Process Engineering

I.

Process modelling and simulation

Endre Rév
associate professor
Department of Chemical and Environmental Process Engineering
Budapest University of Technology and Economics

Adapted from Hungarian by the author in January, 2013

Contents

Chapter 1

Flowsheeting and compound processes

1.1 Flowsheeting

Determining all the (originally missing) data of a *flowsheet* is called *flowsheeting*. This means first of all computation of steady state, detailed material and energy balance, all the flow rates, compositions, pressure and temperature data, and additional design parameters (e.g. reflux ratio, scale of a reactor, etc.) needed to achive some specified product quality or process constrains. This calculation is needed mainly because of two reasons:
1. When the process (the flow sheet) is designed, the unit operations are designed relatively independently, assuming some estimated properties of the streams arriving to the designed unit from other places of the process. These data should be brought to consistence before one may believe the data are acceptable.
2. One would like to know, at least by calculation, the missing, not measured data of an already working plant in order to fit a mathematical model to it before estimating the effects of some planned change in the process.

Solution approaches are demonstarted here on an arbitrary, and small, example flowsheet shown in **Fig. 1.1**. It consists of a unifier (mixer) (U), a reactor (R), and a vapor-liquid phase equilibrium component separator (S). Some reagent A is converted to B according to equation

$$A \xrightarrow{\zeta} 2B$$

where ζ is conversion factor. Let the feed be pure A. Both components occur at the exit of the reactor, stream (3). The product is separated from the reagent in unit (S). Raw materal A is more volatile than B, so that liquid stream (5) will contain the product, and vapor stream (5) is recycled to the unifier unit (U) to mix with

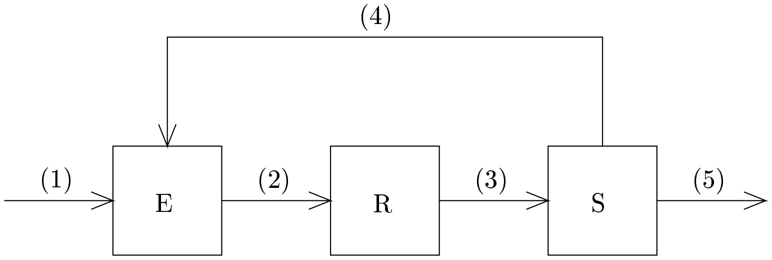

Figure 1.1: Simplified flowsheet

feed stream (1) and give together stream (2), feed of the reactor. Since component separation is not perfect, stream (4) will carry some of the product component B, and the product stream will be contaminated with reagent A, as well.

For simplity, only material balance is considered here. Stream component flow rates are considered as two-entry arrays of components A and B: $\boldsymbol{X}_i$ ($i \in 1, 2, 3, 4, 5$).

Problem:

Given feed stream $X_{1,A} = 100\,\mathrm{kmol/h}$, $X_{1,B} = 0\,\mathrm{kmol/h}$, reactor pressure $p_R = 4\,\mathrm{bar}$, reactor temperature $T_R = 320\,\mathrm{K}$, separation tank temperature $T_S = 345\,\mathrm{K}$, and recovery ratio of A into vapor phase $\eta_A = 0.98$, as well as physico-chemical models detailed below, compute all the component flow rates $X_{i,A}$, $X_{i,B}$, conversion factor ζ in the reactor, and the needed pressure p_S in the separation tank at steady state of the process!

Physico-chemical models:

This empirical model describes conversion in function of temperature T_R and pressure p_R in the reactor, and mole fractions x_A and x_B in the reactor's feed:

$$\zeta = 0.93 \exp\left(-0.76\frac{p_R}{T_R} - 0.22\frac{x_B}{x_A}\right)$$

Vapor-liquid equilibria in the separation tank is modelled with modified *Raoult-Dalton* equation

$$Py_A = \gamma_A x_A p_A^\circ$$
$$Py_B = \gamma_B x_B p_B^\circ$$

where x are mole fractions in liquid, y in vapor, vapor pressures p° of pure components at temperatures T are calculated with

$$\lg p_A^\circ = 2.033 - \frac{77.4246}{T - 230}$$

$$\lg p_B^\circ = 1.0044 - \frac{123.14}{T - 230}$$

and activity coefficients γ are modelled with

$$\lg \gamma_A = \frac{0.176}{\left(1 - \dfrac{x_A}{x_B}\right)^2}$$

$$\lg \gamma_B = \frac{0.176}{\left(1 - \dfrac{x_B}{x_A}\right)^2}$$

1.1.1 Equation solving approach

Collect all the equations describing the behaviour of the system and find its solution with any general mathematical method.

Here we provide such an equation system of the problem that its all unknown variables is expressed to left hand side of one of the equations, i.e. it is given in the form of $\boldsymbol{x} = \boldsymbol{g}(\boldsymbol{x})$:

1. $X_{2,A} = X_{1,A} + X_{4,A}$
2. $X_{2,B} = X_{1,B} + X_{4,B}$
3. $X_{3,A} = (1-\zeta)X_{2,A}$
4. $X_{3,B} = X_{2,B} + 2\zeta X_{2,A}$
5. $\zeta = 0.93 \exp\left(-0.76\dfrac{p_R}{T_R} - 0.22\dfrac{X_{2,B}}{X_{2,A}}\right)$
6. $X_{4,A} = X_{3,A} - X_{5,A}$
7. $X_{4,B} = X_{3,B} - X_{5,B}$
8. $X_{5,A} = (1-\eta_A)X_{3,A}$
9. $X_{5,B} = X_{5,A}\dfrac{p_A^\circ \gamma_A}{p_B^\circ \gamma_B}\dfrac{X_{4,B}}{X_{4,A}}$
10. $p_A^\circ = 10^{2.033 - \dfrac{77.4246}{T_S - 230}}$
11. $p_B^\circ = 10^{1.0044 - \dfrac{123.14}{T_S - 230}}$
12. $\gamma_A = 10^{\dfrac{0.176}{\left(1 - \dfrac{X_{5,A}}{X_{5,B}}\right)^2}}$
13. $\gamma_B = 10^{\dfrac{0.176}{\left(1 - \dfrac{X_{5,B}}{X_{5,A}}\right)^2}}$
14. $p_S = \dfrac{\gamma_A p_A^\circ X_{5,A} + \gamma_B p_B^\circ X_{5,B}}{X_{5,A} + X_{5,B}}$

All the necessary equations are listed here, and even more. For example, equation 14 can be omitted and applied after solving the remaining system of 1 to 13.

Table 1.1: Results with equation solving approach

n	$X_{2,A}$	$X_{2,B}$	$X_{3,A}$	$X_{3,B}$	$X_{4,A}$	$X_{4,B}$	$X_{5,A}$	$X_{5,B}$
0	140.00	240	54.33	470.28	53.24	233.05	1.000	199
5	151.76	249.42	66.96	483.98	56.45	232.22	1.150	200.58
10	166.22	230.71	66.61	434.33	65.88	250.95	1.340	187.35
15	162.58	248.21	63.54	443.58	62.54	247.89	1.277	201.65
20	162.64	244.72	63.87	442.11	62.68	243.08	1.280	198.76
25	162.57	243.52	63.72	441.61	62.48	243.47	1.275	197.79
30	162.40	243.42	63.66	441.08	62.39	243.44	1.273	197.72

Similarly, equations 10 and 11 could be applied once only, before all the others, since temperature of the separation tank is specified.

For solving the system, one has to provide estimation to all the unknowns. We list a possible estimation array for the component flow rates in the 0-th row of **Table 1.1**. Activity coefficients are estimated with 1.0, vapor pressures with 0.5 bar, and conversion with 0.8.

Using iterated substitution with 70 % damping, the calculation converged in 30 steps with stop criterion 0.05 % relative error limit as shown in **Table 1.1**. Calculated pressure of the separation tank is 1.071 bar.

Equation systems are ususally to be solved with applying some numerical procedure. Most of them involve repeatedly substituting the estimated values of unknowns into a function that calculates updated estimates. This series of estimation is expected to converge to the solution.

Basic methods are explained in *Section (3.1)*.

Although solving a large system of equations is a difficult task, modern techniques are avaliable to help us. The real burnen of equation solving approach is collecting all the equations ncessary to model the system, all in proper form, unfailed, and providing good initial estimates. Many unit operations are combined in a real process, some of them are compound units difficult to model in themselves.

1.1.2 Sequential modular approach

Sequential modular approach applies *input to output models* to the moduls (units) of the flowsheet, and tries to solve them in the sequence, i.e. along the paths, assigned by directions of the streams.

How it is done is shown using the small examle problem. First the input→output relations are collected:

Unifier module:

1. $X_{ki,A} = X_{be1,A} + X_{be2,A}$
2. $X_{ki,B} = X_{be1,B} + X_{be2,B}$

Reactor module:

1. $\dfrac{x_B}{x_A} = \dfrac{X_{be,B}}{X_{be,A}}$
2. $\zeta = 0.93 \exp\left(-0.76\dfrac{p_R}{T_R} - 0.22\dfrac{x_B}{x_A}\right)$
3. $X_{ki,A} = (1-\zeta)X_{be,A}$
4. $X_{ki,B} = X_{be,B} + 2\zeta X_{be,A}$

Separation tank:

This is an iterative procedure in itself.

1. $p_A^\circ = 10^{\,2.033 - \dfrac{77.4246}{T_S - 230}}$
2. $p_B^\circ = 10^{\,1.0044 - \dfrac{123.14}{T_S - 230}}$
3. Estimate liquid mole fractions: x_A, $x_B = 1 - x_A$
4. $V_A = \eta_A X_{be,A}$
5. $L_A = (1-\eta_A)X_{be,A}$
6. $L_B = \dfrac{x_B}{x_A}L_A$
7. $V_B = X_{be,B} - L_B$
8. $\gamma_A = 10^{\dfrac{0.176}{\left(1-\dfrac{x_A}{x_B}\right)^2}}$
9. $\gamma_B = 10^{\dfrac{0.176}{\left(1-\dfrac{x_B}{x_A}\right)^2}}$
10. $y_A = \dfrac{V_A}{V_{ki,A} + V_B}$
11. $y_B = 1 - y_A$
12. $\varepsilon = |y_A\gamma_B p_B^\circ x_B - y_B\gamma_A p_A^\circ x_A|$
13. If $\varepsilon <$a small positive number then go to step 17.
14. $x_A^{(\text{new})} = \dfrac{L_{ki,A}}{L_{ki,A} + L_{ki,B}}$; $x_B^{(\text{new})} = 1 - x_A^{(\text{new})}$
15. $x_A = \text{New estimate}(x_A, x_A^{(\text{new})})$; $x_B = 1 - x_A$
16. Return to step 4.
17. $p_S = \gamma_A p_A^\circ x_A + \gamma_B p_B^\circ x_B$

Calculate the flow rates:

Provide an estimate for the recycled stream (4) as $X_{4,A} = 141.64\,kmol/h$, $X_{4,B} = 95.655\,kmol/h$.

Stream (2) is calculated with the unifier module. Stream (3) is calculated with the reactor module. The separation tank module is used to calculate an updated stream (4*), stream (5), and the pressure of the tank. Then we check if the component flow rates of stream (4) and stream (4*) are approximately equal.

Table 1.2: Results with sequential modular approach

n	$X_{2,A}$	$X_{2,B}$	$X_{3,A}$	$X_{3,B}$	$X_{4,A}$	$X_{4,B}$	$X_{5,A}$	$X_{5,B}$
0	241.64	95.66	54.33	470.28	53.29	246.53	1.086	233.76
3	161.06	237.76	62.62	434.64	61.36	238.72	1.252	195.92
6	161.67	239.81	63.01	437.13	61.75	240.09	1.260	197.04
9	161.84	240.39	63.12	437.82	61.86	240.47	1.262	197.35

If not, the updated values are substituted to stream (4), and the calculation starts again.

Results obtained with undamped direct substitution are shown in **Table 1.2**. The iterative calculation stopped after 9 steps. Calculated pressure of the separation tank is 1.072 bar.

Circuits and tearing streams

A preferable property of sequential modular approach is that if there is no recycle in the flowsheet then no iteration is needed, at least on the level of flowsheet (iteration may be needed inside the modules). For example, calculation of the flowsheet shown in **Fig. 1.2** is done by simply calculating units (A), (B), and (C), in this sequence, obtaining values of streams (2), (3), and (4), in this sequence.

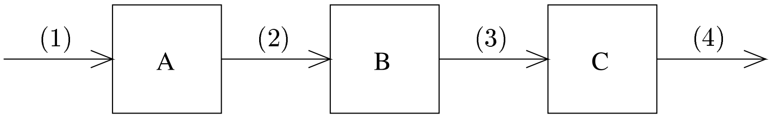

Figure 1.2: An acyclic flowsheet

The process shown in **Fig. 1.1** cannot be so calculated because not all the input streams of unit (U) are known. If the flowsheet is small and simple, one can at once see where the recycling stream flow, and which streams should be assignes as tearing streams. In more complex cases, however, circuits (cycles) of the flowsheet are to be explored, and one has to find a minimum number of tearing streams for easier estimation and calculation.

For exploring cycles, theory of directed graphs can be applied. Units (modules) of the flowsheet, as well as hypothetical sources of the input streams and sinks of the ouput streams of the flowsheet are considered as *nodes*, and streams connecting the modules are considered as (directed) *edges*, or *arcs*, of a direted graph. For example, a directed graph of the example flowsheet is shown in **Fig. 1.3**.

One possible way of exploring cycles is systematically building up a so-called *spanning tree* of the graph. Spanning tree is an acyclic (undirected) graph, or a

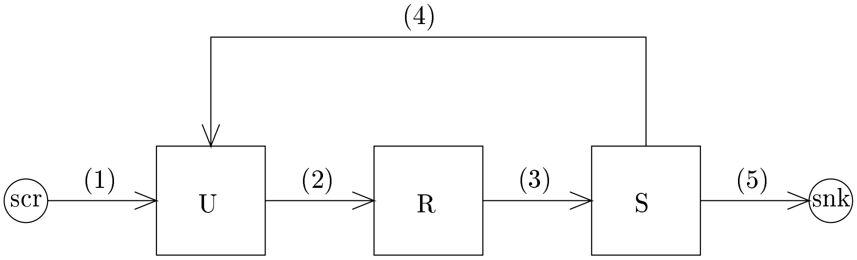

Figure 1.3: Directed graph of the example flowsheet

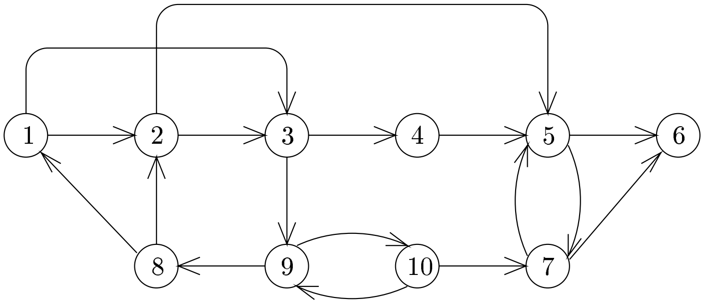

Figure 1.4: A more complex directed graph

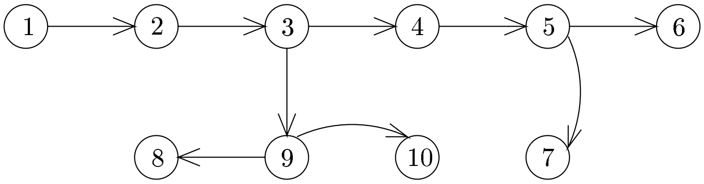

Figure 1.5: Spanning tree of the more complex directed graph

directed graph without directed circuit or by-pass. Spanning tree of a given graph is a graph consisting of all the nodes and a subset of the edges of that graph. For example, the spanning tree of the more complex graph of **Fig. 1.4** is shown in **Fig. 1.5**. This is called *spanning* all the nodes of the original graph with a minimum set of edges so that there is only one path from a source node(s) to any of the nodes.

A spanning tree can be constucted in the way of *Ariadne's thread* (from Greek mythology). In practice this is done by walking along the edges of the graph, turning back if reaching a node already reached earlier, and cutting out edges leading us to such situation. Cutting these nodes cut (eliminate) cycles and by-passes. Cycles of the studied (more complex) graph are: [5,7], [9,10], [2,3,9,8], [2,3,(9,10),8], [1,3,9,8], [1,3,(9,10),8], [1,2,3,9,8], [1,2,3,(9,10),8].

Convergence. Once the cycles are known, tearing streams can be assigned. If not known better, minimum number of tearing streams (to be estimated and updated) are assigned, but if it is the best for converging the calculation is uncertain. The only merit obtained this way is a possible easier estimation.

Converging a complex flowsheet with recycles can be rather difficult with sequential modular approach. One reason of this may be difficulty in converging the moduls themselves if they are compound units. Another reason can be lack of any hint to estimate the tearing streams.

1.1.3 Simultaneous modular approach

According to the simultaneous modular approach, models of moduls are decomposed to linear and non-linear parts. Linear submodel of each module is a homogeneous linear function of the variables in the input stream to that module, and the non-linear submodule calculates the non-constant coefficients of the linear submodel.

The models of our example process are as follow.

Unifier modul:

Linear submodel:

L1. $X_{ki,A} = X_{be1,A} + X_{be2,A}$

L2. $X_{ki,B} = X_{be1,B} + X_{be2,B}$

Non-linear submodel: None.

Reactor module:

Linear submodel:

L3. $X_{ki,A} = (1-\zeta)X_{be,A}$

L4. $X_{ki,B} = X_{be,B} + 2\zeta X_{be,A}$

Non-linear submodel:

1. $\zeta = 0.93 \exp\left(-0.76\dfrac{p_R}{T_R} - 0.22\dfrac{X_{be,B}}{X_{be,A}}\right)$

Separation tank module:

Linear submodel:

L5. $V_A = \eta_A X_{be,A}$

L6. $L_A = (1-\eta_A)X_{be,A}$

L7. $V_B = \eta_B X_{be,B}$

L8. $L_B = (1-\eta_B)X_{be,B}$

Non-linear submodel:

1. $p_A^\circ = 10^{\displaystyle 2.033 - \frac{77.4246}{T_S - 230}}$
2. $p_B^\circ = 10^{\displaystyle 1.0044 - \frac{123.14}{T_S - 230}}$
3. Estimate liquid mole fractions: x_A, $x_B = 1 - x_A$
4. $\gamma_A = 10^{\displaystyle \frac{0.176}{\left(1 - \frac{x_A}{x_B}\right)^2}}$
5. $\gamma_B = 10^{\displaystyle \frac{0.176}{\left(1 - \frac{x_B}{x_A}\right)^2}}$
6. $\alpha = \dfrac{\gamma_A p_A^\circ}{\gamma_B p_B^\circ}$
7. $\eta_B = \dfrac{1}{1 + \alpha\left(\dfrac{1}{\eta_A} - 1\right)}$
8. $V_A = \eta_A X_{be,A}$
9. $L_A = (1 - \eta_A) X_{be,A}$
10. $V_B = \eta_B X_{be,B}$
11. $L_B = (1 - \eta_B) X_{be,B}$
12. $y_A = \dfrac{V_A}{V_A + V_B}$
13. $y_B = 1 - y_A$
14. $\varepsilon = |y_A \gamma_B p_B^\circ x_B - y_B \gamma_A p_A^\circ x_A|$
15. If $\varepsilon <$ a small positive number then go to step 19.
16. $x_A^{(\text{new})} = \dfrac{L_{ki,A}}{L_{ki,A} + L_{ki,B}}$; $x_B^{(\text{new})} = 1 - x_A^{(\text{new})}$
17. $x_A = \text{New estimate}(x_A, x_A^{(\text{new})})$; $x_B = 1 - x_A$
18. Return to step 6.
19. $p_S = \gamma_A p_A^\circ x_A + \gamma_B p_B^\circ x_B$

Results of the non-linear submodel are pressure p_S and recovery ratio η_B.

Linearized flowsheet model

A linearized model of the full flowsheet can be built up from the linear submodels

Table 1.3: Results with simultaneous modular approach

n	$X_{2,A}$	$X_{2,B}$	$X_{3,A}$	$X_{3,B}$	$X_{4,A}$	$X_{4,B}$	$X_{5,A}$	$X_{5,B}$
0	141.64	198.30	42.49	396.60	41.64	198.30	0.850	198.30
1	158.95	241.18	60.16	438.77	58.95	241.18	1.203	197.59
2	162.56	240.62	64.26	438.05	62.97	240.62	1.285	197.43
3	161.56	240.61	62.82	438.09	61.56	240.61	1.256	197.48
4	162.00	240.61	63.27	438.08	62.00	240.61	1.256	197.46

(L1 to L8):

$$\begin{bmatrix} 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 \\ -(1-\zeta) & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ -2\zeta & -1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -\eta_A & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -\eta_B & 0 & 1 & 0 & 0 \\ 0 & 0 & -(1-\eta_A) & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -(1-\eta_B) & 0 & 0 & 0 & 1 \end{bmatrix} * \begin{bmatrix} X_{2,A} \\ X_{2,B} \\ X_{3,A} \\ X_{3,B} \\ X_{4,A} \\ X_{4,B} \\ X_{5,A} \\ X_{5,B} \end{bmatrix} = \begin{bmatrix} X_{1,A} \\ X_{1,B} \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

The simultanous modular calculation:

Simultanous modular calculation is a kind of successive linearization procedure.

A szimultán moduláris eljárás matematikai szempontból nézve fokozatos linearizás. First estimation is given to non-constant parameters in the coefficient matrix. For example, estimate ζ with 0.7, and η_B with 0.5; the coefficient matrix is then fully determined. Solve the system of linear equations to get the first estimation of the component flow rates. The apply the non-linear submodels in any sequence to obtain updated values of the non-constant parameters in the coefficient matrix: ζ and η_B. This pair of linear - nonlinear computation is iterated till convergence.

Results of this calculation are shown in **Table 1.3**. Calculated pressure of the separation tank is 1.071 bar.

Signal-flow graphs

Coefficient matrix of the linearized flowsheet model can be of any shape of sparsity, and the whole thing must be considered in a *Gauss* or *Gauss-Jordan* elimination procedure or in another numerical solution procedure. There is also another elimination technique, that might be of interest.

This technique works on the *signal-flow graph* of the process. For obtaining such a graph, streams are considered as nodes, and transformation of stream properties, made in the modules, are represented by directed arcs. Each arc is attached with a factor applied as a multiplier to its source node to calculate an increment to the ist target node.

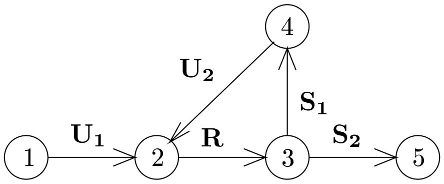

Figure 1.6: Signal-flow graph of the example flowsheet

Signal-flow graph of the example process is shown in **Fig. 1.6**. The streams (nodes) are connected by arcs characterized with matrix-valued factors because streams are characterized with vector-valued data. Factors of the unifier arcs $\mathbf{E_1}$ and $\mathbf{E_2}$ are 2×2 unit matrices. The factor of the reactor $\mathbf{R}$'s arc is:

$$\mathbf{R} = \begin{pmatrix} 1-\zeta & 0 \\ 2\zeta & 1 \end{pmatrix}$$

The factors of the separation tank's arcs $\mathbf{S_1}$ and $\mathbf{S_2}$ are:

$$\mathbf{S_1} = \begin{pmatrix} \eta_A & 0 \\ 0 & \eta_B \end{pmatrix}$$
$$\mathbf{S_1} = \begin{pmatrix} 1-\eta_A & 0 \\ 0 & 1-\eta_B \end{pmatrix}$$

Signal-flow graphs can be constructed component by component, as well, with scalar properties. Such a signal-flow graph of the example process is shown in **Fig. 1.7**.

Elimination is performed with equivalent transformations of the graph. These include three basic transformation steps:

1 **Elinimation of an internal node.** This is shown in **Fig. 1.8**. Naturally, if $y = ax$ and $z = by$ then $z = abx$.

2 **Elimination of a by-pass arc.** This is shown in **Fig. 1.9**. Evidently, if $y = ax + bx$ then $z = (a + b)x$.

3 **Elimination of a loop.** This is shown in **Fig. 1.10**. If $y = ax + by$ then $y = \dfrac{a}{1-b}x$. In case of vectorial data with matrix factors, letters are to be applied in proper order: $\boldsymbol{y} = (\mathbf{I} - \mathbf{B})^{-1}\,\mathbf{A}\ \boldsymbol{x}$.

The solution of **Fig. 1.6** with this method is: $\boldsymbol{X_2} = (\mathbf{I} - \mathbf{E_2}\,\mathbf{S_1}\,\mathbf{R})^{-1}\mathbf{E_1}\,\boldsymbol{X_1}$, $\boldsymbol{X_3} = \mathbf{R}\boldsymbol{X_2}$, $\boldsymbol{X_4} = \mathbf{S_1}\boldsymbol{X_3}$, $\boldsymbol{X_5} = \mathbf{S_2}\boldsymbol{X_3}$.

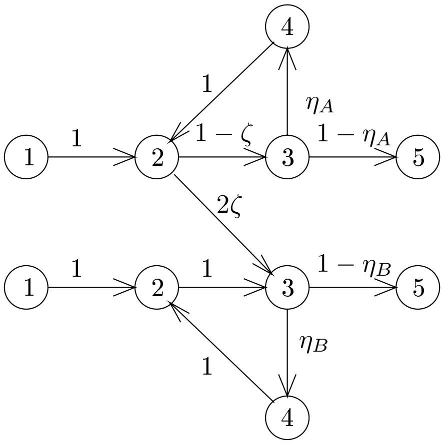

Figure 1.7: Scalar signal-flow graph of the example process

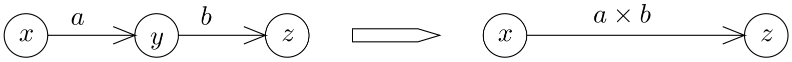

Figure 1.8: Rule 1 of equivalent transformations of signal-flow graphs

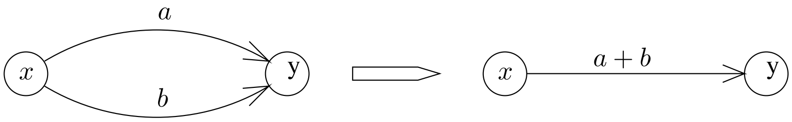

Figure 1.9: Rule 1 of equivalent transformations of signal-flow graphs

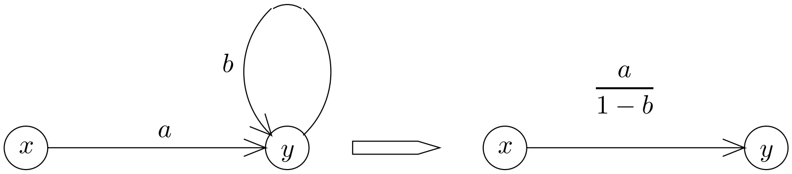

Figure 1.10: Rule 3 of equivalent transformations of signal-flow graphs

1.2 Selection of design and tearing variables

1.2.1 Design variables

Models normally consist of variables and equations, called also constraints. Constraints incorporate specifications as well, for example specifying pressure as 5 bar would give rise an equation $p = 5$. Models of unit operations typically contain more variables than independent equations. Denote number of variables by V, number of independent constraints (equations) by E, then a so-called *degree of freedom* F can be defines as

$$F = V - E$$

Solution of an equation system is normally determined if $F = 0$. If $F > 0$ then F number of variables are undetermined, one has to arbitrarily select such a number of varables and assign some value to them at will, thus constituting F new constraints. These selected variables are called *design variables*.

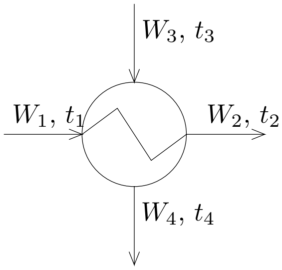

Figure 1.11: Heat exchanger connections

For example, a simplified steady state model of a heat exchange used for cooling down some product stream is given below (**Fig. 1.11**)

Variables:

1. β type of heat exchanger
2. Q heat power
3. A heat transfer area
4. U overall heat transfer coefficient
5. W_1 inlet cold stream mass flow rate
6. W_2 outlet cold stream mass flow rate
7. W_3 inlet hot stream mass flow rate
8. W_4 outlet hot stream mass flow rate
9. t_1 inlet cold stream temperature
10. t_2 outlet cold stream temperature
11. t_3 inlet hot stream temperature
12. t_4 outlet hot stream temperature

13. Δt_{ln} logarithmic approach temperature
14. c_h hot stream specific heat
15. c_c cold stream specific heat

Simplified model equations:

1. $Q = UA\Delta t_{ln}$
2. $\Delta t_{ln} = \dfrac{(t_1 - t_4) - (t_2 - t_3)}{\ln \dfrac{t_1 - t_4}{t_2 - t_3}}$
3. $W_1 = W_2$
4. $W_3 = W_4$
5. $Q = W_1 c_h (t_1 - t_2)$
6. $Q = W_3 c_c (t_4 - t_3)$
7. $U = U(W_1, W_2, W_3, W_4, t_1, t_2, t_3, t_4, \beta)$

Specifications:

8. $W_1 = \ldots$
9. $t_1 = \ldots$
10. $t_2 = \ldots$
11. $t_3 = \ldots$
12. $c_h = \ldots$
13. $c_c = \ldots$

7 model equations and 6 specifiations together add to 13 constraints, i.e. $E = 13$. Number of varables is $V = 15$, hence degree of freedom is $F = 15 - 13 = 2$. Two of the non-specified variables must be selected and specified arbitrarily by the designer. In princple, any two of the non-specified variables β, Q, A, U, W_2, W_3, W_4, t_4, Δt_{ln} can be selected as design variables.

The unit operation may be optimized (according to some objective) against the selected design variables.

Non-continuous variables are expedient to select into the design variables because their value cannot be obtained as a result of numerical calculation. Such a variable in the heat exchange example is type β that can be double pipe heat exchanger, shell-and-tube heat exchanger, plate heat exchanger, and includes many other data, e.g. in case of shell-and-tube type it can be co- or countercurrent, one- or two-pass exchanger, vertical or horizontal, a number of tubes, their arrangement (square or hexagonal), etc.

Once β is selected, the task to select another one remains. It *does* matter which one is selected. Selecting cooling water flow rate W_3 makes the design easy because all the unknown variables can be calculated by simple substitution as long as we keep a good sequence as shown in **Fig. 1.12**. If, on the other hand, heat transfer area A is selected then only W_2 and Q can be calculated directly, and all the other variables must be determined by solving a system of 5 equations simultaneously, as is shown in **Fig. 1.13**.

How can the design variables expediently selected?

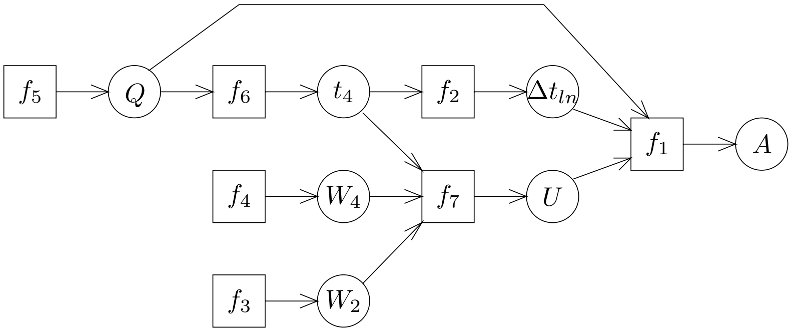

Figure 1.12: Acyclic design sequence with design variables β and W_3

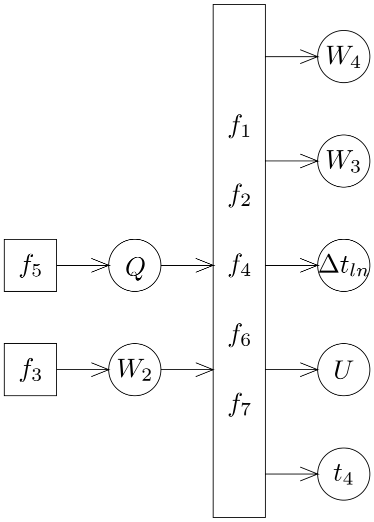

Figure 1.13: Design sequence with design variables β and A

1.2.2 Selecting design variables to obtain acyclic sequence

If all variables occur in each equation then simultanous equation system cannot be avoided. Simpler solutions are possible only because not all variables take place in each equation. Consider, for example, the equation system (1.1) ($V = 8$, $E = 4$, $F = 4$):

$$\begin{array}{lcccccccccl} f_1(& x_1, & x_2, & x_3 & & & & & &) = 0 \\ f_2(& & & x_3, & x_4, & x_5 & & & &) = 0 \\ f_3(& & & & x_4, & x_5, & x_6 & & &) = 0 \\ f_4(& & x_2, & & & & & x_7, & x_8 &) = 0 \end{array} \tag{1.1}$$

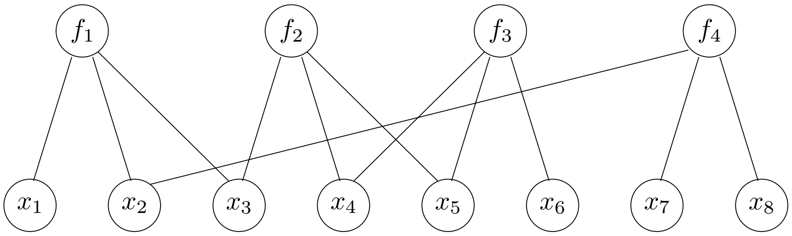

Figure 1.14: Bipartite graph of equation system (1.1)

Variables and equations can be *coupled* according to if the variable take place in the equation. This coupling can be represented with a *bipartite graph* (**Fig. 1.14**), or an *occurence matrix* (1.2).

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
f_1	1	1	1					
f_2			1	1	1			
f_3				1	1	1		
f_4		1					1	1

(1.2)

Finding acyclic sequence is possible if the matrix is sparse enough. The search is based on finding a variable that is coupled to exactly one equation, this can be expressed from this equation only. Such a variable is x_1 is our case, coupled to f_1. Expressing x_1 is indicated by directing the edge toward it. All the other variables in that equation must be known before calculating x_1, thus they must be directed toward f_1 (**Fig. 1.15**).

The analogous step in the occurence matrix is finding a column with exactly one entry '1' ('true'), that entry assignes its row coupled to this column ((1.3)).

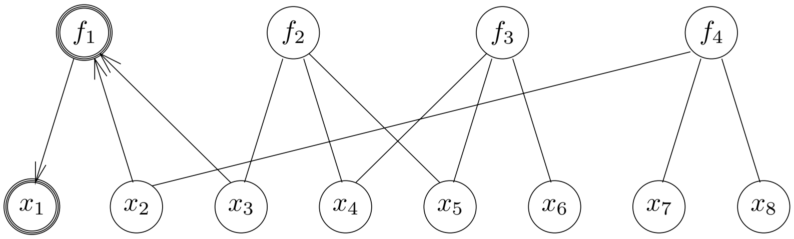

Figure 1.15: Selecing the first variable and equation

	$\boxed{x_1}$	x_2	x_3	x_4	x_5	x_6	x_7	x_8
$\boxed{f_1}$	$\boxed{1}$	1	1					
f_2			1	1	1			
f_3				1	1	1		
f_4		1					1	1

(1.3)

Then delete the assigned variable and equation ((1.4) and **Fig. 1.16**).

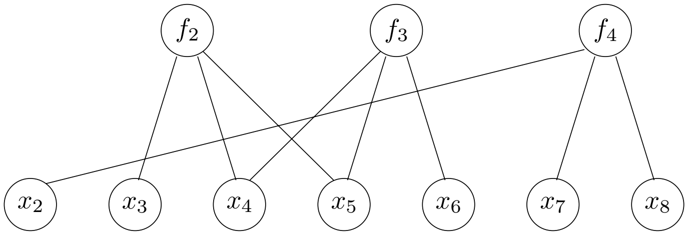

Figure 1.16: Graph that remains after deleting the first assigned variable and equation

		1.							
		~~x_1~~	x_2	x_3	x_4	x_5	x_6	x_7	x_8
1.	~~f_1~~	~~1~~	~~1~~	~~1~~	////	////	////	////	////
	f_2	/////		1	1	1			
	f_3	/////			1	1	1		
	f_4	/////	1					1	1

(1.4)

The same step is repeated in the remaining graph or matrix ((1.5) and **Fig. 1.17**):

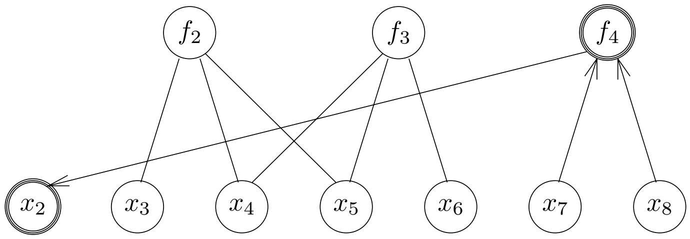

Figure 1.17: The second step

		1.	2.						
		~~x_1~~	$\boxed{x_2}$	x_3	x_4	x_5	x_6	x_7	x_8
1.	~~f_1~~	~~1~~	~~1~~	~~1~~	////	////	////	////	////
	f_2	/////		1	1	1			
	f_3	/////			1	1	1		
2.	$\boxed{f_4}$	/////	$\boxed{1}$					1	1

(1.5)

These steps are repeated until all equations are deleted:

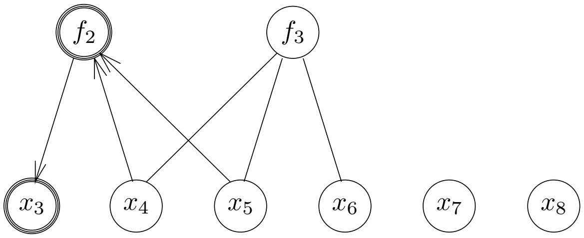

Figure 1.18: Third step

		1.	2.	3.					
		~~x_1~~	~~x_2~~	$\boxed{x_3}$	x_4	x_5	x_6	x_7	x_8
1.	~~f_1~~	~~1~~	~~1~~	~~1~~	////	////	////	/////	/////
3.	$\boxed{f_2}$	/////	/////	$\boxed{1}$	1	1			
	f_3	/////	/////		1	1	1		
2.	~~f_4~~	/////	~~1~~	/////	////	/////	////	~~1~~	~~1~~

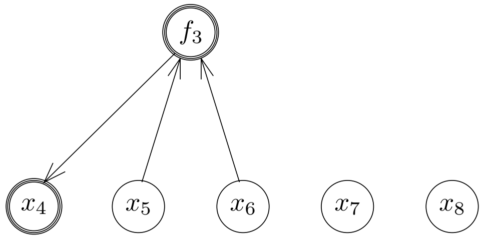

Figure 1.19: Fourth step

		1.	2.	3.	4.				
		~~x_1~~	~~x_2~~	~~x_3~~	$\boxed{x_4}$	x_5	x_6	x_7	x_8
1.	~~f_1~~	~~1~~	~~1~~	~~1~~	/////	/////	////	/////	/////
3.	~~f_2~~	/////	/////	~~1~~	~~1~~	~~1~~	////	/////	/////
4.	$\boxed{f_3}$	/////	/////	/////	$\boxed{1}$	1	1		
2.	~~f_4~~	/////	~~1~~	/////	/////	/////	////	~~1~~	~~1~~

Variables not deleted are not expressed from any equation, therefore these will be selected as design variables: x_5, x_6, x_7, and x_8. Based on the sequence of selections, a proper sequence of calculation can be given (**Fig. 1.20**).

In each step of the algorithm, a variable with exactly one coupling is selected. There can be sever such variables in any step and, accordingly, there can be several different proper selections and acyclic design sequences.

It may also happen that no good selection if found in a step. Then we may find a good sequence by first stepping back to an earlier stage and selecting another variable in that stage.

1.2.3 Tearing variables

Acyclic calculation sequence cannot always be constructed. If each variable is coupled to at least $1 + k$ ($k > 0$) equations, but k is significantly smaller that E

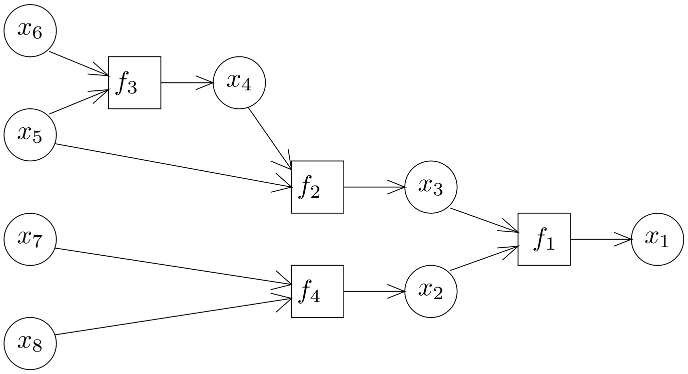

Figure 1.20: Assigned acyclic calculation sequence

then it is worth looking for a calculation sequence with k tearing variables, i.e. k variables to estimate and update in iteration cycles.

Let number of variables be V, number of equations (constraints) be E, i.e. degree of freedom $F = V - E$. Then not only F design variables but additional k tearing variables have also be selected. Az expedient way to accomplish this task is first temporarily deleting k equations, finding an acyclic calculation sequence with $F + k$ hypothetical design variables, and then selecting k of them to be tearing variables, together with re-including the k temporarily deleted equations, leaving the rest F to be real design variables.

