Friedrich-Alexander-Universität
Naturwissenschaftliche Fakultät

Lecture Script

Reactive Transport in porous Media

Summer semester 2010, 2011, 2016, 2017, 2019, 2021, 2022, 2023

Prof. Dr. Serge Kräutle

at

Lehrstuhl für Angewandte Mathematik (Modellierung und Numerik)

AM1

— Version dated 2023 April 29 —

Friedrich–Alexander–Universität Erlangen–Nürnberg

Table of Contents

This is the translation of the lecture notes that I had originally written in German language several years ago. For the translation I used automatic translation by *DeepL*. I had a swift review of what the software produced – mainly because the script is typed in LaTeX, and DeepL doesn't work so well with texts containing LaTeX–commands. Hence, the translation will be rather clumsy. If you find incomprehensible sentences or strange formulations, you may let me know and I can improve the translation.

Serge Kräutle, April 2022

In theory, there is no difference between theory and practice.
In practice, there is.

(unknown author)

1 Introduction: What is a Porous Medium? Examples, Applications.

A *porous medium* consists of a *solid skeleton* (also: solid matrix or soil matrix) and a *pore space*. The structure is of a very small scale compared to the total size of the porous medium. The pores contain one (or several immiscible) fluid(s), and the fluid may move. By fluid we mean both liquids and gases. The most important example of a porous medium is soil or rock. The pore space is usually fully or partially filled with water and/or air. But also e.g. the combinations oil/water ($\rightarrow$ oil production) or also CO_2/water ($\rightarrow$ storage of CO_2 in the soil to reduce the greenhouse effect) are considered. The pore space should be connected to allow movement of the fluid(s).

Please note: I have prepared a set of slides (which lecture participants can find on StudOn) that includes images of various porous media, and in which various technical applications and questions are motivated. I have not included this part in the lecture notes you are reading right now.

Examples

porous medium	application
(1) soil/rock/sand, with groundwater or deep water	(a) contaminated sites (contamination of soil/aquifers by, among other things, industrial operations, accidents, waste dumps, etc.). water from, among other things, industrial operations, unäll, müll dumps). (b) Exploitation of oil/gas fields (c) CO_2 storage in the subsurface (d) nuclear repositories (e) salinization of agricultural soils/aquifers
(2) (historical) buildings	weathering: sulfur in air/rainwater can cause conversion of stone to gypsum
(3) armored concrete	intrusion of water and oxygen causes corrosion of steel
(4) porous burners (in vehicles?)	heat generation by gas (hydrogen) combustion without open flame
(5) filters, filtering facilities (activated carbon, flow through sand fillings)	removal of impurities, by chemical reactions and/or adsorbtion on pore walls.
(6) agricultural soils	investigation of the complicated intertwined processes that effect formation, preservation, amendmend, deterioration of agricultural soils

Topics (1) and (6) were/are in the research focus of parts of the chair *Angewandte Mathematik (Modellierung und Numerik)* (formerly *Angewandte Mathematik 1*).

It is a very characteristic property of porous media that different (spatial) scales (= orders of size) play a role: First of all there is the *pore scale* or *microscale* (about $10^{-4}\text{m} = 10^{-1}\text{mm}$ or smaller), then there is the scale on which the soil properties[1] are 'somewhat constant', let's say about 10^{-2}–10^{2}m. [2] And then there is the *macroscale* (in geo-/hydrology called *field scale*), which describes the size of the considered area, and which is $10^2 - 10^4\text{m}$ or even more.
Not only the large variation of spatial scales, but also different time scales may play a role for processes in porous media: Chemical reactions of sustances that are dissolved in the water and in the same pore may take place on the scale of seconds, but for the formation/dissolution of minerals or for nuclear storage sites times scales of thousands of years may be relevant. Processes that live on different space- and timescales may affect each other. This is a typical difficulty when dealing with porous media. When we want to apply a numerical scheme to the equations, how can we resolve processes that act on such different scales? The next chapter may give some ideas: The derivation and usage of 'effective' equations.

2 Modeling of the flow regime

If we wanted to resolve, in the sense of a numerical simulation (FDM,FEM), the pore scale of a $10^3 \times 10^3 \times 10^1\text{m}^3$ large area $\Omega \subset \mathbb{R}^3$, with a mesh size $h = 10^{-5}\text{m} = 10^{-2}\text{mm}$, we would have to use $10^8 \cdot 10^8 \cdot 10^6 = 10^{22}$ nodes/elements, which is impossible (not to mention the difficulties of generating a three-dimensional FEM grid usable for numerical purposes in pore space in the first place). Therefore we need a *macroscopic* description of the processes, a description by *'effective'* ('averaged') equations. The simplest approach is the volumetric *averaging*. This is the one we will use in the following. (Another approach is the so-called *homogenization*, also called *upscaling*; this is mathematically more precise, but makes the assumption that the pore space is periodic, see Exercise.)
We choose a so-called *representative elementary volume (REV)*, which is a cube or sphere with diameter significantly larger than the pore scale but smaller than the correlation scale. We average entities över REVs around each point x:

[1] By soil properties one can think here of porosity and hydraulic conductivity, see later

[2] It is possible to grasp this length scale more precisely in the framework of *stochastic* models of the soil. The soil property φ at each point x of the domain is then a random variable with some random distribution, and additionally it is required that the *correlation* of the random variables $\varphi(x)$ and $\varphi(y)$ is a (generally monotonically decreasing) function of $|x-y|$, i.e. the farther points are from each other, the less correlated is the soil property at the points. A parameter that specifies how fast this correlation function decays is called *correlation length*. As 'length on which the soil properties are somewhat constant' one can use this correlation length

Porosity $\omega(x) := \frac{V_{por}(x)}{V_{tot}}$, where V_{tot} is the total volume of the REV and $V_{por}(x)$ is the volume of the pore space intersected with the REV. Thus, $\omega \in (0,1)$ holds.
Note: In some situations ω can also be a function of x and t, for example, if mineral dissolution or erosion changes the the pore space, or if biofilms form on the surfaces of the pores, or if pressure changes so much that it can affect the pore space.
(Volumetric) water content $\theta(t,x) := \frac{V_W(t,x)}{V_{tot}}$, where $V_W(t,x)$ denotes the volume of water within the REV. Thus, we have $V_W(t,x) \leq V_{por}(x)$, that is, $0 \leq \theta(t,x) \leq \omega(x)$. If $\theta(t,x) = \omega(x)$, we call the porous medium at location x at time t *saturated*, otherwise *unsaturated*.

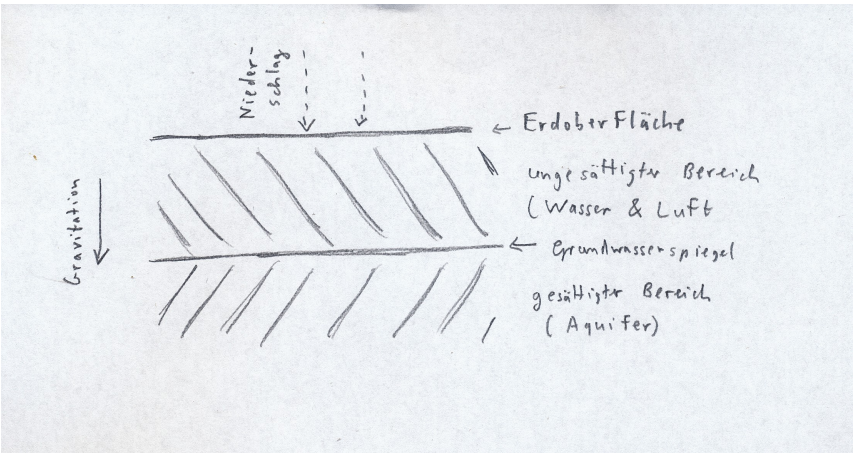


Abbildung 1: Typical behavior of the water content in the uppermost soil layers in temperate climates, schematic.

There are two ways to define the 'average' of the velocity of the fluid phase via the REV: On the one hand, one can consider the so-called *seepage velocity* $\vec{u}(t,\vec{x})$; this is the mean velocity of the water, averaged (only) over the $V_W(t,\vec{x})$-volume of the REV. On the other hand, there is the so-called *Darcy velocity*[3] $\vec{v}(t,\vec{x})$; for this, the velocity is averaged over the total volume of the REV (i.e. pore space and solid matrix). Hence it holds true

$$\vec{v} = \theta\,\vec{u} + (1-\theta)\,\vec{0} = \theta\,\vec{u}\,.$$

The seepage velocity is relevant, for instance, if we want to describe the speed at which the front of a pollution proceeds (at least if we neglect diffusion and if there are no sorption processes). The Darcy velocity, on the other hand, is needed to describe the volume flux through a macroscopic surface F (i.e., the volume of water moving through the surface per unit time): The volume flux is described by the surface integral (of 'second kind', i.e., for vector-valued functions) over $\vec{v}$ over the face F; the Darcy velocity thus represents the *volume flux density*. If one is looking for the *mass* flux or the *mass* flux density, one must multiply $\vec{v}$ by the density ρ, thus consider $\rho\,\vec{v}$ or $\int_F \rho\,\vec{v}\cdot\vec{n}\,do$.

[3]Henry Darcy, 1803-1858, hydraulic engineer, Dijon 1856.

For any macroscopic volume $V \subseteq \Omega$ we assume *mass conservation*. The change in water mass in volume V per time should be equal to the net water flux over the boundary of V per time. Hence,

$$\frac{d}{dt}\int\limits_V \rho\,\theta(t,\vec{x})\,d\vec{x} \stackrel{!}{=} -\int\limits_{\partial V} \rho\,\vec{v}\cdot\vec{n}\,do\,,$$

where $\vec{n}$ is the outward normal unit vector on ∂V. The reader realizes that on the right-hand side, the sign is necessary to describe the net mass flux *into* the volume. On the right-hand side we use the divergence theorem and on the left-hand side, assuming a certain smoothness of the integrand, we draw the differentiation into the integral. We obtain

$$\int\limits_V \frac{\partial}{\partial t}\,\rho\,\theta(t,\vec{x})\,d\vec{x} \stackrel{!}{=} -\int\limits_V \nabla\cdot(\rho\,\vec{v})\,d\vec{x}\,.$$

To get rid of the integrals, we assume V to be a sphere around a point $\vec{x}\in\Omega$ with radius r, we divide the above equation by the volume of the sphere, and then let r go to zero. In the limit, provided the integrands are sufficiently smooth (e.g., if we take the integrals to be Riemann integrals, the continuity of the integrands is sufficient for this purpose), we obtain

$$\boxed{\partial_t(\rho\,\theta) + \nabla\cdot(\rho\,\vec{v}) = 0}\,. \tag{2.1}$$

If we assume that the density of water is constant[4] the equation of mass conservation simplifies to

$$\partial_t(\theta) + \nabla\cdot\vec{v} = 0\,.$$

This is *one* equation for $n+1$ unknowns $\vec{v},\theta$. Hence we need some additional assumptions/equations.

2.1 The saturated case

If we assume that a priori it is known that the medium is saturated (or by restricting our view to those parts of the domain that are saturated), $\theta = \omega$ holds. And the porosity ω can usually be assumed to be constant in time. So in this case we get the equation

$$\nabla\cdot\vec{v} = 0\,. \tag{2.2}$$

We now need one more constitutive law. In 1856 Henry Darcy found experimentally the so-called *Darcy-law* (*Darcy's law*) for saturated porous media

$$\vec{v} = -K_{sat}\,\nabla p$$

where p is the *pressure* and K_{sat} is the *hydraulic conductivity*.
Is it possible to find other reasons or derivations for Darcy's law besides experimental findings?

[4]The density of water may depend a little on temperature, possible impurities, and, if extreme pressure occurs, a little on pressure,

On the microscale (in the pore), the motion of the fluid is described by the Navier-Stokes equations

$$\begin{aligned} \rho\,\partial_t\vec{w} + \rho\,(\vec{w}\cdot\nabla)\vec{w} - \eta\,\Delta\vec{w} + \nabla p &= \vec{f}\,, \\ \nabla\cdot w &= 0 \end{aligned}$$

(or, since the velocities are small, approximated by the Stokes equation) with zero boundary conditions at the pore surfaces. η is the dynamic viscosity of the fluid. The external force (density) f is caused here by the gravitation of the earth, thus by a potential force, i.e., there is a potential $\psi(x_1,...,x_n) = -\rho\,g\,x_n$ to this force field, i.e., $\vec{f} = \nabla\psi = -\rho g\vec{e}_n$, where $\vec{e}_n$ is the n-th standard basis vector and where $g = 9.81N/kg$ is the acceleration of gravity. By introducing $\hat{p} := p - \psi$, we can express the Navier-Stokes equation as

$$\begin{aligned} \rho\,\partial_t\vec{w} + \rho\,(\vec{w}\cdot\nabla)\vec{w} - \eta\,\Delta\vec{w} + \nabla\hat{p} &= \vec{0}\,, \\ \nabla\cdot w &= 0 \end{aligned}$$

where ψ can be called *hydrostatic pressure* and $\hat{p}$ *hydrodynamic pressure*; $p = \psi + \hat{p}$. If we now make the extremely simplifying assumption that all pores are similar Rëars pointing in the same direction with constant radius $R > 0$, we can solve the Navier-Stokes equation analytically. Under the above assumption on the pore geometry, one obtains a parabolic profile, the so-called *Poiseuille profile*. It is described, if the pores point in x_1-direction and the space dimension is $n=3$, by

$$w(r) = -\frac{\partial_{x_1}\hat{p}}{4\eta}\,(R^2 - r^2)\,, \quad r \in [-R, R]\,.$$

By integrating över a cross-sectional surface of the pore and using polar coordinates, calculate the volume flux through the pore:

$$\mathcal{F} = -\frac{\partial_{x_1}\hat{p}}{4\eta}\int_0^R\int_0^{2\pi} r\,(R^2 - r^2)\,d\varphi\,dr = -\frac{\partial_{x_1}\hat{p}}{4\eta}\,2\pi\left(\frac{1}{2}r^2R^2 - \frac{1}{4}r^4\right)\Big|_0^R = -\frac{\pi R^4}{8\eta}\,\partial_{x_1}\hat{p}\,.$$

This is the so-called *Hagen-Poiseuille law*. The flow through a pore of radius R (i.e. cross-sectional area $\sim R^2$) thus, for given pressure gradients, increases with the fourth power of R! By averaging over the cross-sectional area (i.e., by dividing by πR^2) we get the macroscopic seepage velocity

$$\vec{u} = -\frac{R^2}{8\eta}(\partial_{x_1}\hat{p}, 0, 0)^T$$

and the Darcy velocity

$$\vec{v} = \theta\,\vec{u} = -\theta\,\frac{R^2}{8\eta}(\partial_{x_1}\hat{p}, 0, 0)^T = -\theta\,\frac{R^2}{8\eta}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}\nabla\hat{p}\,.$$

We can see that the Darcy velocity depends quadratically on the pore radius. Above all, however, we recognize: The derived relation confirms *Darcy's law*

$$\vec{v} = -K_{sat}\,\nabla\hat{p}\,, \tag{2.3}$$

where here

$$K_{sat} = \theta\,\frac{R^2}{8\eta}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{2.4}$$

This conductivity tensor reflects the extreme anisotropy of the assumed pore geometry; it is such that the resulting $\vec{v}$ – of course – is always oriented in the x_1 direction, even if the pressure gradient has a different direction; only the x_1-component of the pressure gradient acts as the forcing.
If we substitute (2.3) into (2.2), we get the elliptic problem

$$-\nabla\cdot(K_{sat}\,\nabla\hat{p}) = 0\,. \tag{2.5}$$

A real porous medium will be more isotropic than the case calculated out above. In the case of complete isotropy (i.e., as far as the pore orientation is concerned, there are no preferential directions at all), instead of (2.4) one would probably rather assume

$$K_{sat} = \theta\,\frac{R^2}{8\eta}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{2.6}$$

In this case, the elliptic problem (2.5) can also be written with *scalar*

$$K_{sat} = \theta\,\frac{R^2}{8\eta}\,.$$

For a moderately unisotropic porös medium, the matrix can be assumed to be fully filled.
If we express the hydrodynamic pressure $\hat{p}$ again by the physical pressure, Darcy's law (2.3) reads

$$\vec{v} = -K_{sat}\,\nabla(p+\rho g\vec{e}_3)\,. \tag{2.7}$$

To better understand this additional term in the pressure, the following explanation: In a fully saturated porous medium, a groundwater reservoir resting on an impermeable rock layer, and on which no forces act except gravity, p increases continuously in downward direction (i.e. in $-\vec{e}_3$-direction), i.e., there is a pressure gradient. Nevertheless, the fluid is at rest, i.e. $\vec{v}=\vec{0}$. This is only possible by the compensating term $\rho g\vec{e}_3$. The same is true in a swimming pool: although a downward force acts everywhere on the water, it does not cause the water to move. Only the hydrodynamic part of the pressure causes motion of the water, but not the hydrostatic part.
Finally let us note that the conductivity tensor can be written as

$$K_{sat} = \frac{\kappa_{sat}}{\eta}\,,$$

i.e., as a numerator depending only on the geometry of the pore space and a factor depending on the fluid, namely the viscosity. The matrix or scalar κ_{sat} is called the *permeability* of the porous medium.
We will learn about another way to derive Darcy's law, the asymptotic evolution, in the tutorials. Different to our approach in this chapter, the concept of asymptotoc expansion allows for a rather complex pore geometry; however, it assumes that the pore space is periodic.

