INTRODUCTION TO NONLINEAR OPTIMIZATION[1]-

Nonlinear optimization is concerned with methods for finding the minimum or maximum value of nonlinear function F(**x**) of any number of independent variables $x_1, x_2, \ldots, x_n$. The basic problem is *to find vector x* that minimizes* F(**x**) where F(**x**) is called the *Objective Function* or *Performance Index*. The variables $x_1, x_2, \ldots, x_n$ are called *Decision Variables*. The value F(**x***) is called *minimum* of the problem while **x*** is called a *minimizer*.

Unconstrained optimization: there are no special conditions that the independent variables are allowed to be subject to.

Constrained optimization: there are special conditions that the independent variables are allowed to be subject to:

- *Equality* constraints: $c_i(x)=0$, i=1,2,…,m
- *Inequality* constraints: $c_i(x) \geq 0$, i=1,2,…, k

Note: The set of all minimizers of F(x) is denoted by *argmin* F(x) with $x \in S$. Note: *Feasible region* or *feasible set* is a subset **S** of the set of real numbers, **R**.

Note. The maximization of F(x)=minimization (-F(**x**)):

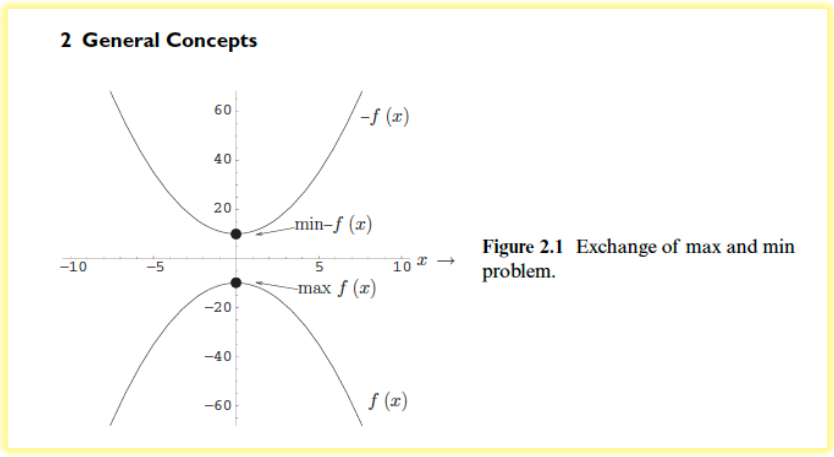

Figure 2.1 Exchange of max and min problem.

[1] Scales, L.E., *Introduction to Non-linear Optimization*, Springer-Verlag, New York, NY, 1985.

[1] Gill, P.E., W. Murray, and M. H. Wright, *Practical Optimization*, Academic Press, London, UK, 1981.

[1] Edgar, T.F. and D.M. Himmelblau, *Optimization of Chemical Processes*, 2nd ed McGraw-Hill, New York, 2001

E-H 7

[1] Rao, S.S., *Engineering Optimization*, 5th ed, Wiley, Hoboken, NJ, 2020.

[1]Theodore, L., K. Behan, Introduction to Optimization for Chemical and Environmental Engineers, CRC Press, Boca Raton, FL, 2018.

First, we discuss Unconstrained optimization (minimization). In this case **S**=**R**
For example, consider the objective $F(x_1,x_2)=(x_1-1)^2+(x_2-3)^2$ defined $\forall x \in R$. Then min F(x)=0 at $x_1^*=1$ and $x_2^*=3$ or *argmin* $F(x)=(1, 3)^T$.

Solution methods. There are two categories of methods to solve the problem:

1. *Gradient methods* (require derivatives of the objective functions). It is assumed that the objective function has continuous second derivatives, whether or not these are explicitly available. *Gradient methods* are still efficient if there are some discontinuities in the derivatives.
2. *Direct search methods* (derivative-free). *These methods* use function values and are more efficient for highly discontinuous functions.

Applications: Nonlinear Parameter Estimation or Nonlinear Regression.