How this is done is demonstrated in the following small example. Let the original problem be:

	x_1	x_2	x_3	x_4	x_5	x_6
f_1	1	1				
f_2	1		1	1		
f_3		1	1	1	1	1
f_4				1	1	1

Here $V = 6$, $E = 4$, and $F = 2$. Acyclic sequence cannot be found since each variable is coupled to at least 2 equations, i.e. $k = 1$. Therefore we temporarily omit one equation; let it be, just by chance, the first one:

	x_1	x_2	x_3	x_4	x_5	x_6
f_2	1		1	1		
f_3		1	1	1	1	1
f_4				1	1	1

An acyclic calculation sequence is shown in **Fig. 1.21** with hypothetical design variables x_3, x_5, and x_6.

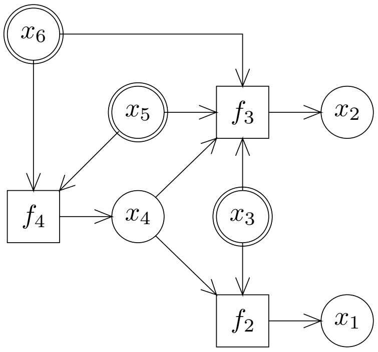

Figure 1.21: Acyclic calculation sequence after dropping f_1

However, the temporarily omitted equation f_1 should also be taken into account. Its connections are also shown in **Fig. 1.22**.

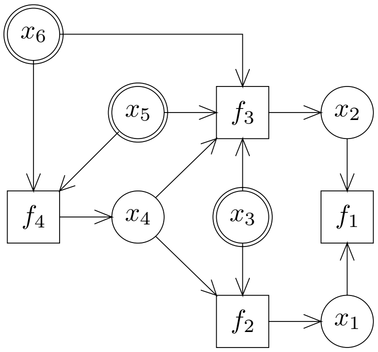

Figure 1.22: Acyclic calculation sequence together with connections of f_1

Now both variables of f_1 are necessarily input variables to it because they are output from some other equation. However, one of them should be turned back to become an output variable from f_1. Let it be, for example, x_1. After redirecting the arc, now x_1 is calculated from two equations, f_1 and f_2, instead of one. Therefore we have to turn back one of the arcs entering to f_2, and this procedure can be evolved aling paths of the graph until arriving to a hypothetical design variable. Such a state is shown in **Fig. 1.23**. Having reached such a state, that hypothetical design variable is re-qualified as a tearing variable: It has to be estimated and, after performing an acyclic calculation, updated repeatedly in an ieration process. In our case x_3 becomes a tearing variable whereas x_5 and x_6 remain design variables.

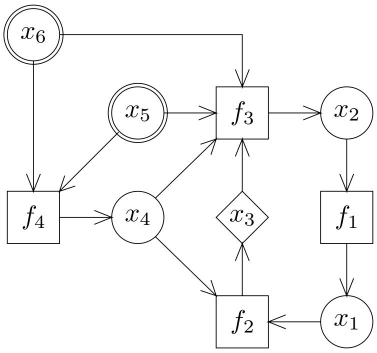

Figure 1.23: Tearing system after redirection of arcs.

1.3 Steady state of separation columns

Detailed models of steady state of countercurrent multistage fluid separation processes like distillation (rectification and stripping), absorption or desorption, countercurrent extraction are so large and complex that they deserve a treatment in their own.

Here we deal with staged models only, even packed columns are modelled stagewise; moreover we assume equilibrium stages.

1.3.1 Steady state models of separation columns

The basic element is the model of a(n equilibrium) stage. Stage efficiency may be taken into account but here we consider *equilibrium stages* only. If otherwise not indicated, we apply here notions of a distillation column. In this sense, an equilibrium stage works as a vapor - liquid equilibrium separation unit with at least two feeds: vapor emerges from below and liquid flows down from above, they mix in the stage, and separate to two output streams. The vapor emerging from the stage is in equilibrium with the liquid flowing down from it.

Connections and variables of such a stage are shown in **Fig. 1.24**.

Stages are conventionally counted top down at modelling (in contrast to the industrial practice where they are counted bottom up). Pressure of stage j is p_j, its temperature is T_j, the molar flow rate of liquid flowing from it is $L_j + s_j$, its mole fraction array is $\boldsymbol{x}_j$, the molar flow rate of the stream (vapor phase or light liquid phase) emerging from it is $V_j + S_j$, its mole fraction array is $\boldsymbol{y}_j$. Molar enthalpy of the downflowing liquid is h_j, that of the emerging phase is H_j. Part of the liquid can be led out of the column as side product with flow rate s_j so that only L_j reaches the next stage $j+1$ below. In the same way, part S_j of the emerging phase might be led out as sideproduct so that only V_j reaches the next stage $j-1$ above. From stage $j-1$ above flows down a stream with flow rate L_{j-1}

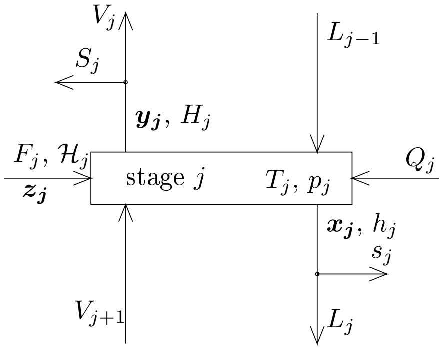

Figure 1.24: Two-phase stage

to stage j, and from the stage $j+1$ below emerges V_{j+1}, characterized with the respective compositions and molar enthalpies. These connections are a bit changed at the two end-points (top and bottom) of the column.

Feed with flow rate F_j, composition $\boldsymbol{z}_j$, and some molar enthalpy $\mathcal{H}_j$ may arrive to any stage j. Heat can be introduced to any stage j as Q_j. (Negative Q_j means cooling.)

If there are C components in the process then any stage is described by $3C+9$ scalar variables.

The steady state model consists of 4 main equation groups, the so-called MESH equations, and additional axiliary equation groups. These are as follow, in case of N stages.

MESH equations These equations are related to stages: (component) **M**aterial balances around each stage, **E**quilibrium relations on each stage, mole fractions **S**um conditions (up to 1) on each stage, and **H**eat balance around each stage:

("**M**"): Component molar balances around stages

$$\begin{aligned} &L_{j-1}x_{i,j-1} + V_{j+1}y_{i,j+1} + F_j z_{i,j} \\ &- (L_j + s_j)x_{i,j} - (V_j + S_j)y_{i,j} = 0 \quad (i = 1\ldots C;\;\; j = 1\ldots N) \end{aligned} \tag{1.6}$$

These are slightly modified at top and bottom since there is no stream 0 or stream $M+1$.

("**E**"): Phase equilibrium equations. These can be of any form, here they are represented by equilibrium ratios $K_{i,j}$:

$$y_{i,j} = K_{i,j}x_{i,j} \quad (i = 1\ldots C;\;\; j = 1\ldots N) \tag{1.7}$$

("**S**"): Summation in both phases (i.e. $\mathrm{S_x}$ and $\mathrm{S_y}$)

$$-1 + \sum_{i=1}^{C} x_{i,j} = 0 \qquad (j = 1 \dots N) \tag{1.8}$$

$$-1 + \sum_{i=1}^{C} y_{i,j} = 0 \qquad (j = 1 \dots N) \tag{1.9}$$

("**H**"): Stage heat balances

$$\begin{aligned} &L_{j-1}h_{j-1} + V_{j+1}H_{j+1} + F_j\mathcal{H}_j + Q_j \\ &- (L_j + s_j)h_j - (V_j + S_j)H_j = 0 \quad (j = 1 \dots N) \end{aligned} \tag{1.10}$$

Total molar balance Total molar balance on the stage can also be written:

$$L_{j-1} + V_{j+1} + F_j - (L_j + s_j) - (V_j + S_j) = 0 \quad (j = 1 \dots N) \tag{1.11}$$

but this is a combination of the component balances, not independent of them.

Auxiliary equations These are mainly detailed equilibrium models for computing $K_{i,j}$, see *chapter 2*, providing a function as

$$K_{i,j} = K(T_j, p_j, \boldsymbol{x}_j, \boldsymbol{y}_j, i) \quad (i = 1 \dots C;\ j = 1 \dots N) \tag{1.12}$$

as well as functions for molar enthalpies:

$$h_j = h(T_j, p_j, \boldsymbol{x}_j) \quad (j = 1 \dots N) \tag{1.13}$$

$$H_j = H(T_j, p_j, \boldsymbol{y}_j) \quad (j = 1 \dots N) \tag{1.14}$$

Not considering additional variables involved in the auxiliary equations, there are altogether $N(4C + 9)$ variables and $N(3C + 5)$ equations, and the degrees of freedom is $N(C + 4)$.

Column pressure is fixed at just one point of the system, and pressure drop over stages formes according to physics. Usually no additional equations describing pressure drop dependence on flow rates, temperature, and composition are introduced, but pressure drop is estimated and pressure p_j on each stage are thus specified (N specified ariables). Usually all the F_j, $\boldsymbol{z}_j$, and $\mathcal{H}_j$ are known, thesa are together $N(C + 2)$ scalar data. Additional specifications are $N - 2$ heat input data (the end-point condensation and boiling powers are usually variable to be calculated), side stream flow rates, and two other data are specified to get a system of equations with equal number of unknowns. Typical selection for the last two data are one product flow rate (e.g. distillate D) and reflux ratio R.

Several solution methods are published and in use for calculating the unknowns. Each has a preferred set of specifications.

1.3.2 Initial estimation

Any solution method is used, unkonwns must be estimated before starting iteration. The simplest estimations are listed below.

Temperature At the endpoint the temperatures can be estimated based on the expected product purity: bubble point and dew points can be estimated or calculated. In between, along the column, a temperature profile linear in stage numbers may be assumed. This is usually far from the solution but provides good initial start.

Equilibrium ratios Equilibrium ratios may be considered as functions of temperature only, for this aim. One may start up from constant relative volatilities, as well.

Compositions Expected mole fractions in the products are usually known. Composition on feed stages are near the feed composition. A linear profile in stage numbers can be applied. Vapor compositions can be obtained in the simplest way with supposed constant relative volatilities $\alpha_i = \alpha_{i/*}$:

$$y_{i,j} = \frac{\alpha_i x_{i,j}}{\sum_{k=1}^{C} \alpha_k x_{k,j}}$$

Internal flow rates Usual design variables are distillate flow rate D and reflux ratio R; hence flow rates in the top of the column can be determined:

$$L_0 = RD$$

$$V_1 = (R+1)D$$

The other flow rates along the column can be calculated with material balance and *constant molar overflow* approximation:

$$\begin{aligned} L_j &= L_{j-1} + q_j F_j - s_j & \quad (j = 2, \dots N) \\ V_{j+1} &= L_j + s_j + V_j + S_j - L_{j-1} - F_j & \quad (j = 1, \dots N-1) \end{aligned}$$

where q_j is estimated liquid phase ratio of the feed to stage j.

1.3.3 Linear subsystem with tridiagonal coefficient matrix

Equations **M** (1.6) are linear in mole fractions, and equations **H** (1.10) in flow rates if the other variables are fixed. If $K_{i,j}$ are (temporarily) also fixed then the vapor compositions can be expressed as $y_{i,j} = K_{i,j} x_{i,j}$, and only the $x_{i,j}$ are unknown variables. These equations, linear after fixing the other variables, have a special form beause at most three stage index values take place in each equation: the stage to which the balance is made, the next upper stage, and the next lower stage: $j-1$,

j, and $j+1$. Thus the coefficient matrix is of a tridiagonal one. This is essential because such systems can be solved fast and easily . Let the coefficients, variables, and right hand sides be denoted as follows:

$$\begin{bmatrix} b_1 & c_1 & & & & & \\ a_2 & b_2 & c_2 & & & & \\ & \ddots & \ddots & \ddots & & & \\ & & a_j & b_j & c_j & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & a_{N-1} & b_{N-1} & c_{N-1} \\ & & & & & a_N & b_N \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_j \\ \vdots \\ x_{N-1} \\ x_N \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_j \\ \vdots \\ d_{N-1} \\ d_N \end{bmatrix}$$

then it can be transformed to bidiagonal, using *Gauss* elimination:

$$\begin{bmatrix} 1 & g_1 & & & & \\ & 1 & g_2 & & & \\ & & \ddots & \ddots & & \\ & & & 1 & g_j & \\ & & & & \ddots & \ddots \\ & & & & 1 & g_{N-1} \\ & & & & & 1 \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_j \\ \vdots \\ x_{N-1} \\ x_N \end{bmatrix} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_j \\ \vdots \\ u_{N-1} \\ u_N \end{bmatrix}$$

From this form the unknowns are at once can be expressed. The solution procedure works with two additional arrays: g_j and u_j:

$$\begin{aligned} g_1 &= \frac{c_1}{b_1} \\ u_1 &= \frac{d_1}{b_1} \\ g_j &= \frac{c_j}{b_j - a_j g_{j-1}} \qquad (j = 2, \ldots, N-1) \\ u_j &= \frac{d_j - a_j u_{j-1}}{b_j - a_j g_{j-1}} \qquad (j = 2, \ldots, N) \\ x_N &= u_N \\ x_j &= u_j - g_j x_{j+1} \qquad (j = N-1, \ldots, 1) \end{aligned}$$

1.3.4 Simultaneous methods

Simultaneous or *global* solution methods solve the whole system of equations in a simultaneous manner, in contrast to *sequential* methods. All simultaneous methods apply some kind of *Newton*-iteration, which is a kind of successive linearization. They differ from each other in technical details and degree of applied simplifying assumptions.

Many unknowns are to be determined in this computation (see *section 1.3.1*): T_j, L_j, V_j, h_j, H_j, $x_{i,j}$, $y_{i,j}$, and $K_{i,j}$ (i=1, 2, …, C) (j=1, 2, …, N), i.e. $N(3C+5)$ variables. This can be a large number. For example, with $C = 5$ and $N = 70$ (a usual, moderate scale problem) this is a system of 1400 variables and equations. This can be decreased by considering the auxiliary equations as functions and substituting them to the MESH equations. In that case the number of unknowns ($x_{i,j}$, $y_{i,j}$, T_j, L_j, V_j)) is decreased to $N(2C+3)$.

Even this is a large system, 910×910 in the above example. Note also that partial derivatives (entries of the *Jacobian* matrix) are needed for performing *Newton*-iteration. Many of those derivatives are zero (we have a rather sparse matrix) but, even so, many derivatives have to be computed, and many of them can be calculated with relatively large numeric uncertainty only. For example, composition dependence of equilibrium temperature is not known as an explitice function but should be determined with bubble point calculation, an iterative procedure, discussed in *2*, using difficult fugacity and / or activity coefficient models.

Computation work load can be decreased by expressing some variables from some equations and substituting them to others. Another opportunity is not updating *Jacobian* in every single iteration but only after several steps. Temporary neglection of some members in computing partial derivatives can also be used to decrease the load. For example, at computing partial derivative of equilibrium ratio

$$K_{i,j} = \frac{\gamma_{i,j}(T_j, \boldsymbol{x}_j) x_{i,j} p_i^\circ(T_j)}{p_j}$$

according to temperature one may neglect temperature dependence of $\gamma_{i,j}$:

$$\frac{\partial K_{i,j}}{\partial T_j} \approx \frac{\gamma_{i,j}}{p_j} \left(\frac{\mathrm{d}p_i^\circ}{\mathrm{d}T_j} \right)_{T_j}$$

Only the two best known methods are drafted here.

Naphtaly-Sandholm **method**

This is the most general and widely applied procdure first publisged by *Naphtaly* and *Sandholm* in 1971. It involves large computation load; its convergence depends havily on initial estimation, but is fast near the the solution.

The method considers liquid and vapor component flow rates

$$\begin{aligned} l_{i,j} &= L_j x_{i,j} \\ v_{i,j} &= V_j y_{i,j} \end{aligned}$$

as basic variables instead of mole fractions. Instead of applying summation constraints $\mathrm{S_x}$ and $\mathrm{S_y}$, it decreases the number of variables and equations by substituting the expressions

$$L_j = \sum_{i=1}^{C} l_{i,j}$$
$$V_j = \sum_{i=1}^{C} v_{i,j}$$

and

$$x_{i,j} = \frac{l_{i,j}}{L_j}$$
$$y_{i,j} = \frac{v_{i,j}}{V_j}$$

to the other equations. The remaining equations **M**, **E**, and **H** are

$$l_{i,j-1} + v_{i,j+1} + F_j z_{i,j} - \left(1 + \frac{s_j}{\sum_{k=1}^{C} l_{k,j}}\right) l_{i,j} - \left(1 + \frac{S_j}{\sum_{k=1}^{C} v_{k,j}}\right) v_{i,j} = 0 \quad (\mathbf{M}_{i,j})$$

$$v_{i,j} \sum_{k=1}^{C} l_k, j - K_{i,j} l_{i,j} \sum_{k=1}^{C} v_{k,j} = 0 \quad (\mathbf{E}_{i,j})$$

$$h_{j-1} \sum_{i=1}^{C} l_{i,j-1} + H_{j+1} \sum_{i=1}^{C} v_{i,j+1} + \mathcal{H}_j F_j + Q_j - h_j \left(s_j + \sum_{i=1}^{C} l_{i,j}\right) - H_j \left(S_j + \sum_{i=1}^{C} v_{i,j}\right) = 0 \quad (\mathbf{H}_j)$$

If auxiliary equations $K_{i,j}$, h_j, and H_j are substituted as functions then there are only $N(2C+1)$ equations to be solved for variables $l_{i,j}$, $v_{i,j}$, and T_j. Simplifying approximations are yet necessary in computing the *Jacobian* because the number of partial derivatives is a large number even after this reduction of scale.

The method can be further simplified by applying constant molar overflow (practically: assuming sectionwise constant L_j and V_j) so that variables v_j are cancelled. In this way the *Jacobian* becomes a tridiagonal hypermatrix, easier to deal with. Constant molar overflow is an acceptable approximation in many cases.

Ishi-Otto **method**

This method, first published by *Ishi* és *Otto* in 1973, considers $x_{i,j}$, T_j, and V_j as basic independent variables. Vapor mole fractions are expressed with equilibrium ratios as (1.7), and liquid flow rates are expressed with material balance:

$$L_j = V_{j+1} - V_1 + \sum_{k=1}^{j} (F_k - s_k - S_k) \quad (*)$$

and then expressions (1.7) and (*) are substituted to the forms (1.6), (1.8), and (1.10) of equations **M**, **S**, and **H**, respectively. In this way the number of independent variables is decreased to $N(C+2)$.

It neglects partial derivatives of equilibrium ratios $K_{i,j}$ by temperature T_j and *same* component mole fractions $x_{i,j}$; it computes partial derivatives of enthalpies h_j and H_j by temperature T_j only, neglects the others. This gives rise to a special shape of *Jacobian*.

Residues $\mathbb{M}_{i,j}$, $\mathbb{S}_j$, and $\mathbb{H}_j$ take place in the linearized equation system of the *Newton* method, applied to increment variables $\Delta x_{i,j}$, ΔT_j, and ΔV_j; the simplest of these is equation group for residues $\mathbb{S}_j$:

$$\sum_{i=1}^{C} \Delta x_{i,j} = -\mathbb{S}_j$$

Equation groups derived from component balances can be sorted according to components, in the following form:

$$\begin{bmatrix} * & * & & & & & \\ * & * & * & & & & \\ \ddots & \ddots & \ddots & & & & \\ & & * & * & * & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & * & * & * \\ & & & & & * & * \end{bmatrix} \times \begin{bmatrix} \Delta x_{i,1} \\ \Delta x_{i,2} \\ \vdots \\ \Delta x_{i,j} \\ \vdots \\ \Delta x_{i,N-1} \\ \Delta x_{i,N} \end{bmatrix} + \begin{bmatrix} * & * & & & & & \\ & * & * & & & & \\ & & \ddots & \ddots & & & \\ & & & * & * & & \\ & & & & \ddots & \ddots & \\ & & & & & * & * \\ & & & & & & * \end{bmatrix} \times \begin{bmatrix} \Delta T_1 \\ \Delta T_2 \\ \vdots \\ \Delta T_j \\ \vdots \\ \Delta T_{N-1} \\ \Delta T_N \end{bmatrix}$$

$$+ \begin{bmatrix} * & * & & & & & \\ * & * & * & & & & \\ * & & \ddots & \ddots & & & \\ * & & & * & * & & \\ * & & & & \ddots & \ddots & \\ * & & & & & * & * \\ * & & & & & & * \end{bmatrix} \times \begin{bmatrix} \Delta V_1 \\ \Delta V_2 \\ \vdots \\ \Delta V_j \\ \vdots \\ \Delta V_{N-1} \\ \Delta V_N \end{bmatrix} = - \begin{bmatrix} \mathbb{M}_{i,1} \\ \mathbb{M}_{i,2} \\ \vdots \\ \mathbb{M}_{i,j} \\ \vdots \\ \mathbb{M}_{i,N-1} \\ \mathbb{M}_{i,N} \end{bmatrix} \qquad (i = 1, 2, \ldots C)$$

Shape of equations derived from enthaly balances:

$$\begin{bmatrix} * & * & & & & & \\ * & * & * & & & & \\ \ddots & \ddots & \ddots & & & & \\ & & * & * & * & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & * & * & * \\ & & & & & * & * \end{bmatrix} \begin{bmatrix} \Delta T_1 \\ \Delta T_2 \\ \vdots \\ \Delta T_j \\ \vdots \\ \Delta T_{N-1} \\ \Delta T_N \end{bmatrix} + \begin{bmatrix} * & * & & & & & \\ * & * & * & & & & \\ * & & \ddots & \ddots & & & \\ * & & & * & * & & \\ * & & & & \ddots & \ddots & \\ * & & & & & * & * \\ * & & & & & & * \end{bmatrix} \begin{bmatrix} \Delta V_1 \\ \Delta V_2 \\ \vdots \\ \Delta V_j \\ \vdots \\ \Delta V_{N-1} \\ \Delta V_N \end{bmatrix} = - \begin{bmatrix} \mathbb{H}_{i,1} \\ \mathbb{H}_{i,2} \\ \vdots \\ \mathbb{H}_{i,j} \\ \vdots \\ \mathbb{H}_{i,N-1} \\ \mathbb{H}_{i,N} \end{bmatrix}$$

Terrible it might look, this is a special form so that first variables $\Delta x_{i,j}$ can be eliminated and then the system of scale $2N$ can be solved.

1.3.5 Sequential methods

Sequential (or *decomposition*) methods assign tearing variables like those in *section 1.2*, and these are updated in iteration cycles. Other variables can be computed in well defined calculation sequence as functions of the actual values of the tearing variables.

These are the simplest, and least demanding, computation methods but do not always applicable to more difficult problems like systems involving azeotropes or more complex processes e.g. distillation with side strippers, or combination of distillation and absorption processes.

The basic procedures are BP method applicable to near boiling mixtures (distillation), SR method applicable to wide boiling mixtures (absorption, desorption), and ISR methods applicable to multistage countercurrent extraction. There are several versions with detailed suggestions for improving convergence or best specification and estimation methods, but only the basic ideas are discussed here.

BP method

Its scheeme is shown in **Fig. 1.25**. The procedure consists of two embedded cycles. Internal flow rates L_j and V_j are modified in the outer cycle, whereas temperatures T_j are updated in the inner cycle. The temperature is updated stage by stage for known (estimated) compositions with bubble point calculations, that is why the method is called BP. Equilibrium ratios $K_{i,j}$ are updated here as well, needed for computing the material balances, then enthalpies in the outer cycle so that mole rates can be updated according to heat balance.

Estimation of equilibrium ratios $K_{i,j}$ is needed because they take place in the component balances. When component balances are solved for a given component i, the mole fraction of that component is computed to all stages: $x_{i,1}$, $x_{i,2}$, $\ldots x_{i,j}$, $\ldots x_{i,N}$, i.e. mole fractions of a particular stage j are computed in C different balance calculations; their sum is not necessarily 1. However, normalized mole fractions are needed for computing bubble points and enthalpies; hence is the normalization step ($\mathrm{S_x}$ equations):

$$x_{i,j}^{\text{normalized}} = \frac{x_{i,j}}{\sum_{k=1}^{C} x_{k,j}} \qquad (j = 1, 2, \ldots N)$$

Bubble points are computed stage by stage, independently. The obtained vapor mole fractions $y_{i,j}$ are normalized; equations $\mathrm{S_y}$ are incorporated in bubble point calculation.

Top vapor V_1 and reflux stream L_0 flow rates are determined by D and R. Flow rates V_{j+1} és L_j (j=1, 2, …, $N-1$) can be calculated stage by stage using total material balance (1.11) and heat balance (1.10):

$$V_{j+1} = L_j + s_j + V_j + S_j - F_j - L_{j-1}$$

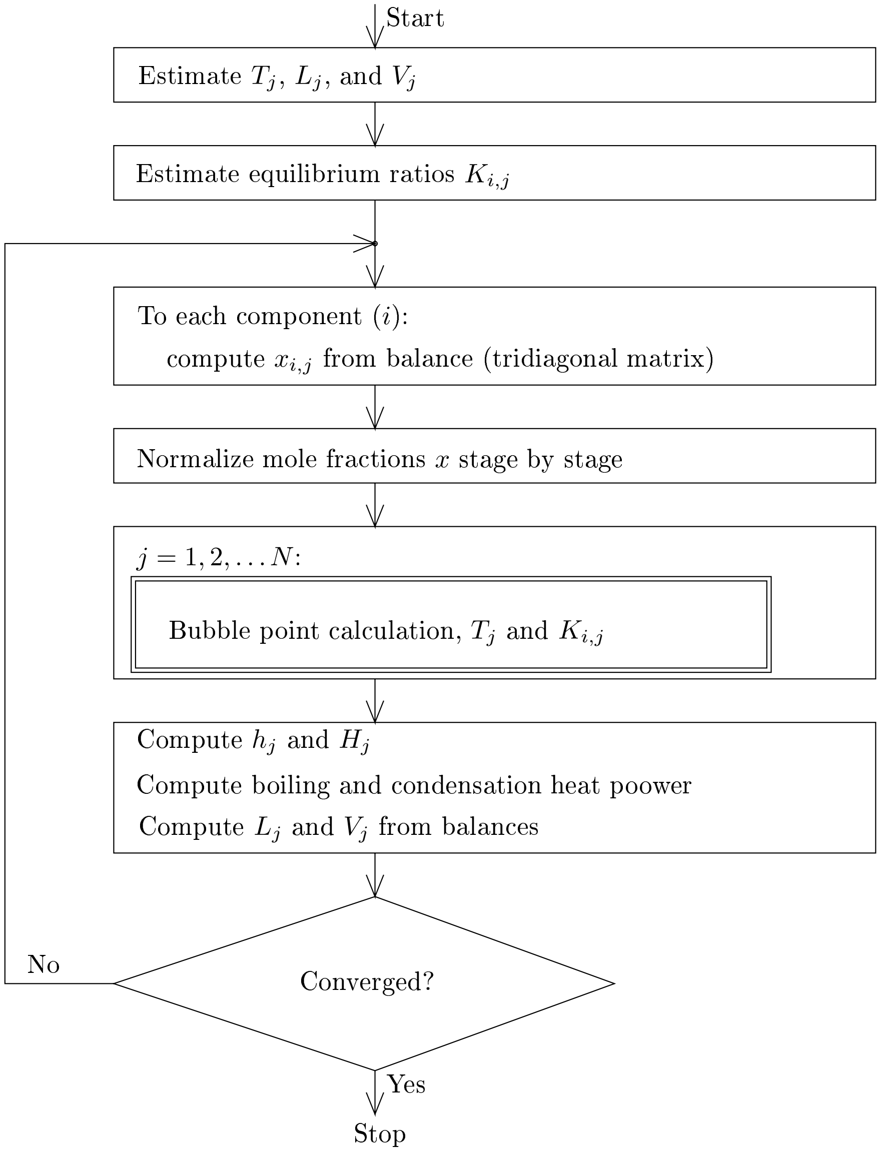

Figure 1.25: Scheeme of BP method

$$L_j = \frac{(H_{j+1} - h_{j-1})L_{j-1} + (H_{j+1} - \mathcal{H}_j)F_j - (H_{j+1} - h_j)s_j - (H_{j+1} - H_j)(V_j + S_j)}{H_{j+1} - h_j}$$

(Stage N is the boiler).

The convergence starts fast but then slows down near the solution, and usually oscillates. Several ideas are published to improve it, with some success, but the general tendency remains.

All the product stream flow rates must be specified; no way to find product flow rate for specified purity, unless embedding the whole procedue in an outer cycle. Either reflux ratio or condenser heat power is to specified, or the similar data at the bottom.

Temperature is sensitive to liquid composition in the bubble point calculations; that is why BP does not converge if componenst with far different boiling points are present (absorption, desorption).

SR method

The SR method, used for computing wide boiling point mixtures, absorption and desorption, updates temperatures with enthalpy balance, and liquid flow rates with summation equations.

Scheeme of the procedure is shown in **Fig. 1.25**. The fractions computed from component balances are used *before or without normalization* to update flow rates L_j:

$$L_j^{\text{updated}} = L_j \sum_{i=1}^{C} x_{i,j}$$

The products $L_j x_{i,j}$ are component flow rates to sum up, hence the name SR: '**S**um of **R**ates'. Mole fractions V_j are expressed from total balance:

$$V_{j+1} = L_j + s_j + V_j + S_j - F_j - L_{j-1}$$

Estimated $K_{i,j}$ and *normalized* liquid mole fractions x are used to calculate vapor mole fractions, and they are also normalized before using them to calculate enthalpies a recompute equilibrium ratios $K_{i,j}$.

Temperatures are updated in the inner cycle with *Newton* iteration, in principle, but only one step is made here. (No need for a full iteration with convergence.) Neighbouring trays take place in the linearized equation system so that the calculation is simple because we have a tridiagonal coefficient matrix. The only difficulty is computing partial derivatives according to temperatures.

According to experience, SR converges well if applied to a proper mixture.

ISR methods

No boiling temperatures have to be calculated in case of extraction processes. Instead, one may specify temperature as a whole or stage by stage. Heat power needed

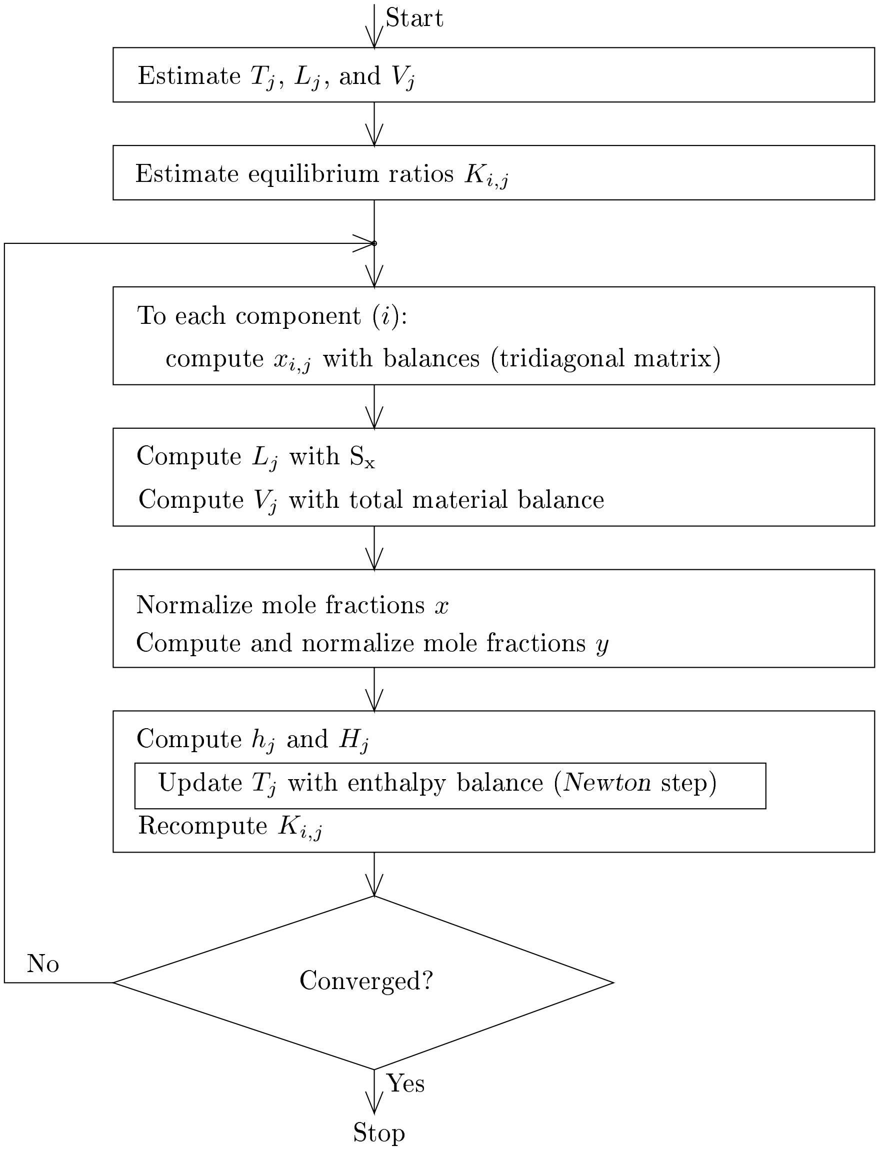

Figure 1.26: Scheeme of SR method

for maintaining the assigned temperatures can be computed afterwards; thus enthalpy balance can be omitted. Thus, MES equations may be solved instead of MESH.

Applying BP method is nonsense because of the same reason. Thus, modified versions of SR, called ISR (Isothermal SR) can be used.

Although two liquid phases are in equilibrium here, we keep the above notation: V and y are used to denote light phase flow rate and composition, respectively.

Feed stream, (and, perhaps, side product flow rates), pressures, and temperatures must be specified. Initial estimates are to be given to flow rates V_j and L_j, and equilibrium ratios $K_{i,j}$. Flow rates can be estimated using sectionwise constant V/L phase ratios, even mutual unsolvability of the phase solvents may be assumed.

No temperature calculating cycle is there in ISR, but there is oanother one, instead, because of the strong composition dependence of liquid-liquid equilibrium ratios.

Scheeme of the *Tsuboka-Katayama* version (1976) of ISR is shown in **Fig. 1.27**. The inner cycle strives for stability of mole fractions $x_{i,j}$ *before normalization*, but activity coefficients should be calculated with normalized mole fractions. Once the mole fractions do not change more, flow rates V_j are updated according to the SR idea, and L_j are calculated from material balance.

Version of *Renon-Assilenau-Cohen-Raimbault* (1971) is shown in **Fig. 1.28**. Not one or the other phase flow rate is updated but their ratio:

$$A_j = \frac{V_j^{\text{new}}}{L_j^{\text{new}}} = \frac{V_j \sigma_j^{(V)}}{L_j \sigma_j^{(L)}}$$

Flow rates of one of the phases (e.g. V_j) can be eliminated by substituting the so updated phase ratios. In this way a linear system with tridiagonal coefficient matrix is obtained for the folw rates of the other phase.

1.4 Cost estimation

1.4.1 Cost estimation levels

Costs are estimated with different precision at different levels of design. The following levels are roughly distinguished (*Pikulik and Diaz, Chem. Eng., 84(21), 106, (1977)*):

1. **Over-of-magnitude estimate.** Or : Ratio estimate. This is based on similar previous data; its propable accuracy is: ±40 %. For making this estimae, rough approximating formulas are applied.

2. **Study estimate.** Or: Factored estimate. This is based on knowning the flowsheet and major items of equipment; its propable accuracy is: ±25 %.

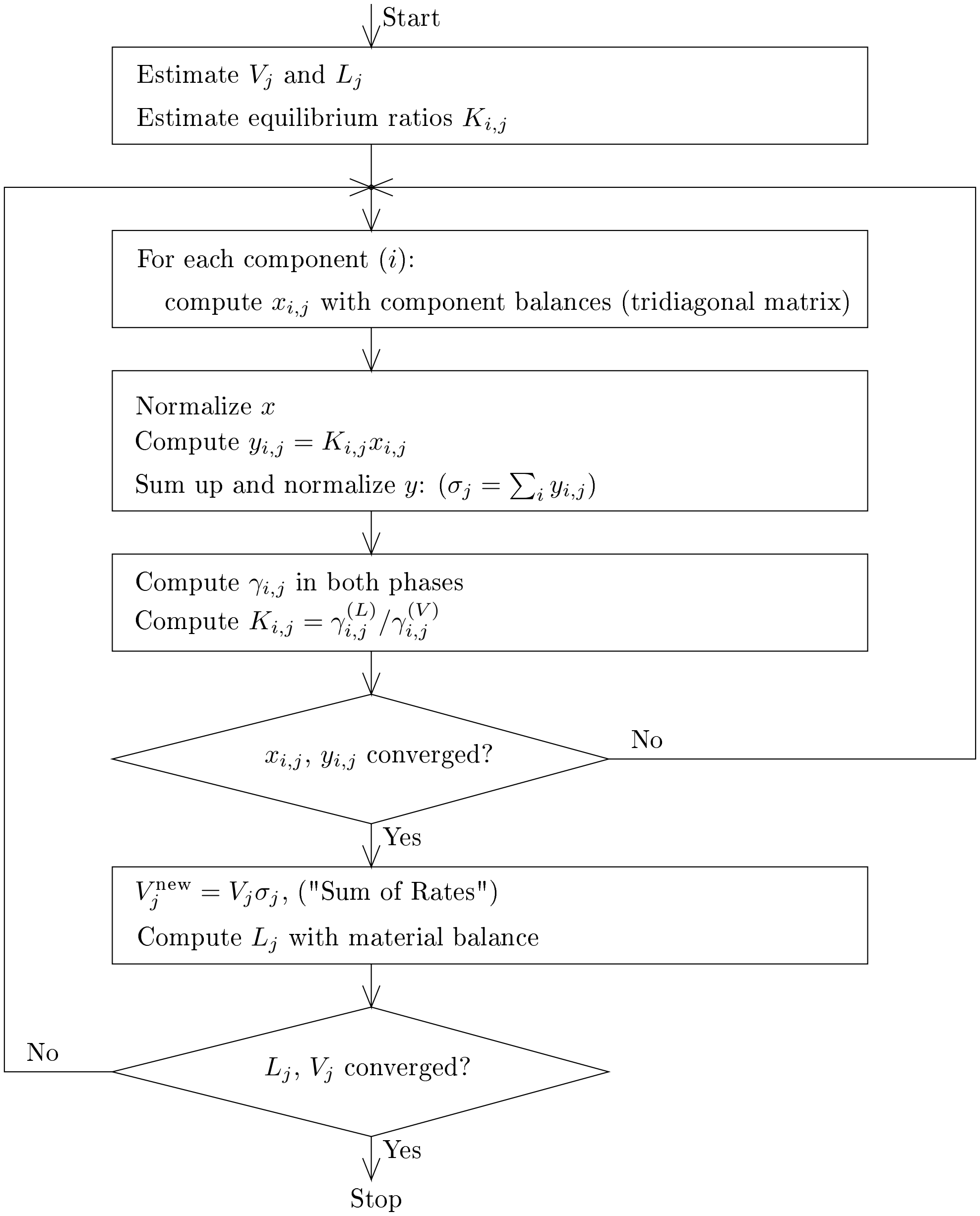

Figure 1.27: Scheem of ISR - *Tsuboka-Katayama* version

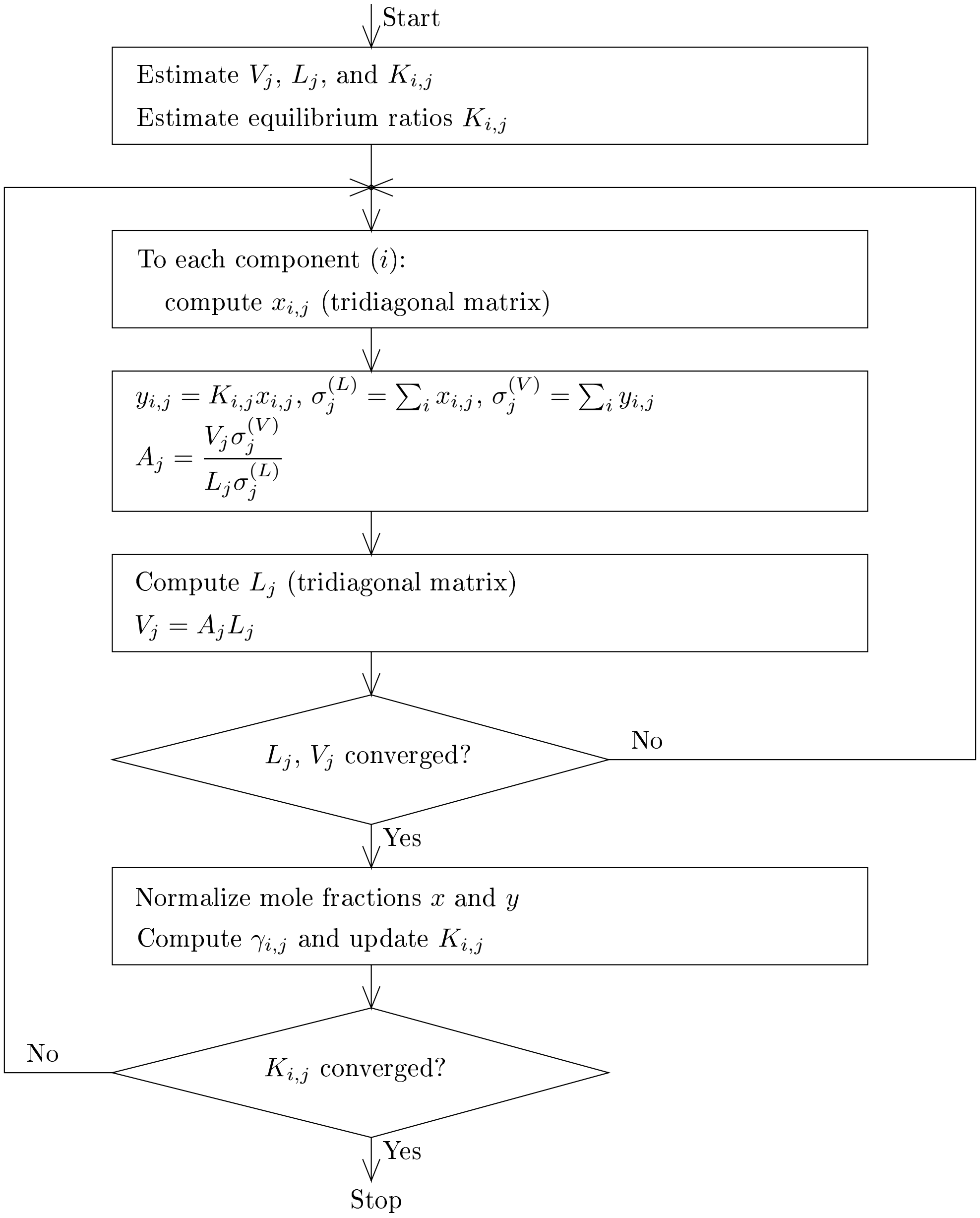

Figure 1.28: Scheeme of ISR - *Renon* version

3. **Preliminary stimate.** Or: Budget authorization estimate, or: Scope estimate. This is based on suffient data to permit the estimate to be budgeted; its propable accuracy is: ±12 %.