2.2 The unsaturated case

This case will be dealt with only very briefly in this lecture.
In the twentieth century (starting in 1908), Darcy's law, which originally referred to saturated media, was also applied to unsaturated media:

$$\begin{aligned} \partial_t(\rho\,\theta) + \nabla\cdot(\rho\,\vec{v}) &= 0 \\ \vec{v} &= -K\,\nabla(p+\rho g x_n) \end{aligned} \tag{2.8}$$

However, some changes appear in the conductivity tensor. This can be easily understood if one visualizes:

- In Chapter 2.1 we found out that the mean velocity in large pores is much larger than in small pores.
- In an unsaturated medium, due to the capillary effect, water is preferentially in the small pores and air in the large pores.

In a non-saturated medium, the seepage velocity, i.e. the conductivity, will be lower than in the saturated case, beause there is less volume avaiable for the transport, and because variability of pore radii in combination with capillary effects hinder movement of the fluid. It is reasonable to assume that the conductivity is a monotonically increasing function of the water content, and that in a very dry porous medium the water phase is no longer connected and thus there is no transport of water at all. Taken together, one thus models

$$K = K(\theta) = k_{rel}(\theta)\,K_{sat}\,, \quad \text{with } k_{rel}(0) = 0\,,\ k_{rel}(\omega) = 1,$$

$k_{rel} : [0,\omega] \to [0,1]$ monotonically increasing. If we put this into our equations (2.8), we are still not done, because unlike the saturated case, our model still contains the unknown θ in addition to the unknown p. So we need another constitutive law. For this purpose, we postulate that the water content can be written as a function of the pressure: $\theta = \theta(p)$
We obtain the model (ρ=const assumed)

$$\partial_t\,\theta(p) - \nabla\cdot[K(\theta(p))\,\nabla(p+\rho g x_n)]\,. \tag{2.9}$$

This is the so-called *Richards equation*[5]. It is nonlinear and of parabolic type. For the (nonlinear) isotherms $p \mapsto \theta$ and $p \mapsto K$ different approaches are in use, which we do not discuss here. Note also that often, a priori, it is not clear whether/where the medium is saturated or unsaturated. So in parts of the computational domain the parabolic Richards equation holds, in other parts the elliptic equation (cf. chap. 2.1) holds, and the location of the interface between the two domains is generally unknown and variable in time. Hence, the situation where saturatedness/unsaturatedness is not a priori know is a rather demanding mathematical problem!

3 Derivation of the transport reaction equation

We assume in this course that there is one 'dominant' chemical species in the fluid phase (water is the dominant species in the water phase), and that all other substances dissolved in the fluid phase have small concentrations.
The *concentration* of a chemical species X dissolved in water we denote by $c(t,x)$, it is ≥ 0, and can be measured in mass per *water* volume $V_W(t,x)$, thus has the unit kg/l or g/l or kg/m^3. When modeling chemical reactions, it is useful to measure in *moles/l* instead. Strictly speaking, this magnitude is called *molality*; however, in what follows I continue to use the term 'concentration', for g/l as well as for mol/l. Furthermore, for the number of moles of X per *mass* of water, there is the term *molarity*.[6]
A mole is that amount of a substance consisting of $6.022 \cdot 10^{23}$ particles. This number is also called *Avogadro's constant*. It is just chosen so that if a particle of the substance consists of n nuclear building blocks (protons, neutrons), then one mole of this substance has the mass n grams.[7].
The concentration of a substance with respect to the *total* volume is $\theta(t,x)c(t,x)$.
Now for the *mass balance* of a substance dissolved in the fluid in a volume $\Omega' \subset \Omega$:

$$\begin{aligned} \frac{d}{dt}\int_{\Omega'} \theta(t,\vec{\vec{x}})\, c(t,\vec{x})\, d\vec{x} &= \text{net inflow through edge } + \text{ sources (chem. react.)} \\ &= -\int_{\partial\Omega'} \vec{Q}\cdot\vec{n}\, d\vec{o} + \int_{\Omega'} \theta(t,\vec{x})\, f(t,\vec{x})\, d\vec{x} \end{aligned}$$

In the representation of the boundary integral, it was assumed that the mass flux can be represented by means of a vector field $\vec{Q}$, whose normal component $\vec{Q}\cdot\vec{n}$ is important when the flux across the boundaries is to be calculated.

[5] Lorenzo Richards, 1931-93, USA, 1931 (note the correct usage of the apostrophe: it is *Richards' equation* or *the Richards equations*, but not *Richard's equation*!)

[6] In addition to these two possibilities, i.e., mass of substance per volume and mole of substance per volume, one can also specify mass of substance per mass of fluid phase (=*mass fraction*, mol/mol) or mole of substance per mole of fluid phase (=*mole fraction*, mol/l) for measuring concentrations.

[7] Strictly speaking, this relation is only valid for carbon C_{12}, for other substances there are minimal deviations due to relativistic mass effects.

The integrand on the left side has the dimension mass per (total) volume, thus the left side has the dimension mass per time. It follows that the vector field $\vec{Q}$ introduced above has dimension mass per area×time.[8]. The quantity $\vec{Q}$ of dimension

$$\frac{\text{mass}}{\text{area} \times \text{time}} = \frac{\text{mass}}{\text{volume}} \cdot \frac{\text{length}}{\text{time}}$$

is called *flux density* of the substance X. It can be seen from the line above that the flux density is the product of a concentration (mass/volume) and a velocity (distance/time); we will revisit this fact in a moment.
The reaction density f has the dimension mass per time×(water) volume. We will examine concrete models for f later. θf describes the reaction density as mass per time×(total)volume.
Three effects/phenomena contribute to the mass flux $\vec{Q}$ in a porous medium: advection, diffusion, and dispersion:

$$\vec{Q} = \vec{Q}_{adv} + \vec{Q}_{diff} + \vec{Q}_{disp}$$

1st, advection: The substance X is entrained by the flow field $\vec{u}$:

$$\vec{Q}_{adv} = \theta \vec{u} c = c\vec{v}\,.$$

Note that the dimension of $\vec{Q}_{adv}$ fits the dimension required above.
2nd, (molecular) diffusion: The intrinsic thermal motion of the particles (Brownian motion) provides, macroscopically, a net migration of the particles of the substance X that are in regions of high concentration to regions of low concentration. It is assumed that the diffusive flux is a function of the concentration gradient, or more precisely, that there is a linear relationship between them. This linear relationship is referred to as Fick's law of diffusion or *Fickian diffusion*:

$$\vec{Q}_{diff} = -d_{diff}\, \theta\, \vec{\nabla} c$$

The diffusion coefficient $d_{diff} > 0$ may depend on other parameters, in particular on the temperature, it is generally a one different for every species. The inclusion of a factor θ in the above law is plausible, since diffusion only takes place in the pore space, and θ measures the pore space volume.
Note: Advection and diffusion also take place in 'free flow' situations, i.e., in fluids outside of porous media. Diffusion also occurs when the fluid is at rest ($\vec{u}=\vec{0}$).
3., (Kinematic) Dispersion: This is an effect that occurs only in porous media. Only in porous media, there are microscopic processes that we have not yet incorporated in our macroscopic model (with the velocities averaged over REVs) (see Fig. 2):
(a) In the center of the pores, the velocity is larger than at the edge of the pores – see also our considerations in Chap. 2.1, where we derived a Poiseuille profile

[8]Note that $\vec{Q} \cdot \vec{n}$ and $\vec{Q}$ have the same dimension – for $\vec{Q} \cdot \vec{n}$ is simply the component of $\vec{Q}$ in the normal direction – since $\vec{n}$ is considered to be dimensionless, which can be motivated by the fact that $\vec{n}$ is declared by dividing a normal field by its norm: length/length

for homogeneous pores. Thus, depending on whether a particle of the substance X is in the vicinity of the pore boundaries or not, it advances faster or slower in the direction of flow than indicated by the advection, leading to an additional diffusion-like effect, especially in the direction of the flow. (b) Particles which are very close to each other, and which in a fluid outside a porous medium would move away from each other only very slowly by diffusion, may occasionally turn into different pores in a porous medium and, following the course of the pores, move further away from each other than predicted by molecular diffusion. This effect has a macroscopic effect like an additional diffusion in directions perpendicular to the low direction $\vec{u}$. (c) There are pores in which particles travel faster (thicker pores allow higher velocities, moreover straight pores) than in others (narrower pores, tortuous pores).

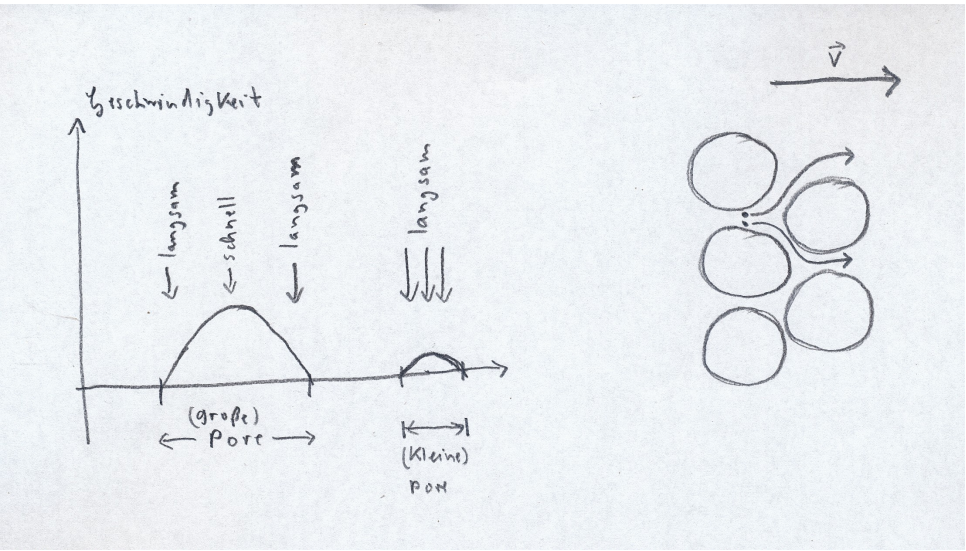


Abbildung 2: Left: profile of the flow within a pore. Right: Main cause of transversal dispersion.

It is difficult to model and quantify the three effects mentioned above with simple formulas, since it strongly depends on the geometry of the pores, on the probability distribution of the pore sizes, etc.[9] A widely used model is to model dispersion analogously to molecular diffusion. However, since the effects in the direction of the flow are different from those perpendicular to the flow direction, it makes sense to use different dispersion parameters for the flow direction and for directions perpendicular to the flow, the dispersion parameters β_l, β_t (l=longitudinal, t=transversal); such a directional dependence is also called *anisotropy*. As a rule of thumb, $\beta_l \approx 10\beta_t$. In the case that

[9]Under the idealizing assumption that the pore structure of the medium is periodic, one can mathematically rigorously derive effective equations that are even more accurate than the so-called Bear–Scheidegger model we aim at here, using the mathematical concept *homogenization / asymptho-tic expansion.*.

the advection $\vec{u}$ goes in x_1-direction, a so-called tensorial diffusion would be

$$\vec{Q}_{disp} = -|\vec{v}|\, D_{disp,0}\, \vec{\nabla} c, \quad D_{disp,0} = \mathrm{diag}(\beta_l, \beta_t, \beta_t)\,;$$

the factor $|\vec{v}|$ seems plausible, since the magnitute of the above-mentioned phenomena seems to be proportional to the advection velocity (in particular, the phenomena vanish for $\vec{v}=\vec{0}$!). To apply this law to the case where $\vec{u}$ does not point in x_1–direction is easy, by means of a principal axis transformation: In the general case

$$\vec{Q}_{disp} = -|\vec{v}|\, D_{disp}\, \vec{\nabla} c, \quad D_{disp} = X D_{disp,0} X^{-1},$$

where X is the orthogonal matrix describing a rotation of the standard basis to the basis $\{\vec{v}, \vec{v}^{\perp}\}$ (for simplicity, I assume the 2-D case in the following calculation; the derivation and the result can be transferred to the 3-D case):

$$X = \frac{1}{|\vec{v}|}\begin{pmatrix} v_1 & -v_2 \\ v_2 & v_1 \end{pmatrix}, \quad X^{-1} = \frac{1}{|\vec{v}|}\begin{pmatrix} v_1 & v_2 \\ -v_2 & v_1 \end{pmatrix}, \quad D_{disp,0} = \begin{pmatrix} \beta_l & 0 \\ 0 & \beta_t \end{pmatrix}.$$

We subtract from $D_{disp,0}$ the summand $\beta_t\,\mathrm{Id}$ and obtain

$$\begin{aligned} D_{disp} &= \beta_t X X^{-1} + \begin{pmatrix} v_1 & -v_2 \\ v_2 & v_1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} v_1 & v_2 \\ -v_2 & v_1 \end{pmatrix} \\ &= \beta_t\,\mathrm{Id} + \frac{\beta_l - \beta_t}{|\vec{v}|^2}\begin{pmatrix} v_1^2 & v_1 v_2 \\ v_1 v_2 & v_2^2 \end{pmatrix} = \boxed{\beta_t\,\mathrm{Id} + \tfrac{\beta_l-\beta_t}{|\vec{v}|^2}\,\vec{v}\vec{v}^T} \end{aligned}$$

The matrix D_{disp} is called the *Bear-Scheidegger tensor*, the dispersion model the *Bear-Scheidegger model* (Scheidegger'61, Bear'72). In 3-D, we get the same formula.
Hence, in total, we get the flux density

$$\vec{Q} = -\underbrace{\Big(\theta\, d_{diff}\,\mathrm{Id} + \beta_t\,|\vec{v}|\,\mathrm{Id} + \frac{\beta_l - \beta_t}{|\vec{v}|}\,\vec{v}\vec{v}^T\Big)}_{=:D(\vec{v})}\vec{\nabla} c + c\vec{v}.$$

with the diffusion–dispersion tensor $D(\vec{v})$.
We put this into the equation of conservation of mass and apply the Gaussian divergence theorem:

$$\begin{aligned} \frac{d}{dt}\int\limits_{\Omega'} \theta\, c\, d\vec{x} &= \int\limits_{\partial\Omega'} (D(\vec{v})\vec{\nabla} c - c\vec{v})\cdot \vec{n}\, d\vec{o} + \int\limits_{\Omega'} \theta\, f\, d\vec{x} \\ &= \int\limits_{\Omega'} \vec{\nabla}\cdot (D(\vec{v})\vec{\nabla} c - c\vec{v})\, d\vec{x} + \int\limits_{\Omega'} \theta\, f\, d\vec{x} \end{aligned}$$

Multiplying this equation by $\frac{1}{|\Omega'|}$ and then letting $|\Omega'| \to 0$, this yields, if the integrands have some regularity (e.g. are continuous in $\vec{x}$), the PDE of mass conservation for the species X

$$\boxed{\partial_t(\theta c) - \vec{\nabla}\cdot (D(\vec{v})\vec{\nabla} c - c\,\vec{v}) = \theta f},$$

the *advection–diffusion–dispersion–reaction equation*, or *equation of reactive transport*. For comparison, the corresponding equation for transport processes outside porous media (i.e., $D_{disp}=0$, $\theta=1$):

$$\partial_t c - \vec{\nabla} \cdot (d_{diff} \vec{\nabla} c - c\vec{v}) = f$$

If d_{diff} is constant, it is particularly easy to write the equation in non-divergence form:

$$\partial_t c - d_{diff} \Delta c + \vec{\nabla} \cdot (c\vec{v}) = f$$

If the flow field is *solenoidal* (=divergence-free), which is true for incompressible fluids,

$$\partial_t c - d_{diff} \Delta c + \vec{v} \cdot \nabla c = f$$

follows. For multispecies problems diffusion is species–dependent, but dispersion is not. Dispersion depends on $|\vec{v}|$, diffusion does not. In the soil, $\vec{v}|$ is usually sufficiently large such that dispersion is usually (significantly) larger than diffusion.[10]. As a consequence, diffusion is often neglected relative to dispersion, and the diffusion–dispersion tensor is assumed to be *species-independent*. We will see later what mathematical/numerical advantages this assumption may have.

4 Chemical reaction rates

4.1 Simple examples

A chemical reaction is described by means of a so-called *chemical equation*, such as

$$\mathrm{X}_1 + \mathrm{X}_2 \longrightarrow \mathrm{X}_3.$$

This means that a particle (molecule, ion,...) of substance X_1 combines with a particle of substance X_2, resulting in a particle of substance X_3. It follows immediately that 1 *mole* of X_1 and 1 mole of X_2 react to give one *mole* X_3.
Thus, for the source terms f_1, f_2, f_3 of the three associated concentrations, it follows that f_1 and f_2 are negative and f_3 is positive; more precisely, that $\vec{f}$ has the form

$$\vec{f} = \begin{pmatrix} f_1 \\ f_2 \\ f_3 \end{pmatrix} = \begin{pmatrix} -R \\ -R \\ +R \end{pmatrix} = \begin{pmatrix} -1 \\ -1 \\ +1 \end{pmatrix} R$$

where $R \geq 0$ is a reaction rate yet to be modeled, and – very important – where concentrations are measured in *moles*, i.e., in the numb er of particles, (not: grams) per liter.

