), 'ENTER RADIUS (R) OF WHEEL IN METERS'
        READ *, R
        WRITE (*,*), 'ENTER MAX ALLOWABLE UNBALANCED MOMENT IN N-
METERS'
        READ *, TOL
        TOL=TOL/(T*RHO*9.81*R**3)
        WRITE (*,*), 'ENTER INITIAL GUESS IN RADIANS FOR THETA'
        READ *, X
        C = M/(T*RHO*9.81*R**3)
C*******EVALUATE F(THETA) AT CURRENT APPROXIMATION
100     F = COS(X)*(2*X-SIN(2*X))
        F=F-C
        WRITE (*,*), 'THETA =',X,'      F(THETA)=',F
C*******EVALUATE DERIVATIVE OF F(THETA) AT CURRENT APPROXIMATION
        FPRM = -SIN(X)*(2*X-SIN(2*X)) + COS(X)*(2-2*COS(2*X))
C*******EVALUATE THE CORRECTION
        CORR = -F/FPRM
C*******ADD CORRECTION TO THETA
        X=X+CORR
C*******TEST FOR TOLERANCE
        IF (ABS(F).GT.TOL) GOTO 100
        FF=ABS(F)*T*RHO*9.81*R**3
        WRITE (*,*), 'THE UNBALANCED MOMENT IS ',FF
        STOP
        END
```

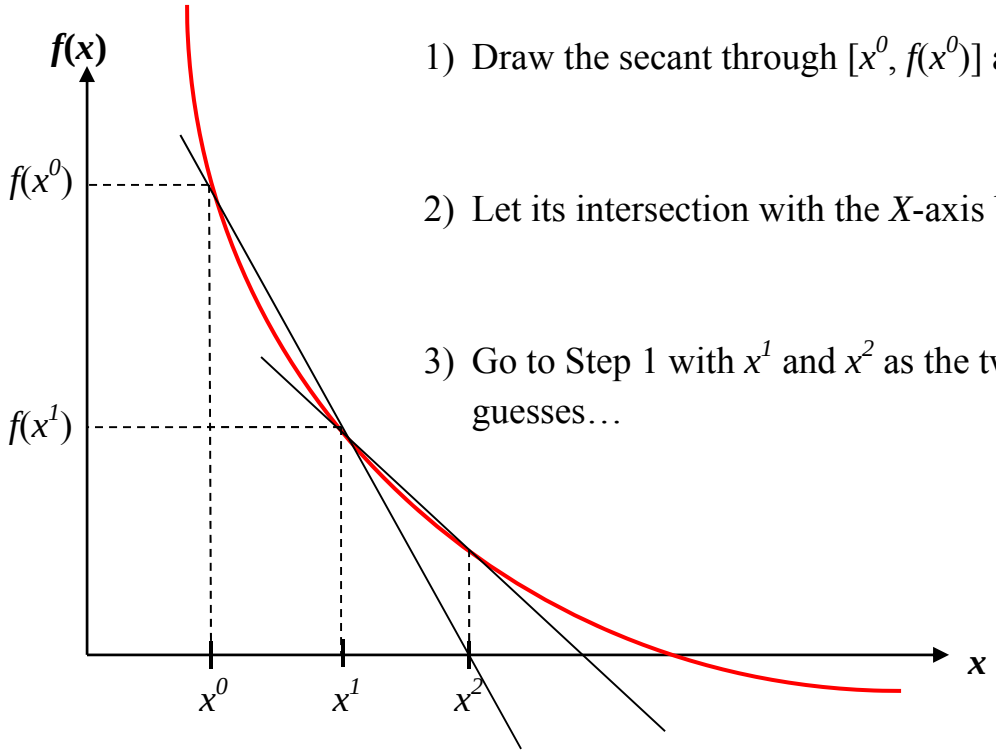
```
$RUN NEWTRAP
ENTER UNBALANCED MOMENT (M) IN NEWTON-METERS
600
ENTER THICKNESS OF COUNTERWEIGHT (T) IN METERS
0.1
ENTER DENSITY (RHO) OF COUNTERWEIGHT IN KG/CU.M
7000
ENTER RADIUS (R) OF WHEEL IN METERS
1
ENTER MAX ALLOWABLE UNBALANCED MOMENT IN N-METERS
0.001
ENTER INITIAL GUESS IN RADIANS FOR THETA
1
THETA = 1.0000000          F(THETA) =  0.5019348
THETA = 0.1805149          F(THETA) = -7.9709038E-02
THETA = 0.8159554          F(THETA) =  0.3468729
THETA = 0.4664296          F(THETA) =  2.8322041E-02
THETA = 0.4237927          F(THETA) =  1.8656477E-03
THETA = 0.4205555          F(THETA) =  1.0982156E-05
THETA = 0.4205362          F(THETA) = -2.2351742E-08
THE UNBALANCED MOMENT IS 1.5348941E-04
FORTRAN STOP
```

```
$RUN NEWTRAP
ENTER UNBALANCED MOMENT (M) IN NEWTON-METERS
600
ENTER THICKNESS OF COUNTERWEIGHT (T) IN METERS
0.1
ENTER DENSITY (RHO) OF COUNTERWEIGHT IN KG/CU.M
7000
ENTER RADIUS (R) OF WHEEL IN METERS
1
ENTER MAX ALLOWABLE UNBALANCED MOMENT IN N-METERS
0.001
ENTER INITIAL GUESS IN RADIANS FOR THETA
1.4
THETA = 1.400000      F(THETA) =  0.3315967
THETA = 1.587457      F(THETA) = -0.1408246
THETA = 1.544450      F(THETA) = -7.3895305E-03
THETA = 1.541928      F(THETA) = -2.5711954E-05
THETA = 1.541919      F(THETA) = -5.2154064E-08
THE UNBALANCED MOMENT IS 3.5814199E-04
FORTRAN STOP
```

SECANT METHOD [TO SOLVE $f(x)=0$]

The Newton-Raphson method requires *derivatives* ($f'(x)$) which are sometimes difficult to compute...

Let x^0 and x^1 be **two** initial "guesses"



1) Draw the secant through $[x^0, f(x^0)]$ and $[x^1, f(x^1)]$

2) Let its intersection with the X-axis be at x^2

3) Go to Step 1 with x^1 and x^2 as the two new guesses…

To Find x^2:

From properties of similar triangles:

$$\frac{x^2 - x^1}{f(x^1) - 0} = \frac{x^1 - x^0}{f(x^2) - f(x^1)}$$

Solving for x^2 we have,

$$x^2 = x^1 - \frac{x^1 - x^0}{f(x^1) - f(x^0)} \cdot f(x^1)$$

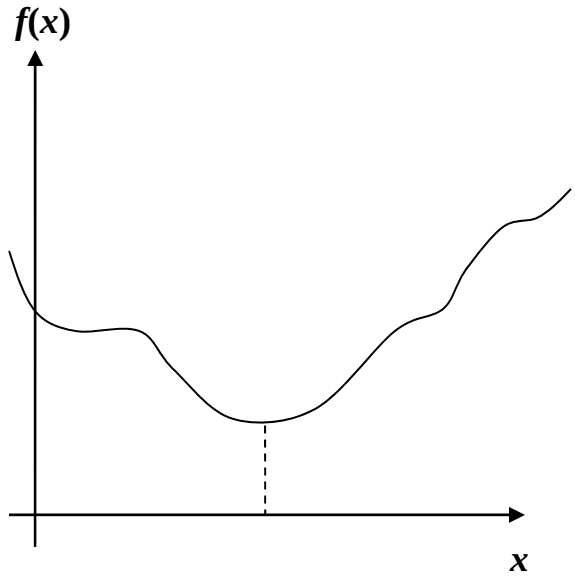
MINIMIZATION OF A FUNCTION OF A SINGLE VARIABLE

(WITHOUT USING DERIVATIVES)

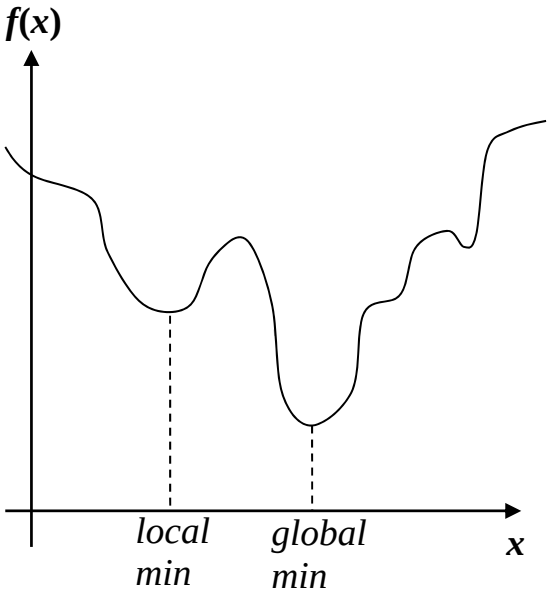
⇒ Analytical expressions for $f(x)$ may be unknown (e.g. $f(x)$ is measured in a laboratory), or $f(x)$ may not be differentiable everywhere

⇒ The result of minimization will be an "interval of uncertainty" which contains the optimum and we assume that an initial interval of uncertainty is specified - say $[a^0, b^0]$. The method stops when the interval is "sufficiently" small.

⇒ We assume that f is **UNIMODAL**.



UNIMODAL



NOT UNIMODAL

⇒ These methods are usually used to solve "subproblems" within larger multi-variable problems.

THREE POINT EQUI-INTERVAL SEARCH

A simple but rather inefficient approach...