4. **Definitive estimate.** Or: Project control estimate. This is based on almost complete data, but completion of drawings and specifications; its propable accuracy is: ±6 %.

5. **Detailed estimate.** Or: Contractor's estimate. It is based on complete engineering drawings, specifications, and site surways; probable accuracy is ±3 %.

Designing the process also comes with costs; it depends on the scale of the process to be designed, measurable roughly with the expected costs themselves. **Table 1.4** gives a hint on costs of design at about 1977:

Table 1.4: Costs of design in USD, in 1977

	< 1 MM $	1 MM - 5 MM $	5 MM – 50 MM $
Study, 10^3 USD	5–15	12–30	20–40
Preliminary, 10^3 USD	15–35	30–60	50–90
Definitive, 10^3 USD	25–60	60–120	100–230

1.4.2 Cost estimation models

Capacity exponent

Simple, though rough, power relation can be given between costs of identical type but different capacity equipments or technologies *installed in the same period* (same year or not far in time). Denote costs with K, and capacity with C then, as a rough estimate, one may apply

$$\frac{K_1}{K_2} = \left(\frac{C_1}{C_2}\right)^a$$

where exponents a of equipment or plant types are fitted to empirical data. For example, cost exponents of some equipments are shown in **Table 1.5**, and of some technologies in **Table 1.6**.

If cost of an equipment of process of some capacity is known, and the same kind has to be designed for another capacity, then its cost *in the period of installing the known item* can be estimated.

Cost index

For comparing and estimating costs of items of *the same kind and same capacty*, technical and economic journals has been listing so-called *cost indexes* for decades.

Table 1.5: Capacity exponents of some equipments

equipment	exponent
bubble cap tray	1.20
vacuum drum dryer	0.20
shell-and-tube heat exchanger	0.44
centrifugal pump	0.61
non-mixed, lined reactor	0.41

Table 1.6: Capacity exponents of some technologies

process	exponent
catalytic cracking	0.85
catalytic polymerization	0.70
oxigene production	0.47
acetylene production	0.75
high pressure polyethylene production	0.90

These indexes relate to cost in a given period (year, quarter, or month) to some reference period, similarly to inflation rates and stock indexes. If the cost of an item was known as K_1 in an earlier period, and the cost index of that earlier period, and the actual period, repspectively I_1 and I_2, are known then the actual expected cost can be calculated as

$$\frac{K_2}{K_1} = \frac{I_2}{I_1}$$

With applying capacity exponents and cost indexes together, at least magnitudes of the costs of a planned item can be estimated, as far as the spanned time does not stretch over 10 years.

There are many cost indexes, the most general ones are as follow:

Marshall & Swift Equipment Index (M&S) Its two main branches are (a) *all-industry index*, formed by averaging 47 different indexes of referring to industry, commerce, houskeeping machines etc., and (b) *process-industry index*, an average of indexes of 8 most important branches (cement, chemicals, clay, glass, dyes, paper and pulp, petrol, rubber) weighed by their total production. Its starting year is 1926.

Engineering News-Record Construction Index (ENR) A compound labor and material cost index of industrial developments, starting in 1976.

Nelson Refinery Construction Index (NRC) Its components are: skilled labor, unskilled labor, iron and steel, construction materials, fittings, etc. Starting year: 1946.

Chemical Engineering Plant Cost Index (CE) Its main components: equipment, machinery, auxuliaries (61 %), set-up (22 %), construction material and labor (7 %), enineering (10 %). Starting year: 1959.

journal	peridicity	M&S	ENR	NRC	CE
Chemical Engineering	week	X			X
Engineering News-Record	week		X		
Oil and Gas Journal	forthnight			X	

Equipment cost functons

In order to obtain cost estimation with more accuracy that order of magnitude, cost functions of main equipment parameters can be fitted to experimental cost data. These are usually provided in plots, but can be transformed to analytical functions as well, usualy in polynomial form. Such functions are listed, for example, in **Table 1.7**, where K: cost, L: length, height, D: diameter, p: pressure load, A: heat transfer area (\$, m, bar, m^2).

Table 1.7: Cost functions of some cylindrical units

equipment	cost function
coat of tray column	$K = 460L^{0.91}D^{0.88}p^{0.18}$
coat of packed column	$K = 1260L^{0.68}D^{0.96}$
settler	$K = 8L^{0.24}D^{0.5}p^{0.18}$
shell-and-tube exchanger	$K = 240A^{0.59}$
catalytic reactor	$K = 100H^{0.57}D^{1.72}p^{0.76}$

These functions underestmate the real costs by far because they do not take into account additional cost elements like set-up, insulation, piping, painting, etc. These items can be taken into account roughly with *factoral technique* or more precisley with *modular technique*.

Factoral technique

Here the additional cost elements are related to the sum of all the equipments in the process:

$$K = f \sum_i K_i$$

The simplest (and the first) factor type is that of *Lang* depending on consistency valid to the whole of the technology (**Table 1.8**).

Table 1.8: *Lang* factors

process type	factor (f)
solid	4.6
solid and liquid	4.9
liquid	5.7

Several ideas have been suggested for decomposing the single factor to a product a partial factors. For example:

$$f = f_t f_p f_m$$

where f_t counts for temperature effects, f_p for pressure effects, f_m for construction material.

Modular technique

Here the additional cost elements are related to the equipments:

$$K = \sum_i f_i K_i$$

so that the additional cost factors can be given as equipment-dependent.

Most known is *Guthrie*'s decomposition to *basic cost elements* as direct labor costs, direct material costs, and indirect costs, all together constituting net module costs, and *secondary cost elements* further breaken down to smaller items like

(1) auxiliary materials: pipeline, substructure, smaller elements, instrumentiation, electric network, insulation, dyeing

(2) construction costs: building in materials, allocation, groundwork etc.

(3) indirect costs: transportation, insurances, taxes, overhead, engineering work, safety factor, subcontractor fees

Basic costs are modified by factors belonging to equipment kinds, for example:

Pressure vessel: $K = K_B f_m f_p$
Air cooler: $K = K_B(f_S + f_t + f_m)$
Distillation tray: $K = K_B(f_S + f_t + f_m)$
Heat exchanger: $K = K_B(f_t + f_p) f_m$

wherel K_B is basic costs, f_m is construction material factor, f_p is pressure factor, f_t is equipment type factor, f_S is tray spacing factor.

1.5 Flowheet design

This is a rather complex activity; here we give some hints only about what points should be kept in mind, and how a process is developed step by step. We apply an example of hydrodealkylation of toluene to benzene; this example is taken from *Douglas, J.M.: Conceptual Desing of Chemical Processes, McGraw Hill, 1988*

This chemical reaction takes place at high temperature an pressure, without catalyst, at about 35 bar and 620 C to 700 C temperature, and a side reaction is formation of diphenyl:

$$\begin{aligned} \text{Toluene} + \text{H}_2 &\rightarrow \text{Benzene} + \text{CH}_4 \\ 2\text{Benzene} &\rightleftharpoons \text{Diphenyl} + \text{H}_2 \end{aligned}$$

Hydrocracking happens at higher temperature. In order to avoid coke forming, 5 to 1 hydrogen excess is applied, and the reaction product is quenched under 620 C.

Available feedstock is pure atmospheric toluene, and hydrogen gas at 38 bar, contaminated with 5 % methane.

The ususal decision hierarchy is the following:

1. Batch or continuous process
2. Preliminary material balance
3. Reactpr and recyclng
4. Separation system
    - *a.* Gas recovery
    - *b.* Liquid recovery
5. Energy supply and recovery
6. Other utilities, and control system
7. Allocation, piping, instrumentation, etc.

Process design stretches roughly up to step 6.

Many process ideas are ruled out at the earlier stages, based on rough estimates. For example, if the cost of raw materials is higher than the income expected from selling the product than any cost of the equipment would only make the situation even worse. Another example is that if the process is not profitable even without considering costs of purifying the product, but just considering costs of a reactor, then again the process is not worth to consider further.

A first idea is shown in **Fig. 1.29**. Diphenyl (D) is a side product beside benzene (B). Hydrogen has to be blown down because methane wold be too expensive to separate, and would be accumulated in the recycle. Profitability is highly dependent on conversion, reacttor scale, compressor costs, etc.

The first real version is shown in **Fig. 1.30**. Another one, with diphenyl recycling and re-reacting is shown in **Fig. 1.31**. Final verion wit energy recovery is shown in **Fig. 1.31**.

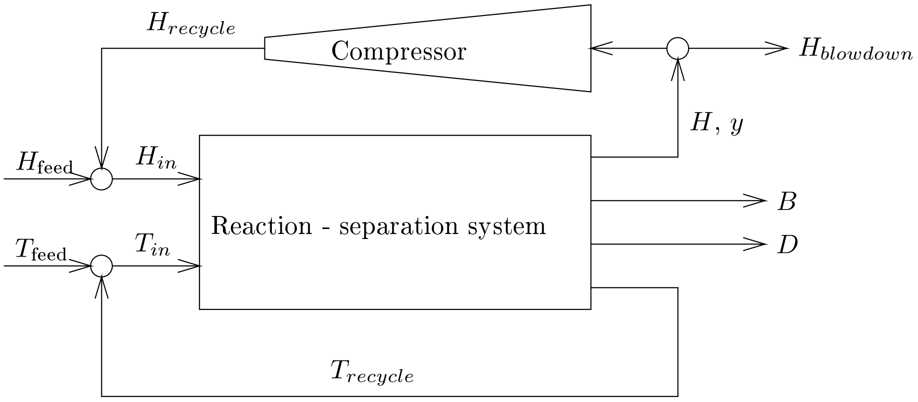

Figure 1.29: First idea

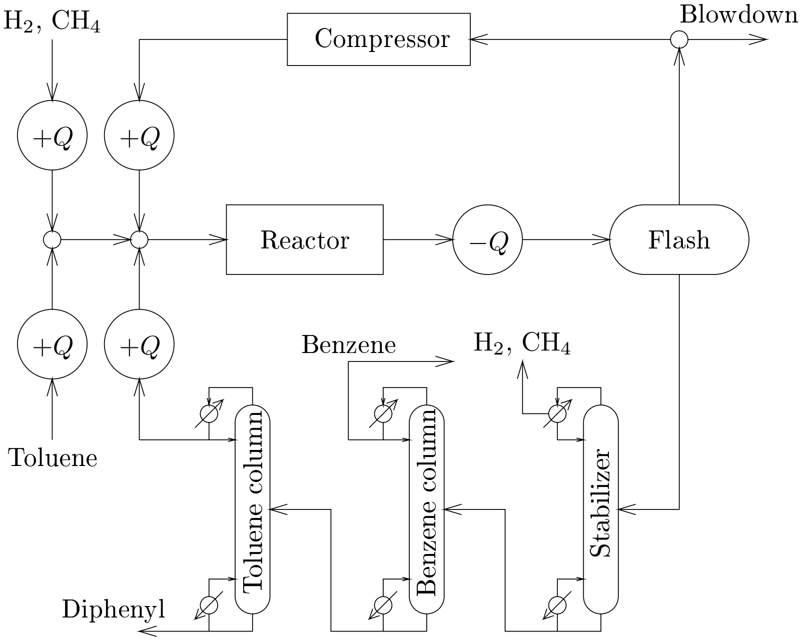

Figure 1.30: First version

1.6 Questions

1. What approaches of flowsheeting can you mention? Give a short description of them!

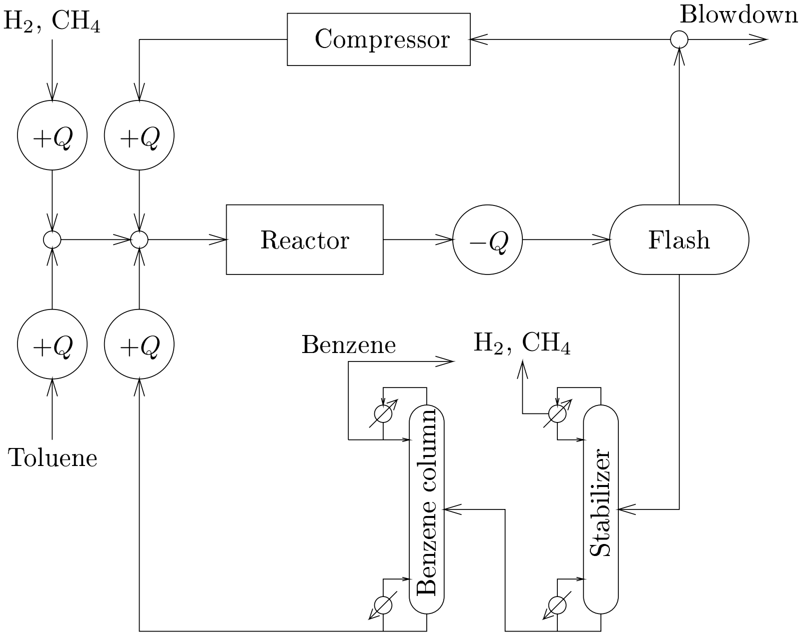

Figure 1.31: Another version

2. What is signal-flow graph? What is it for? How to simplify it?
3. What is simultaneous modular approah in mathematical sense?
4. What are design variables? How many of them can be assigned?
5. Haow can you find an acyclic calculation sequence?
6. Write a system of MESH equations!
7. List the discussed decomposition algorithms for counercurrent separation processes! Wich is applicable for what kind of processes? What are their rough scheemes?
8. What kind of cost estimation technique do you know? Describe them!

1.7 Suggested literaure

Westerberg A. W., Hutchinson H. P., Motard R. L., and Winter P.: *Process Flowsheeting*. Cambridge Universities Press, 1979.

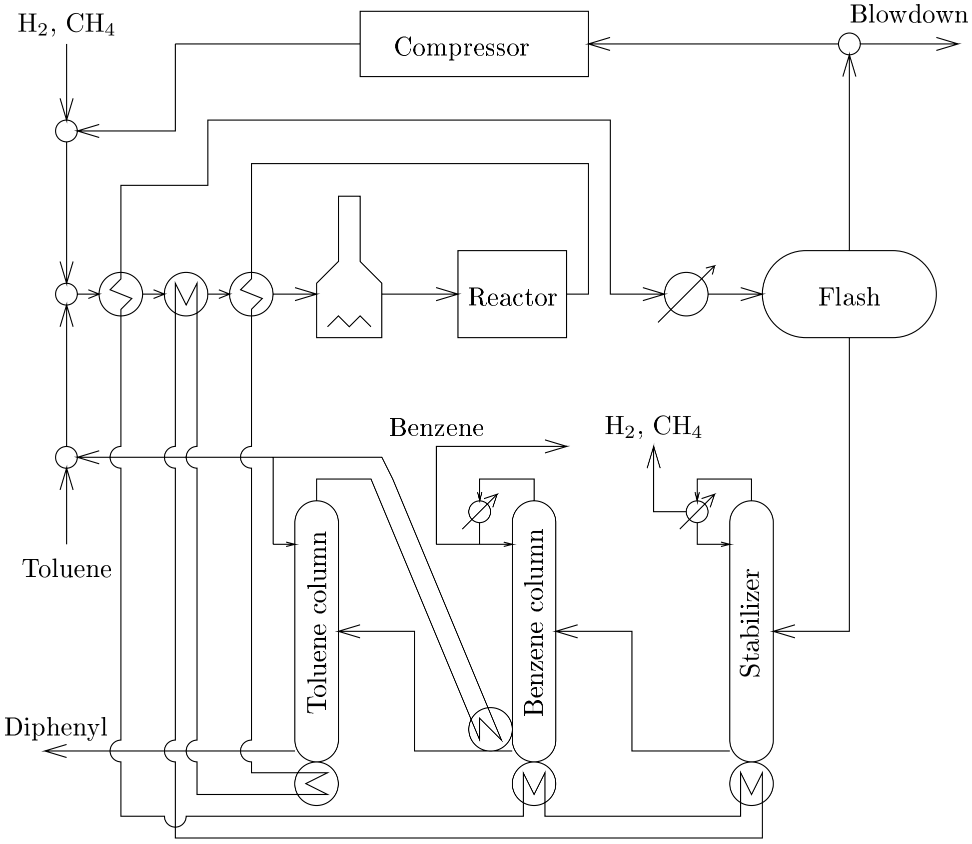

Figure 1.32: Final version

Veverka V. V., and Madron, F.: *Material and Energy balancing in the Process Industries.* Elsevier, 1997.

Peters, M. S., Timmerhaus, K. D.: *Plant Design and Economics for Chemical Engineers.* McGraw-Hill, 1969.

King, C. J.: *Separation Processes.* McGraw-Hill, 1971.

Douglas, J. M.: *Conceptual Design of Chemical Processes.* McGraw-Hill, 1988.

Chapter 2

Phase equilibria

2.1 Conditions of equilibria

Equilibrium of phases in direct touch (not separated by wall) necessarily goes with thermal and mechanical equilibrium. Thus, equality of temperature and pressure is necessary:

$$T^{I} = T^{II} \tag{2.1}$$

$$p^{I} = p^{II} \tag{2.2}$$

Besides fixing variables T and p, thermodynamic equilibrium is also characterized by the extremum of free enthalpy, i.e. Gibbs free energy. Therefore calculation of thermodynamic equilibria is reduced to extremum of molar free enthalpy function $G(T,p)$. Should the molar free energies of the phases in touch differ, the total free energy could be decreased by shifting some material from one phase to the other. Consequently the molar free enthalpies are equal if the total is at minimum:

$$G^{I} = G^{II} \tag{2.3}$$

In case of mixtures, equality of partial molar free enthalpy component by component is also a necessary condition:

$$\mu_i^{I} = \mu_i^{II} \qquad (i = 1, 2, \ldots, c) \tag{2.4}$$

One of its consequence is the equality of molar values (2.3).

Conditions (2.1)- (2.4) are necessary but not always sufficient. In practice only these conditions are usually checked in most cases because checking the condition of minimum is rather difficult and time consuming. However, applying the condition of minimum is applied more and more frequently, first of all for liquid-liquid equilibria and high pressure fluid equilibria calculations.

2.2 Formal thermodynamics of pure materials

Coherent equilibrium data triplets P–V–T of some pure material constitute a so-called surface of equilibrium states in the three-dimensional linear space of that material. The **thermal equations of states** are relationships describing these surfaces of equilibrium states. Value of V is unequivocal at some given p and T; therefore equations of states are usually expressed to pressure:

$$p = p\,(V,T) \tag{2.5}$$

Caloric (heat or energy related) properties (U, S, A, G), not considering entropy S, are defined as difference from a reference state only, and their absolute value is not important. Reference state is selected as soem well known ideal state like perfect (ideal) gas or ideal (Einstein-) cristal.

For reference state, some state of the perfect gas is frequently used. Equation of state of the perfect (ideal) gas is

$$p^{\square} = \frac{NkT}{v} \equiv \frac{n\mathrm{N_A}kT}{v} \equiv \frac{RT}{V}$$

If some caloric property is defined with reference to the perfect gas then equilibrium molar internal energy U, equilibrium molar entropy S, equilibrium molar enthalpy H, equilibrium molar free energy (Helmholtz energy) A, and equilibrium molar free enthalpy (Gibbs energy) G are computed as a sum of the corresponding data of the perfect gas and some calculated departure function:

$$X = X^{\square} + \Delta X$$

where X is any of the above caloric properties. The departure function ΔX can, however, be defined in several ways depending on the selected reference state.

One of them can be the *actual pressure and actual temperature reference convention* where the reference pressure p^* equals the actual pressure ($p^* = p$) and the reference temperature T^* equals the actual temperature ($T^* = T$):

$$X\,(T,p) = X^{\square}\,(T,p) + \Delta_{Tp}X$$

so that the departure function is

$$\Delta_{Tp}X \equiv X\,(T,p) - X^{\square}\,(T,p)$$

To calculate a potential function, one may start from the following general thermodynamic relation:

$$(\mathrm{d}A)_T = -p\,(T,V)\,\mathrm{d}V$$

Free energy departure is obtained if pressure in integrated at constant temperature along molar volume from the reference state up to the actual state of the real gas:

$$\Delta A = -\int\limits_{\text{reference state}}^{\text{actual state}} p\,(T,v)\,\mathrm{d}v$$

Here v is the molar volume index for integration whereas V is the actual molar volume.

This integral goes from an ideal state to a real state and the path from the former to the latter cannot be easily chosen. We utilize the fact that any gas behaves as ideal gas at infinite dilution. The integral is decomposed into two parts: one for a perfect (ideal) gas from the reference state to infinite molar volume and another one for the real gas from infinite molar volume to the actual molar volume.

In case of arbitrarily given pressure and actual temperature reference convention:

$$\Delta_{Tp^*}A = -\int_{\infty}^{V} p\left(T,v\right)\mathrm{d}v - \int_{V^*}^{\infty} p^{\square}\left(T,v\right)\mathrm{d}v$$

where

$$V^* = \frac{RT}{p^*}$$

i.e. we integrate from the perfect gas with volume corresonding to the reference pressure to the actual state real material. However, the integral of the ideal gas is not convergent here. This obstackle is overcome by adding and subtracting the same term

$$\int_{\infty}^{V} \frac{RT}{v}\mathrm{d}v$$

so that we get

$$\Delta_{Tp^*}A = -\int_{\infty}^{V}\left[p\left(T,v\right) - \frac{RT}{v}\right]\mathrm{d}v - \int_{\infty}^{V}\frac{RT}{v}\mathrm{d}v - \int_{V^*}^{\infty}\frac{RT}{v}\mathrm{d}v$$

resulting in

$$\Delta_{Tp^*}A = -\int_{\infty}^{V}\left[p\left(T,v\right) - \frac{RT}{v}\right]\mathrm{d}v - RT\ln\frac{V}{V^*}$$

$$\Delta_{Tp^*}S = -\left(\frac{\partial\Delta_{Tp^*}A}{\partial T}\right)_V = \int_{\infty}^{V}\left[\frac{\partial p\left(T,v\right)}{\partial T} - \frac{R}{v}\right]\mathrm{d}v + R\ln\left(\frac{V}{V^*}\right) - R$$

where

$$\frac{V}{V^*} = \frac{Vp^*}{RT} = Z\frac{p^*}{p}$$

and Z is the so-called compressibility:

$$Z \equiv \frac{pv}{NkT} \equiv \frac{pv}{n\mathrm{N_A}T} \equiv \frac{pV}{RT}$$

In case of actual pressure and actual temperature reference convention, the actual pressure is substituted:

$$\Delta_{Tp}A = -\int\limits_{\infty}^{V} \left[p\left(T,v\right) - \frac{RT}{v}\right] \mathrm{d}v - RT\ln Z \tag{2.6}$$

$$\Delta_{Tp}S = -\left(\frac{\partial \Delta_{Tp}A}{\partial T}\right)_V = \int\limits_{\infty}^{V} \left[\frac{\partial p\left(T,v\right)}{\partial T} - \frac{R}{v}\right] \mathrm{d}v + R\ln Z - R$$

Once we know free energy and entropy, all other caloric state functions can be expressed:

$$\Delta U = \Delta A + T\Delta S$$

$$\Delta H = \Delta U + RT\left(Z-1\right)$$

$$\Delta G = \Delta A + RT\left(Z-1\right)$$

2.2.1 Global instability and phase separation

We use the reduced form of the *van der Waals* equation in our example:

$$\pi = \frac{8\,\vartheta}{3\,\eta - 1} - \frac{3}{\eta^2}$$

Its $\vartheta = 0,75$ reduced temperature isotherm is shown in **Fig. 2.1**. The equilibrium reduced pressure belonging to this isotherm is $\pi^\circ = 0.282463$. A horizontal line intersecting from the isotherm the reduced molar volumes ($\eta_L = 0.489631$ and $\eta_V = 5.64305$) of the two equilibrium phases (L='Liquid', V='Vapor') is also shown in the figure. The connecting horizontal section separates equal areas below and above the isotherm. This is easy to show but only a knowledge obtained after observing the phases: if there are two phases then this rule must be held. On the other hand, this rule is not a reason of separating into phases.

According to the conditions by equilibrium thermodynamics, forming two phases may be caused only by having one of the potential functions better value than without it. Consider now a reduced volume between the two end-points of the horizontal connecting section of the real isotherm of reduced pressure $\pi = 0,282463$.

At given (fixed) temperature and *molar volume*, the equilibrium state is characterized by minimum of (*Helmholtz*) free energy A. The curve

$$\frac{\Delta_{TV}A}{RT} = -\frac{9}{8\vartheta\eta} - \ln\left(\eta - 1/3\right)$$

of the reduced free energy departure ($\Delta_{TV}A/RT$) belonging to reduced temperature $\vartheta = 0.75$ is drawn in **figure 2.2**. The common tangent line drawn to this curve at reduced molar volumes $\eta_L = 0.489631$ and $\eta_V = 5.64305$ is everywhere runs under the curve between the two tangent points, i.e. the free energy of the

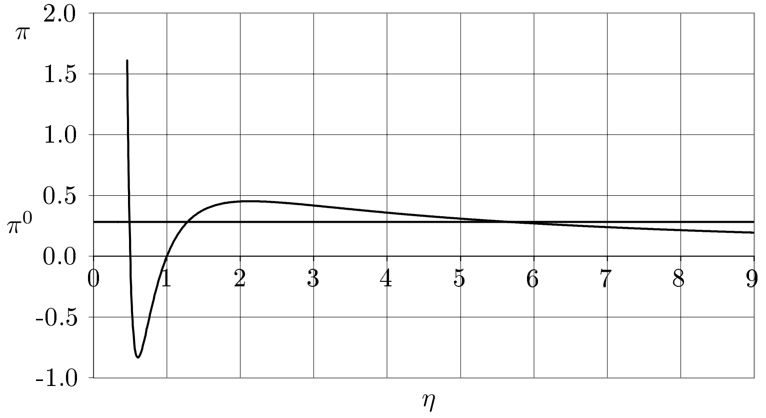

Figure 2.1: Isotherm of reduced EOS

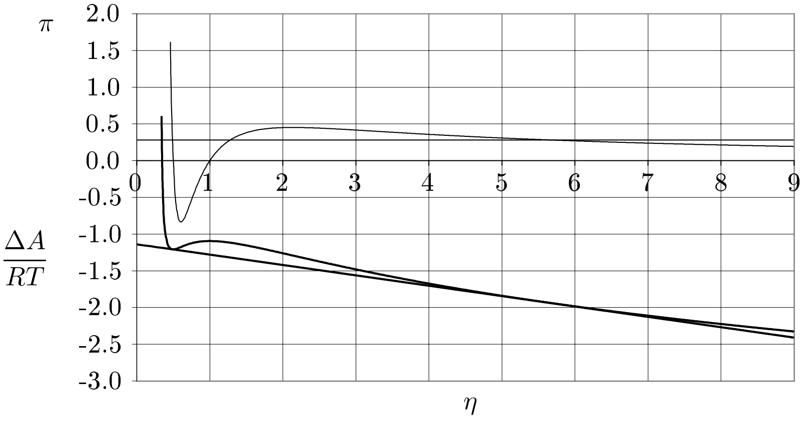

Figure 2.2: Reduced (Helmholtz) free energy and equilibrium distribution

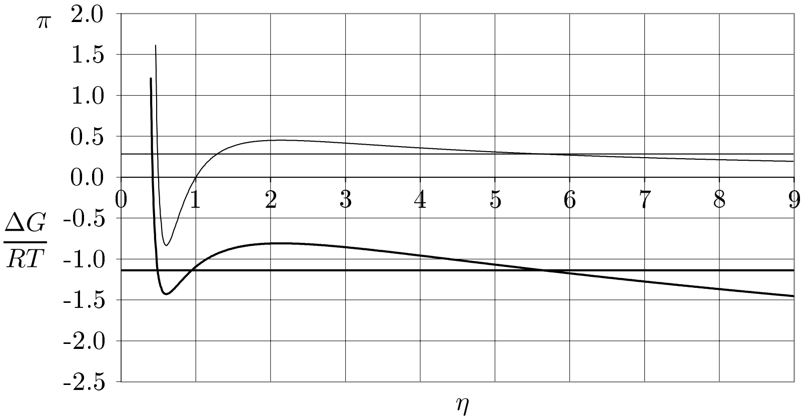

Figure 2.3: Reduced (Gibbs) free enthalpy and equilibrium distribution

homogeneous material is larger than what can be obtained by combining the two values at the molar volume endpoints. That is why the homogeneous system in unstable against densisty fluctuations inside this section. The real equilibrium free energy function runs along the common tangent line between the two end-points.

The common tangent line intersect the vertical axis ($\eta = 0$) at -1.13839. You can check by the general relation

$$p = -\left(\frac{\partial A}{\partial V}\right)_T$$

that this is just the reduced (Gibbs) free enthalpy departure ($\Delta_{TV}G/RT$) belonging to reduced molar volumes $\eta_L = 0.489631$ and $\eta_V = 5.64305$:

$$\frac{\Delta_{TV}G}{RT} = -\frac{9}{4\vartheta\eta} + \frac{3\eta}{3\eta - 1} - \ln\left(\eta - 1/3\right)$$

The curve of reduced free enthalpy departure computed to reduced temperature $\vartheta = 0.75$ and the horizontal line at -1.13839 intersecting the free enthalpy curve just at reduced molar volumes $\eta_L = 0.489631$ and $\eta_V = 5.64305$ are also shown in **Fig. 2.3**. This perfectly justifies the knowledge that, since temperatures and pressures of the distributing phases are equal, the extremum criterion belonging to given pressure and temperature, i.e. molar free enthalpy, has equal values in the two phases. *Equality of the free* ***enthalpies*** in the two phases is *necessary* condition of phase equilibrium, whereas *minimum of total free* ***energy*** at given temperature and molar volume is *satisfactory* condition of phase equilibrium.

2.2.2 Fugacity coefficient

To check conditions (2.3), one has to calculate molar free enthalpies of the phases. Such a *calculation is always based on some equation of state*.

If the same reference state is applied to both phases then checking equality of the departure is enough. Since the temperatures are equal, the following equality should also be satisfied:

$$\frac{\Delta G^{I}}{RT} = \frac{\Delta G^{II}}{RT} \tag{2.7}$$

Change of free enthalpy of perfect gas at constant temperature:

$$\left(\mathrm{d}G^{\square}\right)_T = \frac{RT}{p}\mathrm{d}p = RT\mathrm{d}\left(\ln p\right)$$

Therefore free enthalpy of perfect gas can be written as:

$$G^{\square}\left(T,p\right) = G^{\square}\left(T,p\right) - G^{\square}\left(T,p^{*}\right) = RT\left(\ln p - \ln p^{*}\right) = RT\ln\frac{p}{p^{*}} \tag{2.8}$$

Departure from perfect gas behavior can *formally* be expressed by substituting some property f in place of pressure p. This property f is called **fugacity** nicked by *Lewis* . Equation

$$G\left(T,p\right) \equiv RT\ln\frac{f}{p^{*}} \tag{2.9}$$

is *definition* of fugacity.

Comparison of (2.8) and (2.9) reveals that (logarithm of) fugacity is closely related to free enthalpy departure $\Delta_{TP}G$:

$$\Delta_{TP}G \equiv G\left(T,p\right) - G^{\square}\left(T,p\right) = RT\ln\frac{f}{p} \equiv RT\ln\varphi \tag{2.10}$$

In the same time, relation (2.10) is definition of **fugacity coefficient** φ:

$$\varphi \equiv \frac{f}{p}$$

Necessary condition of equilibrium (2.7) may thus be written as

$$f^{I} = f^{II} \tag{2.11}$$

or

$$\ln\varphi^{I} = \ln\varphi^{II} \tag{2.12}$$

The form (2.12) is used for modelling phase equilibria of pure materials, the form (2.11) is, if properly expressed (see sections *2.5.1 and 2.5.2*), for that of mixtures.

Logarithm of fugacity coefficient can be easily expressed based on departure calculations:

$$\ln\varphi \equiv \ln\frac{f}{p} \equiv \frac{\Delta_{TP}G}{RT} \equiv \frac{\Delta_{TP}A}{RT} + \left(Z-1\right) \tag{2.13}$$

By substituting relation (2.6), one obtains a formula for the logarithm of fugacity coefficient:

$$\boxed{\ln\varphi(V,T) = -\int\limits_{\infty}^{V}\left[\frac{p(T,v)}{RT} - \frac{1}{v}\right]\mathrm{d}v - \ln Z + Z - 1} \tag{2.14}$$

Using this integral to ***properly shaped equation of state*** gives rise to closed formula for computing logarithm of fugacity coefficient in function of mole fraction V and temperature T.

2.3 A few well known equations of states

Cubic equations

The first cubic equation of state is the famous eqaution of *van der Waals, 1873*:

vdW:

$$p = \frac{RT}{V-b} - \frac{a}{V^2} \tag{2.15}$$

This assumes a hard sphere model, with occupied volume ratio b/V. From this data a relation between sphere radius σ and parameter b can be calculated. The second member expresses the effects of attraction potential. The following relation can be given between the parameters and *critical data*:

$$a = \frac{27}{64}\frac{R^2T_c^2}{p_c}$$

$$b = \frac{RT_c}{8p_c}$$

$$Z_c = \frac{3}{8}$$

Substituting these, one obtains the reduced form of the equation:

vdWr:

$$\pi = \frac{8\,\vartheta}{8\,\eta - 1} - \frac{3}{\eta^2}$$

This a practical application of corresponding states' theorem. However, parameter a is temperature dependent according to experiments: $a = a(T)$.

This equation provides with a qualitative description of behaviour of non-polar materials. Several trials have been made to improve its capability and make it able to use for quantitative modelling as well. Most applied are the following two variants:

Equation *Peng és Robinson (1976)* (abbreviated as PR):

PR:
$$p = \frac{RT}{V-b} + \frac{a(T)}{V^2 + 2bV - b^2}$$

Equation *Redlich - Kwong (1949)* modified by *Soave (1972)* (abbreviated as SRK or RKS):

SRK:
$$p = \frac{RT}{V-b} + \frac{a(T)}{V^2 + bV}$$

Function $a = a(T)$ can be decomposed in both equations as

$$a(T) = a_c \alpha(T)$$

and parameters a_c, b can be expressed by critial data. For example, in equation SRK:

$$a_c = 0.42748 \frac{R^2 T_c^2}{p_c}$$
$$b = 0.08664 \frac{R T_c}{p_c}$$

Approximate temperature dependence for non-polar materials is:

$$\alpha(T) = \left[1 + \left(1 - \sqrt{T_r}\right) f(\omega)\right]^2$$

where ω is the so-called *acentricity factor* (a material property), whose definition incorporates a ratio of the vapor pressure measured at 70% of critical temperature to the critical pressure:

$$\omega \equiv -\log_{10}\left(\frac{p^\circ_{T=0.7T_c}}{p_c}\right) - 1$$

and shape of function $f(\omega)$ is:

$$f(\omega) = 0.48 + 1.57\omega - 0.17\omega^2$$

Another function form is applied to polar materials.

Cubic equations serve well around and below critical point and, sometimes, even at higher pressures. Below critical point they are able to trace vapor pressure and mole fraction of vapour but model liquid density with significant error only. Liquid density calculated with them must be corrected somehow.

Virial equations

Not too dense gases behave as approximately perfect gases and their non-ideality can be described by expanding them according to density around the properties of perfect gas. In its ususal form, compressibility

$$Z \equiv \frac{pV}{RT}$$

is expanded according to ρ around $\rho = 0$ ($V = \infty$):

$$Z \equiv \frac{pV}{RT} = 1 + \sum_{n=2}^{\infty} B_n \left(\frac{1}{V}\right)^{n-1}$$

The strange indexing and exponents are explained by the following variant:

Virial:

$$\frac{p}{RT} = \sum_{n=1}^{\infty} B_n \left(\frac{1}{V}\right)^{n} \equiv \frac{1}{V} + \sum_{n=2}^{\infty} B_n \left(\frac{1}{V}\right)^{n}$$

The ***virial coefficients are independent of ρ (of V)*** because of expansion around $\rho = 0$ ($V = \infty$), and thus they are dependent on temperature only. The first coefficient is, naturally, $B_1 = 1$; this is the only coefficient of perfect gas. The other coefficients are

$$B_n \equiv \frac{1}{n-1} \left(\frac{\partial^{n-1}}{\partial \rho^2} \left(\frac{p}{\rho RT} \right) \right)_{\rho = 0}$$

Thus the value and temperature dependence of the coefficients can be determined from empirical p–V–T (i.e. p–ρ–T) data. The coefficients can be calculated in case of simple models like hard sphere. The coefficients are temperature dependent: $B_n(T)$. Second virial coefficient $B_2(T)$ has special significance because this is the first coefficient expressing departure from the behaviour of perfect gases.

Virial equations can be expanded to arbitrary degree, and particular shape variants may be inserted. For example, equation BWR (*Benedict-Webb-Rubin, 1940*) applies even an exponential member:

BWR:

$$p = \frac{RT}{V} + \frac{B_0 RT - A_0 - \frac{C_0}{T^2}}{V^2} + \frac{bRT - a}{V^3} + \frac{a\alpha}{V^6} + \frac{C}{T^2 V^3} \left(1 + \frac{\gamma}{V^2}\right) \exp\left(-\frac{\gamma}{V^2}\right)$$

This equation has 7 parameters and fits well to the data at higher densities and even for polar molecules.

Hard sphere and hard body equations

Liquids and dense fluids are approximately incompressible hence the hard sphere model can be applied to them with success. Equation of state of *Percus and Yevick*:

PY(a):

$$\frac{p}{RT} = \frac{6}{\pi} \left[\frac{\zeta_0}{1 - \zeta_3} + \frac{3\zeta_1 \zeta_2}{(1 - \zeta_3)^2} + \frac{3\zeta_2^3 (1 - \zeta_3)}{(1 - \zeta_3)^3} \right]$$

Another version of the *Percus-Yevick* model:

PY(b):

$$\frac{p}{RT} = \frac{6}{\pi} \left[\frac{\zeta_0}{1 - \zeta_3} + \frac{3\zeta_1 \zeta_2}{(1 - \zeta_3)^2} + \frac{3\zeta_2^3}{(1 - \zeta_3)^3} \right]$$

The *Carnahan and Starling* equation of state:

CS:

$$\frac{p}{RT} = \frac{6}{\pi}\left[\frac{\zeta_0}{1-\zeta_3} + \frac{3\zeta_1\zeta_2}{(1-\zeta_3)^2} + \frac{\zeta_2^3(3-\zeta_3)}{(1-\zeta_3)^3}\right]$$

In the above equations

$$\zeta_n \equiv \frac{\pi\sigma^n}{6V} \equiv \frac{\pi\sigma^n\rho}{6}$$

The only parameter of these equations is the radius σ of the hard sphere.

Non-spherical molecular shapes can be taken into account with some shape factor.

2.4 Vapor-liquid distribution of pure materials

Whether there is phase distribution at given state $[p,T]$ can be easily determined in simple cases based on an equation of state expressed to pressure as (2.5), see (**Fig. 2.4**). If there is no more than one real V molar volume belonging to the given T and P then separation into different molar volumes cannot occur.

Cubic equations of states are called cubic because expressed to zero they are polinomials of third degree in V. For a given pair of T and P, the three molar volume roots can be analytically expressed. One of them is always real; the other two are either real or form a pair of conjugated complex roots. That is, there is either one real root or there are three real roots. Two or three of them may be equal in the latter case; then we say in a sloppy sense that there are only two real roots, or only one.

If there is only one real V root then the material is present in a single homogeneous phase at the given state T and p. If there are three ($V_1 < V_2 < V_3$) or two ($V_1 < V_3$) different real roots then one has to check if the fluid separates to a (liquid) state $[p,T,V_1]$ and a (vapor) state $[p,T,V_3]$; and if yes then in what phase ratio. Whether the two hypothetical phases of states $[p,T,V_1]$ and $[p,T,V_3]$ are in mutual equilibrium can be checked by criteria (2.3) i.e. (2.11) and (2.12), and calculating fugacity coefficients with (2.13). The calculation and check consists of the following steps:

Algorithm VL0: ***Check vapor-liquid phase distribution of pure materal***
Given: temperature T and pressure p
Looked for: *number* and molar volumes V of phases

1. If $T > T_c$ (critical temperature) then there is only one supercritical phase. Stop.
   Otherwise:
2. Roots V are computed at given p and T. In case of a cubic equation, the results is either a single root V or real roots $V_1 < V_2 < V_3$ or only real roots $V_1 < V_3$.

