Process Engineering

II.

Process synthesis

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

Adapted from Hungarian by the author in February, 2013

Contents

1 Process synthesis in general **3**
1.1 Hierahical process design . . . 3
1.2 Questions . . . 5
1.3 Suggested literature . . . 5

2 Energy recovery networks **8**
2.1 Heat exchanger networks (HENs) . . . 8
2.2 Energy targeting . . . 11
2.2.1 Composite curves . . . 12
2.2.2 Composite Curves (CC)–diagram and pinch . . . 13
2.2.3 Unshifted and shifted CC . . . 14
2.2.4 Grand Composite Curve (GCC) . . . 16
2.2.5 Heat cascading . . . 18
2.2.6 General pinch rules . . . 20
2.3 Heat Echanger Network Synthesis (HENS) . . . 22
2.3.1 Fast algorithm heuristic . . . 22
2.3.2 Pinch design . . . 24
2.3.3 Hidden pinch . . . 27
2.4 Area targeting and supertargeting . . . 28
2.4.1 Area targeting . . . 28
2.4.2 Supertargeting . . . 33
2.5 GCC and thermal integration . . . 37
2.5.1 Utility levels . . . 37
2.5.2 Thermal processes along the heat cascade . . . 39
2.5.3 Heat pumps along the heat cascade . . . 43
2.6 Questions . . . 44
2.7 Suggested literature . . . 44

3 Component recovery networks **45**
3.1 Mass Exchane Networks (MENs) . . . 45
3.2 Water use and re-use networks . . . 51
3.2.1 Water profiles . . . 51
3.2.2 Limit Composite Curve (LCC) and pinch . . . 52

3.2.3 Assignment of water using networks . . . 53
3.2.4 Regeneration and re-use . . . 54
3.2.5 Regeneration and recycling . . . 57
3.3 Distributed water treatment . . . 58
3.3.1 Water treatment systems . . . 58
3.3.2 Lumped treatment . . . 60
3.3.3 Distributed treatment . . . 62
3.4 Several contaminating components . . . 63
3.5 Questions . . . 63
3.6 Suggested literature . . . 64

4 Rectification systems without azeotropes 65
4.1 Separation sequencing . . . 65
4.1.1 Heuristics and load factors . . . 67
4.1.2 Sequencing with dynamic programming . . . 68
4.2 Underwood equations . . . 70
4.3 Reversible distillation . . . 74
4.3.1 Thermodynamic efficiency of distillation . . . 74
4.3.2 Reversible distillation of binary mixtures . . . 76
4.3.3 Reversible distillation of multicomponent mixtures . . . 77
4.4 Energy saving distillation systems . . . 78
4.4.1 More feeds and/or more product streams . . . 78
4.4.2 Stepwise heat turnover . . . 79
4.4.3 Thermal coupling . . . 81
4.4.4 Energy integrated columns . . . 82
4.4.5 Heat pump distillation . . . 84
4.5 Questions . . . 87
4.6 Suggested literature . . . 87

5 Optimization 88
5.1 Local minima over unconstrained domain . . . 89
5.2 Equality constraints . . . 95
5.3 Inequality constraints . . . 95
5.4 Dynamic programming . . . 98
5.5 Global optima . . . 100
5.6 Integer variables . . . 101
5.7 Questions . . . 107
5.8 Suggested literature . . . 107

Chapter 1

Process synthesis in general

Process synthesis is a complex activity aiming at either working out the concepts and details of a new process or a new plant, or modifying a part of an already existing process or working plant. Process design is always an optimization task, with *economic optimum* final aim, but competing targets usually make constructing a clear objective difficult. Many time partial targets are only considered when partial processes are designed. Such partial objectives are e.g. costs, capacity, quality, environment etc.

In mathematical sense, process flowsheeting is an optimization problem with both continuous and non-continuous variables, and involved constraints of both equality and inequality types. Solving this problem exactly in practice is not always possible, en even if it is, it might be to expensive, e.g. it would last to years and decades even with using the best supercomputers. Therefore such problems are often solven with heuristics, and in successive step, in an evolutive manner.

1.1 Hierahical process design

Process design itself is expensive both in expenses and time. Therefore it is made in a hierarchical manner, step by step. 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.

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:

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

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.

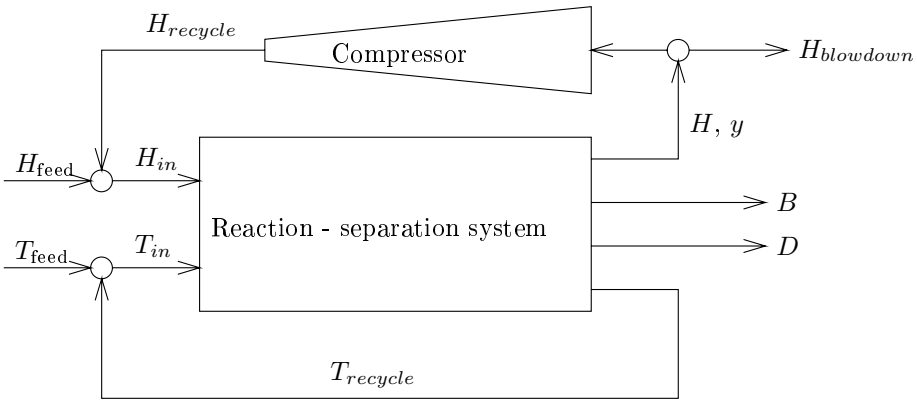


Figure 1.1: First idea

A first idea is shown in **Fig. 1.1**. 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. 1.2**. Another one, with diphenyl recycling and re-reacting is shown in **Fig. 1.3**. Final verion wit energy recovery is shown in **Fig. 1.3**.

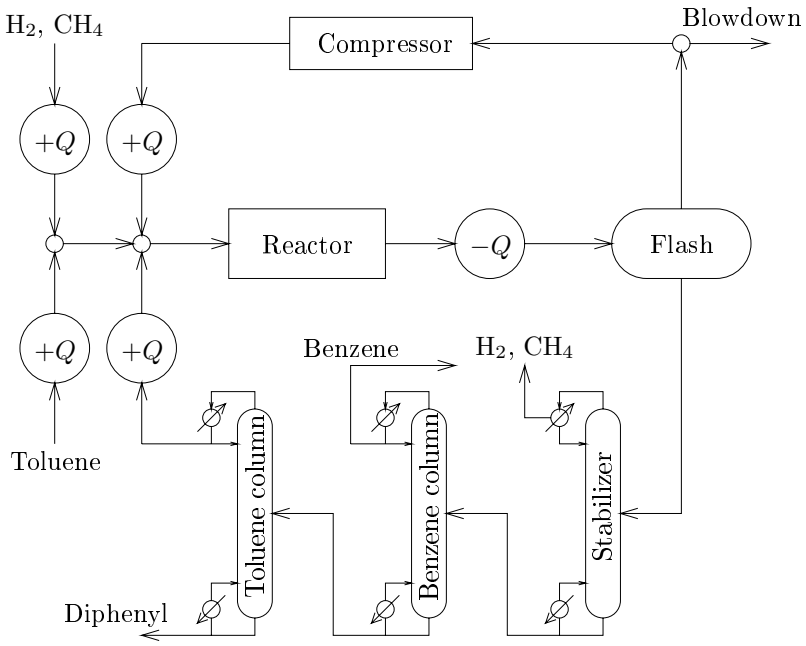


Figure 1.2: First version

1.2 Questions

1. Explain hierarchy of process synthesis! List the levels. Which are the tasks of chemical engineer?

1.3 Suggested literature

Peters, M. S., Timmerhaus, K. D.: *Plant Design and Economics for Chemical Engineers*. McGraw-Hill, 1969.
Douglas, J. M.: *Conceptual Design of Chemical Processes*. McGraw-Hill, 1988.

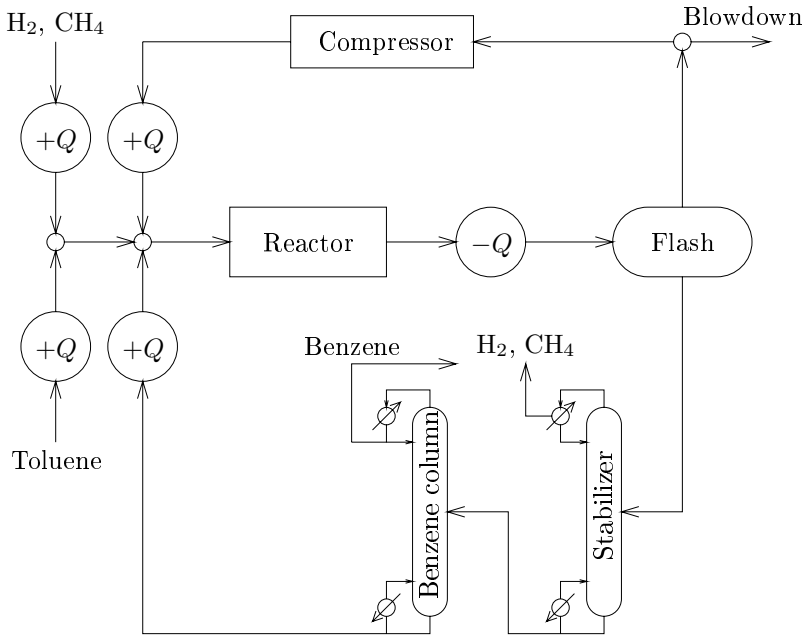


Figure 1.3: Another version

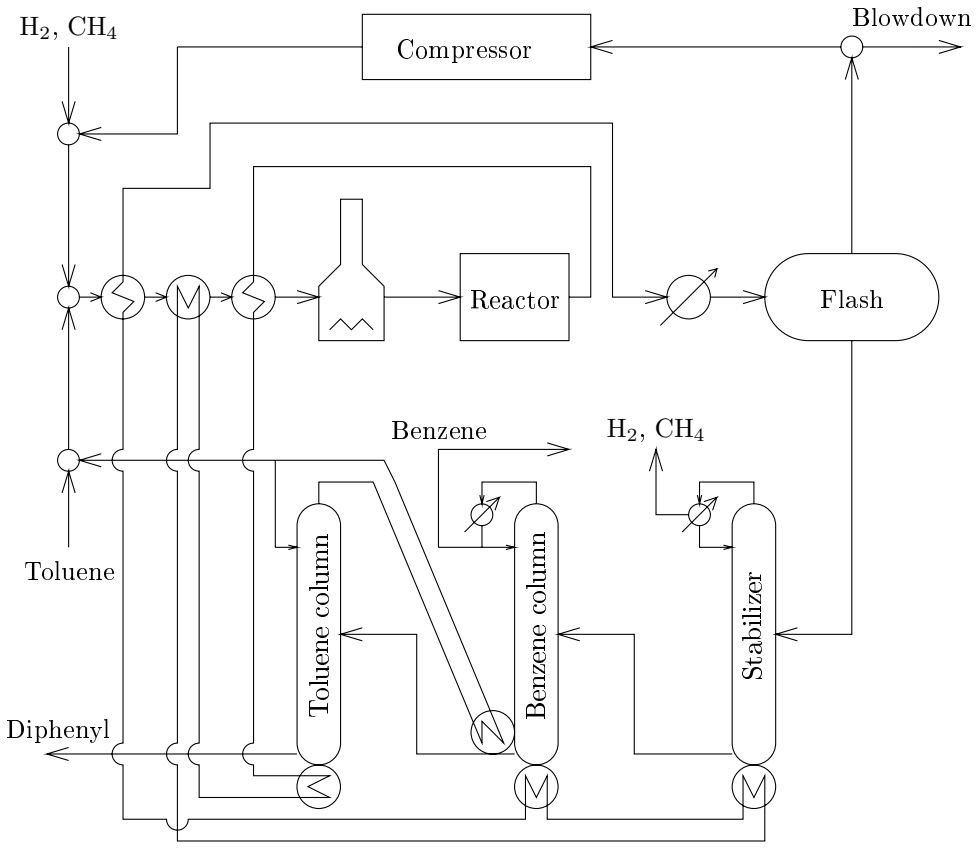


Figure 1.4: Final version

Chapter 2

Energy recovery networks

2.1 Heat exchanger networks (HENs)

Synthesis of heat exchanger networks is the problem that can be formulated the simplest way. Even within this problem family, the simplest case is one without phase change in the heat exchangers, at least in their process stream side. The problem is formulated as follows.

1. Number of cold streams (to be warmed up): N
2. Number of hot streams (to be cooled down): M
3. Cold streams' data: $(i = 1, 2, \dots N)$
   (a) Start temperature: t_i°
   (b) Target temperature: t_i^*
   (c) Heat capacity flow rate: c_i
4. Hot streams' data: $(j = 1, 2, \dots M)$
   (a) Start temperature: T_j°
   (b) Target temperature: T_j^*
   (c) Heat capacity flow rate: C_i

   Heat capacity flow rate, or shortly *heat capacity* is product of specific heat and flow rate [energy/(temperature×time)]. Demanded heating power or cooling power belonging to process streams can be calculated as $(t_i^* - t_i^\circ) \times c_i$, $(T_j^\circ - T_j^*) \times C_j$.
5. Thermal data of utilities and limit of their degradation. For example: specific capacity, temperature, and maximum temperature rise of cooling water, pressure, temperature, and condensation heat of heating steam.

6. Thermal and rheological properties of process streams and utilities (viscosity, density, heat conductivity, all in function of temperature, fouling potential)
7. Type(s) of heat exchangers (including also heaters and coolers), their essential properties (wall thickness, heat conductivity, etc.)
8. Cost of heat exchangers in function of heat transfer area
9. Specific costs of utilities (e.g. \$/m^3 cooling water).
10. Depreciation ratio of expected lifetime

Assigning an optimal HEN is a very diffucult task even in this case, therefore the following simplifying assumptions are usually applied:

1. Heat capacities are independent of temperature (i.e. constant) (or at least constant sectionwise).
2. Heat exchangers are modelled as countercurrent units, with the following general relation:
$$Q = UA\Delta T_{\ln}$$
$$\Delta t_{\ln} = \frac{\Delta T_1 - \Delta T_2}{\ln \frac{\Delta T_1}{\Delta T_2}}$$
where ΔT_1 and ΔT_2 are approach temperatures at the ends of the heat exchanger: hot inlet minus cold outlet, and hot outlet minus cold inlet, for computing logarithmic mean of approach temperatures.
3. Instead of detailed modelling the overall heat transfer coefficient U, either
   (a) an average constant U is applied for modelling all the exchangers in the network, or
   (b) it is computed as
   $$\frac{1}{U} = \frac{1}{\alpha_1} + \frac{1}{\alpha_2}$$
   where each individual stream i is charaterized by a constant heat transport coefficient α_i that incorporates also the wall resistance and the effects of fouling.
4. Costs of heat exchangers are approximated as
$$K = c + aA^b$$
where A is heat transfer area, a, b, and c are constants. Generally $c = 0$, and exponent b is a number smaller than 1.

Network structure has a decisive effect on economics (costs) of HEN. The following variants have to taken into account:

Matchings. There are the simplest alternatives, illustrated in **Figures 2.1 a./ and b./**. Streams A and B are either coupled, respectively, with C and D, or with D and C.

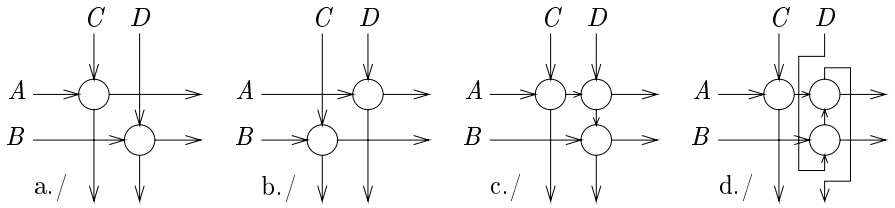


Figure 2.1: Matchings

Match sequences. A stream can match several other streams, and thus it can match them in different sequences along its path. Such alternatives are shown in **Figures 2.1 c./ and d./**. Stream D matches streams A and B in different sequences.

Multiple matches. The same pair of cold and hot streams can match several times in a networks. There is always at least another match bracketed between them, otherwise two neighboring matches could be lumped. Different sequences can also be constructed. Such a simple case is shown in **Fig. 2.2**.

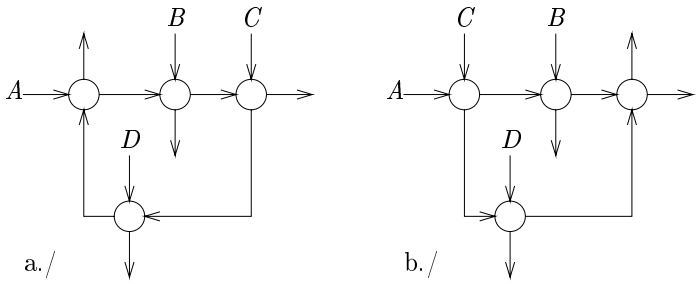


Figure 2.2: Matches and sequences

Branching Streams can be branched; branching results in *branches* of the same stream with smaller capacities. Branches can be matched with different streams, can be further branched and / or unified.

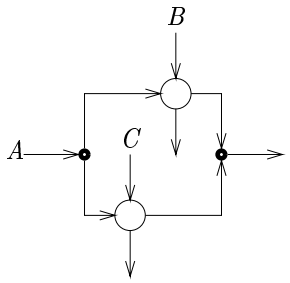


Figure 2.3: Network with a branching

Number of possible network structures is *underestimated* by counting only those structures conceivable of given N cold and M hot streams wihout considering multiple matching or branching, and only one cooling stream and only one heating stream considered. The minimum number of matches approximately is

$$MIN = (N + 1) + (M + 1) - 1$$

the number of possible matches is

$$MAX = (N + 1)(M + 1) - 1$$

because no match is designed between the utilities. The number of possible combinations is

$$\binom{MAX}{MIN} = \frac{MAX!}{MIN!(MAX - MIN)!}$$

This is an astronomical number even with small N and M. Considering multiple matches makes the number its powers, and the possibility of branching makes the number of possible structures practically infinite. From these networks should an optimal one be selected.

2.2 Energy targeting

According to experience, costs of HENs with maximum energy recovery, i.e. minimum utility costs, are much nearer the optimum costs then those with minimum recovery and applying extensive external heating and cooling. Therefore synthesizing an optimal HEN is suggested to start with constructing one with maximum recovery. Even apart from this idea, knowledge on maximum possible energy recovery and, what is the same, minimum external heating and cooling demand, is useful for pre-estimating the expected costs of the HEN even before design.

The other extreme, i.e. heating and cooling with zero recovery, is easy to calculate:

$$C_0 = \sum_{j=1}^{M} (T_j^\circ - T_j^*) C_j$$

$$H_0 = \sum_{i=1}^{N} (t_i^* - t_i^\circ) c_i$$

In a rather lucky case maximum recovery would be the minimum of the two figures. If $H_0 < C_0$ then maximum recovery would be H_0, minimum cooling would be $C_0 - H_0$, and heating would not be needed. In the opposite case we would get respective results wich changing the roles of C and H.

Maximum recovery is generally smaller than above because the given temperatures constinute constraints on heat transfer between hot and cold streams. These constraints are taken into account by *pinch technology* to determine the actual maximum.

2.2.1 Composite curves

Heat power need of cold streams can be visualized in $t-E$ diagram (t: temperature, E: energy). Such a diagram is shown in **Fig. 2.4**. Start and target temperatures t_i° and t_i^* are taken in axis t. Heating power need is denoted by Q in axis E. Directed straight line represents a stream, here a line connecting points A and B.

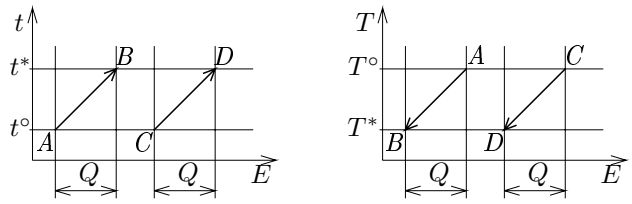


Figure 2.4: Heat curves

The line AB is straight just because heat capacity of the stream is constant. Its slope is proportional to the reciprocal of heat capacity.

Since energy is a quantity related to reference point, and not E but Q is important here, the lines can be shifted to the right or left, along axis E. Line CD represents the same stream, or a stream with the same data, as line AB in **Fig. 2.4**.

Hot stream can be represented the same way. Line CD represents the same hot stream, or a stream with the same data, as line AB in **Fig. 2.4**.

Total heating need of all N cold streams can be represented with a *composite curve*. Reciprocal of its slope is proportional to the sum of the heat capacities of all the cold stream at the actual temperature.

How this can be constructed graphically is shown in **Fig. 2.5**. Two cold streams are present in **Fig. 2.5 a.**/, and they overlap in temperature interval $[t_C, t_B]$.

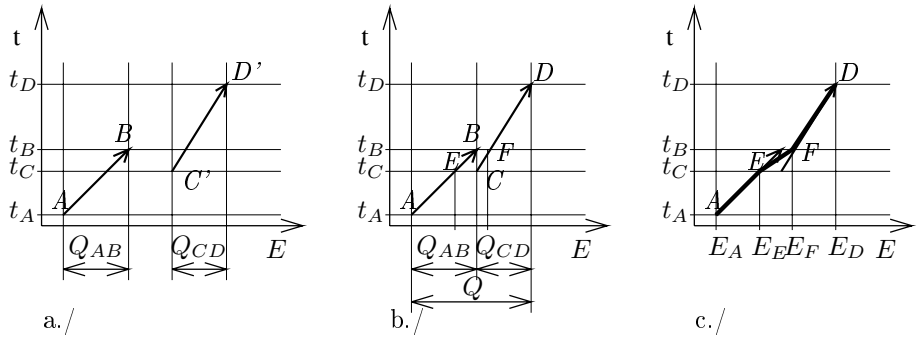


Figure 2.5: Construction of composite curve

First the two streams should be shifted along axis E so that their projected sections to axis E just touch, have one common point, and they form a contiguos interval. This is shown in **Fig. 2.5 b.**/. Length of the section $E_A - E_D$ is just the sum of the individual lengths. Then temperature intervals and energy intervals along which the composite line has constant slope are assigned. These are determined by the characteristic temperatures. This way one obtains points E and F in **Fig. 2.5 b.**/, and the respective projections E_E and E_F in **Fig. 2.5 c.**/.

Third, the straight line belonging to each temperature interval is drawn. This is easy, given the covered energy interval with length Q, and the two end-point temperatures. This is line AE in interval $[t_A, t_C]$, and line FD in interval $[t_B, t_D]$, and a line EF is drawn in the overlapping interval $[t_C, t_B]$. After discarding the original lines, the composite line remains $AEFD$.

2.2.2 Composite Curves (CC)–diagram and pinch

Heat exchange between individual hot and cold streams or their temperature sections can also be represented in t–E diagram. COuntercurrent heat exchange is represented by two heat lines in **Fig. 2.6 a.**/. The covered E sections coincide, and the temperatures of the hot stream is everywhere over that of the cold stream.

Infeasible heat exchange is shown in **Fig. 2.6 b.**/. Heat balance is O.K., but the hot stream is not everywhere warmer than the cold stream.

The same principle can be applied to composite curves, as well. Line sections over each other represent heat exchange, and the Second Low of Thermodynamics must be satisfied.

E-overlapping section BC of hot composite curve CA and cold composite curve BD represents energy recovery. Uncovered section CD represents remaining heat-

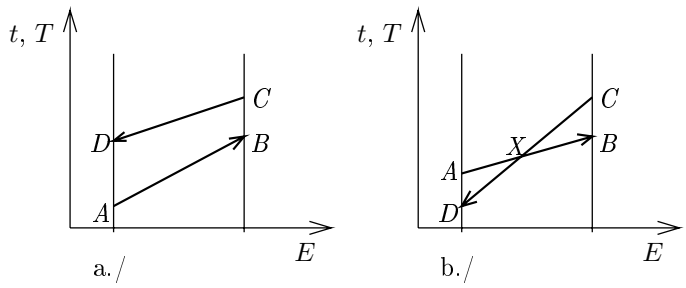


Figure 2.6: Lines of feasible and infeasible heat exchange assignments

ing demand of the cold stream, whereas section BA represents remaining cooling demands of hot streams. Such a diagram, with or without projection lines, is called *Composite Curve Diagram*, abbreviated as *CC Diagram*.

By shifting the two compiste lines toward each other, the overlapping section becomes longer, and uncovered sections shorter, as in (Fig. **2.7 b.**/), involving increased energy recovery and decreased external utility consumption. What kind of HEN would do it is not yet known but such a situation can be achieved somehow because the conditions are satified.

By shifting the curves toward again, they will intersect each other as shown in **Fig. 2.7 d.**/. Hot streams become under cold streams in section XY, so that such a situation is infeasible.

Maximum energy recovery and, in the same time, minimum utility consumption is assigned be a situation that the two curves just touch but do not intersect. Such a *pinch* is shown in **Fig. 2.7 c.**/. The touchpoint is called *Pinch Point*, here labelled with P, the respective temperature is *pinch temperature*. Maximum energy recovery Q_{max}, minimum heating H_{min}, and minimum cooling C_{min} can be read from the figure.

2.2.3 Unshifted and shifted CC

Energy recovery of **Fig. 2.7 c.**/ is infeasible because positive approach temperature is needed to cause heat transport. Any arbitrarily small approach temperature cannot be applied because of resistance that cannot be taken into account exactly. In practice a *minimum approach temperature*, MAT, is (arbitrarily) specified.

The CC diagram in pinch situation is then shown in **Fig. 2.8 a.**/. Instead of a pinch point, there is pinch pair of points, and a pair of pinch temperatures: hot pinch temperature and cold pinch temperature.

For technical reasons, it is usual to shift the curves *vertically*, along axis T, together by MAT so that when they apparently pinch then the real approach temperature is just MAT. Such a situation is shown in **Fig. 2.8 b.**/.

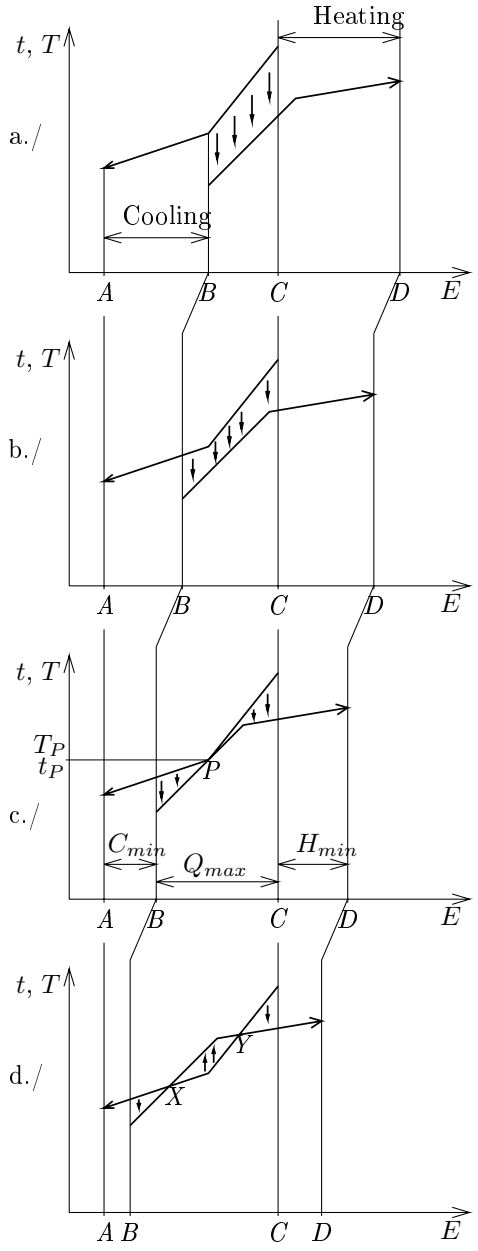


Figure 2.7: CC-diagram and pinch

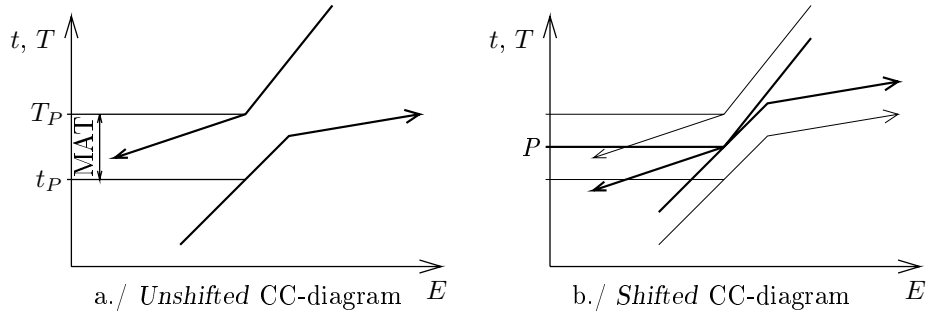


Figure 2.8: Unshifted and shifted CC

2.2.4 Grand Composite Curve (GCC)

Grand Composite Curve (or *GCC*) is applied in more general considerations, discussed later. Here just its definition and basic properties are introduced.