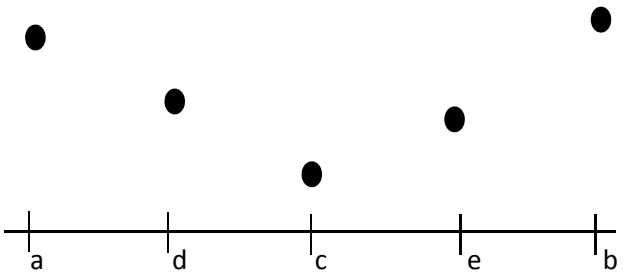
- Let $[a^i, b^i]$ be the interval of uncertainty iteration i. Let $c^i = (a^i + b^i)/2$ -- the midpoint -- and let $f(a^i)$, $f(b^i)$, $f(c^i)$ be known,
- Find the midpoints of the two new subintervals $[a^i, c^i]$ and $[c^i, b^i]$ via

$$d^i = \frac{3a^i + b^i}{4}, \qquad e^i = \frac{a^i + 3b^i}{4}$$

- Evaluate $f(d^i)$ and $f(e^i)$

Consider the relative magnitudes of f at the five points.

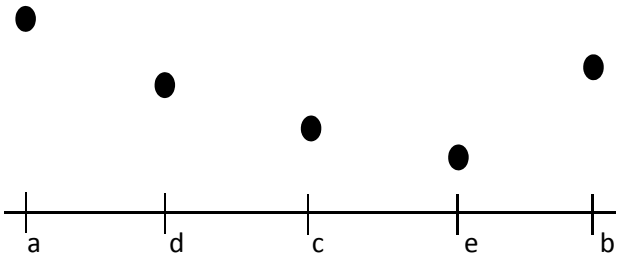
Case 1:



New interval of uncertainty

$a^{i+1}=$ $b^{i+1}=$

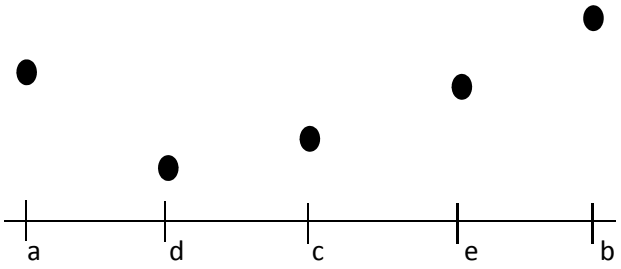
Case 2:



New interval of uncertainty

$a^{i+1}=$ $b^{i+1}=$

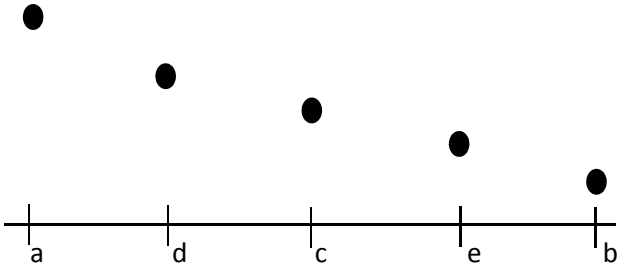
Case 3:



New interval of uncertainty

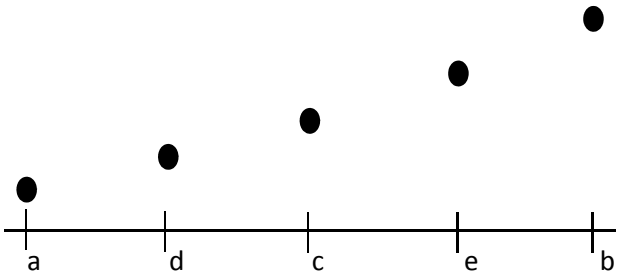
$a^{i+1}=$ $b^{i+1}=$

Case 4: New interval of uncertainty



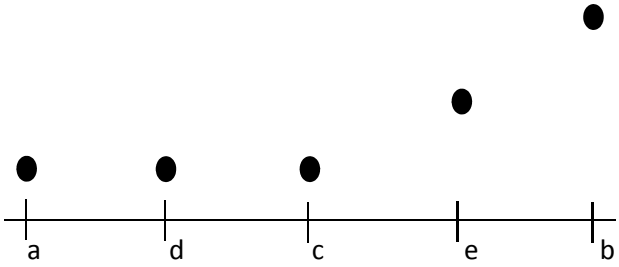
$a^{i+1}=$ $b^{i+1}=$

Case 5: New interval of uncertainty



$a^{i+1}=$ $b^{i+1}=$

Case 6: New interval of uncertainty



$a^{i+1}=$ $b^{i+1}=$

Repeat the procedure until the interval of uncertainty is smaller than some prespecified tolerance ε, i.e., $|b^n - a^n| < \varepsilon$.

At each iteration, the interval of uncertainty is (usually) halved (for example in Cases 1-3); occasionally it may be reduced by more than 50 % (Cases 4-5) or less than 50% (Case 6). HOWEVER, we need TWO function evaluations at each iteration -- this may be quite expensive....

A more efficient search method is the so-called GOLDEN SECTION SEARCH, which tries to reduce the number of function evaluations required.

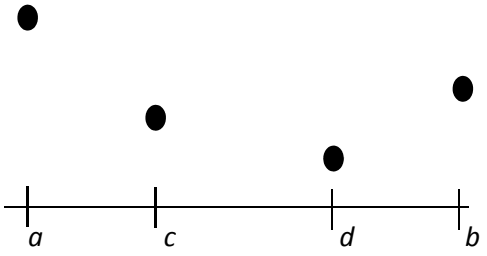
GOLDEN SECTION SEARCH

Again, we have an interval of uncertainty $[a^i, b^i]$ at iteration i, along with a point c^i in the interior. HOWEVER, c^i is NOT the midpoint of the interval...

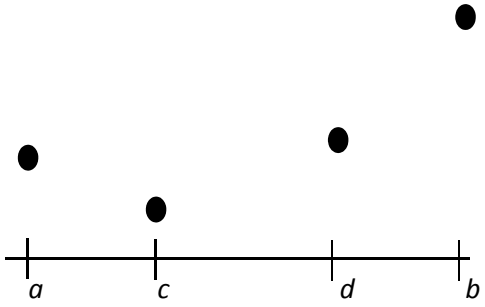
We choose c^i along with another interior point d^i such that they are both symmetrical about the midpoint of $[a^i, b^i]$.

Consider the cases shown below:

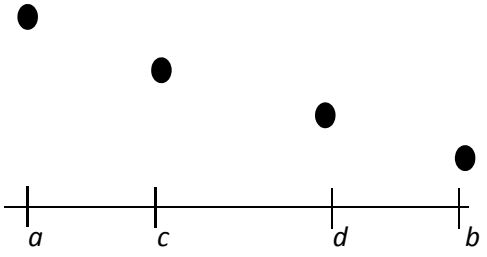
Case 1



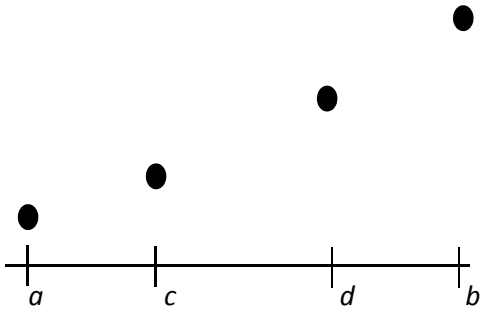
Case 2



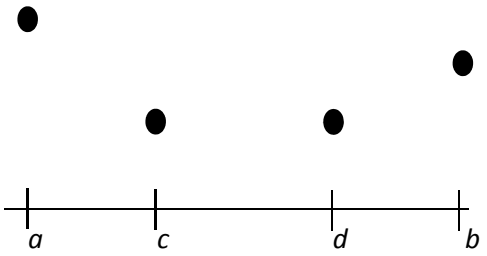
Case 3



Case 4



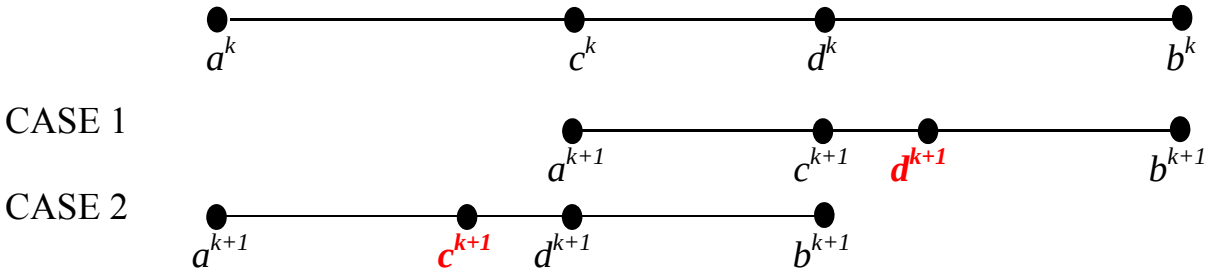
Case 5

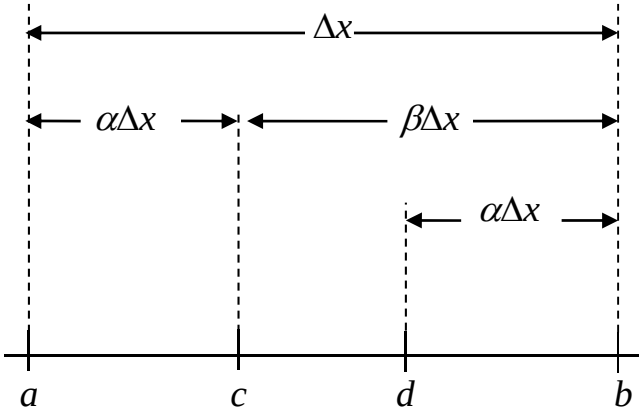


Find $[a^{i+1}, b^{i+1}]$ for each of these five cases

The interval is (usually) less than halved at each step. (i.e., the new interval is larger than in the 3-pt equi-interval search). HOWEVER, only ONE new function evaluation is required!