Suppose that we have a mathematical model describing of physical / chemical / biological process. A special class of optimization problems arises when the objective function is a suitable measure of the overall departure of the model calculated values from the experimental measurements obtained from the process. The objective function is of the form[2]

$$S(\mathbf{x}) = \sum_{i=1}^{N} \boldsymbol{e}_i^T \mathbf{Q}_i \mathbf{e}_i$$

This is the *parameter estimation* or *nonlinear regression problem*. Vector $\mathbf{e}_i$ is the vector of residuals from the i^{th} experiment and $\mathbf{Q}_i$ is a user supplied square weighting matrix. $\mathbf{e}_i=\hat{y}_i - f(\boldsymbol{x}_i, \boldsymbol{k})$. If Q=I then we minimize the sum of squares of errors (SSE) and we have the *least squares* (LS) estimation problem. The objective function is given by

$$S(\mathbf{x}) = \sum_{i=1}^{N} \boldsymbol{e}_i^T \mathbf{e}_i$$

The mathematical model describing the process can be linear (w.r.t. parameters) or nonlinear algebraic equations, ordinary differential equation (ODE), or Partial Differential Equation (PDEs).

---

[2] Englezos, P., and N.E. Kalogerakis, *Applied Parameter Estimation for Chemical Engineers*, Marcel-Dekker, 2001

BASIC CONCEPTS (Gradient vector, Hessian Matrix, Jacobian matrix)

Let F(**x**) be a typical function where **x** is the real n-dimensional vector $(x_1, x_2,...,x_n)^T$. **Vectors** are denoted by using lower case **boldface** letters. Vectors are considered column vectors

$$\mathbf{x}=\begin{pmatrix} x_1 \\ x_2 \\ . \\ . \\ . \\ x_n \end{pmatrix}=(x_1, x_2, \ldots, x_n)^T$$

The gradient vector g(x) and the Hessian matrix G(x) are given by

$$\nabla F(\mathbf{x}) \equiv \frac{\partial F}{\partial \mathbf{x}} = g(\mathbf{x}) = \left[\frac{\partial F}{\partial x_1}, \frac{\partial F}{\partial x_2}, \ldots, \frac{\partial F}{\partial x_n}\right]^T$$

$$G(x) \equiv \nabla^2 F(x) = \begin{bmatrix} \frac{\partial^2 F}{\partial x_1^2}, \frac{\partial^2 F}{\partial x_1 \partial x_2}, \ldots, \frac{\partial^2 F}{\partial x_1 \partial x_n} \\ \frac{\partial^2 F}{\partial x_2 \partial x_1}, \frac{\partial^2 F}{\partial x_2^2}, \ldots, \frac{\partial^2 F}{\partial x_2 \partial x_n} \\ \ldots \\ \ldots \\ \frac{\partial^2 F}{\partial x_n \partial x_1}, \frac{\partial^2 F}{\partial x_n \partial x_2}, \ldots, \frac{\partial^2 F}{\partial x_n^2} \\ \end{bmatrix}$$

Vector sizes[3]. Given an n-dimensional vector the **norms of vector** (infinity norm, Norm-1, Euclidian or Norm-2 and the Norm-k) offer a measure of the "length" of the vector.

The infinity norm is $\|x\|_\infty = \max(x_i)$
Norm-1 is $\|x\|_{k1} = [\sum_{i=1}^{n} |x_i|]$
Norm-k is $\|x\|_k = [\sum_{i=1}^{n} |x_i|^k]^{1/k}$
The Euclidean norm is $\|\mathbf{x}\|_2 = \left(x_1^2 + x_2^2 + \ldots + x_n^2\right)^{1/2} = (\mathbf{x}^T\mathbf{x})^{1/2}$.

[3] Vasiliadis et al. Optimization for Chemical and Biochemical Engineering, CUP, 2020.

Jacobian Matrix.