GCC is a t–Q diagram equivalent to and derived from a *shifted* CC. To each t, difference between the respective E coordinates of cold and hot composite curves are measured on axis Q. One could say that the actual hot $\Rightarrow$ cold turnover is calculated at eact temperature.

In order to explain *heat cascading* introduced in the next section, we illustrate GCC in a series of plots staring with a siyuation where the hottest points of both hot and cold composite curves are fitted to a common E coordinate in *shifted* CC, see **Fig. 2.9 a./-CC**. Such a situation means just zero heating demand. If we are lucky, this is a feasible situation. If not, as in our case because the CC curves intersect, the GCC curve has sections with negative Q. Point A is at $Q = 0$ because of coincidence in E in the CC diagram. Intersection points, too, in CC correspond to $Q = 0$ points in GCC. Negative GCC sections mean heat trasport from lower temperature to higher, this is infeasible.

Figures 2.9 c./–CC and c./–GCC show feasible energy recovery. Shape of GCC is not changed, the line is shifted to the right with the amount the two CC lines are shifted apart with. The actual heating and cooling can be read at points A and D as distance from axis t.

Figures 2.9 d./–CC and d./–GCC show maximum energy recovery and minimum external heat turnover. *Pinch point* of CC is transformed to a touch to the t axis in GCC, here point B.

By shifting GCC to the right from the pinch situation, one gets extra heating and cooling. The two extra's are equal because of the heat balance.

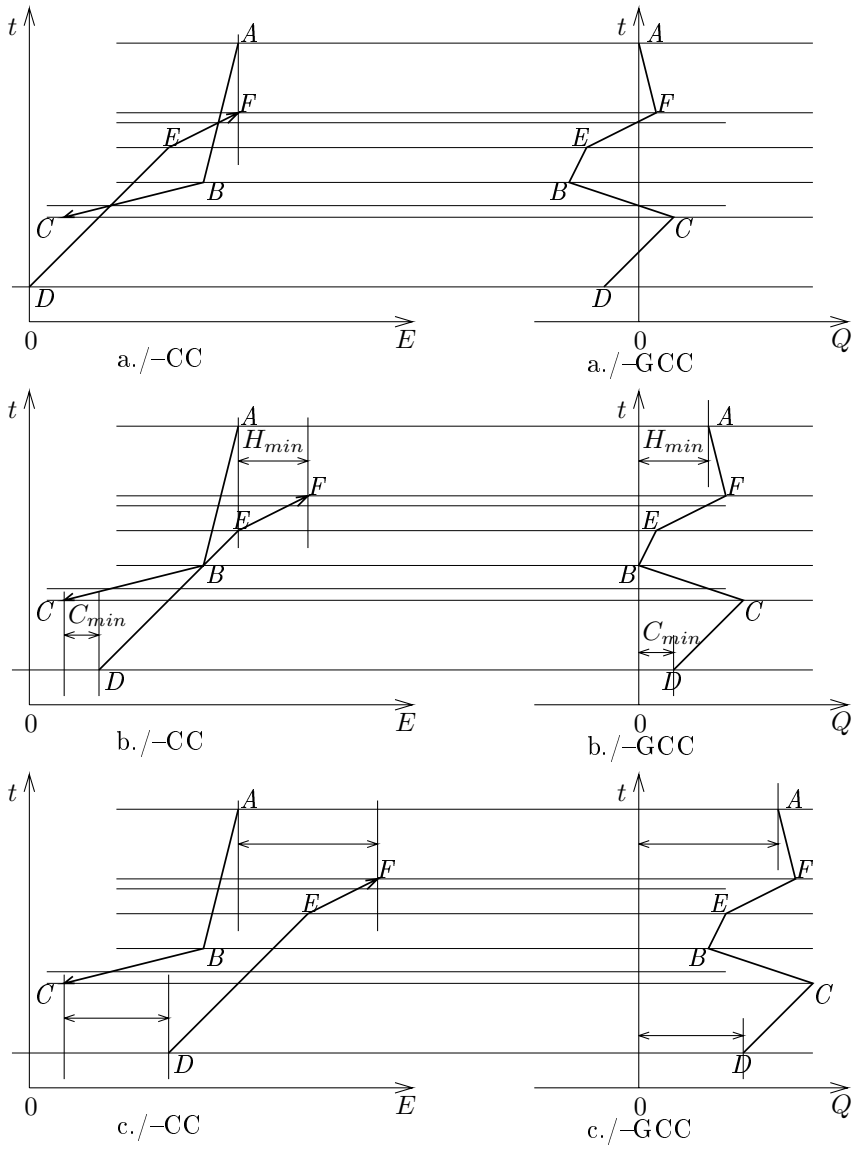


Figure 2.9: GCC-diagram and pinch

Table 2.1: Example data

Stream	Capacity kW/C	Start temperature C	Target temperature C
A	2	150	60
B	8	90	60
C	2.5	20	125
D	3	25	100
MAT=20 C			

Table 2.2: Example data after shifting

Stream	Capacity kW/C	Start temperature C	Target temperature C
A	2	140	50
B	8	80	50
C	2.5	30	135
D	3	35	110
MAT=0 C			

2.2.5 Heat cascading

Heat cascading is determination of pinch point and minimum external heat turnover numerically. It is demonstrated by solving an example given in (**Table 2.1**).

Shifted data are calculated first. We do it arbitrarily by shifting down hot streamw with half a MAT, i.e. 10 C, and shihting up cold stream with the same half MAT, resulting in the data shown in **Table 2.2**.

Characteristic temperatures are determined, i.e. at breaking points of the CC curves. These are, in decreasing order: 140, 135, 110, 80, 50, 35, 30. Then we know which streams are present in each interval, and we can calculate heat supply and heating demand in each interval, and finally a net heat supply of that interva. How it is down is explained in **Fig. 2.10** and **Fig. 2.11**.

Fig. 2.10 is a so-called *heat content diagram*, here the area of a rectangle is proportional to the streams heat supply or heating demand. Any hot stream area can be matched with some cold stream area of the same size if the temperature difference is not negative. If shifted temperatures are used then an apparently zero difference corresponds MAT, so that the exchange is all right. This relation is satisfied along any horizontal line, therefore maximum recovery between a hot stream and a cold stream in the same temperature interval can be calculated by taking their minimum, and the difference will be the remaining supply or demand. Demand will be accounted for a negative supply.

Temperature range of streams and their capacity are visualized simpler way in **Fig. 2.11**. Net supply is calculated by (1) summing up capacities (taking negative

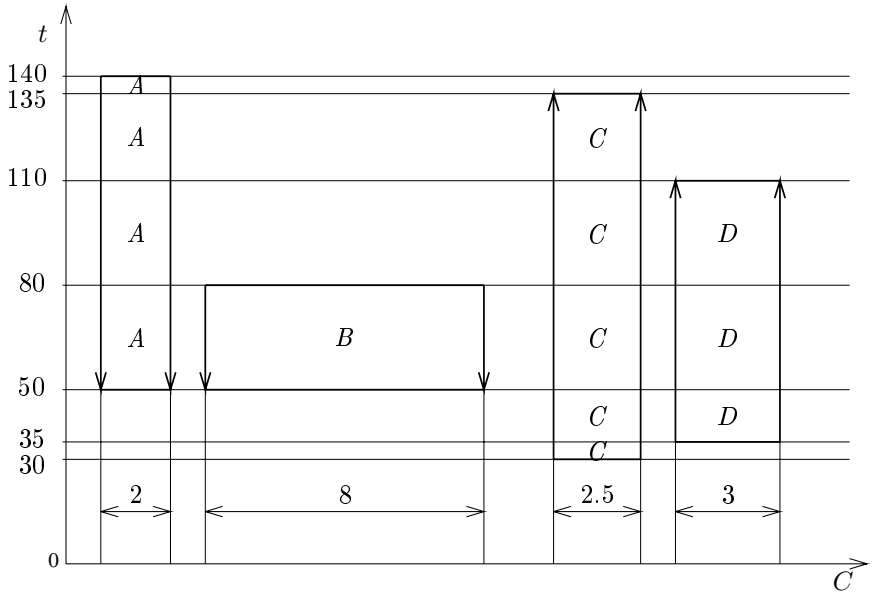


Figure 2.10: Heat content diagram

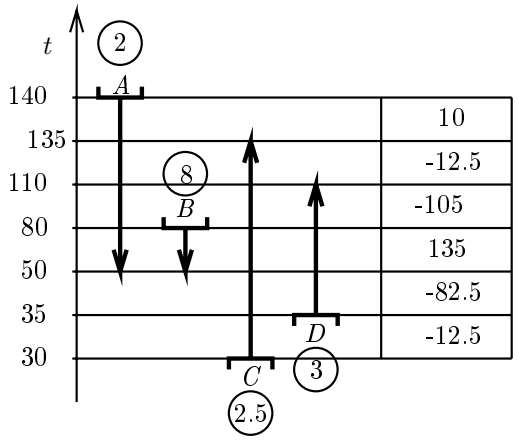


Figure 2.11: Heat cascading – 1.

for cold streams), and (2) multiplying this figure with the lenght of the interval. The so calculated algebraic sums are shown in the right hand side of the figure.

Once these remaining net supplies are known, the heat cascade can be computed along the ideas followed at constructing GCC. First assume no external heating (see also **Fig. 2.9. a.**/). Since maximum recovery is already taken into account inside intervals, one has to count recovery between intervals only. Heat flows downwards like water in stages.

Such a calculation is shown in **Fig. 2.12. a.**/. Each box contains its net supply as computed in **Fig. 2.11**. If it is negative then the interval consumes that amount from the heat flowing down to it from above, and emits the remaining supply only. If it is not satisfied then a negativ emission is calculated. We get inter-interval heat transports as 0, 10, -2.5, -107.5, 27.5, -55, -67.5 kW. The first value 0 kW stand for no heating, the last value -67.5 kW says so much cooling. The so obtained numbers correspond to the Q-values of the GCC.

Unfortunately, there are negative numbers in this array. These negative numbers mean heat transport from below, which is infeasible. GCC must be shifted to the right so that the numbers become non-negative. How much heating is needed?

Just the same amount as much negative number appears in the array, because any X heating will flow downwords the cascade, and inrease all numbers with X (heat balance). The minimum value is -107.5 kW, therefore the minimum heating is 107.5 kW. The resulted cascade is shown in **Fig. 2.12. b.**/. Minimum heating is 107.5 kW, minimum cooling is 40 kW. There is a temperature at which the transport is 0, this is pinch temperature: here 80 °C but, since we applied shifted data, the pinch temperatures are 90 °C for hot streams, and 70 °C for cold streams.

2.2.6 General pinch rules

Both heat cascade and GCC show that minimum external heat turnover comes with no heat transfer through pinch temperature. More can also be said. There are three gereral rules characterizing maximum recovery and minimum heat turnover:

1. **No heat transfer through pinch**
2. **No heating above pinch**
3. **No cooling below pinch**

Note: There are two pich temperature in case of positive MAT. Rule 1 means that no heat exchange is done between any hot stream setcion above hot pinch and any cold stream section below cold pinch. Rule 2 excludes any external heating to hot stream above hot pinch or cold stream above cold pinch. Rule 3 excludes any external cooling of hot stream below hot pinch or cold stream below cold pinch. The rules are explaind with **Fig. 2.13**.

Arrangement corresponding to the pinched GCC or cascade is shown in **Fig. 2.13. a.**/. Applying heat transport X through pinch, as is shown in **Fig. 2.13. b.**/, involves the same amount of extra heating and cooling. Applying external

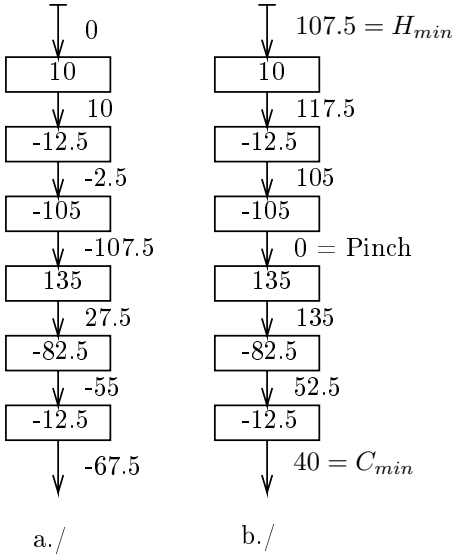


Figure 2.12: Heat cascading – 2.

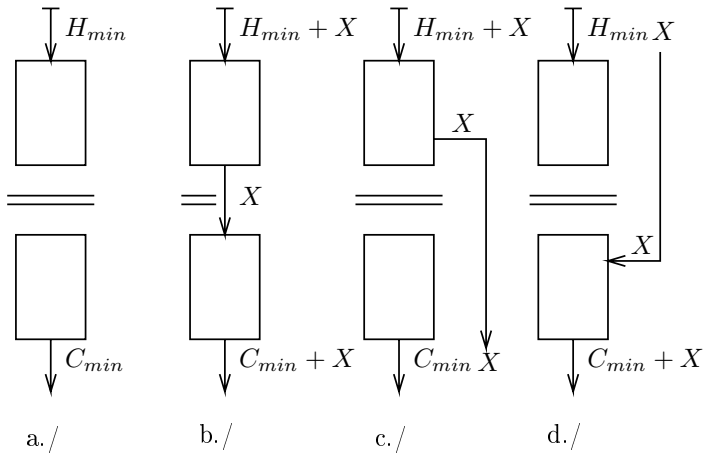


Figure 2.13: General *pinch* rules

cooling X above pinch, shown in **Fig. 2.13. c.**/, involves the same amount if extra heating, and the external cooling is also an extra to minimum cooling. The symmetric case is shown in **Fig. 2.13. d.**/ with external heating below pinch.
Warning Pinch rules are conditions for minimum heating and cooling, but they may be violated if maximum energy recovery is not the objective. For example, minimum cost may belong to smaller energy recovery.

2.3 Heat Echanger Network Synthesis (HENS)

2.3.1 Fast algorithm heuristic

Conventional HENS applied the principle of countercurrenct arrangement. This can be formulated as *Fast algorithm* below.

1. Select hot stream $j^{\square}$ with the highest starting temperature $T^{\square}$, and denote its target temperature with $T^{\square *}$.
2. Select cold stream $i^{\square}$ with the highest target temperature $t^{\square}$, and denote its starting temperature with $t^{\square \circ}$.
3. If $T^{\square} < t^{\square} + MAT$ and $C_j \geq c_i$ then heat up cold stream $i^{\square}$ from $T^{\square} - MAT$ to $t^{\square}$, and return to step 1.
4. If $T^{\square} \leq t^{\square} + MAT$ and $C_j > c_i$ then heat up cold stream $i^{\square}$ from $T^{\square} - (MAT + \Delta)$ to $t^{\square}$, and return to step 1. (Δ is any positive value assigned heuristically and optimized later.)
5. Design a match between $j^{\square}$ and $i^{\square}$ with hot end temperatures $T^{\square}$ and $t^{\square}$, and maximum feasible heat load. Denote cold end temperatures with T_{out} and t_{in}.
6. If $T_{out} > T^{\square *}$ then let the starting temperature of the remaining section of hot stream j be $T_j^{\circ} = T_{out}$. Otherwise hot stream j does not take place in the remaining problem.
7. If $t_{in} > t^{\square \circ}$ then let the target temperature of the remaining section of cold stream i be $t_i^* = t_{in}$. Otherwise cold stream i does not take place in the remaining problem.
8. If there is cold stream in the remaining problem then return to step 1.
9. If there is (are) hot stream(s) in the remaining problem then satisfy it (them) with external cooling till target temperature.
10. Stop.

Example 1
Problem data:

	T°, t°	T^*, t^*	C, c
C1	140	320	14.45
C2	240	500	11.53
H1	320	200	16.70
H2	480	280	20.00
	MAT	=	20.00

The solution is shown in **Fig. 2.14**. $Q_F = 461.2$, $Q_1 = 2536.6$, $Q_2 = 1463.4$, $Q_3 = 1137.6$, $Q_H = 862.4$.

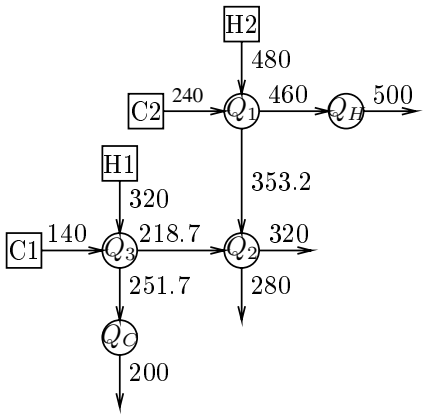


Figure 2.14: Fast algorithm, solution of Example 1

Example 2

Problem data:

	T°, t°	T^*, t^*	C, c
C1	25	420	70
H1	500	110	20
H2	430	230	50
H3	400	210	30
	MAT	=	20.00

The solution is shown in **Fig. 2.15**.

Example 1 is well soved by fast algorithm, but is designed several small exchanger for Example 2. That problem can be solved with only 4, moreove, only 3 exchanger. Fast algoritm sometimes also assignes HEN with much more external heating and cooling than necessary.

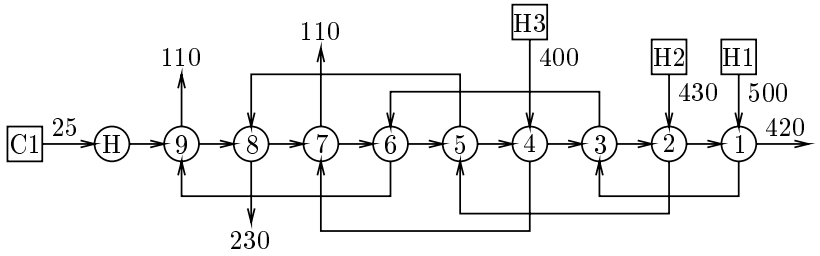


Figure 2.15: Fast algorithm, solution of Example 2

2.3.2 Pinch design

It is clear from the general pinch rules that subnetworks over and below pinch are separated according to heat transport and, therefore, they can be designed independently, if maximum energy recovery is aimed.

At any side of pinch, the demand related to a stream approaching the pinch may be satisfied with opposite streams in the same side only. Above pinch, for example, heat supply of hot stream may be taken up by cold streams above pinch only. Cooling is prohibited above pinch by Rule 2, and its heat taken up be a cold stream section below pinch is also prohibited by Rule 1.

Satisfying streams going away pinch is easier. Considering a cold stream above pinch, for example, it can be heated up either by a hot stream or with external heating.

Most difficult are the streams actually reaching (or crossing) pinch temperature, because they can be satisfied by matching those streams only which leave pinch, i.e. start at the pinch or cross it from the other side. The approach temperature between these matching pairs is just MAT, i.e. the minimum specified. These streams are called *pinch streams*.

Example

Pich design is made with the data of Example in *subsection 2.2.5*: **Table 2.1**.

Design first the subnetwork over pinch (hot side). There are one hot stream and two cold streams there, all three are pinch streams. The hot side subproblem is sketched in **Fig. 2.16**. A double line at the bottom represents the two pinch temperatures. Little square at the start of the streams contains name of the stream; heat capacity data are shown in brackets. Strams are represented with directed sections along an imaginary temperature axis. Characteristic temperatures are also shown.

Here heat exchangers are denoted with two nodes connected by a line, heaters by a single node on the line of a cold stream. This is called *grid representation* of HEN in the literature.

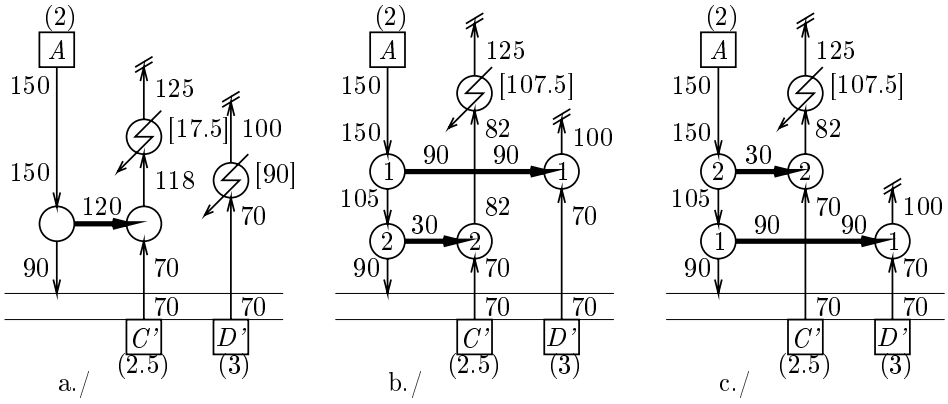


Figure 2.16: Design of hot subnetwork

There is just a single hot stream **above pinch**; this should be cooled down to pinch with the available two cold streams. The most important match is that just by the pinch; there are two possible selections for it: stream A is either matched with C' or D'. he first selection is straigthforward; its results are shown in **Fig. 2.16. a./**. Heating demand of stream C' id far more than heat supply of A, thus extrenal heating is needed for it to reach its target temperature, and of course external heating is needed for the full section of D'. Their sum is just 107,5 kW, the minimum heating calculated earlier.

There are two ways for the other selection. The first one is matching stream D' entering the exchanger at its starting temperature (i.e. pinch temperature) with stream A *entering* the exchanger with its starting temperature. Stream Az A cools down by 45 C, and this remainder is matched with stream C', as is shown in **Fig. 2.16. b./**. The other way is matching the two streams in such a way that stream A *leaves* exchanger at pinch temperature, and then match the remaining *starting* section of stream A with C', as is shown in **Fig. 2.16. c./**.

Version c./ is that complying with the principle os countercurrency, and going with smaller costs. Extra temperature differences over MAT are 0 and 3 in case b./, whereas 15 and 30 in case c./. In more complex cases designing like in case b./ may lead to situations excluding maximum energy recovery; therefore, as a general heuristic, one has to **design the matches by gradually walking away from pinch**.

There are two hot streams and two cold streams **below pinch**. Two match couples can be assigned at the direct neighborhood of pinch, as are ahown in **Fig. 2.17**.

Calculate maximum heat load of match A'-C in **Fig. 2.17. a./** under the constraint that it is a pinched exchanger, i.e. both the hot stream inlet and cold stream outlet temperatures are pinch temperatures! The approach temperature at

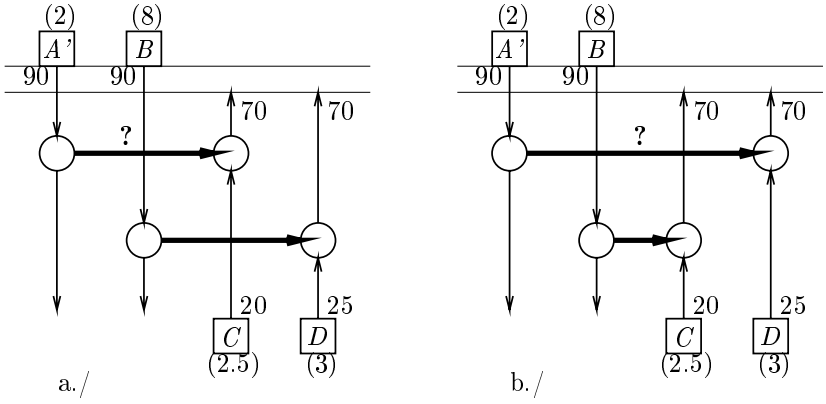


Figure 2.17: Infeasible matches below pinch

the pinch is just MAT. The hot stream cools faster then the cold stream warms up because its heat capacity is smaller: $2 < 2.5$. As a result, the approach temperature at the cold end will be smaller than MAT with any small positive heat load; therefore no heat may be exchanged between them if MAT is a strict constraint. *It follows that this match is infeasible for maximum energy recovery.* The same difficulty arises in **Fig. 2.17. b.**/ with match *A'-D* $(2 < 3)$.

To have feasible pinch matches, **for matching a stream arriving to pinch and leaving pinch the following condition should be satisfied:**

$$C_{\text{leaving}} \geq C_{\text{arriving}}$$

This condition is always satisfied for the sums of heat capacities of the pinch streams; otherwise the pinch would be somewhere else. The condition should be satisfied to each pinch match. If no such a match combination can be found then it can be achieved by **branching some streams**.

Number of streams (branches) is also a factor. Suppose, for example, that stream *C* is branched, say to 0.5 + 2 (streams *C1* and *C2*), then *A'* can match *C2*, and *B* can match *D* (**Fig. 2.18. a.**/). But what will then warm up stream *C1* to pinch?

There are three apparent choices for warming up *C1*:

1. External heating.
2. Hot stream sections above pinch.
3. Remaining section of stream *A'* or stream *B*.

Case 1 is excluded by General Pinch Rule 3. Case 2 is excluded by General Pinch Rule 1. Case 3 is infeasible because starting temperature of either one of remaining stream sections is lower than the pinch, thus the approach temperature at hot end would be smaller than MAT.

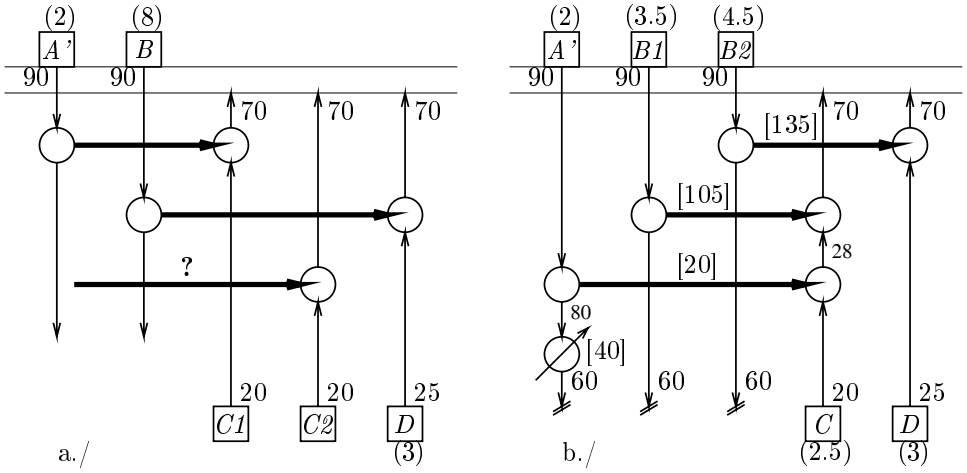


Figure 2.18: Cold subnetwork with branching

Pulling up the cold streams (or branches) to pinch at least so many hot streams (or branches) are needed. Generally, **number of streams arriving to pinch must be at least so many as that of leaving pinch:**

$$N_{\text{leaving}} \geq N_{\text{arriving}}$$

If it is not satisfied then it can be achieved by **branching some stream(s) leaving pinch**.

In our case a good solution can be obtained by branching stream B, not stream C. Branching ratio is an open question, one has some freedom here. A feasible version is shown in **Fig. 2.18. b.**/.

2.3.3 Hidden pinch

How to construct parts of HEN far from pinch is not so straightforward as near the pinch. There are at least three known strategies:

1. *Remaining Problem Analysis* with *Hidden Pinch*.
2. Heuristic of *Driving Force Plot*.
3. Design in separate intervals, then applying *Shifts and Loops* or *Loop Breaking* technique

Only the technique of *Remaining Problem Analysis* with *Hidden Pinch* is discussed here.

Fast algorithm can be applied to design far away from pinch with two modifications:

1. Below the pinch the design goes hot to cold direction, just as is originally stated, but above pinch it goes from cold to hot, gradually going away from pinch.

2. Suitability of each temporarily planned match is checked by heat cascading the problem remaining after that match. It such a cascading results in extraneous heating or cooling then heat load of the planned match is decreased.

Example
Let the remained problem at some point of design be the following:

	T°, t°	T^*, t^*	C, c
1	150	90	8
2	70	130	4
3	90	110	2
4	90	110	2
	MAT	=	20

Heat cascading of thei sproblem is shown in **Fig. 2.19. a.**/. No heating is needed (this is a cold side subproblem); pinch is really at this point because there just MAT between the hottest hot and hottest cold temperatures.

Fast algorithm suggest assigning maximum heat load match between streams 1 and 2. Heat cascade of the problem remaining after such a match is shown in **Fig. 2.19. b.**/. (This maximum heat load matching consumes all heating demand of stream 2.) According to the cascade, minimum heating is now 40 units though it was 0 before. It was 0 because we are below the original pinch.

It follows that the assigned match is either an improper selection or of too large heat load. The first assumption is surely wrong because this is just *the single possible* pinch match. One can but decrease the heat load to a limit value that involves just 0 heating of the remaining problem. The maximum such heat load can be determined either analytically or numerically. Applying this maximum acceptable heat load leads to the remaining cascade shown in **Fig. 2.19. c.**/.

A new pinch situation emerged at 120 C. According to pinch design rules, stream 1 has to be branched and the branches matched parallel with streams 3 and 4. Such emerging pinch situations are called *hidden pinch*.

2.4 Area targeting and supertargeting

2.4.1 Area targeting

Energy targeting. discussed in *section 2.2*, can be used to determine minimum heating and cooling, and their expected costs, even before syntesizing the network. The calculated values depend on the arbitrarily chosen MAT.