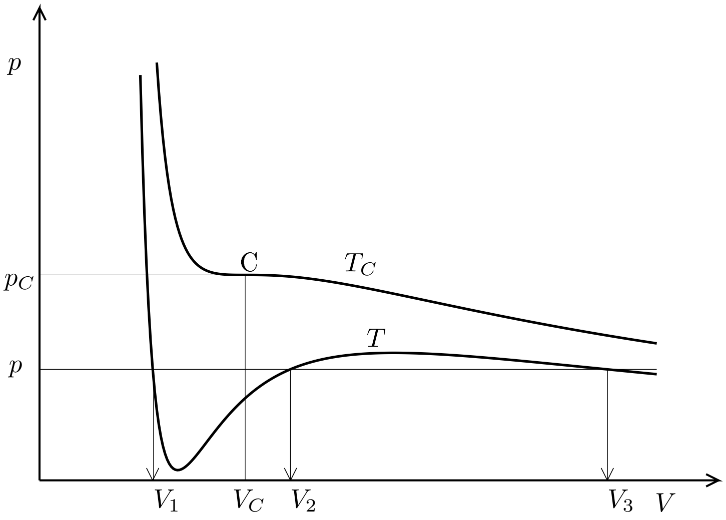

Figure 2.4: Roots V of an isotherm of a cubic EOS at a given p

3. If there is only a single root V then there is no phase distribution. If V is smaller than the critical V_c then the homegeneous phase is considered liquid, otherwise as gas or vapor. Stop.
   If there are at least two real roots then:
4. Root V_1 is considered as a hypothetical liquid molar volume, root V_3 as a hypothetical vapor molar volume.
5. Usng the actual form of equation (2.13), calculate the logarithm of the fugacity coefficients belonging to the hypothetical phases:

$$\ln\varphi_1 \Leftarrow \ln\varphi\left(V_1, T\right) \tag{2.16a}$$

$$\ln\varphi_3 \Leftarrow \ln\varphi\left(V_3, T\right) \tag{2.16b}$$

6. If the two calculated fugacity coefficients are approximately equal (their absolute difference is smaller than a predefined small positive number) then roots V_1 and V_3 are accepted as molar volumes of the respective phases.
   Molar ratio of the phases is not determined by p and T themselves. Another property of the system, for example gross molar volume or enthaly or inernal energy, should also be known for determining phase ratio. Stop.
7. If, otherwise, the two coefficients are different then the material does not separate into two phases at the given state p and T, according to the

model. (The obtained molar volume roots are not in mutual equilibrium.) If $\varphi_1 < \varphi_3$ then then the material is considered as homogeneous liquid, otherwise gas or vapor. Stop.

For a non-cubic equation of states the molar volume roots are determined numerically.

***Algorithm* VL0** is usually not applied in itself. (No need for that.) It may occur as a part of determining boiling point or dew point. The boiling point temperature T_B at given pressure p can be determined by ***Algorithm* T0** :

Algorithm T0: *Boiling point of pure material*
Given: pressure p
Looked for: boiling (dew) point temperature T_B

1. If $p > p_C$ (critical pressure) then there is no boiling point. Stop.
2. Guess the boiling point with some $T < T_C$.
3. Carry out ***Algorithm* VL0**.
4. If equilibrium phase distribution is verified by ***Algorithm* VL0** then $T_B \Leftarrow T$. Stop.
5. Otherwise modify value of T, and return to step 3.

Modification of T in step 5 of ***Algorithm* VL0** is an open question. Difference or ratio or difference of logarithms of the fugacity coefficients calculated wit equations (2.16a) and (2.16b) changes monotically in T near the root; thus the simplest way is applying bisection root search method.

Boiling (or dew) point pressure p_B can also be looked for at a given temperature T. In this case value of p_B is guessed with some $p < p_C$, and equality of fugacities is checked in function of p:

Algorithm P0: *Vapor tension of pure liquid*
Given: temperature T
Looked for: boiling (dew) point pressure p_B

1. If $T > T_C$ then there is no boiling point. Stop.
2. Guess boiling pressure by some $p < p_C$.
3. Carry out ***Algorithm* VL0**.
4. If equilibrium phase distribution is verified by ***Algorithm* VL0** then $p_B \Leftarrow p$. Stop.
5. Otherwise modify value of p, and return to step 3.

These algorithms are rarely used in practice because the vapor pressure function of T is normally known and can be approximated by some function. For example, the *Antoine* equation (derived from eqation *Clausius-Clapeyron*):

Antoine:

$$\ln p = A - \frac{B}{T + C}$$

where A, B, and C are parameters fitted to measured data.

However, these algorithms are applied in modified forms for computing bubble point and dew point temperatures and pressures of mixures.

2.5 Formal thermodynamics of mixtures

2.5.1 Ideal mixtures

Ideal mixture of real materials (non-ideal i.e. imperfect gases) is *defined* by the property that each component's departure from perfect gas behaviour is independent of presence of the other components. In other words, for a given total number N of all molecules behaviour of the N_i molecules of kind i does not depend on the kinds of the other $N - N_i$ molecules. They produce the same (partial) pressure and internal energy as in pure form, given the molar volume and temperature. Therefore, internal energy and pressure are sums of the properties of the pure constituents, and the specific (molar) internal energy is the average of that of pure components:

$$U^{id}\left(V, T, \boldsymbol{y}\right) = \sum_{i}^{c} y_i U_i^{\circ}\left(V, T\right)$$

$$p^{id}\left(V, T, \boldsymbol{y}\right) = \sum_{i}^{c} y_i p_i^{\circ}\left(V, T\right)$$

where superscript ° refers to pure material as a *reference state of mixing. This is usually the pure material in case of mixing non-electrolites or rather small solubilitiy, but can be a different state when dealing with electrolite solutions, absorption, or solid materials.* (see also: *2.5.4 and 2.5.5*)

At given pressure the molar volumes sum up:

$$V^{id}\left(p, T, \boldsymbol{y}\right) = \sum_{i}^{c} y_i V_i^{\circ}\left(p, T\right)$$

Entropy has a mixing contribution even in ideal mixture of perfect gases. This has a consequence to free energy and free enthalpy as well:

$$S^{id} = \sum_i^c y_i S_i^\circ - R \sum_i^c y_i \ln y_i$$
$$A^{id} = \sum_i^c y_i A_i^\circ + RT \sum_i^c y_i \ln y_i$$
$$G^{id} = \sum_i^c y_i G_i^\circ + RT \sum_i^c y_i \ln y_i$$

Since free enthalpy (Gibbs energy) is sum of, and the molar free enthalpy is the average of, the chemical potentials, the chemical potential of a component in an ideal mixture is:

$$\mu_i^{id} = \mu_i^\circ + RT \ln y_i$$

where, according to relation (2.9) valid for the chemical potential of pure material:

$$\mu_i^\circ \equiv G_i^\circ \equiv RT \ln \frac{f_i^\circ}{p^*}$$

(The numerator contains fugacity of the pure component at the actual temperature and pressure.)

Therefore, **partial fugacity** f_i^{id} of components ***in ideal mixture*** can be defined as:

$$\mu_i^{id} \equiv RT \ln \frac{f_i^{id}}{p^*} \equiv RT \ln \frac{f_i^\circ}{p^*} + RT \ln y_i \equiv RT \ln \frac{y_i f_i^\circ}{p^*} \qquad (2.17)$$

Thus, partial fugacity ***in ideal mixture*** is:

$$f_i^{id} = y_i f_i^\circ \equiv y_i \varphi_i^\circ p$$

Hence **partial fugacity coefficient** φ_i^{id} ***in ideal mixture*** equals fugacity coefficient of the pure material:

$$\varphi_i^{id} \equiv \frac{f_i^{id}}{y_i p} = \frac{f_i^\circ}{p} \equiv \varphi_i^\circ$$

In ideal mixture of perfect gases pure component's fugacity equals the pressure, and then the value of fugacity coefficient is unity:

$$f_i^{\square,id} = y_i p$$
$$\varphi_i^{\square,id} = \varphi_i^{\square,\circ} \equiv 1$$

2.5.2 Real mixtures of non-electrolites

Chemical potential of a component ***in real mixture*** is also written according to (2.17):

$$\mu_i \equiv RT \ln \frac{f_i}{p^*} \qquad (2.18)$$

but now omitting superscript ^{id} means that **partial fugacity** f_i valid in real mixtures is applied. This is definition of f_i. Pressure p^* is always the reference pressure. (This can be the actual pressure as well.) What kind of model is used to obtain from data p–V–T–$\boldsymbol{y}$ a partial fugacity function of the same variables is essential in practice. Partial fugacity defined by (2.18) can be derived in two different ways:

1. *Instead of* fugacity coefficient φ_i° of pure material and fugacity coefficient φ_i^{id} valid in ideal mixture, a **partial fugacity coefficient** φ_i valid in real mixture is introduced:

   $$f_i \equiv y_i \varphi_i p$$

   By introducing partial fugacity coefficient φ_i, we cannot refer to ideal mixture as a reference point. Instead, the reference is the perfect gas; the same idea that used for pure real gases.

   Partial fugacity coefficient depends not just on molar volume and temperature but composition as well. This is why it is introduced. Therefore, for calculating partial fugacity coefficient, composition should take place in the equation of state:

   $$p = p\,(V, T, \boldsymbol{y})$$

   The equation analogous to (2.14) is:

   $$\ln \varphi_i\,(V, T, \boldsymbol{y}) = -\int\limits_{\infty}^{V} \left[\frac{1}{RT} \left(\frac{\partial\,[np\,(T, v, \boldsymbol{y})]}{\partial n_i} \right)_{T,(nv),\widehat{\boldsymbol{n_i}}} - \frac{1}{v} \right] \mathrm{d}v - \ln Z \quad (2.19)$$

   where expression

   $$y_i \equiv \frac{n_i}{\sum\limits_{j=1}^{c} n_j} \equiv \frac{n_i}{n}$$

   is, by definition, substituted to the place of mole fractions. Not molar volume v but total volume nv is kept constant while differentiating; here v is molar volume index of integration.

   Applicability of fugacity coefficient depends on the applied equation of states. Equations of states are usually based on ideas independent of composition. Apart from a very few exceptions, like hard sphere equations, they do not give any hint about composition dependence. The usual way to follow is introducing composition dependent parameters. That is, should the equation contain fitted parameters a, b, c, …etc., i.e. its form is

   $$p = p\,(V, T; a, b, c, \dots) \quad (2.20)$$

   then parameters a, b, c, …etc. are considered composition dependent:

   $$a = a(\boldsymbol{y}), \quad b = b(\boldsymbol{y}), \quad c = c(\boldsymbol{y}), \quad \text{etc.} \quad (2.21)$$

Shape of these functons can be approximately estimated, and even calculated for some component pairs at moderate density from the values of pure component parameters a, b, c, …etc. The approach given by formulas (2.20) and (2.21) essentially links the actual values of parameters a, b, c, …etc. to composition $\boldsymbol{y}$ and thus forms a fictious pure material with the same equation of state (2.20) but with the such calculated parameters. That is why this approach is also called **single fluid model**. It is also called ***model of conformal mixtures*** because is is applicable to mixture of such components only that can be modelled with the same equation of states.

2. As a measure of departure from ideality, **activity coefficient** γ_i is introduced:

$$f_i \equiv \gamma_i f_i^{id} \equiv \gamma_i y_i f_i^{\circ} \tag{2.22}$$

Since activity coefficient expresses deviation from ideal mixture, its purpuse is to account for the effect of composition only and not for the relation p–V–T, i.e. not for the effect of change in molar volume to the other properties of the mixture. This is a simplification or neglection applicable to ***liquid mixtures*** only (*and, sometimes to solid mixtures*) because their molar volume is approximately constant.

Effects of temperature cannot be neglected but, since according to experience pressure dependence of activity coefficient at low pressure or fairly narrow pressure domain is small, activity coefficient models usually express the effects of composition and temperature only:

$$\gamma_i \approx \gamma_i\left(\boldsymbol{x}, T\right) \tag{2.23}$$

where liquid mole fractions are denoted by $\boldsymbol{x}$ to distinguish them from vapor mole fractions $\boldsymbol{y}$. This notation is applied all over our text.

Activity coefficient models can be derived from models of excess mixing free enthalpy (Gibbs energy) models

$$\Delta^E G \equiv G - G^{id} = G - \sum_{i=1}^{c} x_i G_i^{\circ} - RT \sum_{i=1}^{c} x_i \ln x_i$$

where superscript ^{E} refers to *excess*.

Partial fugacity coefficient and activity coefficient serve for modelling the same partial fugacity; their values are interdependent:

$$y_i \gamma_i f_i^{\circ} \equiv f_i \equiv y_i \varphi_i p$$

Hence, activity coefficient can be calculated from equation of states, and partial fugacity can be calculated from an activity coefficient model and a model of pure component fugacity.

Activity itself is a term defined as ratio of partial fugacity to the reference state fugacity:

$$a_i \equiv \frac{f_i}{f_i^\circ}$$

This can be expressed as a product of activity coefficient and mole fraction:

$$a_i \equiv \gamma_i x_i$$

Thus, chemical potential can be written this way as well:

$$\mu_i = \mu_i^\circ + RT \ln a_i$$

2.5.3 Corrections to the approximating activity coefficient

When activity coefficient is applied to express fugacity of a component in a liquid mixture, two kinds of neglections are made:

1. Because of the form (2.23) of activity coefficient models, they do not express pressure dependence of the coefficient. Instead, the model parameters are determined at some reference pressure p^{aref}, e.g. atmosphesric pressure. The activity coefficient is valid at this pressure only. The fugacity calculated at the reference pressure

   $$f_i^L\left(T, p^{aref}, \boldsymbol{x}\right) \equiv \gamma_i\left(T, \boldsymbol{x}\right) x_i f_i^{\circ,L}\left(T, p^{aref}\right) \tag{2.24}$$

   can be corrected to the actual pressure as

   $$f_i^L\left(T, p, \boldsymbol{x}\right) \equiv f_i^L\left(T, p^{aref}, \boldsymbol{x}\right)\xi \tag{2.25}$$

   i.e. by multiplying it with a correction factor ξ that in principle can be calculated as

   $$\ln \xi = \int\limits_{p^{aref}}^{p} \frac{V_i^L\left(T, \pi, \boldsymbol{x}\right)}{RT} \mathrm{d}\pi \tag{2.26}$$

   An equation of states would be needed to calculate correction ξ but an approximation is applied instead, see below.

2. Activity coefficients are usually applied to liquids. Fugacity of pure component *in liquid form* is to be computed. This is achieved so that it is first determined at the temperature of the system and at the vapor presure of the pure component. The pure liquid is in equilibrium with its vapor at the systems's temperature T and at ***equilibrium vapor pressure*** p_i°, thus their fugacities are equal. This fugacity is then corrected to the reference pressure p^{aref} applied in relations (2.24) - (2.26):

   $$\begin{aligned} f_i^{\circ,L}\left(T, p_i^\circ\left(T\right)\right) &= f_i^{\circ,V}\left(T, p_i^\circ\left(T\right)\right) = \varphi_i^{\circ,V}\left(T, p_i^\circ\left(T\right)\right) p_i^\circ\left(T\right) \\ f_i^L\left(T, p^{aref}\right) &= \eta f_i^{\circ,L}\left(T, p_i^\circ\left(T\right)\right) \end{aligned}$$

where correction factor η could, in principle, be calculated with the following integral:

$$\ln \eta = \int\limits_{p_i^\circ}^{p^{aref}} \frac{V_i^{\circ L}(T, \pi)}{RT} \mathrm{d}\pi \tag{2.27}$$

An equation of states would be needed to calculate correction η but an approximation is applied instead, see below.

Molar volumes occuring in relations (2.26) $\ln \xi$ and (2.27) for $\ln \eta$ can be considered approximately constant in liquid state; thus the integral can be analytically expressed. An additional approximation is that partial molar volume V_i^L occuring in (2.26) does not significantly differ from pure component molar volume $V_i^{\circ L}$ used in (2.27); thus pure component molar volume $V_i^{\circ L}$ may be substituted in the form of $\ln \xi$. In this way a combined correction is obtained, called ***Poynting*** correction after its inventor:

$$\ln \eta\xi \approx \int\limits_{p_i^\circ}^{p} \frac{V_i^{\circ L}(T, \pi)}{RT} \mathrm{d}\pi \approx \frac{(p - p_i^\circ(T))\, V_i^{\circ L}(T, p)}{RT}$$

$$\ln \mathcal{P} \equiv \frac{(p - p_i^\circ)\, V_i^{\circ L}(T, p)}{RT}$$

Partial fugacity of a component in liquid mixture is thus:

$$f_i^L(T, p, \boldsymbol{x}) \approx \gamma_i(T, \boldsymbol{x})\, x_i \varphi_i^{\circ, V}(T, p_i^\circ(T))\, p_i^\circ(T)\, \mathcal{P} \tag{2.28a}$$

or in short:

$$\boxed{f_i^L \approx \gamma_i x_i \varphi_i^\circ p_i^\circ \mathcal{P}} \tag{2.28b}$$

Correcting the neglection applied in deriving *Poynting* correction would be corretion of a correction, that is why the first corretion works well. Reference pressure p^{aref} is cancelled out at integration; it does not occur in realations (2.28a)-(2.28b). Thus it is not needed, but only he actual pressure and the vapor pressure of the actual component at the systems's temperature. Vapor fugacity is to be taken at system's temperature and the vapor pressure; its value is approximately unity at low pressure (near perfect gas behaviour and no vapor phase association). It can be neglected under 5 bar and non-associating vapor in practice.

2.5.4 Assymmetric activity reference and Henry constant

For computing partial fugacity in liquid mixture with activity coefficient, equilibrium vapor pressure p_i° of pure components as functions of temperature must be known. This vapor pressure exists under critical temperature only. It happens, however, that a liquid mixture contains components with lower critical temperature that the actual system temperature (*non-condensing gases*). Such typical case is solution of supercitical gases in liquids, e.g. air components in water at atmospheric circumstances.

In these cases the partial properties of the non-condensing components cannot be related to their pure form because they do not exist in liquid form as pure materials at the system's temperature and pressure. Instead, we choose the other extreme as reference state: the lack of that component. In practice this is considered as a limit, called infinite dilution. Thus, in contrast to the so-called *symmetric convention* reference applied in (2.22):

$$\lim_{x_i \to 1} \gamma_i = 1 \tag{2.29}$$

we use here an *asymmetric convention* reference:

$$\lim_{x_i \to 0} \hat{\gamma}_i = 1 \tag{2.30}$$

The pure component reference state denoted by superscript $^\circ$ is substituted by the other, infinite dilution, reference state denoted by superscript $^\infty$.

Since this convention is used just for dilute solutions, the activity coefficient can frequently be considered as unity.

For describing solution of non-condensing gases in liquids, instead of $f_i^{\circ,L}$ we apply $f_i^{\infty,L}$ conventionally denoted as H_i. Its name is *Henry* 'constant', although it is temperature dependent and slightly composition dependent. H_i is defined as a limit:

$$f_i^{\infty,L} \equiv H_i \equiv \lim_{x_i \to 0} \frac{f_i^L}{\hat{\gamma}_i x_i} \equiv \lim_{x_i \to 0} \frac{f_i^L}{x_i}$$

Composition dependence of *Henry* 'constant' does not mean a dependence on the mole fraction of the actual component, being anyhow a rather small value, but on the mole ratios of the ***condensing*** solvent components if there are more than one present simultaneously. In the same way as at the symmetric case, the *Henry* 'constant' could be determined first at some reference pressure p^{aref} and then the partial fugacity would be

$$f_i^L \left(p^{aref}\right) = \hat{\gamma}_i x_i H_i \left(p^{aref}\right) \tag{2.31}$$

But *Henry* 'constant' is measured at some given temperature, and this temperature determines the ***total partial pressure as a sum of partial pressures of all the condensing components*** j. Thus, *Henry* 'constant' is known at this pressure:

$$H_i(T) = H_i\left(p^\circ_{\forall j}\right)$$

where

$$p^\circ_{\forall j}(T) \equiv \sum_{\forall j\text{: cond.}} x_j p_j^\circ(T)$$

(j is a running index for all the condensing components.) In spite of its small concentration, the non-condensing gas component may give rise to large pressure in the vapor space over the liquid, i.e. total vapor pressure $p^{\circ}_{\forall j}$ can be significantly different from the system's pressure. Therefore the *Henry* 'constant' measured or calculated at temperature T must be recalculated from $p^{\circ}_{\forall j}$ to the reference pressure p^{aref} occuring in (2.31):

$$\ln f_i^L\left(p^{aref}\right) = \ln\hat{\gamma}_i + \ln x_i + \ln H_i\left(T\right) + \int\limits_{p^{\circ}_{\forall j}}^{p^{aref}} \frac{V_i^{\infty}}{RT}\mathrm{d}\pi$$

This should again be recalculated to the actual pressure p and lumping the integrals leads to:

$$\ln f_i^L\left(p\right) = \ln\hat{\gamma}_i + \ln x_i + \ln H_i\left(T\right) + \int\limits_{p^{\circ}_{\forall j}}^{p} \frac{V_i^{\infty}}{RT}\mathrm{d}\pi$$

p^{aref} is again cancelled, and finally:

$$\boxed{\ln f_i^L\left(p\right) \approx \ln x_i + \ln H_i\left(T\right) + \frac{\left(p - p^{\circ}_{\forall j}\right)V_i^{\infty}}{RT}}$$

where the last member is *Poynting* correction of assymmetric convention.

2.5.5 Electrolite solutions. Osmotic coefficient

Concentration of electrolites, first of all in water, are conventionally expressed not in mole fractions but in molalities m_i:

$$m_i \equiv \frac{n_i}{n_W M_W}$$

where $_W$ is index of the solvent, letter W referring to water (Wasser), and M_W is molar weight (molar mass) of he solvent expressed in kg/mol. That is, unit of molality is (mol solute)/(kg solvent). Activity in chemical potential

$$\mu_i = \mu_i^{ref} + RT\ln a_i$$

is now expressed in the form of

$$a_i = \tilde{\gamma}_i m_i \tag{2.32}$$

where upper tide denotes that the activity belongs to molality and not to mole fraction. Reference state of the chemical potential of the solute is not pure component but infinite dilution:

$$\lim_{m_i\to 0}\tilde{\gamma}_i = 1 \tag{2.33}$$

Hence activity in ideal mixture can be expressed as:

$$\mu_i^{id} = \mu_i^{\infty} + RT \ln m_i$$

where superscript $^{\infty}$ refers to state of infinite dilution.

Molality of the solvent is nonsense because it would be constant $1/M_W$ according to the definition. Therefore activity of the solvent is expressed in the symmetric convention, bond to mole fractions, according to (2.29):

$$\mu_W^{id} = \mu_W^{\circ} + RT \ln x_W$$

Logarithm of $1 - h$ when $|h| \ll 1$ can be well approximated with $-h$:

$$\ln x_W = \ln \left(1 - \sum_{i \neq W} x_i\right) \approx - \sum_{i \neq W} x_i \qquad [= -(1 - x_W)]$$

thus chemical potential of the solvent in the solution at low concentrations can be written as

$$\mu_W^{id} = \mu_W^{\circ} - RT\,(1 - x_W) = \mu_W^{\circ} - RT \sum_{i \neq W} x_i$$

The following equality is approximately true at low concentrations and exactly satisfied in the limit of infinite dilution:

$$x_i \approx \frac{n_i}{n_W}$$

hence chemical potential of the solvent in dilute electrolite solutions can be written as

$$\mu_W^{id} = \mu_W^{\circ} - RT \frac{1}{n_W} \sum_{i \neq W} n_i$$

and Gibbs energy (free enthalpy) of such a solution as a function of mole numbers can be written as

$$g^{id} = n_W \left[\mu_W^{\circ} - RT \frac{1}{n_W} \sum_{i \neq W} n_i\right] + \sum_{i \neq W} \left[n_i \left(\mu_i^{\infty} + RT \ln \frac{n_i}{n_W M_W}\right)\right]$$

If the solution is neither infinitely dilute or not ideal then activity coefficients defined in relations (2.32) and (2.33), using to molality, and denoted by upper tide, are used for solute components and a so-called **osmotic coefficient** ϕ is used for the solvent:

$$\boxed{\mu_i = \mu_i^{\infty} + RT \ln\left(\tilde{\gamma}_i m_i\right)}$$

$$\boxed{\mu_W = \mu_W^{\circ} - RT \frac{\phi}{n_W} \sum_{i \neq W} n_i}$$

$$\boxed{g = n_w \left[\mu_W^\circ - RT\frac{\phi}{n_W} \sum_{i \neq W} n_i \right] + \sum_{i \neq W} \left[n_i \left(\mu_i^\infty + RT \ln \frac{\tilde{\gamma}_i n_i}{n_W M_W} \right) \right]}$$

Temperature and composition dependence of these acticity coefficients and osmotic coefficient can be derived from excess mixing Gibbs energy models.

2.6 Mixture equations of states

Hard sphere mixtures

The equations can be derived for mixtures of hard spheres of different radiuses. Only the expression of ζ is changed:

$$\zeta_n = \frac{\pi}{6} \sum_i \frac{\sigma_i^n}{V_i} \equiv \frac{\pi}{6} \sum_i (\sigma_i^n \rho_i)$$

Conformal mixtures, mixing rules

Conformal mixtures or conformal solutions are those whose components can be well modelled with the same form of equation of states. Let parameters a_i, b_i, c_i, etc. occur in the equation of state of each component i, then the so-called mixture parameters a, b, c, etc. occuring in the equation of states of the mixture can be given as functions of composition and the parameters of all the components. Such functions are called **mixing rules**.

Such a model can be successful if interactions between components are similar for like and unlike component pairs.

Dimension of arameter b of *cubic equations* is molar volume; this parameter is characteristic to the volume of the molecules; therefore arithmetic mean is a good approximation, as suggested by *van der Waals*:

$$b = \sum_i x_i b_i \tag{2.34}$$

If the molecules do not behave as hard spheres but as if they were compressible then parwise parameters b_{ij} are to be introduced, e.g. like this:

$$b = \sum_i \sum_j x_i x_j b_{ij}$$

For parameter a of cubic equations the following rule was suggested again by *van der Waals*:

$$a = \sum_i \sum_j x_i x_j a_{ij} \tag{2.35}$$

where a_{ij} is pairwise attraction parameter.

Pairwise indexed parameters can be estimated for *near ideal mixtures*; several suggestions are known for them. For example, b can be estimated by quadratic or cubic mean:

$$b_{ij} = \left(\frac{\sqrt{b_i} + \sqrt{b_j}}{2} \right)^2$$

$$b_{ij} = \left(\frac{\sqrt[3]{b_i} + \sqrt[3]{b_j}}{2} \right)^3$$

or some similar estimation. Applying arithmetic mean gives rise to rule (2.34).

The usual estimation parameter a is

$$a_{ij} = \sqrt{a_i a_j}$$

(although other forms could also be used here). If the mixture behaves a bit differently then the cross-effects can be taken into account with a so-called *interaction parameter* k_{ij}:

$$a_{ij} = (1 - k_{ij}) \sqrt{a_i a_j} \tag{2.36}$$

This interaction parameter is fitted to measurement data.

For computing phase equilibria, effects of composition is especially important. Therefore several suggestions are known for describing composition dependence of the interaction parameter. Such suggestions are, for example:

$$k_{ij} = k_{ij}^{\circ} + m(x_i - x_j)$$

$$k_{ij} = x_i h_{ij} + x_j h_{ji}$$

$$k_{ij} = k_{ij}^{\circ} + n_{ij}(x_i + x_j) \frac{h_{ij} x_i - h_{ji} x_j}{h_{ij} x_i + h_{ji} x_j}$$

The latter one proved be a good rule for calculating vapor-liquid equilibria of ternary systems.

Another technique for describing composition dependence is application of so-called local concentrations by modifying rule (2.35):

$$a = \sum_i \sum_j x_i w_{ji} a_{ji}$$

$$w_{ji} \equiv \frac{\tau_{ij} x_j}{\sum_k \tau_{ik} x_k}$$

Form of the departure functions necessarily change with the applied mixing rules. For example, if rules (2.34) and (2.36) are applied in the vdW equation of states (2.15) then the form of fugacity coefficient is

$$\ln \varphi_i = -\ln \left(Z \frac{V - b}{V} \right) + \frac{b_i}{V - b} - \frac{2\sqrt{a_i}}{VRT} \sum_j \left[x_j (1 - k_{ij}) \sqrt{a_j} \right]$$

Virial equation can also be given in mixture form but applying the idea of conformal mixtures is easier. The second virial coefficient can be given like (2.35):

$$B_2(T) = \sum_i \sum_j x_i x_j B_{2,ij}(T)$$

2.7 Mixing excess models

Excess properties and activity coefficient are linked by the following general thermodynamic relation:

$$RT \ln \gamma_i = \left(\frac{\partial n \Delta^E G}{\partial n_i} \right)_{T,p,\widehat{N}}$$

According to a widely applied idea, density dependence of excess Gibbs energy *can be neglected in first approximation* because liquid density dependence on pressure and temperature is much smaller that that of gases. *In this case mixing Gibbs energy excess can be approximated by mixing Helmholtz energy excess.* Both are related to mixing heat and spatial arrangement of molecules. Mixing heat is caused by interaction of unlike molecules being diferent from interaction of like molecules. Spatial arrangement means that concentration of molecules of different kinds can be different around a molecule of a kind, although the multitude of molecules are arranged in the space with homogeneous density. This concentration in the neighborhood of a selected species can be perfectly uniform (*random arrangement*) or something else (*partially or perfectly ordered arrangement*). Perfect order cannot occur in liquids (but is achieved as an average in crystals).

2.7.1 Numerical expansions

According to a possible approach, excess free enthalpy is determined by ratios of so-called *effective molar volumes* q_i. This is quantity proportional to co-ordination number; it increases with the diameter of the molecules bot not necessarily proportionally. By weighting these effective molar volumes with mole fractions one obtains *effective volume fractions*:

$$\Phi_i \equiv \frac{x_i q_i}{\sum_j x_j q_j}$$

According to *Wohl (1946)*, the excess can be expanded according to these fractions:

$$\frac{\Delta^E G}{RT \sum_i q_i \Phi_i} = \sum_{ij} a_{ij} \Phi_i \Phi_j + \sum_{ijk} a_{ijk} \Phi_i \Phi_j \Phi_k + \cdots$$

where besides effective volume fractions, interaction parameters a of two, three, etc. indexes are to be fitted to measurement data. Stopping at three indexes, the equation for binary mixtures is:

$$\frac{\Delta^E G}{RT} = (q_A x_A + q_B x_B)(2\Phi_A \Phi_B a_{AB} + 3\Phi_A^2 \Phi_B a_{AAB} + 3\Phi_A \Phi_B^2 a_{ABB})$$

or by transforming the parameters:

$$\frac{\Delta^E G}{RT} = A_{AB}\Phi_B^2 x_A + A_{BA}\Phi_A^2 x_B$$

Ratio of effetive molar volumes can be estimated. By approximating q_B/q_A with A_{BA}/A_{AB}, one obtains equation (2.38) of *van Laar*. By approximating q_B/q_A with the ratio of real volumes V_B/V_A, one obtains *Scatchard-Hammer* equation (*1935*):

$$\begin{aligned}\ln\gamma_A =& \Phi_B^2\left[A_{AB} + 2\Phi_A\left(A_{BA}\frac{\rho_B}{\rho_A} - A_{AB}\right)\right]\\ \ln\gamma_B =& \Phi_A^2\left[A_{BA} + 2\Phi_B\left(A_{AB}\frac{\rho_A}{\rho_B} - A_{BA}\right)\right]\end{aligned}$$

By approximating q_B/q_A with 1, one obtains *Margules* equation (*1895*):

Margules:

$$\boxed{\begin{aligned}\ln\gamma_A =& x_B^2\left[A_{AB} + 2x_A\left(A_{BA} - A_{AB}\right)\right]\\ \ln\gamma_B =& x_A^2\left[A_{BA} + 2x_B\left(A_{AB} - A_{BA}\right)\right]\end{aligned}}$$

In contrast to the above models *Redlich and Kister (1948, 1952)*, after *Guggenheim*, applied a mathematical expansion without any physical reasoning. For binary mixture:

$$\frac{\Delta^E G}{RT} = x_A x_B\left[A_{AB} + B_{AB}(x_A - x_B) + C_{AB}(x_A - x_B)^2 + D_{AB}(x_A - x_B)^3 + \cdots\right] \tag{2.37}$$

For multicomponent mixture the series is similar but contains much more parameters to be fitted to measured data.

2.7.2 Derivation from equation of states

Van Laar derived a formula for excess free energy from the *van der Waals* equation of state (2.15):

van Laar:

$$\boxed{\begin{aligned}\ln\gamma_A &= \frac{A_{AB}x_B^2}{\left(\dfrac{A_{AB}}{A_{BA}}x_A + x_B\right)^2}\\ \ln\gamma_B &= \frac{A_{BA}x_A^2}{\left(x_A + \dfrac{A_{BA}}{A_{AB}}x_B\right)^2}\end{aligned}} \tag{2.38}$$

This model is based on rather rough assumptions. Calculating parameters A_{AB} és

A_{BA} does not give good results; they must be fitted to measured data instead. In that case the equation serves as a good model in many cases and can be used for calculating phase equilibria.

2.7.3 Lattice (grid) models

Liquid structure is similar to crystal structure because of high liquid density. Partial ordering of the molecules in short distances can be observed. Liquids are similar to gases, too, in attenuation of spatial ordering in larger distance and in fast change of location of molecules, but similarity to crystals is stronger. Thus, liquids can be modelled as if they were solid. Location of molecules can be modelled as whole multiples of a given distance (grid spacing). As an approximation, only interaction between neighboring molecules is taken into account in lattice models. Number of neighbours is called ***co-ordination number***; this can be different for different species in the lattice, and ratio of molecules of different species can also be different. As an approximation, *internal state of the molecules is considered independent of the kinds of neighbours*, and thus change of free energy is a consequence of spatial rearrangement only, i.e. of changing the number of interacting molecule pairs of a given pair of species.

Mixture of equal size molecules

This is the simplest case. Co-ordination number is the same in the pure material and in the mixture. The spatial lattice is illustrated in **Fig. 2.5** as a square plain lattice with co-ordination number $z = 4$. (Diagonal neighbours are not counted.)

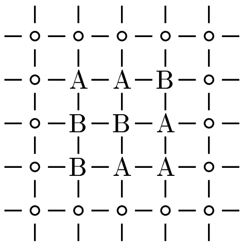

Figure 2.5: Square plain lattice with components A and B

Consider, for simplicity, binary liquid mixture consisting of N_A molecules of kind A, and N_B molecules of kind B. Then spatial arrangement of molecules in the lattice determines the pair numbers N_{AA}, N_{BB}, and N_{AB}; thus, the interaction energies u_{AA}, u_{BB}, and u_{AB} of pairs can be summed up to obtain lattice energy.

Random arrangement (zero-th approximation by *Guggenheim*) This approximation assumes that spatial arrangement is independent of pair energies (*Guggenheim, 1952*). Evidently this cannot be exactly true; and can be applicable only if the pair energies are rather similar (near ideal mixture). According to the model, probability of locating a molecule at a gird point is independent of the kind, but probalility of finding a kind A in a grid point is proportional to the number of molecules of kind A. The mixing excess is obtained as:

$$\Delta^E G \approx W x_A x_B \tag{2.39}$$

Hence the activity coefficient is:

$$\begin{aligned}\gamma_A =& \exp\left(\frac{W x_B^2}{RT}\right)\\ \gamma_B =& \exp\left(\frac{W x_A^2}{RT}\right)\end{aligned}$$

This is the simplest theoretical model based on combinatorics. Relation (2.39) is also considered as an interpretation of the first member of *Redlich-Kister* expansion.

Non-random arrangement, quasi-chemical model (first approximation by *Guggenheim*) Non-randomness is measured by a parameter κ:

$$\begin{aligned}N_{AB}^* =& \kappa\, z \frac{N_A N_B}{N_A + N_B}\\ \kappa \equiv& \frac{N_{AB}^*(N_A + N_B)}{z N_A N_B}\end{aligned}$$

According to the *quasi-chemical model* (*Guggenheim, 1952*), pairing of molecules is considered as a weak chemical reaction, and the low of mass action can be applied (with a quasi 'chemical' equilibrium ratio):

$$\frac{N_{AB}^2}{(zN_A - N_{AB})(zN_B - N_{AB})} = \exp\left(-\frac{2w}{zRT}\right)$$

In case of binary mixtures, the equation is quadratic in N_{AB} and can be solved analytically:

$$\frac{2}{\kappa} = 1 + \left(1 + 4x_A x_B\left[-1 + \exp\left(\frac{2w}{zRT}\right)\right]\right)^{1/2}$$

and hence the excess is

$$\Delta^E G \approx RT\frac{z}{2}\left(x_A \ln\frac{1-\kappa x_B}{x_A} + x_B \ln\frac{1-\kappa x_A}{x_B}\right)$$

$$\ln\gamma_A = \frac{z}{2}\ln\frac{1-\kappa x_B}{x_A}$$
$$\ln\gamma_B = \frac{z}{2}\ln\frac{1-\kappa x_A}{x_B}$$

In case of multicomponent mixtures the pair numbers cannot be analytically expressed; they must be determined numerically.

Mixture of different size molecules

The molecules are considered as consisting of several like or unlike but equal sized building segments. For example, n-hexane is built up of 6 almost equal segments if the difference between gropus CH_3 and CH_2 is neglected. A molecule of 2-propanone (dimethyl-ketone or acetone) is built up of three segments: two CH_3 groups and one CO group. These groups or segments (e.g. CH_2, CO, NH_2, COO) take place in the grid points. Flexible molecules can take up different shapes in the grid. For example, a molecule of three segments can take a straight or a broken shape.

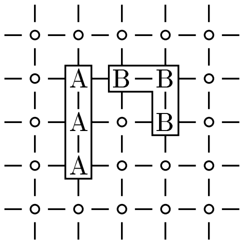

Figure 2.6: Three-segmented molecules AAA and BBB in the lattice

The number of possible neigbouring segments can change with the shape. For example, a straight three-segment molecule has 8 neighbours whereas there are only 7 for a broken one, see **Fig. 2.6**. This number depends on both size and shape of the molecule, and constitutes an additional property besides the grid co-ordination number z.

Spatial arrangement and randomness - non-randomness is more difficult to model for molecules of different size that those for equal size. To overcome the problem, the effects are decomposed to a part describing effect present in random arrangement and another part for those effects due to non-random arrangement. There is a ***combinatorial (or athermal or thermoneutral) part*** and a so-called ***residual part***:

$$\Delta^E G = \Delta^E_{comb} G + \Delta^E_{res} G \tag{2.40a}$$

$$\ln \gamma_i = \ln \gamma_i^{comb} + \ln \gamma_i^{res} \tag{2.40b}$$

In case of **molecules consisting of uniform segments (homogeneous molecules)**, the following form can be derived for the *athermal (combinatorial, $\kappa = 1$) part*:

$$\Delta_{comb}^E G = RT \left(\sum_i x_i \ln \frac{\Phi_i}{x_i} + \frac{z}{2} \sum_i q_i x_i \ln \frac{\Phi_i}{\theta_i} \right) \tag{2.41}$$

$$\ln \gamma_i^{comb} = 1 - \frac{\Phi_i}{x_i} + \ln \frac{\Phi_i}{x_i} - \frac{z}{2} q_i \left(1 - \frac{\Phi_i}{\theta_i} + \ln \frac{\Phi_i}{\theta_i} \right)$$

Here Φ_i is a so-called segment fraction, and θ_i is a so-called surface fraction, where r_i is the number of (homogeneous) segments in molecules of kind i:

$$\Phi_i \equiv \frac{x_i r_i}{\sum\limits_j x_j r_j} \tag{2.42}$$

$$\theta_i \equiv \frac{x_i q_i}{\sum\limits_j x_j q_j} \tag{2.43}$$

For estimating the residual part, the form of the quasi-chemical equilibrium is

$$\frac{N_{AB}^2}{(zN_A q_A - N_{AB})(zN_B q_B - N_{AB})} = \exp\left(-\frac{2w_{AB}}{zRT} \right) \tag{2.44}$$

Hence the residual part can be derived for a binary mixture:

$$\Delta_{res}^E G \approx RT \frac{z}{2} \left(x_A q_A \ln \frac{1 - \kappa\theta_B}{\theta_A} + x_B q_B \ln \frac{1 - \kappa\theta_A}{\theta_B} \right)$$

$$\ln \gamma_A^{res} = \frac{z}{2} q_A \ln \frac{1 - \kappa\theta_B}{\theta_A}$$

$$\ln \gamma_B^{res} = \frac{z}{2} q_B \ln \frac{1 - \kappa\theta_A}{\theta_B}$$

where κ is expressed from the quadratic equation (2.44).

Molecules consisting of uniform segments are usually fictious; they are mostly simplified, averaged, models of real, inhomogeneous, molecules.

In case of **molecules of different segments (inhomogeneous molecules)**, relative frequencies of occuring different segment pairs ought to be taken into account. That would be so difficult that *group contribution models* are used instead (see *section* 2.7.5).