QUESTION: How to locate c^i inside $[a^i, b^i]$?





Δx = length of current interval

$\overline{\Delta x}$ = length of next interval

α, β are constants; $\alpha + \beta = 1$

α and β must satisfy

$\leftarrow \alpha\overline{\Delta x} \rightarrow\leftarrow \beta\overline{\Delta x} \rightarrow$ $\quad \beta\Delta x = \overline{\Delta x} \quad \Rightarrow \quad \beta = 1 - \alpha = \overline{\Delta x}/\Delta x$

$$\beta\Delta x = a\overline{\Delta x} + \beta\overline{\Delta x} \qquad (\blacklozenge)$$

$$\beta = a + a(\overline{\Delta x}/\Delta x) \Longrightarrow \frac{\beta - \alpha}{\alpha} = \frac{\overline{\Delta x}}{\Delta x} \Longrightarrow \frac{1 - 2\alpha}{\alpha} = \frac{\overline{\Delta x}}{\Delta x} \qquad (\blacklozenge\blacklozenge)$$

So from ($\blacklozenge$)and ($\blacklozenge\blacklozenge$) it follows that $(1-2\alpha)/\alpha = (1-\alpha)$; i.e., $\alpha^2 - 3\alpha + 1 = 0$

$\alpha = \frac{3 \pm \sqrt{5}}{2}$. Choose "-" sign; "+" leads to $\alpha > 1$

$\alpha = (3-\sqrt{5})/2$	$= 0.381966$
$\beta = 1-\alpha$	$= 0.618034$

Q. WHAT IS THE IMPLICATION?

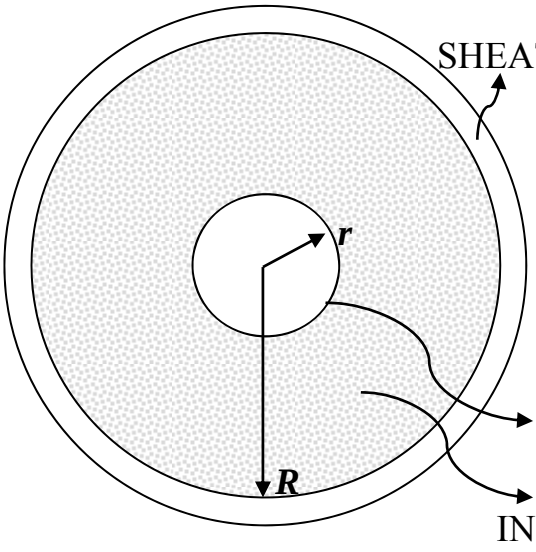
A. If c^i (d^i) is placed at a distance of **0.381966*($b^i - a^i$)** units from a^i (b^i), then at the beginning of iteration n, c^n (d^n) will automatically be at a distance of **0.381966*($b^n - a^n$)** from a^n (b^n).

Here at each iteration, the interval of uncertainty Δx is reduced by approx 38% and there is only ONE new function evaluation per iteration.

Compare Golden Section to 3-pt Equi-Interval Search...

No. of evaluations of $f(x)$	Interval of Uncertainty 3-pt. Equi-Interval	Interval of Uncertainty Golden Section
3	1.00	1.00
4		$(0.618)^1$=0.618
5	$(0.5)^1$=0.5	$(0.618)^2$=0.382
6		$(0.618)^3$=0.236
7	$(0.5)^2$=0.25	$(0.618)^4$=0.146
8		$(0.618)^5$=0.090
9	$(0.5)^3$=0.125	$(0.618)^6$=0.0557
10		$(0.618)^7$=0.0344
11	$(0.5)^4$=0.0625	$(0.618)^8$=0.0213
12		$(0.618)^9$=0.0132
13	$(0.5)^5$=0.01325	$(0.618)^{10}$=0.00813
14		$(0.618)^{11}$=0.00503
15	$(0.5)^6$=0.0156	$(0.618)^{12}$=0.00311
16		$(0.618)^{13}$=0.00192
17	$(0.5)^1$=0.0078	$(0.618)^{14}$=0.00119

ELECTRICAL CABLE INSULATION



Line Voltage applied = V

$V = \frac{\sigma \ln\left(\frac{R}{r}\right)}{2\pi\epsilon}$. where

σ = charge/unit length

ϵ = dielectric constant of the insulation

Maximum Electrical Stress = $\sigma/(2\pi\epsilon r)$

Utilization = $\boldsymbol{U}$ =(Line Voltage) / (Maximum Electrical Stress)

$$\Longrightarrow \quad U = \left.\frac{\left(\sigma \ln(\frac{R}{r})\right)}{2\pi\epsilon}\right/ \frac{\sigma}{2\pi\epsilon r} \quad = \quad r \ln(\frac{R}{r})$$

PROBLEM: Proportion R with r so as to maximize the "utilization" for a given volume (cross-sectional area) of the insulation --- this allows a 'maximum' safe voltage for the cable for a given stress (and hence presumably the most power).

A = area = $\pi R^2 - \pi r^2 = \pi r^2(R^2/r^2 - 1) = \pi r^2(x^2 - 1)$, where we define $x = R/r$

Also, $U = r \ln(R/r)$ $\Rightarrow$ $r = U / (\ln x)$. Therefore

$$A = \pi \frac{U^2}{(\ln x)^2}(x^2 - 1) \quad \Longrightarrow \quad U = \sqrt{\frac{A(\ln x)^2}{\pi(x^2 - 1)}}$$

To maximize U, we perform the equivalent task of minimizing $1/U$...

$$Minimize\ 1/U = \sqrt{\frac{\pi(x^2-1)}{A(\ln x)^2}}$$

i.e., $\boxed{Minimize\ V = \left(\sqrt{c(x^2-1)}\right) / \ln x}$, where $V = 1/U$ and $c = \pi/A$

(It turns out that this function is unimodal in x)

Thus one could use (1) Three point Equi-interval Search, or (2) Golden Section Search.

FIBONACCI SEARCH

Named after Leonardo of Pisa, son of Bonacci (hence Fibonacci!), in 1202 A.D.

Demonstrated initially with rabbit-breeding...

ASSUME

- Each pair of mature rabbits produces one pair per litter,
- Exactly ONE litter per month,
- Maturation requires 1 month,

How many pairs of rabbits are there at the end of each month, starting with one new born pair in January?

Month	J	F	M	A	M	J	J	A	S	O
Immature	1	0	1	1	2	3	5	8	13	
Mature	0	1	1	2	3	5	8	13	21	
TOTAL	1	1	2	3	5	8	13	21	34	

FIBONACCI NUMBERS: 1, 1, 2, 3, 5, 8, 13, 21,...

In general, $F_0 = F_1 = 1$, $F_{N+1} = F_N + F_{N-1}$,

The Fibonacci numbers are used for the Fibonacci Search. It also starts with 2 initial evaluations, and with only ONE subsequent evaluation per iteration. HOWEVER, the interval of uncertainty is *not reduced by the same amount each time*. Also, the number of function evaluations planned (say n) is determined *before* the search commences.

Suppose at iteration k we have the interval $I_k = [a_k, b_k]$ of length $\bar{I}_k = b_k - a_k$. Let

$$c_k = a_k + \frac{F_{n-(k+1)}}{F_{n-(k-1)}}(b_k - a_k)$$

$$d_k = a_k + \frac{F_{n-k}}{F_{n-(k-1)}}(b_k - a_k)$$

Then the new interval of uncertainty I_{k+1} will usually be $[a_k, d_k]$ or $[c_k, b_k]$,

In the first case, $\bar{I}_{k+1} = d_k - a_k = (F_{n-k}/F_{n-(k-1)})\,\bar{I}_k$

In the second case, $\bar{I}_{k+1} = b_k - c_k = (F_{n-k}/F_{n-(k-1)})\,\bar{I}_k$ (VERIFY!!)

In either case $\bar{I}_k$ is reduced by a factor of $(F_{n-k}/F_{n-(k-1)})$

<u>EXERCISE</u>: At iteration k+1, $[a_{k+1}, b_{k+1}]$ = (i) $[c_k, b_k]$ or (ii) $[a_k, d_k]$. Show that,

(i) $c_{k+1} = d_k$ or (ii) $d_{k+1} = c_k$,

Thus only ONE new function evaluation is required at the next step.

QUESTION: How should we choose *n*?

Let *l* = required final interval of uncertainty (Tolerance). We know that

$$\bar{I}_2 = \frac{F_{n-1}}{F_n}\bar{I}_1$$

$$\bar{I}_3 = \frac{F_{n-2}}{F_{n-1}}\bar{I}_2 = \frac{F_{n-2}}{F_{n-1}}\frac{F_{n-1}}{F_n}\bar{I}_1 = \frac{F_{n-2}}{F_n}\bar{I}_1$$

$$\bar{I}_4 = \frac{F_{n-3}}{F_{n-2}}\bar{I}_3 = \frac{F_{n-3}}{F_{n-2}}\frac{F_{n-2}}{F_n}\bar{I}_1 = \frac{F_{n-3}}{F_n}\bar{I}_1$$

etc., etc., etc…

$$\bar{I}_n = \frac{F_{n-(n-1)}}{F_{n-(n-2)}}\bar{I}_{n-1} = \frac{F_1}{F_2}\bar{I}_{n-1} = \frac{F_1}{F_2}\frac{F_2}{F_n}\bar{I}_1 = \frac{F_1}{F_n}\bar{I}_1 = \frac{1}{F_n}\bar{I}_1$$

So, if we want $\bar{I}_n$ to be < *l*, then we must choose $\frac{1}{F_n}\bar{I}_1 < l$, i.e., $F_n > \bar{I}_1 / l$.

Thus we pick the first *n* such that this condition is satisfied.