Normally HEN with minimum costs is to be synthesized. Beside energy costs, investments costs would also be worth to estimate in advance. Investment costs,

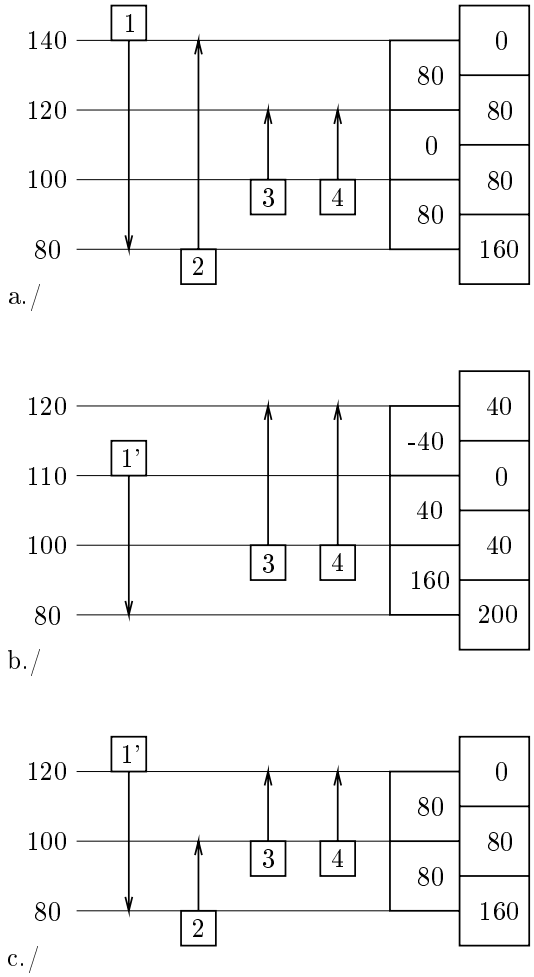


Figure 2.19: Hidden pinch, check of remaining problem

projected to a time period, can be taken proportional to the built in heat transfer area, in first estimate. How much heat transfer area is actually needed depends on the actual structure and unit design, but a minimum heat transfer area can be calculated irrespectively to them. This area minimum is also a function of the arbitrarily chosen MAT, and can be determined using the idea of so-called *vertical heat transfer*.

Area minimum and 'vertical' heat transfer

Minimum area at maximum energy recovery belongs to a network based on principle of countercurrency.

It is a basic unit operation knowledge that countercurrent arrangement is more efficient in thermodynamic sense than co-current one.

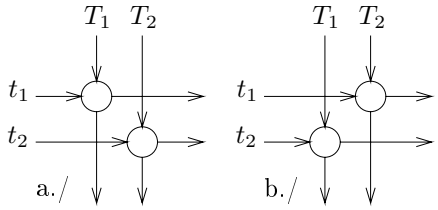


Figure 2.20: Match alternatives

In order to verify preferability of countercurrent structures, consider two alternatives shown in **Fig. 2.20**.

Apply notation

$$\Delta_{j,i} = T_j - t_i$$

$$\alpha = \frac{Q}{c_i}$$

$$\beta = \frac{Q}{C_j}$$

Change in total heat transfer area A at a fix total heat load Q when changing from arrangement a/ to arrangement b/ can be written, after algebraic manipulations, as

$$\Delta A = \frac{(\alpha-\beta)^{-1}}{U}\left[\left(\ln\frac{\Delta_{1,2}-\beta}{\Delta_{1,2}-\alpha}+\ln\frac{\Delta_{2,1}-\beta}{\Delta_{2,1}-\alpha}\right)-\left(\ln\frac{\Delta_{1,1}-\beta}{\Delta_{1,1}-\alpha}+\ln\frac{\Delta_{2,2}-\beta}{\Delta_{2,2}-\alpha}\right)\right]$$

Suppose, for example, that area A of the originally arrangement a/ was minimal; then the calculated change must be positive. Hence

$$\frac{(\Delta_{1,2}-\beta)(\Delta_{2,1}-\beta)(\Delta_{1,1}-\alpha)(\Delta_{2,2}-\alpha)}{(\Delta_{1,1}-\beta)(\Delta_{2,2}-\beta)(\Delta_{1,2}-\alpha)(\Delta_{2,1}-\alpha)} > 1$$

Take into account that

$$\Delta_{1,1} + \Delta_{2,2} = \Delta_{1,2} + \Delta_{2,1}$$

then by rearrangement:

$$(T_2 - T_1)(t_2 - t_1)\left[(\Delta_{1,1} - \alpha) + (\Delta_{2,2} - \beta)\right] \geq 0$$

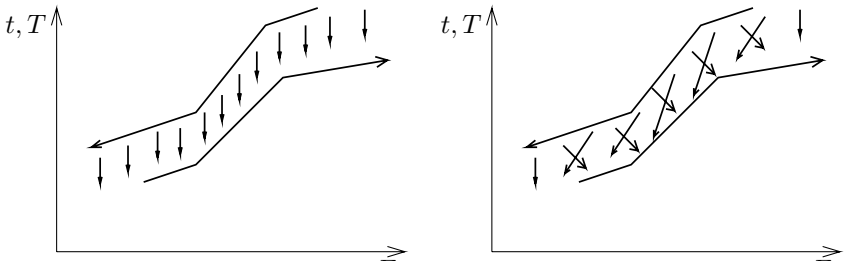
Members in square brackets are positive, thus

$$(T_2 - T_1)(t_2 - t_1) \geq 0$$

It follows that signs of the bracketed terms are the same. In other word, warmer hot stream is matched with the warmer cold stream, and vice versa.

This proof can be applied even if the considered hot 'streams' are just sections of the same single hot stream, and the same is true for cold streams. In that case contercurrent cascade structure is proved be preferable in case of multiple matches, as well.

Perfect countercurrency would mean that cold stream sections with equal temperature t are matched with hot stream sections with equal temperature T. This can be visualized in a CC diagram by vertical heat transfer directions, i.e. common temperature hot sections emit heat to common temperature cold sections over a common energy section, as is shown in **Fig. 2.21. a.**/.



a./ Vertical heat transfer b./ Crossing heat transfer

Figure 2.21: Vertical vs. crossing heat transfer

Such a vertical heat transport could be achieved by branching the stream heavily, and designing many small heat exchangers. Cost exponent b is usually a number smaller than 1, and thus applying fewer exchangers with cris-crossing, as is shown in **ig. 2.21. b.**/ is cheaper. Then total area A increases but costs decrease.

A single match is shown in **Fig. 2.22**. All four temperatures are known, hence the logarithmic mean driving force Δ can be readily calculated. Projection of axis E is heat load Q. If overal heat transfer coefficient U is estimated then the heat transfer area can be at once calculated as

$$A = \frac{Q}{U\Delta}$$

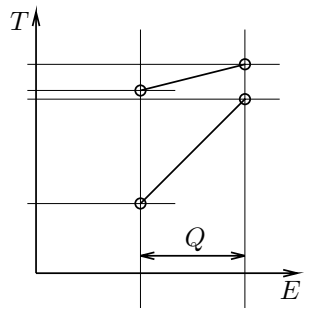


Figure 2.22: Egyetlen csatolás CC-diagramja

Suppose now that there are several heat exchangers with heat loeads Q_i, and with the same for temperatures, and equal overal heat transfer coefficients U. Then their total area can be alculated as

$$A = \sum_i A_i = \sum_i \frac{Q_i}{U\Delta} = \frac{1}{U\Delta} \sum_i Q_i = \frac{Q}{k\Delta}$$

where

$$Q = \sum_i Q_i$$

That is, if **Fig. 2.22** represents composite curve sections with, say 2 hot and 3 cold streams, then they can be branched to become 6 hot and 6 cold branches connected with 6 exchangers, and their total area can be calculated as they were a single exchanger with heat load Q and the four temperatures.

Area minimum

Consider now a full CC diagram, and subdivide its projection on the energy axis to intervals with constant slopes, as is shown in **Fig. 2.23**. Average logarithmic approach temperature Δ_j can be calculated to each j, thus minimum area A_j belonging to each energy interval j with heat load Q_j can be calculated as discussed above. Hence the minimum area is

$$A_{min} = \sum_j A_j = \frac{1}{U} \sum_j \frac{Q_j}{\Delta_j}$$

Slightly better approximation can be given by approximating U between two streams i and m as

$$\frac{1}{U_{i,m}} = \frac{1}{\alpha_i} + \frac{1}{\alpha_m}$$

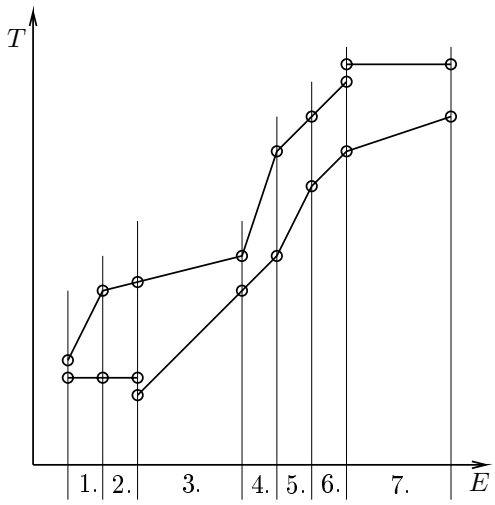


Figure 2.23: Subdivision of CC diagram to energy intervals

Then the estimate for minimum total area is

$$A_{min} = \sum_j \left(\frac{1}{\Delta_j} \sum_i \frac{Q_{i,j}}{\alpha_i} \right)$$

where $Q_{i,j}$ is heat load demand of stream i over energy interval j.

2.4.2 Supertargeting

Since both minimum energy turnover and minimum heat transfer area can be calculated as function of MAT, both cost items can be estimated, and they and their sum can be plotted against MAT before synthesizing a network. Such a plot is shown in **Fig. 2.24**.

Thus one has an estimate to the minimum total costs. If that is acceptable then the design is worth to perform with a MAT at or near the optimal one.

This diagram can be transformed to another form that compares area costs and energy costs, or simply total built in area and annual energy consumption, as is ahown in **Fig. 2.25**. The continuous line consists of points of suboptimal designs. MAT changes continuously along this line: High energy consumption and low area belongs to small MAT, ow energy consumption and high area to larger MAT. Economic optimum is located somewhere in between.

Retrofit targeting

Retrofit costs are different from costs of building a new plant.

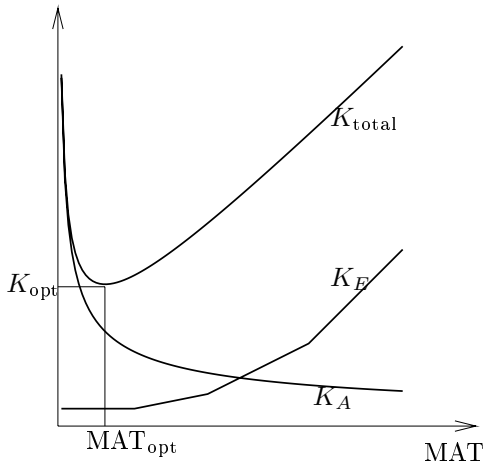


Figure 2.24: Supertargeting

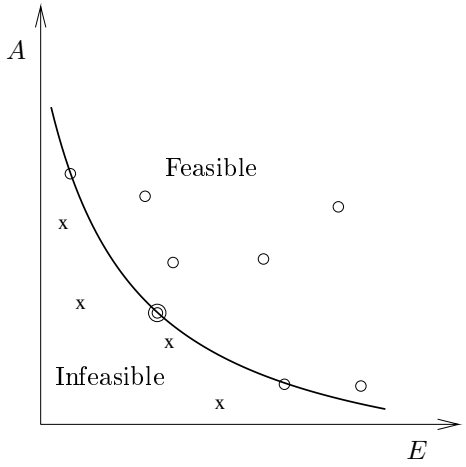


Figure 2.25: Feasible and infeasible design domains

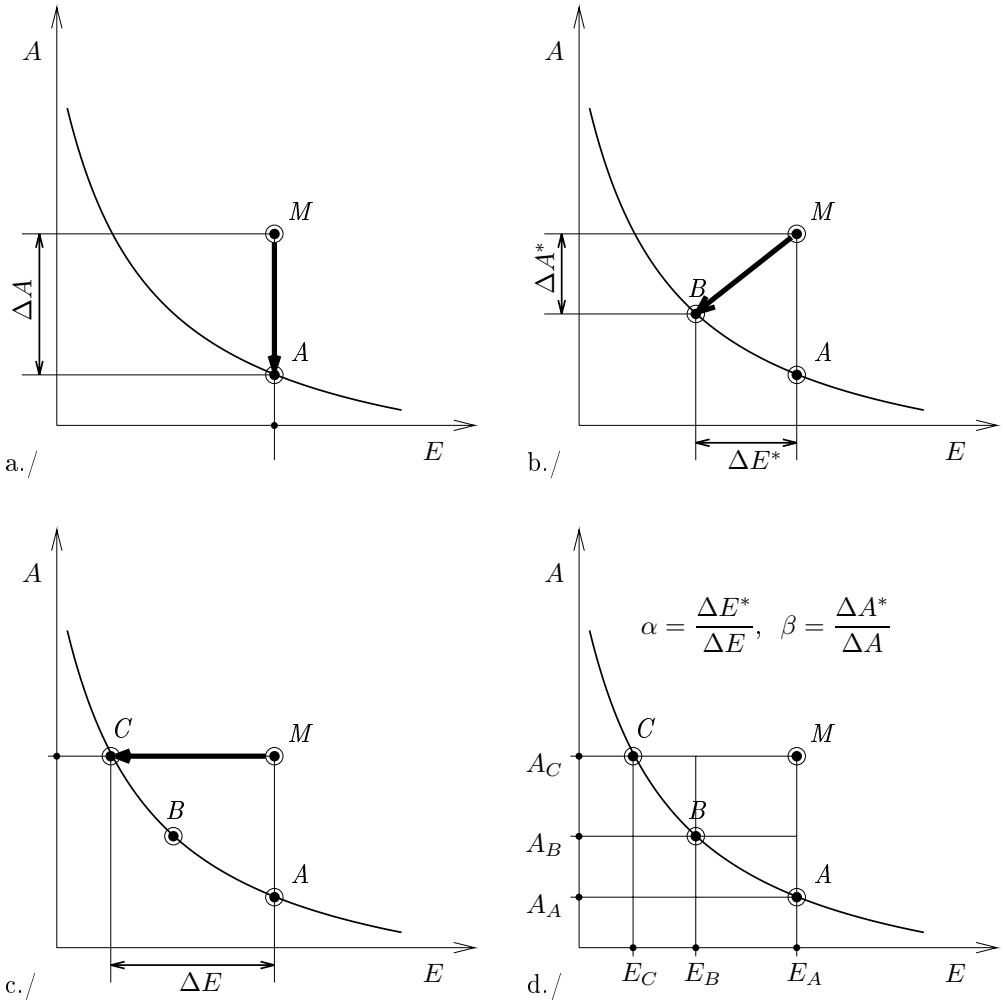


Figure 2.26: Retrofit and cost points

Consider point M representing an already working plant in **Fig. 2.26. a.**/. Minimum area design belonging to given energy consumption is located in point A. Naturally, one would like to decrease also energy consumption. Global economic optimum is found is point B, shown in **Fig. 2.26. b.**/. Arrow M-B gives us the direction of change. Its projection to axis E shows the expected decrease in energy consumption, the projection to axis A shows the expected savings in area costs.

But so much area is already built in the plant as is represented by point M. Given that area, even more energy saving could be achieved. The energy saving that could be achieved in some ideal case is shown in **Fig. 2.26. c.**/. In principle, projetion of this horizontal arrow to axis E shown savings in energy consumption **without additional investment**, thus immidiate return. Unfortunately, this ideal rearrangement cannot be achieved because the already installed and working system is designed with some shape and ratios. To form a new, different system from it, it had to be melted and new exchangers with pipelines and instrumentation fabricated.

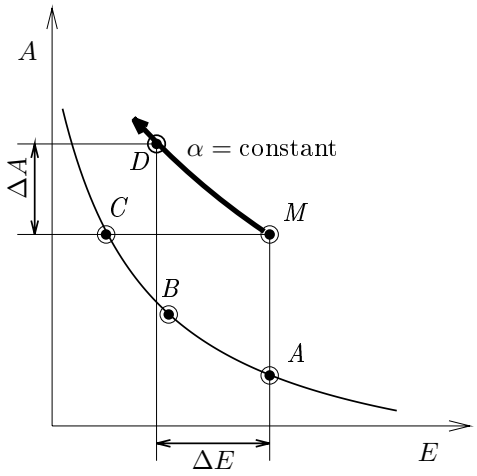


Figure 2.27: Retrofit targeting

Ratios

$$\alpha = \frac{E_C - E_B}{E_C - E_A}$$

$$\beta = \frac{A_A - A_B}{A_A - A_C}$$

read from **Fig. 2.26. d.**/ are characteristic to the constructed HEN. With retrofitting, these ratios can change a little only. Assuming these ration remain constant, the area invenstment and energy consumption effects of any smaller change, like inserting a new exchanger, repiping etc., can be expected as shown in **Fig.**

2.27. Optimal retrofitting points are located along the wide curved arrow from M toward left and up. Retrofit to HEN in point D involves investment $K = A_D - A_M$ and energy consumption savings $E = E_M - E_D$.

2.5 GCC and thermal integration

2.5.1 Utility levels

GCC can be used to optimally allocate utility temperatures and, sometimes, make a compromise between utility levels and HENS.

A simple GCC, of pinch situation, is shown as broken line CDFHIJ in **Fig. 2.28**. Lines of (minimum) heating and cooling are also shown so that a total GCC line ABCDFHIJKL starts and ends at $Q = 0$, this is called *balanced*) GCC. The heating and cooling are imagined at the top and bottom temperatures just because with this allocation the introduced heat can reach any point of the system, and the cooling can be applied to any point of the system. This was preferable for accounting, i.e. calculating heat balance while taking into account Second Low.

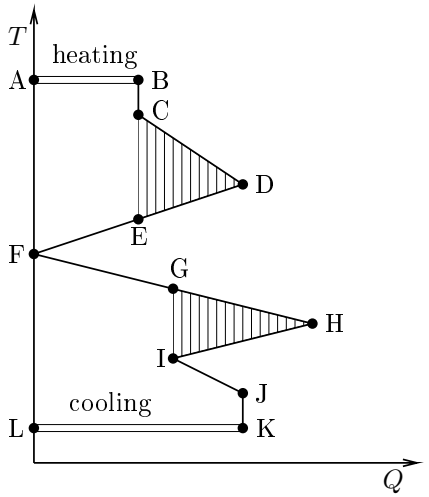


Figure 2.28: Balanced GCC

Applyinh the utilities at these temperatures are, however, not necessary. The balanced GCC can be subdivided along the points ABCDEFGHIJKL because line sections with positive slope correspond to heating demand whereas sections with negative slope represent heat supply. Section CD has enough supply to satisfy the need of section ED; this is shown by vertical hatching. Above pinch point F, only the need of section FE is to be satisfied with external utility, and a much lower temperature heating medium can be applied, instead of section AB. For example,

middle or small pressure steam, or heating oil, can be used instead of high pressure steam, as is shown in **Fig. 2.29**.

In the same way, section IH can cool section GH. Thus, instead of expensive refrigeration cooling LK one can apply cooling water or even air cooling under section FG. If pinch temperature is high enogh then even secondary steam generation can be designed under it.

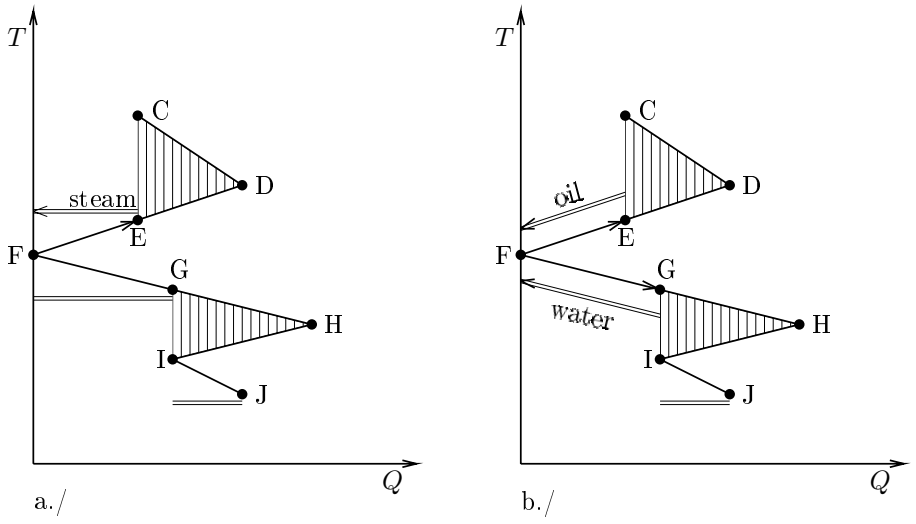


Figure 2.29: Utility allocation

Parts denoted with vertical hatching represent internal energy recovery. At optimal location of utilities, these *pockets* are not considered; they can be omitted. This gives rise to a so-called *reduced GCC* as shown in **Fig. 2.30. a.**/.

Corner points of reduced GCC indicate temperatures that are constraints to temperatures of applicable utilities. Utility lines touch the reduced GCC in ideal case; these points are called *utility-pinch*. For example, reduced GCC of another system is shown in **Fig. 2.30. b.**/, here flue gas utility line just touches the reduced GCC.

Larger chemical plants, so-called chemical complexes, usually work in symbiosis with a power plant producing both electric power and heating steam. Such power plants are designed to produce a set of steam streams at specified pressure levels. When they are designed, engineers take into account expected seasonal need fluctuations and the need of chemical plants. Actual used steam in the chemical plant is produced from higher pressure steam with throttling valves, giving rise to energy loss. In ideal case the steam production levels and the chemical plant energy system are to be co-designed to some optimum.

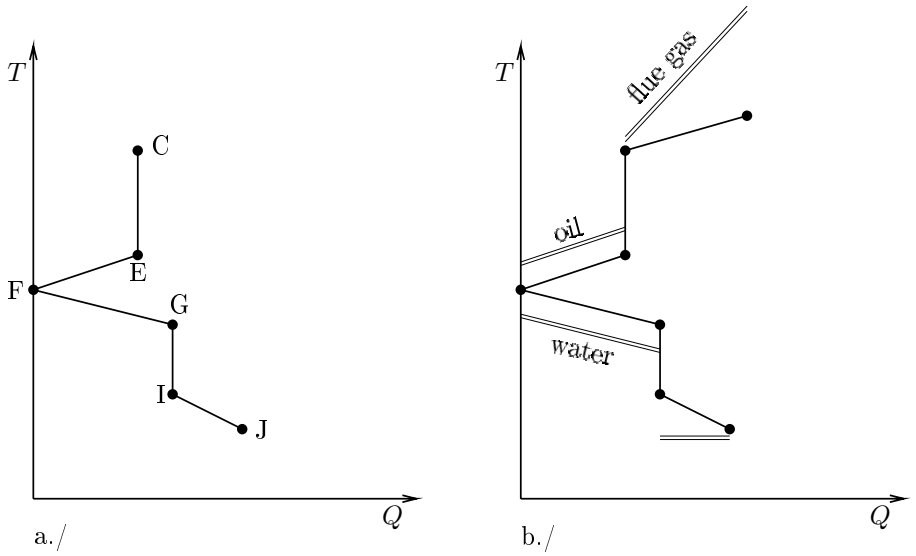


Figure 2.30: Reduced GCU

2.5.2 Thermal processes along the heat cascade

Suppose that a new distillation column is to be installed in a system already complex enough to characterize with a heat cascade. (Any other heat degrading process can be considered instead of distillation. It can be a small power plant. It can be some other separation process, e.g. extraction using distillation as a regeneration part. In can be evaporation heated with steam, and followed by a condenser working with cooling water.) The general question here is which streams in the original system can be utilized to accomplich the heating and / or cooling of the new process. For this aim, consider first the heat cascade *without* the new process and analyze how it can be integrated to the cascade. At first glance one can assume that pinch temperature is not changed as a result of integration.

If the new distillation process is not integrated in the cascade then the heat duties of the boiler and the condenser increase the minimum heating and cooling, respectively, of the system, as is shown in **Fig. 2.31**.

Roughly three version can be conceived, as is depicted in the following figures.

Temperatures of both the reboiler and the condenser are above pinch in **Fig. 2.32. a.**/. Heat load Q_C removed from the condenser is led to interval 5; thus the heat transport from interval 4 to interval 5 must be decreased by the same amount. So must be decreased the transport above because of heat balance. Above the upper connection, in the same time, the demand for boiling heat load Q_B is also to be covered; therefor the transport figures change with $Q_B - Q_C$ there. Finally, this is the value the minimum heating changes with. That is, if both integration

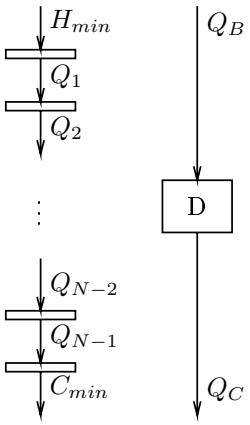


Figure 2.31: Heat cascade with non-integrated distillation

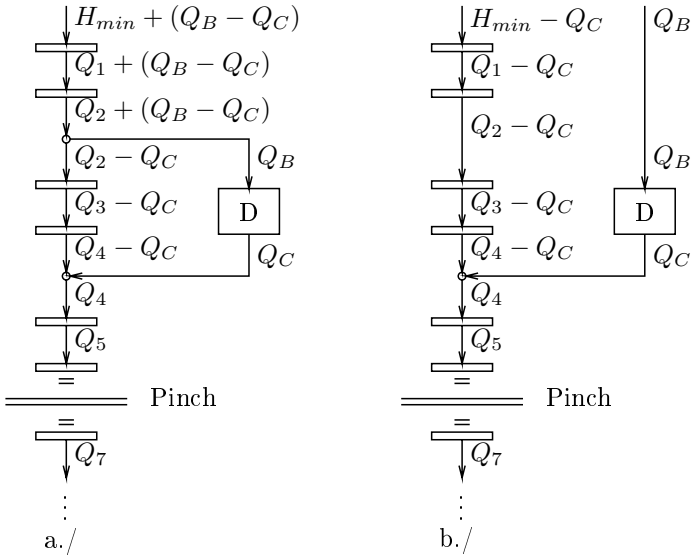


Figure 2.32: Distillation integrated to the cascade above pinch

points are above pinch then, instead of Q_B, the extra heating is only $Q_B - Q_C$. This difference is normally a magnitude smaller than either heat load at the ends of the colomn. A condition for this preferable integration is, of course, that neither of $Q_i - Q_C$ or $Q_i + Q_B - Q_C$ become negative.

The case when only the condenser is integrated to the cascade is shown in **Fig. 2.32. b.**/. This is equivalent to the case of a./ because if the boiler load is covered with external heating then the change of total heating is the same.

Fig. 2.33 is just a mirror image of **Fig. 2.32**, with similar results for integration below pinch.

The two figures together show that this way a significan savings in energy consumption can be achieved.

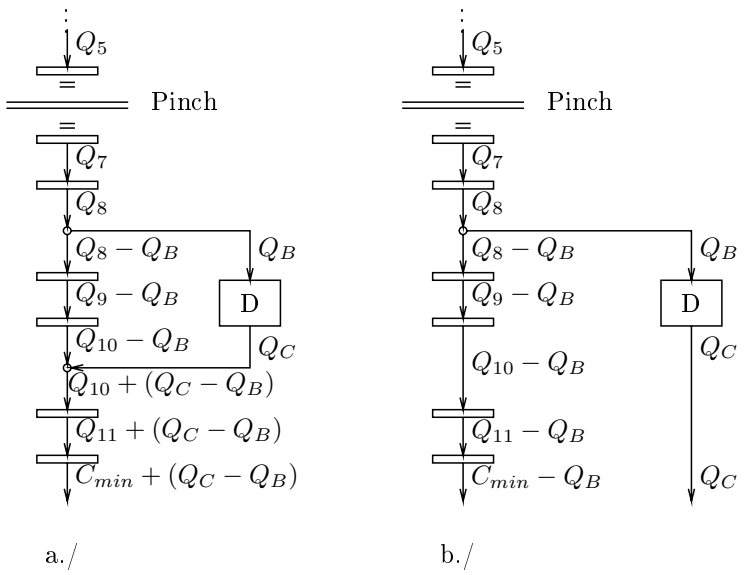


Figure 2.33: Distillation integrated to the cascade below pinch

Fig. 2.34. a./ shows the case where the boiler is integrated above, and condenser below, the pinch. In this case the inter-interval transports increase with Q_B above pinch, and increase with Q_C below pinch; that is no saving is achieved. No wonder because in this case the heat energy is carried along the distillatiom tower from the boiler to the condenser through pinch temperature. Partial integrations are shown in **Figs. 2.34. b.**/ **and c.**/, with the same results.