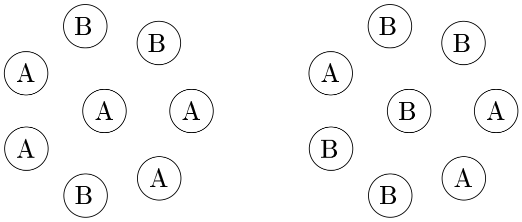

Figure 2.7: Cellular model: cells around A and B

2.7.4 Cellular (two-liquid) models

If distribution of species around A is independent of that of B then the distribution is uniform (random), otherwise there is some kind of spatial ordering. This approach can be maintained even if no lattice is applied. One can consider distribution of different species in a small neighbourhood of a randomly selected molecule of species A, and analyze it statistically. The same can be done around a molecule of species B. Thus, there are cells with center A and cells with center B **(figure 2.7)**. A cell with center A consists of a central molecule A and a shell containing both A and B molecules, and the central molecule A moves in the potential field of the shell. The cells overlap since any molecule A or B in the shell can be considered as a center of another cell. However, properties of the mixture is modelled as if it were be an ideal mixture of independent A and B cells; that is why it is also called *two-liquid model.*

Distributions of kinds around different centers can be different from each other and from the gross composition of the mixture. This is expressed with the term ***local composition***. *In case of perfect randomness the distribution of species around the centers is independent of the center; and thus the local mole fractions equal the gross mole fractions.* Particular models are invented according to the assumptions made on local compositions like local mole fractions or local volume fractions. If interaction potential of AA is lower than that of AB then one may expect larger fraction of molecules A around a center A than the gross (average) fraction, etc.

Mixture of equal size molecules

In case of perfect randomness the number of pairs N^*_{AB} equals the numbers N_{AB} and N_{BA} in the cells; thus

$$N_{AB} = N^*_{AB} = N_{BA} = z\frac{N_A N_B}{N_A + N_B}$$

After a tedious derivation one gets

$$\Delta^E G \approx \Delta^E A = RT\frac{z}{2}\left[x_A \ln\left(x_A + x_B A_{BA}\right) + x_B \ln\left(x_B + x_A A_{AB}\right)\right]$$

$$\ln\gamma_A = -\frac{z}{2}x_A \ln\left(x_A + x_B A_{BA}\right) + \frac{z}{2}x_B\left[\frac{A_{BA}}{x_A + x_B A_{BA}} - \frac{A_{AB}}{x_B + x_A A_{AB}}\right]$$

$$\ln\gamma_B = -\frac{z}{2}x_B \ln\left(x_B + x_A A_{AB}\right) + \frac{z}{2}x_A\left[\frac{A_{AB}}{x_B + x_A A_{AB}} - \frac{A_{BA}}{x_A + x_B A_{BA}}\right]$$

Mixture of different size molecules

(*Wilson, 1964*) suggested to calculate the ratio of volumes occupied by the molecules from ratio of *Boltzmann* factors. With additional assumtions he got:

$$\frac{\Delta G^E}{RT} = -x_A \ln\left(x_A + \Lambda_{AB}x_B\right) - x_B \ln\left(x_B + \Lambda_{BA}x_A\right)$$

Wilson, binary:

$$\boxed{\begin{aligned}\ln\gamma_A &= -\ln\left(x_A + \Lambda_{AB}x_B\right) + x_B\left(\frac{\Lambda_{AB}}{x_A + \Lambda_{AB}x_B} - \frac{\Lambda_{BA}}{x_B + \Lambda_{BA}x_A}\right)\\ \ln\gamma_B &= -\ln\left(x_B + \Lambda_{BA}x_A\right) + x_A\left(\frac{\Lambda_{BA}}{x_B + \Lambda_{BA}x_A} - \frac{\Lambda_{AB}}{x_A + \Lambda_{AB}x_B}\right)\end{aligned}}$$

Its multicomponent version is

Wilson:

$$\frac{\Delta G^E}{RT} = -\sum_i x_i \ln\sum_j \Lambda_{ij}x_j$$

Wilson:

$$\boxed{\ln\gamma_i = -\ln\left(\sum_j \Lambda_{ij}x_j\right) + 1 - \sum_k \frac{\Lambda_{ki}x_k}{\sum\limits_j \Lambda_{kj}x_j}}$$

where

$$\Lambda_{ij} \equiv \exp\left(-\frac{u_{ii} - u_{jj}}{RT}\right)$$

In practice, parameters $A_{ij} = u_{ij} - u_{ii}$ are fitted to measured data. There are versions of *Wilson* model, like that of three parameters, a version related to enthalpies etc., but the two-parameter version is the one widely known and applied.

Renon és Prausnitz (1968) applied enthalpy in *Boltzmann* factors, and an additional parameter α measuring non-randomness. They nicked it as *'Non-random two-liquid model', or NRTL*. Intruducing notations

$$\tau_{ij} \equiv \frac{g_{ij} - g_{jj}}{RT}$$

$$G_{ij} \equiv \exp\left(-\alpha_{ij}\tau_{ij}\right)$$

the equation looks as

NRTL:

$$\frac{\Delta G^E}{RT} = \sum_i x_i \frac{\sum\limits_j \tau_{ji} G_{ji} x_j}{\sum\limits_k G_{ki} x_k}$$

NRTL:

$$\boxed{\ln\gamma_i = \frac{\sum\limits_j \tau_{ji} G_{ji} x_j}{\sum\limits_k G_{ki} x_k} + \sum_j \frac{G_{ij} x_j}{\sum\limits_k G_{kj} x_k} \left[\tau_{ij} - \frac{\sum\limits_m \tau_{mj} G_{mj} x_m}{\sum\limits_k G_{kj} x_k}\right]}$$

Its parameters to fit are $A_{ij} = g_{ij} - g_{jj}$ and α, but α is usually fixed at about 0.2 - 0.3.

Prausnitz and co-workers (1975, 1978) applied decomposition to combinatorial and residual parts ((2.40a)–(2.40b)). They used equation (2.41) for the combinatorial part, and the quasi-chemical approximation of *Guggenheim* to the residual part. Their model is called UNIQUAC (universal quasi-chemical model) because it incorporates most of the known other models, i.e. they can be derived from this one with appropriate substitutions. Its form is:

UNIQUAC:

$$\Delta^E G = \Delta^E_{comb} G + \Delta^E_{res} G$$

$$\Delta^E_{comb} G = RT\left(\sum_i x_i \ln\frac{\Phi_i}{x_i} + \frac{z}{2}\sum_i q_i x_i \ln\frac{\Phi_i}{\theta_i}\right)$$

$$\frac{\Delta G^E_{res}}{RT} = -\sum_i\left[q_i x_i \ln\left(\sum_j \tau_{ji}\theta_j\right)\right]$$

$$\ln\gamma_i = \ln\gamma_i^{comb} + \ln\gamma_i^{res}$$

$$\ln\gamma_i^{comb} = 1 - \frac{\Phi_i}{x_i} + \ln\frac{\Phi_i}{x_i} - \frac{z}{2}q_i\left(1 - \frac{\Phi_i}{\theta_i} + \ln\frac{\Phi_i}{\theta_i}\right)$$

$$\ln\gamma_i^{res} = q_i\left[-\ln\left(\sum_j \tau_{ji}\theta_j\right) + 1 - \sum_k \frac{\tau_{ki}\theta_k}{\sum\limits_j \tau_{jk}\theta_j}\right]$$

where

$$\tau_{ji} \equiv \exp\left(-\frac{u_{ji} - u_{ii}}{RT}\right)$$

An alternative form of the combinatorial part is:

$$\ln \gamma_i^{comb} = \ln \frac{\Phi_i}{x_i} - \frac{z}{2} q_i \ln \frac{\Phi_i}{\theta_i} + l_i - \frac{\Phi_i}{x_i} \sum_j x_j l_j$$

where

$$l_i = \frac{z}{2}(r_i - q_i) - (r_i - 1)$$

(z is co-ordination number.)

Material parameters r és q of UNIQUAC belong to components; these parameters are fairly independent of what kind of mixture is the actual component in. Co-ordination number z is usually taken as 10, and parameters $A_{ji} = u_{ji} - u_{ii}$ are fitted to data.

Modeling mixtures of *heterogeneous molecules* is difficult in the same way as discussed at lattice models. Group contribution models are used instead.

2.7.5 Group contribution models

Different segments in a heterogeneous molecule are characterized by different volume (R) and surface (Q). Apply index k to distinguish between segment types in the molecule, and denote the number of segments of type k in molecule i with ν_{ki}; then volume and surface of molecule i can simply be defined by

$$r_i \equiv \sum_k \nu_{ki} R_k \tag{2.45a}$$

$$q_i \equiv \sum_k \nu_{ki} Q_k \tag{2.45b}$$

Total surface of segments k in molecule i is

$$Q_k^i \equiv \nu_{ki} Q_k$$

Group contribution models refer to atomic groups in the molecules that can be characterized with uniform, molecule-independent, R_k and Q_k values, and interactions between these groups also independent of the actual molecule they are found in. Such groups can be, for example, CH_2, NH_2 connected to an aliphatic chain, CO, OH, COO, aromatic ring, etc. Of course, behaviour of these groups depend strongly on the other groups they are in direct contact or simply near to them; thus, their independence is fictious. However, good approximation is obtained in many cases by applying universal group properties R and Q. Although these groups are environment-dependent, their application makes possible a more detailed characterization of the molecule than by applying models of homogeneous molecules. The uniform parameters R_k, Q_k, and A_{km} are fitted simultaneously to data of many equilibrium mesaurements made on several different mixtures.

Wilson és Deal (1962) defined the principles of the method of so-called *group solutions*. The system is considered as a mixture of functional groups. The principles are as follow.

1. Activity coefficient is expressed as product of combinatorial and residual parts:
$$\ln\gamma_i = \ln\gamma_i^{comb} + \ln\gamma_i^{res}$$
2. Functional group k increments γ_i of each component i by a term dependent on composition and temperature but independent of i. The increments Γ_k are taken as functions of temperature T and group fraction $\boldsymbol{w}$:
$$w_k \equiv \frac{\sum\limits_i \nu_{ki}x_i}{\sum\limits_i\sum\limits_m \nu_{mi}x_i}$$
$$\Gamma_k = f_k(\boldsymbol{w}, T)$$
3. For computing the resisular part, the residual effects in pure components are subtracted from the rsidual effects in the mixture
$$\ln\gamma_i^{rez} = \sum_k \nu_{ki}\left(\ln\Gamma_k - \ln\Gamma_k^{(i)}\right)$$
where superscript (i) refers to pure component i:
$$\Gamma_k^{(i)} = f_k\left(\boldsymbol{w}^{(i)}; T\right)$$
$$w_k^{(i)} \equiv \frac{\nu_{ki}}{\sum\limits_m \nu_{mi}}$$

A model named ASOG (analytical solution of groups) was developed on these principles but it, unfortunately, could not gain ground. However, *Frendenslund* and co-workers (in several articles from 1975) applied UNIQUAC in following the same principles to develop a new model they nicked to UNIFAC:
UNIFAC:

$$\ln\gamma_i^{comb} = \ln\frac{\Phi_i}{x_i} - \frac{z}{2}q_i\ln\frac{\Phi_i}{\theta_i} + l_i - \frac{\Phi_i}{x_i}\sum_j x_j l_j$$
$$l_i = \frac{z}{2}(r_i - q_i) - (r_i - 1)$$

where co-ordination number z is taken as 10, the volume and surface fractions Φ and θ are defined by (2.42) and (2.43), molecule properties r és q given by relations (2.45a) to (2.45b). The residual part is:

$$\ln\Gamma_k^{res} = Q_k\left[-\ln\left(\sum_m \Psi_{mk}\Theta_m\right) + 1 - \sum_m \frac{\Psi_{km}\Theta_m}{\sum\limits_n \Psi_{nk}\Theta_n}\right]$$

where indexes m és n run on groups, and the surface fraction is:

$$\Theta_m = \frac{w_m Q_m}{\sum\limits_n w_n Q_n}$$

and *Boltzmann* factor is computed as

$$\Psi_{nm} = \exp\left(-\frac{A_{nm}}{RT}\right)$$

UNIFAC is widely used in practice. Its parameters are group properties R_k and Q_k and group interaction parameters $A_{mn} \neq A_{nm}$. Several parameter sets are in use. A common property of them is that the first members of homologous series do not fit well to the measured data if the general groups are applied; therefore special groups has to be defined for describing these first members. For example, longer alcohols containing hydroxil groups connected to methyl and methylene groups are described well with these groups but methanol cannot be described this way; methanol has to be considered a separate group.

2.7.6 Models widely applied in practice, and their capability

The most applied and suggested models for which fitted parameters are available in the literature are *van Laar*, *Margules*, *Wilson*, *NRTL*, *UNIQUAC*, and *UNIFAC*.

Several similar equations exist but they are applied most. Common in them that parameters fitted to binary vapor-liquid equilibria can be applied to calculate multicomponent vapor-liquid equilibria as well. *NRTL* and *UNIQUAC* can be successfully applied to liquid-liquid equilibria, too, but frequently the parameters fitted to vapor-liquid equilibria cannot be used to liquid-liquid equilibria and vice versa; separate data set must be applied. The two-parameter *Wilson* equation cannot be applied to model liquid-liquid equilibria because its form excludes concave $\Delta^E G$ functions. For calculating ternary liquid-liquid equilibrium, one cannot use binary interaction parameter values because in many cases some species pairs do not have such data. For example, there is not liquid phase separation between water and ethanol and between ethanol and benzene, and thus ternary data are needed for fitting parameter to ternary mixture water - ethanol - benzene. Those parameters can then be used to calculate multicomponent mixtures. *UNIFAC* serves with good approximation in some cases, but rather rough estimations in other cases.

2.7.7 Electrolite solutions

All the models explained above are for non-electrolites. Modeling electrolite solutions is much more difficult. There are well behaving models but we do not treat them here.

2.8 Phase equilibrium calculations

2.8.1 Vapor liquid equilibria

Equilibria of mixtures is more coplex than that of pure materials because composition of the phases in equilibrium are different.

Phase relations of binary mixtures can be visualized in plain because the mole fraction of one component determines the mole fraction of the other one in the same phase. *Phase diagrams* (or *phase plots*) are drawn for constant pressure or constant temperature as are shown in **Fig.s 2.8 and 2.9**. In these figures mole fractions of the more volatile component (i.e. of the higher vapor pressure and lower boiling temperature) are measured on the horizontal axes. That is, the point of $x = 1$ and $y = 1$ is the place of pure more volatile (light) component, and the point of $x = 0$ and $y = 0$ is the place of pure less volatile (heavy) component.

Simple equilibrium calculations problems

We speak about ***simple equilibrium calculations*** if the composition of one of the phases ($\boldsymbol{x}$ or $\boldsymbol{y}$) as well as one of the intensive state variables (T is p) is given, and values of the other variables as assigned by the phase curve or phase surface (or hypersurface) are to be found.

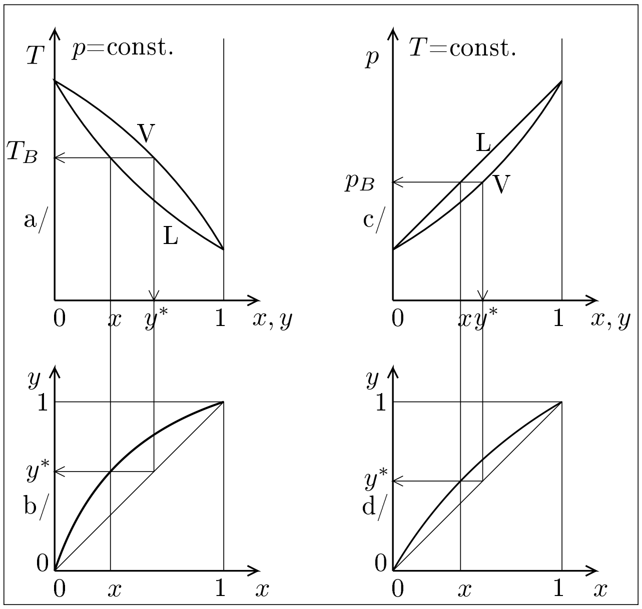

Figure 2.8: Bubble point in phase plots and equilibrium plots

Curves of equilibria at constant pressure p are shown in **Fig.s 2.8a-b**. Given a liquid composition x, the vertical projection line intersects the lower (liquid) phase curve at heigth of its buble point temperature T_B. The horizontal projection line from this intersection intersects the upper (vapor) phase curve over the vapor composition y^* that keeps equilibrium with x.

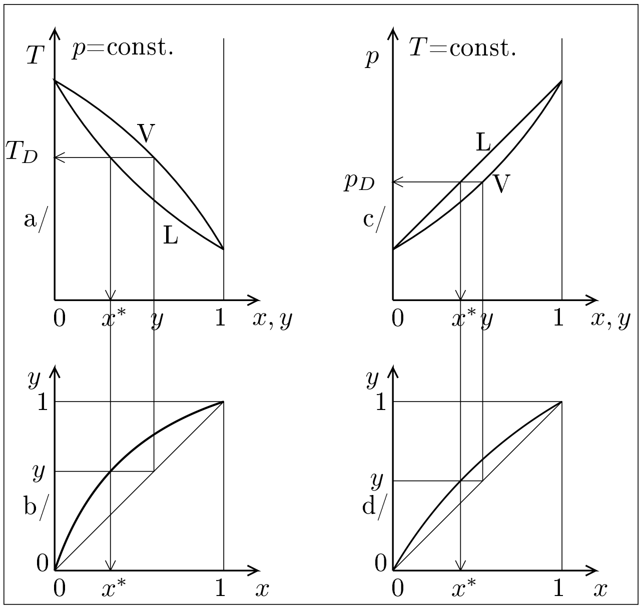

Figure 2.9: Dew point in phase plots and equilibrium plots

Curves of equilibria at constant temperature T are shown in **Fig.s 2.8c-d**. Given a liquid composition x, the vertical projection line intersects the upper (liquid) phase curve at heigth of its buble point pressure p_B. The horizontal projection line from this intersection intersects the lower (vapor) phase curve over the vapor composition y^* that keeps equilibrium with x.

Inversely:

Given a vapor composition y (**Fig.s 2.9a-b**) at constant pressure p, the vertical projection line intersects the upper (vapor) phase curve at heigth of its dew point temperature T_D (). The horizontal projection line from this intersection intersects the lower (liquid) phase curve over the liquid composition x^* that keeps equilibrium with y.

Given a vapor composition y (**Fig.s 2.9c-d**) at constant temperature T, the vertical projection line intersects the lower (vapor) phase curve at heigth of its dew point pressure p_D (). The horizontal projection line from this intersection intersects the upper (liquid) phase curve over the liquid composition x^* that keeps equilibrium with y.

In case of multicomponent mixtures the equilibrium points assign surfaces and hypersurfaces; the task of searching for the intersection points is more difficult and usually is performed with numerical iterative calculations, but there are just these four calculation tasks: determine bubble point temperature, dew point temperature, bubble point pressure, and dew point pressure, each with the equilibrium composition of the other phase.

Phase distribution (flash) calculation problems

We speak about **Phase distribution (or: 'flash') calculations** if gross composition $\boldsymbol{z}$ of the mixture and one or more intensive thermodynamic state variables are given and total state of the system, including compositions and ratio of the separating phases format at the given state variables are to be found. Combinations of compositions and temperature at given pressure are shown in **Fig. 2.10**. Location of point p-T-$\boldsymbol{z}$ in the phase diagram does not give us all possible knowledge on the state of the system. As changing T or p involves shifting point p-T-$\boldsymbol{z}$ in the diagram, so change the system's gross molar energy, enthalpy, entropy, free energy, and Gibbs energy as well. Conventionally q is used to denote the so-called *heat state* on condensation degree of the material at composition $\boldsymbol{z}$ and given intensive variables. This measures with how many times of the molar condensation (vaporization) heat does the molar enthalpy of the material deviates from its value at dew point. Thus, its value is $q = 0$ at dew point, and $q = 1$ at bubble point.

The calculation problems shown in **Fig. 2.10** can also be interpreted according to **Fig. 2.11**

Given the gross composition $\boldsymbol{z}$, pressure p and temperature T, we search for the equilibrium state of the system, i.e. whether it is liquid, gas, or forms two phases and, in the latter case, in what ratio, and what compositions are there in the phases, as well as what molar enthalpies (not shown in the figure) are there in the system and in the phases. This is, however, only one of the possible problem kinds. Instead of pressure or temperature, molar enthaly may be given, and the missing prssure or temperature might be a variable to determine.

Only a part of possible flash probles are nterpreted in **Fig. 2.11**; more general interpretation is shown in **Fig. 2.12**. Flash originally means the operation that the pressurized liquid is heated up and let through a throttling valve to suddenly vaporize.

Flash problems can be of various kinds. One can specify demanded phase ratio, needed mole fraction of a component in one of the phases, recovery ratio and so on. Therefore, all equilibrium calculations are, in wide sense, called flash calculations.

Typical flash probles are listed in **Table 2.1**. Here J is molar enthalpy of feed, h is liquid molar enthalpy, H is vapor molar enthalpy, f_i and v_i are component molar flow rates in feed and vapor, respectively.

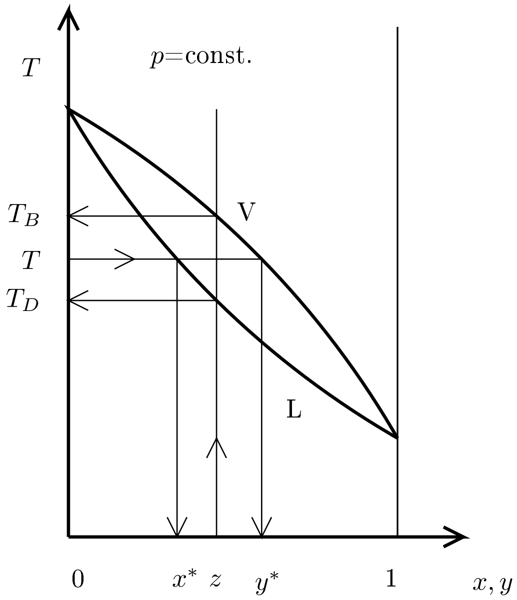

Figure 2.10: Flash problem in phase plot

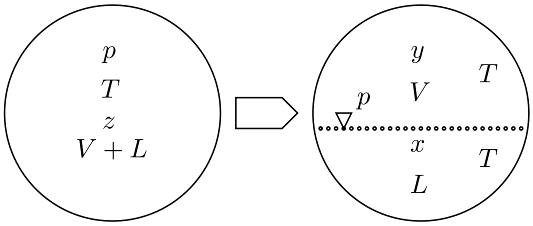

Figure 2.11: Static flash problem

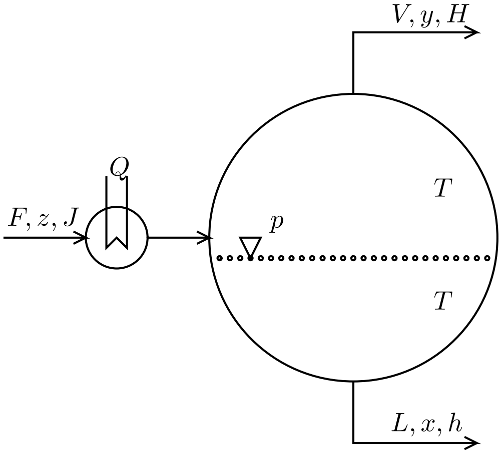

Figure 2.12: Steady state flash problem

Table 2.1: Typical flash problems

Given (specified)	Looked for	Notes
$\boldsymbol{z}$, p, T	J, V/L, $\boldsymbol{x}$, $\boldsymbol{y}$, h, H	isothermal-isobaric ('isothermal') flash
$\boldsymbol{z}$, p, J	T, V/L, $\boldsymbol{x}$, $\boldsymbol{y}$, h, H	isobaric-adiabatic ('adiabatic') flash
$\boldsymbol{z}$, J, T	p, V/L, $\boldsymbol{x}$, $\boldsymbol{y}$, h, H	isothermal-adiabatic flash
$\boldsymbol{z}$, p, V/L, J	T, $\boldsymbol{x}$, $\boldsymbol{y}$, h, H	temperature for phase ratio
$\boldsymbol{z}$, T, V/L, J	p, $\boldsymbol{x}$, $\boldsymbol{y}$, h, H	pressure for phase ratio
$\boldsymbol{z}$, p, x_i, J	T, V/L, $\boldsymbol{x}$, $\boldsymbol{y}$, h, H	for given purity
$\boldsymbol{z}$, T, x_i, J	p, V/L, $\boldsymbol{x}$, $\boldsymbol{y}$, h, H	for given purity
$\boldsymbol{z}$, p, y_i, J	T, V/L, $\boldsymbol{x}$, $\boldsymbol{y}$, h, H	for given purity
$\boldsymbol{z}$, T, y_i, J	p, V/L, $\boldsymbol{x}$, $\boldsymbol{y}$, h, H	for given purity
$\boldsymbol{z}$, p, v_i/f_i, J	T, V/L, $\boldsymbol{x}$, $\boldsymbol{y}$, h, H	for given recovery
$\boldsymbol{z}$, T, v_i/f_i, J	p, V/L, $\boldsymbol{x}$, $\boldsymbol{y}$, h, H	for given recovery
$\boldsymbol{z}$, y_i, v_i/f_i, J	p, T, V/L, $\boldsymbol{x}$, $\boldsymbol{y}$, h, H	for given purity and recovery

2.8.2 φ/φ, φ/γ and γ/γ methods

When comparing fugacities of two phases, they can be modelles in three possible combinations:

1. ***φ/φ method***: If both phases are modelled with equation of states then (the common) pressure is cancelled algebraically:

   $$y_i^I \varphi_i^I = y_i^{II} \varphi_i^{II} \tag{2.46a}$$

   Fugacity coefficient in each phase is calculated at the given (common) pressure and temperature and at the **composition of that phase**. Then the equilibrium ratio is:

   $$K_i^{II/I} \equiv \frac{y_i^{II}}{y_i^I} = \frac{\varphi_i^{II}}{\varphi_i^I} \equiv \exp\left(\ln \varphi_i^{II} - \ln \varphi_i^I\right) \tag{2.46b}$$

   In case of vapor-liquis equilibrium the equation's form is:

   $$y_i \varphi_i^V = x_i \varphi_i^L \tag{2.46c}$$

   where y_i mole fraction in vapor, x_i is mole fraction in liquid, and

   $$K_i \equiv K_i^{V/L} \equiv \frac{y_i}{x_i} = \frac{\varphi_i^L}{\varphi_i^V} \equiv \exp\left(\ln \varphi_i^L - \ln \varphi_i^V\right) \tag{2.46d}$$

2. ***φ/γ method***: If one of the phases is modelled with equation of states, and the other one with mixing excess model then the forms are different at the two sides:

   $$y_i \varphi_i^V p = \gamma_i x_i \varphi_i^\circ p_i^\circ \mathcal{P} \tag{2.47a}$$

   The right hand side is applied to a condesed phase, usualy liquid, otherwise vapor pressure p_i° has no meaning. The left hand side is applied to a vapor phase. Partial fugacity coefficient in the left hand side is taken at the system's pressure and temperature and at vapor composition $\boldsymbol{y}$. Activity coefficient in the right hand side is taken at the systems's temperature and at liquid composition $\boldsymbol{x}$; there is a vapor pressure of pure component i at systems's temperature, preceded by a fugacity coefficient of the pure component's vapor at the same vapor pressure and the system's temperature, and there is the *Poynting* correction. Equilibrium ratio is then:

   $$K_i \equiv \frac{y_i}{x_i} = \frac{\gamma_i \varphi_i^\circ p_i^\circ \mathcal{P}}{\varphi_i^V p} \tag{2.47b}$$

3. ***γ/γ method***: If both phases are modelled with mixing excess then vapor tensions and corrections are algebraically cancelled out:

   $$x_i^I \gamma_i^I = x_i^{II} \gamma_i^{II} \tag{2.48a}$$

Activity coefficient in each phase is taken at system's temperature and the phase's composition $\boldsymbol{x}$. Equilibrium ratio is then:

$$K_i^{II/I} \equiv \frac{x_i^{II}}{x_i^I} = \frac{\gamma_i^{II}}{\gamma_i^I} \equiv \exp\left(\ln \gamma_i^{II} - \ln \gamma_i^I\right) \tag{2.48b}$$

Methods φ/φ and φ/γ can be used for modelling vapor-liquid and liquid-liquid equilibria as well, but method γ/γ to liquid-liquid equilibria only. Application of methods φ/φ and γ/γ are clear and straightforward. In contrast, method φ/γ contains so many different variables and models that usually, if possible, its simplified versions are applied. At moderate pressure, and if no components associating in vapor phase are present, partial fugacity coefficient at the vapor side is omitted; thus the vapor is considered as ideal mixture of perfect gases. *Poynting* correction and fugacity coefficient of pure component vapor are usually omitted in the liquid side at moderate temperatures. Thus we obtain the so-called *modified Raoult-Dalton* equation:

$$\boxed{y_i p = \gamma_i x_i p_i^\circ} \tag{2.49a}$$

$$K_i \equiv \frac{y_i}{x_i} = \frac{\gamma_i p_i^\circ}{p} \tag{2.49b}$$

All equations (2.46a)-(2.49b) are expounded versions of the general necessary condition

$$f_i^I = f_i^{II} \quad (i = 1, 2, \dots, c)$$

where c is number of components.

Composition of one of the phases and one intensive state variable (T or p) is given in simple equilibrium calculation problems, and the other intensive state variable is looked for together with the equilibrium composition of the other phase. That is, the number of unknown variables is $c+1$. But the unknown mole fractions must add up to 1; thus the number of unknowns is decreased to c. In case of a 5-component mixture, for example, the problem is finding the zero(s) of a 5-variable system of equations. In case of phase distribution problems, the specified data are an intensive state variable or two of them besides the gross composition, and the variables looked for include mole fractons of both equilibrium phases; therefore the number of unknowns is even higher.

It does not follow, however, that we really have to solve a system of equations by numerically iterating in c or even more variables, because the equation system can be transformed to a function of much less variables, see below. The numerical search methods can be selected arbitrarily, but the procedure shown below are simple, clear, and easy to perform.

2.8.3 Vapor-liquid equilibria with φ/γ

Ideal mixture

In this case the equilibrium ratio does not depend on composition but temperature and pressure only:

$$K_i = K_i(T,p) = \frac{p_i^\circ}{p}$$

where vapor tension is function of temperature. The equilibrium calculations are very simple. The simplest one is finding bubble point pressure at given temterature:

Algorithm BP-id: ***Bubble point pressure***
Given: temperature T and liquid composition $\boldsymbol{x}$
Looked for: bubble point pressure p_B and equilibrium vapor composition $\boldsymbol{y}$
Guess: -

1. Compute vapor pressures p_i° using known vapor pressure curves.
2. Compute total pressure p_B:

$$p_B \Leftarrow \sum_i x_i p_i^\circ$$

3. Compute equilibrium vapor mole fractions:

$$y_i \Leftarrow \frac{x_i p_i^\circ}{p_B}$$

4. Stop.

This is simply a series of substitutions; no iteration is here. In each other problem some iteration is to be applied, but not here.

Algorithm DP-id-a: ***Dew point pressure***
Given: temperature T and vapor composition $\boldsymbol{y}$
Looked for: dew point pressure p_D and equilibrium liquid composition $\boldsymbol{x}$
Guess: p_D is estimated with an appropriate p.

1. Compute vapor pressures p_i° using known vapor pressure curves.
2. Compute liquid mole fractions:

$$x_i \Leftarrow \frac{y_i p}{p_i^\circ}$$

3. Compute sum of the obtained liquid mole fractions:

$$\sigma \Leftarrow \sum_i x_i$$

4. Normalize the mole fractions:

$$x_i \Leftarrow \frac{x_i}{\sigma}$$

5. Compute bubble point pressure to the calculated liquid composition:

$$p_B \Leftarrow \sum_i x_i p_i^\circ$$

6. If estimated p and calculated p_B are equal up to a predefined small error then the estimated pressure p is accepted as a dew point pressure: $p_D \Leftarrow p$, and stop.
7. Otherwise: Modify p, and return to step 2.

The above algorithm can be re-formulated as follows:

Algorithm DP-id-b: *Dew point pressure*
Given: temperature T and vapor composition $\boldsymbol{y}$
Looked for: dew point pressure p_D and equilibrium liquid composition $\boldsymbol{x}$
Guess: p_D is estimated with an appropriate p.

1. Compute vapor pressures p_i° using known vapor pressure curves.
2. Compute equilibrium ratios K_i:

$$K_i \Leftarrow \frac{p_i^\circ}{p}$$

3. Compute liquid mole fractions:

$$x_i \Leftarrow \frac{y_i}{K_i}$$

4. Compute sum of the obtained liquid mole fractions:

$$\sigma \Leftarrow \sum_i x_i$$

5. Normalize the mole fractions:

$$x_i \Leftarrow \frac{x_i}{\sigma}$$

6. If estimated σ equals 1 up to a predefined small error then the estimated pressure p is accepted as a dew point pressure: $p_D \Leftarrow p$, and stop.
7. Otherwise: Modify p, and return to step 2.

If the caclucated pressure is higher than the estimated one, or the calculated sum of mole fractions is higher than 1, then the estimated pressure must be decreased, otherwise increased. This is so because too high calculated pressure means too high mole fractions of the volatile components in the liquid.

Algorithm BT-id-a: ***Bubble point temperature***
Given: pressure p and liquid composition $\boldsymbol{x}$
Looked for: dew point temperatureT_B and equilibrium vapor composition $\boldsymbol{y}$
Guess: T_B is estimated with some appropriate T.

1. Compute tensions p_i° at temperature T.
2. Compute partial pressures:
$$p_i \Leftarrow x_i p_i^\circ$$
3. Sum up partial pressures:
$$\pi \Leftarrow \sum_i p_i$$
4. If π equals p up to a predefined small error then accept T as bubble point temperature: $T_B \Leftarrow T$, and go to step 6.
5. Otherwise: Modify T, and return to step 1.
6. Compute vapor mole fractions:
$$y_i \Leftarrow \frac{p_i}{\pi}$$
Stop.

If $\pi < p$ then T must be increased, otherwise decreased. The above algorithm can be re-formulated as follows:

Algorithm BT-id-b: ***Bubble point temperature***
Given: pressure p and liquid composition $\boldsymbol{x}$
Looked for: dew point temperatureT_B and equilibrium vapor composition $\boldsymbol{y}$
Guess: T_B is estimated with some appropriate T.

1. Compute tensions p_i° at temperature T.
2. Compute equilibrium ratios:
$$K_i \Leftarrow \frac{p_i^\circ}{p}$$
3. Compute vapor mole fractions:
$$y_i \Leftarrow K_i x_i$$
4. Sum up vapor mole fractions:
$$\sigma \Leftarrow \sum_i x_i$$
5. If σ equals 1 up to some predefined small error then accept T as bubble point tempereture: $T_B \Leftarrow T$, and go to step 6.
7. Otherwise: Modify T, and return to step 1.

6. Normalize vapor mole fractions:

$$x_i \Leftarrow \frac{x_i}{\sigma}$$

Stop.

If $\sigma < 1$ then T must be increased, otherwise decreased.

Calculating dew point temperature is done is a similar way.

Non-ideal liquid phase

The relations are supplemented with acticity coefficient. Since this depends on liquid composition and temperature, bubble point calculations are easier that dew point calculations. Bubble point iterations are simply supplemented with a step for calculating γ_i.

Algorithm BT-γ: *Bubble point temperature*
Given: pressure p and liquid composition $\boldsymbol{x}$
Looked for: dew point temperatureT_B and equilibrium vapor composition $\boldsymbol{y}$
Guess: T_B is estimated with some appropriate T.

1. Compute vapor pressures p_i° at T and activity coefficients γ_i at T and $\boldsymbol{x}$.
2. Compute equilibrium ratios:

$$\boldsymbol{K_i \Leftarrow \frac{p_i^\circ}{p}}$$

3. Compute vapor mole fractions:

$$\boldsymbol{y_i \Leftarrow K_i x_i}$$

4. Sum up vapor mole fractions:

$$\boldsymbol{\sigma \Leftarrow \sum_i x_i}$$

5. If $\boldsymbol{\sigma}$ equals 1 up to some predefined small error then accept $\boldsymbol{T}$ as bubble point tempereture: $\boldsymbol{T_B \Leftarrow T}$, and go to step 6.
7. Otherwise: Modify $\boldsymbol{T}$, and return to step 1.
6. Normalize vapor mole fractions:

$$\boldsymbol{x_i \Leftarrow \frac{x_i}{\sigma}}$$

Stop.

If $\sigma < 1$ then T must be increased, otherwise decreased.

Algorithm DT-γ: *Dew point temperature*
Given: pressure p and vapor composition $\boldsymbol{y}$
Looked for: dew point temperature T_D and equilibrium liquid composition $\boldsymbol{x}$
Guess: T_D is estimated with and appropriate T; γ_i is estimated with 1.

1. Compute tensions p_i° at temperature T.
2. Compute liquid mole fractions:
$$x_i \Leftarrow \frac{y_i p}{\gamma_i p_i^\circ}$$
3. Sum up liquid mole fractions:
$$\sigma \Leftarrow \sum_i x_i$$
4. Normalize liquid mole fractions:
$$x_i \Leftarrow \frac{x_i}{\sigma}$$
5. Compute γ_i at T and the normalized liquid mole fractions.
6. Compute total pressure over the calculated liquid composition:
$$\pi \Leftarrow \sum_i \gamma_i x_i p_i^\circ$$
7. If p equals π up to a predefined small error then accept T as dew point temperature: $T_H \Leftarrow T$, and stop.
8. Otherwise: Modify T, and return to step 1.

If $\pi < p$ then T must be increased, otherwise decreased.

General case φ/γ

If composition dependent property, like fugacity coefficient, turns up at vapor side, i.e. there is such a thing at both sides, then embedded iteration cycles are needed in the calculations, the same way as in the next subsection.

2.8.4 Vapor-liquid equlibria with φ/φ

Figacity coefficients must be computed at the system's pressure and temperature, at the composition of the actual phase, and with the properly selected molar volume root. If bubble point is calculated then the smallest one must be selected from the (at most) three roots, and the highest root cannot be used for computing fugacity coefficient of the vapor phase because that root must be calculated at vapor composition, not liquid composition. For computing vapor phase properties, first you have to estimate vapor composition $\boldsymbol{y}$ and use this data to find the roots and then calculate vapor fugacity coefficient. The situation is just opposite at dew point calculations.

Since the models for liquid and vapor are equal, the four problems' calculation procedures can be lumped.

Algorithm BDTP-φ: *Equilibrium calculation with* φ/φ
Given: (T or p) and (liquid or vapor) composition $\boldsymbol{y^I}$ (folyadék vagy pára)
Looked for: (p^* or T^*) and (vapor or liquid) composition $\boldsymbol{y^{II}}$
Guess: (p^* or T^*) is estimated with some appropriate (p or T); φ^I and φ^{II} are estimeted with 1 or some more precise method.

1. Compute fugacity coefficients in the given phase φ_i^I.
2. Set counter n as $n \Leftarrow 0$.
3. Compute equilibrium ratios:
$$K_i \Leftarrow \frac{\varphi_i^I}{\varphi_i^{II}}$$
4. Compute equilibrium mole fractions in the other phase:
$$y_i^{II} \Leftarrow K_i y_i^I$$
5. Sum up equilibrium mole fractions in the other phase:
$$\sigma \Leftarrow \sum_i y_i^{II}$$
6. Normalize equilibrium mole fractions in the other phase:
$$y_i^{II} \Leftarrow \frac{y_i^{II}}{\sigma}$$
7. Compute fugacity coefficients in the other phase φ_i^{II}.
8. Compute equilibrium ratios:
$$K_i \Leftarrow \frac{\varphi_i^I}{\varphi_i^{II}}$$
9. If $n = 0$ then go to step 12.
10. Otherwise: Compute a norm of deviation, between two consecutive iterations, of other phase calculated mole fractions:
$$\Delta \Leftarrow \sum_i \left| y_i^{II} - ye_i^{II} \right|$$
11. If Δ is smaller that a predefined small positive error δ then go to step 14 (leave the inner cycle).

12. Otherwise: Store the calculated mole fractions:
$$ye_i^{II} \Leftarrow y_i^{II}$$
13. Increment n:
$$n \Leftarrow n+1$$
and return to step 3 (inner cycle).
14. If σ equals 1 up to a predefined small positive error ε then accept the actual estimation of (p or T): $p^* \Leftarrow p$ or $T^* \Leftarrow T$, and stop.
15. Otherwise: Modify (p or T), and return to step 1 (outer cycle).

2.8.5 Vapor-liquid distribution calculations

Composition of none of the phases is known, thus estimation of equilibrium ratios is not enough to start the calculations. But if phase ratio is also known then the phases can be calculated.