The **Fibonacci search** then proceeds as follows:

1) Start with a_1, b_1 (where $b_1-a_1=\bar{I}_1$). Then find c_1 and d_1 using the formulas on the previous page. Set k=0.

2) Set $k=k+1$. Evaluate $f(a_k)$, $f(b_k)$, $f(c_k)$ and $f(d_k)$ and eliminate a portion of the interval. In general, we will have one of two cases:

   (i) $a_{k+1}=a_k$, $b_{k+1}=d_k$, $d_{k+1}=c_k$ and c_{k+1} is computed via the formula on page 30

   (ii) $a_{k+1}=c_k$, $b_{k+1}=b_k$, $c_{k+1}=d_k$ and d_{k+1} is computed via the formula on page 30

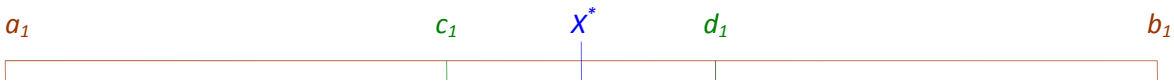
3) If $k\neq n$-1 go to Step 2.

4) When $k=n$-1 the two formulas for c_k and d_k yield the same value $c_k = d_k = a_k + (½)(b_k-a_k)$, since $F_0 = F_1 = 1$. So no further interval reduction is possible. Consequently, we place c_{n-1} (or d_{n-1} as the case may be…) at a distance ε from the existing interior point.

5) Repeat the procedure of Step 2 for this last interval and obtain the final interval of uncertainty $[a_n, b_n]$

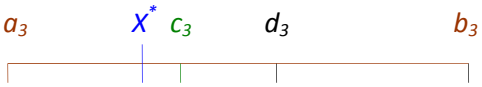
NOTE: The quantity ε is referred to as a "distinguishability constant."

We will have $b_n-a_n = (b_1-a_1)/F_n + \delta$, where $\delta = 0$ or $\delta=\varepsilon$.

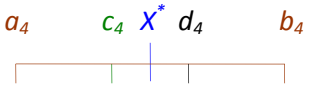


$\bar{I}_1=13l$ ITER 1

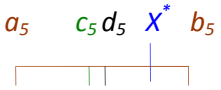
$\bar{I}_2=8l$ ITER 2



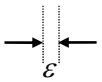
$\bar{I}_3=5l$ ITER 3



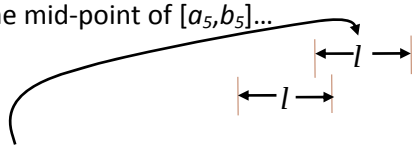
$\bar{I}_4=3l$ ITER 4



$\bar{I}_5=2l$ ITER 5



Note that d_5 is the mid-point of $[a_5,b_5]$...



$\bar{I}_6= l$ (Final Interval) ITER 6

Note that the length of the final interval is $l = (1/F_6)*(b_1-a_1)+\varepsilon$ since the final interval is $[c_5,b_5]$ for this example. If the optimum had been in the final interval $[a_5,d_5]$ then its length would have been $l = (1/F_6)*(b_1-a_1)$.

Also, note that $\bar{I}_k = \bar{I}_{k+1} + \bar{I}_{k+2}$. Thus we start with $\bar{I}_1 = \bar{I}_2 + \bar{I}_3$

Then $F_n = F_{n-1} + F_{n-2}$ $\Rightarrow$ $\frac{F_{n-1}}{F_n} + \frac{F_{n-2}}{F_n} = 1$ $\Rightarrow$ $\bar{I}_1\left(\frac{F_{n-1}}{F_n}\right) + \bar{I}_1\left(\frac{F_{n-2}}{F_n}\right) = \bar{I}_1 = \bar{I}_2 + \bar{I}_3$

EXAMPLE: Suppose that the initial interval is [1.05, 4.00], so that $\bar{I}_1$=2.95 and we need to reduce the final interval of uncertainty to l =0.01

Since $\bar{I}_1/l = 295$, the smallest value of n for which F_n exceeds 295 is given by n=13 (…F_{13}=377). So we need 13 iterations. The search might then proceed as follows:

a_1= 1.05

c_1= 1.05+ F_{11}/F_{13}(2.95) = 1.05+(144/377)(2.95) = 2.1768

d_1= 1.05+ F_{12}/F_{13}(2.95) = 1.05+(233/377)(2.95) = 2.8732

b_1= 4.00

a_2= 1.05

c_2= 1.05+ F_{10}/F_{12}(2.8732-1.05) = 1.05+(89/233)(1.8232) = 1.7464

d_2= 2.1768

b_2= 2.8732

a_3= 1.7464

c_3= 2.1768

d_3==1.7464 + F_{10}/F_{11}(2.8732-1.7464) = 1.05+(89/144)(1.268) = 2.4428

b_3= 2.8732

etc. etc. etc…

LAGRANGE'S INTERPOLATING POLYNOMIALS

Assume that we are given the n+1 values $x_0, x_1,\ldots,x_n$.

Define

$$l_j(x) = \prod_{\substack{k=0 \\ k\neq j}}^{n} \left(\frac{x - x_k}{x_j - x_k}\right)$$

$$= \left(\frac{x - x_0}{x_j - x_0}\right)\left(\frac{x - x_1}{x_j - x_1}\right)\ldots\left(\frac{x - x_{j-1}}{x_j - x_{j-1}}\right)\left(\frac{x - x_{j+1}}{x_j - x_{j+1}}\right)\ldots\left(\frac{x - x_n}{x_j - x_n}\right)$$

Some properties of $l_j(x)$:

1. $l_j(x)$ is a polynomial of degree n
2. $l_j(x_j) = 1$
3. $l_j(x_i) = 0$ for $i\neq j$,

Lagrange's Interpolating polynomial is

$$p(x) = \sum_{j=0}^{n} f(x_j) l_j(x)$$

Note that

$$p(x_0) = \sum f(x_j)\, l_j(x_0) = f(x_0)$$

$$p(x_1) = \sum f(x_j)\, l_j(x_1) = f(x_1) \qquad etc.\ \ etc.$$

QUADRATIC INTERPOLATION OF A MINIMUM

Given a, b, c and $f(a)$, $f(b)$, $f(c)$ to find a point d and $f(d)$.

The interpolating polynomial is

$$p(x) = \left[f(a)\frac{(x-b)(x-c)}{(a-b)(a-c)}\right] + \left[f(b)\frac{(x-a)(x-c)}{(b-a)(b-c)}\right] + \left[f(c)\frac{(x-a)(x-b)}{(c-b)(c-a)}\right]$$

Note: $p(a) = f(a)$, $p(b) = f(b)$, $p(c) = f(c)$.

Differentiating $p(x)$ w.r.t. x and equating to zero, we obtain

$$p'(x) = \frac{f(a)}{(a-b)(a-c)}(2x-b-c) + \frac{f(b)}{(b-a)(b-c)}(2x-a-c) + \frac{f(c)}{(c-b)(c-a)}(2x-b-a) = 0$$

Solving for x:

$$x = \frac{1}{2}\ \frac{\frac{f(a)(b+c)}{(a-b)(a-c)} + \frac{f(b)(a+c)}{(b-a)(b-c)} + \frac{f(c)(a+b)}{(c-a)(c-b)}}{\frac{f(a)}{(a-b)(a-c)} + \frac{f(b)}{(b-a)(b-c)} + \frac{f(c)}{(c-a)(c-b)}}$$

$$\Longrightarrow x = \frac{1}{2}\ \frac{f(a)(b^2-c^2) + f(b)(c^2-a^2) + f(c)(a^2-b^2)}{f(a)(b-c) + f(b)(c-a) + f(c)(a-b)}$$

Now, let $d=x$, then find $f(d)$ and proceed as usual…

MULTI-DIMENSIONAL OPTIMIZATION *WITHOUT* DERIVATIVES: THE HOOKE-JEEVES METHOD

Used to solve the general problem:

MINIMIZE $f(\boldsymbol{x})$, where $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]^T \in \mathbb{R}^n$,

At iteration i suppose we have the points $\boldsymbol{x}^i$ and $\boldsymbol{y}^1$ ($\in \mathbb{R}^n$) and a set of search directions $\boldsymbol{d}^1, \boldsymbol{d}^2, \ldots, \boldsymbol{d}^n$.

(NOTE: Usually these directions are the coordinate directions, but this is not necessary…)

STEP 1

Using $\boldsymbol{y}^1$ find the optimum solution to the problem:

Minimize $f(\lambda) = f(\boldsymbol{y}^1 + \lambda \boldsymbol{d}^1)$ to obtain λ_1.

Let $\boldsymbol{y}^2 = \boldsymbol{y}^1 + \lambda_1 \boldsymbol{d}^1$,

Using $\boldsymbol{y}^2$ find the optimum solution to the problem:

Minimize $f(\lambda) = f(\boldsymbol{y}^2 + \lambda \boldsymbol{d}^2)$ to obtain λ_2.

Let $\boldsymbol{y}^3 = \boldsymbol{y}^2 + \lambda_2 \boldsymbol{d}^2$, etc. etc…

Continue until $\boldsymbol{y}^{n+1} = \boldsymbol{y}^n + \lambda_n \boldsymbol{d}^n$

Let $\boldsymbol{x}^{i+1} = \boldsymbol{y}^{n+1}$

Compare $\boldsymbol{x}^{i+1}$ and $\boldsymbol{x}^i$; IF $\left\| \boldsymbol{x}^{i+1} - \boldsymbol{x}^i \right\| < \varepsilon$ then STOP, ELSE go to STEP 2.

STEP 2

Let $\boldsymbol{d} = \boldsymbol{x}^{i+1} - \boldsymbol{x}^{i}$; solve the subproblem:

Minimize $f(\lambda) = f(\boldsymbol{x}^{i+1} + \lambda \boldsymbol{d})$ to obtain λ^{*}.