Let $\mathbf{h}(\mathbf{x})=[h_1(\mathbf{x}), h_2(\mathbf{x}),\ldots,h_m(\mathbf{x})]$ (a vector-valued function) where $\mathbf{x}=[x_1,x_2,\ldots,x_n]^T$. Then the Jacobian matrix, is given by

$$J(\mathbf{x})=\frac{\partial \mathbf{h}}{\partial \mathbf{x}}=\begin{bmatrix} \frac{\partial h_1}{\partial x_1},\frac{\partial h_1}{\partial x_2},\ldots,\frac{\partial h_1}{\partial x_n} \\ \ldots. \\ \ldots. \\ \ldots \\ \frac{\partial h_m}{\partial x_1},\frac{\partial h_m}{\partial x_2},\ldots,\frac{\partial h_m}{\partial x_n} \end{bmatrix}$$

The Hessian matrix of a scalar function $F(\mathbf{x})$ is the Jacobian matrix of the vector function $\mathbf{g}(\mathbf{x})$, the gradient vector i.e.

$$\mathbf{g}(\mathbf{x})=\nabla F(\mathbf{x})=\left[\frac{\partial F}{\partial x_1}=h_1,\frac{\partial F}{\partial x_2}=h_2,\ldots,\frac{\partial F}{\partial x_n}=h_n\right]^T \text{ and}$$

$$J(\mathbf{x})=\partial \mathbf{g}/\partial \mathbf{x}=\nabla^2 F(\mathbf{x})=G(\mathbf{x})$$

NOTE: For continuously differentiable functions, H is symmetric

CONVEXITY AND CONCAVITY (multivariable functions)

Consider the multivariable function F(**x**) which has continuous second partial derivatives

- F(**x**) is *concave* if and only if the Hessian matrix, **G**(**x**), is negative semi-definite. For F(**x**) to be *strictly concave*, **G**(**x**) must be negative definite. **G**(**x**) is negative definite if and only if $\mathbf{x}^T\mathbf{G}\mathbf{x}$ is <0 for all $\mathbf{x} \neq \mathbf{0}$.
- F(**x**) is *convex* if **G**(**x**) is positive semidefinite. For F(**x**) to be *strictly convex*, **G**(**x**) must be positive definite. **G**(**x**) is positive definite if and only if $\mathbf{x}^T\mathbf{G}\mathbf{x}$ is >0 for all $\mathbf{x} \neq \mathbf{0}$.
- Note that **G**(**x**) is indefinite if and only if $\mathbf{x}^T\mathbf{G}\mathbf{x}$ is <0 for some **x** and >0 for other **x**.

$(1\times n)(n\times n)(n\times 1) : 1\times 1$

Note: It is noted that for *convexity* and *concavity*, the strict inequalities > or <, respectively, in the tests are replaced by ≥ or ≤, respectively.

Test for strict concavity: All the *eigenvalues* of **G**(**x**) are negative (<0).

Test for strict convexity: All the *eigenvalues* of **G**(x) are positive (>0).[5]

Note: If a function has a *stationary point* where the Hessian has eigenvalues of mixed signs, the function is neither convex nor concave.

[5] Bellman, R Introduction to Matrix Analysis, 2nd edition, SIAM< 1997

Eigenvalues and Singular Value decomposition (SVD)[6]:

Let **A** be a *square* (n x n) matrix. If there exists a scalar **λ** and a nonzero vector v such that $\mathbf{Av}=\lambda\mathbf{v}$

Then λ is called an *eigenvalue* of **A** and **v** is the corresponding *eigenvector.* All eigenvalues λ_i (some of which may be equal) can be obtained by solving the characteristic equation of **A** i.e. $\det(\mathbf{A}-\boldsymbol{\lambda}\mathbf{I})=0$.

If the matrix A is *normal* i.e. $\mathbf{A}\mathbf{A}^T=\mathbf{A}^T\mathbf{A}$, then it can be factorized into

$$\mathbf{A}=\mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T$$

Where **V** is an (n x n) orthogonal matrix and **Λ** is a diagonal matrix which has all the eigenvalues of **A** as its diagonal elements. We say that matrix A is orthogonally diagonizable and $\boldsymbol{\Lambda} = \mathbf{V}^T\mathbf{A}\mathbf{V}$.