[10] Exceptions could be seepage through rocks or low-permeability media such as clay

A plausible assumption is that the reaction rate R is proportional to the encounter probability of the reacting particles, and that this in turn is proportional to the product of the two concentrations:

$$R(\vec{c}) = k\, c_1\, c_2, \quad k>0$$

Such a model is quite simple and does not take into account the exact effects of forces between particles (Van der Waals force,...). Note that our PDEs of reactive transport are now *coupled* via the source term, and that this term is *nonlinear*, so we are dealing with *systems of nonlinear (more precisely: semilinear) partial differential equations.*
A generalization to somewhat more complicated chemical reactions can be done as follows:
Let us consider the reaction

$$\mathrm{X}_1 + 2\,\mathrm{X}_2 \longrightarrow \mathrm{X}_3 + \mathrm{X}_4$$

one particle of X_1 combines with two particles of X_2, i.e., the source term here must be

$$\vec{f} = \begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \end{pmatrix} = \begin{pmatrix} -R \\ -2R \\ +R \\ +R \end{pmatrix} = \begin{pmatrix} -1 \\ -2 \\ +1 \\ +1 \end{pmatrix} R.$$

The coefficients in the chemical equation are called *stoichiometric coefficients*; they occur as prefactors in front of the reaction rate, but they are signed depending on which side of the reaction arrow they are on. The species on the left-hand side of a reaction equation are called *reactants*, those on the right-hand side are called *products.*
As for the modeling of the rate itself, even if more than two particles react with each other, it is still modeled by the encounter probability of all particles involved, in the example this would be

$$R(\vec{c}) = k\, c_1\, c_2^2\,.$$

Hence, the stoichiometric coefficients occur as prefactors and as exponents as well in the source terms.
It makes sense to validate such rate laws by experiments, but this is an ambitious task, because experiments are affected by how well and how fast reactants are mixed. Often the rate coefficient k that is valid for a reaction in the field (in the subsurface) is not known or only very roughly known. Also note that in the real world, complicated reactions are often composed of several (sometimes very fast) successive partial reactions, which may lead to rate models different from the one above.

4.2 Some minimum requirements for reaction rates

Let us consider a rather general chemical reaction with I be the number of species:

$$s_1^e\,\mathrm{X}_1 + ... + s_I^e\,\mathrm{X}_\mathrm{I} \longrightarrow s_1^p\,\mathrm{X}_1 + ... + s_I^p\,\mathrm{X}_\mathrm{I}$$

Let $s_1^e, ..., s_I^e,\ s_1^p, ..., s_I^p \in \mathbb{R}_0^+$ (usually: $\in \mathbb{N}_0$). Educts are those $\mathrm{X_i}$ for which $s_i^e > 0$ holds, and products are those $\mathrm{X_i}$ for which $s_i^p > 0$. Note that here we allow a species to appear as both reactant and product ($\rightarrow$'catalyst').
As a minimum requirement for rate functions $R : (\mathbb{R}_0^+)^I \rightarrow \mathbb{R}$ we postulate: Let R be continuous, and let it hold true that

$$R(\vec{c}) = 0, \text{ if } \exists\, i : \ s_i^e \neq 0 \ \wedge \ c_i = 0$$

(i.e., if any of the reactants has concentration=0, the reaction cannot take place), and

$$R(\vec{c}) > 0, \text{ if } \forall\, i : \ s_i^e = 0 \ \vee \ c_i \neq 0$$

(i.e., if all required reactants have concentration> 0, then the reaction will definitly take place).
If the *support* of a vector is denoted by

$$\begin{aligned} \text{supp } \vec{c} \quad &:= \quad \{i \in \{1, ..., I\} \mid c_i \neq 0\}\,, \\ \text{supp } \vec{s}^e \quad &:= \quad \{i \in \{1, ..., I\} \mid s_i^e \neq 0\} \end{aligned}$$

then the above conditions can be written in short as

$$\begin{aligned} R(\vec{c}) > 0 \quad &\Longleftrightarrow \quad \text{supp } \vec{s}^e \subseteq \text{supp } \vec{c}, \\ R(\vec{c}) = 0 \quad &\text{otherwise} \end{aligned}$$

We have not yet defined the rate function for negative concentrations, which is physically understandable; however, for mathematical reasons, it may be useful to continue the rate function for negative concentrations as well, for example by setting $R(u) := R(u^+)$ for $u \in \mathbb{R}^I \backslash (\mathbb{R}^+)^I$, where $u^+ := \max\{u, 0\}$, where the maximum of a vector is taken componentwise; such a continuation is continuous, since $x \mapsto \max\{x, 0\}$ is continuous).[11]

4.3 The law of mass action

The *law of mass action (LMA)* is nothing more than the principle we applied to the examples in Chap. 4.1:

$$\boxed{R(\vec{c}) = k \prod_{i=1}^{I} c_i^{s_i^e}}, \quad k > 0,$$

[11] Why should we define reaction rates for negative concentrations? Negative concentrations cannot occur! Well, *in the real world*, negative concentrations cannot occur. But if we consider *a mathematical model*, then it is not obvious that solutions to this model are always nonnegative. Hence, a proof of nonnegativity of the solution is frequently desired. Before such a proof is completed, i.e., as a precondition to conduct such a proof, the well-definedness of reaction rates also for negative concentrations is required. Another reason might be the consideration of numerical algorithms to compute a solution of a reactive transport problem. The equations are nonlinear, and so iterative solution algorithm such as *Newton's method* will be used. What to do if an iterate is negative (even if the solution will be positive)? We do not want the iteration to crash due to an undefined reaction rate for negative iterates. See for example [Kr21] where the behavior of algorithms with respect to the occurence of negative iterates is one of the topics.

with the source term

$$\boxed{\vec{f} = (\vec{s}^{\,p} - \vec{s}^{\,e})\, R(\vec{c}).}$$

Now we model a *system* of J chemical reactions in the same way:
The chemical equations are

$$\begin{array}{ccc} s^e_{11}\,\mathrm{X}_1 + ... + s^e_{I1}\,\mathrm{X}_\mathrm{I} & \longrightarrow & s^p_{11}\,\mathrm{X}_1 + ... + s^p_{I1}\,\mathrm{X}_\mathrm{I} \\ \vdots & \vdots & \vdots \\ s^e_{1J}\,\mathrm{X}_1 + ... + s^e_{IJ}\,\mathrm{X}_\mathrm{I} & \longrightarrow & s^p_{1J}\,\mathrm{X}_1 + ... + s^p_{IJ}\,\mathrm{X}_\mathrm{I} \end{array}$$

The stoichiometric coefficients now form matrices $S^e = (s^e_{ij}), S^p = (s^p_{ij}) \in (\mathbb{R}^+_0)^{I\times J}$. Note that the convention in this lecture is: Every reaction corresponds to a *column*, and every species corresponds to a *row* in the stoichiometric matrix. Some authors do it the other way round. The total source term for a muti-species multi-reaction problem is the *sum* of the individual source terms of the reactions,

$$\vec{f} = \sum_{j=1}^{J} \vec{f}_j = \sum_{j=1}^{J} R_j(\vec{c})\,(\vec{s}^{\,e}_j - \vec{s}^{\,p}_j) = \boxed{(S^p - S^e)\,\vec{R}(\vec{c})}$$

where $\vec{s}^{\,e}_j$ and $\vec{s}^{\,p}_j$ denote the j-th column of the matrix in question, and $\vec{R} = (R_j)_{j=1..J}$ is the vector of reaction rates. This vector obviously has the components

$$\boxed{R_j(\vec{c}) = k_j \prod_{i=1}^{I} c_i^{s^e_{ij}}}, \quad k_j > 0, \quad j = 1, ..., J.$$

4.4 Reversible systems

Often the assumption is made that all reactions can be bidirectional, can run 'forward' and 'backward', i.e., in our set of reactions $j = 1, ..., J$, all reactions occur in pairs; we can thus identify a pair of reactions by the chemical equation[12]

$$s^e_{1j}\,\mathrm{X}_1 + ... + s^e_{Ij}\,\mathrm{X}_\mathrm{I} \longleftrightarrow s^p_{1j}\,\mathrm{X}_1 + ... + s^p_{Ij}\,\mathrm{X}_\mathrm{I}, \quad j = 1, ...\tilde{J},$$

where j now, to keep consistency with the earlier chapters, runs from 1 to $\tilde{J} := \frac{J}{2}$. Such reactions or systems of reactions are called *reversible*. If we also summarize the rates in pairs, we get

$$R_j(\vec{c}) = R^f_j(\vec{c}) - R^b_j(\vec{c}), \quad j = 1, ..., \tilde{J},$$

[12] The choice which side is 'left' and which is 'right', i.e., which direction is 'forward' and which is 'backward', is arbitrary.

and the source term

$$\begin{aligned}\tilde{f} &= \sum_{j=1}^{\tilde{J}} R_j^f(\vec{c})\,(\vec{s}_j^{\,p}-\vec{s}_j^{\,e}) + R_j^b(\vec{c})\,(\vec{s}_j^{\,e}-\vec{s}_j^{\,b}) \\ &= \sum_{j=1}^{\tilde{J}} (R_j^f(\vec{c}) - R_j^b(\vec{c}))\,(\vec{s}_j^{\,p}-\vec{s}_j^{\,e}) = \boxed{(S^p-S^e)\,(\vec{R}^{\,f}(\vec{c})-\vec{R}^{\,b}(\vec{c}))} \\ &= \boxed{S\,\vec{R}(\vec{c})}\end{aligned}$$

where now $S := S^p - S^e$ and $\vec{R}(\vec{c}) := \vec{R}^{\,f}(\vec{c}) - \vec{R}^{\,b}(\vec{c})$.

4.5 Further rate laws beyond the LMA

4.5.1 Law of mass action with activity correction

A closer investigation shows that the reaction velocity is not exactly proportional to the concentration, but to the so-called *activity* of the species. Hence, the LMA

$$R(\vec{c}) = k \prod_{i=1}^{I} c_i^{s_i}$$

is corrected to

$$R(\vec{c}) = k \prod_{i=1}^{I} a_i(\vec{c})^{s_i},$$

where $\vec{a} = (a_1, ..., a_I)^T$ is the vector of the activities of the species $\mathrm{X}_1, ..., \mathrm{X}_\mathrm{I}$. Each activity a_i in turn is a function of the vector of concentrations $\vec{c}$. This relation between activity and concentration is often written in the form

$$a_i(\vec{c}) = \gamma_i(\vec{c})\, c_i.$$

Here γ_i is called *the activity coefficient* of the species $\mathrm{X_i}$. It is a number that is mostly close to one, which justifies the 'simplified' LMA from Chap. 4.3. The smaller the so-called ionic strength of the solution, the more valid is the assumption from Chap. 4.3 that $a_i \approx c_i$ ($\gamma_i \approx 1$); the ionic strength (see below) is the weighted sum of the concentrations of charged particles. A widely used *activity correction* in geosciences is after *Debye* and *Hückel*:

$$\gamma_i = \exp\left(-\frac{A z_i^2 \sqrt{H}}{1 + r_i B \sqrt{H}}\right),$$

where $z_i \in \mathbb{Z}$ is the charge number of the particles, $H = H(\vec{c}) = \frac{1}{2}\sum_{i=1}^{I} c_i z_i^2$ is the ionic strength of the solution, r_i is the effective diameter of the particles (see tables,

in the order of 10^{-10}m), and, at 25 degrees Celsius, $A = 0.509\,(\mathrm{l/mol})^{1/2}$ and $B = 0.328 \cdot 10^{10}\,(\mathrm{m/mol})^{1/2}$. [13] Thus, $\gamma_i \leq 1$, and $\gamma_i \longrightarrow 1$ for $\vec{c} \longrightarrow \vec{0}$.
If, as in Chap. 4.3, one uses the approximation $a_i = c_i$, one speaks of the LMA with *ideal activities*.

4.5.2 The Monod model for biological decay

The metabolism of microorganisms leads to the degradation of certain organic substances, including 'pollutants', which is why this degradation is of particular interest for the desired attenuation of contaminated sites. The metabolism of an organism consists of thousands of chemical reactions that are largely unexplored. However, it turns out that when these reactions are grouped together, essentially a redox reaction occurs, i.e., an *electron acceptor=oxidant A* (such as O_2, Fe(III), NO_3^- (=nitrate), SO_4^{2-} (=sulfate),...) and an *electron donor=reducing agent D* (e.g. an organic pollutant like the chlorinated hydrocarbon chloroperethene=tetrachloroethene C_2Cl_4) react with each other. The net reaction is a *redox reaction*; the reducing agent is oxidized and the oxidizing agent is reduced. Oxidation (reduction, resp.) means an increase (decrease, resp.) of the so-called electronegativity, which occurs by dragging electrons from one partner towards the other. A microbial population X has a role comparable to a catalyst: it is neither degraded nor built up by the reaction. However, the microbial population benefits from the reaction in that its reproduction rate depends on it. Let us call the degradation product P. We may decide whether we introduce an unknown and a differential equation for P in the model, depending on whether the substance is classified as 'harmless' or if the substance plays a role in further interesting reactions.
The metabolism of a microorganism consists of a bunch of reactions; greatly simplified, the mechanism can be represented as

$$s_A\,\mathrm{A} + s_B\,\mathrm{B} + \mathrm{X} \longrightarrow s_C\,\mathrm{P} + \mathrm{X}$$

In the *Monod model* the associated rate is assumed to be of the form

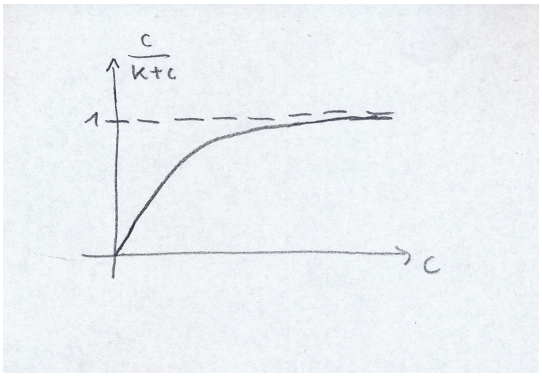
$$R(c_A, c_B, c_X) = k\,c_X\,\frac{c_A}{k_A + c_A}\,\frac{c_D}{k_D + c_D}, \quad \text{with } k, k_A, k_D > 0,$$

has. The source terms of the system of differential equations are

$$\begin{aligned}
f_A &= -s_A\,R(c_A, c_B, c_X), \\
f_D &= -s_D\,R(c_A, c_B, c_X), \\
f_X &= \underbrace{\left(1 - \frac{c_X}{c_{X,max}}\right)}_{\text{biomass limitation}}\,R(c_A, c_B, c_X) - \underbrace{k_d\,c_X}_{\text{mortality term}}, \\
f_P &= +s_P\,R(c_A, c_B, c_X) \quad (\text{if } c_P \text{ is included in the model})
\end{aligned}$$

[13] For the derivation of the Debye–Hückel model the Boltzmann model of thermodynamics (statistical model describing the probability distribution of the velocities of particles) with electrostatics (Poisson equation describing electrostatic fields of a point charge), gives $B = \sqrt{2e^2 N_A/(\epsilon k_B T)}$ and $A = e^2 B/(8\pi\epsilon k_B T \ln 10)$, where k_B is Boltzmann's constant, e is the elementary electric charge, N_A is Avogadro's constant, ϵ is the dielectric conductivity of the solution, and T is the absolute temperature.

There are variants of the model. For example, one can omit the mortality term, or omit the biomass limit (which guarantees $c_X < c_{X,max}$), or one can incorporate other factors into the rate R, called *inhibition factors*, of the form $k_Y/(c_Y+k_Y)$ to model if a substance Y inhibits the process or harms the microorganism. The substance Y might be oxygen for microorganisms that prefer an anaerobic way of like.
To motivate Monod's model: At first, it is plausible that when food is scarce, the metabolic rate is proportional to food supply, but when there is an oversupply, there is – quite literally – a saturation effect, which causes terms of the form $c/(c+k)$ to be plausible.



As for the deeper explanation of the model, this is somewhat controversial (there is a paper discussing five different theories explaining the behavior). However, the most widely accepted explanation of the Monod model is the following: The rate of degradation is essentially determined by the slowest partial reaction, and this is an enzyme reaction subject to so-called Michaelis-Menten kinetics. An enzyme reaction according to Michaelis-Menten (1911) has the form

$$\mathrm{A} + \mathrm{E} \underset{k_2}{\overset{k_1}{\rightleftarrows}} K \xrightarrow{k_3} \mathrm{E} + \mathrm{P}.$$

Where E is the (uncomplexed) enzyme, K is the enzyme complex, A is the reactant, P is the product. It is assumed that all three reactions proceed according to the law of mass action:

$$\begin{aligned} c'_K &= k_1 c_A c_E - (k_2+k_3) c_K, \\ c'_E &= (k_2+k_3)\, c_K - k_1 c_A c_E, \\ c'_P &= k_3 c_K \end{aligned}$$

By addition two equations, it follows that $c'_K + c'_E = 0$, so $c_K + c_E = \text{const} =: c_{E0}$. Now, the assumption of an equilibrium state is made, i.e., let $c'_K = c'_E = c'_A = 0$.[14] We obtain

[14] which is, for $k_3 > 0$, only possible if the substance A is supplied 'from outside' at a rate $R_0 > 0$; it is $c'_A = k_2 c_K - k_1 c_A c_E + R_0 = 0$

the equations

$$\begin{aligned} k_1 c_A c_E &= (k_2+k_3)\, c_K\,, \\ c_K + c_E &= c_{E0}\,, \\ c_P' &= k_3 c_K\,. \end{aligned}$$

Eliminating c_E in the first equation using the second equation, we get $k_1 c_A(c_{E0}-c_K) = (k_2+k_3)c_K$, which is

$$c_K = c_{E0}\,\frac{c_A}{c_A + \frac{k_2+k_3}{k_1}}\,.$$

This leads to

$$c_P' = k_3 c_{E0}\,\frac{c_A}{c_A + \frac{k_2+k_3}{k_1}}\,.$$

This has the form of a Monod term, with Monod parameter $k = (k_2+k_3)/k_1$, which can serve as a motivation for adopting such terms in the Monod model.