Let $\boldsymbol{y^1} = \boldsymbol{x}^{i+1} + \lambda^{*} \boldsymbol{d}$ and return to STEP 1

NOTE: At iteration 1, there is no $\boldsymbol{y^1}$; so use $\boldsymbol{y^1} = \boldsymbol{x^1}$)

==========X==========X==========

AN EXAMPLE

Minimize $f(\boldsymbol{x}) = f(x_1, x_2) = (x_1 - 2)^4 + (x_1 - 2x_2)^2$

Start with $\boldsymbol{x^1} = \begin{bmatrix} 0 \\ 3 \end{bmatrix}$, and use $\boldsymbol{d^1} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $\boldsymbol{d^2} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$

ITERATION 1

STEP 1 $\quad \boldsymbol{y^1} = \begin{bmatrix} 0 \\ 3 \end{bmatrix}$, $\quad \text{Min } f\left\{ \begin{bmatrix} 0 \\ 3 \end{bmatrix} + \lambda \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right\} = f(\lambda, 3)$

i.e., $\quad \text{Min } (\lambda - 2)^4 + (\lambda - 2 \cdot 3)^2 \Rightarrow \lambda_1 = 3.13$

Therefore $\boldsymbol{y^2} = \boldsymbol{y^1} + \lambda_1 \boldsymbol{d^1} = \begin{bmatrix} 0 \\ 3 \end{bmatrix} + 3.13 \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 3.13 \\ 3.00 \end{bmatrix}$.

Minimize $f\left\{ \begin{bmatrix} 3.13 \\ 3.00 \end{bmatrix} + \lambda \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\} = f(3.13, 3 + \lambda)$

i.e. Min $(3.13 - 2)^4 + [3.13 - 2(3 + \lambda)]^2 \Rightarrow \lambda_2 = -1.44$,

Therefore $\boldsymbol{y^3} = \boldsymbol{y^2} + \lambda_2 \boldsymbol{d^2} = \begin{bmatrix} 3.13 \\ 3.00 \end{bmatrix} + (-1.44) \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 3.13 \\ 1.56 \end{bmatrix}$.

Let $\boldsymbol{x^2} = \begin{bmatrix} 3.13 \\ 1.56 \end{bmatrix}$, $\quad \| \boldsymbol{x}^2 - \boldsymbol{x}^1 \|$ is not negligible; go to Step 2.

STEP 2

$\boldsymbol{d} = \begin{bmatrix}3.13\\1.56\end{bmatrix} - \begin{bmatrix}0\\3\end{bmatrix} = \begin{bmatrix}3.13\\-1.44\end{bmatrix}$

Minimize $f(\lambda) = f\left\{\begin{bmatrix}3.13\\1.56\end{bmatrix} + \lambda\begin{bmatrix}3.13\\-1.44\end{bmatrix}\right\} \Rightarrow \lambda^* = -0.10$

Therefore $\boldsymbol{y^1} = \boldsymbol{x^2} + \lambda^* \boldsymbol{d} = \begin{bmatrix}3.13\\1.56\end{bmatrix} - (0.10)\begin{bmatrix}3.13\\-1.44\end{bmatrix} = \begin{bmatrix}2.82\\1.70\end{bmatrix}$

ITERATION 2

STEP 1 $\boldsymbol{y^1} = \begin{bmatrix}2.82\\1.70\end{bmatrix}$, Min $f\left\{\begin{bmatrix}2.82\\1.70\end{bmatrix} + \lambda\begin{bmatrix}1\\0\end{bmatrix}\right\} = f(2.82+\lambda, 1.7)$

This yields $\lambda_1 = -0.12$

Therefore $\boldsymbol{y^2} = \boldsymbol{y^1} + \lambda_1 \boldsymbol{d^1} = \begin{bmatrix}2.70\\1.70\end{bmatrix}$.

Minimize $f\left\{\begin{bmatrix}2.70\\1.70\end{bmatrix} + \lambda\begin{bmatrix}0\\1\end{bmatrix}\right\}$

$\Rightarrow$ Min $f(2.7, 1.7+\lambda)$ $\Rightarrow \lambda_2 = -0.35$,

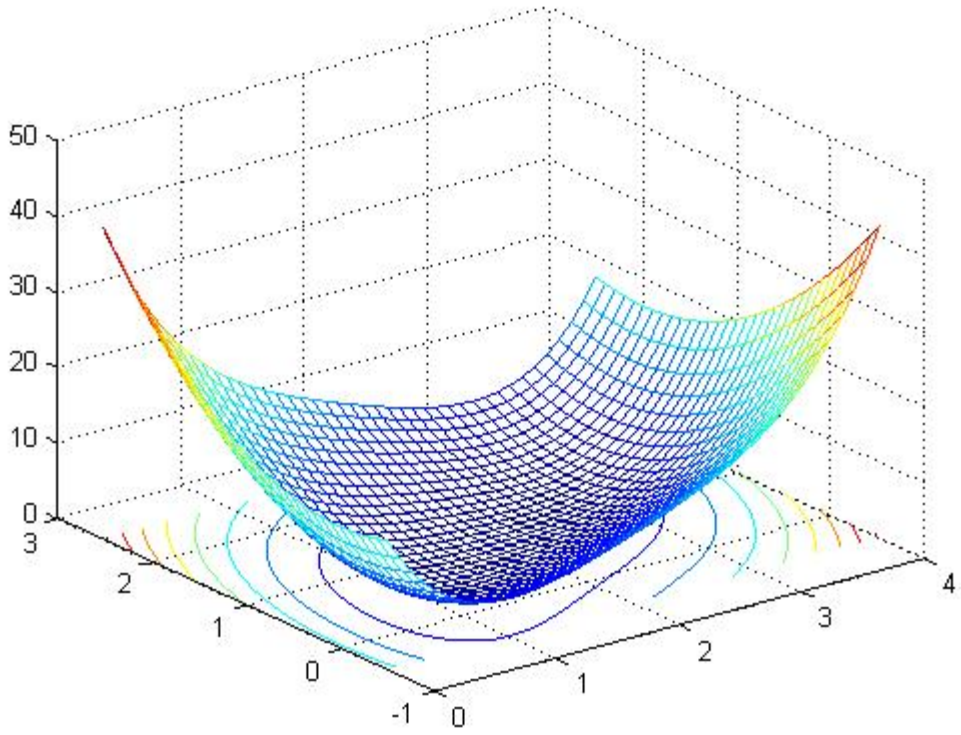
Therefore $\boldsymbol{y^3} = \boldsymbol{y^2} + \lambda_2 \boldsymbol{d^2} = \begin{bmatrix}2.70\\1.35\end{bmatrix}$.

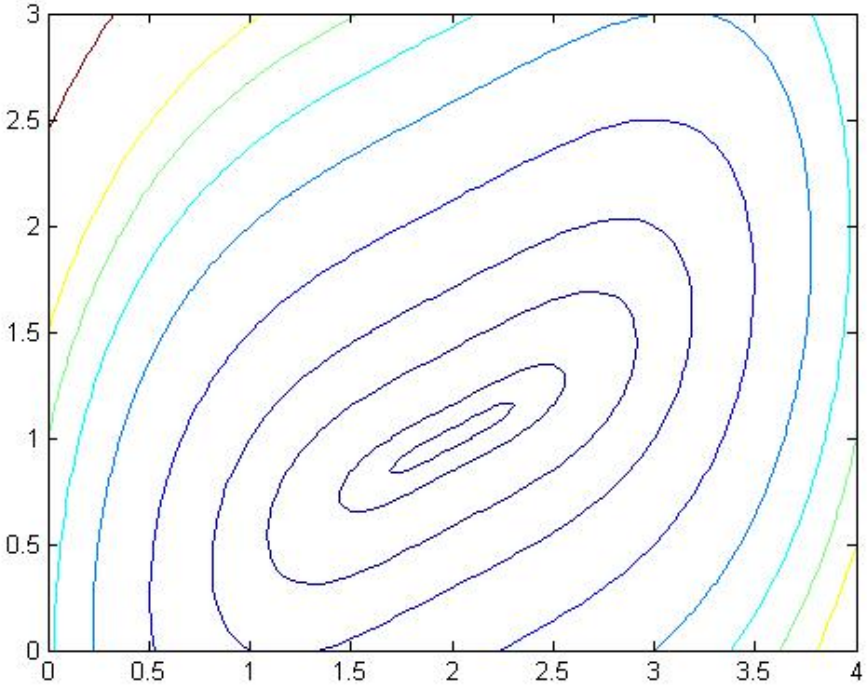
Let $\boldsymbol{x^3} = \begin{bmatrix}2.70\\1.35\end{bmatrix}$, $\|\boldsymbol{x}^3 - \boldsymbol{x}^2\|$ is not negligible; go to Step 2.