All these considerations have important consequenced in design of distillation (and other thermal) processes. For example, *increaing or decreasing column pressure inreases or decreases temperatures, respectively, and thus such processes overlapping pinch can be shifted above or below pinch so that integration becomes preferable. If it does not work then properties of the cascade might be changed for enabling preferable integration.*

Note also that integrating the thermal process to one side of pinch means putting in inside a pocked of GCC. Thus one puts such a process in between a heat recovery

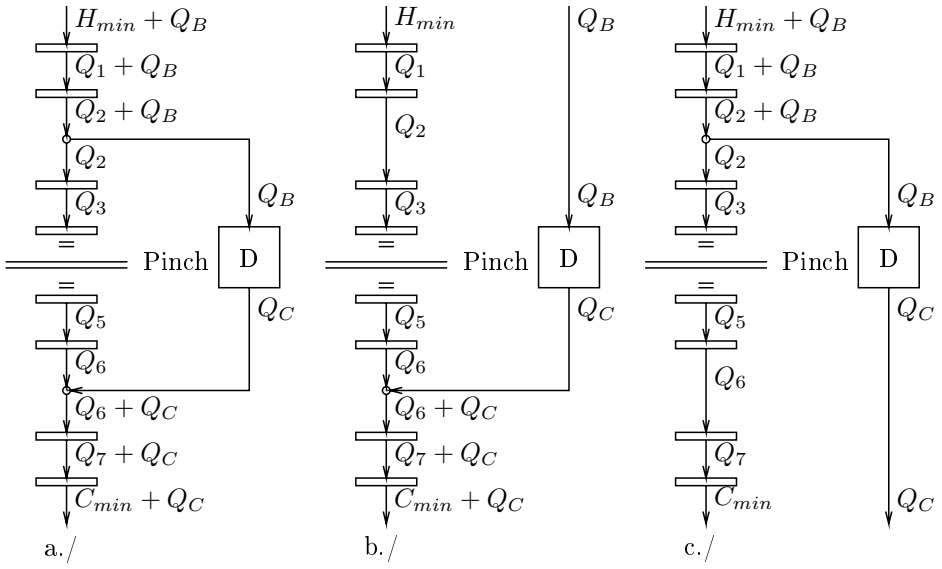


Figure 2.34: Distillation integrated to the cascade across pinch

path, as is ahown in **Fig. 2.35**. Fig. a/ depicts putting a distillation process there, whereas a turbine in sketched in Fig. b/.

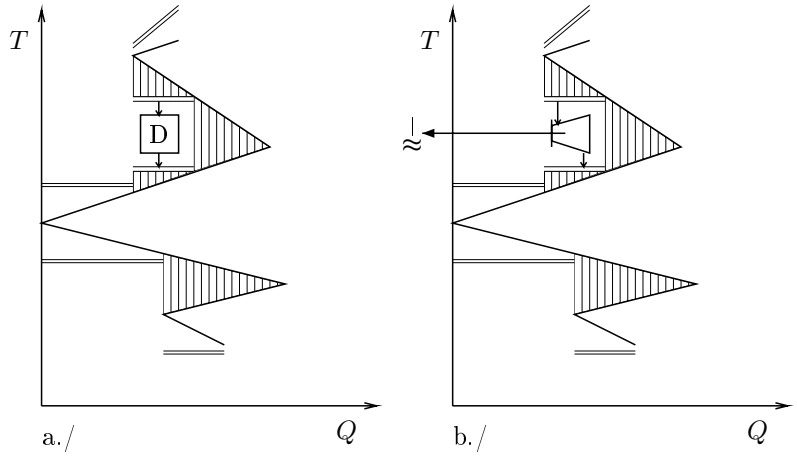


Figure 2.35: Thermal operations in the pockets of GCC

2.5.3 Heat pumps along the heat cascade

Heat pumps and refrigerators (cooling cycles) work on the same principles. They transport heat from lower temperature to higher temperature, and consume work for this. Denote the heat taken up at lower temperature with Q_L, the consumed work with W, then the heat given over at higher temperature Q_H equals the sum of the other two.

Just as distillation or turbine, a heat pump can also be integrated to heat cascade in three basic ways: over, under, or across pinch, as shown in **Fig. 2.36**.

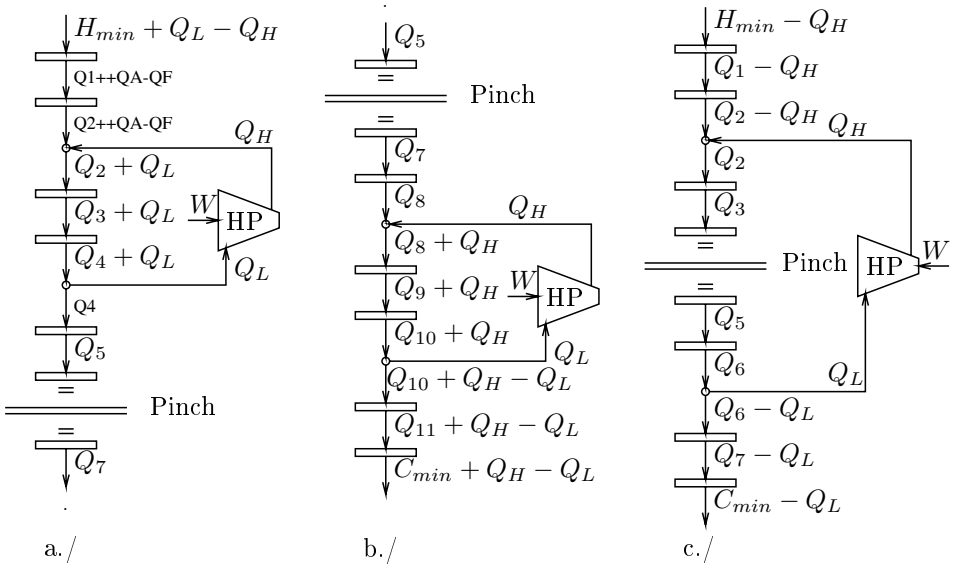


Figure 2.36: Heat pump integrated to heat cascade

Figure a/ shows integration above pinch. Net heat supply must necesserily be increased with Q_L between two pints of integration. From the upper integration point upwards the change is $Q_L - Q_U$, i.e. minimum heating decreases just with W, but that amount in introduced to the compressor itself. Thus, there is not any saving. Figure b/ is almost its mirror. The only difference is that minimum cooling is increased by W, the introduced work; this is even worse than without applying heat pump.

Heat pump is accrosed pinch in Figure c. In this case, however, minimum heating is decreased with Q_U, minimum cooling is decreased with Q_L, as a result of applying W work load. This is a preferable integration.

The rules for applying heat pump against heat cascade is just the opposite to applying a heat power unit or heat consuming unit against

the cascade. The proper use of heat pump along a cascade is applying it across the pinch.

2.6 Questions

1. Construct the CC diagram of the problem given in **Table 2.1**!
2. Construct GCC of the problem given in **Table 2.1**!
3. Calculate heat cascade of the problem given in **Table 2.1** and determine its pinch data!
4. What are general pinch rules? What are the pinch design rules?
5. How can you determin minimum heat transfer area of a HENS problem?
6. What is supertargeting? What is measured in its independent variable axis?
7. What is GCC for? How can it be used for that?
8. What is preferred integration of heat consuming processes to a heat cascade? What is preferred integration of heat pumps to a heat cascade?

2.7 Suggested literature

SHENOY, U.V.: *Heat Exchanger Network Synthesis: Process Optimization by Energy and Resource Analysis.* Gulf Publishing Company, Houston, TX, USA, 1995.

LINNHOFF, B.; ET AL.: *Pinch Analysis and Process Integration: A User Guide on Process Integration for the Efficient Use of Energy.* Butterworth, 1982.

Chapter 3

Component recovery networks

3.1 Mass Exchane Networks (MENs)

Entrainer are applied for recovery, removal, separation etc. of some chemical components from gases and liquids. Such entrainers are used in unit porcesses like extraction, extractive distillation, extractive crystallization, absorption, or they can be carriers at chromatographic separation processes. These entrainers, sometimes called solvers, are utilities just like heat carriers in HEN. Water and air are frequently applied as entrainers. However, even the capacity of these entrainers are not infinite.

The same entrainer can be applied several times in complex processes, or in larger chemical sites. The same entrainer stream can be re-used at different solute concentrations, just like a hot or a cold stream can be used at different temperatures. Systems in which entrainers are used as utility streams, and some specific chemical components are to be recovered, are called *mass exchange networks (MENs)*. In analogy with HENs, one's target can be minimum entrainer consumption, and MAN Synthesis. Role of temperature in HENS is taken up by concentration in MANS.

Energy recovery systems (HENs)	Component recovery systems (MENs)
Heat energy	Contamination
Heat load	Component flow rate
Temperature	Concentration

In the same way as heat energy flows (conduction of internal energy happens) from higher temperature to lower temperature only, conductive component flow (diffusion) happens spontaneously from higher to lower concentration only at common pressure.

Basic concept are best visualized considering a simple process of countercurrent absorption or extraction. There are to inlet and two oulet streams, there is a dilute end and a dense or concentrated end of the unit. Such an arrangement is shown in

Fig. 3.1; here the two counterflowing flow rates are that of gas G and liquid L, and respetive concentrations of the carried contaminating or valueable component to be recovered are denoded by y and x.

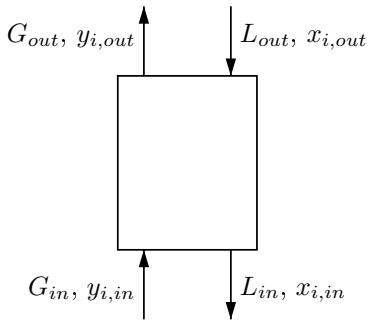


Figure 3.1: Process streams of an absorber

Conventional equilibrium diagrams (with straight equilibrium lines, for simplicity) are shown in **Fig. 3.1**. *equilibrium line* co-ordinates respective concentrations of equilibrium phases; this is independent of the operation. The operations are characterized with the slope of *operating line* depending on the ratio of flow rates of counterflowing streams. The nearer the operating line to the equilibrium line, the nearer the operation to the ideal reversible process, and the less flow rate of entrainer is used although the number of equlibrium stages, as well as invenstment costs, are higher.

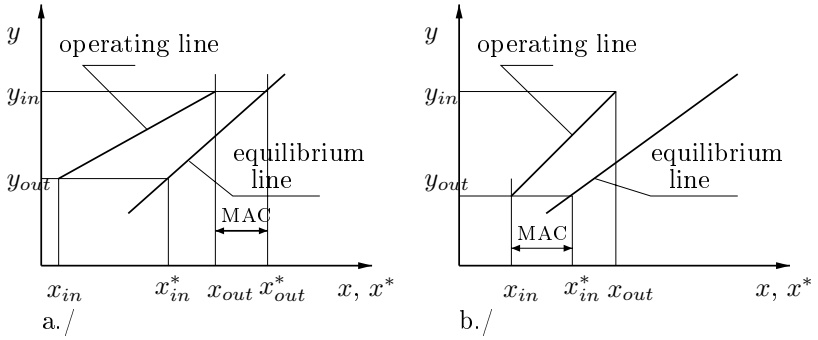


Figure 3.2: Operating line, MAC, equilibrium composition

Just line MAT in HENS, minimum approach concentration, MAC, is an arbitrarily chosen value that has a practical use.

Table 3.1: Example problem: H_2S removal

Stream name	Mass flow rate kg/s	mol % in	mol % out
G1	0.9	7.00	0.03
G2	0.1	5.10	0.01
L1 (ammonia)	2.3	0.06	3.10
L2 (methanol)	?	0.02	0.35

Although there can be several contaminating or valueable components to be recovered in MANs, here we discuss the case of a single component only, and show hoe composite curves are constructed, and what pinch means in this ase.

The task is removing hydrogen sulphide from fresh coke gas and from partially purified coke gas. Products are different; a larger recovery is needed from the partially purified gas. Contaminating component (i.e. hydrogen sulphide) is first of all absorbed in aquous ammonua solution but contamination remaining in the gas aftre this treatment is finally removed with cooled methanol. Ammonia is obtained from the coking process, its flow rate is given. Methanol is considered as utility, its flow rate is to be minimized. Data are given in **Table 3.1**.

MAC is uniformly 0.01 mol %. Equlibrium relation is also a part of the problem, a difference from HENS. Equilibrium mole fractions of hydrogen sulphide between the given gas and aquous ammonia can be well approximated with a straigh line:

$$y = 1.45x_1$$

Equilibrium between the gas and cooled methanol is described with another straight line :

$$y = 0.26x_2$$

Double concentratios are used when calculating equilibrium with ammonia at constructing CC and cascade: to each gas mole fraction there is a particular **equilibrium liquid mole fraction** as

$$x_1^* = \frac{y}{1.45} - 0.0001$$

where 0.0001 is the MAC subtracted, i.e. the streams are *shifted* together vertically (along concentration axis). The 6 concentration couples bracketing 5 concentration intervals are shown in **Table 3.2**.

Stream L2, i.e. data of the utility stream, do not take place in this table. Only concentrations x^* are used in targeting; values y a listed here just for explanation.

Net component supplies can be summed up interval by interval, as is shown in **Table 3.3**, and cascading is done in the same way as cascading at HENs, see **Table 3.4**.

Table 3.2: Interval borders in mol %

note	y	x^*	note
$G1_{in}$	7.00	4.82	x^*
$G2_{in}$	5.10	3.51	x^*
y^*	4.51	3.10	$L1_{out}$
y^*	0.10	0.06	$L1_{in}$
$G1_{out}$	0.03	0.01	x^*
$G2_{out}$	0.01	0.00	x^*

Table 3.3: Net supplies, in g/s

Interval	Net supply g/s
1	17.10
2	5.90
3	-25.84
4	0.72
5	0.02

Negative of the smallest algebraic inter-interval supply should be introduced from above to get a non-negative array in th elast column.

Upper value 2.84 g/s is unexploited washing capacity of aquous ammonia at minimum methanol consumption. Lower value 0.74 g/s is the amount of hydrogen sulphide to be removed with methanol. Minimum methanol consumption can be calculated from here (see also methanol specification):

$$L_2 = \frac{0.74}{0.0035 - 0.0002} = 224.2\,\text{g/s}$$

Demands for increasing and decreasing concentrations can be visualized with straight line setions in concentration against mass flow rate digrams. Uptaken and overgiven amounts of contaminating component are measured over the horizontal axis, equivalent equlibrium concentrations shifted with MAC are shown in the vertical axis. In the particular case, stream G1 gives over 62.73 g/s, stream G2 5.09 g/s contamination, while stream L1 takes up 69.92 g/s. Just like CC diagram at HENS, *concentration composite curves* (CCC) can be constructed in case of MENS. These lines an be shifted along mass flow rate axis, and can be *pinched* as limit arrangement.

Due to the large variations in concentrations, shape of the diagram would visualize well in large scale only. *Concentration GCC diagram* (CGCC) may be a better plot, see **Fig. 3.3**. CCC of just a fictious problem is shown in **Fig. 3.4**.

Table 3.4: Concentration cascading

y mol %	x mol %	Supply g/s	Supply g/s
7.00	4.82	0.00	2.84
5.10	3.51	17.10	19.94
4.51	3.10	23.00	25.84
0.10	0.06	-2.84	0.00
0.03	0.01	-2.12	0.72
0.01	0.00	-2.10	0.74

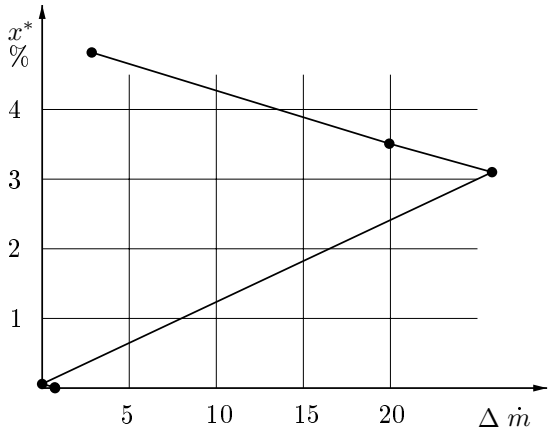


Figure 3.3: CGCC of Example problem

One can read from the figure:

- Unexploited recovery potential in the process
- Mass flow rate to be taken up by entrainer
- Internal recovery
- Pinch point in the concentration scale

For designing MEN, similar rules are applicalbe as for HEN. Solution of the example with minimum external utility (entrainer) is shown in **Fig. 3.5. a.**/. The network can be simplified, of course, if more entrainer is used. In that case outlet concentration of methanol can also change. Methanol consumption is 5.6667 kg/s in the system shown in **Fig. 3.5. b.**/.

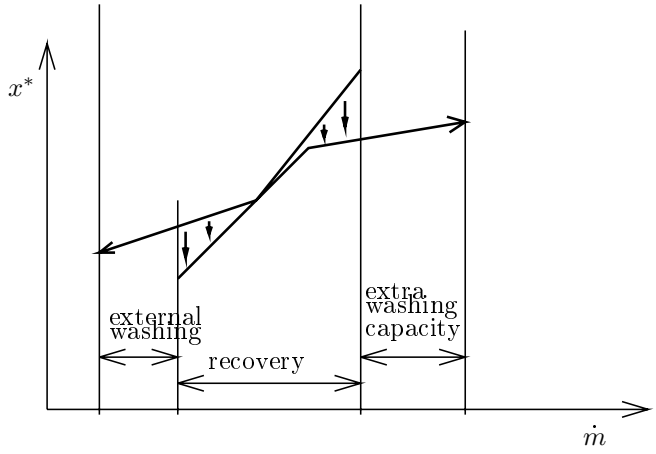


Figure 3.4: Interpretation of CCC

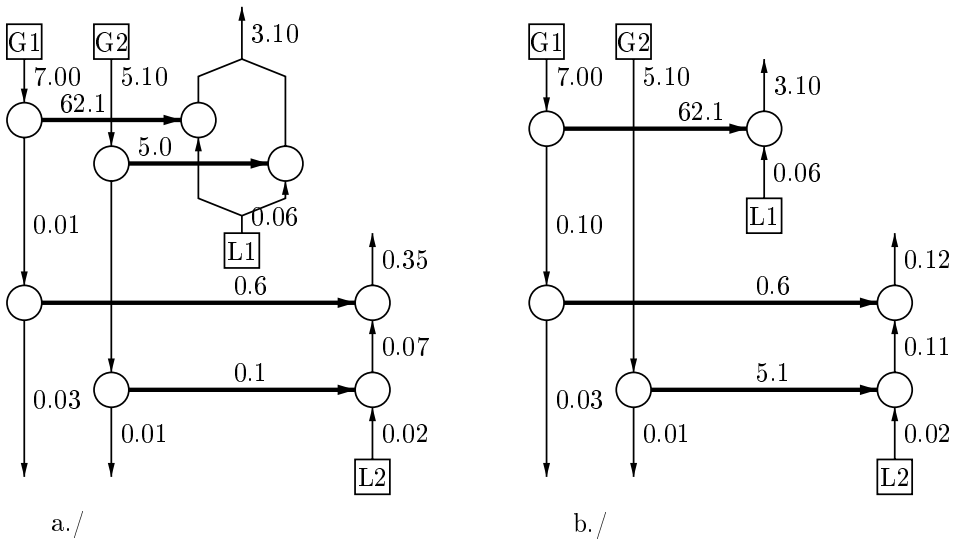


Figure 3.5: Mass exchange networks

3.2 Water use and re-use networks

3.2.1 Water profiles

Water is considered as a special utility in MENs. Water can be used and re-used, and even re-generated in a water cleaning process. Water is used for washing out some contaminating components or for stripping, for example. Constraints are put on the purity of the applied water stream and on the maximum enabled concentration of contaminating component when it leaves the process. Depending on these constraints, the amount of contamination to be removed, and on the actual purity of the applied water, the consumption (water flow rate) can be different. The total consumed fresh water flow rate at a complex site can be decreased by carefully selecting inlet and outlet concentration so that water used in a process can be re-used in another process.

Here we deal with utility water only, and the treatment applied in *section 3.1* can be simplified. Demands against water are visualized in *C-m* diagrams. Horizontal axis shown the amount of contamination taken up by the water stream. For simplicity, only one contaminating component is considered here.

Straight line section directed from right to left in **Fig. 3.6** represents concentrations in the process stream, as a function of contamination amount given over.

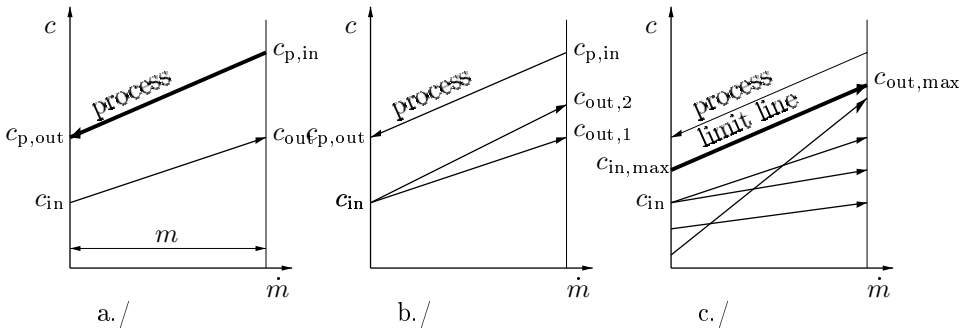


Figure 3.6: Water lines and limit line

The same task can be accomplished with different water flow rates. The higher water flow rate applied, the lower outlet concentration the water has, see **Fig. 3.6. b./** . Initial water concentration can also change. There are, however, limiting inlet and outlet concentrations, assigning a *limiting water profile* or *limit line*, also shown in **Fig. 3.6. c./** .

Once this is constructed, **process profile is no more needed**, that was used only for defining the limiting line.

3.2.2 Limit Composite Curve (LCC) and pinch

Data of four water using processes are collected in **Tble 3.5** (*Wang & Smith, 1994*). V is water mass flow rate in t/h, i.e. approximately m^3/h. Measurement units are selected such a way that $V \times \Delta c$ provides us with m in g/h. For example, $20 \times (100 - 0) = 2000$ g/h, i.e. 2 kg/h, in Process 1. Total load of the four processes is 41 kg/h. Total water need would be 170 t/h without re-use of water, but used water stream from Processes 1 and 2 can be re-used in Process 4, so that the consumption can be less.

Table 3.5: Water use example (*Wang* and *Smith*, 1994)

Process	m (kg/h)	c_{in} (ppm)	c_{out} (ppm)	V (1000 kg/h)
1	2	0	100	20
2	5	50	100	100
3	30	50	800	40
4	4	400	800	10

Data in the Table refer to limit lines only. Water lines belonging to minimum consumption using uncontaminated fresh water are shown in **Fig. 3.7**. If fresh water is used in all four processes then their minimum flow rates are smaller than given in the Table. The consumptions in this case are: 20 t/h, 50 t/h, 37.5 t/h, and 5 t/h, altogether 112.5 t/h only.

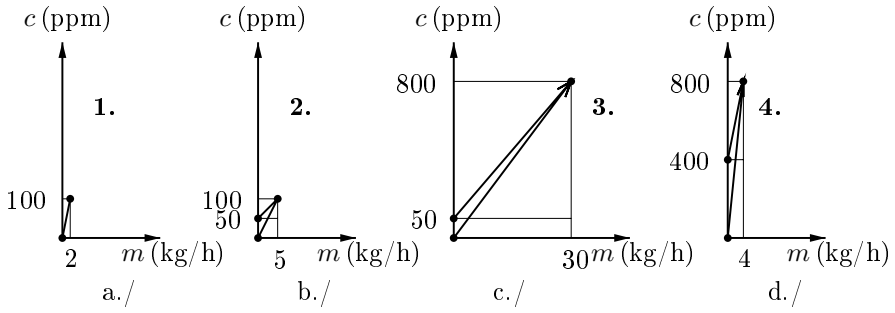


Figure 3.7: Processes in Example of Wang and Smith

Limiting Composite Curve can be constructed in the same way as composite curves are made in the case of HENS. This is shown in **Figs. 3.8. a./ and b./**. Line of used water must lie below (or at least not above) this LCC. The steepest the line, the less water is consumed. Theoretical minimum of consumption

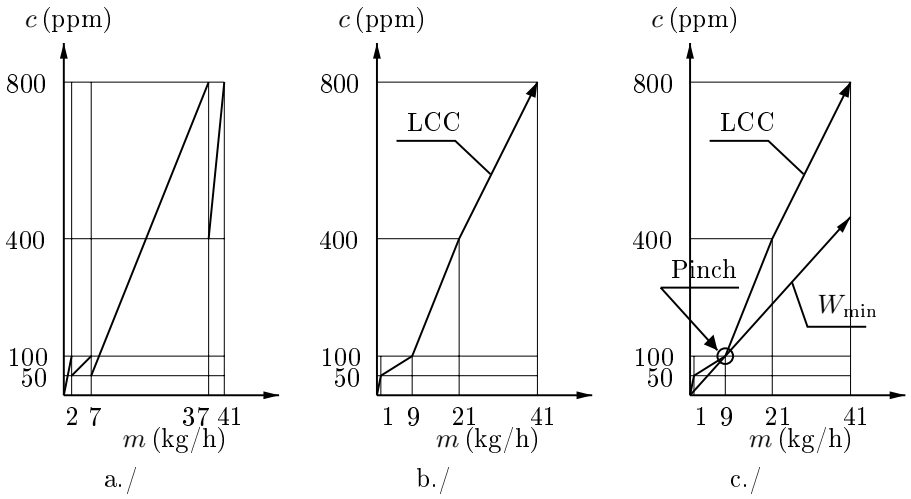


Figure 3.8: Construction of LCC, and line of minimum water consumption

is determined by pinching the two lines, as is shown in **Fig. 3.8. c.**/. From the slope of the pinched *water line*, one can read that minimum consumption is 90 t/h only.

3.2.3 Assignment of water using networks

Water using network with minimum consumption can be synthesized in several ways. One of them follows a technique originally developed for HENS: synthesize subnetworks in temperature intervals, and then simplify the network in successive development steps. However, there are two ways the first step can be accomplished in case of water using networks.

One of them is subdividing the m axis at breaking points of LCC, and synthesing subnetworks in m-intervals, as is shown in **Fig. 3.9**. This is in analogy with vertical heat transport, here it is vertical component transport, and produces a design with small unit costs. The initial network obtained this way can be simplified by lumping neighboring matches, and shifting matches along a process in feasible concentration range, to get some matches neighboring so that they can be lumped (loop-breaking).

The other one is subdividing the c axis at breaking points of LCC, and synthesing subnetworks in c-intervals, as is shown in **Fig. 3.10**. This method produces small number of matches, and then this initial network can also be further simplified by the method of shiftings and loop-breaking. A first step, for example, can be lumping two matches of Process 3 at two sides of border $c = 400$, and lumping two matches of Process 1 at two sides of border $c = 50$.

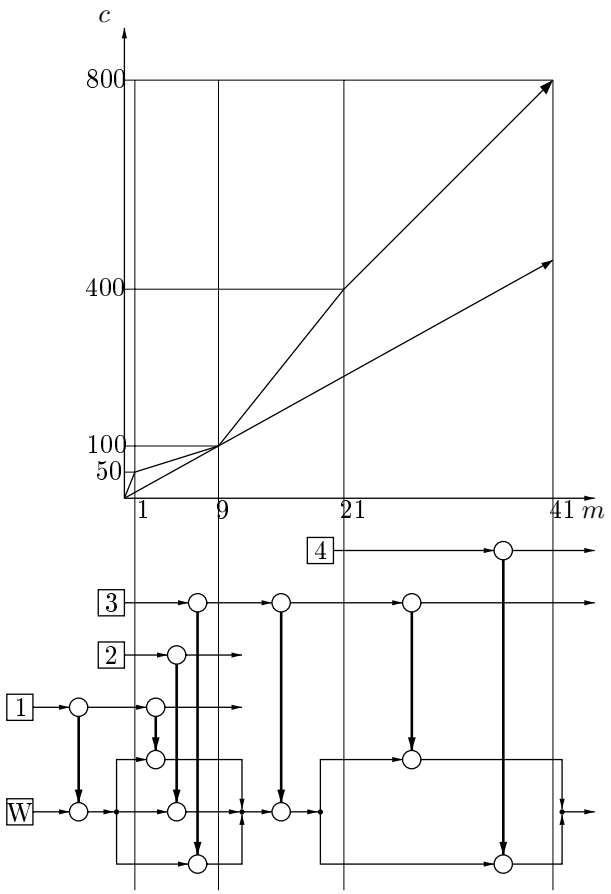


Figure 3.9: Assignment of water using network according to mass flow rate intervals

3.2.4 Regeneration and re-use

Used, therefore contaminated, water has to be cleaned before emitting it to the environment. The cleaned water is either let out or re-used. Re-use of cleaned water is, in principle, the same as using fresh water, and regeneration is cleaning, in essence. However, the concentration after regeneration may be different from that after cleaning before emission, and can be defferent from that of the fresh water.

In the consideration that follow in this section, we suppose that concentration c_0 after regeneration is specified. The real question is which used stream is worth to regenerate and re-use.