Let F, V, and L denote feed, vapor, and liquid amount (or flow rate), respectively, and let λ denote liquid ratio $\lambda = L/F$; then the following material balances

$$Fz_i = Vy_i + Lx_i \qquad (i = 1, 2, \ldots, c)$$

and

$$F = V + L$$

can be used to construct a material balance expressed with λ:

$$z_i = (1-\lambda)y_i + \lambda x_i \qquad (i = 1, 2, \ldots, c)$$

that is

$$z_i = (1-\lambda)K_i x_i + \lambda x_i \qquad (i = 1, 2, \ldots, c) \tag{2.50}$$

Both phase mole fractions can be expressed from here, using estimated equilibrium ratios K_i:

$$x_i = \frac{z_i}{\lambda + (1-\lambda)K_i} \qquad (i = 1, 2, \ldots, c) \tag{2.51}$$

$$y_i = \frac{K_i z_i}{\lambda + (1-\lambda)K_i} \qquad (i = 1, 2, \ldots, c) \tag{2.52}$$

Mole fractions should sum up to 1:

$$\sum_i \frac{z_i}{\lambda + (1-\lambda)K_i} = 1 \tag{2.53}$$

$$\sum_i \frac{K_i z_i}{\lambda + (1-\lambda)K_i} = 1 \tag{2.54}$$

If K_i are given or estimated then a liquid ratio λ satisfying one of the mole fraction sum conditions can be numerically determined. One can find the root of

Table 2.2: Interpretation of $\psi(\lambda)$ defined in (2.55)

$\psi(0)$	$\psi(1)$	fázisállapot	$\psi\psi \equiv \psi(0)\psi(1)$
< 0	< 0	liquid below boiling point	> 0
< 0	$= 0$	boiling point liquid	$= 0$
< 0	> 0	two phases	< 0
$= 0$	> 0	dew point vapor	$= 0$
> 0	> 0	superheated vapor	> 0

either (2.53) or (2.54), and this can be substituted back to (2.51) and (2.52) to get the phase mole fractions. One of them (which was used to determine λ) sums up to 1; the other one is to be normalized. Functions (2.53) or (2.54) are not monotonic in $0 \le \lambda \le 1$; that may cause numerical difficulty. That is why both are used to determine λ:

$$\psi(\lambda) \equiv \sum_i y_i - \sum_i x_i = \sum_i \frac{(K_i - 1)z_i}{\lambda + (1-\lambda)K_i} = 0 \tag{2.55}$$

None of the mole fractions obtained this way sum up to unity; both has to be normalized; but function $\psi(\lambda)$ defined in (2.55) is strictly monotonic in $0 \le \lambda \le 1$ because

$$\frac{\mathrm{d}\psi(\lambda)}{\mathrm{d}\lambda} = \sum_i \frac{(K_i - 1)^2 z_i}{[\lambda + (1-\lambda)K_i]^2} > 0 \tag{2.56}$$

always positive. This makes root finding much easier.

A formal condition for having two phases, i.e. root of (2.55) to fall between 0 and 1 is that function $\psi(\lambda)$ has negative value at $\lambda = 0$ and positive value at $\lambda = 1$, because the function is monotonic increasing. The possible cases are collected in **Table 2.2**.

One cannot know if the system in equilibrium has more than one phase; and if not then if its state is liquid or vapor. In any case, only estimated equilibrium ratios K_i are known before convergence, and that might tell us a state different from the real one.

Calculation of isothermal-isobaric flash

In this case the distribution is determined by temperature and gross composition together; enthalpies can be calculated afterwards. (If heat state of the feed is known than a necessary heating or cooling power needed to reach the specified temperature and pressure can be calculated.)

Algorithm VL-TP: ***Isothermal-isobaric flash with*** φ/φ ***or*** φ/γ
Given: T, p, and $\boldsymbol{z}$
Looked for: λ, $\boldsymbol{x}$ and/or $\boldsymbol{y}$, and $\boldsymbol{K}$ in case of distribution
Guess: $\boldsymbol{K}$

1. Compute composition independent data (interaction parameters if they are temperature dependent, and vapor pressures).
2. Compute $\psi(0)$ and $\psi(1)$, as well as $\psi\psi \Leftarrow \psi(0)\psi(1)$.
3. If $\psi\psi \geq 0$ then there is only a single phase, stop.

   Otherwise: There is a root of (2.55) in the needed domain (there may be two phases).
4. Save the earlier (last) estimation of $\boldsymbol{K}$:

$$Ke_i \Leftarrow K_i$$

5. Find the root λ of (2.55) between 0 and 1.
6. Compute mole ractions in both phases with (2.53) and (2.54).
7. Normalize the mole fractions in both phases.
8. Compute phase properties with the normalized mole fractions. (In case of equation of state: mixed parameters, volume roots, root selection, and finally partial fugacity coefficients. In case of excess model: activity coefficients, and corrections if used.)
9. From data obtained in step 8, and depending on whether φ/φ or φ/γ is applied, compute the difference:

$$\sigma \Leftarrow \left| \frac{y_i^I \varphi_i^I - y_i^{II} \varphi_i^{II}}{y_i^{II} \varphi_i^{II}} \right|$$

   or

$$\sigma \Leftarrow \left| -p + \sum_i \frac{\gamma_i x_i \varphi_i^{\circ,L} p_i^\circ \mathcal{P}}{y_i \varphi_i^V} \right|$$

   or a simplified epression.
10. If σ is smaller than a predefined small error ε then there are two phases with computed data $\boldsymbol{x}$, $\boldsymbol{y}$, and λ, stop.
11. Otherwise: From data obtained in step 8, and depending on whether φ/φ or φ/γ is applied, re-compute equilibrium ratios K_i.
12. Compute $\psi(0)$ and $\psi(1)$, as well as $\psi\psi \Leftarrow \psi(0)\psi(1)$.
13. If $\psi\psi \geq 0$ then there is only a single phase, stop.
14. Otherwise: There is a root of (2.55) in the needed domain (there may be two phases). Compute the norm of difference between the new and the earlier equilibrium ratio estimates:

$$\Delta \Leftarrow \sum_i |Ke_i - K_i|$$

15. If Δ is smaller than a predefined small error δ then there are two phases with computed data $\boldsymbol{x}$, $\boldsymbol{y}$, and λ although precision specified by ε is not reached, stop.
16. Otherwise: Return to step 4.

Calculation of isobaric-adiabatic flash

Inthis case the equilibrium state is determined by the enthalpy balance.

The simplest method is embedding the isothermal-isobaric flash calculation in a cycle with unknown variable T:

Algorithm VL-PQ-embd: ***Isobaric-adiabatic flash with embedded cycle***
Given: p, J, and $\boldsymbol{z}$
Looked for: λ, T, $\boldsymbol{x}$ and/or $\boldsymbol{y}$, and $\boldsymbol{K}$ in case of distribution
Guess: $\boldsymbol{K}$ és T

1. Execute **Algorithm VL-TP**.
2. Compute H and/or h at the calculated compositions.
3. Compute the error of the enthalpy balance:

$$\sigma \Leftarrow |\lambda h + (1-\lambda)H - J|$$

4. If error σ is smaller than a predefined limit ε then stop.
5. Otherwise: Modify T, and return to step 1.

.

Faster computation can be achieved with *Newton*-iteration in two simultaneous variables: λ and T. The two equations are

$$\psi(\lambda, T) \equiv \sum_i \frac{(K_i - 1)z_i}{\lambda + (1-\lambda)K_i} = 0 \tag{2.57a}$$

$$\zeta(\lambda, T) \equiv \lambda h + (1-\lambda)H - J = 0 \tag{2.57b}$$

The first equation is a variant of (2.55), the second one is a heat balance. Four partial derivatives (entries of the *Jacobian* matrix) are needed for performing *Newton*-iteration. Partial derivative of (2.57a) by λ is given by (2.56). Partial derivative of (2.57b) by λ can be approximated by neglecting λ-dependence of H and h:

$$\left(\frac{\partial \zeta(\lambda, T)}{\partial \lambda}\right)_T \approx h - H$$

Partial derivative of (2.57b) by T can be obtained from derivatives of the enthalpy functions:

$$\left(\frac{\partial \zeta(\lambda, T)}{\partial T}\right)_\lambda = \lambda \frac{\partial h}{\partial T} - (1-\lambda)\frac{\partial H}{\partial T}$$

Partial derivative of (2.57a) by T can either be neglected or obtained with numerical differentiation.

Algorithm VL-PQ-Nwtn: ***Isobaric-adiabatic flash with Newton-iteration***
Given: p, J, and $\boldsymbol{z}$
Looked for: λ, T, $\boldsymbol{x}$ and/or $\boldsymbol{y}$, and $\boldsymbol{K}$ in case of distribution
Guess: $\boldsymbol{K}$, T, and λ

1. Save the earlier (last) estimations:

$$Ke_i \Leftarrow K_i$$
$$\lambda e \Leftarrow \lambda$$

2. Compute a $\psi(\lambda, T)$, $\zeta(\lambda, T)$, and their partial derivatives.
3. Update λ and T according to *Newton*'s iteration method.
4. Compute both compositions with (2.53) and (2.54).
5. Normalize the compositions.
6. Compute phase properties with the normalized mole fractions. (In case of equation of state: mixed parameters, volume roots, root selection, and finally partial fugacity coefficients. In case of excess model: activity coefficients, and corrections if used.)
7. From data obtained in step 8, and depending on whether φ/φ or φ/γ is applied, compute the difference:

$$\sigma \Leftarrow \left| \frac{y_i^I \varphi_i^I - y_i^{II} \varphi_i^{II}}{y_i^{II} \varphi_i^{II}} \right|$$

or

$$\sigma \Leftarrow \left| -p + \sum_i \frac{\gamma_i x_i \varphi_i^{\circ,L} p_i^\circ \mathcal{P}}{y_i \varphi_i^V} \right|$$

or a simplified epression.

8. If σ is smaller than a predefined small error ε then there are two phases with computed data $\boldsymbol{x}$, $\boldsymbol{y}$, and λ, stop.
9. Otherwise: From data obtained in step 6, and depending on whether φ/φ or φ/γ is applied, re-compute equilibrium ratios K_i.
10. Compute $\psi(0)$ and $\psi(1)$, as well as $\psi\psi \Leftarrow \psi(0)\psi(1)$.
11. If $\psi\psi \geq 0$ then there is only a single phase, stop.
12. Otherwise: There is a root of (2.55) in the needed domain (there may be two phases). Compute the norm of difference between the new and the earlier equilibrium ratio estimates:

$$\Delta \Leftarrow \sum_i |Ke_i - K_i|$$

13. If Δ is smaller than a predefined small error δ then there are two phases with computed data $\boldsymbol{x}$, $\boldsymbol{y}$, and λ although precision specified by ε is not reached, stop.
14. Otherwise: Return to step 1.

2.8.6 Liquid-liquid(-vapor) distribution calculations

Problems analogous to bubble point and dew point calculations do not typically occur in liquid-liquid equilibria. The question is usually whether ther is distribution and if yes then how the liquid is distributed. Therefore the computations are similar to the isothermal-isobaric flash calculations. Method γ/γ can also be used, and this is the typical case. The two liquid phases are denoted by superscripts ^{I} and ^{II} in the following algorithm.

Algorithm LL: *Isothermal-isobaric liquid distribution*
Given: T, p, and $\boldsymbol{z}$
Looked for: (phase^{I} to total) ratio λ, $\boldsymbol{x^I}$ and/or $\boldsymbol{x^{II}}$, and liquid/liquid equilibrium ratios $\boldsymbol{K}$ in case of distribution
Guess: $\boldsymbol{K}$

1. Compute composition independent data (interaction parameters if they are temperature dependent, and vapor pressures).
2. Compute $\psi(0)$ and $\psi(1)$, as well as $\psi\psi \Leftarrow \psi(0)\psi(1)$.
3. If $\psi\psi \geq 0$ then there is only a single phase, stop.

   Otherwise: There is a root of (2.55) in the needed domain (there may be two phases).
4. Save the earlier (last) estimation of $\boldsymbol{K}$:

   $$Ke_i \Leftarrow K_i$$

5. Find the root λ of (2.55) between 0 and 1. (Apply $\boldsymbol{x^I}$ instead of $\boldsymbol{x}$.)
6. Compute mole ractions in both phases with (2.53) and (2.54). (Apply $\boldsymbol{x^I}$ and $\boldsymbol{x^{II}}$ instead of $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively.)
7. Normalize the mole fractions in both phases.
8. Compute phase properties with the normalized mole fractions. (In case of equation of state: mixed parameters, volume roots, root selection, and finally partial fugacity coefficients. In case of excess model: activity coefficients; no corrections are needed.)
9. From data obtained in step 8, and depending on whether φ/φ or γ/γ is applied, re-compute equilibrium ratios K_i.
11. Compute $\psi(0)$ and $\psi(1)$ értékeket, as well as $\psi\psi \Leftarrow \psi(0)\psi(1)$.
12. If $\psi\psi \geq 0$ then there is a single liquid phase, stop.
13. Otherwise: (2.55) has a root in the domain; there may be two phases. Compute a norm of difference between the new and the last estimations:

    $$\Delta \Leftarrow \sum_i |Ke_i - K_i|$$

14. If Δ is smaller than a predefined error limit δ then there are two phases with calculated data $\boldsymbol{x^I}$, $\boldsymbol{x^{II}}$ and λ althouh precision limit ε could not be achieved, stop.
15. Otherwise: Return to step 4.

A week point of this method is that is may end up with a result of a single phase (no distribution) even if there *is* distribution according to the model. This is a **good** solution, see *Section 2.10.* There are multiple solutions of the equation system we solve here, and the one phase is the so-called ***'trivial solution'***. This **is** the right solution if really there is no phase distribution. If there **is** phase distribution then trivial solution can be avoided with a good chance by overestimating the equilibrium ratios $\boldsymbol{K}$: If $K_i > 1$ is expected then let it be very high, e.g. 10^4; if $K_i < 1$ is expected then let it be very low, e.g. 10^{-4}.

A frequent **vapor-liquid-liquid** distribution problem is finding bubble point of some liquid mixture at constant pressure in the case when the components only partially solvable in each other in liquid phase. If really there is liquid phase distribution then the vapor is not in equilibrium with the liquid phase in gross liquid composition but with each separate liquid phases in its particular composition. Therefore, liquid-liquid distribution should be taken into account at bubble point calculations.

Consequences of neglecting this note is demonstrated on the example of atmospheric isobutanol - water system modelled with modified *Raoult-Dalton* equation, UNIQUAC model with inetractio parameters 493.100 and 87.888 cal/(mol K).

If only gross liquid composition is considered then the results are shown in **Fig. 2.13**. There is some overlap of the phase curves in a short interval, and there is a nonsense equilibrium slope in a long interval. If, on the other hand, liquid distribution is taken into account then the results are shown in **Fig. 2.14**, revealing the real structure of the equilibrium system: heterogeneous azeotrope.

Since there are three mutual equilibria here, checking equilibria of vapor with only one of the liquid phases is enough. A possible algorithm is what follows:

Algorithm VLL: *Bubble point temperature with possible liquid distribution*
Given: pressure p and gross liquid composition $\boldsymbol{z}$
Looked for: bubble point temperature T_B, equilibrium liquid phase compositions $\boldsymbol{x^I}$ and $\boldsymbol{x^{II}}$ in case of liquid distribution, and equilibrium vapor composition $\boldsymbol{y}$
Guess: T_B is estimated with an appropriate T; liquid-liquid equilibrium ratios are estimated with some appropriate $\boldsymbol{K}$ values.

1. Compute temperature dependent data at T.
2. Execute **Algorithm LL**.
3. Initialize counter n to 0.
4. Compute equilibrium ratios $\boldsymbol{K}^*$ between one of the liquid phases and the vapor phase.
5. Compute vapor phase mole fractions:
$$y_i \Leftarrow K_i^* x_i^I$$
6. If $n = 0$, then go to step 10. Otherwise:
7. Sum up vapor mole fraction:
$$\sigma \Leftarrow \sum_i y_i$$

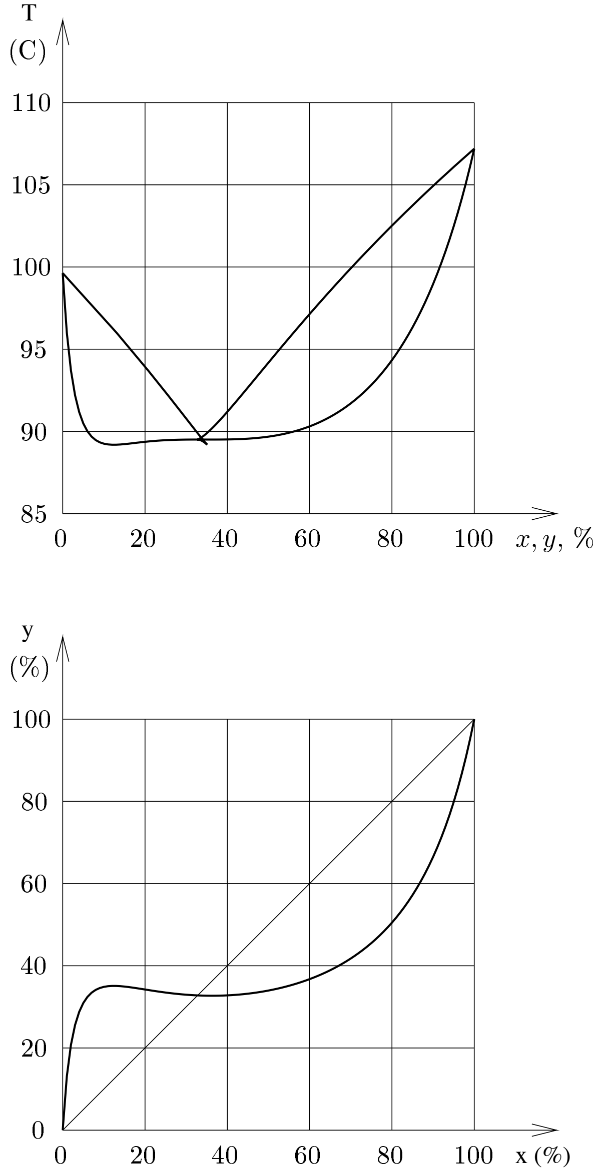

Figure 2.13: Isobutanol - water, 1 bar, with gross liquid composition

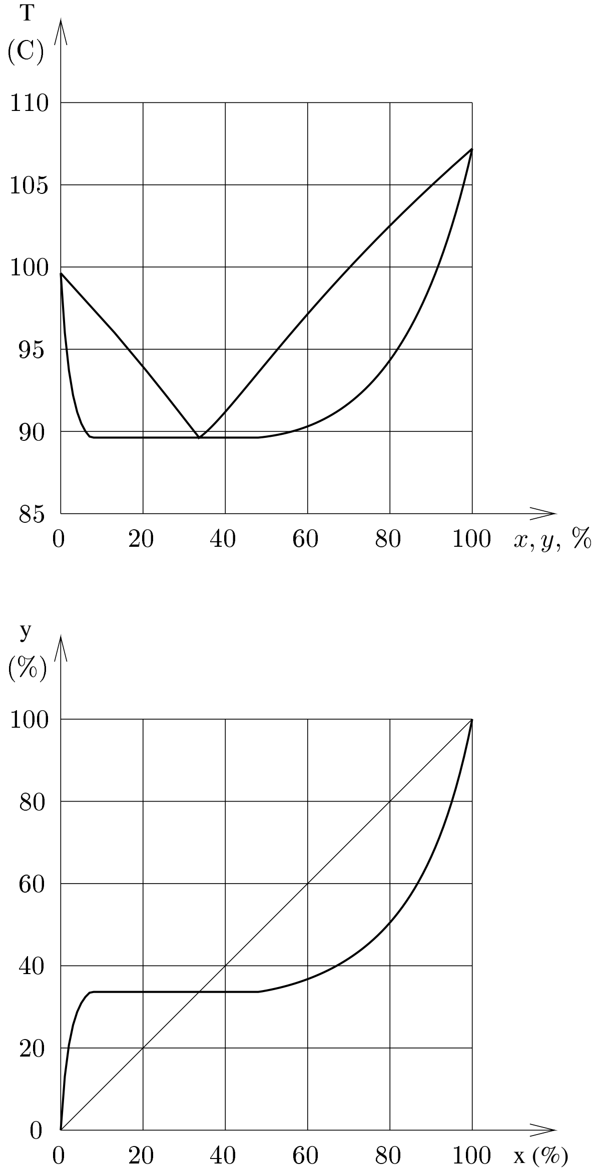

Figure 2.14: Isobutanol - water, 1 bar, with liquid distribution

8. Count norm of differences between consecutive vapor phase estimations:

$$\Delta \Leftarrow \sum_i |y_i - ye_i|$$

9. If Δ is small enough then go to step 15. Otherwise:
10. Store the calculated (not yet normalized) vapor mole fractions:

$$ye_i \Leftarrow y_i$$

11. Normalize vapor mole fractions:

$$y_i \Leftarrow \frac{y_i}{\sigma}$$

12. Increment counter:

$$n \Leftarrow n + 1$$

13. Re-compute composition dependent vapor phase properties if there are such variables in the model.
14. Return to step 4. (Inner cycle)
15. If $\sigma \approx 1$ then accept T a bubble point temperature: $T_B \Leftarrow T$, and go to step 17.
16. Otherwise: Modify T, and return to step 1. (Outer cycle)
17. Normalize vapor mole fractions:

$$y_i \Leftarrow \frac{y_i}{\sigma}$$

and stop.

2.8.7 Thermodynamic consistency of measured data

Before fitting parameters to data, it is highly suggested to check thermodynamic consistency of the data. That is, check if the data system satifies general thermodynamic relations independent of the applied particular model. Such relations are, for example:

$$RT \ln \gamma_i \equiv \left(\frac{\partial n \Delta^E G}{\partial n_i} \right)_{T,\, p,\, \hat{\boldsymbol{n}}_{\boldsymbol{i}}} \tag{2.58}$$

$$\sum_i x_i \mathrm{d} \ln \gamma_i = \frac{\Delta^E V}{RT} \mathrm{d}p - \frac{\Delta^E H}{RT^2} \mathrm{d}T \tag{2.59}$$

Of course, these relations never fit perfectly but only up to uncertainty in measurement data. Consistency check methods mainly look for systematic errors, but are sometimes able to point out large random errors as well.

Consistency check of only binary vapor-liquid equilibria are treated in this text. The most applied methods are the following two.

Area integral consistency check

Relations $x_2 = 1 - x_1 \equiv 1 - x$ are valid in binary mixtures. Applying (2.59), one can write

$$\mathrm{d}\ln\frac{\gamma_1}{\gamma_2} = \frac{\Delta^E V}{RT}\mathrm{d}p - \frac{\Delta^E H}{RT^2}\mathrm{d}T \tag{2.60}$$

Mixing volume effect are usually small in common liquids and can be neglected.

Isothermal case (*Hála et al, 1967*). In this case the second member of (2.60) is zero; the first member is rather small; therefore integration according to mole fraction should give approximately zero:

$$\int_0^1 \ln\frac{\gamma_1}{\gamma_2}\mathrm{d}x = \int_{p(x=0)}^{p(x=1)} \frac{\Delta^E V}{RT}\mathrm{d}p \approx 0$$

Activity coefficients can be well approximated with ideal gas at low pressure:

$$\gamma_i = \frac{y_i p}{x_i p_i^\circ}$$

After plotting logarithm of ratio of activity coefficients against x, numerical or graphical integration can be performed. Normally, the plot is positive at one section inside interval [0,1], and negative in the complementary section. The integral is zero if the areas under the plot of the two sections are equal. That is why this method is called ***area integral***. *King (1969)* analyzed two measurement data sets of system benzene - n-heptane with this method, and obtained results shown in **Fig. 2.15**. The data shown with empty circles are wrong; the other one is acceptable.

Izobaric case (*Herington, 1951*). In this case mixing heat is also to be taken into account but usually no measured excess heat data are available. However, molar mixing heat data of organic materials are rather similar; as well as water - organic data, and rough estimation of the expected range of the integral value can be given.

Residues consistency check

Applying a general thermodynamic relation, one of the interdependent variables can be expressed, and compared to its measured value. The deviation is called residue. These resudues ought to be zero; their distribution can be analyzed. An example is shown below.

Activity coefficients of binary mixture can be expressed in isothermal case as

$$\ln\gamma_1 = \frac{\Delta^E G}{RT} + (1-x)\frac{\mathrm{d}\left(\dfrac{\Delta^E G}{RT}\right)}{\mathrm{d}x} \tag{2.61a}$$

$$\ln\gamma_2 = \frac{\Delta^E G}{RT} \quad - \quad x\ \frac{\mathrm{d}\left(\dfrac{\Delta^E G}{RT}\right)}{\mathrm{d}x} \tag{2.61b}$$

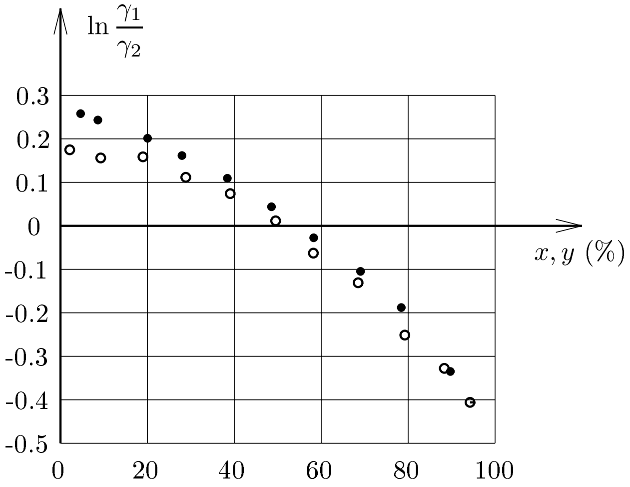

Figure 2.15: benzene – n-heptane, area integral data consistency check, *King, 1969*

Then total pressure can be calculated with modified *Raoult-Dalton* equation:

$$p(x) = x p_1^\circ \exp\left[\frac{\Delta^E G}{RT} + (1-x)\frac{\mathrm{d}\left(\frac{\Delta^E G}{RT}\right)}{\mathrm{d}x}\right] + (1-x) p_2^\circ \exp\left[\frac{\Delta^E G}{RT} - x\frac{\mathrm{d}\left(\frac{\Delta^E G}{RT}\right)}{\mathrm{d}x}\right] \tag{2.62}$$

Members in square bracket are for activity coefficients according to (2.61a) and (2.61b). Equation (2.62) is an ordinary differential equation whose solution is the function $\frac{\Delta^E G}{RT} \equiv \mathcal{G}(x)$. Total pressure in function of x, $p(x)$, is measured; at the endpoints of the interval it equals pure component's vapor pressure; vapor pressures are known; but function $\frac{\Delta^E G}{RT} \equiv \mathcal{G}(x)$ is unknown. This unknown function is obtained by some numerical solution of the differential equation.

Once $\frac{\Delta^E G}{RT} \equiv \mathcal{G}(x)$ is obtained, activity coefficients γ_1 and γ_2 can be computed with (2.61a) and (2.61b), then the vapor mole fractions can also be computed:

$$y_i^* = \frac{\gamma_i x_i p_i^\circ}{p}$$

These y_i^* vapor mole fractions are computed merely from the measured x, p, and T data. (T is used in calculating vapor tensions.) In principle, these y_i^* data ought to equal the measured y_i data:

$$\Delta y \equiv y_i^* - y_i \approx 0$$

After plotting residues Δy_i against x, its distribution can be analyzed and evaluated. Uniform distribution refers to random errors; any shift in expectation along x refers to systematic errors.

2.9 Parameter fitting and extrapolation

Only fitting of binary intaraction parameters are discussed in this text.

Before fitting, thermodynamic consistency should be checked, see *section 2.8.7*.

The form of the model expressed to zero is usually something like this:

$$f(T, p, \boldsymbol{x}, \boldsymbol{y};\, a, b, \dots) = 0$$

Here a, b, etc. are the parameters to be fitted to data. Fitting is done numerically; this is a special kind of optimization problem. Once the optimal parameters are found, their variances around expectations are worth to analyze.

Binary interaction parameters are strongly correlated; the pair of data (e.g. $A_{i,j}$ and $A_{j,i}$) can be shifted along the longer axis of an elongated ellipse (of the level lines of error function to be minimized at fitting). Apparently essentially different parameter pairs can be obtained with different fitting methods, and they may work similarly well or badly when substituted into the thermodynamic model.

Fitted parameters of the thermodynamic models usually valid only in some range of temperature and pressure; they are *sensitive for temperature, first of all*, particularly those applied to liquid-liquid equilibria.

Parameters fitted to **binary** systems ***can be extrapolated*** for calculating equilibria of **multicomponent** mixtures according to experience, probably because effects of ternary and higher order interactions on the equilibria are smaller by decades than those of binary effects. The other way, i.e. applying parameters fitted to ternary systems to calculating equilibria of binary systems, do not work well because measured data inside the composition triangle does not account for behavior of binary systems. For example, for fitting a binary system, the phase curves should precisely reach the pure component's vapor pressure and/or boiling temperature.

Ternary liquid-liquid equilibria cannot always be calculated based on paramaters fitted to binary systems because usually not all the three component pairs are partially mixing; those calculations must be based on parameters fitted to ternary measurements, but those parameters can be used for calculating mixtures of more components.

Parameters fitted to **vapor-liquid** equilibria usually ***cannot be used*** for calculating **liquid-liquid** equilibria, and vice versa. The empirical models discussed in this texts are not based on such firm theoretical background; thus separate parameter sets are needed for calculating both kinds. It is usually so even in case of modelling vapor-liquid-liquid equilibria.

2.10 Phase stability

Checking stability of calculated phases against composition and density fluctuations may be important in simple liquid-liquid equilibria and high pressure phase equilibria. (The usual calculation methods work well for ordinary vapor-liquid equilibria.)

The problem forming at liquid-liquid equilibria can be well visualized on binary cases although the real problem occurs mainly in multicomponent cases. The equilibria are studied at given temperature **and** pressure; thus mole fraction x of one of the two components is the only independent variable against which the calculated excess Gibb energy $\Delta^E G$ is plotted (**Fig. 2.16**). The tangent line drawn to a point of this function intersects vertical axes (at $x = 0$ and $x = 1$) the respective chemical potentials.

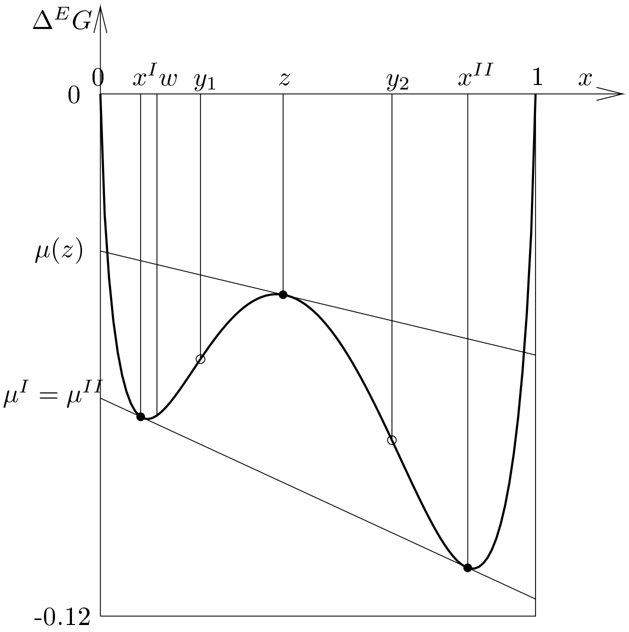

Figure 2.16: $\Delta^E G$ function of methanol – cyclohexane system at 20 C

Function $\Delta^E G(x)$ shows excess Gibbs energy of a (hypothetical) homogeneous mixture with composition x. According to the minimum condition of equilibrium, the phase with composition z is stable if it cannot separate to two phases with compositions x^I and x^{II} being in mutual equilibria in such a way that

$$\lambda \Delta^E G(x^I) + (1 - \lambda)\Delta^E G(x^{II}) < \Delta^E G(z) \tag{2.63}$$

where $z = \lambda x^I + (1 - \lambda)x^{II}$ and $0 < \lambda < 1$.

If inequality (2.63) holds then the homogeneous phase with composition z is unstable. z, x^I, and x^{II}, together with material balance determine the phase ratio. Homogeneous phase with composition z is stable only if inequality (2.63) canot occur with any feasible combination of x^I and x^{II}.

Instability of some compositions is easy to show. Homogeneous phase with composition z is evidently unstable if function $\Delta^E G(x)$ is concave there, i.e. if

$$\left(\frac{\mathrm{d}^2\Delta^E G(x)}{\mathrm{d}x^2}\right)_{x=z} < 0 \tag{2.64}$$

This does not work inversely. Point z in **Fig. 2.16** is unstable because condition (2.64) is staisfied there, but although this condition is **not** satisfied at composition w, this composition is also unstable because the common tangent line drawn to the function in points x^I, and x^{II} runs below $\Delta^E G(w)$, i.e. the Gibb energy is smaller as a combination of the two separate phases than that of the homogeneous phase.

Denote with y_1 and y_2 those compositions at which the sign of the second derivative changes:

$$\left(\frac{\mathrm{d}^2\Delta^E G(x)}{\mathrm{d}x^2}\right)_{x=y_1,y_2} = 0$$

Condition (2.64) is satisfied in all points between y_1 and y_2; we say that the homogeneous mixture is ***locally unstable*** in this domain. Between x^I and y_1, and between x^{II} and y_2, the system is ***locally stable but globally unstable***; the points out of these ranges, near the pure components, are ***locally stable***. Locally stable but globally unstable compositions are called ***metastable***.

If the homogeneous phases x^I and x^{II} are stable in themselves and are in mutual equilibrium then their chamical potentials are equal component by component; this is what the common tangent line points out by intersecting the two vertical axes. This is why the two tangent lines coincide. ***Equality of the chemical potentials, i.e. the common tangent line, is a necessary but not always sufficient condition of phase equilibrium.*** If, for example, function $\Delta^E G(x)$ has more than three local extrema (more than one maximum and two minima, as in **Fig. 2.16**), then there are more common tangent lines (*suboptimal solutions*) whose value in point z is higher than the minimum.

Equality of the chemical potentials is also satisfied by a 'common' tangent line drawn to $x^I = z$ and $x^{II} = z$; ***this is the trivial solution***. Any suboptimal solution is optimal in its small neighborhood. *Sure solution of the stability problem is a global optimum only.*

2.11 Questions

1. What is fugacity?
2. List some well known equations of states! How can they be classified?

3. For chacking phase equilibrium of pure material at given temperature and pressure, how can we calculate fugacity coefficients? Which state variables are their arguments?
4. How can partial fugacity be expressed using equation of states, and how with an excess function model?
5. What is *Poynting* correction for? When may it be neglected?
6. What are mixing rules?
7. What is a lattice model and what is a cell modell?
8. List some frequently applied activity coefficient models!
9. What is modified *Raoult–Dalton* equation?
10. Describe a simple algorithm for calculating bubble point at give pressure and liquid composition with modified *Raoult–Dalton* equation!
11. How does isothermal-isobaric flash calculation work? How is isothermal liquid-liquid distribution calculated? How is it for three phase bubble point?
12. What is area integral check?
13. How can fitted interaction parameters be used for more general equilibria?

2.12 Suggested literature

General works

HÁLA, E.; PICK, J.; FRIED, V.; VILIM, O. *Vapor-liquid equilibrium.* Pergamon press, New York, 1968.

REID, R. C.; PRAUSNITZ, J. M.; SHERWOOD, T. K. *The Properties of Gases and Liquids*, 3rd ed. McGraw-Hill, Inc., New York, 1977.

KING, M. B. *Phase equilibrium in mixtures.* Pergamon Press, Oxford, 1969.

PRAUSNITZ, J.; ANDERSON, T.; GRENS, E.; ECKERT, C.; HSIEH, R.; O'CONNELL, J. *Computer Calculations for Multicomponent Vapor-Liquid and Liquid-Liquid Equilibria.* Prentice-Hall, Inc., NJ USA, 1980.

OONK, H. *Phase Theory. The Thermodynamics of Heterogenous Equilibria*, vol. 3 of *Studies in Modern Thermodynamics.* Elsevier Sci. Pulb., Amsterdam, 1981.

GUGGENHEIM, E. *Mixtures. The theory of the equilibrium properties of some simple classes of mixtures solutions and alloys.* Oxford Univ. Press, London, 1952.

Kemény, S.; Thury, É. Fluid phase equilibria for separation modelling. In *Studies in Computer-Aided Modelling, Design and Operation. Part A: Unit Operations.*, P. I. and F. Zs., Eds. Akadémiai Kiadó - Elsevier, Budapest, 1992, ch. 6.

Prigogine, I.; Bellemans, A.; Mathot, V. *The molecular theory of solutions.* North-Holland Publ. Co., Amsterdam, 1957.

Rowlinson, J.; Swinton, F. *Liquids and Liquid Mixtures*, 3rd ed. Butterworth Sci., London, 1982.

Wei, Y. S.; Sadus, R. Equations of state for the calculation of fluid phase equilibria. *AIChE Journal*, vol. 46, (2000), 169-196.

Christopher G. Gray, Keith E. Gubbins, Christopher G. Joslin Theory of Molecular Fluids, I: Fundamentals, II: Applications Oxford University Press

Other works

Gmeling, U.; Onken, U.; Arlt, W., et al. *Vapor-Liquid Equilibrium Data Collection.* Chemistry Data Series, vol I. Eds: Behrens, D.; Eckermann, R. DECHEMA, Frankfurm a.M., 1977 ...

McDonald, C. M.; Floudas, C. A. Global optimization and analysis for the Gibbs free energy function using the UNIFAC, Wilson, and ASOG equations. *Ind. Eng. Chem. Research*, vol. 34, (1995), 1674-1687.

McDonald, C. M.; Floudas, C. A. Global optimization for the phase and chemical equilibrium problem: Application to the NRTL equation. *Comp. Chem. Eng.*, vol. 19, (1995), 1111-1139.

Michelsen, M. L. The isothermal flash problem. Part I. Stability. Part II. Phase-split calculations. *Fluid Phase Equil.*, vol. 9, (1982), 1-19, 21-40.

Péneloux, A.; Rauzy, E. A consistent correction for Redlich-Kwong-Soave volumes. *Fluid Phase Equil.*, vol. 8, (1982), 7-23.

Chapter 3

Numerical methods

3.1 Equation solving

3.1.1 Zeroes

The task is finding values of variables in equations that satisfy the equations. For example, a 2-variable system is:

$$\begin{aligned} sin(x+2y)/x^2 &= -28.3 \\ x^2 + cos(y) &= 3.5xy \end{aligned}$$

We say that x and y together constitute a solution of the system if both equations are satisfied with them. Introduce two new variables, f and g, functions of x and y:

$$\begin{aligned} f(x,y) &= sin(x+2y)/x^2 + 28.3 \\ g(x,y) &= x^2 - 3.5xy + cos(y) \end{aligned}$$

Values of functions $f(x,y)$ and $g(x,y)$ are generally not zero. x^* and y^* together are called *solution* of the equation system if

$$\begin{aligned} f(x^*,y^*) &= 0 \\ g(x^*,y^*) &= 0 \end{aligned}$$

are satisfied. In that case, array $[x^*,\ y^*]$ is also called *zero of the (system of) equations*. Not all equations have zero, and an equation may have several zeroes. Given x and y, the values $f(x,y)$ and $g(x,y)$ are called *residues* of these functions. x^* and y^* together constitute a solution of the equation (system) if the residues $f(x,y)$ are $g(x,y)$ are zeroes. Thus, zeroes of the equation are also called zeroes of the function (pair) $\{f(x,y)$,$g(x,y)\}$.

General form of an equation (system) is:

$$\begin{aligned} f_1(x_1, x_2, \dots x_n) &= 0 \\ f_2(x_1, x_2, \dots x_n) &= 0 \\ &\vdots \\ f_m(x_1, x_2, \dots x_n) &= 0 \end{aligned}$$

Here each variable $f_1, f_2, \dots f_m$ (altogether m *dependent variables*) may, in principle, depend on all variables $x_1, x_2, \dots x_n$ (altogether n *independent variables*).

Sometimes a dense notation is applied, using arrays (vectors) of x and f:

$$\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{0}$$

Zeroes of an equation are always looked for inside a given domain, e.g. in an interval, in case of a single variable. Generally any domain is enabled, even the whole space. If otherwise not said, however, we always assume continuous independent variables, and their contiguous domain.

The equation may have a solution if $m < n$, $m = n$, or $m > n$.
Example 1. $n = 1$, $m = 2$:

$$f(x, y) = 3x(y - 4)$$

At values $[x, y] = [0, 4]$: $f = 0$, hence the pair $[0, 4]$ is a zero of equation $f(x, y) = 0$.
Example 2. $n = 2$, $m = 1$:

$$f(x) = 3(x + 2)$$

$$g(x) = 1 + x/2$$

Both $f(x) = 0$ and $g(x) = 0$ at $x = -2$, thus -2 is a zero of the system $\{f(x) = 0, g(x) = 0\}$.

Case $m = n$ is the most important in engineering practice. A point in the n-dimension space of the independent variables can be assigned by n independent equations. For example, a smooth curve in the $(x,\ y)$ plain $(n = 2)$ is described by a two-variate $f(x, y) = 0$ equation (that might be transformed to the form of $y = f^*(x)$). Another curve can be described by equation $g(x, y) = 0$. If the two curves intersect then the intersection point is a zero of the system $\{f(x, y) = 0, g(x, y) = 0\ \}$.

One of the equation removed (e.g. remove g, and let remain $f(x, y) = 0$ only), the system has intinite many solutions (all the points of the curve).