STEP 2

$\boldsymbol{d} = \boldsymbol{x^3} - \boldsymbol{x^2} = \begin{bmatrix}2.70\\1.35\end{bmatrix} - \begin{bmatrix}3.13\\1.56\end{bmatrix} = \begin{bmatrix}-0.43\\-0.21\end{bmatrix}$

Minimize $f(\lambda) = f\left\{\begin{bmatrix}2.70\\1.35\end{bmatrix} + \lambda\begin{bmatrix}-0.43\\-0.21\end{bmatrix}\right\} \Rightarrow \lambda^* = 1.50$


50
40
30
20
10
0
3
2
1
0
-1
0
1
2
3
4


3
2.5
2
1.5
1
0.5
0
0
0.5
1
1.5
2
2.5
3
3.5
4

Therefore $\boldsymbol{y^1} = \boldsymbol{x^3} + \lambda^* \boldsymbol{d} = \begin{bmatrix} 2.70 \\ 1.35 \end{bmatrix} + (1.50) \begin{bmatrix} -0.43 \\ -0.21 \end{bmatrix} = \begin{bmatrix} 2.06 \\ 1.04 \end{bmatrix}$

ITERATION 3,

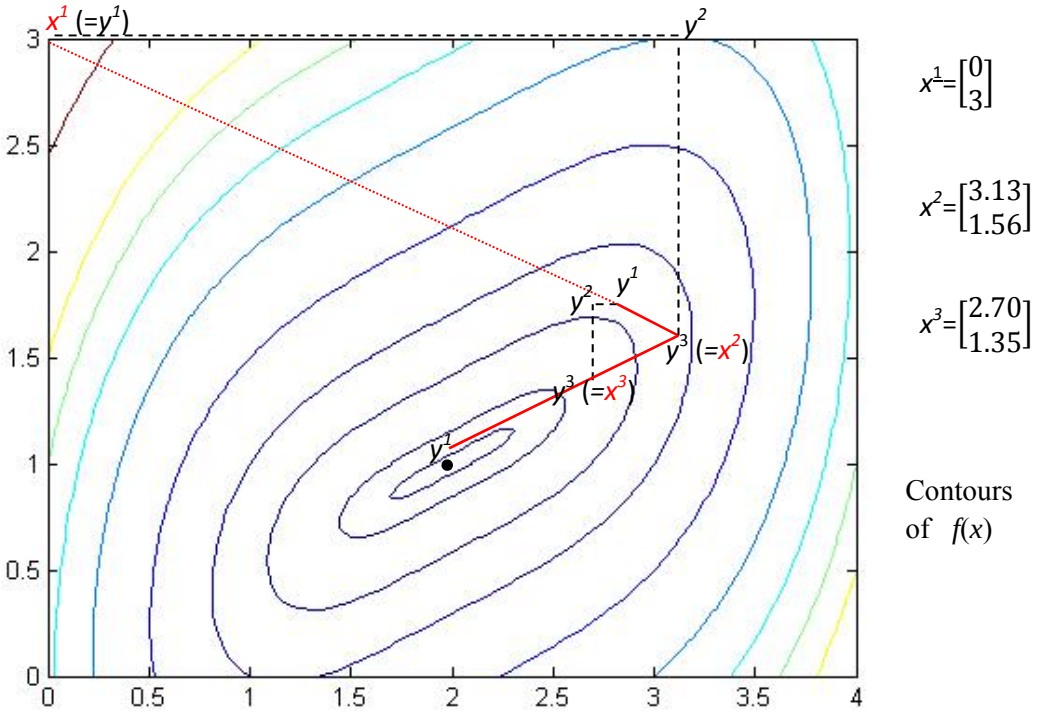
STEP 1 $\boldsymbol{y^1} = \begin{bmatrix} 2.06 \\ 1.04 \end{bmatrix}$, Min $f\left\{\begin{bmatrix} 2.06 \\ 1.04 \end{bmatrix} + \lambda \begin{bmatrix} 1 \\ 0 \end{bmatrix}\right\}$ ETC. ETC.

Note that the optimal solution to the problem is, $\boldsymbol{x}^* = [x_1 = 2,\ x_2 = 1]$, and we're headed that way...

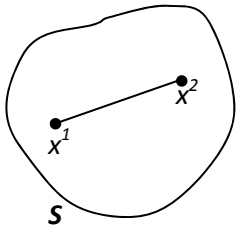
<u>GEOMETRICAL INTERPRETATION</u>

STEP 1: Exploratory search to get $\boldsymbol{x^{i+1}}$ (via $\boldsymbol{y^1}, \boldsymbol{y^2}$) ----------

STEP 2: Line search along $\boldsymbol{x^{i+1}} - \boldsymbol{x^i}$ (to $\boldsymbol{y^1}, \boldsymbol{y^2}$ for the next iteration) ____



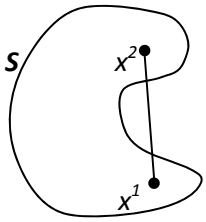
CONVEXITY



A set $\boldsymbol{S}$ is said to be CONVEX if for any two points $\boldsymbol{x^1}$ and $\boldsymbol{x^2}$ in $\boldsymbol{S}$, the line segment joining $\boldsymbol{x^1}$and $\boldsymbol{x^2}$ also lies entirely in the set $\boldsymbol{S}$.

i.e., If $\boldsymbol{x^1} \in \boldsymbol{S}$, $\boldsymbol{x^2} \in \boldsymbol{S} \Rightarrow \lambda\, \boldsymbol{x^1} + (1-\lambda)\boldsymbol{x^2} \in \boldsymbol{S}$ for all $\lambda \in [0,1]$, then $\boldsymbol{S}$ is a convex set.

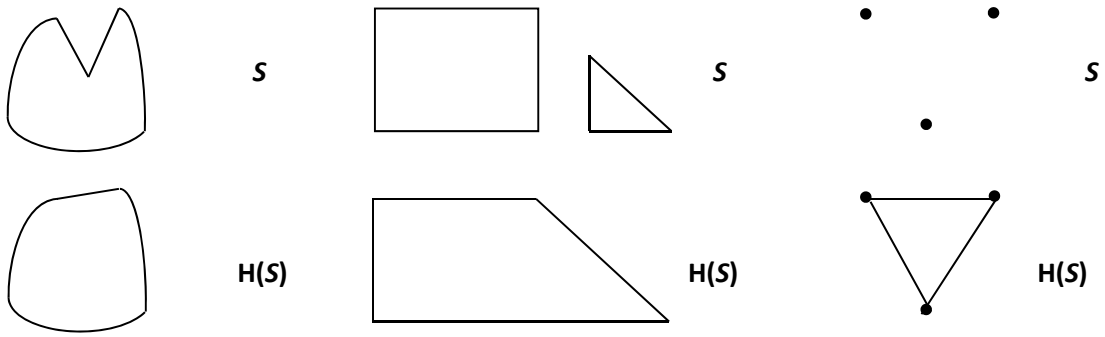
In general, $\lambda_1\, \boldsymbol{x^1} + \lambda_2 \boldsymbol{x^2}$ is said to be a convex combination of $\boldsymbol{x^1}$ and $\boldsymbol{x^2}$ if $\lambda_1, \lambda_2 \geq \boldsymbol{0}$ and $\lambda_1 + \lambda_2 = 1$.



This set $\boldsymbol{S}$ is a NONCONVEX set since the line segment joining $\boldsymbol{x^1}$ and $\boldsymbol{x^2}$ is not entirely in the set $\boldsymbol{S}$.

CONVEX HULL: The convex hull of the set $\boldsymbol{S}$ denoted by $\boldsymbol{H(S)}$ 1is the smallest convex set that contains $\boldsymbol{S}$. It is the collection of all convex combinations of points in $\boldsymbol{S}$.

e.g.

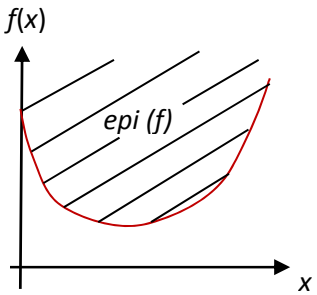


CONVEX FUNCTIONS

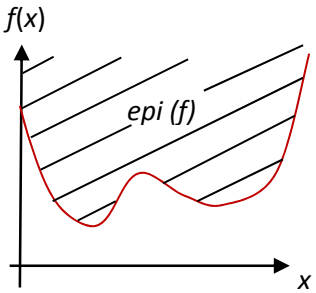
The **EPIGRAPH** of a function f - denoted by $epi(f)$ - is the set of points on or above the graph of $f(x)$.

The **HYPOGRAPH** of a function f - denoted by $hyp(f)$ - is the set of points on or below the graph of $f(x)$.

A function f is convex if, and only if, the epigraph of f is a convex set.



convex

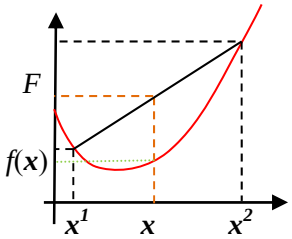


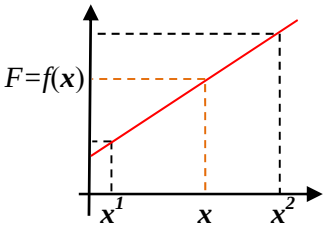
nonconvex

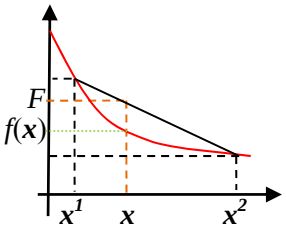
Equivalently, for $\lambda \in [0,1]$

$$\overbrace{(1-\lambda)f(\mathbf{x}^1) + \lambda f(\mathbf{x}^2)}^{F} \geq f(\overbrace{(1-\lambda)\mathbf{x}^1 + \lambda \mathbf{x}^2}^{\mathbf{x}})$$

if, and only if, f is convex.







Some Convex Functions

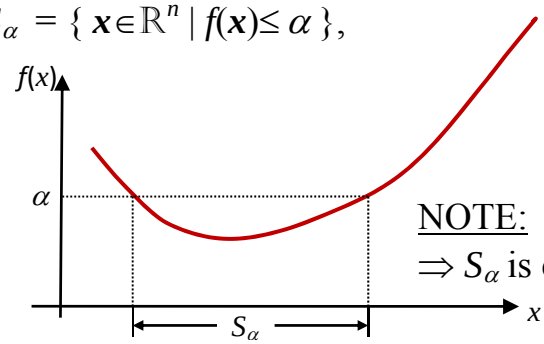
SOME TERMINOLOGY AND FURTHER EXTENSIONS...