5 The batch problem/ODE model

In this chapter we deal with a general mass action system neglecting spatial transport, i.e. our model is a (nonlinear) *ODE* system. In the geosciences, a reactive problem in which transport is neglected and all substances are considered to be in one place is also called a batch problem.
This consideration is justified for example if one thinks of spatially homogeneous (i.e. constant with respect to x) situations. The reason for this simplification is that the analysis for an ODE system is much simpler (less 'technical') than for a PDE system, which we will also examine later in chapter 7. Many of the properties, which one can show for the ODE system, can be transferred, if one makes the necessary effort, also to the PDE model.

5.1 Positivity of solutions

Theorem. The solution of the batch problem

$$\vec{c}\,'(t) = S\,\vec{R}(\vec{c}(t))$$

with positive initial value $\vec{c}(0) > \vec{0}$ and mass action kinetics (not necessarily reversible) and stoichiometric coefficients $s_{ij}^e, s_{ij}^p \in \mathbb{N}$[15] is strictly positive on any interval of existence of the solution.
Proof. Let $[0,T)$ be an interval of existence of the solution. Suppose there exists an $i \in \{1,...,n\}$ and a $t_1 \in [0,T)$ with $c_i(t) \leq 0$. Since the initial value is strictly positive, the compact set $\{t \in [0,t_1] \,|\, \exists j : c_j = 0\}$ is nonempty and thus has a smallest element

[15] This can be weakened to $s_{ij}^e, s_{ij}^p \in \{0\} \cup [1,\infty)$

t_2, and $t_2 > 0$. Thus on $[0, t_2)$ all $c_i(t)$ are strictly positive, and there is an i such that $c_i(t_2) = 0$.
The i-th component of our ODE system is

$$c_i'(t) = \sum_{j=1}^{J} (s_{ij}^p - s_{ij}^e) \underbrace{R_j(\vec{c}(t))}_{\geq 0}, \quad R_j(\vec{c}) = k_j \prod_{k=1}^{J} c_k^{s_{kj}^e}.$$

On $[0, t_2]$, $R_j(\vec{c}(t))$ is obviously nonnegative. We distinguish the terms by their sign. A summand can be negative if and only if $s_{ij}^e > s_{ij}^p$, that is, if $s_{ij}^e \geq 1$. But then the $R_j(\vec{c}(t))$ contains a factor $c_i(t)^{s_{ij}^e}$ with $s_{ij}^e \geq 1$. We denote the sum of all terms with positive sign $\alpha_0(t) \geq 0$ (this term does not contain c_i, see exercises, but this is not essential here). The negative terms, on the other hand, depend on c_i in polynomial form. They can thus be written as $\sum_{r=1}^{m} \alpha_r(t)\, c_i(t)^r$, where the $\alpha_r(t)$ includes the dependencies on the c_j, $j \neq i$, and thus are also ≥ 0 on $[0, t_2]$.[16]
On the compact interval $[0, t_2]$ the continuous functions α_0, α_r are bounded, so there is $C > 0$ with $0 \leq \frac{c_i(t)}{C} \leq 1$. Hence, $0 \leq (\frac{c_i(t)}{C})^r \leq \frac{c_i(t)}{C} \forall\, r, i$, hence $c_i(t)^r \leq C^{r-1} c_i(t)$. We can thus estimate the powers $c_i(t)^r$ by linear terms:

$$c_i'(t) = \underbrace{\alpha_0(t)}_{\geq 0} - \sum_{r=1}^{m} \underbrace{\alpha_r(t)}_{\geq 0}\, c_i(t)^r \geq - \underbrace{\sum_{r=1}^{m} C^{r-1} \alpha_r(t)}_{\leq \tilde{C}}\, c_i(t)$$

The term $\sum\limits_{r=1}^{m} C^{r-1} \alpha_r(t)$ is bounded by a constant $\tilde{C} > 0$ because of its continuous dependence on $t \in [0, t_2]$. Since also $c_i(t) \geq 0$ on $[0, t_2]$, we end up with

$$\frac{c_i'(t)}{c_i(t)} \geq -\tilde{C} \quad \forall t \in [0, t_2).$$

Integration yields, because of the monotonicity of the integral,

$$\int_0^t \frac{c_i'(t)}{c_i(t)}\, dt \geq -\tilde{C} t$$

for $t \in [0, t_2]$, thus (substitution $u := c(t)$)

$$c_i(t) \geq c_i(0)\, \exp(-\tilde{C} t), \quad t \in [0, t_2).$$

This is a contradiction to the continuity of c_i, since $c_i(t_2) = 0$. □

[16] It is important here that the summation starts at $r = 1$ and not at $r = 0$.

5.2 Boundedness of solutions, existence of global solutions

In this chapter, two different ways to show boundedness of solutions of the batch problem with mass action kinetics are given.
Exercise task 1 showed us one way how to prove boundedness of solutions using invariants together with the nonnegativity from Chap. 5.1.

Theorem. Let the assumptions of the previous theorem hold, and let $\mathcal{S}^\perp \cap (\mathbb{R}^+)^I \neq \emptyset$. Then the solution of the batch problem with mass action kinetics (not necessarily reversible) on any interval of existence $[0, T)$ is bounded by a constant. The constant depends on the initial values and the stoichiometry, but is independent of T.
Note that $\mathcal{S} := \text{image}(S)$ is the column space of the matrix S and $\mathcal{S}^\perp$ its orthogonal complement.

Proof. Let $\vec{s^\perp} \in \mathcal{S}^\perp \cap (\mathbb{R}^+)^I$. We form the linear combination

$$\Phi(t) := \sum_{i=1}^{I} s_i^\perp \, c_i(t).$$

This Φ is a conserved quantity:

$$\Phi'(t) = \langle s^\perp, \vec{c}'(t) \rangle = \langle s^\perp, S\vec{R}(\vec{c}(t)) \rangle = \langle \underbrace{S^T s^\perp}_{=\vec{0}}, \vec{R}(\vec{c}(t)) \rangle = 0\,,$$

because $s^\perp \in \mathcal{S}^\perp = \text{image}(S)^\perp = \text{core}(S^T)$. Hence, $\Phi(t) = \Phi(0) = \text{const} = \langle s^\perp, \vec{c}(0) \rangle$.
Since $s_i^\perp \neq 0$, we can solve the equation $\sum_{i=1}^{I} s_i^\perp \, c_i(t) = \langle s^\perp, \vec{c}(0) \rangle$ to $c_i(t)$:

$$c_i(t) = \frac{1}{s_i^\perp} \left(\langle s^\perp, \vec{c}(0) \rangle - \sum_{j \neq i} s_j^\perp \, c_j(t) \right)$$

Since $s_j^\perp \geq 0$ and $c_i(j) \geq 0$, we obtain

$$c_i(t) \leq \frac{1}{s_i^\perp} \, \langle s^\perp, \vec{c}(0) \rangle \,. \qquad \square$$

Corollary. Under theassumptions of the above theorem, the solution of the batch problem with mass action kinetics exists on all of $[0, \infty)$; i.e., the batch problem has a *global* solution.

The reason for this is a theorem found, for example, in the book [Ha64], on ODE systems with locally Lipschitz-continuous right-hand side:
If there exists a function $f : [0, \infty) \to \mathbb{R}$ such that every local solution of the ODE system on its interval of existence satisfies the condition $|\vec{c}(t)| \leq f(t)$, then the

solution exists on all of $[0,\infty)$. (Background: There is always a 'maximal' solution, i.e., a solution which extends 'to the boundary' of the (possibly unbounded) domain, e.g. $[0,\infty)\times\mathbb{R}^n$, on which the right-hand side is locally Lipschitz-continuous. The assumption that the solution is 'blocked' (bounded) by the graph of $f(t)$ ensures that the maximal solution must go to $t=\infty$).

Physical interpretation. In 'real' chemical systems each particle (molecule) is composed of a strictly positive number of atoms. If we set $s_i^\perp$ as the number of atoms composing a molecule of $\mathrm{X_i}$ (which is obviously a strictly positive number), $\vec{s}^\perp$ is orthogonal to each column of the matrix S (this says nothing but that the number of atoms on the left and on the right-hand side of a chemical equation must be equal: conservation of the number of atoms). Thus, the vector $\vec{s}^\perp$ formed in this way satisfies the condition of the above theorem. So the assumption of the theorem $\mathcal{S}^\perp\cap(\mathbb{R}^+)^I\neq\emptyset$ is quite 'natural'/realistic.

Other invariants are, e.g., the number of atoms of a certain atomic variety (e.g., the C-atoms) or the charge number; however, their components are usually only nonnegative instead of strictly positive, i.e., their use and can only show the boundedness of some of the components of the solution vector.

Note that here we used the particles' composition of atoms for the first time.

There is another approach to the boundedness of solutions, which is based on so-called *Lyapunov functions*, and which does not use the composition of the particles from atoms. However, the straight-forward application of this method requires *reversible* systems; any extension to non-reversible systems is difficult. As a motivation to pursue this new approach may serve that it
1. will be quite useful also for the PDE model and
2. that it is also applicable to so-called *peculiar* chemical systems. To see what a peculiar system is and why it might be considered, let us look at the following scenarios: 2.a.: In reactions with water (or with substances that are present in large quantities and whose quantity is hardly affected by reactions) the substance 'water' is often eliminated from the system: Instead of the chemical equation $2\mathrm{H^+}+\mathrm{OH^-}\longleftrightarrow\mathrm{H_2O}$, to which belong the rates $R^v(\vec{c})=k^f\,c_1^2c_2$, $R^r(\vec{c})=k^r\,c_3$, one draws the (de facto constant) concentration of water, c_3, into the reaction constants and gets $\tilde{R}^v(\vec{c})=k^f\,c_1^2c_2$, $\tilde{R}^r(\vec{c})=\tilde{k}^r$, which now corresponds to the chemical equation $2\mathrm{H^+}+\mathrm{OH^-}\longleftrightarrow$ 'nothing' in such a chemical equation there is no conservation of atoms , and there is no $\vec{s}^\perp$ with positive entries orthogonal to column $(1,1)^T$ of the associated new stoichiometric matrix.
2.b.: Also *inflow* and *outflow* from a chemical reactor can be modelled by mass action kinetics for (pseudo-)reactions 'nothing' $\longrightarrow\mathrm{X_i}$ and $\mathrm{X_i}\longrightarrow$ 'nothing', respectively.

The Lyapunov technique (for ODEs) For an ODE system $\vec{y}'(t)=\vec{f}(t,\vec{y}(t))$ find a

functional φ such that

1. $\frac{d}{dt}\varphi(\vec{y}(t)) \leq 0$, where $t \to \vec{y}(t)$ is the solution of the ODE ('φ decreases along solutions'),

2. from the boundedness of $t \to \varphi(\vec{y}(t))$ follows the boundedness of $t \to \vec{y}(t)$ (hence, if, for example, an estimate of the form $|\vec{y}| \leq c_1\varphi(\vec{y}) + c_2$ holds for φ).

The reasoning is then as follows:
From 1. it follows that $\varphi(\vec{y}(t)) \leq \varphi(\vec{y}(0)) = \text{const}$, then from 2. it follows $|\vec{y}(t)| \leq c_1\varphi(\vec{y}(t)) + c_2 \leq c_1\varphi(\vec{y}(0)) + c_2$. From such a bound then the existence of a *global* solution follows.

Remarks.

- Variants/improvements of the precondition 1. are conceivable. Thus, it is also sufficient to 'control' a possible *growth* of φ along solutions, e.g., by weakening condition 1. to $\frac{d}{dt}\varphi(\vec{y}(t)) \leq g(t)$ or to $\frac{d}{dt}\varphi(\vec{y}(t)) \leq c\,\varphi(\vec{y}(t))$.

- If an ODE system is given without a concrete application background, it is often very difficult to find a suitable function φ. For application problems it is often promising to use physical quantities like some energy as φ; for processes including friction (motion in a gravitational field, motion of a spring oscillator...) perhaps the sum of kinetic and potential energies.
  Mathematically, the calculation

$$\frac{d}{dt}\varphi(\vec{y}(t)) = \langle \nabla\varphi(\vec{y}(t)), \vec{y}'(t)\rangle \overset{\text{ODE}}{=} \langle \nabla\varphi(\vec{y}(t)), f(t,\vec{y}(t))\rangle\,,$$

  shows that the requirement for φ is that $\langle \nabla\varphi(\vec{y}(t)), f(t,\vec{y}(t))\rangle$is non-positive or at least 'not too big'.

In the present case of a reversible reactive problem with mass action kinetics, a suitable functional is

$$\varphi(\vec{c}) := \sum_{i=1}^{I}(\mu_i - 1 + \ln c_i)\, c_i + e^{1-\mu_i}, \quad (\mathbb{R}_0^+)^I \longrightarrow \mathbb{R},$$

where the vector $\vec{\mu}$ is a solution of the linear system of equations

$$S^T\vec{\mu} = -\ln\vec{K}$$

and $S = S^p - S^e$, $\vec{K} \in \mathbb{R}_+^{\tilde{J}}$, $k_j = \frac{k_j^v}{k_j^r}$.

Motivation/physical meaning of φ for chemical reactions: The construction of φ is inspired by a quantity (a 'potential') from the thermodynamics of mixtures, the so-called *Gibbs free energy*. It can be thought of as a kind of chemical energy; the system tries to minimize its chemical energy by letting the reactions proceed ($\to$ expectation:

φ monotonically decreasing along solutions). The additive constant $e^{1-\mu_i}$ merely shifts the zero level of φ, which leads to having $c_2 = 0$ in item 2 (one can also work without an additive constant). Existence of a solution $\vec{\mu}$ of the above LGS: For this we assume that the columns of S are linearly independent; it is then $\mathrm{rang}(S) = \mathrm{rang}(S) = \tilde{J}$, thus the image of S is the entire $\mathbb{R}^{\tilde{J}}$, i.e., the linear system of equations is solvable for any right-hand side (the solution is in general not unique; one solution is obviously $\vec{\mu} = -S(S^TS)^{-1} \ln \vec{K}$; the full set of solutions is $\mathrm{kernel}(S^T) - S(S^TS)^{-1} \ln \vec{K}$).

Theorem. For the batch problem with reversible mass action kinetics, the solution on each interval of existence $[0, T)$ is bounded by a constant (which is independent of T).

Proof. We compute $\frac{\partial \varphi}{\partial c_i} = \mu_i + \ln c_i$, i.e.,

$$\nabla \varphi(\vec{c}) = \vec{\mu} + \ln \vec{c},$$

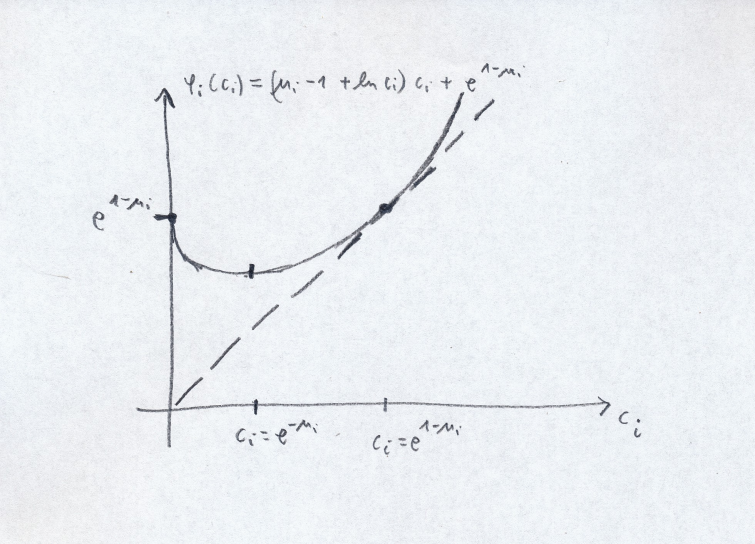
where the ln is to be understood component-wise. Since the solution $t \mapsto \vec{c}(t)$ is strictly positive, we can form $\varphi(\vec{c}(t))$ and even $\nabla\varphi(\vec{c}(t))$. (Note: φ is also defined by continuous continuation on the *edge* of the positive rectand, but $\nabla\varphi$ is not). We obtain

$$\begin{aligned}
\frac{d}{dt}\varphi(\vec{c}(t)) &= \langle \nabla\varphi(\vec{c}(t)), \vec{c}\,'(t)\rangle \\
&= \langle \vec{\mu} + \ln \vec{c}(t), S\vec{R}(\vec{c}(t))\rangle \\
&= \langle S^T(\vec{\mu} + \ln \vec{c}(t)), \vec{R}(\vec{c}(t))\rangle \\
&= \langle -\ln \vec{K} + S^T \ln \vec{c}(t), \vec{R}(\vec{c}(t))\rangle \\
&= \sum_{j=1}^{J} [-\ln K_j + (S^T \ln \vec{c}(t))_j]\, R_j(\vec{c}(t)) \\
&= \sum_{j=1}^{J} \underbrace{\left[-\ln \overbrace{K_j}^{=\frac{k_j^v}{k_j^r}} + \sum_{i=1}^{I} (s_{ij}^p - s_{ij}^e) \ln c_i(t) \right]}_{\text{(I)}} \underbrace{\left[k_j^v \prod_{i=1}^{I} c_i(t)^{s_{ij}^e} - k_j^r \prod_{i=1}^{I} c_i(t)^{s_{ij}^p} \right]}_{\text{(II)}}
\end{aligned}$$

where

$$\begin{aligned}
\text{(II)} \gtreqless 0 \quad &\Longleftrightarrow \quad \ln k_j^v + \sum_{i=1}^{I} s_{ij}^e c_i \gtreqless \ln k_j^r + \sum_{i=1}^{I} s_{ij}^p c_i \\
&\Longleftrightarrow \quad \text{(I)} \gtreqless 0
\end{aligned}$$