If, on the othar hand, the system is supplemented with a randomly chosen third equation, say $h(x, y) = 0$, so that $m > n$, then the system of three equations having a zero is not excluded but its chance is very low. Why would curve $h(x, y) = 0$ go through the zero of $\{f(x, y) = 0,\ g(x, y) = 0\}$? It is possible to find such a curve, but it would be redundant beside $f(x, y) = 0$ and $g(x, y) = 0$. Even if all three equations apply to the same real engeneering or scientific problem, and well written, any small measurement error or numerical error due to inevitable rounding

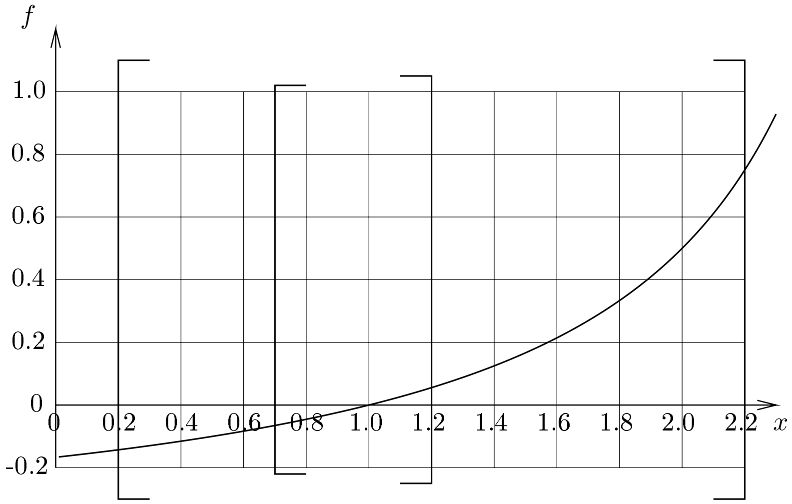

Figure 3.1: Bisection

at multiplication or division can cause them loosing a common zero, even if any two of them would serve us with a good approximation of the proper solution.

Here we discuss *numerical solution* of non-linear equations and equation systems in case $m = n$. We assume you are familiar with the methods of solving linear equations.

3.1.2 Bisection

This is a method applicable to univariate and monotonic functions only.

For example, $f(x) = 1/(3-x)-0.5$ is a strictly monotonian increasing function. We a looking for its zero in the closed interval $[0.2, 2.2]$ (**Fig. 3.1**). Exact value of the zero is 1.0 . The function value at the left hand side end-point of the interval is negative (approximately -0.142857143), and is positive (exactly 0.75) at the right hand side end-point of the interval. We compute the middle point: $x = (0.2 + 2.2)/2 = 1.2$, and evaluate the function here as well. The computed value is $0.05555555\ldots$, positive. Since the function is strictly increasing, it is clear that the zero cannot be found to the right of the middle point, therefore the right hand side end-point of the interval can be shifted left to the present midpoint. From now the zero is looked for in the closed interval $[0.2, 1.2]$. Its midpoint is $(0.2+1.2)/2 = 0.7$; here the function value is negative, hence the midpoint becomes a left hand side end-point, and so on.

In general, first we need somehow bracketing the solution. Once it is done, we compare the signs of the function values calculated in the two end-points and in the midpoint. The midpoint becomes that endpoint with which their signs equal. This works because the function is monotonic increasing or decreasing in the interval. If two end-points of an interval are of the same sign then that interval can be omitted from the search.

This is one simple member of the more general family of domain excluding methods, the only member of them that is easy to organize and perform. All the others are much more demanding.

This simple method works for single variables only, and it divides the length of the interval by 2 in each step; thus the search domain shrinks to 2^{-n} times its original in n steps.

3.1.3 Successive approximation methods

These methods try to approximate the solution $\boldsymbol{x}^*$ by successively approximating it with generating a point series $\boldsymbol{x}^{(0)}$, $\boldsymbol{x}^{(1)}$, $\boldsymbol{x}^{(k-1)}$, $\boldsymbol{x}^{(k)}$, $\boldsymbol{x}^{(k+1)}$, … Point $\boldsymbol{x}^{(0)}$ is the *initial estimate* of the solution and, generally, each $\boldsymbol{x}^{(k)}$ point is its k-th estimate. Methods of successive approximations provide some procedure for generating new approximations from earlier ones. The simples methods utilize just the actual $k-1$-th approximation for generating the k-th one:

$$\boldsymbol{x}^{(k)} = \boldsymbol{g}(\boldsymbol{x}^{(k-1)}) \tag{3.1}$$

Other methods may utilize several earlier approximations:

$$\boldsymbol{x}^{(k)} = \boldsymbol{g}(\boldsymbol{x}^{(k-1)}, \boldsymbol{x}^{(k-2)}, \boldsymbol{x}^{(k-3)}, \dots) \tag{3.2}$$

The methods differ in these procedures (3.1) and (3.2) only.

Any equation $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{0}$ can be transformed to the form $\boldsymbol{x} = \boldsymbol{g}(\boldsymbol{x})$. If no other idea crosser your mind, you can apply transformation $\boldsymbol{g}(\boldsymbol{x}) \equiv \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{x}$, but there are a lot of possibilities. After such a transformation, form (3.1) can be applied.

Any procedure can be used to generate a new approximation. The ususal way is approximating the function with a simpler one with can be solved analytically. Such a function is, e.g. a second order polinomial. The parameters of the approximating functions are fitted to the known points and their function values, and then the zero is calculated; This zero is taken as the new approximation.

In this section only linear approximatios are discussed, i.e. all the methods discussed here are so-called *successive linearization* methods.

Convergence and damping

Successive approximation is successful if the generated series $\boldsymbol{x}^{(0)}$, $\boldsymbol{x}^{(1)}$, $\boldsymbol{x}^{(k-1)}$, $\boldsymbol{x}^{(k)}$, $\boldsymbol{x}^{(k+1)}$, … converges to a zero $\boldsymbol{x}^*$. It is successful in practice if, started from an initial estimate $\boldsymbol{x}^{(0)}$, converges to the demanded zero and fast enough.

Convergence problems are best visualized in univariate case, applying form (3.1). Let $y = g(x)$, and plot $y = g(x)$ in a right angle $y-x$ system. At x-es where

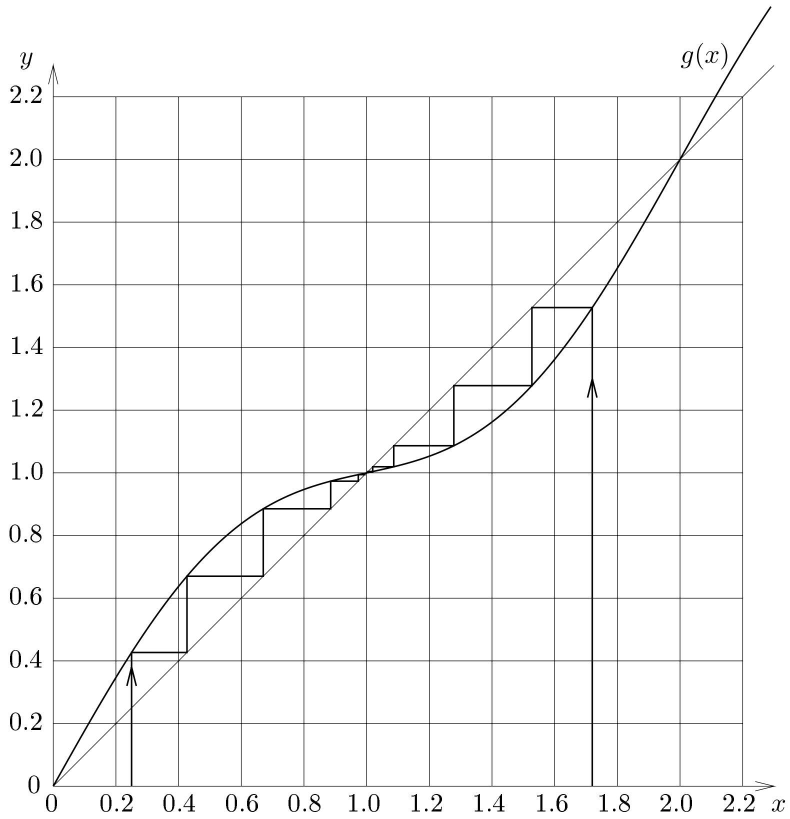

Figure 3.2: $g(x) = x + sin(\pi x)/4$

$f(x) = 0$, $x = g(x)$ is also satisfied, and the curve $y = g(x)$ crosses the straight line $y = x$.

Search, for example, zeroes of $sin(\pi x) = 0$ in the interval $x \in [-0.5, 2.5]$. (The zeroes are $x = 0$, $x = 1$, and $x = 2$.) In this case $f(x) = sin(\pi x)$, and let us choose, for example, $g(x) = x + sin(\pi x)/4$. Function $g(x)$, the straight line $y = x$, and successive estimations with two different initial estimates, namely $x_0 = 0.25$ and $x_0 = 1.72$, are shown in **Fig. 3.2**. Both series converges to zero $x = 1$. Zeroes $x = 0$ and $x = 2$ cannot be approached, the series according to (3.1) are a going away from their neighborhoods. Zero $x = 1$ is called a stable zero, whereas zeroes $x = 0$ and $x = 2$ are unstable with (3.1). Stability depends on the slope of $g(x)$ in the neighborhood of the zero.

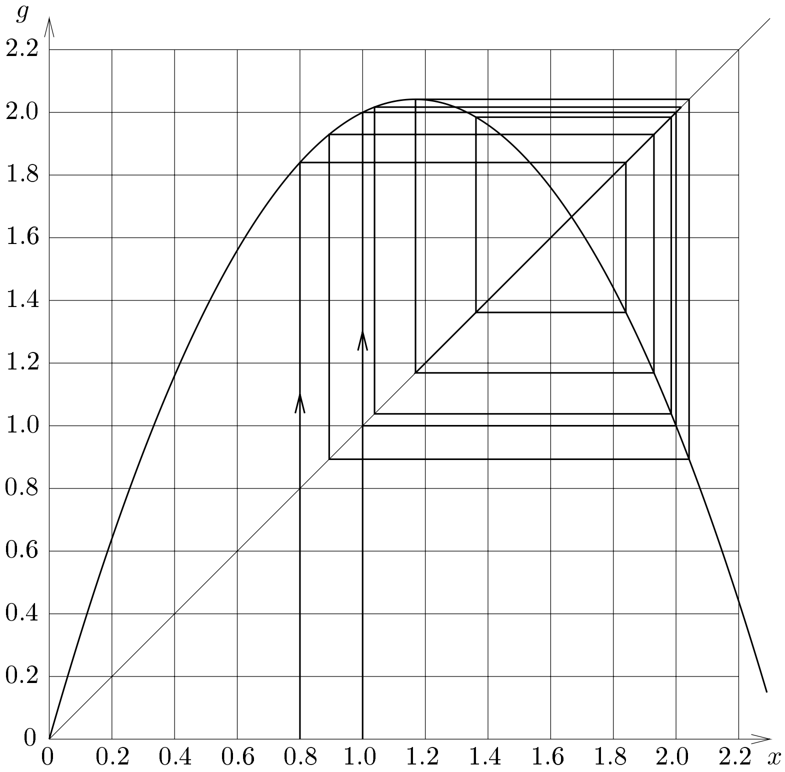

Figure 3.3: $g(x) = 3.5x - 1.5x^2$

A more complex situation is shown in (**Fig. 3.3**) with $g(x) = 3.5x - 1.5x^2$. Start the search anywhere, you cannot reach near the zero $x^* = 1.6666\ldots$ unless start exactly there. There is a couple of oscillating points $x = 1$ and $x = 2$: $g(1) = 2$ and

$g(2) = 1$. Start the calculation from $x = 0.8$, then the series series 'convereges' toward and oscillating system of four points: $[1.168729824, 2.041660282, 0.893245927, 1.929528315]$. One can construct $g(x)$ functions producing oscillating series with two, three, four, five, etc. points, up to apparently acyclic but disordered series.

These considerations are valid to any successive approximation method even if it is not transparent at once. Even if none one of the forms (3.2) is directly applied but residues of $f(x)$ taken at earlier approximations $x^{(k-1)}$, $x^{(k-2)}$, etc. take place in the form of calculating a new approximation $x^{(k)}$, the right hand side of $x^{(k)} = \ldots$ is the actual function $g(x^{(k-1)}, x^{(k-2)}, \ldots)$.

In case of multivariate functions (i.e. multivariate equation systems), any component's behaviour can be considered as a single variable above. Depending on the actual form of function $\boldsymbol{x}^{(k)} = \boldsymbol{g}(\boldsymbol{x}^{(k-1)})$ or $\boldsymbol{x}^{(k)} = \boldsymbol{g}(\boldsymbol{x}^{(k-1)}, \boldsymbol{x}^{(k-2)}, \boldsymbol{x}^{(k-3)}, \ldots)$, and of the initial estimate, the generated series can be convergent or not, and if convergent then either to the demanded solution or not. **No general well behaving, good-for-all, successive approximating method exists. An appropriate search method together with an initial estimator has to be found to each problem experimentally.**

If the selected method does not converge, or is too slow, or oscillates, then *damping* is a cheap way to try mending it with modifying the actual form of $\boldsymbol{x}^{(k)} = \boldsymbol{g}(\boldsymbol{x}^{(k-1)})$ or $\boldsymbol{x}^{(k)} = \boldsymbol{g}(\boldsymbol{x}^{(k-1)}, \boldsymbol{x}^{(k-2)}, \boldsymbol{x}^{(k-3)}, \ldots)$. Let

$$\boldsymbol{x}^{(k)} = \alpha \boldsymbol{x}^{(k-1)} + (1-\alpha)\boldsymbol{g}(\boldsymbol{x}^{(k-1)}, \boldsymbol{x}^{(k-2)}, \boldsymbol{x}^{(k-3)}, \ldots) \tag{3.3}$$

Here α is a so-called *damping factor*. If $\alpha = 0$ then the original method is selected without damping; if $\alpha = 1$ then the last approximation is repeated, the damping is of 100%. If $0 < \alpha < 1$ then the new approximation is somewhere in the straight line section between the last $\boldsymbol{x}^{(k-1)}$ and that suggested by $\boldsymbol{g}(...)$.

Secant method

This is again a method for a single variable only, and a kind of successive linearization. The actual function $f(x)$ is approximated, in each iteration step, by a linear one, $h(x) = ax + b$, fitted to two points. Then the solution of $h(x) = 0$ is the new approximation.

Consider two earlier approximations $x^{(k-1)}$ and $x^{(k-2)}$, together with their respective function values (residues) $f^{(k-1)} \equiv f(x^{(k-1)})$ and $f^{(k-2)} \equiv f(x^{(k-2)})$; these data assign two points in the $f-x$ plain: $(x^{(k-1)}, f^{(k-1)})$ and $(x^{(k-2)}, f^{(k-2)})$. Fit a straight line $h(x) = ax + b$ to these two points:

$$\begin{aligned} f^{(k-1)} &= ax^{(k-1)} + b \\ f^{(k-2)} &= ax^{(k-2)} + b \end{aligned}$$

Subtract the equations: $f^{(k-1)} - f^{(k-2)} = a(x^{(k-1)} - x^{(k-2)})$, hence

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

Table 3.1: Secant methord, first example

k	$x^{(k)}$	$f^{(k)}$
0	0.2	-0.142857143
1	2.2	0.75
2	0.52	-0.096774194
3	0.712	-0.062937063
4	1.06912	0.017898575
5	0.99004672	-0.002475998
6	0.999656015	$-8.6 * 10^{-5}$
7	1.000001712	$4.3 * 10^{-7}$
8	1	$-7.4 * 10^{-11}$

$$b = f^{(k-1)} - ax^{(k-1)}$$

Zero of $h(x) = ax + b = 0$ is $x = -b/a$, thus:

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

This form expresses the ideas incorporated in the method but usually a less informative but simpler form (obtaind by a little algebra) is used:

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

It is transparent that two initial estimates are needed to start the iteration.

Consider, for example, function $f(x) = 1/(3-x) - 0.5$.

Let the two initial estimates first be $x^{(0)} = 0.2$ and $x^{(1)} = 2.2$. The respetive residues are $f^{(0)} = -0.142857143$ and $f^{(1)} = 0.75$. The straight line fittes to these points intersect the x $x^{(2)} = 0.52$, this is the new approximation (**Fig. 3.4**). The series of approximation obtained this way is shown in **Table 3.1**. Let now the two initial estimates be $x^{(0)} = 2.2$ and $x^{(1)} = 1.6$. Then the series of approximations is shown in **Table 3.2**, illustrated in **Fig. 3.5**.

A version of secant method is *regula falsi* or *(double) false position* method. This does not apply always the last two approximations but maintain the approximations so that they bracket the solution, just like in the bisection method.

Multivariate secant method

For N equations and variables, residues of $N+1$ different points can be calculated. Here we would like to use points $\boldsymbol{x}^{(0)}$, $\boldsymbol{x}^{(1)}$, …, $\boldsymbol{x}^{(N)}$ and residues $\boldsymbol{f}^{(0)}$, $\boldsymbol{f}^{(1)}$, …, $\boldsymbol{f}^{(N)}$ to generate a new approximation $\boldsymbol{x}^{(N+1)}$. We suppose that approximations $\boldsymbol{x}^{(0)}$, $\boldsymbol{x}^{(1)}$, …, $\boldsymbol{x}^{(N)}$ span the N-dimension linear space (they give rise to a simplex),

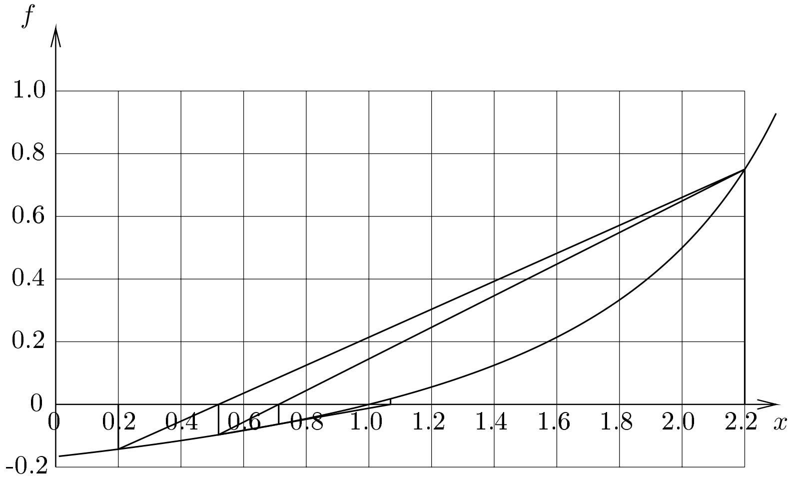

Figure 3.4: Secant methord, first example

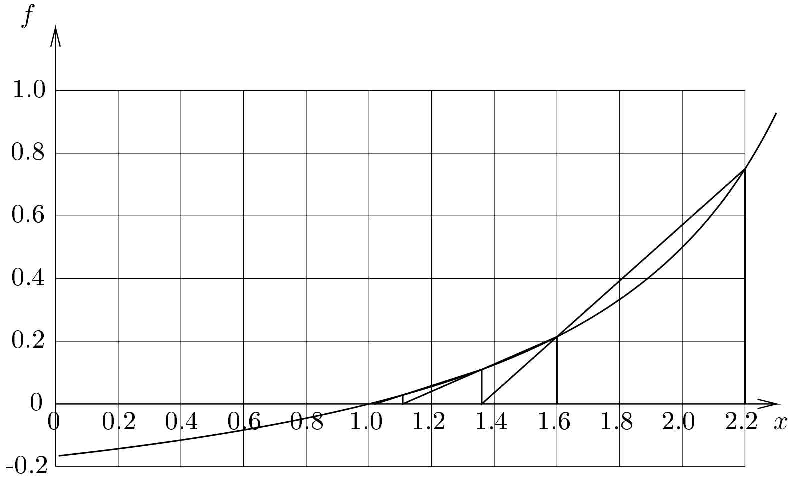

Figure 3.5: Secant methord, second example

Table 3.2: Secant methord, second example

k	$x^{(k)}$	$f^{(k)}$
0	2.2	0.75
1	1.6	0.214285714
2	1.36	0.109756098
3	1.108	0.028541226
4	1.01944	0.004907703
5	1.00104976	0.000262578
6	1.000010204	$2.551 * 10^{-6}$
7	1.000000005	$1.34 * 10^{-9}$

so that an N-dimension hyperplain can be fitted to the $N+1$ points $(\boldsymbol{x}^{(0)},\boldsymbol{f}^{(0)})$, $(\boldsymbol{x}^{(1)},\boldsymbol{f}^{(1)})$, …, $(\boldsymbol{x}^{(N)},\boldsymbol{f}^{(N)})$ in the form of $\boldsymbol{h}(\boldsymbol{x}) = \mathbf{A}\boldsymbol{x} + \boldsymbol{b}$:

$$\begin{aligned} h_0(x_0, x_1, \dots x_N) &= \sum_{i=0}^{N} a_{0,i} x_i + b_0 \\ h_1(x_0, x_1, \dots x_N) &= \sum_{i=0}^{N} a_{1,i} x_i + b_1 \\ &\vdots \\ h_N(x_0, x_1, \dots x_N) &= \sum_{i=0}^{N} a_{N,i} x_i + b_N \end{aligned}$$

The hyperplain should intersect all $N+1$ points, therefore

$$\begin{aligned} \boldsymbol{f}^{(0)} &= \mathbf{A}\boldsymbol{x}^{(0)} + \boldsymbol{b} \\ \boldsymbol{f}^{(1)} &= \mathbf{A}\boldsymbol{x}^{(1)} + \boldsymbol{b} \\ &\vdots \\ \boldsymbol{f}^{(N)} &= \mathbf{A}\boldsymbol{x}^{(N)} + \boldsymbol{b} \end{aligned}$$

(Superscripts denote approximations, indexes denotes space components.)

Subtract each neighboring equations to obtain N vetorial equations where $\boldsymbol{b}$ is already cancelled. Denote differences $\boldsymbol{f}^{(k)} - \boldsymbol{f}^{(k-1)}$ by $\Delta\boldsymbol{f}^{(k)}$, and $\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}$ by $\Delta\boldsymbol{x}^{(k)}$, to obtain:

$$\begin{aligned} \Delta\boldsymbol{f}^{(1)} &= \mathbf{A}\Delta\boldsymbol{x}^{(1)} \\ \Delta\boldsymbol{f}^{(2)} &= \mathbf{A}\Delta\boldsymbol{x}^{(2)} \\ &\vdots \\ \Delta\boldsymbol{f}^{(N)} &= \mathbf{A}\Delta\boldsymbol{x}^{(N)} \end{aligned} \tag{3.4}$$

Combine column vectors $\Delta \boldsymbol{f}^{(k)}$ $(k = 1, 2, \ldots N)$ to obtain an $N * N$ matrix $\mathbf{F}$, and combine respectively column vetors $\Delta \boldsymbol{x}^{(k)}$ to obtain matrix $\mathbf{X}$. In this way equation system (3.4) can be considered as matrix equation

$$\mathbf{F} = \mathbf{A}\mathbf{X}$$

From this matrix equation, $\mathbf{A}$ can formally be expressed as $\mathbf{A} = \mathbf{F}^{-1}\mathbf{X}$, and from equation $\boldsymbol{f}^{(N)} = \mathbf{A}\boldsymbol{x}^{(N)} + \boldsymbol{b}$, formally $\boldsymbol{b} = \boldsymbol{f}^{(N)} - \mathbf{A}\boldsymbol{x}^{(N)}$. By substituting these into $\boldsymbol{h}(\boldsymbol{x}^{(N+1)}) = \mathbf{A}\boldsymbol{x}^{(N+1)} + \boldsymbol{b} = \mathbf{0}$, a new approximation can be expressed:

$$\boldsymbol{x}^{(N+1)} = \boldsymbol{x}^{(N)} - \mathbf{A}^{-1}\boldsymbol{f}^{(N)} = \boldsymbol{x}^{(N)} - \mathbf{X}\mathbf{F}^{-1}\boldsymbol{f}^{(N)}$$

However, computing the inverse $\mathbf{F}^{-1}$ is not needed. Instead, let

$$\boldsymbol{x}^{(N+1)} = \boldsymbol{x}^{(N)} + \boldsymbol{\Delta}^{(N)}$$

where $\boldsymbol{\Delta}^{(N)}$ is obtained by solving the linear equation system

$$\mathbf{F}\boldsymbol{\Delta}^{(N)} = -\mathbf{X}\boldsymbol{f}^{(N)}$$

Wegstein method

To start multivariate secant iteration, $N + 1$ initial estimates are needed, and a linear equation system is to be solved in each step. *Wegstein* suggested, instead, to apply single variate secant formulas for multivariate case, too:

$$x_i^{(k)} = \frac{f_i^{(k-1)} x_i^{(k-2)} - f_i^{(k-2)} x_i^{(k-1)}}{f_i^{(k-1)} - f_i^{(k-2)}} \quad , \quad (i = 1, 2, \ldots N)$$

This means coupling variables with equations, and neglecting effects of other variables when updating them.

Newton method

If the function is differentiable then linearization can be based on a single point instead of two, or $N + 1$. In the approximation $x^{(k-1)}$, let the residue and the derivative be, respectvely, $f^{(k-1)}$ and $\left(\frac{\mathrm{d}f(x)}{\mathrm{d}x}\right)^{(k-1)} = f'^{\,(k-1)}$, then the updating formula is

$$x^{(k)} = x^{(k-1)} - \frac{f^{(k-1)}}{f'^{\,(k-1)}}$$

Solution of the problem presented at secant method, with initial estimate $x_0 = 2.2$, is shown in **Fig. 3.6** and **Table 3.3**.

Newton method is a limit of secant method. It can be obtaine by approaching the two initial estimates to each other. If the function cannot be analytically

Table 3.3: Newton method

k	$x^{(k)}$	$f^{(k)}$
0	2.2	0.75
1	1.72	0.28125
2	1.2592	0.074448529
3	1.03359232	0.008541545
4	1.000564222	0.000141095

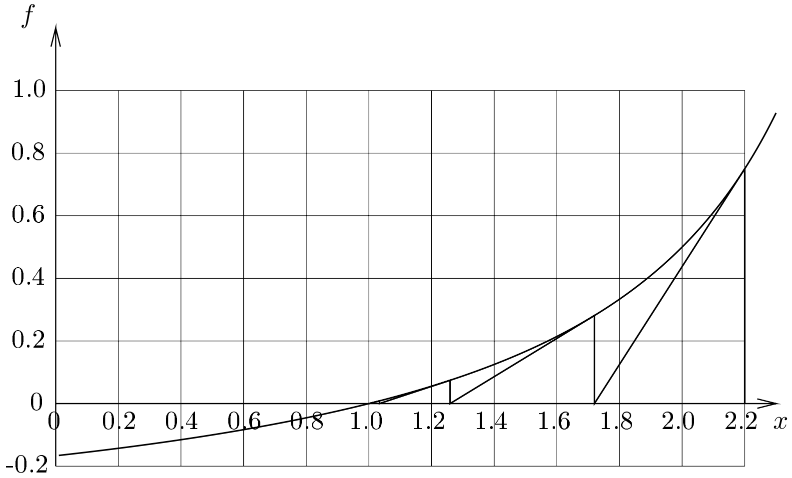

Figure 3.6: *Newton* method

differentiated, for example not given in explitice form, or if its computation is difficult, then the derivative can be approximated numerically:

$$\frac{\mathrm{d}f}{\mathrm{d}x} \approx \frac{f(x+\delta x) - f(x)}{\delta x}$$

In this case the *Newton* method coincides with secant method with the substitution of $x^{(k-2)} = x^{(k-1)} + \delta x$.

Inverse of the matrix *Jacoban*, containing partial derivatives, takes place in the formula in the **multivariate** case:

$$\mathbf{J} \equiv \begin{bmatrix} \dfrac{\partial f_1}{\partial x_1} & \cdots & \dfrac{\partial f_1}{\partial x_N} \\ \vdots & \vdots & \vdots \\ \dfrac{\partial f_N}{\partial x_1} & \cdots & \dfrac{\partial f_N}{\partial x_N} \end{bmatrix}$$

$$\boldsymbol{x}^{(\boldsymbol{N}+\mathbf{1})} = \boldsymbol{x}^{(\boldsymbol{N})} - \mathbf{J}^{-\mathbf{1}} \boldsymbol{f}^{(\boldsymbol{N})}$$

No need, again, for computing inverse of $\mathbf{J}^{-1}$ but simply apply

$$\boldsymbol{x}^{(N+1)} = \boldsymbol{x}^{(N)} + \boldsymbol{\Delta}^{(N)}$$

where $\boldsymbol{\Delta}^{(N)}$ is solution of the linear equation system

$$\mathbf{J}\boldsymbol{\Delta}^{(N)} = -\boldsymbol{f}^{(N)}$$

When applying numerical differentiation, multivariate *Newton* method coincides with multivariate secant method.

Basic features of linearizing methods of successive approximation

The series of approximations generated by *Newton* method coverges to the zero if function $f(x)$ or $\boldsymbol{f}(\boldsymbol{x})$ and its derivatives are monotonian near the zero, and the initial estimate is near the zero. This is well illustrated in **Fig. 3.6** and **Table 3.3. táblázat**.

If, however, the function is not monotonic or the initial estimate is pure then the series of approximations may diverge away from the zero. Such a case is shown in **Fig. 3.7**, where the zero is $x^* = 0.1$. If $x^{(k-1)} < 0.4$ then *Newton* method works well. If $x^{(k-1)} > 0.6$ then the updated approximation is farther from the zero. The figure shows $x^{(k-1)} = 1$ with $x^{(k)} \approx 1.346$. If $x^{(k-1)} \approx 0.55$ then $x^{(k)}$ cannot be computed because $f' \approx 0$. In multivariate case a similar situation is signalled by a (near) singular *Jacobian*.

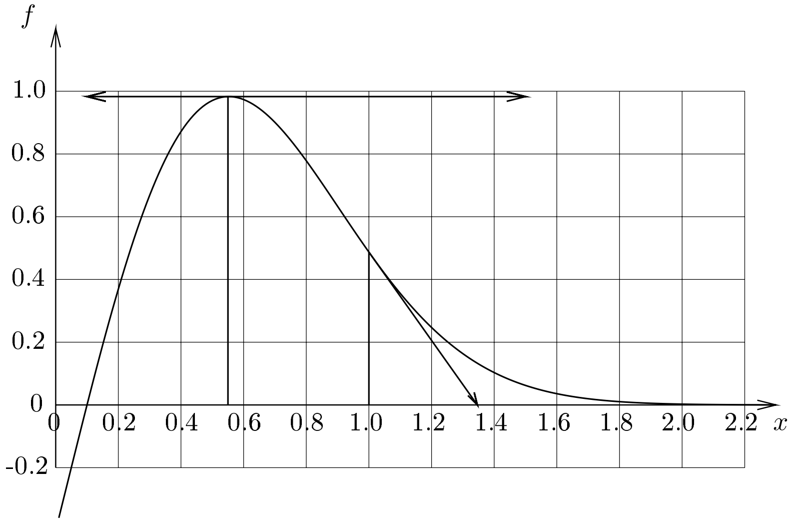

Figure 3.7: Failures with *Newton* method

3.1.4 Transformation to minimization problem

Any zero finding problem $f_i(x_1, x_2, \ldots x_N) = 0$ $(i = 1, 2, \ldots N)$ can be transformed to minimization problem:

$$\min_{\{x_1, x_2 \ldots x_N\}} \sum_{i=1}^{N} f_i^2(x_1, x_2, \ldots x_N) \tag{3.5}$$

In this case the objective is never negative, an since the objective's value is 0 ate the zero of the equation, the zero is the minimal point.

In practice, however, functions are several local minima, and the function value of most of them is not 0. As an example, a univariate function and its square are plotted in in **Fig. 3.8**. Function $f(x)$ has a zero at $x \approx 0.08$; function $f^2(x)$ has local minima at $x \approx 0.08$ and $x \approx 1.21$, and the latter is not a zero of $f(x)$. Many local minima can occur in multivariate functions.

Applying the usual methods of minimum search, only one of the local minima, belonging to the actual initial estimate, is found. That is why this method is not suggested. Just the contrary: search for local minima are formulated as search for zeroes. For example, multivariate *Newton* method was originally suggested as a part of minimizing multivariate non-linear functions: Equations are obtained be equaling partial derivatives with 0, and any solution of such equation is a local minimizer of the original function. This is the *Newton-Raphson* method for finding local minima.

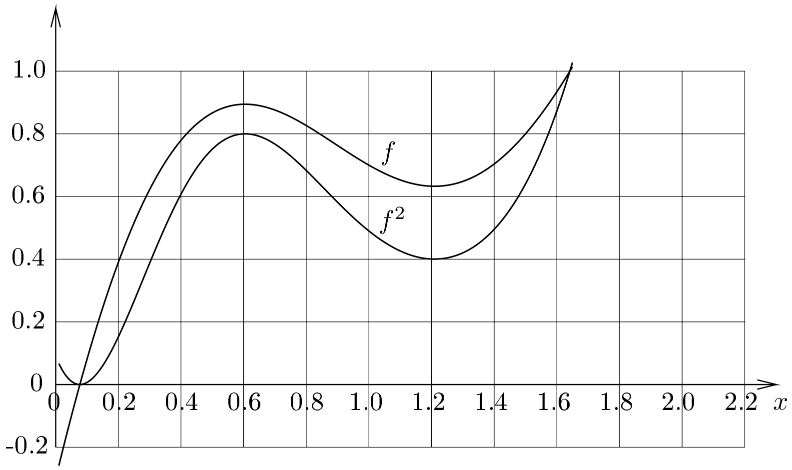

Figure 3.8: Zeroes and minima

3.2 Parameter fitting

The problem is, in mathematical sense, fitting a continuous manifold to some (at least two, but can be many) points of an N-dimension space.

Simplest formulation is the following. Consider an $N-1$ dimension real Euclidean space whose components describe some properties of a physical object. Such properties can be, for example, length, width, temperature, pressure, etc., together forming a property array $\boldsymbol{x}$. We are given a series of such property arrays, $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, $\ldots \boldsymbol{x}_n$, altogether n points in the $N-1$ dimension space. To each of these points an other, originally unknown, property y is determined by measurement: y_1, y_2, $\ldots y_n$. For example, we measure intensity of heat radiation. This is an N-th property. Such an $N-1$-variate real function $f(\boldsymbol{x})$ is looked for that fits to this set of n points. *We already know the form of* f, and it contains some unknown parameters p_1, p_2, $\ldots p_m$; thus the task is to find those parameters that make $f(\boldsymbol{x}; p_1, p_2, \ldots p_m)$ best fitting to the points.

For example, a straight line should be fitted to a set of points in the $y-x$ plain; its slope and itercept are to be determined.

Two main classes of fitting problems are distuingushed.

If exact fitting is needed then the problem is called *collocation.* n equations are obtained for n points if these points are substituted into the function; thus n parameters can be determined by solving an equation system. For example, a straight line can be fitted to two points, a circle to three points in a plain, or a plain to three points in higher space, etc. We do such a thing in case of interpolation, for example.

If the measurement errors are involved, and exact fitting is impossible, then the task is to find a fitting best in some sense. In that case the task is called *regression*. Ususally the number of parameters is less than the number of points. For example, a straight line is fitted to 25 measured points.

In this section *regression* is discussed only. Here the parameter values are selected so that the function obtained this way be as near the 'genuine' function, that would be obtained if no error were present, as possible. However, this genuine function is not known (if it were then no need for regression would occur), and what 'error' is is also a question.

3.2.1 Simple regression

In this case we assume that exact values of the independent variables $\boldsymbol{x}$ are known, i.e. they are not distorted with error. For example, temperature is measured in exactly known time moments, and only temperature is prone to error.

To find optimal fitting, shape of probality distribution of measurable values belonging to each point ought to be known. If this were the case then one could estimate distribution parameters, e.g. expectation and variance, and then find the most probable parameter values. This is called *maximum likelihood method.*

Normal distribution, with expectation φ and variance σ, may frequently be assumed because measurement errors are caused by many, small, independent effects. The task is to determine parameter values so that function values f_1, f_2, $\dots f_n$ calculated in points $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, $\dots \boldsymbol{x}_n$ best approach expectations φ_1, φ_2, $\dots \varphi_n$.

Least squares

Suppose, as a first case, that variance σ is independent on the place of measurement, i.e. the same variance σ is valid in all the $\boldsymbol{x}_i$ points. Then the probability density function of f_i measurable at points $\boldsymbol{x}_i$ is

$$\phi_i(f_i) = \exp\left(-\frac{(f_i-\varphi_i)^2}{2\sigma^2}\right)$$

Since the measured values, and their errors, are independent, the combined density function of all the measurements is their product:

$$\phi(\boldsymbol{f}) = \prod_{i=1}^{n} \exp\left(-\frac{(f_i-\varphi_i)^2}{2\sigma^2}\right)$$

Optimal regression is obtained at maximum of this probability density. It is easier, however, to minimize the negative logarithm of this product:

$$\min_{\{a,b,c,\dots\}} \sum_{i=1}^{n} \frac{(f_i-\varphi(\boldsymbol{x}_i,a,b,c,\dots))^2}{2\sigma^2}$$

where parameters $a, b, c, \dots$ of function $\varphi(a, b, c, \dots)$ are fitted, and we assume that respective values of function $\varphi(\boldsymbol{x}_i, a, b, c, \dots)$ would have been obtained by 'precise' measurement. (I.e., the unknown expectations are determined by fitting parameters $a, b, c, \dots$.)

Factor σ^2 is constant, it can be lifted up from the braces. Its value does not influence the minimizer point, i.e. the point $a, b, c, \dots$ at which the minimum is located, thus the following minimization promlem is equivalent to the original one:

$$\min_{\{a,b,c,\dots\}} \sum_{i=1}^{n} (f_i - \varphi(\boldsymbol{x}_i, a, b, c, \dots))^2$$

This is *method of least squares.* If the minimum is searched for in an infinite space of parameters $a, b, c, \dots$, or if we can be sure that the minimum is inside the search domain, then searching for local minima is enough. These minima are amongst the zeroes (solutions) of the following equation system:

$$\begin{aligned} 2\sum_{i=1}^{n}\left((f_i - \varphi(\boldsymbol{x}_i, a, b, c, \dots))\frac{\partial\varphi(\boldsymbol{x}_i, a, b, c, \dots)}{\partial a}\right) &= 0 \\ 2\sum_{i=1}^{n}\left((f_i - \varphi(\boldsymbol{x}_i, a, b, c, \dots))\frac{\partial\varphi(\boldsymbol{x}_i, a, b, c, \dots)}{\partial b}\right) &= 0 \\ &\vdots \end{aligned} \tag{3.6}$$

In case of linear function $\varphi(a, b, c, \dots)$, e.g. fitting straight line or plain, the objective is a sum of terms quadratic in parameters $a, b, c, \dots$. As a result, the first member in system (3.6) is a linear function of the parameters, and the derivatives *do not* contain the unknown parameters, thus a linear equation system in the parameters is obtained. For fitting a plain, for example, the equation system is

$$\begin{aligned} \sum_{i=1}^{n}(f_i - (a + bx_i + cy_i)) \quad &= 0 \\ \sum_{i=1}^{n}(f_i - (a + bx_i + cy_i))\, x_i &= 0 \\ \sum_{i=1}^{n}(f_i - (a + bx_i + cy_i))\, y_i &= 0 \end{aligned}$$

Here values f_i, x_i, y_i $(i = 1, 2, \dots, n)$ are known and values of parameters a, b, and

c are looked for, i.e. the equation can be writtes as

$$
\begin{aligned}
(n)\,a \quad + \quad \left(\sum_{i=1}^{n} x_i\right) b + \quad \left(\sum_{i=1}^{n} y_i\right) c &= \left(\sum_{i=1}^{n} f_i\right)\\
\left(\sum_{i=1}^{n} x_i\right) a + \left(\sum_{i=1}^{n} x_i^2\right) b + \left(\sum_{i=1}^{n} x_i y_i\right) c &= \left(\sum_{i=1}^{n} f_i x_i\right)\\
\left(\sum_{i=1}^{n} y_i\right) a + \left(\sum_{i=1}^{n} x_i y_i\right) b + \left(\sum_{i=1}^{n} y_i^2\right) c &= \left(\sum_{i=1}^{n} f_i y_i\right)
\end{aligned}
$$

The bracketed sums are known, these are the coefficients and the right hand side of the equations.

Generalized linear regression. Fit, to a set of points, a function $\varphi(\boldsymbol{x}) = \sum_{k=1}^{m} a_k g_k(\boldsymbol{x})$ where functions $g_k(\boldsymbol{x})$ are known, given, do not contain unknown parameters. (Usually members of a so-called othogonal series. For example: $g_k(\boldsymbol{x}) = \boldsymbol{x}^{k-1}$.) Then $\varphi(\boldsymbol{x})$ depends linearly on parameters a_k, and the method of least squares can be used.

Weighted least squares

Generally, variances $\sigma_1, \sigma_2, \ldots \sigma_n$ at the different measurement points are different, and the problem to be solved is

$$
\min_{\{a,b,c,\ldots\}} \sum_{i=1}^{n} \frac{(f_i - \varphi(\boldsymbol{x}_i, a, b, c, \ldots))^2}{\sigma_i^2}
$$

i.e.

$$
\begin{aligned}
\sum_{i=1}^{n} \left(\frac{f_i - \varphi(\boldsymbol{x}_i, a, b, c, \ldots)}{\sigma_i^2} \; \frac{\partial \varphi(\boldsymbol{x}_i, a, b, c, \ldots)}{\partial a} \right) &= 0\\
\sum_{i=1}^{n} \left(\frac{f_i - \varphi(\boldsymbol{x}_i, a, b, c, \ldots)}{\sigma_i^2} \; \frac{\partial \varphi(\boldsymbol{x}_i, a, b, c, \ldots)}{\partial b} \right) &= 0\\
&\vdots
\end{aligned}
$$

If a linear function is fitted then linear equation system is obtained in this case, too.

In this way the values with smaller variance are taken into account with larger weights, the more uncertain ones with smaller weights.