Regeneration and re-use has economic and environmental significance in two

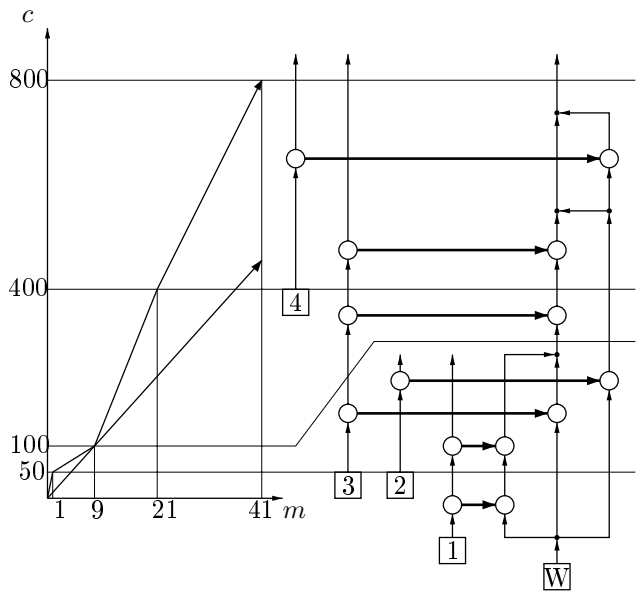


Figure 3.10: Assignment of water using network according to concentration intervals

cases:

1. Cleaning (regeneration) is made before water reaches it most contaminated state. Cleaning from this mediate concentration is less expensive.
2. The targeted concentration after cleaning is higher (smaller purity) that that of final cleaning befor emission. Cleaning to this less purity is less expensive. This is applicable if such a slightly contaminated water can be used in the process.

From the viewpoint of network synthesis, there are two network kinds with regeneration:

1. Regeneration and re-use: Use of the partially cleaned water at some downstream point of the network.
2. Regeneration and recycling: Use of the partially cleaned water at some upstream point of the network.

Here we discuss **regeneration and re-use**, without recycling. Check how minimum consumption changes if regeneration is done at concentration c_R, below or above pinch concentration c_P.

Regeneration is represented with a sudden vertical change in concentration show in **Fig. 3.11. a.**/. Slope of the water consumption line is unchanged after the

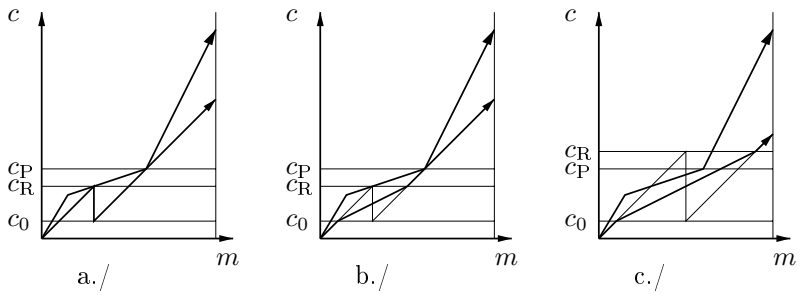


Figure 3.11: Regeneration below pinch

break because the flow rate is the same, but fresh water consumption is decreased. Concentration of regeneration and consumption rate are selected so that the line just touches LCC, i.e. a pich situation is formed. In spite of this, the consumption calculated this way is not at minimum.

To justify this claim, first construct a composite line from the sections before and after regeneration, see **Fig. 3.11. b.**/. This again shows a pinch situation, but if the regeneration concentration were higher at unchanged flow rate then the second section of the composite line would look as shown in **Fig. 3.11. c.**/, i.e. it would not reach LCC. It follows that smaller consumption can be achieved by letting the water reach higher contamination before regenerating it.

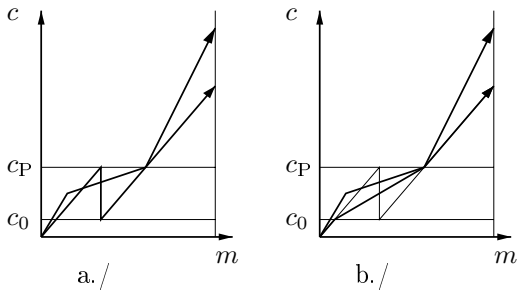


Figure 3.12: Regeneration at pinch

Minimum consumption if regeneration is done at pinch is shown in **Fig. 3.12**. In this case the composite line touches LCC. The water line is steeper than in **Fig. 3.11**, the consumption is $\frac{6}{7}$ of the former one.

Keeping this consumption of pinch regenerated case ($\frac{6}{7}$ of the earlier one) and regenerating above pinch, like shown in **Fig. 3.13. a.**/, the situation can be analyzed in the same way. The water line is subdiveded in **Fig. 3.13. b.**/, and

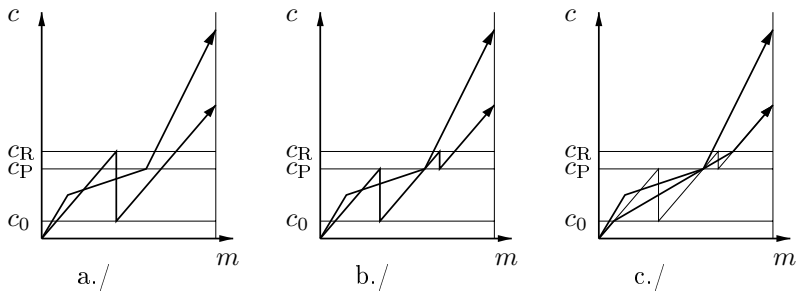


Figure 3.13: Regeneration above pinch

a composite line is constructed, shown in **Fig. 3.13. c./**. It can be seen that (i) water consumption remains at minimum, (ii) part of the regeneration above pinch is redundant, and thus it is even harmful because of useless costs involved.

Summary:

1. **Water consumption is at minimum if regeneration is done at or above pinch;**
2. **Regeneration above pinch does not decrease or increase consumption related to minimum, but is redundant and, therefore, is a waste.**

Minimum consumption can be calculated taking into account that flow rate of regenerated water equals that of water to be regenerated, i.e. their slopes are equal. It can be read from **Fig. 3.14** that

$$\frac{c_P - c_0}{m_P - m_R} = \frac{c_P}{m_R}$$

wherefrom m_R can be expressed, and then minimum calculated from the slope:

$$W_{\min} = \frac{m_R}{c_P}$$

Partial regeneration is also possible; in that case only a part of the water is regenerated, i.e. flow rate of the regenerated water is smaller than the minimum consumption, and the calculation is different.

3.2.5 Regeneration and recycling

Fresh water consumption can be decreased even more by using the regenerated water upstream.

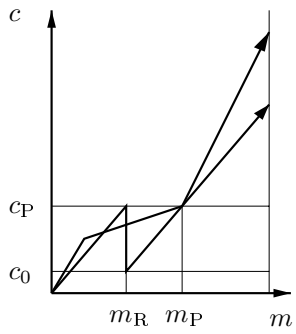


Figure 3.14: Determination of minimum consumption with regeneration

If concentration c_0 is equal to or smaller than minimum needed by any process then, in principle, no fresh water is necessary at all, but regenerated water is recycled in the process just like a regenerated solvent. Only replacement of inevitable losses sre then applied. Profitability of such a process is a question, however, because regeneration has also its costs. Economic optimum is usually looked for.

If c_0 is larger then at least so much fresh water is needed that assigend by the section LCC below c_0. As for the parts above c_0, the paragraphs above applies.

3.3 Distributed water treatment

3.3.1 Water treatment systems

Several differently contaminated waste water streams are formed in complex chemical sites. These streams has to be treated (cleaned) before emission. Treatment usually includes mechanicala dn chemcal pretreatment, biological treatment, and aftercleaning. Irrespectively to the particular details of these processes, the following general considerations are worth to take into account at assignments of treatment:

1. Contamination is decreased in subsequent treatment processes.
2. Treatment processes are normally characterized by a maximum inlet concentration $c_{\text{in, max}}$ above which the process does not work, and a minimum outlet concentration $c_{\text{out, min}}$ that can be produced, as a limit, by the process. The actual inlet and outlet concentrations can be somewhere in this range (**Fig. 3.15**).
3. Task of treatment is to reach a specified concentration c_0. If this is significantly larger than what can be reached by the process, i.e. the process is capable to produce significantly purer water, then it is sometimes more economical, and thus preferable, to clean only a part of the stream, reach

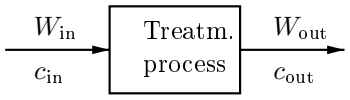


Figure 3.15: Treatment process in itself

better purity this way, and then mix it with the by-passing branch as is shown in **Fig. 3.16**.

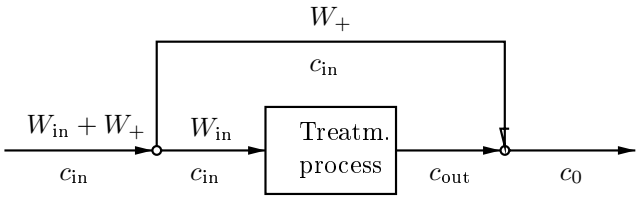


Figure 3.16: Treatment process with by-pass

The load on the treatment process is:

$$W_{\text{in}} = \frac{m}{c_{\text{in}} - c_{\text{out}}} \tag{3.1}$$

4. Annual cost of treatment depends on inlet and outlet concentrations, and mass load (flow rate):

$$K = f_{ccW}(c_{\text{in}}, c_{\text{out}}, W_{\text{in}})$$

Concentrations are specified for by-passing, so that the cost can be considered as function of the flow rate only. Let the amount of contamination to be removed be m, then flow rate can be calculated with (3.1). Thus the cost function is

$$K = f_W(W_{\text{in}})$$

5. Shape of the cost function $K = f_W(W_{\text{in}})$ depends on the particular treatment process. In can be incresing or decreasing along W_{in}, and can even be a function with minimum.

The streams forming in the chemical complex can be unified and treated together (*lumped treatment*), or can be fed to different points of the treatment system (*distributed treatment*). One may apply branchings and by-passes. *The problem considered in this section is assigning a minimum cost water treatment network.*

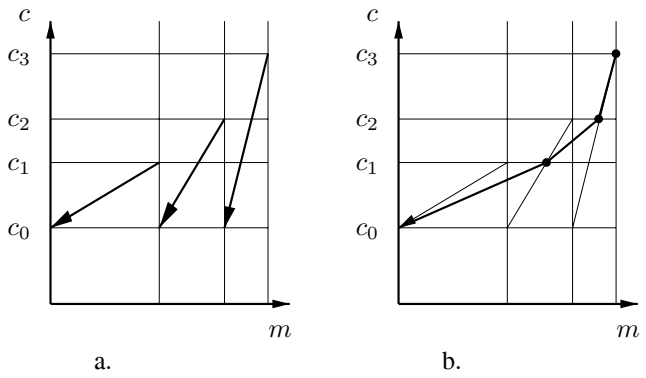


Figure 3.17: Waste water lines and their composite curve

3.3.2 Lumped treatment

Treatment tasks can be visualized in *C-m* diagram as is shown in **Fig. 3.17**. Several such lines (a/) can be unified to form a composite waste water curve (b/).

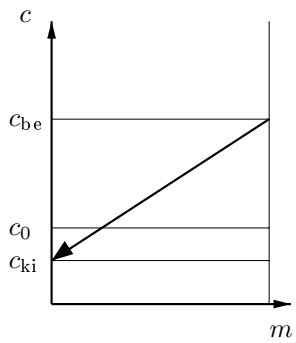


Figure 3.18: Treatment line

A single treatment can be visualized in *C-m* diagram in the same way as shown in **Fig. 3.18**. Concentration c_{out} is preferably smaller than the specified c_0, so that only a part of the water is treated and by-pass is applied.

At given waste water flow rate W, the higher concentrations, the smeller the costs are. Hence minimum cost is obtained by pushing the treatment line at high as feasible, i.e. up to pinch, as is ahown in **Fig. 3.19. a**. Applying higher concentrations, as in **Fig. 3.19. b**, the specified concentration c_0 could not be

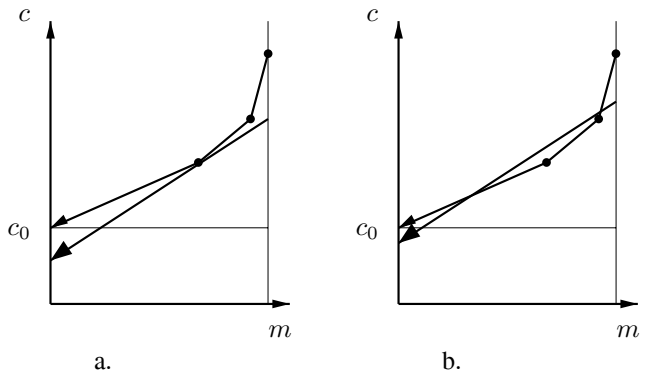


Figure 3.19: Treatment line at pinch (a.) and above (b.)

achieved after mixing.

Example:

Consider two waste water streams: one with $W_1 = 100\,\mathrm{t/h}$, $c_{1,\mathrm{in}} = 100\,\mathrm{ppm}$, and another one with $W_2 = 50\,\mathrm{t/h}$, $c_{2,\,\mathrm{in}} = 200\,\mathrm{ppm}$. Environmental specification is $c_0 = 50\,\mathrm{ppm}$. Thus, contamination to be removed from Stream 1 is 5 kg/h, from Stream 2 it is 7.5 kg/h, together 12.5 kg/h. Waste water lines and composite curve are shown in **Figs. 3.20. a. and b.** Concentration of the unified stream is $133\frac{1}{3}\,\mathrm{ppm}$. This point can be intercested from the vertical line over 12.5 kg/h by a line obtained with extending the lower section of the composite curve, as is shown in **Fig. 3.20. c.** If only a part, e.g. $x\,\mathrm{t/h}$ of Stream 1 is mixed to Stream 2, and th eother part by-passes the treatment process, then the concentration of the mixture is between $133\frac{1}{3}\,\mathrm{ppm}$ and 200 ppm, and its flow rate is not $150\,\mathrm{t/h}$ but $50 + x\,\mathrm{t/h}$ (**Fig. 3.20. d.**) Concentration c_{out} obtained after removing 12.5 kg/h contamination from the mixture can be easily calculated, but first consider what concentration is reached after removing the first 5 kg/h contamination, i.e. at $m = 7.5\,\mathrm{kg/h}$ on the axis. Concentration of the mixture fed to the treatment process is

$$c_{\mathrm{in}} = \frac{50 \times 200 + x \times 100}{50 + x}$$

Material balance is

$$5 = \frac{(c_{\mathrm{in}} - c_{7.5}) \times (50 + x)}{1000}$$

where $c_{7.5}$ is the concentration after removing 5 kg/h contamination, and 1000 is a factor connecting the units ppm, t, and kg. Hence

$$c_{7.5} = 100\frac{50 + x}{50 + x} = 100\,\mathrm{ppm}$$

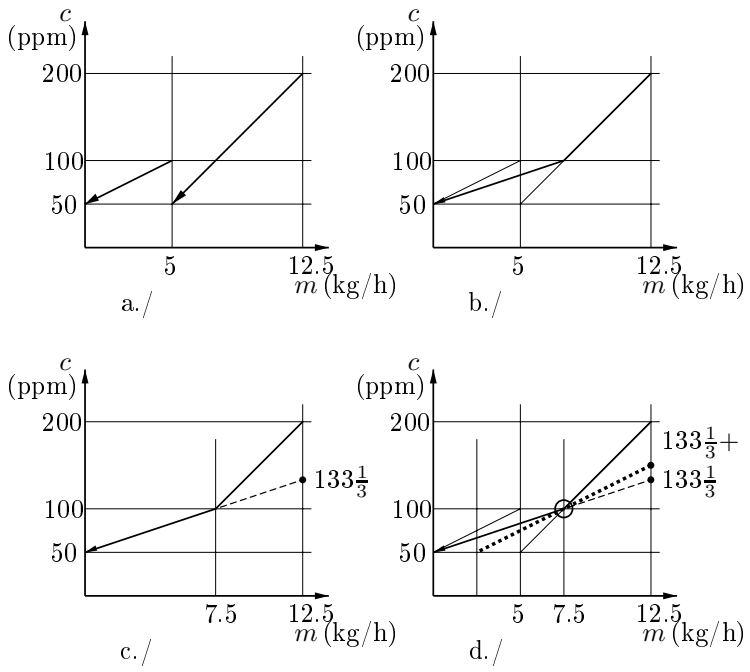


Figure 3.20: Example problem for treatment

i.e. the treatment line of such mixture always intersects point (7.5, 100). Since the smallest inlet concentration is $133\frac{1}{3}$ ppm, the treatment line never intersects but only touches the composite curve. If also Stream 2 were used only partially then both minimum and maximum of the inlet concentration would be lower, and the treatment line would not touch the composite curve. This is true generally, maximum of reachable concentrations is limited by pinch.

By varying the mixing and by-passing ratios, slope of the treatment line varies as well, it turns around the pinch point but remains touching it. This turning is constraind by:

1. Material balance (of mixing)
2. Minimum outlet concentration $c_{\text{out, min}}$ of the treatment process
3. Minimum inlet concentration $c_{\text{in, max}}$ of the treatment process

3.3.3 Distributed treatment

Optimum design depend on the shape of the cost function.

Decreasing cost function

With decreasing cost function the best case is maximum treated water flow rate, i.e. no by-pass. Slope of the treatment line is determined the lowest section of the composite curve. All the streams are mixed and treated together (**lumped treatment**).

Inreasing or minimum passing cost function

With increasing cost function the best case is minimum treated water flow rate, this belongs to the steepest treatment line. This slope is determined by either the minimum outlet concentration or the maximum inlet concentration. In this case *one has to treat differently the streams starting above, at, and below, pinch.*

All the streams starting above pinch are led to the treatment unit. All the streams starting below pinch by-pass the treatment unit. For removing the needed amount of contamination, working up the streams above pinch and a part of the streams strating at the pinch is just enough. Therefore part of the streams starting at the pinch by-passes the treatment, the other part is led to it. By-passing ratio is determined by material balance. This is (distributed treatment).

If the cost function goes through a minimum then optimal treatment line is somewhere between the two limits. In this case by-pass ratio is a decision variable for optimization.

Summary (see **Fig. 3.21**):

1. **Streams above pinch are treated.**
2. **Streams below pinch are not treated.**
3. **Streams started at pinch are partially treated.**

3.4 Several contaminating components

Dealing with several different contaminations with different specification is over the capability of such simple tools. Mathematical programming can be used in such cases.

3.5 Questions

1. Discuss analogy and differences between MAN and HEN, and their visualization.
2. What is water line, limit line, limit composite line?
3. How can minimum fresh water consumption be determined?
4. At which concentration is worth using regeneration?

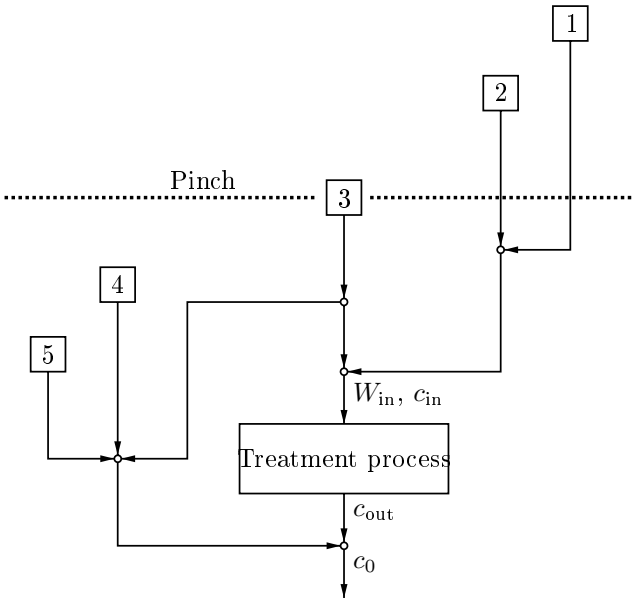


Figure 3.21: Distributed treatment

5. Which streams are to be treated and in what ratio if there are several streams with different concentration?

3.6 Suggested literature

El-Halwagi, M. M.; V. Manousiouthakis, V.: Synthesis of Mass Exchange Networks. *AIChE Journal*, **35**(8): 1233-1244 (1989).

Wang, Y. P. ; Smith, R.: Wastewater Minimisation. *Chemical Engineering Science*, **49**: 981-1006 (1994).

Hallale, N.: A New Graphical Targeting Method for Water Minimisation. *Advances in Environmental Research*, **6**(3): 377-390 (2002).

Chapter 4

Rectification systems without azeotropes

In this chaper we discuss separation sequences for zeotropic systems and energy saving rectification variants. We assume that volatility order of the components do not change with pressure, at least in the considered range.

Components can be ranked according to volatility. They are denoted with sequence numbers (1, 2, 3, ..., C), or with capital letters (A, B, C, ...), **always in order or decreasing volatility** (i.e. increasing boiling point).

4.1 Separation sequencing

Suppose the following:

1. Components of the mixture are to be separated in some *relatively sharp* sense: in high purity and/or with high recovery ratio.
2. We apply *conventional* distillation columns with one feed and two products each.
3. Separated components are *neighboring according to volatility order*.

The last assumption means that, for example, a mixture of components A, B, and C cannot be separated to a product mixture of A and B, and another product mixture of B and C. Such a mixture can be separated in two ways:

Direct cut: Distillate is enriched in component A, and bottom product is enriched in components B and C. We denote this cut as $[A/BC]$, i.e.:

$$[ABC] \Longrightarrow [A], [BC]$$

Indirect cut: Distillate is enriched in components A and B, and bottom product is enriched in component C. We denote this as $[AB/C]$, i.e.:

$$[ABC] \Longrightarrow [AB], [C]$$

As mentioned, **intermediate** case with significant distribution of a middle component between the produts, denoted as $[ABC] \Longrightarrow [AB], [BC]$, is excluded.

It may happen, however, that some *neighboring* components are not to be separated. In that case these components are lumped to form a pseudo component.For example, if butane isomers are not to be separated from a mixture of propane, n-butane, i-butane, and i-pentane, then they can be lumped to a pseudo-component 'butanes', and we consider pseudo three-component mixture.

A ternary mixture $[ABC]$ can be cut to two submixtures at 2 places, hence there ate 2 separation sequences:

1. **Direct sequence:** $[A/BC]$, $[B/C]$
2. **Indirect sequence:** $[AB/C]$, $[A/B]$

A four-component mixture $[ABCD]$ can be cut at 3 places, hence there are 5 separation sequences:

1. $[A/BCD]$, $[B/CD]$, $[C/D]$
2. $[A/BCD]$, $[BC/D$, $[B/C]$
3. $[AB/CD]$, $[A/B]$, $[C/D]$
4. $[ABC/D]$, $[AB/C]$, $[A/B]$
5. $[ABC/D]$, $[A/BC]$, $[B/C]$

The first of these are called **direct sequence**, the others are called **indirect sequences**.

Generally, a C-component mixture can be cut at $C-1$ places. In any separation sequence there are $C-1$ **cuts**. Number of possible separation sequences is $\frac{2(C-1)!}{C!(C-1)!}$, and the one which produces a most volatile component in the distillate of each cut is called *direct sequence*.

For selecting the optimal sequence, they have to be compared, but before comparison each must be optimized so that their best version is used in comparison. Optimizing a sequence is a complex task with the following decision variables: number of theoretical stages, place of feed, column pressure, reflux ratio, feed heat state to each column, as well as flow rates and purity of inter-column streams. Even if number of stages and pressures are temporarily fixed, this is a large problem, and to each decision variable array belongs a flowsheeting problem to be solved to obtain the unknown concentrations, boiling and condensation heat loads, internal vapor and liquid streams, column diameters. Finding exact solution of such an optimization problem is rather difficult in practice.

4.1.1 Heuristics and load factors

Based on long term engineering experience, *heuristics* (or, in other name, *rules of thumb*) can be set down with which near optimal separation sequence can be constructed in many cases, or at least the surely wrong sequences can be excluded. Several such rules can be found in the literature, a short summary is the following set:

1. In case of near equimolar feed composition and near equal relative volatilities of neighboring component pairs, the direct sequence is preferable.
2. In case of near equal relative volatilities of neighboring component pairs, the omponent(s) with large concentration is preferable to separate first. For example, let the composition of mixture $[A\ B\ C\ D\ E]$ be x $= [0.1, 0.5, 0.2, 0.15, 0.05]$, i.e. let B dominate the feed, then cutting first between B and C is preferable because only one component remains beside B in the distillate, whereas in any other case more components and in larger concentration remain beside it.
3. Cuts that are 'difficult' or expensive in any sense is better left dowstream, toward the end of the sequence. The cut can be difficult
   (a) if the relative volatility is near 1 (near boling components);
   (b) if either of the two components is to be produced in extreme purity or recovery (significantly higher purity or recovery than the others);
   (c) if the two components form azeotrope.
4. Near equal distillate and bottom product molar flow rates (approximately bisecting the feed) is preferable.

These rule are not always unequivocal. For example, if composition is x =[0.7, 0.2, 0.1] in mixture $[ABC]$, and relative volatilities are $\alpha_{A,B} = 1.03$, $\alpha_{B,C} = 1.7$, and purity and recovery specifications are approximately uniform then rule 1 prefers direct sequence, whereas rule 3 suggests indirect sequence.

The basic idea behind most rules is that more expensive cuts are to be left toward the end of the sequence. Cost of the cut is paid in number of stages (investment) or boiling and condensation (energy consumption), and there is a compromise between them. If an expensive cut between two components is made at the beginning then the other comopnents are involved in boiling and condensation, and that increases the costs. If, on the other hand, an easiy separation but with a large amount of component is left toward the end then this component is present in all the earlier cuts, increasing their costs.

Rule 1 is applied just because distillate is usually purer than bottom product, and with this selection temperature in the boiler becomes high at the end of the sequence, not earlier.

Uncertainty and contradictions of heuristics can be decreased by applying so-called *load factors*. Such factors are, e.g. the following expressions:

Application of rule 4 is helped by measures

$$\Delta_k^{(1)} = \frac{D_k}{W_k} \qquad (k = 1, 2, \ldots, C-1)$$

These numbers are to be determined to each possible sequences, and the one nearest 1 is selected.

Both variants of another factor measure difficulty of cuts:

$$\Delta_k^{(2)} = \left(\frac{F}{T_{HK} - T_{LK}} \right)_k$$

$$\Delta_k^{(3)} = \left(\frac{F}{\alpha_{HK/LK} - 1} \right)_k$$

where F is feed of the column, HK and LK are heavy key component and light key component, the neighboring components to be separated in the cut, T is normal boiling point, α is relative volatility. F is estimated to each columns in the sequence. The factors are summed up:

$$\Delta^{(n)} = \sum_{k=1}^{C-1} \Delta_k^{(n)}$$

and the sequence with the smallest sum is selected.

Even better is determining minimum vapor flow rate to each cut, and the sequence with the least total minimum vapor flow rate is selected. Minimum vapor flow rate can be easiest estimated with *Underwood*'s equations.

4.1.2 Sequencing with dynamic programming

Consider determination of cut places of mixtures with C, $C-1$, ..., 2 component mixtures, together with optimizing according to the othes decision variables (reflux ratio, number of stages etc.) as subsequent decisions. Then, e.g. in a four-component mixture problem, a desicion structure depicted in **Fig. 4.1** is obtained.

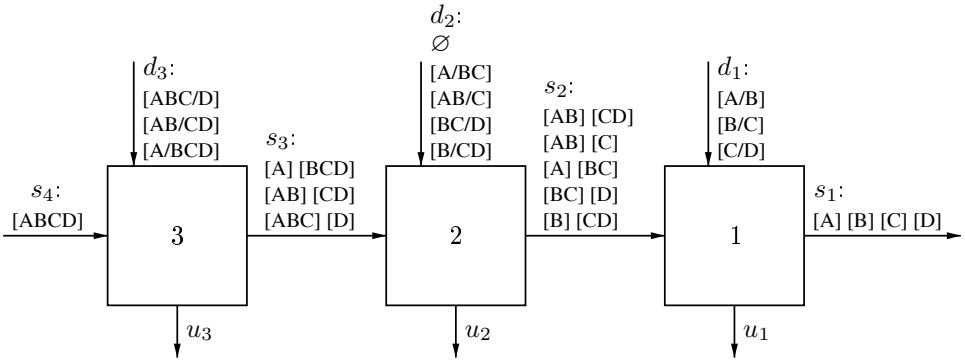


Figure 4.1: Decision sequence

Table 4.1: Example results. Step 1

Number of components	mixture	cut	cost of cut	cost of sequence	cost of optimal sequence
2	AB	A/B	0.752	0.752	0.752
	BC	B/C	0.899	0.899	0.899
	CD	C/D	6.026	6.026	6.026

Table 4.2: Example results. Step 2

Number of components	mixture	cut	cost of cut	cost of sequence	cost of optimal sequence
2	AB	A/B	0.752	0.752	0.752
	BC	B/C	0.899	0.899	0.899
	CD	C/D	6.026	6.026	6.026
3	ABC	A/BC	0.997	0.997	
		AB/C	0.990	1.742	1.742
	BCD	B/CD	1.331	7.357	
		BC/D	6.108	7.007	7.007

Table 4.3: Example results. Step 3

Number of components	mixture	cut	cost of cut	cost of sequence	cost of optimal sequence
2	AB	A/B	0.752	0.752	0.752
	BC	B/C	0.899	0.899	0.899
	CD	C/D	6.026	6.026	6.026
3	ABC	A/BC	0.997	0.997	
		AB/C	0.990	1.742	1.742
	BCD	B/CD	1.331	7.357	
		BC/D	6.108	7.007	7.007
4	ABCD	A/BCD	1.248	8.255	
		AB/CD	1.399	8.178	
		ABC/D	6.134	7.876	7.876

Main steps of the calculation are demonstrated, as example, in **Tables 4.1-3**. Minimum cost, here 7.876 is obtained in the last row. Cost of optimal separation sequence belonging to submixtures is written to row of the respective first cut, so that the optimal sequence can be backtracked: [ABC/D], [AB/C], [A/B].