Hence, $\varphi \circ \vec{c}$ monotonically decreasing, and therefore $\varphi(\vec{c}(t)) \le \varphi(\vec{c}(0)) =: \varphi_0$.
It remains to infer from the boundedness of $\varphi \circ \vec{c}$ from the boundedness of $\vec{c}$. For this purpose a short analysis of the curve φ:



We write

$$\varphi(\vec{c}) = \sum_{i=1}^{I} \varphi_i(c_i) \quad \text{mit } \varphi_i(x) := (\mu_i - 1 + \ln x)\, x + e^{1-\mu_i}\,.$$

Obviously, $\lim\limits_{x\to 0} \varphi_i(x) = e^{1-\mu_i}$, $\lim\limits_{x\to\infty} \varphi_i(x) = \infty$. We obtain the only extremal (=minimal) point by

$$\varphi_i'(x) = 0 \quad \Leftrightarrow \quad \mu_i + \ln x = 0 \quad \Leftrightarrow \quad x = e^{-\mu_i}\,.$$

Thus, the minimum value of φ_i is $\varphi_i(e^{-\mu_i}) = e^{-\mu_i}(e-1)$.
Thus $\varphi : (\mathbb{R}_0^+)^I \to \mathbb{R}$ has a positive lower bound. Furthermore, $\frac{d}{dx}(\varphi_i(x) - x) = \mu_i + \ln x - 1 = 0 \iff x = e^{1-\mu_i}$ and $(\varphi_i(x) - x)|_{x=\exp(1-\mu_i)} = 0$, that is, the graph of φ_i touches the identity $f(x) = x$ exactly once (at the point $x = e^{1-\mu_i}$) and runs above this line otherwise. Thus

$$\varphi_i(x) \geq x \quad \forall x \in \mathbb{R}_0^+.$$

Together with the non-negativity of φ_j we obtain

$$\varphi(\vec{c}) \geq \sum_{i=1}^{I} c_i, \quad \text{in particular } \varphi(\vec{c}) \geq c_i \quad \forall i = 1, ..., I.$$

□

The stoichiometric space and illustration of the two approaches to the boundedness of solutions.
Integration of the ODE system yields

$$\vec{c}(t) - \vec{c}(0) = S \int_0^t \vec{R}(\vec{c}(\tau))\, d\tau\,,$$

thus the vector $\vec{c}(t) - \vec{c}(0)$ always lies in the space $\text{image}(S) = \text{kernel}(S^T)^\perp$. Thus, the solution cannot leave the affine subspace

$$\vec{c}(t) \in \vec{c}(0) + \text{image}(S) = \vec{c}(0) + \text{kernel}(S^T)^\perp\,.$$

The vector space $\mathcal{S} := \text{image}(S) = \text{kernel}(S^T)^\perp$ is called *stoichiometric space*. The affine space $\vec{x} + \mathcal{S}$ is called the *stoichiometric class* of the vector $\vec{x} \in \mathbb{R}^I$.

If there is a vector $s^\perp \in \mathcal{S}^\perp$ with all entries being strictly positive (which, see above, is the case, e.g., if the problem has an inherent 'conservation of the number of atoms'), then the intersection of the space $\vec{x} + \mathcal{S}$ with the positive rectand *must be bounded* (see Fig. 3, left). Since the solution must lie in both sets simultaneously, boundedness of solutions follows provided $\mathcal{S}^\perp \cap (\mathbb{R}_+)^I \neq \emptyset$, which geometrically illustrates the first approach to prove boundedness of solutions.
The second approach (Lyapunov technique) shows that the existence of such a vector is not necessary for *reversible* systems: All level lines of φ are bounded because of $\varphi(\vec{x}) \geq |\vec{x}|_1$, and since $\varphi(\vec{c}(t))$ can never be larger than $\varphi(\vec{c}(0))$ for monotonicity reasons, the solution cannot leave the bounded domain $(\vec{c}(0)+\mathcal{S}) \cap \{\vec{x} \mid \varphi(\vec{x}) \leq \varphi(\vec{c}(0))\}$ (see Fig. 3, right).

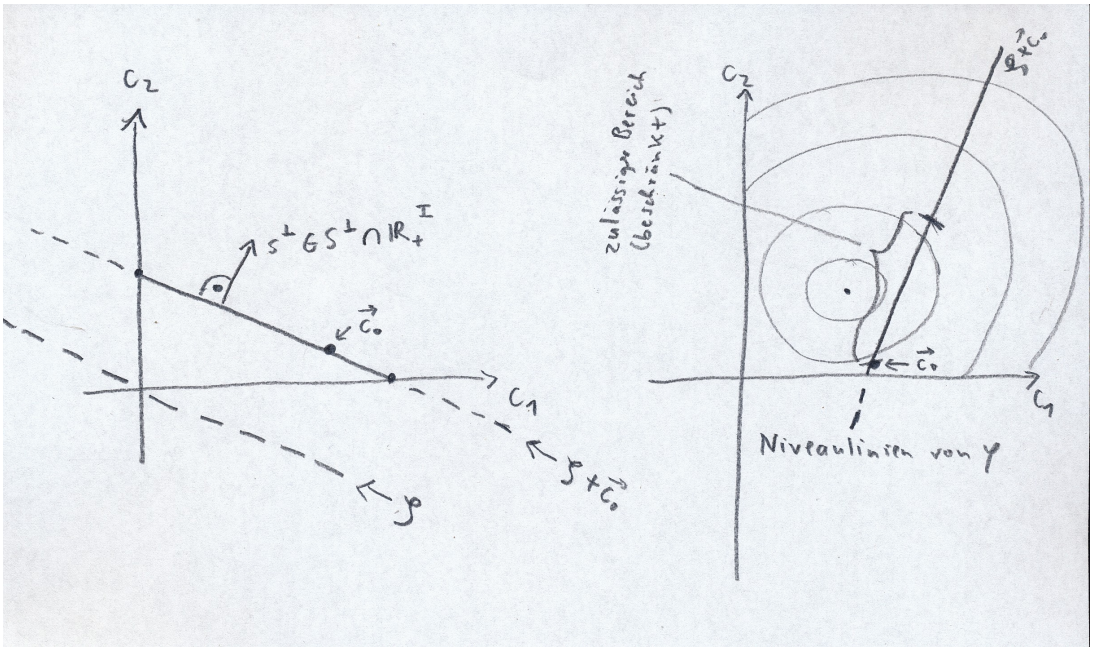


Abbildung 3: Left: Stoichiometric class in the case that $\exists \vec{s}^\perp \in \mathcal{S}^\perp \cap (\mathbb{R}_+)^I \neq \emptyset$ (i.e., in the case of an 'conservation of atomic entities' inherent to the reactive system, for example). Right: contour lines of φ in the case of 'arbitrary' orientation of $\vec{s}^\perp$. In both cases we get the boundedness of solutions.

5.3 Reaction invariants

For the sake of simplicity we assume in this chapter that the columns of S are linearly independent; however, this requirement can also be weakened; see remark below.
We decompose each vector $\vec{c} \in \mathbb{R}^I$ into a direct sum consisting of its projection $P_{\mathcal{S}}\vec{c}$

onto the space $\mathcal{S}$ and its projection $P_{\mathcal{S}^\perp}\vec{c}$ onto the space $\mathcal{S}^\perp$:

$$\vec{c} = P_{\mathcal{S}}\vec{c} \oplus P_{\mathcal{S}^\perp}\vec{c}$$

Since $\mathcal{S} = \text{image}(S)$ is spanned by the columns of S, one can write $P_{\mathcal{S}}$ as the linear combination of these columns with coefficients ξ_i; analogously for $P_{\mathcal{S}^\perp}$:

$$P_{\mathcal{S}}\vec{c} = S\vec{\xi}, \quad P_{\mathcal{S}^\perp}\vec{c} = U\vec{\eta}, \quad \vec{\xi}\in\mathbb{R}^J, \vec{\eta}\in\mathbb{R}^{I-J}$$

Where U is a $(I-J)\times I$ matrix whose columns form a basis of the $I-J$-dimensional space $\mathcal{S}^\perp$; thus, in particular, the matrix equations

$$U^T S = 0, \quad S^T U = 0$$

hold. Hence, we have the representation

$$\boxed{\vec{c} = S\vec{\xi} + U\vec{\eta}}, \qquad \vec{\xi}\in\mathbb{R}^J, \vec{\eta}\in\mathbb{R}^{I-J} \qquad (*)$$

A formula for the reversal of this coordinate transformation can be found in two ways. Either one knows the fact that an orthogonal projection onto a space $\mathcal{S} = \text{image}(S)$ can be written as $P_{\mathcal{S}}\vec{c} = S(S^TS)^{-1}S^T\vec{c}$ (note that S^TS is invertible due to the assumption that S has maximal column rank). Similarly, $P_{\mathcal{S}^\perp}\vec{c} = U(U^TU)^{-1}U^T\vec{c}$. It follows

$$\vec{c} = S\underbrace{(S^TS)^{-1}S^T\vec{c}}_{=\vec{\xi}} + U\underbrace{[U^TU]^{-1}U^T\vec{c}}_{=\vec{\eta}}$$

from which

$$\boxed{\vec{\xi} = (S^TS)^{-1}S^T\vec{c}}, \quad \boxed{\vec{\eta} = (U^TU)^{-1}U^T}$$

'can be read off'. Or, apply S^T to equation (*), then exploit $S^TU = 0$, and then apply the matrix $(S^TS)^{-1}$ to get the formula for $\vec{\xi}$; analogously for $\vec{\eta}$.

The ODE system

$$\frac{d}{dt}\vec{c}(t) = S\vec{R}(\vec{c}(t))$$

(not necessarily LMA, not necessarily reversible) can be transformed by multiplication on the one hand by $(S^TS)^{-1}S^T$, and on the other hand by $[U^TU]^{-1}U^T$ into the equivalent system

$$\begin{aligned}
\frac{d}{dt}\underbrace{[U^TU]^{-1}U^T\vec{c}}_{=\vec{\eta}(t)} &= [U^TU]^{-1}\underbrace{U^TS}_{=0}\vec{R}(\vec{c}(t)) \\
\frac{d}{dt}\underbrace{(S^TS)^{-1}S^T\vec{c}}_{=\vec{\xi}(t)} &= \underbrace{(S^TS)^{-1}S^TS}_{=\text{Id}}\vec{R}(\vec{c}(t))\,.
\end{aligned}$$

Using the new variables ξ, η this can be written

$$\begin{aligned} \frac{d}{dt}\,\vec{\eta}(t) &= \vec{0} \\ \frac{d}{dt}\,\vec{\xi}(t) &= \vec{R}(\vec{c}(t))\,. \end{aligned}$$

Hence, we have η known:

$$\vec{\eta}(t) = \text{const} = \vec{\eta}(0) = (U^T U)^{-1} U^T\, \vec{c}(0) =: \vec{\eta}_0\,.$$

We just have to solve the *smaller* system consisting only of J (instead of I) ODEs

$$\boxed{\tfrac{d}{dt}\,\vec{\xi}(t) = \vec{R}(S\vec{\xi}(t) + U\vec{\eta}_0)}.$$

The components η_i are called *reaction invariants*, the ξ_i *extents of reaction*. Geometrically, the ξ_i denote coordinates in the direction of the stoichiometric class, and the η_i denote coordinates perpendicular to the stoichiometric class. The fact that the η_i are constant fits to the previously derived fact that the solution never leaves the stoichiometric class of the initial value.

A note on a relaxation of the assumptions: If the columns of S are linearly dependent, then introduce a matrix S^* consisting of a maximal linearly independent subsystem of columns of S; we have $\text{image}(S^*) = \text{image}(S)$. Then there is[17] a matrix A such that $S = S^* A$. The transformation can then be performed on $\vec{\xi}$-$\vec{\eta}$-coordinates as well.

A benefit of the $\vec{\xi}$-$\eta\eta$-reformulation of this section is that numerically solving the smaller 'reduced' problem is faster. This is especially interesting when it comes to (time-consuming) solution of *PDE* systems (transferring chap. 5.3 to PDEs: see later).

5.4 Equilibrium reactions

We first consider the ODE problem without specifying initial values. We look for *equilibrium solutions*, i.e., solutions with $\vec{0} \stackrel{!}{=} \vec{c}\,'(t) = S\vec{R}(\vec{c}(t))$, thus vectors $\vec{c} \in (\mathbb{R}_0^+)^I$ with

$$S\vec{R}(\vec{c}) = \vec{0},$$

thus $\vec{R}(\vec{c}) \in \text{kernel}(S)$. We now assume that the columns of S are linearly independent. Then by multiplication[18] with $(S^T S)^{-1} S^T$ the necessary condition

$$\vec{R}(\vec{c}) = \vec{0}$$

[17] Multiply the equation $S = S^* A$ by $((S^*)^T S^*)^{-1} (S^*)^T$ to see that $A = ((S^*)^T S^*)^{-1} (S^*)^T S$.

[18] alternative reasoning: Then $\vec{R}(\vec{c}) \in \text{kernel}(S) = \{\vec{0}\}$

can be derived. If we assume in addition that the system *is reversible*, we obtain

$$\vec{c}\,\text{is equilibrium solution} \quad \Longleftrightarrow \quad \forall\, j=1,...,j: R_j^v(\vec{c}) = R_j^r(\vec{c})\,,$$

i.e., each individual forward-backward reaction must then be at equilibrium. In the case of law of the mass action, the equilibrium condition in logarithmic form is

$$\ln \vec{k}^v + (S^v)^T \ln \vec{c} = \ln \vec{k}^r + (S^r)^T \ln \vec{c},$$

which can be written

$$\boxed{\ln \vec{K} + S^T \ln \vec{c} = 0}\,.$$

The solution of this eqution is of course usually not unique, since there are J equations for I unknowns.
However, if we now additionally postulate a concrete initial value $\vec{c}_0 \in \mathbb{R}^I_+$, this determines the stoichiometric class in which we are looking for an equilibrium solution, which corresponds to $I-J$ additional conditions $\eta_i = \eta_{i,0}$. Although the equilibrium conditions are nonlinear, under the assumptions made, we can show that now the equilibrium solution in a stoichiometric class (determined by the initial value) exists and is uniquely determined:

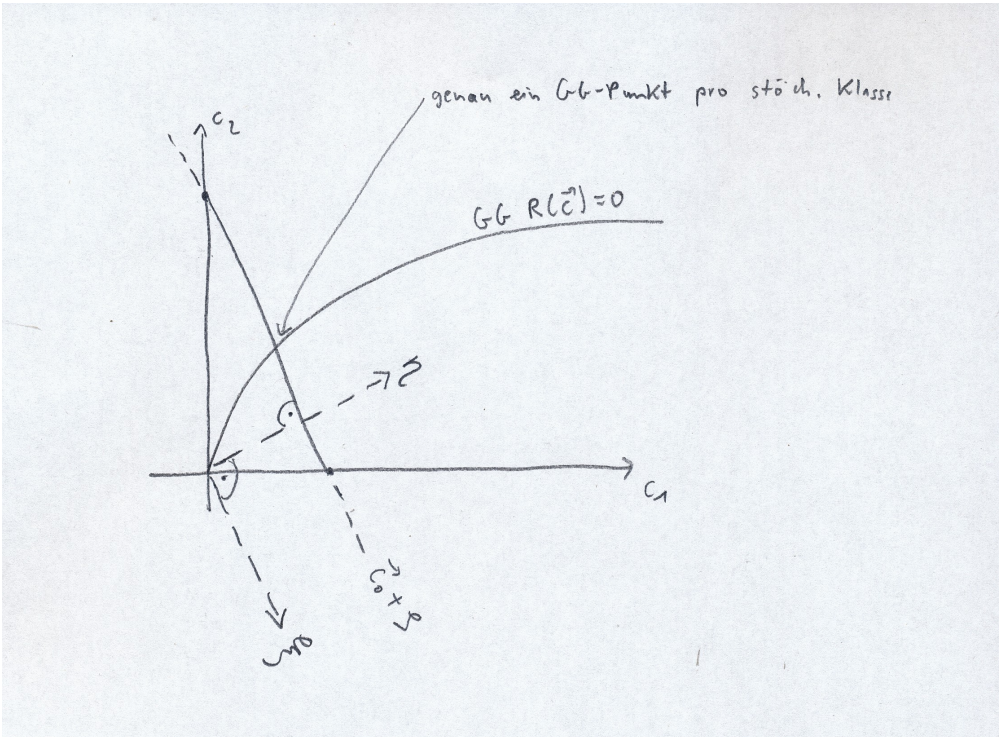


Abbildung 4: For the reversible system $\mathrm{X}_1 \longleftrightarrow 2\mathrm{X}_2$, i.e., $S = (-1, 2)^T$, and the equilibrium condition $k_1^v c_1 = k_1^r c_2^2$: Location of stoichiometric class(es) $(\vec{c}_0 + \mathcal{S}) \cap \mathbb{R}^I_+$ and location of equilibrium points; there is exactly one equilibrium point in each class..