1) EPIGRAPH: $EPI(f) = \{ (\boldsymbol{x},y) \mid \boldsymbol{x}\in\mathbb{R}^n, y\in\mathbb{R}, y\geq f(\boldsymbol{x}) \} \subseteq \mathbb{R}^{n+1}$,

2) HYPOGRAPH: $HYP(f) = \{ (\boldsymbol{x},y) \mid \boldsymbol{x}\in\mathbb{R}^n, y\in\mathbb{R}, y\leq f(\boldsymbol{x}) \} \subseteq \mathbb{R}^{n+1}$

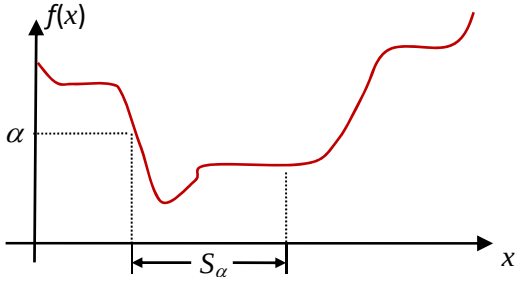
3) LEVEL SET: $S_\alpha = \{ \boldsymbol{x}\in\mathbb{R}^n \mid f(\boldsymbol{x})\leq \alpha \}$,

e.g.



NOTE: f: $\mathbb{R}^n\rightarrow\mathbb{R}$ convex $\Rightarrow S_\alpha$ is convex for all α.

4) QUASICONVEXITY: Given $\lambda\in[0,1]$, $\boldsymbol{x}^1\neq\boldsymbol{x}^2$, the function f is said to be **quasiconvex** if $f(\lambda\boldsymbol{x}^1+(1-\lambda)\,\boldsymbol{x}^2) \leq \text{Max}\{f(\boldsymbol{x}^1), f(\boldsymbol{x}^2)\}$. Note: ($f(\boldsymbol{x})$ quasiconvex $\Leftrightarrow$ level sets are **all** convex).



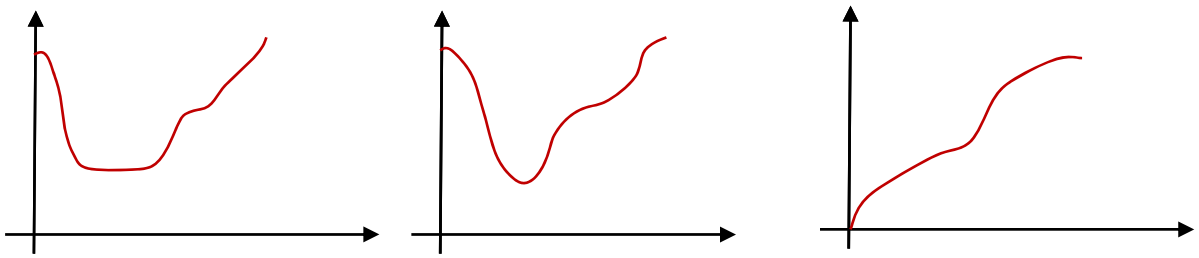
Quasiconvex, but not convex

4a) Given $\lambda\in[0,1]$, $\boldsymbol{x}^1\neq\boldsymbol{x}^2$, the function f is said to be **strongly quasiconvex** if

$f(\lambda\boldsymbol{x}^1+(1-\lambda)\,\boldsymbol{x}^2) < \text{Max}\{f(\boldsymbol{x}^1), f(\boldsymbol{x}^2)\}$

4b) Given $\lambda\in[0,1]$, $\boldsymbol{x}^1\neq\boldsymbol{x}^2$, $f(\boldsymbol{x}^1)\neq f(\boldsymbol{x}^2)$, the function f is said to be **strictly quasiconvex** if $f(\lambda\boldsymbol{x}^1+(1-\lambda)\,\boldsymbol{x}^2) < \text{Max}\{f(\boldsymbol{x}^1), f(\boldsymbol{x}^2)\}$

A local minimum for a strictly quasiconvex function is also a global one, and a unique global minimum for a strongly quasiconvex function.

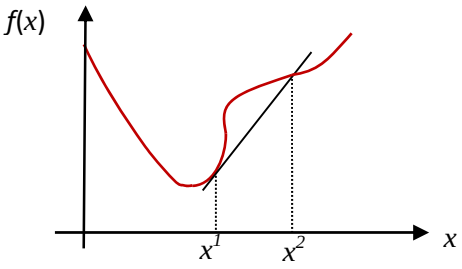


Strict (but not Strong) Strong as well as strict quasimonotone

4c) A function f is **quasiconcave** if $-f$ is quasiconvex

4d) A function that is both quasiconvex as well as quasiconcave is said to be **quasimonotone**.

6) PSEUDOCONVEXITY: f is said to be pseudoconvex on its domain if it is differentiable everywhere and $\nabla \boldsymbol{f}^{T}(\boldsymbol{x^1})\,(\boldsymbol{x^2}-\boldsymbol{x^1}) \geq 0 \Rightarrow f(\boldsymbol{x^2}) \geq f(\boldsymbol{x^1})$, where $\nabla \boldsymbol{f}$ is the gradient vector of f (i.e., $\nabla f_i = \partial f/\partial x_i$)



$$\nabla \boldsymbol{f}(x^1) \equiv \frac{f(x^2) - f(x^1)}{x^2 - x^1}$$

Note that if the gradient vector of a pseudoconvex function is equal to zero at some point, the point must be a (global) minimizer.

THEOREM: Suppose S is a nonempty open convex set in $\mathbb{R}^n$ and let $f: S \to \mathbb{R}$ be differentiable on S. Then f is convex if, and only if, for any $\boldsymbol{x}' \in S$, we have

$$f(\boldsymbol{x}) \geq f(\boldsymbol{x}') + \nabla \boldsymbol{f}^{\mathrm{T}}(\boldsymbol{x}')\,(\boldsymbol{x} - \boldsymbol{x}') \quad \text{for each } \boldsymbol{x} \in S$$

Also f is strictly convex if, and only if, for any $\boldsymbol{x}^* \in S$, we have

$$f(\boldsymbol{x}) > f(\boldsymbol{x}') + \nabla \boldsymbol{f}^{\mathrm{T}}(\boldsymbol{x}')\,(\boldsymbol{x} - \boldsymbol{x}') \quad \text{for each } \boldsymbol{x} \in S$$

Note that if we are minimizing a convex f over $\boldsymbol{x} \in S$ then given any point $\boldsymbol{x}' \in S$, the affine function $f(\boldsymbol{x}') + \nabla \boldsymbol{f}^{\mathrm{T}}(\boldsymbol{x}')\,(\boldsymbol{x} - \boldsymbol{x}')$ bounds f from below. So the minimum of this affine function over $\boldsymbol{x} \in S$ yields a lower bound on the optimum value of f. This fact is often used in optimization algorithms.

POSITIVE (NEGATIVE) SEMIDEFINITE (DEFINITE) MATRICES

A matrix $\boldsymbol{H}$ is said to be POSITIVE SEMIDEFINITE at $\boldsymbol{x}' \in S$ if $\boldsymbol{x}^{\mathrm{T}} \boldsymbol{H}(\boldsymbol{x}') \boldsymbol{x} \geq 0$.

Similarly, $\boldsymbol{H}$ is said to be NEGATIVE SEMIDEFINITE if $\boldsymbol{x}^{\mathrm{T}} \boldsymbol{H}(\boldsymbol{x}') \boldsymbol{x} \leq 0$.

A matrix $\boldsymbol{H}$ is said to be POSITIVE DEFINITE at $\boldsymbol{x}' \in S$ if $\boldsymbol{x}^{\mathrm{T}} \boldsymbol{H}(\boldsymbol{x}') \boldsymbol{x} > 0$ for all $\boldsymbol{x} \neq \boldsymbol{0}$, and NEGATIVE SEMIDEFINITE if $\boldsymbol{x}^{\mathrm{T}} \boldsymbol{H}(\boldsymbol{x}') \boldsymbol{x} < 0$ for all $\boldsymbol{x} \neq \boldsymbol{0}$.

If the above conditions cannot be satisfied the matrix $\boldsymbol{H}$ is said to be INDEFINITE. (Also, note here that if all entries in $\boldsymbol{H}$ are ≥ 0, it does not imply **anything** -- this is not a sufficient condition for $\boldsymbol{H}$ to be positive semidefinite).

Examples:

- $\boldsymbol{H} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \implies \boldsymbol{x}^T\boldsymbol{H}\boldsymbol{x} = [x_1 \;\; x_2]\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \; 2x_1^2 + 2x_2^2$

- > 0 for all $\boldsymbol{x} \neq \boldsymbol{0}$

  Thus $\boldsymbol{H}$ is **positive definite**.

- $\boldsymbol{H} = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix} \implies \boldsymbol{x}^T\boldsymbol{H}\boldsymbol{x} = [x_1 \;\; x_2]\begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 2(x_1 + x_1)^2$

- ≥ 0 for all $\boldsymbol{x} \neq \boldsymbol{0}$

  Thus $\boldsymbol{H}$ is **positive semidefinite**,

- $\boldsymbol{H} = \begin{bmatrix} 2 & 0 \\ 0 & -2 \end{bmatrix} \implies \boldsymbol{x}^T\boldsymbol{H}\boldsymbol{x} = [x_1 \;\; x_2]\begin{bmatrix} 2 & 0 \\ 0 & -2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} =$

  $2x_1^2 - 2x_2^2$ may be of any sign.

  Thus $\boldsymbol{H}$ is **indefinite**.

- $\boldsymbol{H}(\boldsymbol{x}') = \begin{bmatrix} 6(x_1')^2 + 2 & 0 \\ 0 & 12(x_2')^2 + 1 \end{bmatrix} \implies$

  $\boldsymbol{x}^T(\boldsymbol{x}')\boldsymbol{x} = [x_1 \;\; x_2]\begin{bmatrix} 6(x_1')^2 + 2 & 0 \\ 0 & 12(x_2')^2 + 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

  $= \; x_1^2\{6(x_1')^2 + 2\} + x_2^2\,\{12(x_2')^2 + 1\} > 0$ for all $\boldsymbol{x} \neq \boldsymbol{0}$

  Thus $\boldsymbol{H}$ is **positive definite** everywhere.