- The set of all eigenvalues of a matrix is its *spectrum*.
- The *spectral radius* of matrix **A** is defined as $\boldsymbol{\rho}(\mathbf{A})=\max|\lambda_i(\mathbf{A})|$.
- If matrix A is *symmetric* ($\mathbf{A}=\mathbf{A}^T$) then all eigenvalues are real numbers.
- If matrix A is *non-singular* (its inverse $\mathbf{A}^{-1}$ exists) then all its eigenvalues are nonzero and the eigenvalues of $\mathbf{A}^{-1}$ are the reciprocals of the eigenvalues of **A**.

Note: A *square matrix* with *Determinant* equal to zero is a *singular matrix*.

[6] Cichoki A. and R. Unbehauen, *Neural Networks for Optimization and Signal Processing*, Wiley, 1993, New York, NY.

Singular Value Decomposition (SVD). Let **A** be an (m x n) matrix. Then real *orthogonal matrices* **U** (m x m) and **V** (n x n) exist such that

$$\mathbf{U}^T\mathbf{A}\mathbf{V}=\mathbf{S}$$

where **S** is a diagonal matrix *diag* ($\sigma_1, \sigma_2,....,\sigma_p$) and p=*min* (m,n). The real non-negative numbers σ_i are called *singular values* of A and matrix **A** can be written as $\mathbf{A}=\mathbf{U}\mathbf{S}\mathbf{V}^T$. It is noted that $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_p \geq 0.$

The singular values of matrix **A** are the square roots of the nonzero eigenvalues of $\mathbf{A}^T\mathbf{A}$ or $\mathbf{A}\mathbf{A}^T$

A *square* (n x n) matrix is called *orthogonal* if $\mathbf{Q}^T\mathbf{Q}=\mathbf{Q}\mathbf{Q}^T=\mathbf{I}$. For $m \neq n$, an (m x n) matrix is called orthogonal if $\mathbf{Q}^T\mathbf{Q}=\mathbf{I}$. All the diagonal elements of the identity matrix (**I**) are equal to 1 and the rest are zeros.

A specific and important case of the singular value decomposition (SVD) is obtained when **A** is a *symmetric* ($\mathbf{A}=\mathbf{A}^T$) nonnegative definite matrix. In this case the matrix **S** is $\mathbf{S}=diag\ (\lambda_1, \lambda_2,....,\lambda_p)$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0.$ are the real eigenvalues of **A**.

One important use of the SVD is the solution of a system of linear equations: $\mathbf{A}\mathbf{x}=\mathbf{b}$ where **A** is an (m x n) matrix, **x** an n-dimensional vector and **b** an m-dimensional vector. Substituting the SVD of **A** we obtain $\mathbf{U}\mathbf{S}\mathbf{V}^T\mathbf{x}=\mathbf{b}$. Then $\mathbf{x}=\mathbf{V}\mathbf{S}^{-1}\mathbf{U}^T\mathbf{b}$.

NECESSARY AND SUFFICIENT CONDISTIONS FOR AN EXTREMUM

(EH P. 135-142)

Now Consider a multivariable function F(**x**). If F(**x**) has continuous second partial derivatives,

The **Necessary** and **Sufficient Conditions** for an *extremum* (x*) of **the** unconstrained Function F(**x**) where $\mathbf{x}=(x_1, x_2,\ldots,x_n)^T$ are as follows[7,8,6]

Necessary Conditions:

1. F(**x**) is twice differentiable at **x***
2. $\nabla F(\mathbf{x}^*)=\mathbf{0}$, that is, a *stationary point* exists at **x***.

Sufficient Condition:

$H(\mathbf{x}^*)=\nabla^2 F(\mathbf{x}^*)$ is positive definite for a minimum to exist at **x*** or negative definite for a maximum to exist at **x***.

[7]Bertsekas, D.P. Nonlinear Programming, 3rd edition, 2016. Pages 6-12

[8]Gill, P.E., W. Murray, and M. H. Wright, *Practical Optimization*, Academic Press, London, UK, 1981 p. 59-82

[6]Scales, L.E., *Introduction to Non-linear Optimization*, Springer-Verlag, New York, NY, 1985. P 14-23