Theorem. For reversible mass action systems where the stoichiometric matrix has maximum column rank, there exists exactly one equilibrium point in each nonempty

stoichiometric class $(\vec{c}_0+\mathcal{S}) \cap \mathbb{R}^I_+$.

Proof. We first consider the auxiliary problem

"Minimize the functional φ from Sec. 5.2 under the constraint $\vec{\eta} = (U^T U)^{-1} U^T \vec{c} \stackrel{!}{=} \vec{\eta}_0$"

(It is suggests itself to conjecture that this problem is equivalent to our equilibrium problem). Using the Lagrangian formalism, we obtain the equivalent[19] system

$$\nabla\varphi(\vec{c}) = (U^T U)^{-1} U^T]^T \vec{\lambda}, \quad (U^T U)^{-1} U^T \vec{c} = \vec{\eta}_0$$

which is equivalent to

$$\mu + \ln \vec{c} = U\,(U^T U)^{-1} \vec{\lambda}, \quad (U^T U)^{-1} U^T \vec{c} = \vec{\eta}_0\,.$$

The first of the two equations is equivalently transformed by multiplication, on the one hand by U^T, on the other hand by S^T. We obtain the three equations

$$\underbrace{S^T(\mu + \ln \vec{c}) = 0}_{= \text{ equilibrium condition}}\,, \quad U^T(\mu + \ln \vec{c}) = \vec{\lambda}\,, \quad \underbrace{(U^T U)^{-1} U^T \vec{c} = \vec{\eta}_0}_{\Leftrightarrow\ \vec{c} \in \vec{c}_0 + \mathcal{S}}\,.$$

We can drop the second of the three equations since it only defines the (uninteresting) $\vec{\lambda}$.
We see: The above constrained minimization problem is equivalent to finding equilibrium points in the stoichiometric class.
For the constrained minimization problem, in turn, one can show existence and uniqueness of the solution relatively easily: The "admissible set" (i.e., the set described by the constraint) is convex. The objective function φ is strictly convex, because the Hessian matrix

$$H\varphi(\vec{c}) = \mathrm{diag}(\frac{1}{c_1}, ..., \frac{1}{c_I})$$

is positive definite. Thus, the constrained minimization problem has at most one solution. To show the existence of a solution, it is sufficient that there exists a nonempty compact level set

$$M_l := \{\vec{c} \in (\mathbb{R}^+_0)^I \mid \varphi(\vec{c}) \leq l\}\,.$$

That this is true follows from the estimate $\varphi(\vec{c}) \geq |\vec{c}|$ from Sec. 5.2, since this estimate involves $M_l \subseteq K_l(\vec{0})$, thus yielding boundedness, and since $\vec{c}$ is continuous on the nonnegative closed rectand, yielding closedness of M_l. □

[19] Generally, the Lagrangian equations are only a *necessary* criterion for solutions of constrained optimization problems. But since our minimization problem is *convex*, every critical point must be a minimum point.

6 Feinberg's network theorie

6.1 Introduction

The main goal of Feinberg's network theory is to provide statements about **existence and uniqueness of equilibrium states** of the batch problem. We have already seen in Chap. 5.4 that under the assumption of reversibility existence and uniqueness of an equilibrium state can be shown in any nonempty stoichiometric class. Feinberg succeeded in weakening the assumptions; an essential role is played by the notion of *weak reversibility*. Interestingly, in Feinberg's network theory, a mathematical graph is set up for each reactive system, and criteria for existence/uniqueness of equilibrium states are formulated using graph–theoretic notions.
Let us learn to set up the graph for a reactive system by using an example:
Let us consider the chemical reactions

$$\begin{aligned} X_1 + X_2 &\rightleftarrows X_3 \\ X_3 &\to X_4 + X_5 \\ X_6 &\rightleftarrows X_4 + X_5 \\ 2X_2 &\to X_2 + X_7 \\ X_8 &\to 2X_2 \\ X_8 &\rightleftarrows X_2 + X_7 \end{aligned} \qquad I = 8 \text{ species}$$

What is on one side of a chemical reaction arrow, we call a *complex*, e.g., $X_2 + X_7$. A complex has *no* concentration. Complexes are introduced only as a mathematical tool; they do not correspond to any physical substance.

Representation as a directed graph:

$$\begin{aligned} \text{complexes} &\mathrel{\widehat{=}} \text{nodes} \\ \text{reactions} &\mathrel{\widehat{=}} \text{edges} \end{aligned}$$

In the example:

$$\begin{array}{ccc} X_1 + X_2 & \rightleftarrows & X_3 \to X_4 + X_5 \rightleftarrows X_6 \\ 2X_2 & \to & X_2 + X_7 \\ \nwarrow & & \swarrow\nearrow \\ & X_8 & \end{array}$$

Even if a complex occurs more than once in the reactions, it is represented in the graph by *one* node. Here we have $n = 7$ nodes/complexes and $J = 9$ edges/reactions (counting "$\rightleftarrows$" as two edges).

We can take the graph to be a "structure" in $\mathbb{R}^I$ ($I = 8$):
Nodes/complexes can be taken as elements of $(\mathbb{R}_0^+)^I$; e.g., $X_1 + X_2 \mathrel{\widehat{=}} \vec{e}_1 + \vec{e}_2 = (1, 1, 0, 0, 0, 0, 0)^T$, where $\vec{e}_1, \ldots, \vec{e}_I$ are the standard basis vectors of $\mathbb{R}^I$.

Thus: node set/complex set $\mathcal{C} \subseteq (\mathbb{R}_0^+)^I$
Note: The components of the vector are just the *stoichiometric coefficients* s_{ij}^e, s_{ij}^p from the previous chapters!

On the one hand, the set of edges/reactions can be understood as a *relation* $\mathcal{R} \subseteq \mathcal{C} \times \mathcal{C}$ on the set $\mathcal{C}$. Besides, the edges/reactions have also an interpretation as direction vectors=the difference of the endpoints, thus to be taken as elements of $\mathbb{R}^I$: For example, the reaction"$2X_2 \rightarrow X_2 + X_7$" corresponds to $\vec{e}_2 + \vec{e}_7 - 2\vec{e}_2 = \vec{e}_7 - \vec{e}_2 = (0, -1, 0, 0, 0, 0, 1, 0)^T \in \mathbb{R}^I$
Note: The components of the vector are just the stoichiometric coefficients $s_{ij} = s_{ij}^p - s_{ij}^e$ from the previous chapters. Since each node/complex can act as both reactant and product, the notation $(\vec{s}_i, \vec{s}_j) \in \mathcal{R}$ for a reaction proceeding from node/complex $\vec{s}_i$ to node/complex $\vec{s}_j$ makes more sense; the associated rate function is denoted $R_{(\vec{s}_i, \vec{s}_j)}$ or $R_{\vec{s}_i \to \vec{s}_j}$.

Our model, which reads

$$\vec{c}\,'(t) = (S^p - S^e)\, \vec{R}(\vec{c}(t)) = \sum_{j=1}^{J} (\vec{s}_j^{\,p} - \vec{s}_j^{\,e})\, R_j(\vec{c}(t))$$

in our old notation, now reads

$$\vec{c}\,'(t) = \sum_{(\vec{s}_i, \vec{s}_j) \in \mathcal{R}} (\vec{s}_j - \vec{s}_i)\, R_{\vec{s}_i \to \vec{s}_j}(\vec{c}(t))$$

or

$$= \sum_{(\vec{s}, \vec{s}\,') \in \mathcal{R}} (\vec{s}\,' - \vec{s})\, R_{\vec{s} \to \vec{s}\,'}(\vec{c}(t))$$

in Feinberg's notation.

Def. (Chemischal Reaction Network, Reactive System) A chemical reaction network (RNW) is a tripel $(\mathcal{M}_\mathcal{S}, \mathcal{C}, \mathcal{R})$, where
$\mathcal{M}_\mathcal{S} = \{1, 2 \ldots, I\}, I \in \mathbb{N}$, is the *set of species*,
$\mathcal{C} \subset \bar{\mathbb{R}}_+^N$ where $|\mathcal{C}| =: n \in \mathbb{N}$ is the (finite) *set of complexes*,
and where $\mathcal{R} \subset \mathcal{C} \times \mathcal{C}$, the set of reactions, is a *relation* on $\mathcal{C}$ with the properties $(\vec{s}, \vec{s}) \notin \mathcal{R}\ \forall\ \vec{s} \in \mathcal{C}$; (generally the relation is not symmetric, i.e., the graph is *oriented*). If, forthermore, we have a rate function given for each rate/edge $(\vec{s}, \vec{s}\,') \in \mathcal{R}$, and the rate functionis continuously differentiable $R_{\vec{s} \to \vec{s}\,'} : (\mathbb{R}_0^+)^I \longrightarrow \mathbb{R}_0^+$ and satisfies the minimum requirement of Sec. 4.2, $R(\vec{c}) > 0 \Leftrightarrow \text{supp}\ \vec{s}^{\,e} \subseteq \text{supp}\ \vec{c}$, Feinberg calls the RNW a *reactive system.*

6.2 Weak reversibility, linkage classes, rank, deficiency

The definition of reversibility in Feinberg's notation reads:

A RNW is *reversible*, if and only if

$$\forall \vec{s}, \vec{s}\,' \in \mathcal{C} : \ (\vec{s}, \vec{s}\,') \in \mathcal{R} \ \Rightarrow (\vec{s}, \vec{s}\,') \in \mathcal{R}.$$

(This corresponds to the symmetry of the relation $\mathcal{R}$.)

Def. (weak reversibility). An RNW is called *weakly reversible* if for all $\vec{s}, \vec{s}\,' \in \mathcal{C}$ for which there is a directed path in the graph from $\vec{s}$ to $\vec{s}\,'$ there is a directed path from $\vec{s}\,'$ to $\vec{s}$:

$$\begin{aligned} \forall \vec{s}, \vec{s}\,' \in \mathcal{C} : \big(\exists : m \in \mathbb{N}_0, \vec{s}_1, ..., \vec{s}_m \in \mathcal{C} : \ & (\vec{s}, \vec{s}_1), (\vec{s}_1, \vec{s}_2), ..., (\vec{s}_{m-1}, \vec{s}_m), (\vec{s}_m, \vec{s}\,') \in \mathcal{R} \\ \Longrightarrow \exists : m' \in \mathbb{N}_0, \vec{s}_1\,', ..., \vec{s}\,'_{m'} \in \mathcal{C} : \ & (\vec{s}\,', \vec{s}_1\,'), (\vec{s}_1\,', \vec{s}_2\,'), ..., (\vec{s}\,'_{m'-1}, \vec{s}\,'_{m'}), (\vec{s}\,'_{m'}, \vec{s}) \in \mathcal{R} \big) \end{aligned}$$

Obviously, weak reversibility follows from reversibility.

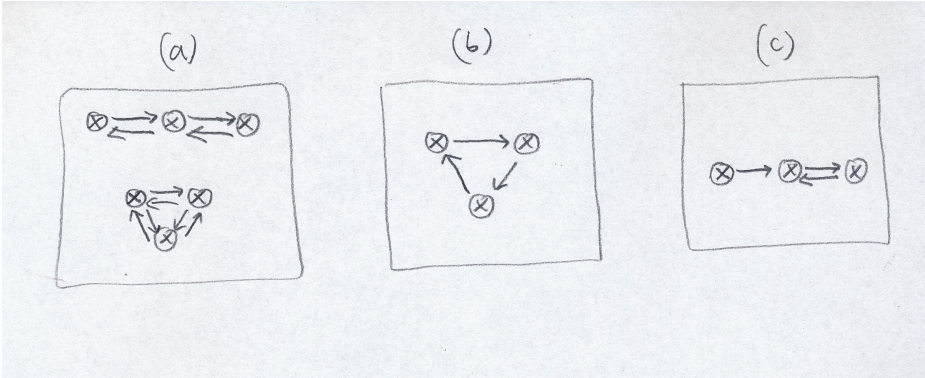


Abbildung 5: Network (a) is reversible, network (b) is weakly reversible but not reversible, network (c) is not weakly reversible but connected.

To the directed graph one can obviously make an *undirected graph* $(\mathcal{M}_\mathcal{S}, \mathcal{C}, \tilde{\mathcal{R}})$ by 'omitting the arrowheads', which thus has the same set of nodes and which has the set of edges

$$\tilde{\mathcal{R}} \subset \mathcal{C} \times \mathcal{C}, \quad (\vec{s}_1, \vec{s}_2) \in \tilde{\mathcal{R}} :\Longleftrightarrow (\vec{s}_1, \vec{s}_2) \in \mathcal{R} \ \vee \ (\vec{s}_2, \vec{s}_1) \in \mathcal{R};$$

hence, the relation $\tilde{\mathcal{R}}$ is symmetric.

Def. (linkage classes). We define another relation, $\sim$, on $\mathcal{C}$:

$$\begin{aligned} \vec{s} \sim \vec{s}\,' \ &:\Longleftrightarrow \ \text{in the undirected graph there is a path}^{20} \text{ from } \vec{s} \text{ to } \vec{s}\,': \\ &\Longleftrightarrow \ \exists : m \in \mathbb{N}_0, \vec{s}_1, ..., \vec{s}_m \in \mathcal{C} : \ (\vec{s}, \vec{s}_1), (\vec{s}_1, \vec{s}_2), ..., (\vec{s}_{m-1}, \vec{s}_m), (\vec{s}_m, \vec{s}\,') \in \tilde{\mathcal{R}} \end{aligned}$$

'$\sim$' is obviously an equivalence relation. The equivalence classes of $\mathcal{C}$ under '$\sim$' are called *linkage classes* (LCs; German: Zusammenhangskomponenten, ZHKs) of the RNW. We denote the number of LCs by l.

[20] By 'path' we want to understand here and in the following 'path of length ≥ 0', i.e., each node shall be related to itself with respect to '$\sim$'

Example: The RNW

$$s_1 \to s_2 \leftarrow s_3$$
$$s_4 \leftrightarrows s_5$$

has $l = 2$ LCs.

Def. (Rank). The *rank* $s \in \mathbb{N}_0$ of an RNW is the dimension of the vector space spanned by the reactions, i.e., the stoichiometric space:

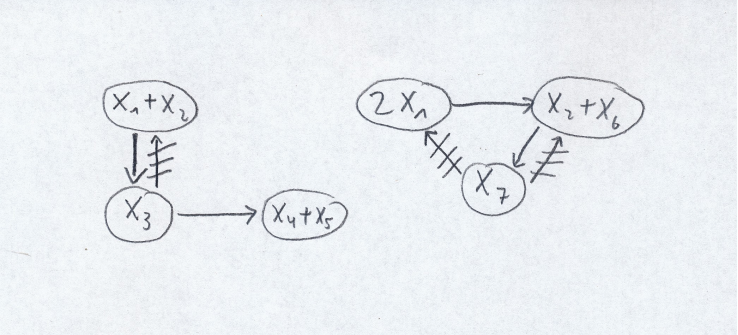
$$s := \dim(\mathrm{span}\{\vec{s} - \vec{s}' \mid (\vec{s}, \vec{s}') \in \mathcal{R}\}) = \dim(\mathrm{range}(S^p - S^e)) = \dim(\mathrm{range}(S)) = \dim(\mathcal{S})$$

Since $S \in \mathbb{R}^{I \times J}$, we have $0 \leq s \leq \min\{I, J\}$.

Obviously, the rank of an RNW *invariant* with respect to

- reversal of the direction of an edge ($\widehat{=}$ multiplication of a spanning vector by (-1)),
- insertion/removal of an edge, as long as this does not change the LCs ($\widehat{=}$ removal/addition of linearly dependent vectors in the spanning set[21] of $\mathcal{S}$).

Example. To the RNW



belongs the stoichiometric matrix

$$S = \begin{pmatrix} -1 & 1 & 0 & -2 & 0 & 0 & 2 \\ -1 & 1 & 0 & 1 & -1 & 1 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & -1 \end{pmatrix}$$

The first two columns are linearly dependent, as are the last four. Three columns/reactions can be deleted here without changing the LCs/rank of the RNW. If we delete the second and the last two columns from S, this corresponds to the deleted reactions in the graph shown in the sketch. The remaining four columns/reactions are linearly independent; thus, it is $s = 4$ here.

[21] German: Erzeugendensystem

The rank of an RNW depends on the edges only insofar as they define the LCs; one can always 'maximally thin' graphs without changing the rank. In a 'maximally thinned' graph, in each LC the number of edges is equal to the number of nodes $|\mathcal{C}_i|$ minus one. Thus, an upper bound on the rank is:

$$s \leq \text{Number of edges in the maximally thinned graph} = \sum_{i=1}^{l}(|\mathcal{C}_i|-1) = \underbrace{|\mathcal{C}|}_{=n} - l\,,$$

hence[22]

$$s \leq n - l.$$

Def. (deficiency). The number

$$\delta := n - l - s \;\; (\geq 0)$$

is called the *deficiency* of the RNW.

In the example above: $n=6$ complexes, $l=2$ LCs, $s=4 \;\; \Rightarrow \;\; \delta=6-2-4=0$

6.3 The Deficiency-Zero Theorem

Der zentrale Satz der Feinberg'schen Netzwerktheorie ist das Deficiency-Zero Theorem:

Theorem (Deficiency-Zero Theorem) For every RNW [23] with deficiency $\delta=0$ the following holds true:

(a) If the RNW *is not weakly reversible*, then the system has no positive stationary solution.

(b) If the RNW *is weakly reversible* and mass action kinetics is assumed, then in any stoichiometric class that has a nonempty intersection with $\mathbb{R}^I_+$, there *exactly one* positive stationary solution, and this is asymptotically stable ('with respect to the stoichiometric class'[24]).