TWICE DIFFERENTIABLE CONVEX FUNCTIONS

Suppose that f has 2nd partial derivatives. Then $\nabla^2 \boldsymbol{f(x)} = \boldsymbol{H}(\boldsymbol{x})$ is called the HESSIAN matrix of f, where $H_{ij}(\boldsymbol{x}) = [\nabla^2 f(\boldsymbol{x})]_{ij} = \dfrac{\partial^2 f(\boldsymbol{x})}{\partial x_i \partial x_j}$.

QUADRATIC FORM: The quadratic form is a convenient way of representing quadratic functions (polynomials of order no higher than 2). Specifically it has the form $f(\boldsymbol{x}) = \frac{1}{2}\, \boldsymbol{x}^{\mathrm{T}} \boldsymbol{H} \boldsymbol{x} + \boldsymbol{b}^{\mathrm{T}} \boldsymbol{x}$, where $\boldsymbol{H}=\nabla^2 f(\boldsymbol{x})$ is the **Hessian matrix** of the function $f(\boldsymbol{x})$ with $H_{ij} = \dfrac{\partial^2 f(\boldsymbol{x})}{\partial x_i \partial x_j}$.

Example 1: $f(x_1, x_2) = ax_1^2 + bx_2^2 + cx_1x_2$

$$\Longrightarrow f(\boldsymbol{x}) = \frac{1}{2}[x_1\ x_2]\begin{bmatrix} 2a & c \\ c & 2b \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

Example 2: $f(x_1, x_2) = ax_1^2 + bx_2^2 + cx_3^2 + kx_1x_2 + lx_1x_3 + mx_2x_3 + p$

$$\Longrightarrow f(\boldsymbol{x}) = \frac{1}{2}[x_1\ \ x_2\ \ x_3]\begin{bmatrix} 2a & k & l \\ k & 2b & m \\ l & m & 2c \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + p$$

THEOREM

a) f: $S \subseteq \mathbb{R} \to \mathbb{R}$; f convex $\Leftrightarrow \dfrac{d^2x}{dx^2} \geq 0\ \forall\ x \in S$

b) f: $S \subseteq \mathbb{R}^n \to \mathbb{R}$; f convex $\Leftrightarrow \boldsymbol{\nabla^2 f(x)}$ is POSITIVE SEMIDEFINITE $\forall\ \boldsymbol{x} \in S$

c) f: $S \subseteq \mathbb{R}^n \to \mathbb{R}$; $\boldsymbol{\nabla^2 f(x)}$ is POSITIVE DEFINITE $\forall\ \boldsymbol{x} \in S \Rightarrow$ f strictly convex

f strictly convex $\Rightarrow$ $\boldsymbol{\nabla^2 f(x)}$ is POSITIVE SEMIDEFINITE $\forall\ \boldsymbol{x} \in S$

EIGENVALUES AND EIGENVECTORS

Given an n row by n column ($n\times n$) symmetric matrix $\boldsymbol{A}$, if we have a scalar λ and a nonzero vector $\boldsymbol{v}\in\mathbb{R}^n$ satisfying $\boldsymbol{Av}=\lambda\boldsymbol{v}$, i.e., $[\boldsymbol{A}-\boldsymbol{I}\lambda]\boldsymbol{v}=0$, then $\boldsymbol{v}$ is called an *eigenvector* for $\boldsymbol{A}$, and λ is called the corresponding *eigenvalue* for $\boldsymbol{A}$.

To compute the eigenvalues, we solve

$$\text{Det}(\boldsymbol{A}-\lambda\boldsymbol{I}) = |\boldsymbol{A}-\lambda\boldsymbol{I}| = 0 \text{ for } \lambda$$

Then for each λ_i (in general we have n of them) we solve

$$(\boldsymbol{A}-\lambda_i\boldsymbol{I})\boldsymbol{v} = 0 \text{ to find the eigenvector } \boldsymbol{v}$$

THEOREM: A matrix is positive (negative) semidefinite if all its eigenvalues are nonnegative (nonpositive). A matrix is positive (negative) definite if all its eigenvalues are positive (negative).

THEOREM: A real $n\times n$ symmetric matrix has n (not necessarily distinct) eigenvalues, and at most n linearly independent eigenvectors.

Example: $\boldsymbol{A}=\begin{bmatrix}2 & 2\\ 2 & 2\end{bmatrix} \implies [\boldsymbol{A}-\lambda\boldsymbol{I}] = \begin{bmatrix}2-\lambda & 2\\ 2 & 2-\lambda\end{bmatrix}$

Then $|\boldsymbol{A}-\lambda\boldsymbol{I}| = (2-\lambda)^2 - 4 = 0 \implies \lambda^2 - 4\lambda = 0$.

$\implies \lambda(\lambda-4) = 0 \implies$ the eigenvalues are $\lambda_1=4$, $\lambda_2=0$.

(Note that these being nonnegative, A is positive semidefinite). Then the eigenvectors are $\boldsymbol{v}^{\mathrm{T}} = [v_1 \;\; v_2]$, where $\begin{bmatrix}2-\lambda & 2\\ 2 & 2-\lambda\end{bmatrix}\begin{bmatrix}v_1\\ v_2\end{bmatrix} = 0$

i.e., $(2-\lambda)\boldsymbol{v}_1 + 2\boldsymbol{v}_2 = 0, \quad 2\boldsymbol{v}_1 + (2-\lambda)\boldsymbol{v}_2 = 0$

$$\text{i.e., } \boldsymbol{v^1} = \begin{bmatrix} t \\ t \end{bmatrix} \text{ for } \lambda_1 = 4, \qquad \boldsymbol{v^2} = \begin{bmatrix} -t \\ t \end{bmatrix} \text{ for } \lambda_2 = 0,$$

where *t* is any arbitrary real number.

THEOREM: Two eigenvectors of a real *symmetric* matrix corresponding to different eigenvalues are orthogonal, i.e., if $\lambda_1 \neq \lambda_2$ then $(\boldsymbol{v}^1)^T\boldsymbol{v}^2 = 0$.

Further, the eigenvalues of such a matrix are always real.

DEFINITION: A principal minor is the determinant of a (square) sub-matrix obtained by deleting from a (square) matrix, a row and a column (or multiple rows and columns) that share the same index. The *leading* principal minors of a square matrix are the determinants of the square sub-matrices along the main diagonal.

THEOREM: A real symmetric matrix has eigenvalues that are positive if, and only if, all the *leading* principal minors are positive. If the eigenvalues are all nonnegative then the leading principal minors are all nonnegative, but for the eigenvalues to be nonnegative, the determinants of ALL possible principal minors must be nonnegative.

The above theorem indicates that if we have $H_{ii} \leq 0$ ($H_{ii} < 0$) for some i , then $\boldsymbol{H}$ can never be positive definite (semidefinite).

THEOREM: A real symmetric matrix $\boldsymbol{A}$ has eigenvalues that are negative if, and only if, $(-1)^i|\boldsymbol{A_i}|$ (where $|\boldsymbol{A_i}|$ is the i^{th} leading principal minor) is positive for i=1,..,n. If the eigenvalues are nonpositive then the corresponding $(-1)^i|\boldsymbol{A_i}|$ are all nonnegative, but for the eigenvalues to be nonpositive, $(-1)^i|\boldsymbol{A_i}|$ must be nonnegative for ALL possible principal minors of order i.

The last two theorems provide another way to check the definiteness of a matrix (and thus to check for convexity...).

EXAMPLE 1: $f(x_1, x_2) = -x_1^2 - 5x_2^2 + 2x_1x_2 + 10x_1 - 10x_2$

$$\boldsymbol{\nabla f} = \begin{bmatrix} -2x_1 + 2x_2 + 10 \\ -10x_2 + 2x_1 - 10 \end{bmatrix} \quad \& \quad \boldsymbol{\nabla^2 f} = \boldsymbol{H} = \begin{bmatrix} -2 & 2 \\ 2 & -10 \end{bmatrix}$$

$|\boldsymbol{H}-\lambda\boldsymbol{I}| = 0 \Rightarrow (-2-\lambda)(-10-\lambda) - 4 = 0 \Rightarrow \lambda = -6 \pm 2\sqrt{5}$

Since $\lambda_1 = -6-2\sqrt{5} < 0$ and $\lambda_2 = -6+2\sqrt{5} < 0$ both eigenvalues are negative, and therefore the Hessian $\boldsymbol{H}$ is negative definite, implying f is strictly concave.

EXAMPLE 2: $f(\boldsymbol{x}) = x_1 e^{-(x_1+x_2)}$. Here $\boldsymbol{H} = e^{-(x_1+x_2)} \begin{bmatrix} (x_1-2) & (x_1-1) \\ (x_1-1) & x_1 \end{bmatrix}$

Let us call $e^{-(x_1+x_2)} = Q$. then $|\boldsymbol{H}-\lambda\boldsymbol{I}| = 0 \Rightarrow \{Q(x_1-2)-\lambda\}(Qx_1-\lambda) - Q^2(x_1-1)^2 = 0$

Solving for λ, we get (VERIFY…),

$$\lambda_1 = e^{-(x_1+x_2)}\left\{(x_1-1) + \sqrt{(x_1-1)^2+1}\right\} \geq 0$$

$$\lambda_2 = e^{-(x_1+x_2)}\left\{(x_1-1) - \sqrt{(x_1-1)^2+1}\right\} \leq 0$$

Therefore $\boldsymbol{H}$ is indefinite and f is neither convex nor concave.