In order to start the calculation in the first row, one has to estimate flow rates and composition of the feed streams at the end of the sequence. This is possible only if sharp purity and/or recovery specifications are applied. In that case perfect separation is a godd approximate.

4.2 Underwood equations

Constant relative volatilities are usually applied in the short-cut calculations. Definition of relative volatility is

$$\alpha_{i/j} \equiv \frac{K_i}{K_j}$$

where K is equilibrium ratio:

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

In order to simplify notation, we will not define relative volatility between neighboring components but all will be related to the same reference component *, so that a single index can be used to refer them:

$$\alpha_i \equiv \frac{K_i}{K^*}$$

where K^* is equlibrium ratio of the reference component.

This notation simplifies expressions with relative volatilities. For example, equilibrium vapor mole fractions can be expressed in the following way if liquid mole fractions are known:

$$y_i = \frac{\alpha_i x_i}{\sum_j \alpha_j x_j}$$

This can be justified as

$$y_i = \frac{K_i x_i}{1} = \frac{K_i x_i}{\sum_j y_j} = \frac{K_i x_i}{\sum_j K_j x_j} = \frac{\frac{K_i}{K^*} x_i}{\sum_j \frac{K_j}{K^*} x_j}$$

Approximating design of *binary* mixtures can be performed graphically applying the assumption of constant molar overflow and the well known *McCabe-Thiele* diagram. A similar design for multicomponent case can be performed if it is considered pseudo-binary. However, better calculation is achieved with *Underwood*'s equations. In a series of articles, *Underwood (1946-47-48)* generalized the operating line of *McCabe and Thiele* to multicomponent rectification, as follows.

Upper operating line is nothing else than a component balance around the upper (enriching) part of the column down to some stage $n-1$, including the borders,

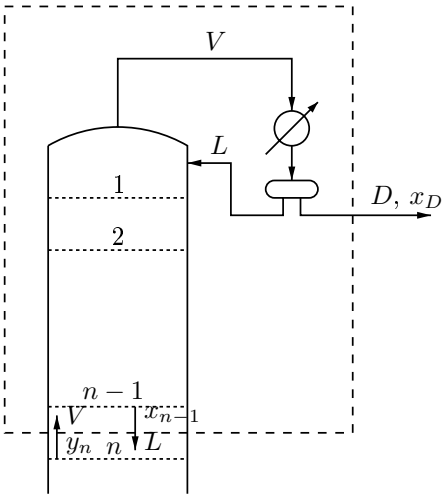


Figure 4.2: Upper encirclement

as is ahown in **Fig. 4.2**. Assuming constant molar overflow, and taken constant liquid and vapor flow rates L and V, respectively, the balance is written as

$$\frac{L}{V}x_{i,n-1} + \frac{D}{V}x_{i,D} = y_{i,n} = \frac{\alpha_i x_{i,n}}{\sum_j \alpha_j xj,n}$$

Multiply both sides with

$$\frac{\alpha_i}{\alpha_i - \varphi}$$

where φ is a temporarily unknown parameters ($\varphi \neq \alpha_i$), then sum up the equations obtained this way (i=1, 2, ..., C):

$$\frac{L}{V}\sum_{i=1}^{C}\frac{\alpha_i x_{i,n-1}}{\alpha_i - \varphi} + \frac{D}{V}\sum_{i=1}^{C}\frac{\alpha_i x_{i,D}}{\alpha_i - \varphi} = \frac{1}{\sum_{j=1}^{C}\alpha_j x_{j,n}}\sum_{i=1}^{C}\frac{\alpha_i^2 x_{i,n}}{\alpha_i - \varphi}$$

Chose φ so that

$$\boxed{\frac{D}{V}\sum_{i=1}^{C}\frac{\alpha_i x_{i,D}}{\alpha_i - \varphi} = 1} \tag{4.1}$$

Substitue and rearrange to get

$$\frac{L}{V}\sum_{i=1}^{C}\frac{\alpha_i x_{i,n-1}}{\alpha_i - \varphi} = \frac{\varphi}{\sum_{j=1}^{C}\alpha_j x_{j,n}}\sum_{i=1}^{C}\frac{\alpha_i x_{i,n}}{\alpha_i - \varphi} \tag{4.2}$$

Equation (4.1) has C zeroes, all are positive, and separated by relative volatilities α_i according to **Fig. 4.3**. Denote the zero falling between α_{i+1} and α_i by φ_i and denote he zero falling between 0 and α_C by φ_C. Each of the C zeroes satisfies

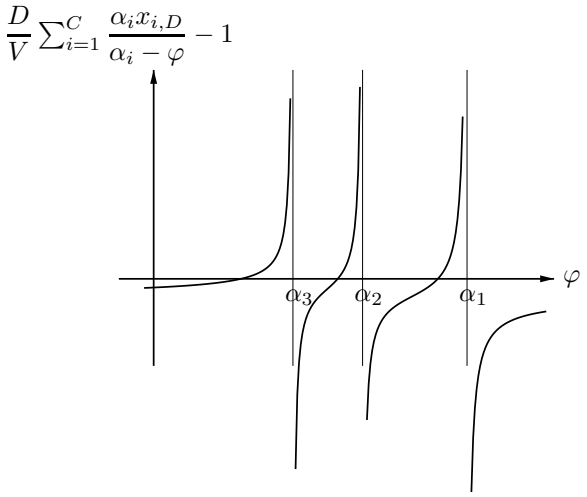


Figure 4.3: Shape of upper part's *Underwood* equation

equation (4.2). Select any two of them, e.g. φ_{LK} és a φ_{HK} gyököket, substitute them to equation 4.2, and then divide one with the other:

$$\left(\frac{\sum_i \frac{\alpha_i x_{i,n-1}}{\alpha_i-\varphi_{LK}}}{\sum_i \frac{\alpha_i x_{i,n-1}}{\alpha_i-\varphi_{HK}}}\right)=\left(\frac{\varphi_{LK}}{\varphi_{HK}}\right)\left(\frac{\sum_i \frac{\alpha_i x_{i,n}}{\alpha_i-\varphi_{LK}}}{\sum_i \frac{\alpha_i x_{i,n}}{\alpha_i-\varphi_{LK}}}\right)$$

The expressions in big brackets differ in stage numbers only, and are related to neighboring stages. Write up the same equation with stage numbers shifted by 1, and substitue. By repeating this action, one obtains an expression valid to the enrichng section:

$$\left(\frac{\sum_i \frac{\alpha_i x_{i,D}}{\alpha_i-\varphi_{LK}}}{\sum_i \frac{\alpha_i x_{i,D}}{\alpha_i-\varphi_{HK}}}\right)=\left(\frac{\varphi_{LK}}{\varphi_{HK}}\right)^{N_R}\left(\frac{\sum_i \frac{\alpha_i x_{i,f}}{\alpha_i-\varphi_{LK}}}{\sum_i \frac{\alpha_i x_{i,f}}{\alpha_i-\varphi_{HK}}}\right)$$

where az f is feed tray sequence number, N_R is number of stages in the enriching section. Both numerator and denominator of the left hand side are equal to V/D

according to equation (4.1), so that finally

$$1 = \left(\frac{\varphi_{LK}}{\varphi_{HK}}\right)^{N_R} \left(\frac{\sum_i \frac{\alpha_i x_i}{\alpha_i - \varphi_{LK}}}{\sum_i \frac{\alpha_i x_i}{\alpha_i - \varphi_{HK}}}\right)_f \tag{4.3}$$

Similarly in the stripping section chose the parameter so that

$$\boxed{\frac{W}{V'} \sum_{i=1}^{C} \frac{\alpha_i x_{i,W}}{\alpha_i - \psi} = -1} \tag{4.4}$$

and get:

$$1 = \left(\frac{\psi_{HK}}{\psi_{LK}}\right)^{N_S} \left(\frac{\sum_i \frac{\alpha_i x_i}{\alpha_i - \psi_{LK}}}{\sum_i \frac{\alpha_i x_i}{\alpha_i - \psi_{HK}}}\right)_f \tag{4.5}$$

where V' is vapor flow rate in the stripping section, and W is weight product flow rate. Zeroes ψ_i of equation (4.4) fall between α_i és α_{i+1}, and zero ψ_1 is higher than α_1, see **Fig. 4.4**.

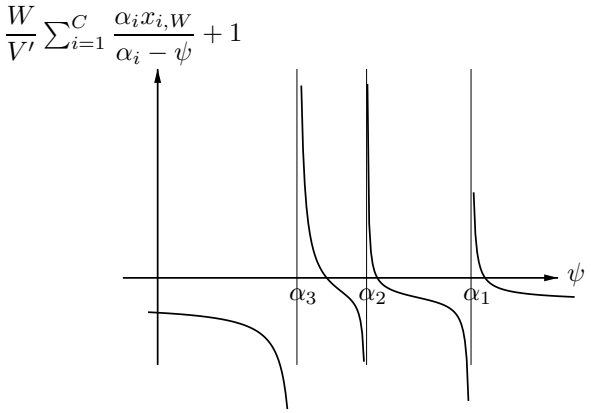


Figure 4.4: Shape of lower part's *Underwood* equation

Main use of *Underwood*'s equations is that minimum reflux ratio can be determined with them. There is a common root of equations (4.1) and (4.4) at minimum internal flow rates, this can be proved (though not easy). The common root falls between the relative volatilities of distributing key components HK and LK. Denote

this common root by ϕ, and substitute it to both equations:

$$V = \sum_{i=1}^{C} \frac{\alpha_i D x_{i,D}}{\alpha_i - \phi}$$
$$-V' = \sum_{i=1}^{C} \frac{\alpha_i W x_{i,W}}{\alpha_i - \phi}$$

Add them together:

$$V - V' = \sum_{i=1}^{C} \frac{\alpha_i (D x_{i,D} + W x_{i,W})}{\alpha_i - \phi}$$

Take into account that $Dx_{i,D} + Wx_{i,W} = Fx_{i,F}$ and $Fq = L' - L$, where q is heat state of the feed. Then its form is

$$\boxed{1 - q = \sum_{i=1}^{C} \frac{\alpha_i x_{i,F}}{\alpha_i - \phi}} \tag{4.6}$$

If composition and heat state of the feed is known then the key zero of (4.6) can be determined (numerically), and by substituting it to equations (4.1) and (4.4) one obtains minimum values of V and V'.

4.3 Reversible distillation

4.3.1 Thermodynamic efficiency of distillation

Materials of different composition mix spontaneously, but work has to be invested to separate a mixture of given composition to two parts with different compositions. Minimum work needed to eccomplish the separation under isothermal - isobaric environment is called **net work of separation**.

Net work of separation is intimately connected with entropy change of mixing. (Spontaneous) mixing involves entropy increase, whereas the reverse process, i.e. separation, is linked to entropy decrease.

If some stream with composition $\boldsymbol{x}_D$ and molar flow rate D is mixed isothermally with another stream of composition $\boldsymbol{x}_W$ and molar flow rate W to produce a third stream with flow rate $F = D + W$ and composition $\boldsymbol{x}_F$ where

$$x_{i,F} = \frac{D x_{i,D} + W x_{i,W}}{F}$$

then the involved entropy increase is

$$\Delta S_{\text{mix}} = -R \left[F \sum_i x_{i,F} \ln \gamma_{i,F} x_{i,F} - D \sum_i x_{i,D} \ln \gamma_{i,D} x_{i,D} - W \sum_i x_{i,W} \ln \gamma_{i,W} x_{i,W} \right]$$

For simplicity, non-ideality will be neglected here so that a simplified form is used:

$$\Delta S_{\text{mix}} = -R\left[F\sum_i x_{i,F}\ln x_{i,F} - \left(D\sum_i x_{i,D}\ln x_{i,D} + W\sum_i x_{i,W}\ln x_{i,W}\right)\right]$$

If that stream of F and $\boldsymbol{x}_F$ is separated to the same streams D, $\boldsymbol{x}_D$, W, and $\boldsymbol{x}_W$ then a process with opposite sign entropy change has to be performed:

$$\Delta S_{\text{sep}} = +R\left[F\sum_i x_{i,F}\ln x_{i,F} - \left(D\sum_i x_{i,D}\ln x_{i,D} + W\sum_i x_{i,W}\ln x_{i,W}\right)\right]$$

Since any complete real process involves entropy increase, according to Second Low, another process must be simultaneously taking place with entropy increase that at least compensates the decrease due to separation. In 'ideal' (reversible) case the two entropy changes are just equal in absolute value: $\Delta S_{\text{mix}} = -\Delta S_{\text{sep}}$. Entropy produced in this second subprocess 'drives' the entropy consuming separation subprocess. The minimum entropy production is $\Delta S_{\text{mix}} = -\Delta S_{\text{sep}}$, and the net separation work at a given temperature T is

$$P_{\text{net sep}} = \frac{-\Delta S_{\text{sep}}}{T}$$

Thermodynamic efficiency of the whole process is the ratio of netw separation work utilited in separation to the invested work:

$$\eta_E = \frac{P_{\text{net sep}}}{P_{\text{invested}}} \leq 1$$

The same ratio can be expressed in entropy terms, as well:

$$\eta_S = \frac{-\Delta S_{\text{sep}}}{\Delta S_{\text{driving}}} \leq 1$$

The driving subprocess in case of conventional distillation is degradation of heat power taken up at hight temperature in the boiler. The degraded energy leaves the system at lower temperature in the condenser. If the effects of preheaters and product coolers are neglected then the entropy production can be approximated as

$$\Delta S_{\text{driving}} \approx \frac{Q_{\text{cond}}}{T_{\text{cooling}}} - \frac{Q_{\text{boil}}}{T_{\text{heating}}}$$

Normally Q_{boil} and Q_{cond} are heat loads of the same magnitude, they are approximately equal, so one may speak about degradation of some common Q heat load between temperatures T_{heating} és T_{cooling}. These temperatures are *not* the boling and condensation temperatures but the temperatures of the utilities.

Approximate efficiency of a few industrial processes are collected in the next table:

Process	efficiency
Cryogenic air separation	0.18
Atmospheric and vacuum distillation unit	0.12
Separation of pyrolysis gases	0.05
Isotope separation	0.0001

4.3.2 Reversible distillation of binary mixtures

The reason of rather low efficiency of conventional distillation processes is that they are far from an ideal process. A reversible process is done in many very small steps, with infinitely small driving forces, through a series of equilibrium subprocesses.

Therefore, there is thermodynamic equilibrium everywhere in a reversible distillation process. Not only vapor and liquid from the same stage are in equilibrium but counterflowing vapor and liquid as well. This is (in theory) possible only if the operating line verywhere coincides with the equilibrium line. This involves a curved operationg line, and infinite number of stages at any concentration. Such a continuous curve operating line can be produced by infinitely small heat introduction (below feed) to, and heat removel (above feet) from, each stage. This continuous haeting and cooling involves a continuously changing vapor flow rate along column height, as is shown in **Fig. 4.5**.

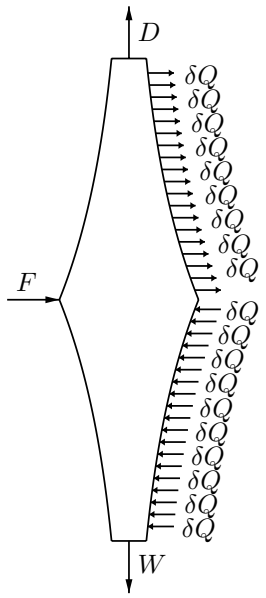


Figure 4.5: Reversible distillation column, infinite number of stages

Partial condenser and partial boiler must be applied to maintain equilibrium

at the ends of the column. Re-mixing should be avoided, therefore the feed must be directed to the stage of the same composition, and its heat state must be set accordingly. Finally, no pressure drop must be assumed on stages.

Table 4.4: Thermodynamic efficiency of rectification

modell	efficiency
$N = \infty$, reversible	1.00
$N = \infty$, adiabatic	0.35
N finite, adiabatic	0.83×0.35
finite, adiabatic, pressure drop	$0.97 \times 0.83 \times 0.35$
finite, adiabatic, Δp, ΔT	$0.825 \times 0.97 \times 0.83 \times 0.35$
total efficiency:	0.232

Fonyó published results of model calculations on rectifying an ethylene - ethane -propane - butane mixture under 30 bar, and cut at C_2/C_3. His results are shown in **Table 4.4**. The column is called adiabatic if heat exchange is done in the boiler and the condenser only, the column itself is insulated. The largest drop in enfficiency is from reversible to adiabatic configuration. Real reflux ratio is the second term, and the third largest factor is due to the finite temperature differences in the heat exchangers. Hydraulic effects may be neglected beside these larger effects.

4.3.3 Reversible distillation of multicomponent mixtures

Washing down all the components heavier than the lightest one is more difficult than washing down all the components heavier than the second lightest. If there are C components in the feed then washing down $C-1$ components from the top requires larger reflux ratio than washing down the heavier $C-2$ components only. In the same way, for producing a bottom product of pure C requies more boiling then producing a bottom prodoct of $C-1$ and C.

Generally, only one component may be removed from a column section to achieve reversible distillation. This can be (in theory) achieved as is shown in **Fig. 4.6** for a 4-component mixture. Each column should be imagined as is shown in **Fig. 4.5**, i.e. infinite number of stages and infinitely small heat load to or from each stage. There are two, four, and six column sections in column 1, 2, and 3, respectively, each serves for removing only one components from its product stream.

Each internal connection is organized so that vapor and liquid streams counter-flowing between the columns are in equilibrium. The vapor from the first column arrives to a stage of the second column where the vapor emerging from the lower stage is of the same composition. Hence the liquid flowing down from the stage above is again in equilibrium with the vapor, and there is equilibrium at top of the first column. Such an equilibrium is achieved in each intercolumn connection.

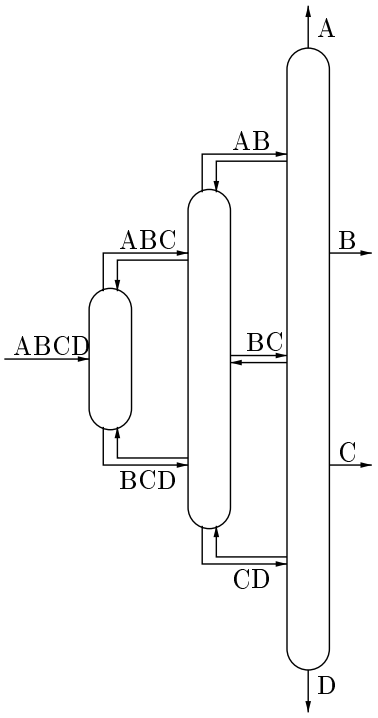


Figure 4.6: Reversible distillation structure, $N = \infty$, with many small heat loads

4.4 Energy saving distillation systems

Multistage distillation is the most energy consuming operation of chemical industry. $R+1$ times of distillate flow rate is to be produced at the top of the column (R is reflux rate) and, beside the frequent use of this separation process in all branches of the industry, almost the whole chemical industry starts with the atmospheric and vacuum fractionation of crude oil.

Reversible model of distillation is a target that helps us finding methods to decrease enery consumption of distillation processes. Most of the energy saving distillation system variants are based on principles borrowed from this model. Heat pump distillation processes are those variants, for example, which apply mechanical work to invest net work of consumption, instead of heat degradation.

4.4.1 More feeds and/or more product streams

Similar (but different) composition mixtures may be worth to separate in one column, but without unifying the feed streams. Such a case is shown in **Fig. 4.7 a.**/

assuming both feeds are at boiling point ($q = 1$). Even if there is just one stream to separate, its partial boiling or condensation can be used to produce two different feed streams. As a result, the operating line becomes nearer to the equilibrium curve, and minimum vapor flow rate is decreased. This is illustrated in **Fig. 4.7 b.**/ in case of binary mixture. Approaching the equilibrium curve with the operating line involves an increase in the number of theoretical stages, but minimum vapor flow rate decreases at given number of stages.

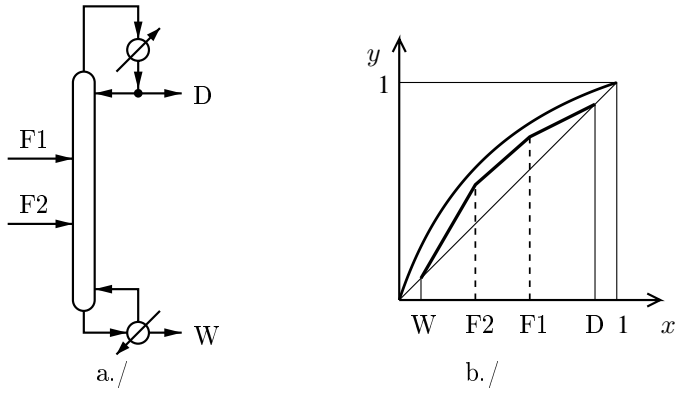


Figure 4.7: Operating line with two feeds

Several cuts can be performed in a single column in some cases, thus taking sideproducts beside distillate and weight products. Such a case is shown in **Fig. 4.8** (again with ($q = 1$)). The wing sections of the operating line are steep, but they belong to decreased product flow rates, therefore the vapor flow rate is decreased. Compositions z of the side products in the example are

$$z_2 = \frac{Dx_D + S1x_{S1}}{D + S1}$$
$$z_2 = \frac{Wx_W + S2x_{S2}}{W + S2}$$

4.4.2 Stepwise heat turnover

Instead of working with infinitely small heat turnovers at infintely many stages, a few number of heat introductions and heat removals along the column with finite number of stages can be applied in practice.

Below the feed the liquid flowing down from a stage is collected, a part of it is led out of the column to a **side bolier** where it is vaporized, and then the vapor is led back below the stage. This results is heating the stage (**Fig. 4.9 a.**/).

Above the feed taking out vapor and condensing it would be too difficult technically because of the pressure drop. Instead, liquid is taken out here, too, and

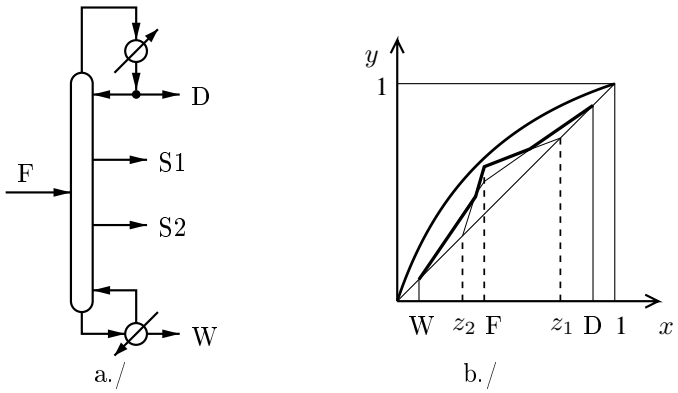


Figure 4.8: Working line with two sideproducts

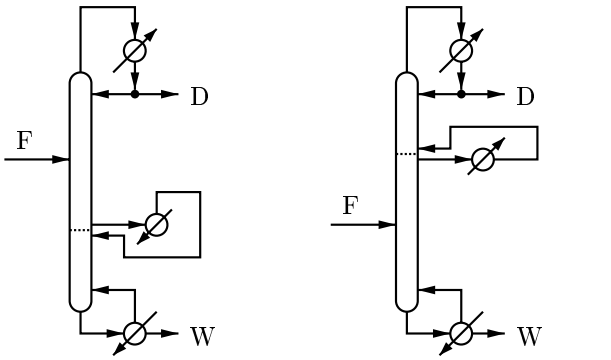


Figure 4.9: Stepwise heat turnover

cooled below its boiling point before leading back above the same stage. This is called **circulation reflux** or **pumparound**. In this way heat is taken from the stage, and part of the vapor emerging from below is condensed (**Fig. 4.9 b.**/).

Operating line is broken in both points, and approaches the equilibrium curve.

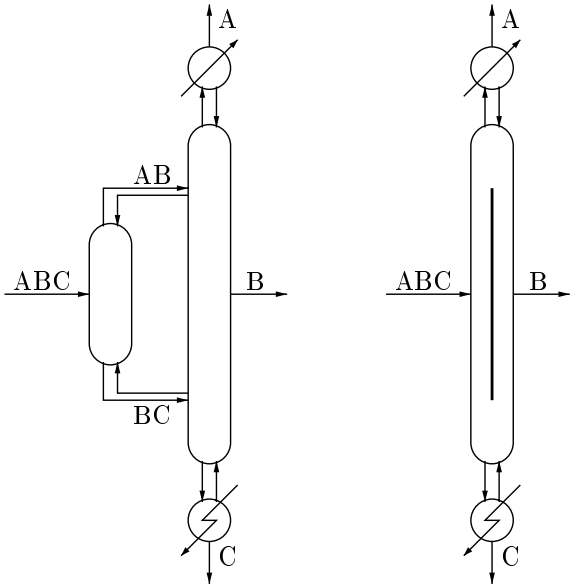
The heating and / or cooling utility applicable along the column is/are less expensive than those applied in the boiler and the condenser.

Side boliers are applied, for example, in natural gas fractionator processes. There are approximately 5 or 6 pumparounds in atmospheric and vacuum distillation plants in the oil industry.

4.4.3 Thermal coupling

These structures are based in the idea of **Fig. 4.6** but used with adiabatic columns with finite number of stages. Heat input and removal is applied at the ends of the final column only, so that the vapor produced in a single boiler carries heat to all columns, and except tha last column, each column gets its reflux stream from the next column. Thus, the columns a **thermally coupled**.

Such a full system for three component is called **Petlyuk system** after its inventor, shown in the left part of **Fig. 4.10**. This system can be achieved in a single vessel, too, and that unit, shown in the right hand side of the figure, is called **Kaibel column** or **dividing wall column**.



a./ *Petlyuk* b./ *Kaibel*

Figure 4.10: *Petlyuk* system and *Kaibel* column

Controlling such system is difficult. Liquid can be collected and distributed at will, but collecting and distributing vapor is technically out of question. Such systems are inflexible, and economical in a narrow range of feed composition only.

Partial coupling is, on the other hand, applied frequently. For example, side strippers applied in atmospheric and vacuum distillation units are such partially coupled systems, see **Fig. 4.11**. Its reverse, side enrichers are applied in cryogenic distillation units.

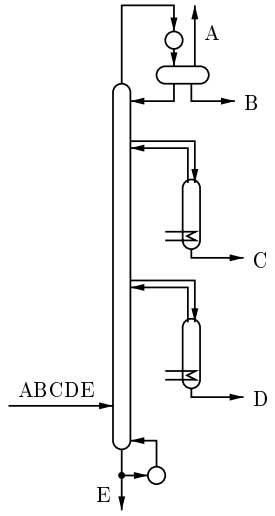


Figure 4.11: Side strippers

4.4.4 Energy integrated columns

Whereas process streams carry energy from one column to the other in thermally coupled systems, heat energy is transported through a wall of heat exchanger in energy integrated distillation systems. Such a heat exchanger serves simultaneously as a condenser and and as a boiler, as is shown in **Fig. 4.12**. For such connec-

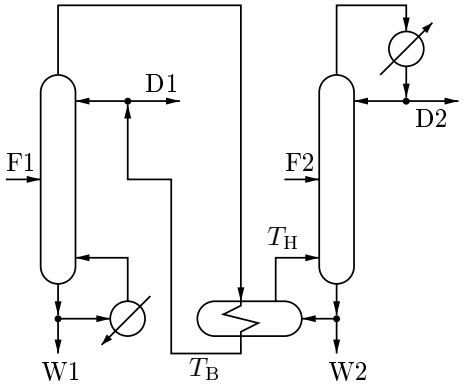


Figure 4.12: Integrated columns

tion, bubble point of the condensing vapor must be higher than dew point of the evaporating liquid.