Absolute values of the variances are need not be known, their ratio is enough. If we know how all the variances σ_i are related to variance σ^* in an arbitrarily selected point $\boldsymbol{x}^*$, and write $\sigma_i^* \sigma^*$ instead of σ_i. Factor σ^* can be lifted up from the brackets, and omitted.

Non-normal distribution of erros

If error distribution is not normal then method of least squares is not the optimal regression. A formula may be derived if the shape of the distribution is known, otherwise approximating methods must be applied.

If the distribution is not known but the measurements fit to some function according to visual observation then method of least squares may be approximately optimal. Otherwise, e.g. if there are many errors due to mistype or unusual reasons with distribution far from normal then minimizing sum of absolute errors or some even simpler statistic is suggested.

Expectable error, and reliability

Usually one has to calculate reliability of the obtained results. This is an important question if we are not sure whether the selected function shape well describes the relation represented by the measured data.

Even if this is not the question, and we know the shape of the function to fit, knowledge on the applicable set of parameter values may be important. Particularly interestig is the relation of possible values of different parameters, and the shape of level lines or surfaces. This is related to the question of if and how much are the parameters correlated.

If the points of equal probability form a circle around the optimal parameter vector, in case of two parameters, then the parameters are not correlated. If they form an elongated ellips then the parameters are strongly correlated.

3.2.2 Statistical regression

Generally even the independent variables are subject to error. One cannot be sure, for example, in the time moment when a temperature measurement was made. Thsu both temperature and time are uncertain. At measuring vapor pressure, not just the measure pressure but the temperature is also uncertain. All these errors must be taken into account at fitting.

In such cases there is no reason to make distinction between independent and dependent variables. Instead, we consider measured points in an N-dimension space of all components, including both 'independent' and 'dependent' variables, and look for a hypesurface well fitting on the points. The least squares method means in this case means the task

$$\min_{\{a,b,c,\dots\}} \sum_{i=1}^{n}\sum_{j=1}^{N} \frac{(x_{i,j}-\xi_{i,j}(a,b,c,\dots))^2}{\sigma_{i,j}^2}$$

$$\boldsymbol{f}(\boldsymbol{\xi}) = \boldsymbol{0}$$

where $\boldsymbol{x}$ are the measured points, $\xi_{i,j}$ is expectation of variable j in measurement point i, and equations $f_k(\boldsymbol{\xi}) = 0$ $k = 1, 2, ..., M)$ are relations of the fitted hypersurface. This is a minimization task with equation type constraints.

3.3 Initial value problems

A typical problem is modelling change of a property or some properties along a single variable, ususally time, and the change can be modelled by a differential equation. For example we study time dependence of temperature in a cooling process, or concentration's change in time in a mixed tank. Time takes place as independent variable in dynamic balances.

The independent variable is not always time. For example, one can study material properties along the length of an tubular reactor. But the independent variable can be anything else, even concentration, depending on the studied physical problem.

Such changes are modelled by ordinary differential equations (ODE-s). Their general form is

$$F(t, x, \frac{\mathrm{d}x}{\mathrm{d}t}, \frac{\mathrm{d}^2x}{\mathrm{d}t^2}, \frac{\mathrm{d}^3x}{\mathrm{d}t^3}, \dots) = 0 \tag{3.7}$$

where t is independent variable, x is dependent variable, $x(t)$ is the unknown function, and $F(t, x, x', x'', \dots)$ is a given, known, expression. If several variables (x_1, x_2, x_3, etc.) depend simultaneously on the same independent variable t, modelled with an ODE, and in expressions F_i not only x_i, x_i', x_i'', etc. dependent variables take place by other variables and their derivatives x_j, x_j', x_j'', etc., too, then these differential equation together form a system of ordinary differential equations.

Higher order ODEs can be transformed to first order ODE systems by introducing new dependent variables. For examples, let $y \equiv \frac{\mathrm{d}x}{\mathrm{d}t}$, $z \equiv \frac{\mathrm{d}^2x}{\mathrm{d}t^2}$, etc., then ODE (3.7) is equivalent to

$$\begin{aligned} F(t, x, y, z, \dots) &= 0 \\ \frac{\mathrm{d}x}{\mathrm{d}t} &= y \\ \frac{\mathrm{d}y}{\mathrm{d}t} &= z \\ &\vdots \end{aligned} \tag{3.8}$$

An ODE generally has infinitely many solutions. Subsets of these solutions are assigned by *boundary conditions*. For assigning particular solution of an N-th order ODE (i.e. an system of N first order ODEs), a system of N boundary conditions are needed.

The N boundary conditions may contain diferent values of t, the independent variable. In other words, the boundary conditions may express some constraints on the values and derivatives of the dependent variables at different t values. If, however, all boundary constraints refer to the same independent value t_0 then they are called *initial conditions*. If these conditions are in the form of $G_j(t_0, x, x', \dots) = b_j$ where values b_j are given then these are called *initial values*, the problem itself is known as *initial value problem*.

From now on we assume that the ODE is given in explicite first oder form. In the case of a single dependent variable this means:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = f(t, x)$$

Its form in several dependent variable is:

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = f_i(t, x_1, x_2, \dots) \qquad (i = 1, 2, \dots)$$

In this section we discuss numerical solution of ODEs.

3.3.1 Euler method

Discretize values of t in equal gaps (stepsize) h (this is also called equidistant discretization), obtaining points $t_0 = 0$, $t_1 = h$, $t_2 = 2h$, etc. The respecting (but yet unknown) function values are $x_0 = x(t_0)$, $x_1 = x(t_1)$, $x_2 = x(t_2)$ etc. Expand the unknown function $x(t)$ around some point t_i in the forward direction (*forward differences*), and stop after the first member:

$$x_{i+1} = x_i + h \left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)_i + \mathfrak{o}\left(h^2\right)$$

where $\mathfrak{o}\left(h^2\right)$ means that the remaining error can be approximated with a polinomial second degree in h. Index i from here on refers to point t_i.

Substitute the right hand side of the explocite ODE to the place of the derivative, to obtain *Euler*'s method:

$$x_{i+1} \approx x_i + h f(t_i, x_i)$$

or, with simpler notation:

$$x_{i+1} \approx x_i + h f_i$$

This the simplest method with constant h stepsize. *Euler*'s method approximates the differential equation with an equation of finite differences.Since value of x_1 obtained in the first step is not exactly precise, it does not lie on the exact particular solution. The next approximate function value, x_2, is cmputed from this, already imprisise, point, so that its inaccuracy is higher. Numerical errors accumulate step by step. There are two main consequences: (1) The numerical solution obtained this way does not coincide to any of the particular solutions, because the actual function value jumps to another particular solution in each step. (2) The more steps are made, the larger the numerical error at the end of the calculation.

Euler's method is unstable in some cases. Consider, for example, the first order linear ODE

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -B^2 x$$

Its exact solution is

$$x(t) = -x(0)B^2 \exp(-B^2 t)$$

where $x(0)$ is the initial value. Each solution (the particular solution belonging to a particular value of $x(0)$) approaches 0 at infinite time. Steps of the *Euler* method are

$$x_{i+1} = x_i - B^2 x_i h = \left(1 - B^2 h\right) x_i$$

Since B and h are constant,

$$x_{i+n} = \left(1 - B^2 h\right)^n x_i = A^n x_i$$

where $A = \left(1 - B^2 h\right)$. Limit of the obtained geometric series depends on A.

$A^n \to \dfrac{1}{1-A}$ if $|A| < 1$, and $A^n \to \infty$ if $|A| > 0$. If $h > \dfrac{2}{B^2}$ then $1 - B^2 h < -1$, $\left|1 - B^2 h\right| > 1$, and the computed solution diverges toward infinity. In order to keep the exponential function in its real track numerically, a step size far below $\dfrac{2}{B^2}$ must be applied. Applying such a small stepsize has two inconvenient consequences. A smaller one is the long computation time it gives rise to. The larger one is the accumulation of rounding errors this way, so that the computed function will be impresice at higher t values.

Such an unstability can be avoided by applying *backwards differences*, i.e. expanding the function backwards in time:

$$x_i = x_{i+1} - h \left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)_{i+1} + \mathfrak{o}\left(h^2\right)$$

By rearrangement one gets the *implicite Euler method*:

$$x_{i+1} \approx x_i + h f_{i+1}$$

that is, by substituting the ODE:

$$x_{i+1} = x_i - h B^2 x_{i+1}$$

and finally:

$$x_{i+1} = \frac{x_i}{1 + h B^2}$$

This is stable irrespectively to h because $1 + hB^2 > 1$.

3.3.2 Stiff equations

There are initial value problems that either need very small step size to calculate with acceptable precision or even impossible to do it with any small stepsize. This stifness is an inherent property of such problems.

Example 1: Find solution of

$$\frac{\mathrm{d}^2 x}{\mathrm{d}t^2} = 100x$$

with initial values

$$\begin{aligned} x(0) &= 1 \\ \frac{\mathrm{d}x}{\mathrm{d}t}(0) &= -10 \end{aligned}$$

and increasing t. The exact general solution is

$$x(t) = A\exp(-10t) + B\exp(10t)$$

where A and B are integrating constants depending on the particular initial values. By substitution one gets $A = 1$, $B = 0$, i.e. the exact particular solution is:

$$x(t) = \exp(-10t)$$

Value of the solution function a 0 is 1, and is exponentially attenuating toward 0 with increasing t. When the ODE is solved numerically, however, there is always some small, any small, rounding error. This manifests in some small but positive ε value of B:

$$x(t) = \exp(-10t) + \varepsilon\exp(10t)$$

Any small ε is, the second member dominates at higher t values, and the numerial solution turns upward.

Example 2: Find solution of the first order linear ODE system

$$\begin{aligned} \frac{\mathrm{d}x}{\mathrm{d}t} &= 998x + 1998y \\ \frac{\mathrm{d}y}{\mathrm{d}t} &= -999x - 1999y \end{aligned}$$

with initial values $x(0) = 1$ and $y(0) = 0$. The exact solution is

$$\begin{aligned} x(t) &= 2e^{-t} - e^{-1000t} \\ y(t) &= -e^{-t} + e^{-1000t} \end{aligned}$$

The coefficient matrix is weakly conditioned (its determinant is near zero), and hence the solutions contains exponents different with magnitudes. The same consequence occurs as in the first example. To obtain a numerically stable solution, the step size must be far lower than 0.001. Otherwise the solution switches to and from different particulat solutions rapsodically.

In some cases stiffness can be overcome by applying well selected varable transformation. By substituting $x = \frac{1}{y}$ in Examle 1, the ODE for y is already not stiff.

One first solve this equation numerically, and then transforms the solution batck to x.

Unstability similar to that in Example 2 occurs in modelling dynamics of countercurrent multistage separation systems (distillation, absorption) because dynamics of vapor velocity with pressure, liquid flow rate with hold-ups, and finally concentration changes are characterized with time constants different with magnitudes. Step size of dynamic simulation should be selected to the fastest process (pressure and vapor flow, some seconds) but then very long simulation is needed to obtain a picture on concentraion changes (hours). Thus, their dynamic modelling is rather difficult although their control can be detached because of the same reason.

3.3.3 Runge-Kutta methods

Accuracy, and thus step lentght can be increased by expanding the series to higher order.

Runge-Kutta methods are such higher order approximations, and can be considered as improvements of the *Euler* method. According to the *Euler* method, change of the unknown function x during a step h is determined by the slope at the beginning of the interval $f(t_{i-1})$, thought the slope can be rather different at the end of the interval$f(t_{i-1}+h)$, i.e. $f(t_i)$. Naturally, there is an integral average slope $\hat{f}$ that would result in the exact value:

$$x(t+h) = x(t) + h\hat{f}$$

The *Runge-Kutta* methods differ from each other in how this average slope $\hat{f}$ is approximated. *Euler* method approximates it with the slope at the beginning of the interval.

The simplest *Runge-Kutta* method is the following. First approximate the change of x along the interval of lengt h with the slope computed at the beginning of the interval (this is $k = f(t, x(t)h)$), and make a half-step along the straight line connecting point $t, x(t)$ to point $t+h, x(t)+k)$ to get to point $(t+h/2, x(t)+k/2)$. Then approximate the step with the slope at this half-step point. This is conventionally denoted as

$$k_1 = hf(t_i, x_i)$$

$$k_2 = hf(t_i + \frac{1}{2}h, x_i + \frac{1}{2}k_1)$$

$$x_{i+1} = x_i + k_2$$

Hanbbooks list several *Runge-Kutta* methods. A preferable one is the so-called fourth-order RK method because of its symmetric nature that makes it almost fifth order in practice:

$$
\begin{aligned}
k_1 &= hf(t_i, x_i)\\
k_2 &= hf(t_i + \frac{h}{2}, x_i + \frac{k_1}{2})\\
k_3 &= hf(t_i + \frac{h}{2}, x_i + \frac{k_2}{2})\\
k_4 &= hf(t_i + h, x_i + k_3)\\
x_{i+1} &= x_i + \frac{k_1}{6} + \frac{k_2}{3} + \frac{k_3}{3} + \frac{k_4}{6}
\end{aligned}
$$

3.4 Boundary value problems

If there are several independent variables in a differential equation and the boundary conditions are not related to just a single point but to more ponts in the space of the independent variables the we speak about a *boundary value problem.*

Practical examples.

1 Distribution of concentration and temperature is looked for along a cross section, situated some distance from the feed point, of a tubular reactor filled with catalyst on solid carrier, given either wall temperature or heating power or insulation. At the boundary of the cross setion the consditions may be taken symmetric. Frequently even the inside is symmetric in a cylindrical device, and then the problem becomes univariate, alond radius. In case of structured packing, and diffusion and heat conductivity changes according to place and direction, then the problem can be rather difficult.

2 Distributions are to be computed along a cross section of a tube in a multitube reactor. In this case the boundary conditions are not symmetric at the tubes situated near the outer part of the bundle.

3 Temperature distribution is looked for in a shell-and-tube heat exchanger.

4 Modelling flam, furnaces, exoterm rections in three dimension space with asymmetric boundaries, gas flow patterns etc.

5 Modelling trickling or diffusion of contamination in soild and air.

Such problems are discussed in conventional chameical engineering with simplified models applying the principle of similarity and analogy (*Reynolds*, *Nusselt*, *Prandtl*, *Peclet*, *Damkoehler* etc. numbers). More precise modelling is possible with numerical solution of differential equations, and sometimes they are really needed.

These problems are modelled with partial differential equations. Here we will denote the independent variables with x, or their array with $\boldsymbol{x}$. A single dependent variable is considered, for simplicity, denoted by U. Here U denotes the (unknown) exact solution, whereas u will denote its computed approximation.

The domain of independent variables can be of any shape. For example, if there are two intependent variables then we may look for an unknown function over a rectangle, or over a full circle area, or over just the contour line of the circle area, or over a non-regural polynom, etc. This **domain** will be denoted here by Ω. *In case of several variables, the* **boundary** *of domain Ω is not a set of discrete points but an extended shape: a line in the plain, a surface in the three-dimension space, etc. Boundary conditions are given over this boundary. Hence boundary conditions can be differential equations as well as ordinary equations.*

3.4.1 Target shooting

The simplest boundary value problem conceivable is a second order ordinary differential equation

$$\frac{\mathrm{d}^2U}{\mathrm{d}x^2} = f(x, U, \frac{\mathrm{d}U}{\mathrm{d}x})$$

to be solved over interval $a \leq x \leq b$, with boundary conditions $U(a) = U_a$ and $U(b) = U_b$. Expression $f(x, U, U')$ is known, and the unknown function $U(x)$ is looked for.

Its simplest numeric solution method is *target shooting*. If we new the derivative $U'(a)$ instead of U_b then we could find $u(x)$ numerically with a method of *section 3.3* solving the initial value problem obtained. Since this initial value of U' is not known, we estimate it: $U' \approx V$, and solve the initial value problem with this estimate. As a result, we obtain numerically a function $u(x)$, and its final value $u(b)$ denoted with u_b. This u_b is a function of the initial value of the derivative $U' \approx V$; one can denote this function with $u_b(V)$.

If V is a good estimate of $U'(a)$ then the computed value equals the specified boundary value: $u_b(V) \approx U_b$. If, by chance, the estimate is precise then $V = U'(a)$, and $u_b(V) = U_b$. That is, the problem can be transformed to solving an equation of form $u_b(V) = U_b$, and we can apply the methods discussed in *section 3.1.1*.

3.4.2 Finite differences

Not only function $U(x)$ but its derivatives are also unknown. In *method of finite diferences* the independent variable x is discretized and $U(x)$ is calculated in discrete x_i points only. These values are denoted by u_i, and the derivatives are estimated with whem. The so estimated derivatives are substituted to the differential equation to obtain an ordinary equation system solvable for u_i.

Single variable finite differences

Forward differences are obtained on an interval $[x_i,\ x_{i+1}]$ of length h by expanding the unknown function from one end:

$$U_{i+1} = U_i + h\left(\frac{\mathrm{d}U}{\mathrm{d}x}\right)_i + \mathfrak{o}(h^2)$$

Hence approximation of the first derivative is

$$\left(\frac{\mathrm{d}U}{\mathrm{d}x}\right)_i = \frac{U_{i+1} - U_i}{h} + \mathfrak{o}(h)$$

i.e.

$$\left(\frac{\mathrm{d}U}{\mathrm{d}x}\right)_i \approx \frac{\Delta_i^{(1)}}{h} = \frac{u_{i+1} - u_i}{h}$$

Higher order derivatives are approximated in the same way:

$$\begin{aligned}
\left(\frac{\mathrm{d}^2U}{\mathrm{d}x^2}\right)_i &\approx \frac{\Delta_i^{(2)}}{h^2} = \frac{\Delta_{i+1}^{(1)} - \Delta_i^{(1)}}{h^2} = \frac{u_{i+2} - 2u_{i+1} + u_i}{h^2} \\
\left(\frac{\mathrm{d}^3U}{\mathrm{d}x^3}\right)_i &\approx \frac{\Delta_i^{(3)}}{h^3} = \frac{\Delta_{i+1}^{(2)} - \Delta_i^{(2)}}{h^3} = \frac{u_{i+3} - 3u_{i+2} + 3u_{i+1} - u_i}{h^3} \\
&\text{stb.}
\end{aligned}$$

Central differences are obtained on an interval $[x_i,\ x_{i+1}]$ of length h by expanding the unknown function from the midpoint to both directions:

$$\begin{aligned}
U_{i+\frac{1}{2}} &= U_i + \frac{h}{2}\left(\frac{\mathrm{d}U}{\mathrm{d}x}\right)_i + \frac{(+h)^2}{8}\left(\frac{\mathrm{d}^2U}{\mathrm{d}x^2}\right)_i + \mathfrak{o}(h^3) \\
U_{i-\frac{1}{2}} &= U_i - \frac{h}{2}\left(\frac{\mathrm{d}U}{\mathrm{d}x}\right)_i + \frac{(-h)^2}{8}\left(\frac{\mathrm{d}^2U}{\mathrm{d}x^2}\right)_i + \mathfrak{o}(h^3)
\end{aligned}$$

Difference of the two equations is:

$$U_{i+\frac{1}{2}} - U_{i-\frac{1}{2}} = h\left(\frac{\mathrm{d}U}{\mathrm{d}x}\right)_i + \mathfrak{o}(h^3)$$

where $\mathfrak{o}(h^3)$ is an expression of degree of error. Hence

$$\left(\frac{\mathrm{d}U}{\mathrm{d}x}\right)_i = \frac{U_{i+\frac{1}{2}} - U_{i-\frac{1}{2}}}{h} + \mathfrak{o}(h^2)$$

Error of the forward differences scheeme is of first degree, whereas that of the central differences is of second degree. Applying central differences is prefereable because on can reach higher occuracy with the same stepsize (mesh or grid spacing).

Central differences are technically easier to work with if mesh size $2h$ is applied, then no half indexes are used:

$$\left(\frac{\mathrm{d}u}{\mathrm{d}x}\right)_i = \frac{u_{i+1} - u_{i-1}}{2h}$$

$$\left(\frac{\mathrm{d}^2u}{\mathrm{d}x^2}\right)_i = \frac{u_{i+2} - 2u_i + u_{i-2}}{4h^2}$$

$$\left(\frac{\mathrm{d}^3u}{\mathrm{d}x^3}\right)_i = \frac{u_{i+3} - 3u_{i+1} + 3u_{i-1} - u_{i-3}}{8h^2}$$

Multivariate finite differences

Pure and mixed partial derivatives are to be approximated. The simplest case is a right angle grid or mesh parallel with the axes of the co-ordinate system, and points to be evaluated are allocated at the eintersection point of the grid. Equal mesh size h in each direction makes the problem easiest to deal with. If possibel, central differences are applied.

Firs, second, third etc. derivatives according to either variable are formed in the same way as in the case of a single variable. For example, second derivative according to the first variable is approximated as

$$\left(\frac{\partial^2u}{\partial x_1^2}\right)_{i,j} = \frac{u_{i+2,j} - 2u_{i,j} + u_{i-2,j}}{4h^2}$$

Mixed partial derivatives can be formed by first differentiating according to one variable, and the according to the other one. For example:

$$\left(\frac{\partial^2u}{\partial x_1 \partial x_2}\right)_{i,j} = \frac{u_{i+1,j+1} - u_{i-1,j} - u_{i,j-1} + u_{i-1,j-1}}{4h^2}$$

More complex expressions can be made in the same way. For example, *Laplacian* operator is approximated as

$$\nabla^2 u_{i,j} \equiv \left(\frac{\partial^2u}{\partial x_1^2}\right)_{i,j} + \left(\frac{\partial^2u}{\partial x_2^2}\right)_{i,j} = \frac{u_{i,j+1} + u_{i-1,j} + u_{i,j-1} + u_{i+1,j} - 4u_{i,j}}{4h^2}$$

or the same rotated with 45 degrees:

$$\nabla^2 u_{i,j} = \frac{u_{i+1,j+1} + u_{i-1,j+1} + u_{i-1,j-1} + u_{i+1,j-1} - 4u_{i,j}}{8h^2}$$

Formulas can be derived for derivatives if different mesh sizes are applied in different directions, in slanted (not right-angle) systems, or even in curved systems. Ready made approximating formulas are provided in handbooks for 2 and 3-dimension systems.

The method of finite differences

By substituting the formulas of finite differences into the differential equation(s) one obtains ordinary equations for the unknown variables u_i or $u_{i,j,...}$. Linear differential equations give rise to a linear equation system that can easily be solved.

As an **example**, solve *ordinary* differential equation

$$\frac{\mathrm{d}U}{\mathrm{d}x} + 2U = 1$$

over domain $\Omega = \{x|0 \leq x \leq 1\}$ with condition $U(0) = 1$. This is an initial value problem but method of finite differences can be applied here as well. (We solve this very simple problem so that it is easy to follow up.)

Interval $[0, 1]$ is subdivided to 4, equal size $h = 0.25$, parts, and look for the solutions in the grid points only, i.e. at $x_1 = 0.25$, $x_2 = 0.5$, $x_3 = 0.75$, and $x_4 = 1$. *Generally, method of finite differences approximates the solution in finite number of discrete points only.* The unknown function must afterward be interpolated between the calculated points.

Function value at $x_0 = 0$ is already given by the boundary condition. Applying forward differences and substituting them to the differential equation leads to:

$$\frac{u_1 - u_0}{0.25} + 2u_0 = 1$$
$$\frac{u_2 - u_1}{0.25} + 2u_1 = 1$$
$$\frac{u_3 - u_2}{0.25} + 2u_2 = 1$$
$$\frac{u_4 - u_3}{0.25} + 2u_3 = 1$$

This is a linear equation system. Taking into account $u_0 = 1$, its matrix form is:

$$\begin{pmatrix} 4 & 0 & 0 & 0 \\ -2 & 4 & 0 & 0 \\ 0 & -2 & 4 & 0 \\ 0 & 0 & -2 & 4 \end{pmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix} = \begin{bmatrix} 3 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

Its solution is

$$\left(u_1, u_2, u_3, u_4\right) = \left(\frac{3}{4}, \frac{5}{8}, \frac{9}{16}, \frac{17}{32}\right)$$

Non-linear differential equation, or a differential equation with coefficients depending on U and/or its derivatives, give rise to non-linear equation system on u. Its solution is more difficult but possible.

The method is not so simple if a derivative is to be calculated at the boundary of Ω because in that case usually points *outside the domain* are also to be applied.

Because of the boundary conditions, grid points should be located along the boundary. In case of two or more independent variables, the boundary can be a rather complex shape. If Ω is as simple as a rectangle or a circle then this is not a difficulty, but setting points in the boundary is not always so simple, and sometimes it is even impossible, in case of strangely shaped Ω domains. This difficulty comes together with the method of finite differences.

3.4.3 Weighted residues

In contrast to the finite differences, the method of weighted residues finds an approximate solution *all aver the domain*, not just in grid points. For this aim, the unknown function $U(\boldsymbol{x})$ is approximated as an expansion in the space of some (preferable complete and orthonormal) function series of so-called *basis functions* $\varphi_i(\boldsymbol{x})$ with unknown coefficients α_i:

$$U(\boldsymbol{x}) = \sum_{i=1}^{\infty} \alpha_i \varphi_i(\boldsymbol{x})$$

This expansion with proper coefficients α_i exactly satisfies the differenctial equation, thus the task is to find the coefficients.

In practive one stops after a finite steps of expansion, and an approximate solution u is looked for:

$$u(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_i \varphi_i(\boldsymbol{x})$$

Denote a differential equatio (system) generally as $\mathbf{D}\left[U(\boldsymbol{x})\right] = 0$ where $\mathbf{D}\left[.\right]$ is a so-called differential operator. The exact solution $U(x)$ satisfies the differential equation in all points of Ω, but $u(x)$ does not:

$$\mathbf{D}\left[\sum_{i=1}^{\infty} \alpha_i \varphi_i(\boldsymbol{x})\right] = 0$$

$$\mathbf{D}\left[\sum_{i=1}^{N} \alpha_i \varphi_i(\boldsymbol{x})\right] \neq 0$$

Method of weighted residues looks for such coefficients α_i that makes an average of the errors over Ω zero. For this aim one has to select N different *weitght functions* $w_j(\boldsymbol{x})$ $(j = 1, 2, \ldots, N)$ arbitrarily, and integrate over the domain with these weights:

$$\int\limits_{\Omega} \mathbf{D}\left[\sum_{i=1}^{N} \alpha_i \varphi_i(\boldsymbol{x})\right] w_j(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = 0 \qquad (j = 1, 2, \ldots, N)$$

This is a system of N ordinary quations with N unknown coefficients α_i.

Consider, for example:

$$U(x)\frac{\mathrm{d}^2 U}{\mathrm{d}x^2} + \left(\frac{\mathrm{d}U}{\mathrm{d}x}\right)^2 + \sin(x)\frac{\mathrm{d}^2 U}{\mathrm{d}x^2} = 0$$

and approximate the solution $U(x)$ of this non-linear ordinary differential equation with $u(x) = \sum\limits_{i}^{N} \alpha_i \varphi_i(x)$ where functions $\varphi_i(x)$ are known. Select some also known weight functions $w_j(x)$ $((j = 1, 2, \dots, N))$, express the average error with weight function j, and equal it to zero:

$$\sum_{i=1}^{N}\sum_{k=1}^{N} \alpha_i \alpha_k \int\limits_{\Omega} \varphi_i(x) \left(\frac{\mathrm{d}^2\varphi}{\mathrm{d}x^2} + \frac{\mathrm{d}\varphi}{\mathrm{d}x} \right) w_j(x)\mathrm{d}x +$$

$$+ \sum_{i=1}^{N} \alpha_i \int\limits_{\Omega} \sin(x)\varphi_i(x) w_j(x)\mathrm{d}x = 0$$

The integrals in this equation can be computed because all the functions in them are known. Thus a non-linear ordinary equation system is obtained with unknowns α_i.

In case of a **linear differential equation** the ordinaty equations obtained this way are linear in unknowns α_i, and the system is easy to solve. Denote a linear differential operator with $\mathbf{L}\,[.]$, the differential equation with $\mathbf{L}[U(x)] = p(x)$, where $p(x)$ is a known function. Then the 'average error is null' equation is:

$$\int\limits_{\Omega} \mathbf{L} \left[\sum_{i=1}^{N} \alpha_i \varphi_i(\boldsymbol{x}) \right] w_j(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int\limits_{\Omega} p(\boldsymbol{x}) w_j(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

Sequence of integration and summation can be swapped, and the coefficients lifted outside the integral, we obtain an equation system linear in α_i

$$\sum_{i=1}^{N} a_{j,i}\alpha_i = b_j \quad (j = 1, 2, \dots, N)$$

where

$$a_{j,i} = \int\limits_{\Omega} \mathbf{L}\left[\varphi_i(\boldsymbol{x})\right] w_j(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

$$b_j = \int\limits_{\Omega} p(\boldsymbol{x}) w_j(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

These integrals do not contain any unknown, they can be calculated in advance either analytically or numerically.

The weight functions are usually again selected as members of some series. Most known selections are the following three:

Method of point collocations This is the simplest method, and no integration is practically needed with it. One has to assign in the domain N different points $\boldsymbol{x}_j$, and select *Dirac* delta functions around these points as weight functions:

$$w_j(\boldsymbol{x}) = \delta(\boldsymbol{x} - \boldsymbol{x}_j)$$

(These functions can be defined several ways, one of them is limits of the normal distribution density functions with expectations $\boldsymbol{x}_j$ as their variance goes to 0.)

For any continuous function $f(\boldsymbol{x})$:

$$\int_{-\infty}^{\infty} f(x)\delta(\boldsymbol{x} - \boldsymbol{x}_j)\mathrm{d}\boldsymbol{x} = f(\boldsymbol{x_j})$$

therefore, instead of integration, one can simply substitute $\boldsymbol{x_j}$ to the function:

$$\int\limits_{\Omega} \mathbf{L}\left[u(\boldsymbol{x})\right]\delta(\boldsymbol{x} - \boldsymbol{x}_j)\mathrm{d}\boldsymbol{x} = \mathbf{L}\left[u(\boldsymbol{x}_j)\right]$$

Method of moments In this method the weight functions are

$$[w_j(\boldsymbol{x})] = x^{j-1}$$

i.e. the members of the series 1, x, x^2, x^3, ...

Method of Galerkin *Galerkin* suggested to select just the basis functions $\varphi_i(\boldsymbol{x})$ as weight functions:

$$w_j(\boldsymbol{x}) = \varphi_j(\boldsymbol{x})$$

Theoretical considerations show that this can be an efficient selection.

Example

For simplicity and transparency, here we solve the ordinary and linear equation

$$\frac{\mathrm{d}^2U}{\mathrm{d}x^2} + U = -x$$

over domain $\Omega = \{0 \le x \le 1\}$ with boundary conditions $U(0) = 0$ and $U(1) = 0$. Here, with our notation,

$$\mathbf{L}[U] = \frac{\mathrm{d}^2U}{\mathrm{d}x^2} + U$$

$$p(x) = -x$$

The exact solution is

$$U(x) = \frac{\sin(x)}{\sin(1)} - x$$

For numerical sulution, let the basis functions be

$$\varphi_i(x) = x^i - x^{i+1} \equiv x^i(1-x) \qquad (i = 1, 2, \dots) \tag{3.9}$$

because these satify the boundary conditions, and stop after two members:

$$\varphi_1(x) = x - x^2$$
$$\varphi_2(x) = x^2 - x^3$$
$$u(x) = \alpha_1\varphi_1(x) + \alpha_2\varphi_2(x)$$

Effect of the linear differential operator on the basis functions is:

$$\mathbf{L}\left[\varphi_1(x)\right] = -2 + x - x^2$$
$$\mathbf{L}\left[\varphi_2(x)\right] = 2 - 6x + x^2 - x^3$$

Solution with point collocation. We select arbitrarily two points. Let them be $x_1 = 0.25$ and $x_2 = 0.5$. With these values we get:

$a_{1,1} = -2 + 0.25 - 0.25^2$	$a_{2,1} = 2 - 6 * 0.25 + 0.25^2 - 0.25^3$
$a_{1,2} = -2 + 0.5 - 0.5^2$	$a_{2,2} = 2 - 6 * 0.5 + 0.5^2 - 0.5^3$
$b_1 = -0.25$	$b_2 = -0.5$

That is

$$\begin{pmatrix} -\frac{29}{16} & \frac{35}{64} \\ -\frac{7}{4} & -\frac{7}{8} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} -\frac{1}{4} \\ -\frac{1}{2} \end{pmatrix}$$

The solution in concise form is:

$$u(x) = \frac{x(1-x)(42+40x)}{217}$$

Solution with moments. The weight functions are:

$$w_1(x) = 1, \qquad w_2(x) = x$$

Elements of the linear system are:

$a_{1,1} = \int_0^1 \left(-2 + x - x^2\right) * 1 \,\mathrm{d}x$	$a_{2,1} = \int_0^1 \left(2 - 6x + x^2 - x^3\right) * 1 \,\mathrm{d}x$
$a_{1,2} = \int_0^1 \left(-2 + x - x^2\right) * x \,\mathrm{d}x$	$a_{2,2} = \int_0^1 \left(2 - 6x + x^2 - x^3\right) * x \,\mathrm{d}x$
$b_1 = \int_0^1 (-x) * 1 \,\mathrm{d}x$	$b_2 = \int_0^1 (-x) * x \,\mathrm{d}x$

The integrals can be computed analytically in this case. Results:

$$\begin{pmatrix} -\frac{11}{6} & -\frac{11}{12} \\ -\frac{11}{12} & -\frac{19}{20} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} -\frac{1}{2} \\ -\frac{1}{3} \end{pmatrix}$$

Solution in concise form:

$$u(x) = \frac{x(1-x)(122+110x)}{649}$$

Solution with method of Galerkin. Weight functions:

$$w_1(x) = x - x^2, \qquad w_2(x) = x^2 - x^3$$

Elements of the linear system:

$a_{1,1} = \int\limits_0^1 \left(-2 + x - x^2\right)\left(x - x^2\right)\ \mathrm{d}x$	$a_{2,1} = \int\limits_0^1 \left(2 - 6x + x^2 - x^3\right)\left(x - x^2\right)\ \mathrm{d}x$
$a_{1,2} = \int\limits_0^1 \left(-2 + x - x^2\right)\left(x^2 - x^3\right)\ \mathrm{d}x$	$a_{2,2} = \int\limits_0^1 \left(2 - 6x + x^2 - x^3\right)\left(x^2 - x^3\right)\ \mathrm{d}x$
$b_1 = \int\limits_0^1 (-x)\left(x - x^2\right)\ \mathrm{d}x$	$b_2 = \int\limits_0^1 (-x)\left(x^2 - x^3\right)\ \mathrm{d}x$

The integrals can be computed analytically in this case. Results:

$$\begin{pmatrix} -\frac{3}{10} & -\frac{3}{20} \\ -\frac{3}{20} & -\frac{13}{105} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} -\frac{1}{12} \\ -\frac{1}{20} \end{pmatrix}$$

Solution in concise form:

$$u(x) = \frac{x(1-x)(71+63x)}{369}$$

Comparison of the results. The results calculated in points $x = 0.25$, $x = 0.5$, and $x = 0.75$ are compared:

	0.25	0.5	0.75
Exact U	0.044014	0.06975	0.06006
Collocation	0.044931	0.07143	0.06221
Moments	0.043191	0.06818	0.05908
Galerkin	0.044080	0.06944	0.06009

Dealing with the boundary conditions

There are two different techniques for dealing with boundary conditions.

Proper basis functions. According to this technique, one has to select basis functions that all satisfy the boundary conditions. This one we applied in the above example. Unfortunately, finding proper basis functions is not an easy task, and is not even always possible.

Involving the boundary conditions in the equation system. This technique looks for such coefficients α_i that not only average error of the main differential equations $\mathbf{D}[U(\boldsymbol{x})] = 0$ makes zero over Ω but the average error of that one **plus** those of the boundary conditions $\mathbf{P}_1[U(\boldsymbol{x})] = 0$, $\mathbf{P}_2[U(\boldsymbol{x})] = 0$ etc. over the boundary $\delta\Omega$:

$$\int\limits_{\Omega} \mathbf{D}\left[\sum_{i=1}^{N} \alpha_i \varphi_i(\boldsymbol{x})\right] w_j(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \sum_k \int\limits_{\delta\Omega} \mathbf{P}_k\left[\sum_{i=1}^{N} \alpha_i \varphi_i(\boldsymbol{x})\right] w_j(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 0$$

$$(j = 1, 2, \ldots, N)$$

Here any basis functions can be selected but even the boundary conditions will not exactly satisfied.

3.4.4 Finite elements

Method of finite elements is a version of weighted residues, especially with Galerkin, but applied on such problems that are defined over highly involved domains and material properties. For example, heat loss and temperature distribution in a reactor under pressure is to be computed point by point. The vessel stands on three legs down from an oval bottom. The cover is again an oval body, fixed to the vessel with flange and bolts. There is a stirrer of anchor type inside the vessel. There are tubes comming from the top down into the liquid in the vessel. There are even more sensors, a window, insulation, etc. What kind of basis function can be applied to model the whole body? How can we take into account of the phases, different materials inside the domain?

For this aim, we subdivide domain Ω to many small, and regularly shaped, subdomains interfacing at the borders of the subdomains only: sides, edges, and vertices. These subdomains are called finite elements.

One basis function is constructed to each vertex in such a way that its value is 1 in that vertex, is zero at the other vertices of the finite elements neighboring this vertex, and is zero over any other finite element.

The integrals occuring in the method of weighted residues can be decomposed to sub-integrals over the finite elements. Since each basis function is non-zero over a very few elements only (2 in one-dimension, 4 in 2 dimensions, 8 in 3 dimensions), one does not need integrate over the whole Ω domain. Even integral of a single basis function can be decomposed as a sum of integrals over its few elements.

Moreover, since these functions have short range, low degree polinomials can be used as basis functions. This low degree polinomial over a single finite element is called *shape function*. Over each finite element there are as many shape functions as the number of vertices of the element. Each has value 1 in one of the vertices and 0 in others; each has 1 in a different vertex of that element.

How it is done is illustrated on a one-dimension case. Here an interval is subdivided into smaller intervals, and we apply (for simplicity) linear shape functions, according to **Fig. 3.9**.

Shape functions generally are curved functions, and are selected to satisfy some smoothness conditions. Linear element provide continuity but they are not differentiable at the boundaries of the elements. Second degree functions can be applied to constuate solutions that are differentiable at least once, etc.

This method usually results in a linear equation system of gigantic size, but its coefficient matrix is sparse, and usually of special pattern that make the solution easier.

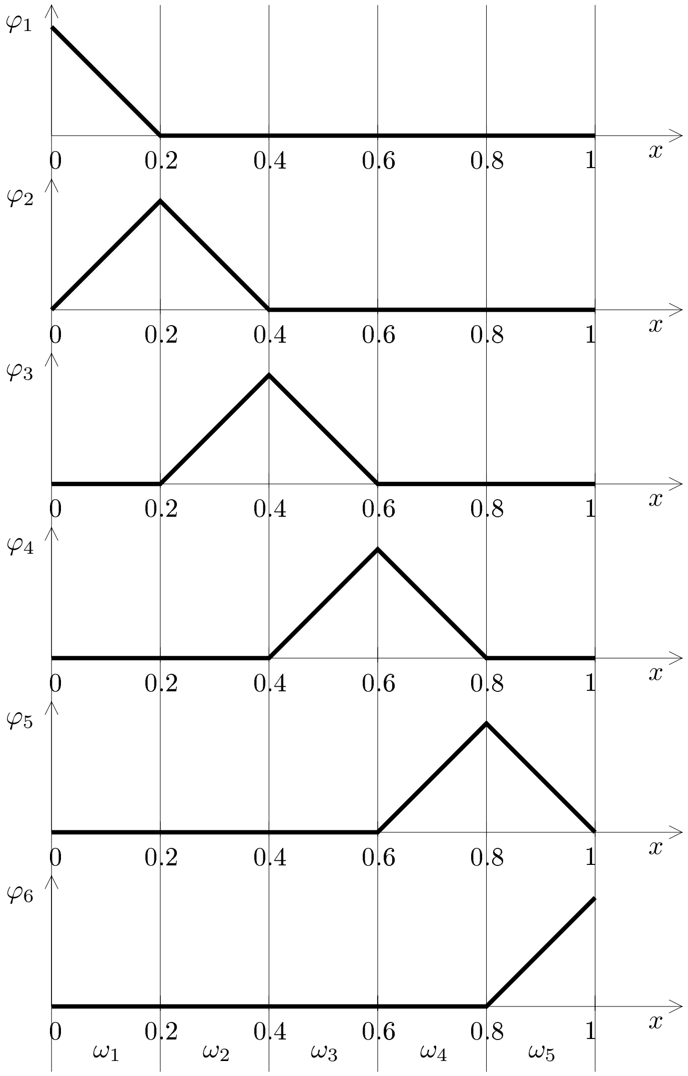

Figure 3.9: Finite elements

3.5 Suggested literature

KORN, R. C. ÉS KORN, J. M.: *Mathematical Handbook for Scientists and Engineers.* McGraw-Hill, New York, 1961.

PRESS, W.H, ET AL.: *Numerical Recipes in C / Pascal / Fortran. (The Art of Scientific Computing).* Cambridge University Press, Cambridge, 1988.

LEDERMANN, W. (ED.): *Handbook of Applicable Mathematics. Volume III: Numerical Methods (ed.: Churchhouse, R.F.).* John Wiley & Sons, New York, 1981.

CARNAHAN, B.; LUTHER, H.A.; WILKES, J.O.: *Applied Numerical Methods.* John Wiley & Sons, New York, 1969.