Sketch of the proof[25] The theorem is about existence/uniqueness of equilibria, more precisely, about *'species equilibria'*, i.e., vectors $\vec{c}$ with

$$0 \overset{!}{=} \frac{d\vec{c}}{dt} = \sum_{(\vec{s},\vec{s}')\in\mathcal{R}} (\vec{s}'-\vec{s})\, R_{\vec{s}\to\vec{s}'}(\vec{c}).$$

[22] Even without the above argument about thinning the graph should be clear: The space spanned by vectors connecting $|\mathcal{C}_i|$ many points can be at most $|\mathcal{C}_i|-1$-dimensional.

[23] not necessarily assuming mass action kinetics, but the 'minimum requirements' for rates should be met

[24] What stability of the stationary solution *with respect to the stoichiometric class* exactly means is clarified in an exercisein the tutorials

[25] More details from the proof: See my lecture notes 'Reaktive Netzwerke'.

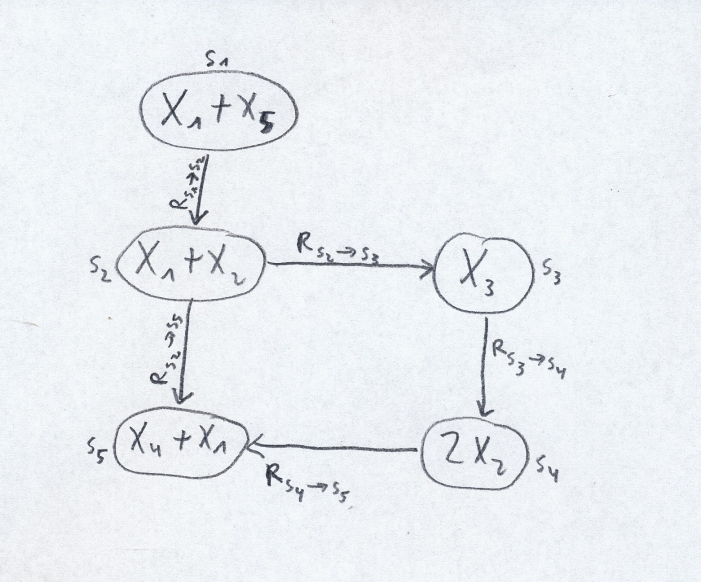


Abbildung 6:

In addition, there is the notion of *complex* or *nodal equilibrium*. These are vectors $\vec{c}$ for which holds: for each node $\vec{s} \in \mathcal{C}$, the rates leading toward and away from the node cancel, hence[26].

$$\forall \vec{s} \in \mathcal{C} : \; r_{\vec{s}}(\vec{c}) := \underbrace{\sum_{\substack{\vec{s}', \\ (\vec{s}', \vec{s}) \in \mathcal{R}}} R_{\vec{s}' \to \vec{s}}(\vec{c})}_{\text{towards node } \vec{s}} - \underbrace{\sum_{\substack{\vec{s}', \\ (\vec{s}, \vec{s}') \in \mathcal{R}}} R_{\vec{s} \to \vec{s}'}(\vec{c})}_{\text{away from } \vec{s}} = 0$$

Graphically, it is easy to see (and it is easy to show) that since any $\frac{d}{dt} c_i(t)$ can be written as a linear combination of 'nodal rates' (see above), that any nodal equilibrium is a species equilibrium. However, the inversion is generally wrong! For an example (see Fig. 6): There $\frac{d}{dt} c_2 = r_{\vec{s}_2}(\vec{c}) + 2 r_{\vec{s}_4}(\vec{c})$ (since X_2 occurs in these two complexes), but $r_{\vec{s}_2}(\vec{c}) + 2 r_{\vec{s}_4}(\vec{c})$ become zero without $r_{\vec{s}_2}(\vec{c})$ and $r_{\vec{s}_4}(\vec{c})$ both being zero (complex rates can be negative).
Feinberg shows, however, that *under the assumption* $\delta = 0$, the reverse direction also holds, i.e., when $\delta = 0$, $\vec{c}$ is species equilibrium if and only if it is complex equilibrium. (As a tool to obtain this intermediate assertion, he uses linear algebra; he constructs a matrix A that depends on a parameter $\vec{\rho} \in \mathbb{R}^J$ such that $\vec{c}$ is node equilibrium if and only if $(1, ..., 1)^T \in \text{core}(A_{\vec{\rho}})$, and $\vec{c}$ is species equilibrium if and only if $(1, ..., 1)^T \in \text{kernel}(S A_{\vec{\alpha}})$, where $\vec{\rho} := \vec{R}(\vec{c})$; see lecture script 'Reaktive Netzwerke' p. 36-38. [27])

[26] Note that nodes/complexes have no physical concentrations; thus, node rates are purely theoretical tools, but their consideration is quite obvious due to the mathematical concept of 'graph': edges are considered as 'pipelines', nodes as 'depots'.

[27] The linear mapping $A_{\vec{\rho}} : \mathbb{R}^n \to \mathbb{R}^n$ reads: $\vec{x} \longmapsto \sum_{\substack{\vec{s}', \\ (\vec{s}, \vec{s}') \in \mathcal{R}}} \rho_{\vec{s}'} x_{\vec{s}} (\vec{e}_{\vec{s}'} - \vec{e}_{\vec{s}})$, thereby we used component

After having that equivalence of species equilibrium and complex equilibrium, it is sufficient to investigate the existence and uniqueness of *complex equilibria*. The proof is very elaborate (lecture script 'Reaktive Netzwerke' p. 38-44); the proof of existence and uniqueness of this equilibrium for (b) is done in three steps:
(1.) Prove that (not necessarily in every class) in $\mathbb{R}^I_+$ there exists a complex equilibrium $\vec{c}_*$. (This is, after preliminary consideration, also a species equilibrium).
(2.) Prove the equality of the sets $\{\vec{c}\in\mathbb{R}^I_+ \mid \vec{c} \text{ is equiibrium}\}$ and $\{\vec{c}\in\mathbb{R}^I_+ \mid \ln\vec{c}-\ln\vec{c}_*\in \mathcal{S}^\perp\} =: E_{\vec{c}_*}$
(3.) Prove: $E_{\vec{c}_*}\cap(\vec{c}_0+\mathcal{S})$ consists of exactly one point.

Remark to point (2.): In the case of *'strong' reversibility* this set equality is immediately clear: Let $\vec{c}_*$ be an equilibrium point, so $\ln\vec{K}+S^T\ln\vec{c}_*=\vec{0}$. It is then $\vec{c}$ an equilibrium point if and only if $\ln\vec{K}+S^T\ln\vec{c}=\vec{0}$, i.e., if and only if $S^T(\ln\vec{c}-\ln\vec{c}_*)=\vec{0}$, which can be written as $\ln\vec{c}-\ln\vec{c}_*\in\text{kernel}(S^T)=\mathcal{S}^\perp$. (In the case of only weak reversibility it must be argued differently, since such a $\vec{K}$ does not exist then.)
Remark to point (3.): The existence and uniqueness of an element of $E_*\cap(\vec{c}_0+\mathcal{S})$ is shown, as already in the case of strong reversibility, with the help of a functional φ by showing that the element we are looking for is a solution of a minimization problem for φ under the constraint $\vec{c}\in\vec{c}_0+\mathcal{S}$. However, in the absence of a "$\vec{K}$" there is also no "$\vec{\mu}$", which in Sec. 5.4 (see also Sec. 5.2) for the definition of φ there; but here one can take instead the functional $\varphi(\vec{c}) := \sum_{i=1}^I(-\ln c_i^*-1+\ln c_i)\,c_i$. □

6.4 Other graph-theoretic terms and the Deficiency-One Theorem

We define a new relation on the directed graph:

$$\vec{s}_1\equiv\vec{s}_2 \iff \text{There is a directed path from } \vec{s}_1 \text{ to } \vec{s}_2 \text{ and one from } \vec{s}_2 \text{ to } \vec{s}_1$$

Obviously, (1) if $\vec{s}_1\equiv\vec{s}_2$, then $\vec{s}_1$ and $\vec{s}_2$ are in the same LC.
(2) '$\equiv$' is an equivalence relation.

The equivalence classes formed with respect to '$\equiv$' are called *strong linkage classes (strong LCs or SLCs)*. Because of (1), every LC is the union of one or more whole, disjoint strong SLCs. An SLC is called *terminal* if no path leads out of it. In the example of the sketch with $l=2$, there are 4 SLCs, of which $t=3$ are terminal.
One can easily show that a LC always must contain at least one strong terminal LC.

Since every LC can be considered as an independent RNW, each LC $Z_1,...,Z_l$ can be assigned a deficiency δ_k, $k=1,...,l$:

$$\delta_j := n_j - s_j - 1$$

notation: $\vec{x}=(x_{\vec{s}})_{\vec{s}\in\mathcal{C}}$, $\vec{\rho}=(\rho_{\vec{s}})_{\vec{s}\in\mathcal{C}}$

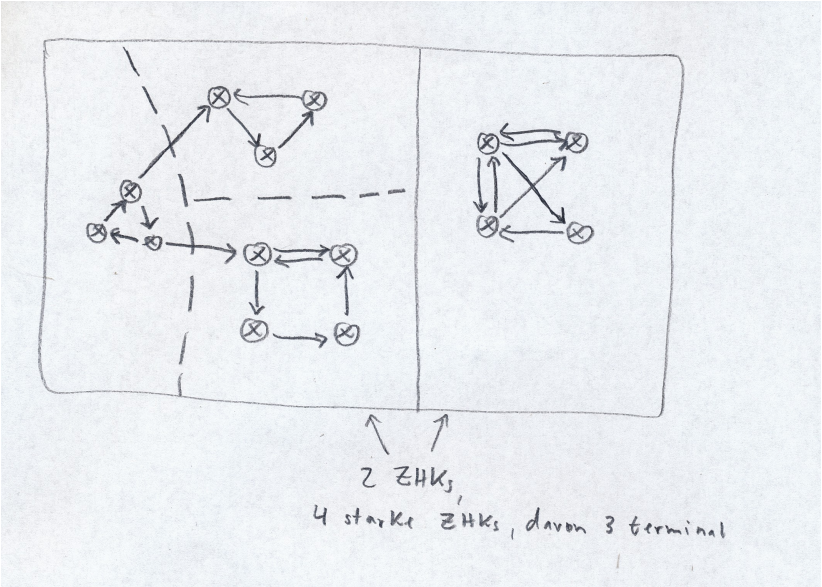


Abbildung 7: RNW with 2 LCs, 4 strong LCs, from which 3 are terminal.

where n_j is the number of nodes of the LC; s_j is the dimension of the space spanned by reactions going from nodes of the LC to nodes of the same LC, and the 1 is just the number of LCs that make up one LC. Addition yields

$$\sum_{k=j}^{l} \delta_j = n - \sum_{j=1}^{l} s_j - l \leq n - s - l = \delta, \quad \text{as } \delta_j \leq \delta \forall\, j.$$

Now another important theorem. Contrary to what the name suggests, it does not deal with networks that have $\delta = 1$:

Theorem (Deficiency-One Theorem) (Feinberg, 1987). We consider an RNW with mass action kinetics and $\delta \in \mathbb{N}_0$ arbitrary. Let $\mathcal{C}_1, ..., \mathcal{C}_l$ be the LCs with deficiencies $\delta_1, ..., \delta_l \in \mathbb{N}_0$. Let us assume that

(V1) $\delta_j \leq 1 \; \forall\, j = 1, ..., l$,

(V2) $\sum_{j=1}^{l} \delta_j = \delta$,

(V3) every LC contains exactly one strong terminal LC.

Then it holds true:

(a) Each stoichiometric class contains *at most one* positive equilibrium state.

(b) If the system has a positive equilibrium state (in any stoichiometric class), then *every* stoichiometric class $\vec{c}_0 + \mathcal{S}$ (which has nonempty intersection with $\mathbb{R}^I_+$) contains *exactly one* positive equilibrium state.

(c) If the RNW is additionally (V3') *weakly reversible*, then the system has a positive equilibrium state, so (b) comes into effect.

Remarks:

- (V3') is a tightening of the condition (V3), since

  (V3') $\Leftrightarrow$ weakly reversible $\Leftrightarrow$ every LC *is equal to a* terminal strong LC $\Rightarrow$ (V3)

- The preconditions of the Deficiency-One Theorem are relaxations of the precondition of the Deficiency-Zero Theorem:

  $$\delta = 0 \quad \begin{array}{c} \Longrightarrow \\ \not\Longleftarrow \end{array} \quad \text{(V1) \& (V2)}$$

  (Indeed Feinberg shows that (V1)–(V3) are sufficient for the equivalence of species equilibrium and complex equilibrium; the stronger requirement $\delta = 0$ from the Deficiency-Zerol Theorem is not really required for this).

- The two theorems only say something about existence/uniqueness of strictly positive equilibrium solutions; beyond this there are considerations by Feinberg also about the existence of equilibrium solutions which are *on the edge* of the positive rectand.

Here one could insert a chapter about equilibrium points at the edge of the positive rectand.

7 The PDE model

In this chapter, we will apply the ideas from Chap. 5 to the PDE model. We make the following assumptions throughout the chapter:

- All reactions are according to the law of mass action.
- All reactions are reversible.
- Dispersion-diffusion is species-independent: $L = \text{diag}(L_1, ..., L_I)$ with $L_1 = ... = L_I$
- Just to simplify the presentation: let the diffusion-dispersion coefficient be scalar and constant, also let the porosity be constant
- homogeneous Neumann boundary conditions (typical outflow boundary condition)
  (result can be transferred to so-called (inhomogeneous) flow boundary conditions $\rightarrow$ typical inflow boundary condition))

We start with a proof of the *uniqueness* of solutions. This is done using energy methods and Gronwall's lemma. The existence of solutions is more difficult to show. To prepare the existence result we first show the non-negativity of solutions. The subsequent proof of the existence of a (global!) solution of the PDE problem (in a function space still to be determined) uses a so-called *fixed point theorem*. The essential condition for the application of a fixed point theorem is that every potential solution satisfies a so-called *a priori bound*. The principle of a priori bounds/fixed point theorems is roughly as follows: If one can prove that any solution to the PDE (initial boundary value problem) satisfies a bound that depends only on the data, then the fixed point theorem provides the existence of a solution.
To prove the existence of an a priori bound, we reuse the Lyapunov functional from Chap. 5.2.
By 'existence of a *global* solution' we mean here that for *arbitrary* $T>0$ a solution can be found on the time interval $[0,T]$.

As function space, in which we search for solutions, we choose (see [WYW] p. 13) the 'anisotropic' Hölder space[28] $C^{2+\alpha,1+\frac{\alpha}{2}}(Q_T)^I$,

$$C^{2+\alpha,1+\frac{\alpha}{2}}(Q_T) := \{u \in C^{2,1}(\overline{Q}_T) \,|\, \sup_{(t,x),(s,y)\in Q_T} \frac{|u(t,x)-u(s,y)|}{(|t-s|^2+|x-y|)^{\frac{\alpha}{2}}} < \infty\}, \quad 0<\alpha<1.$$

Here $C^{2,1}(Q_T)$ denotes functions differentiable twice with respect to x and once with respect to t.[29] Other spaces with lower regularity are also usable[30] (which then also require weaker assumptions on the data), such as the Sobolev space

$$W_p^{2,1}(Q_T)^I := \{u : Q_T \longrightarrow \mathbb{R}^I \,|\, \|u_i\|_{W_p^{2,1}(Q_T)} < \infty\},$$
$$\|u_i\|_{W_p^{2,1}(Q_T)} := \left[\|u\|^p_{L^p(Q_T)} + \|\partial_t u\|^p_{L^p(Q_T)} + \sum_i \|\partial_{x_i} u\|^p_{L^p(Q_T)} + \sum_{i,j} \|\partial_{x_i}\partial_{x_j} u\|^p_{L^p(Q_T)}\right]^{\frac{1}{p}}$$
$$Q_T := (0,T]\times\Omega.$$

where $p \geq n+1$ is required for $W_p^{2,1}(Q_T) \subset W_p^1(Q_T)$ to be compactly embedded in $C(\overline{Q}_T)$; for domains M in the m-dimensional, the space $W_p^1(M)$ is compactly embedded in $C(M)$ for $p>m$ (note that $Q_T \subset \mathbb{R}^{n+1}$, i.e. $m=n+1$).

The choice of space becomes relevant only in Ch. 7.3-7.4; in Ch. 7.1-7.2 it suffices that 'all occurring terms' exist.

---

[28] vgl. [MiSi04]

[29] Classically, the Hölder space $C^{m+\alpha}(\overline{\Omega})$ for $m\in\mathbb{N}_0$ and $\alpha\in[0,1)$ is the space $C^{m+\alpha}(\overline{\Omega}) := \{u \in C^m(\overline{\Omega}) \,|\, \|u\|_{m+\alpha} < \infty\}$ with the norm $\|u\|_{m+\alpha} := \sum_{\substack{|\beta|\leq m \\ \beta\in(\mathbb{N}_0)^n}} \|D^\beta u\|_{L^\infty(\overline{\Omega})} + \sup_{\substack{x\neq y \\ x,y\in\overline{\Omega}}} \frac{|u(x)-u(y)|}{|x-y|^\alpha}$, $D^\beta := \partial^{\beta_1}_{x_1}...\partial^{\beta_n}_{x_n}$, $|\beta| := \sum_{i=1}^n \beta_i$.