Bubble point and dew point depend on pressure and composition. Composition is specified, but pressure is a design parameter, in hand of the designer or the operating engineerer. Pressure is constrained by the following factors:

1. Safety pressure limit in case of an erected column, or construction costs' pressure dependence in case of a planned column.
2. For making vacuum, one needs a vacuum punp or a barometric condenser, and auxiliary devices.
3. The material can be sensitive to heat, and the higher the pressure, the larger the temperature in the boiler.
4. Cost of the heating medium. For example, applying a furnace instead of a given state steam is expected much more expensive.
5. For applying air cooling (a relatively low cost solution), bubble point of the top vapor should not come below some limit temperature. For applying cooling water, there is another (smaller) limit. Keeping top temperature below this limit is possible with applying much more expensive cooling cycle only. Even the cooling cycles are more expensive at lower temperatures.

But if the conditions are satisfied then, in ideal case, approximately 50 % saving can be reached this way, and practically in may go up to 40 to 45 %. That is why these systems are also called *double effect distillation* systems, because the introduced heat is utilized twice: in both columns.

Energy integration can be applied even if a binary mixture is to be separated. One possible way is shown in **Fig. 4.13**. The left hand side column works at higher

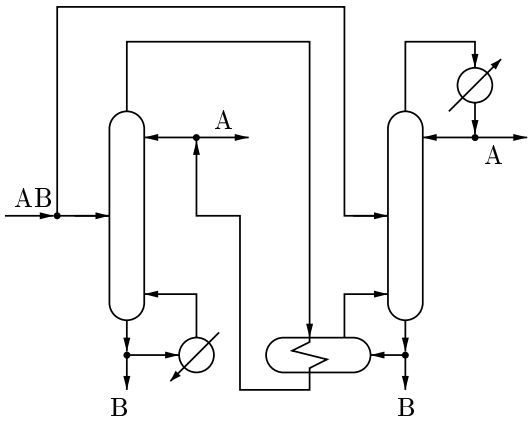
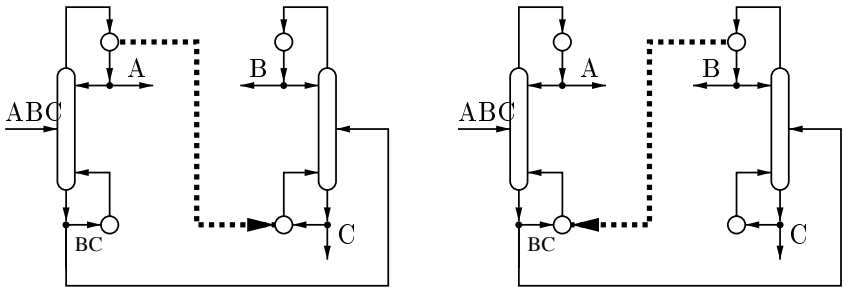


Figure 4.13: Two columns may be cheaper than one.

pressure, so that bubble point of A is higher than dew point of B. If the two columns could be operated in the same states then the feed could be branched to exactly equal streams and the savings were 50 %. However, volatilities and vaporization heats are different at different pressures, that is why the savings are a little bit smaller.

If integration is made between column of s separation sequence then there is a choice to integrate forward or backward along the sequence. Considering a ternary system, for example, the two columns cannot be on the same pressure in either case. If the system is of the direct sequence then the condenser–boiler unit works either between mixtures $[A] \rightarrow [C]$ or $[B] \rightarrow [BC]$ (Fig. **4.14**). The farther the



a./ Forward integrated b./ Backward integrated

Figure 4.14: Two kinds of integration in direct sequence

boling–condensation temperatures of the two mixtures, the larger shift in pressure is needed, hence the backward integration is expected be less expensive. In case of the indirect sequence, direction of integration is either $[AB] \rightarrow [B]$ or $[A] \rightarrow [C]$, and the forward integration seems preferable at the first look.

4.4.5 Heat pump distillation

Energy integration in a single column, i.e. boiling up a column with its own vapor is impossible because the top temperature is always smaller than the bottom, however is the column pressure shifted. In order to use the condensation heat of the vapor to boil up the column at the bottom, the pressure of the vapor have to be increased *outside* the column, or the heat must be mediated through an external heat carrier.

Heat pump distillation with closed cooling cycle

A closed cooling cycle can be attached to the distillation column as is shown in **Fig. 4.15**. Condensator of the distillation unit is, in the same time, the evaporator of the heat pump cycle; boiler of the column is the condenser of the heap pump cycle. Boiling point of the cooling medium, at the applied pressure, is smaller that bubble

point of the top product, therefore the medium is evaporated in the condenser, and becomes saturated vapor (1). This vapor is compressed, consuming P mechanical power, to superheated vapor at higher pressure (2). Such a pressure is to be reached that the medium's boiling point is higher than dew point of the bottom mixture. In the boiler of the column the vapor condenses while giving over its condensation heat. Thus it becomes saturated liquid (3). This liquid is let through a throttling valve into a space of the starting pressure. Its boiling point decreases; a small part of the liquid evaporates and the evaporation heat cools down the stream to the boling point valid at that pressure. Thus largest part of the stream is saturated liquid here (4). This liquid evaporates in the condenser, and so on.

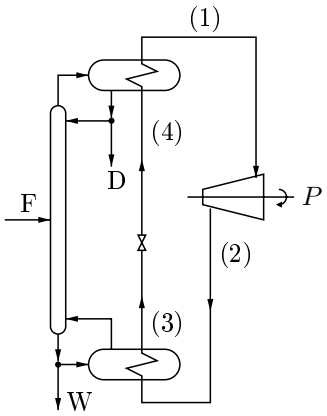


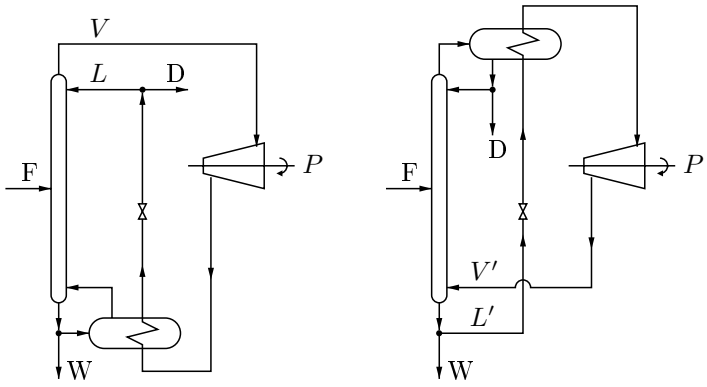
Figure 4.15: Heat pump distillation with closed cooling cycle

Vapor compression and bottom flash

Open cycle heat pumps are formed if a process stream is used for cooling medium, shown in **Fig. 4.16**. In case of **Vapor compression distillation** the vapor is compressed, condensed in the boiler, let through the throttling valve, and then branched to distillate and reflux stream. In case of **Bottom flash distillation** the liquid to be boiled up is first let through the throttling valve, evaporated in the condenser, compressed, and then the superheated vapor is led back to the column.

Energy balance, COP, and application

Heat pump distillation is more that 'double effet', more than 'tripple effect', and so on, because, in principle, no heat is anyhow applied. The process is not driven by heat degradation but by mechanical power. This power can be as small as 10% to 15% of the boiling or condensation load.



a./ Vapor compression b./ Bottom flash

Figure 4.16: Open cycle heat pump distillation variants

The introduced mechanical power is usually more than what is needed to maintain energy balance, and the extra heat is removed in a trim cooler. Thsi is not shown in the figures that concentrate on the main principles.

Condensation of the superheated vapor can be enhanced with spraying in liquid.

In case of open loop cycles, compression may case partial condensation, depending on the process material. (For example, saturated hydrocarbons over butane behaves so.) In this case superheating the vapor before compressing it is a feasible solution.

Coefficient of performance of a heat pump

$$COP = \frac{Q_{\text{forr}}}{P}$$

$$COP^* = \frac{Q_{\text{kond}}}{P}$$

is higher is the temperature gap is smaller. COP decreases radically with increasing gap. Moreover, larger and multistage compressors are needed. Cost of a compressor is comparable to that of a distillation column. Hence, its application is expected economical if near boiling components are to be separated.

In this case, however, their relative volatility is small, and many stages are built in the column. Consequently, the pressure drop between bottom and top is high, and the temperature gap increases with this pressure drop, making COP smaller, thus the mechanical energy consumtion higher, and even the compressor more expensive. That is why in such cases tray column structures are changed to structured packing that is characterized with small pressure drop per equivalent theoretical stages.

4.5 Questions

1. What is indirect distillation sequence?
2. List some heuristics and load factors for distillation sequencing!
3. When is dynamic programming applicable for distillation sequencing?
4. How to calculate equlibrium vapor composition if liquid composition and relative volatilities are given?
5. Where are located the seroes of *Underwood* equations?
6. What is net work of separation?
7. Describe the model of reversible distillation!!
8. What is circulating reflux or pumparound?
9. Draw a *Petlyuk* system and explain its concept! How can it be installed in a single column?
10. What is the ondition of inter-column energy integration? Haw can this condition be accomplished?
11. Draw all the variants of heat pump distillation!
12. Is there any connection between application of heat pump distillation and the column internals?

4.6 Suggested literature

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

Smith, R.: *Chemical Process Design and Integration*, John Wiley & Sons Ltd.m (2005).

Chapter 5

Optimization

A usual notation of a minimum problem is

$$\min_{\{\,\boldsymbol{x}\,\}} f(\boldsymbol{x},\boldsymbol{p})$$
$$\boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{0}$$
$$\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}$$

Here $f(\boldsymbol{x},\boldsymbol{p})$ is **objective function**, to be minimized over a **feasible domain** of solutions $\boldsymbol{x}$. Feasible domain is assigned by a system of conditions, here by **inequality constraints** $\boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{0}$ and equality constraints $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}$. Decision variables are listed below 'min' because the objective function can contain other arguments as well, for exampe $\boldsymbol{p}$ in our case. These other argumenst are parameters, their value is fix in a run of minimum search, but can be changed before a new search.

Any value of $\boldsymbol{x}$ is called a **solution**, even if it is outside of feasible domain. Solutions inside feasible domains are **feasible solutions**, outside they are **infeasible solutions**. A **feasible** solution $\boldsymbol{x}$ for which any other **feasible** solution $\boldsymbol{y}$ $\boldsymbol{f(x) \leq f(y)}$ is called **minimizer**, and the value of the objective in this point, $\boldsymbol{f(x)}$ is the minimum. We speak about maximizer and maximum in the opposite case, or we speak about extremal point and extremum generally.

There can be several minimizers, all producing the same minimum.

The definition of minimizers and minimum above are **global minimizers** and **global minimum**. In case of continuous domain and continuous objective, there can be **local minimizers** and several **local minima**. A **feasible** solution $\boldsymbol{x}$ is **local minimizer** if it is a global minimizer in its small neighborhood as feasible domain.

Any plan for the future is called a 'program', therefore planning is also called programming. Planning is always some kind of searching for optimal plans or optimal designs, extremum problems are also called *programming*.

Thus we speak about *linear programming* (denoted with LP) if all the objective and the constraint equations involve only linear functions of the design variables

$\boldsymbol{x}$. (The equations may be nonlinear in other variables, e.g. parameters $\boldsymbol{p}$.) We speak about *non-linear programming* (denoted with NLP) if any of the functions is not linear. If all the design variables are integer valued (integer domain) than it is *integer programming* (denoted with IP). We distinguish between linear and non-linear integer problems (ILP and NILP), as well as *mixed integer linear programming* (MILP) and *mixed integer non-linear programming* (MINLP) if both continuous and integer design variables take place in the problem. Optimization itself is also called *mathematical programming*.

Huge volumes and complete libraries deal with optimization, but here we discuss just the basic principles and mainly NLP problems.

5.1 Local minima over unconstrained domain

Only internal points are there in unconstrained domain, therefore all its global minimizers are as well local minimizers. (If the domain is constrained then a global minimizer falling to the border of the domain is not always a local minimizer.)

Most frequently the numerical search goes through a series of points $\boldsymbol{x^{(k)}, k = 0, 1, 2, \ldots}$ with decreasing objective: $\boldsymbol{f(x^{(k+1)}) < f(x^{(k)}) < f(x^{(k-1)})}$. Generally an *initial estimate* $\boldsymbol{x^{(0)}}$ is given and, based on some earlier estimate(s) $\boldsymbol{k-1.}$ or $\boldsymbol{k-1., k-2., \ldots}$ together with their objective(s) $\boldsymbol{f_{k-1}, f_{k-2}, \ldots}$, a direction $\boldsymbol{r}$ is calculated so that in that direction the objective ($\boldsymbol{f}$) is expected to decrease. Starting out from estimate $\boldsymbol{x^{(k-1)}}$, and measuring product of direction vector $\boldsymbol{r}$ with an arbitrary multiplier $\boldsymbol{\lambda}$, a new estimate is obtained:

$$\boldsymbol{x^{(k)} = x^{(k-1)} + \lambda r} \tag{5.1}$$

Particular methods differ in the form of function $\boldsymbol{r = r(x^{(k-1)}, x^{(k-2)}, \ldots)}$ and how multiplier $\boldsymbol{\lambda}$ is constructed. $\boldsymbol{\lambda}$ is constant in the simplest case. If the objective has a local minimum, and if condition $\boldsymbol{f_k < f_{k-1}}$ is achieved then such a series is expected to approach a local minimizer.

In this chapter we try using examples with two design variables only because their objective can be visualized with level lines in a plain. We also assume that minimization in a single dimension or a single direction (straight line) is not a great difficulty.

Relaxation

This is the simplest search method. Start from a point $\boldsymbol{x^{(k-1)}}$, first search for a minimum along $\boldsymbol{x_1}$ only, i.e. solve problem

$$\min_{\boldsymbol{x_1}} \boldsymbol{f(x_1, x_2^{(0)}, \ldots x_N^{(0)})}$$

to obtain approximate $\boldsymbol{x^{(1)}}$, then solve problem

$$\min_{\boldsymbol{x_2}} \boldsymbol{f(x_1^{(1)}, x_2, \ldots x_N^{(1)})}$$

etc., then after solving problem

$$\min_{x_N} f(x_1^{(N-1)}, x_2^{(N-1)}, \dots x_N)$$

start again with x_1.

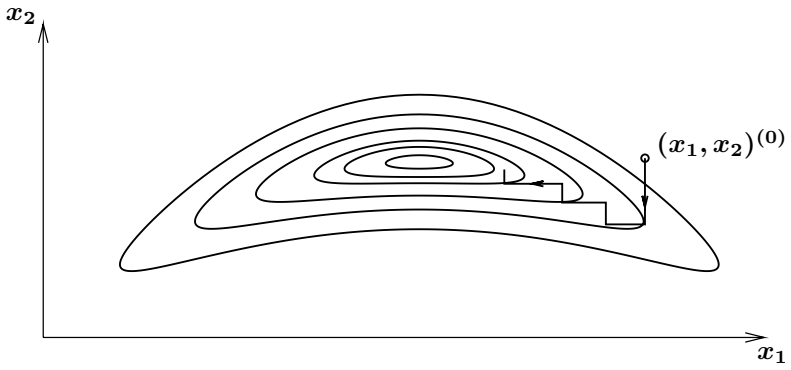


Figure 5.1: Relaxation method

For example, level lines of a two-variate objective are shown in **Fig. 5.1**. The first 7 steps of the search started from initial estimate $(x_1, x_2)^{(0)}$ is visualized. The steps are parallel with the axes because a search along an axis is made in each step.

Steepest descent

In this method first a preferable direction is determined and a one-dimension search is made along that direction in each step. The direction is determined by the *gradient* of the objective.

$$gradf \equiv \nabla f(\mathrm{x}) \equiv \begin{bmatrix} \frac{\partial f(\mathrm{x})}{\partial x_1} \\ \vdots \\ \frac{\partial f(\mathrm{x})}{\partial x_N} \end{bmatrix}$$

The gradient vector is directed to the steepest increase, hence the search is made in the opposite direction. The first three steps are shown in **Fig. 5.2**. The gradient is perpendicular to the actual level line.

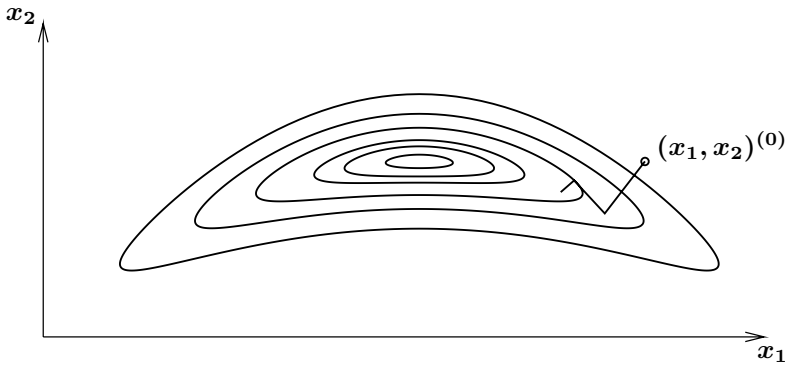


Figure 5.2: First 3 steps of method of steepest descent

Gradient methods

In principle more efficient search can be made if the search is not perfected in each actual direction but only a small step is made and then a new direction is determined in the new point. The step length can be assigned with a positive multiplier $\boldsymbol{\lambda}$, that can be constant in the simplest case:

$$\boldsymbol{x^{(k)} = x^{(k-1)} - \lambda\, gradf(x^{(k-1)})}$$

Results obtained with well selected $\boldsymbol{\lambda}$-s are shown in **Figs. 5.3. 5.4**.

If $\boldsymbol{\lambda}$ is very small then almost a continuous gradient curve is found, along which the minimum is searched for, but this procesure is rather slow because of the many steps and a gradient calculation in each of them. This is shown in **Fig. 5.5**.

If, on the other hand, $\boldsymbol{\lambda}$ is very small then the method cannot adapt to the curvature of the objective. Such a case is shown in **Fig. 5.6**.

The computation work (or time) can be decreased by changing $\boldsymbol{\lambda}$ in course. For this aim the objective is approximated with some simple function around the actual point, and then 'optimal' value of $\boldsymbol{\lambda}$ can be expressed analytically. Several such methods are known. The *conjugate gradient* methods assume quadratic approximation, and determine $\boldsymbol{\lambda}$ according to the difference between a $\boldsymbol{gradf(x^{(k-1)})}$ and $\boldsymbol{gradf(x^{(k-2)})}$. Such methods are applied in *quadratic programming* as well.

Simplex method and stochastic methods

Computation of the gradient, or even the objective itself, is too expensive in some cases. For example, if a full flowsheet is to be optimized then calculation of objective is equavalent to solving a flowsheeting problem in each point, and numeric

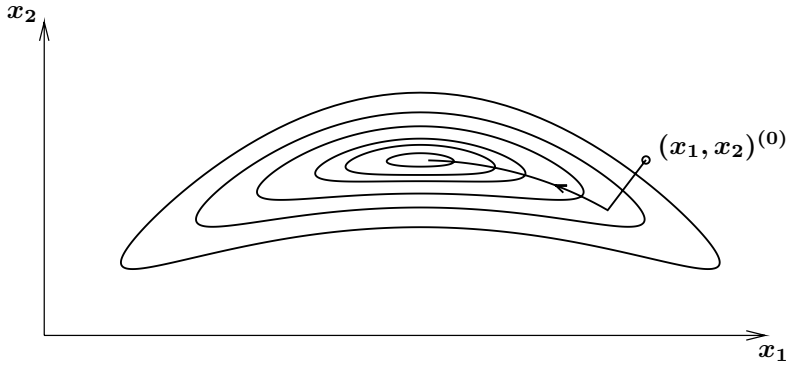


Figure 5.3: Gradient method with well selected multiplier –1

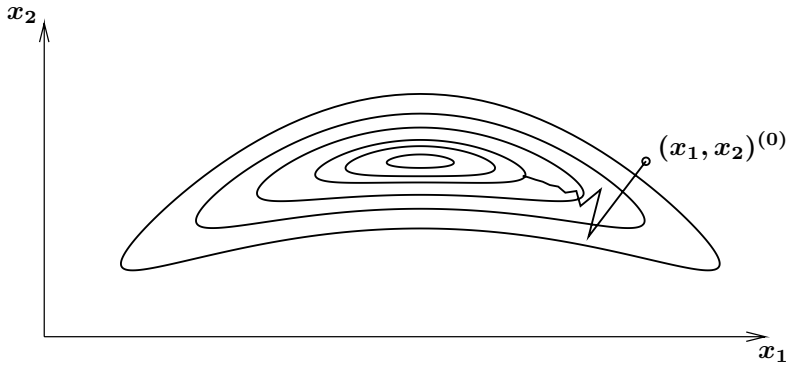


Figure 5.4: Gradient method with well selected multiplier –2

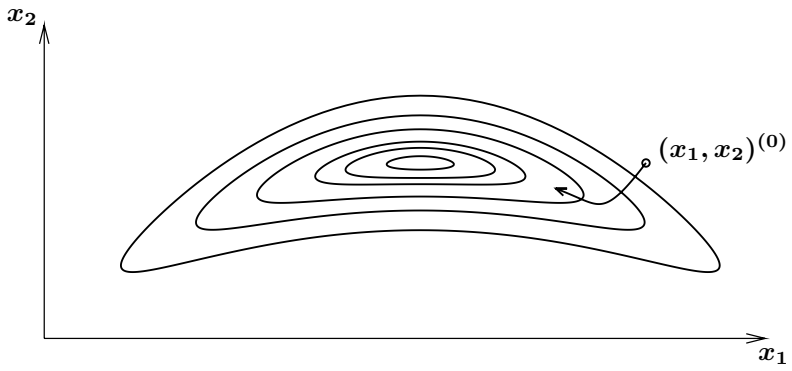


Figure 5.5: Gradient method with too small multiplier

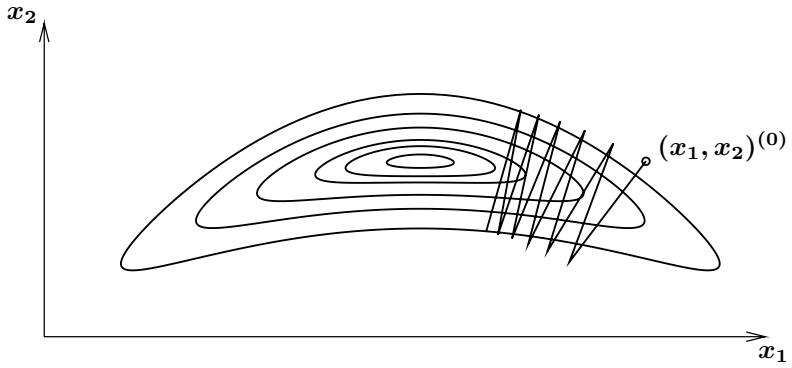


Figure 5.6: Gradient method with too large multiplier

differentiation ought to be applied that involves a huge number of flowsheet calculations in each point.

In such cases the number of evaluation points are preferably kept at possible minimum. A good search direction can be determined from even a small number of evaluations. The *simplex method* for $\boldsymbol{N}$ variables determines a direction based on $\boldsymbol{N+1}$ earlier evaluations that span a simplex in the $\boldsymbol{N}$-dimension space.

Select the point with largest objective, let it be e.g. point indexed $\boldsymbol{N+1}$, for simplicity, and project this point accross the the simplex face opposite this point. That face is spanned by the other $\boldsymbol{N}$ points of the simplex. Proper direction of projection can be obtained with some weighting of the points. For example, if the objective is always positive then a good center point can be:

$$\boldsymbol{x^{\diamond}} = \frac{\sum_{k=1}^{N} \boldsymbol{x^{(k)}} / \boldsymbol{f(x^{(k)})}}{\sum_{k=1}^{N} \boldsymbol{1} / \boldsymbol{f(x^{(k)})}}$$

Point $\boldsymbol{x^{(N+1)}}$ is projected through point $\boldsymbol{x^{\diamond}}$:

$$\boldsymbol{x^{(N+2)}} = \boldsymbol{x^{(N+1)}} + \boldsymbol{\lambda}(\boldsymbol{x^{\diamond}} - \boldsymbol{x^{(N+1)}})$$

Such a projection is shown in **Fig. 5.7**.

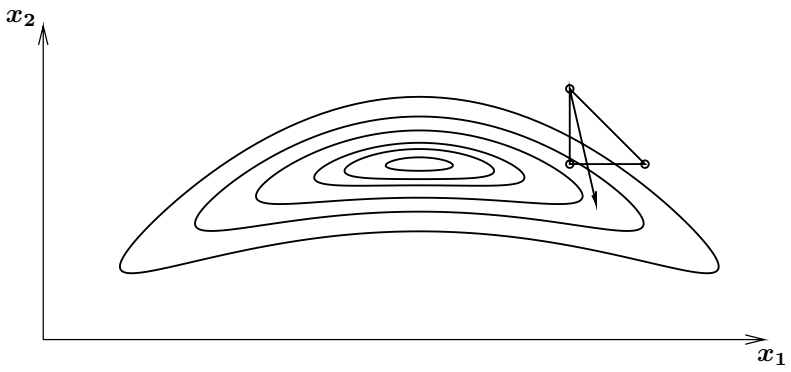


Figure 5.7: Simplex method

Search an also be made by randomly generating a few number of points around the actual point, evaluating them, and fitting an approximate function to this set. These *stochastic search methods* are preferred if there are many variables ($\boldsymbol{N}$ is large) and evaluation of just one point is very expensive.

5.2 Equality constraints

Equality constraints can be utilized to decrease the number of independent variables. For this aim, some variables must be expressed in function of others, or some varibale transformation must be used.

Let. for example, the search domain is Euclidean plain ($\{(\boldsymbol{x},\boldsymbol{y}), \boldsymbol{x} \in \mathbb{R}, \boldsymbol{y} \in \mathbb{R}\}$), and let a constraint be $\boldsymbol{h(x,y) = ax + b - y = 0}$, where $\boldsymbol{a}$ and $\boldsymbol{b}$ are real numbers. This is the equation of a straight line, i.e. the optimum is searched over a straight line. In this case substitute variable $\boldsymbol{y}$ in two-variate objective $\boldsymbol{f(x,y)}$ with expression $\boldsymbol{y = ax + b}$ to obtain a single variate objective $\boldsymbol{f^{\circ}(x)}$. Minimizer of this objective coincides with minimizer of the original objective, and the respective $\boldsymbol{y}$ equals to $\boldsymbol{y = ax + b}$.

Let, as another example, consider constraint $\boldsymbol{h(x,h) = x^2 + y^2 - r^2 = 0}$ instead, where $\boldsymbol{r}$ is a positive real number. This is the equation of a circle, an none of the two variables $\boldsymbol{x}$ or $\boldsymbol{y}$ can be expressed as a function of the other. You can apply, insted, the parametric form. Substitute expressions $\boldsymbol{x = r\cos(t)}$ and $\boldsymbol{y = r\sin(t)}$ into $\boldsymbol{f(x,y)}$ to obtain $\boldsymbol{f^{\circ}(t)}$, and optimize it over interval $\boldsymbol{[0, 2\pi)}$.

In many cases, however, the equations are so complicated that the above transformation methods cannot be applied. In such cases the method of *Lagrange factors* can be used. Let the objective be $\boldsymbol{f(x,y,\dots)}$, and let constraints be $\boldsymbol{h_1(x,y,\dots) = 0}$, $\boldsymbol{h_2(x,y,\dots) = 0}$, $\dots$ etc. According to the method of *Lagrange*, the problem

$$\begin{aligned} \min_{\{\boldsymbol{x},\boldsymbol{y},\dots\}} \ & \boldsymbol{f(x,y,\dots)} \\ & \boldsymbol{h_1(x,y,\dots) = 0} \\ & \boldsymbol{h_2(x,y,\dots) = 0} \\ & \vdots \end{aligned} \tag{5.2}$$

is transformed to the form of

$$\min_{\{\boldsymbol{x},\boldsymbol{y},\dots,\boldsymbol{\lambda_1},\boldsymbol{\lambda_2},\dots\}} \boldsymbol{f(x,y,\dots) + \lambda_1 * h_1(x,y,\dots) + \lambda_2 * h_2(x,y,\dots) + \dots} \tag{5.3}$$

It can be shown that local optimizers $[\boldsymbol{x},\boldsymbol{y},\dots]$ of problem (5.3) are also local optimizers of problem (5.2).

5.3 Inequality constraints

Aktív korlát és redukált irányok

If the actual point is feasible then the direction of search can be determined without respect to the inequality constraints. During the step or the search along that direction, however, one has to check if the new point is yet feasible. (I.e., one hase to check if all conditions $\boldsymbol{g(x) \leq 0}$ are satisfied. If yes then the serach can go

on without respect to the inequalities. Otherwise at least one of them is violated. Which constraint(s) is (are) violated is then has to be explored. Even if several

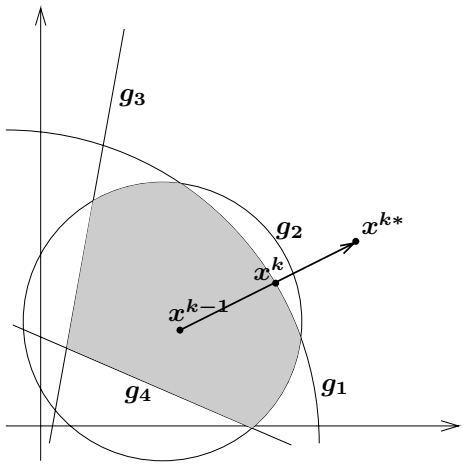


Figure 5.8: Inequality constraints

constraints are violated, there is one violated first or the largest extent, as a step is made from point $\boldsymbol{x^{(k-1)}}$ toward point $\boldsymbol{x^{(k*)}}$. For example, both inequalities $\boldsymbol{g_1}$ and $\boldsymbol{g_2}$ are violated in **Fig. 5.8** by the step to point $\boldsymbol{x^{(k*)}}$, but the feasible domain is bordered in this direction by $\boldsymbol{g_1}$, and this called the **active constraint**. The new approximation is selected as the intersection point of the active constraint and the straight line section connecting points $\boldsymbol{x^{(k-1)}}$ and $\boldsymbol{x^{(k*)}}$, i.e. the new point is $\boldsymbol{x^{(k)}}$.

The search is then has to go on along the active constraint as if it were an equality constraint, as fas as the actual search direction does not show toward the inside of the feasible domain and this constraint remains active.

Direction of search is determined in the original (feasible) domain. If this direction shows out from the feasible domain from its border then one possibility of search is projecting the direction to the border, according to **Fig. 5.9**. The direction obtained this way is called *reduced direction*. Point $\boldsymbol{x^{(k+1,*)}}$ is outside the feasible domain. Projection of this step to the active border assignes the new point $\boldsymbol{x^{(k+1)}}$.

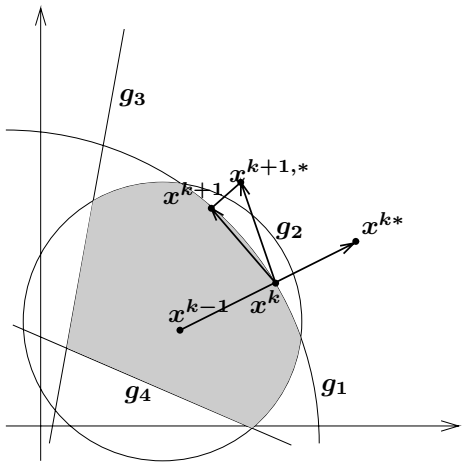


Figure 5.9: Inequality constraints and reduced direction

Penalty function

Another way of dealing with inequality constraints is incorporating a *penalty function* in the objective. Instead of solving directly the problem

$$\begin{aligned} \min_{\{x\}} \; & f(x) \\ & g_1(x) \leq 0 \\ & g_2(x) \leq 0 \\ & \vdots \end{aligned}$$

another one

$$\min_{\{x\}} f(x) + P_1(g_1) + P_2(g_2) + \dots$$

is solved where functions P_i are zero valued if their arguments are negative, and steeply increase in the positive domain. I.e., if the conditions are satisfied then the original problem is solved, but the more they are violated the larger increment is applied to the objective. If penalty functions P_i large (steep) enough then they guide the search toward the feasible domain.

5.4 Dynamic programming

Consider the problem:

$$\min_{\{d_1,d_2,\ldots,d_M\}} \quad U = \sum_{k=1}^{M} u_k(d_k, s_k, s_{k-1})$$
$$s_k = f_k(s_{k-1}, d_k) \quad (k = 1, 2, \ldots, M)$$
$$[0 = h_k(s_{k-1}, d_k) \quad (k = 1, 2, \ldots, M)]$$
$$[0 \geq g_k(s_{k-1}, d_k) \quad (k = 1, 2, \ldots, M)]$$

Such a problem occurs if a series of successive decisions connected with cause-effect relations are to be optimized, as is shown in **Fig. 5.10**. For example, optimal

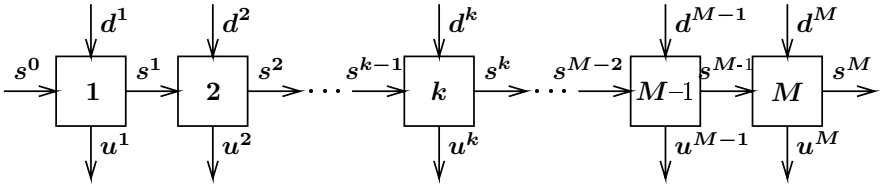


Figure 5.10: Dynamic programming

government as series of commands in discrete time moments is searched for. In this case the actual state of the system is described with array s_{k-1}, and the commands are described with array d_k of decision variables. The state s_k before the next decision is calculated by function $f_k(s_{k-1}, d_k)$. Effect of the decisions to the objective member is described by function $u_k(d_k, s_k, s_{k-1})$. The square bracketed terms $h_k(s_{k-1}, d_k)$ and $g_k(s_{k-1}, d_k)$ are listed just to illustrate that each subproblem has its own system of equality and inequality constraints. These are not essential part of the concepts of dynamic programming.

A chemical engineering example is a series of reactors. States are thermodynamic properties of the streams, and decision variables are reactor parameters like temperature, pressure, volume (or residense time). Tubular reactors can also be modelled as a series of mixed reactors. Another example is repeated extraction where decison variables are the applied entrainer flow rates.

Let the number of decison variables in each stage is N, then the total number of decision variables is $V = M * N$. This can be a large number. Computation work (or time) is rougly proportional to an expression ab^V where a and b are constant; that is the work or time increases exponentially with the number of decision variables. Dynamic programming can decrease this work if $M > 2$.

Discretize feasible domain of possible states s_{M-1} before the last decision with a mesh or grid, and consider discrete points as possible states: $s_{M-1,j}$ ($j =$

$\mathbf{1, 2, \ldots, L}$). For each of these points, solve the respective problem

$$\begin{aligned} \min_{\{d_M\}} \quad & u_M(d_M, s_M, s_{M-1}) \\ s_M &= f_M(s_{M-1}, d_M) \\ 0 &= h_M(s_{M-1}, d_M) \\ 0 &\geq g_M(s_{M-1}, d_M) \end{aligned}$$

Results of these minimizations are *suboptimal decisions* $d^*_{M,j}$ $(j = 1, 2, \ldots, L)$ and respective objective values $U^*_{M,j} \equiv u^*_{M,j}$ $(j = 1, 2, \ldots, L)$, belonging to the states $s_{M-1,j}$ $(j = 1, 2, \ldots, L)$. These values are stored in tables or in graphs as functions of states s_{M-1}: $d^*_M = D_M(s_{M-1})$, and $U^*_M = W_M(s_{M-1})$.

Now discretize feasible domain of possible states s_{M-2} before the one before the last decision with a mesh or grid, and consider discrete points as possible states: $s_{M-2,j}$ $(j = 1, 2, \ldots, L)$. For each of these points, solve the respective problem

$$\begin{aligned} \min_{\{d_{M-1}\}} \quad & u_{M-1}(d_{M-1}, s_{M-1}, s_{M-2}) + W_M(s_{M-1}) \\ s_{M-1} &= f_{M-1}(s_{M-2}, d_{M-1}) \\ 0 &= h_{M-1}(s_{M-2}, d_{M-1}) \\ 0 &\geq g_{M-1}(s_{M-2}, d_{M-1}) \end{aligned}$$

Results of these minimizations are *suboptimal decisions* $d^*_{M-1,j}$ $(j = 1, 2, \ldots, L)$ and respective objective values $U^*_{M-1,j} \equiv (u_{M-1,j} + U_{M,j})^*$ $(j = 1, 2, \ldots, L)$, belonging to the states $s_{M-2,j}$ $(j = 1, 2, \ldots, L)$. These values are stored in tables or in graphs as functions of states s_{M-2}: $d^*_{M-1} = D_{M-1}(s_{M-2})$, and $U^*_{M-1} = W_{M-1}(s_{M-2})$.

Go on this way from the last stage up to the first one. In any stage k, discretize the domain of states s_{k-1}, get discrete points $s_{k-1,j}$ $(j = 1, 2, \ldots, L)$, and solve the respective problems

$$\begin{aligned} \min_{\{d_k\}} \quad & u_k(d_k, s_k, s_{k-1}) + W_{k+1}(s_k) \\ s_k &= f_k(s_{k-1}, d_k) \\ 0 &= h_k(s_{k-1}, d_k) \\ 0 &\geq g_k(s_{k-1}, d_k) \end{aligned}$$

Results of these minimizations are *suboptimal decisions* $d^*_{k,j}$ $(j = 1, 2, \ldots, L)$ and respective objective values $U^*_{k,j} \equiv (u_{k,j} + U_{k+1,j})^*$ $(j = 1, 2, \ldots, L)$, belonging to the states $s_{k-1,j}$ $(j = 1, 2, \ldots, L)$. These values are stored in tables or in graphs as functions of states s_{k-1}: $d^*_k = D_k(s_{k-1})$, and $U^*_k = W_k(s_{k-1})$.

State s_0 before the first stage $(k = 1)$ is usually known, specified, and only one suboptimization task is to be solved. Once this is done, the optimum is already obtained: the optimum of this subproblem. Go back through the stages, i.e. in sequence of stages 1, 2, ..., M, and use the stored functions D_k and W_k, as well as

the specified functions f_k to find all the optimal (minimizer) decisions d_k^* together with interstage states s_k^*.

Computation work of the partical suboptimization problems is ab^N. Apart from stage 1, L such computation is performed in each stage; the total computation work is approximately $a\,[1+(M-1)L]\,b^N$. I.e., whereas the computation work of the original problem would be proportional to b^{M*N}, computation work of dynamic programming is only proportional to b^N. The increasing effect of factor $1+(M-1)L$ can be neglected beside the decreasing exponent.

5.5 Global optima

Convexity. Branch and Bound.

Some set Ω of a continuous space is called *convex set* if all the points of the straight line section connecting its any two points x_1 and x_2 ($x_1 \in \Omega, x_1 \in \Omega$) is also in this set: $\alpha x_1 + (1-\alpha)x_2 \in \Omega$, $(0 \leq \alpha \leq 1)$. For example, a (circular) disc and an ellips are convex sets. Any straight line is also a convex set. A polygon with an internal angle greater than 180 degree is not convex.

The straigh line section connecting two points is called their *convex linear combination*.

Similarly, a function $f(x)$ is called *convex function* if in all the points of the straight line section connecting its any two points x_1 and x_2 in its domain Ω ($x_1 \in \Omega, x_1 \in \Omega$) the function is defined and its value is not higher than the respective point of the straight line fitted on those two points (x_1, $f(x_1)$) and (x_2, $f(x_2)$). More precisely, let $f_1 = f(x_1)$, $f_2 = f(x_2)$ and $x = \alpha x_1 + (1-\alpha)x_2$, where $0 \leq \alpha \leq 1$, then the function is convex if $f(x) \leq \alpha f_1 + (1-\alpha)f_2$ is satisfied.

Hasonlóképpen, egy $f(x)$ függvényt *konvex függvénynek* nevezünk, ha összefüggő értelmezési tartományának bármely két x_1 és x_2 pontját összekötő egyenes szakasz pontjaiban a függvény értéke nagyobb, mint az (x_1, $f(x_1)$) és az (x_2, $f(x_2)$) pontokat összekötő egyenes szakasz f-koordinátája. Pontosabban legyen $f_1 = f(x_1)$, $f_2 = f(x_2)$, és legyen $x = \alpha x_1 + (1-\alpha)x_2$, ahol $0 \leq \alpha \leq 1$. Ekkor teljesülnie kell, hogy $f(x) \leq \alpha f_1 + (1-\alpha)f_2$.

The opposite is the concave function.

For example, exponential function is convex, whereas logarithmic function is concave. Sinusoid function is neither convex or concave, but is concave between 0 and π, and is convex between π and 2π.

Local minimizer of a convex function over a convex domain is also a global minimizer.

Separate the feasible domain Ω of the objective $f(x)$ into convex subdomains $\Omega = \cup_i \omega_i$ contacting along their borders only, and approximate the objective *from below* with appropriate **convex** functions $\phi_i(x)$ over the subdomains. Minima of these convex functions can be easily determined, and they are surely not greater than minima of the original function $f(x)$ over the same subdomain ω_i. Evaluate

$f_i \equiv f(x)$ in an arbitrary point x of subdomain ω_i, and compare it to minimum ϕ_j^* of function $\phi_j(x)$ over another subdomain ω_j. If $f_i < \phi_j^*$ then the objective function in any point of subdomain ω_j is surely ot greater than f_i, therefore subdomain ω_j does not contain a global minimum, and can be excluded from the search.

The **Branch and Bound** method is based on the above idea. The domain is subdivided to two or more subdomains, and the objective function is approximated with some convex function from below, and subdomains are exlcuded with comparisons. The remaining subdomainds are further subdivided to smaller subdomains, and the procedure is repeated until the subdomain(s) containing global minimizer(s) is (are) small enough to consider it (them) identical to its (their) central point(s).

A widely applied variant of **Branch and Bound** considers domains and subdomains containing infeasible solutions as well. Minimizer of the convex lower bound functions have to be checked if they are feasible or not, because exclusion of a subdomain may be based on feasible local minima only. However, this method is rather successful because local minimization can be done over well behaving domains.

5.6 Integer variables

In some cases the integer variables can be embedded into a continuos domain. That is, the whole numbers are considered as real ones, and optimization is made over a real domain, and finaly a whole number near the continuous minimizer is selected. This method is, however, subject to large errors. For example, continuous minimizer is between 3 and 4 (e.g. 3.26), but the integer optimizer is 13 (because of the actual shape of the objective function), and the minimum also can be largely different.

The non-continuous decision variable is not always a numeric quantity. For example, color, or construction material has to be selected, or device type. Even in these cases one can number the possible selections: 1, 2, 3, etc., but their order is arbitrary, any other numbering can be chosen.

In case of several *independent* non-continuous variables, the number of feasible selections is the product of those of the individual variables. For example if the possible values of variable x_1 is 8, and that of x_2 is 5, then the number of possible couples (x_1, x_2) is $\mathbf{8 \times 5 = 40}$. The number of combinations with several non-continuous variables can be very large.

Natural way of solving such problems is evaluating the combinations one by one, and then comparing them. This is called **enumeration**. Since this is a combinatorial task, it is also called **combinatorial explosion**.

If there are continuous variables beside the non-continuous ones then a continuous optimization has to be performed for evaluating each combination.

In some (frequent) cases the number of combinations is so large that merely collecting and listing them is impossible in practice. In some cases their listing and evaluation is possible but so tedious that impractical. For example, it would last 10

Table 5.1: Implicit enumeration

variant	approx. minimum	exact minimum
1	126.0	1277.2
2	127.8	2013.1
3	674.2	1321.8
4	1021.9	**1253.6**
5	1203.0	1843.2
6	**1297.5**	-
7	1506.2	-
8	1506.8	-
9	2017.4	-
etc.	etc.	etc.

years in the modern computers. That is why other methods have been developed to avoid enumeration.

Implicit enumeration

Consider a problem in which the number of possible combinations is 100 000, and optimizing one of them according to the continuous variables takes approximately 30 minutes in average. Enumeration would then last for 5 net years. Suppose, however, that you have found a fast approximation that reliably is always lower than the real minimum of the actual combination. This approximation can be computed within 0.1 second. You compute this approximation to each of the combinations in 1.5 hour.

Sort the combinations according the computed approximation in decreasing order, and then start really optimizing them in this order, taking 30 minutes to each of them. If the approximation is not too bad then you will get, as a result, a table similar to that shown in **Table 5.1**. The real optima are everywhere greater than those obtained with approximation.

You may stop in this particular case after minimizing the 5-th variant, because minima of any other variant is reliably greater than the smallest one of hitherto found minima. Optimum is 1253.6, this belongs to variant 4. Thus the global minimum is found in $\mathbf{1.5 + 5 \times 0.5 = 3}$ hour time. Overhelming majority of the variants are not optimized exactly, but approximated only. We say they are *enumerated implicitely*.

Variants of implicite enumeration are applied widely for solving integer and mixed integer programming problems, usually together with Branch and Bound. That is, not a single linear sorting is applied but branching structure is build up.

Here we show the solution of an assignement problem with this method. *Note: There are much more effective methods for solving this kind of problem, for example the so-called Hungarian method. Here we apply Branch and Bound with implicit*

enumeration in order to demonstrate its principles. Incorporation of infeasible solutions in the domain is also demonstrated in this example.

The task is seleting monogamic couples between four women $\{\boldsymbol{W_1}, \boldsymbol{W_2}, \boldsymbol{W_3}, \boldsymbol{W_4}\}$ and four men $\{\boldsymbol{M_1}, \boldsymbol{M_2}, \boldsymbol{M_3}, \boldsymbol{M_4}\}$. We have four decision variables: $\boldsymbol{M_1}$, $\boldsymbol{M_2}$, $\boldsymbol{M_3}$,and $\boldsymbol{M_4}$, and each of them can take as value the name of a women, i.e. one of $\boldsymbol{W_1}$, $\boldsymbol{W_2}$, $\boldsymbol{W_3}$, and $\boldsymbol{W_4}$.

If any man could select a woman independently of the selection of the others then there would be $\mathbf{4^4 = 256}$ possible selections. Most of these selections are, however, ineasible because of the specified monogamy. There are only $\mathbf{4 \times 3 \times 2 \times 1 = 24}$ feasible selections.

The task is selecting marriages with minimum total number of quarrels in 10 years. The expected number of quarrels in 10 years couple by couple is collected in the cost matrix $\boldsymbol{C}$ shown in the upper part of **Fig. 5.11**.

Any four selections (any value of array $[\boldsymbol{M_1}, \boldsymbol{M_2}, \boldsymbol{M_3}, \boldsymbol{M_4}]$) assigns four different entries of the cost matrix $\boldsymbol{C}$, and the objective to be minimized is the sum of these four values.

Costs are approximated from below so that some constraints are neglected. The less constraints applied, the smaller the objective can be.

All the constraints are omitted for start: each man can choose any women, irrespectively to the selection of others. The minimizer in this case is obtained by selecting a woman to each man inpdependently. This is shown in Matrix 1 in **Fig. 5.11**. The estimated cost is 41, and this is the smallest possible value, not any other selection can be smaller.

If this selection were feasible (F) then the problem would be solved but, unfortunately, this selection is infeasible (I) because woman $\boldsymbol{W_4}$ is selected by two men. This selection is infeasible, but feasible selection might be found if some constraint are taken into account. *Separate* the solutions according to the possible selections of man $\boldsymbol{M_1}$, and take into account his selection by the other men while keep neglection between selections of the three other men.

Possible selections of $\boldsymbol{M_1}$ are shown in the second row of the figure: matrices 2, 3, 4, and 5. The constraints taken into account are indicated with hatching.

If $\boldsymbol{M_1 = W_1}$, i.e. if $\boldsymbol{M_1}$ engages $\boldsymbol{W_1}$, then the independent optimal selections of the other three men from the remaining women is, by chance, a feasible solution as indicated with a letter F under matrix 2. This is the first found feasible solution, with objective 133. Until better feasible solution is not found, this is considered as *candidate* solution.

If $\boldsymbol{M_1 = W_2}$, i.e. if $\boldsymbol{M_1}$ engages $\boldsymbol{W_2}$, then the independent optimal selections of the other three men from the remaining women is, by chance, again a feasible solution as indicated with a letter F under matrix 3. Since this solution is better (104) than the hitherto best feasible solution, this is the candidate from this moment.

If $\boldsymbol{M_1 = W_3}$, i.e. if $\boldsymbol{M_1}$ engages $\boldsymbol{W_3}$, then the independent optimal selections of the other three men from the remaining women are infeasible, as indicated with a letter I under matrix 3. Its objective, 138, is greater than that of the candidate.

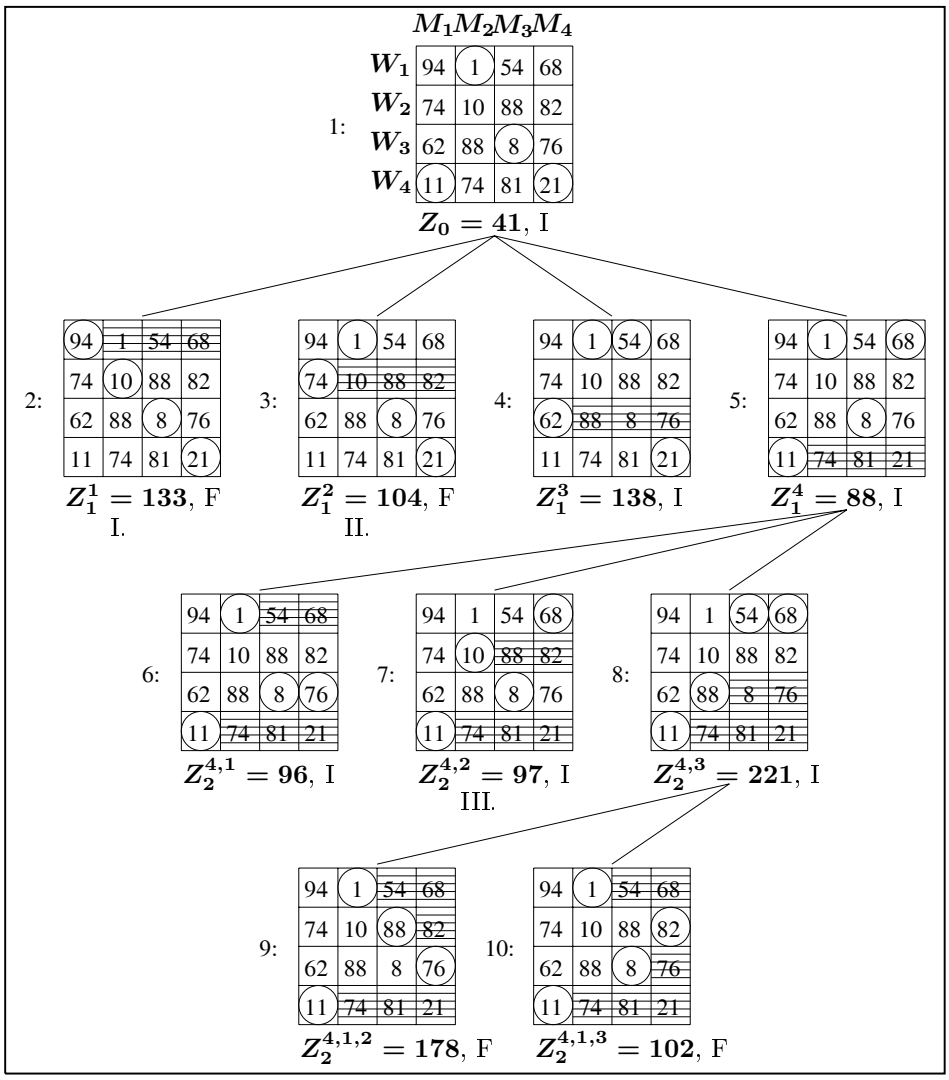


Figure 5.11: Solving assignment problem with *Brand and Bound*

There is no chance that some other, feasible, selection by the remaining three men could be smaller (since this objective is a lower bound to them) than that of the candidate. Therefore this branch of the tree can be closed (excluded).

If $\boldsymbol{M_1 = W_4}$, i.e. if $\boldsymbol{M_1}$ engages $\boldsymbol{W_4}$, then the independent optimal selections of the other three men from the remaining women are again infeasible, as indicated with a letter I under matrix 4. Its objective, 88, is *smaller* than that of the candidate (104). *There is* chance that we find a feasible solution with objective below 104 by expounding this branch accourding to the seletions of $\boldsymbol{M_2}$. This is shown in the third row of the figure.

After selecting $\boldsymbol{M_1 = W_4}$, $\boldsymbol{M_2}$ can choose from the remaining three woman only. These selections are fixed in matrices 6, 7, and 8. The constraints taken into account are again indicated with hatching. The remaining two men can select from the remaining two women only.

As selections of more and more men are taken into account, and the problem tree is branched accordingly, more and more constraints, and finally all of them are taken into account.

Optimal solution of our problem is shown in matrix 7; this is the third candidate. Altogether 10 solutions have been evaluated instead of all the feasible 24 solutions, though the 10 evaluated solutions are not seleted from the set of feasible solutions but from all the 256 solutions including infeasible ones, too.

Metaheuristics

There are several metaheuristic methods to find or at least approach global optima of non-continuous problems. The best known metaheuristics are *Simulated annealing*, *Genetic algorithms*, *Ant colony optimization*, and *Tabu search*. Two of them are discussed shortly.

Simulated Annealing. *Annealing* is an operation in metal working. Metal is first heated up and then slowly let it cooled down with controlled temperature decrease. If the metal is cooled down suddenly then there is no time for achieving thermodynamic equilibrium, and microcrystalline structure with high energy levels (stress, tension) is formed in the material. The controlled, slow coolig makes approaching the equilibrium possible, and one can get rid of extra stress. Heating up is applied to get out of the local minimizer of energy and then find a global minimizer.

According to the theory of statistical ensembles, probability $\boldsymbol{P(s)}$ of a microstate $\boldsymbol{s}$ is proportional to expression $\boldsymbol{\exp(-\frac{E(s)}{kT})}$ where $\boldsymbol{E}$ is energy of the microstate, $\boldsymbol{T}$ is thermodynamic temperature, and $\boldsymbol{k}$ is a universal constant. The higher $\boldsymbol{T}$ is, the higher energy microstates have the greater probability.

Role of energy $\boldsymbol{E}$ is taken up by objective $\boldsymbol{f(d)}$ in the case of *Simulated Annealing*, and values of decision variables $\boldsymbol{d}$ take the role of states. At some arbitrariliy taken $\boldsymbol{T}$, actual values of $\boldsymbol{d}$ are randomly changed and then the change is accepted with probability $\boldsymbol{P(d) \sim \exp(-\frac{f(d)}{kT})}$, and rejected with probability $\boldsymbol{(1 - P(d))}$.

In this way the the objective mainly decreases but it can also increase with some probability. This process is repeated many times to let the energy decrease, but then $\boldsymbol{T}$ is decreased according to a prescribed program. The smaller $\boldsymbol{T}$ is, the smaller is the probability of increasing the objective value, but any small probability for objective increase gives a chance to jump out from a local minimum and go toward a global equilbrium state.

Genetic algorithms. these methods imitate natural selection of biological evolution. Decision variables are considered as genom of an individual in a *population*. The methods deal with the population in which the individuals are competing each other. Objective function is evaluated and considered as *fittness* of the individuals. After evaluating fittness of individuals, the algorithms calculate a next *generation*.Components of the individuals with higher fittness are inherited in the next generation with greater probability.

A rough sketch of the algorithm is:

1. An initial population is randomly generated.
2. Calculate fittness values. Calculate their minimum, maximu, average, or perhaps other statistical features.
3. Generate a new population (next generation). *This is the essential step, detailed below.*
4. If no change is found in the statistical features or if a number of steps is over then stop.
5. Repeat from step 2.

Generating the new population is the key step. The usual calculation of inheritance and change of genoms are the following:

- **Reproduction.** A few number (e.g. 3) of the best individuals are taken over to the next generation without any change. The others form a pool from which stochastic selection is performed.
- **Mutation.** A prescribed number of individuals are randomly selected, and their data are changed with some prescribed low probability. The simplest change is flipping bits (0/1).
- **Crossover.** This mode imitates sexual reproduction. Individual couples are randomly selected, and two children are computed in a way of swapping some genom parts. Let the genom sequence of individual $\boldsymbol{i}$ be $\boldsymbol{A_i}$, $\boldsymbol{B_i}$, $\boldsymbol{C_i}$, ..., $\boldsymbol{Z_i}$, that of individual $\boldsymbol{j}$ be $\boldsymbol{A_j}$, $\boldsymbol{B_j}$, $\boldsymbol{C_j}$, ..., $\boldsymbol{Z_j}$, and select randomly a place to cut them, e.g. between $\boldsymbol{K}$ and $\boldsymbol{L}$, and link them together again but in the wrong way: Let one of them be $\boldsymbol{A_i}$, $\boldsymbol{B_i}$, ..., $\boldsymbol{K_i}$, $\boldsymbol{L_j}$, $\boldsymbol{M_j}$, ..., $\boldsymbol{Z_j}$, and the other one $\boldsymbol{A_j}$, $\boldsymbol{B_j}$, ..., $\boldsymbol{K_j}$, $\boldsymbol{L_i}$, $\boldsymbol{M_i}$, ..., $\boldsymbol{Z_i}$.

The 'random' selections refer to fittness, i.e. individuals with higher fittness are selected with greater probability to reproducate their genom. Weak individuals, i.e. their genom parts, drop out with high probability, but there always some chance for them to be reserved, so that 'good' parts of even weak individuals can slowly spread in the population.

The method is sensitive to population scale and probability parameters .

5.7 Questions

1. Explain the problem of minimization, its terms, and various problem types!
2. What is local minimum, local minimizer, and global minimum?
3. Explain the difference between relaxion method, steepest descent, and gradient method!
4. How can equality constraints be taken into account in optimization?
5. How can inequality constraints be taken into account in optimization?
6. How can integer variables be taken into account in optimization?

5.8 Suggested literature

KORN, R. C. ÉS KORN, J. M.: *Mathematical Handbook for Scientists and Engineers*. McGraw-Hill, New York, 1961.

WILDE, D. J.: *Optimums Seeking Methods*. Prentice-Hall, New York, 1964.