[30] see [Habil]

7.1 Uniqueness of solutions

Let us start with the uniqueness of solutions, since this is much easier to show than the existence. We proceed similar as in the so-called *energy method*, where the PDE is 'tested' (i.e., multiplied) with the solution itself to get an 'energy estimate' of the solution. To show uniqueness, we use the difference of two solutions in the following.

Theorem Let $u, v \in W_p^{2,1}(Q_T)^I$, with $p > n+1$, $T > 0$ be solutions of the PDE system $\theta\partial_t u - d\Delta u + q \cdot \nabla u = SR(u)$ with $\theta = \text{const} > 0$, $d = \text{const} > 0$, $q \in C^0(Q_T)$, the reactions are according to the law of mass action and reversible. Let an initial value $u_{t=0} = u_0 \in L^2(\Omega)$ be given and Dirichlet or Neumann boundary values $u|_{[0,T]\times\partial\Omega} = g$ or $\frac{\partial u}{\partial \nu}\big|_{[0,T]\times\partial\Omega} = g$. Then $u = v$ holds.

Proof. Let $w := u - v$. Then w satisfies the PDE

$$\theta\partial_t w - d\Delta w + q \cdot \nabla w = SR(u) - SR(v), \quad w|_{t=0} = 0 \quad \text{with hom. b.c.}$$

Test the i-th PDE with w_i:

$$\frac{\theta}{2}\frac{d}{dt}\int_\Omega |w_i|^2\,dx + d\int_\Omega |\nabla w_i|^2\,dx - d\underbrace{\int_{\partial\Omega} w_i \frac{\partial w_i}{\partial \nu}\,do}_{=0} + \int_\Omega q\cdot\nabla w_i\, w_i\,dx = \int_\Omega [[SR(u)]_i - [SR(v)]_i] w_i\,dx$$

We now want to estimate the right-hand side by a constant times the L^2 norm of w squared, where the constant is allowed to depend on u and v. For simplicity, we perform this, somewhat unmathematically, using an *example* term; one can see that this is feasible for LMA rates in general: for the single rate $u \mapsto k u_1^3 u_2^2$, as an example, we expand the term

$$\begin{aligned}(k u_1^3 u_2^2 - k v_1^3 v_2^2)\, w_i &= k[(u_1^3u_2^2 - u_1^3u_2v_2) + (u_1^3u_2v_2 - u_1^3v_2^2) + (u_1^3v_2^2 - u_1^2v_1v_2^2) \\ &\quad + (u_1^2v_1v_2^2 - u_1v_1^2v_2^2) + (u_1v_1^2v_2^2 - v_1^3v_2^2)]\, w_i \\ &= k\,[u_1^3u_2w_2 + u_1^3v_2w_2 + u_1^2v_2^2w_1 + u_1v_1v_2^2w_1 + v_1^2v_2^2w_1]\, w_i\,.\end{aligned} \tag{7.1}$$

For the inserted terms, a v_i-power was successively increased by one and the corresponding u_i-power was increased by one at the same time. In general, the right-hand side can thus be written as $\int_\Omega \sum_{j=1}^I w_i w_j f_{ij}(t,x)\,dx$, where the f_{ij} are composed of the components of u and v. Since the $u_i, v_j \in W_p^{2,1}(Q_T) \subset C^\infty(\overline{Q}_T)$, we can estimate this using a constant c that depends on the $L^\infty(Q_T)$-norm of the f_{ij} (i.e., on the u_i, v_i):

$$\frac{\theta}{2}\frac{d}{dt}\int_\Omega |w_i|^2\,dx + d\int_\Omega |\nabla w_i|^2\,dx \le \|q\|_{L^\infty(Q_T)} \int_\Omega \sum_{i=1}^n |\nabla w_i|\,|w_i|\,dx + c\,\|w_i\|_{L^2(\Omega)}\|w\|_{L^2(\Omega)^I}$$

The 'mixed' term on the right-hand side is estimated using the inequality $ab \le \epsilon a^2 + \frac{1}{4\epsilon}b^2$ (which follows directly from $(\sqrt{\epsilon}a - \frac{1}{2\sqrt{\epsilon}}b)^2 \ge 0$):

$$\frac{\theta}{2}\frac{d}{dt}\int_\Omega |w_i|^2\,dx + d\int_\Omega |\nabla w_i|^2\,dx \le \|q\|_{L^\infty(Q_T)}\,\epsilon \int_\Omega |\nabla w_i|^2\,dx + \frac{\|q\|_{L^\infty(Q_T)^n}}{4\epsilon}\int_\Omega |w_i|^2\,dx + c\,\|w_i\|_{L^2(\Omega)}\|w\|_{L^2(\Omega)^I}$$

Summation over i yields

$$\frac{\theta}{2}\frac{d}{dt}\|w\|^2_{L^2(\Omega)^I} + (d - \epsilon\,\|q\|_{L^\infty(Q_T)^n})\sum_{i=1}^{I}\|\nabla w_i\|^2_{L^2(\Omega)^n} \le c\,\|w\|^2_{L^2(\Omega)^I}$$

We now choose $\epsilon > 0$ sufficiently small so that the coefficient of the 'energy term' is positive (or non-negative) (this technique is called *absorption*) and with $h(t) := \|w(t,\cdot)\|^2_{L^2(\Omega)^n}$ we get the differential inequality

$$h'(t) \le \frac{2c}{\theta}\,h(t)$$

with initial value $h(0)=0$, since $w|_{t=0}=0$.
Using the *Gronwall lemma*[31] it follows that

$$h(t) \le 0 \;\forall\, t\in[0,T]\,.$$

Hence, $h(t)\equiv 0$, thus $w\equiv 0$, thus $u\equiv v$. □

7.2 Nonnegativity of solutions

Even though we still don't know anything about the existence of solutions, we can show the nonnegativity of (potentially existing) solutions. We again use a form of the energy method; however, we test the PDE with the negative part of the solution.
We decompose (also vector-valued functions, then component-wise) $u = u^+ - u^-$ with $u^+ := \max\{0,u\}$, $u^- := \max\{0,-u\} = -\min\{u,0\}$; thus, we have $u^+, u^- \ge 0$. Let $\Omega_i^-(t) := \{x\in\Omega \,|\, u_i(t,x)<0\}$. For continuous $u_i : Q_T \longrightarrow \mathbb{R}$ this is a well-defined open set. In the following theorem we consider a slightly modified problem with right-hand side $SR(u^+)$ instead of $SR(u)$:

Theorem. Let $u\in W^{2,1}_p(Q_T)^I$ be a solution of the PDE problem $\theta\partial_t u + Lu = SR(u^+)$ with $u|_{t=0}\ge 0$ and homogeneous Neumann boundary values, on an interval of existence $[0,T]$. Let q be bounded in L^∞. Then $u\ge 0$ holds on Q_T.

Proof. Let u be a solution of the modified problem on an interval of existence $[0,T]$. We test the i-th equation of the system with u_i^- and get

$$\theta\int_\Omega u_i^-\partial_t u_i\,dx - d_i\int_\Omega u_i^-\,\delta u_i\,dx + \int_\Omega u_i^-\,\nabla u_i\cdot q\,dx = \sum_{j=1}^{J}(s^p_{ij}-s^e_{ij})\,[R^v_j(u^+)-R^r_j(u^+)]\,u_i^-$$

$$= \int_\Omega \sum_{j=1}^{J}\overbrace{\Big[\underbrace{s^p_{ij}R^v_j(u^+)}_{=s^p_{ij}k^v_j\prod_{k=1}^{I}(u_k^+)^{s^e_{kj}}} + \underbrace{s^e_{ij}R^r_j(u^+)}_{=s^e_{ij}k^r_j\prod_{k=1}^{I}(u_k^+)^{s^p_{kj}}}\Big]u_i^-}^{=:(\mathrm{I})} - \sum_{j=1}^{J}\overbrace{\Big[\underbrace{s^e_{ij}R^v_j(u^+)}_{=s^e_{ij}k^v_j\prod_{k=1}^{I}(u_k^+)^{s^e_{kj}}} + \underbrace{s^p_{ij}R^r_j(u^+)}_{=s^p_{ij}k^r_j\prod_{k=1}^{I}(u_k^+)^{s^p_{kj}}}\Big]u_i^-}^{=:(\mathrm{II})}\,dx$$

[31] Or, simpler, just use the technique 'separation of the variables' to solve this differential inequality.

All appearing factors and exponents are nonnegative. In the front term of (II), whenever the prefactor s_{ij}^e does not vanish, there is a factor $(u_i^+)^{s_{ij}^e}$ with nonvanishing exponent in the product, i.e., because of multiplication by u_i^-, the product is zero; analogously for the back term of (II). Thus, it is (II)$=0$. The term (I), on the other hand, does not vanish in general; it is in general (I)≥ 0. On the left-hand side we take advantage of the fact that $u_i^- = -u_i$ on $\Omega_i^-(t)$ and $=0$ otherwise, to obtain after multiplication by (-1)

$$\theta \int\limits_\Omega u_i^- \partial_t u_i^- \, dx - d_i \int\limits_{\Omega_i^-(t)} u_i \, \delta u_i \, dx + \int\limits_{\Omega_i^-(t)} u_i \, \nabla u_i \cdot q \, dx = -(I) \leq 0 \, .$$

Now we continue as in the proof of uniqueness in Sec. 7.1 and we get

$$\frac{\theta}{2} \frac{d}{dt} \int\limits_\Omega |u_i^-(t,x)|^2 \, dx \leq \text{const} \int\limits_\Omega |u_i^-(t,x)|^2 \, dx \, .$$

It follows, together with the nonnegativity of the initial value, i.e., $\int_\Omega |u_i^-(0,x)|^2 \, dx = 0$, and with Gronwall's lemma that $\int_\Omega |u_i^-(t,x)|^2 \, dx \leq 0$, thus $u_i^- \equiv 0$, thus $u_i \geq 0$. □

Any solution of the modified problem (with nonnegative initial values) is thus nonnegative, and is thus also solution of the original non-modified problem: $\mathbb{L}_{mod} \subseteq \mathbb{L}$. In Sec. 7.1 we have shown that $\mathbb{L}$ contains at most one element. Thus, we only need to show that $\mathbb{L}_{mod}$ contains at least one element to show $\mathbb{L}_{mod} = \mathbb{L}$ and thus the existence and uniqueness of the solution of the original problem, i.e., in the existence proof we can focus on the modified problem.
The existence of a solution of the modified problem is shown in Sec. 7.4; Sec. 7.3 is necessary as preparation.

7.3 A priori estimates

In preparation for the proof of the existence of a solution of the modified problem, we need a so-called a priori bound. We use the Lyapunov functional from Sec. 5.2 again here for this purpose. As a first step, we prove:

Lemma. Let $\varphi : (\mathbb{R}_0^+)^I \longrightarrow \mathbb{R}$ be as in Sec. 5.2. Let $g : Q_T \longrightarrow \mathbb{R}$ be defined by $g := \varphi \circ u$, where u is a solution of the modified problem on an interval of existence $[0, T]$. Then[32] holds

$$\theta \partial_t g + Lg \leq 0$$

[32] For ∇g to be well-defined, one needs the strict positivity of solutions u, whereas in Sec. 7.2 we have shown only the nonnegativity of solutions. As a way out, one can consider $g_\delta := \varphi \circ u_\delta$ instead of g, where $u_\delta(t,x) := u(t,x) + \delta$, for $\delta > 0$; it is ∇g_δ well-defined. The proof of the lemma then shows that $\theta \partial_t g_\delta + L g_\delta \leq f(\delta)$ holds, where f is a function with $f(\delta) \overset{(\delta \to 0)}{\longrightarrow} 0$; this bound is also sufficient for the following considerations.

where

$$Lu := (-d_i \delta u_i + q \cdot \nabla u_i)_{i=1,\dots,I}.$$

Proof.[33] With μ from chap. 5.2 and the chain rule we have

$$\begin{aligned}
\partial_t g &= ((\nabla\varphi)\circ u)\cdot \partial_t u = (\mu+\ln u)\cdot \partial_t u \\
\partial_{x_i} g &= (\mu+\ln u)\cdot \partial_{x_i} u = \sum_{k=1}^{I} (\mu_k+\ln u_k)\, \partial_{x_i} u_k \\
\partial_{x_i}^2 g &= \sum_{k=1}^{I} \underbrace{\frac{1}{u_k}(\partial_{x_i} u_k)^2}_{\geq 0} + \sum_{k=1}^{I} (\mu_k+\ln u_k)\, \partial_{x_i}^2 u_k \geq (\mu+\ln u_k)\cdot \partial_{x_i}^2 u \\
&\Longrightarrow -\Delta g \leq -(\mu+\ln u)\cdot \Delta u
\end{aligned}$$

Hence,

$$\begin{aligned}
\theta\partial_t g + Lg \leq (\mu+\ln u)\cdot(\theta\partial_t u + Lu) &\overset{\text{(PDE)}}{=} (\mu+\ln u)\cdot SR(u) = S^T(\mu+\ln u)\cdot R(u) \\
&= (-\ln K + S^T \ln u)\cdot R(u) \leq 0
\end{aligned}$$

where the nonpositivity follows at the end as in Sec. 7.2. □

Note that in the proof we used the species independence of the diffusion-dispersion coefficient.

We now use:

Theorem ((parabolic, weak) Maximum Principle). Suppose a function $g \in C^2(Q_T) \cap C(\overline{Q}_T)$ satisfies

$$\partial_t g - \sum_{i,j} a_{ij}(t,x)\partial_{x_i}\partial_{x_j} g + \sum_i b_i(t,x)\,\partial_{x_i} g \leq 0$$

with $\sum_{i,j} a_{ij}(t,x)\xi_i\xi_j \geq 0\ \forall\,\xi$ ('ellipticity'). Then

$$\max_{Q_T} g = \max_{\partial_p Q_T} g\,,$$

where $\partial_p Q_T := \overline{Q}_T \setminus Q_T$ is the so-called *parabolic boundary* of Q_T; $Q_T = (0,T]\times\Omega$. (That is, the maximum of g is assumed for $t=0$ or for $x\in\partial\Omega$.)[34]

.

Proof: See [Evans], Sec. 7.1.4.

[33] The proof roughly follows the argument in [MiSi04].

[34] No regularity is assumed for the coefficients of the PDE. It is essential, however, that g be 'smooth'; "$g\in H^2(Q_T)$" would *not* suffice!

Application of the maximum principle to g yields, provided we assume 'suitable' boundary conditions (see below), that g assumes its maximum at $t=0$, thus g (and thus also u!) can be estimated by the initial data $u_0 = u|_{t=0}$:

Theorem. Under the conditions of the above lemma, u on Q_T has a bound that depends only on the data:

$$0 \le u(t,x) \le c \ \ \forall\, (t,x) \in \overline{Q}_T$$

Proof. We form $\tilde{g}(t,x) := g(t,x) - \epsilon t/\theta$. From the above lemma it follows immediately that $\theta\tilde{g} + L\tilde{g} \le -\epsilon$. Application of the maximum principle to the function $\tilde{g}$ yields that (a) $\tilde{g}$ takes its maximum at a $(t_0,x_0) \in \{0\}\times\overline{\Omega}$ or (b) at a $(t_0,x_0) \in (0,T]\times\partial\Omega$. Suppose the case (b) occurs.
Since $\frac{\partial\tilde{g}}{\partial\nu} = \frac{\partial g}{\partial\nu} = (\mu + \ln u)\,\frac{\partial u}{\partial\nu}$ and since homogeneous Neumann boundary conditions were assumed for u, $\frac{\partial\tilde{g}}{\partial\nu} = 0$. Since (t_0,x_0) is supposed to be maximal, it follows $\frac{\partial\tilde{g}}{\partial t}(t_0,x_0) = 0$ and $\frac{\partial\tilde{g}}{\partial\tau}(t_0,x_0) = 0$ (with τ being an arbitrary tangent direction to $\partial\Omega$ in the point (t_0,x_0)), and $-\Delta\tilde{g}(t,x) \ge 0$ in an Ω-neighborhood[35] of (t_0,x_0). So it follows $\theta\tilde{g} + L\tilde{g} \ge 0$ in a Ω-neighborhood of (t_0,x_0). Contradiction. Thus, only case (a) can occur.
Thus, it is $\max\limits_{(t,x)\in\overline{Q}_T} \tilde{g}(t,x) \le \max\limits_{x\in\overline{\Omega}} \tilde{g}(0,x) = \max\limits_{x\in\overline{\Omega}} g(0,x) = \max\limits_{x\in\overline{\Omega}} \varphi(u_0(x))$. Furthermore, as stated in Sec. 5.2, $u(t,x) \le \varphi(u(t,x)) = g(t,x) \le \tilde{g}(t,x) + \epsilon T/\theta$. It follows the boundedness of u with the bound being $\frac{\epsilon T}{\theta} + \max\limits_{x\in\Omega} \varphi(u_0(x))$. Since this works for all $\epsilon > 0$, with $\epsilon \to 0$ the ϵ-term can even be dropped. [36] □

Alternative Strategies/Spaces. In weaker spaces,in which the maximum principle is not applicable, the function[37]

$$\psi_r : W \to \mathbb{R}_+, \quad \psi(u) := \int\limits_\Omega \varphi(u(\cdot,x))^r\, dx, \quad r \in \mathbb{N},$$

can be considered. One can show that $t \longmapsto (\psi_r \circ u)(t)$, where u is a solution, is monotonically decreasing (for inhomogeneous flux boundary conditions: that is has limited growth). For $r \ne 1$, the proof is similar to the above lemma, but more tedious, since additional terms appear with the differentiation (see [Habil]). As a reward for the effort, one gets a bound for the $L^\infty([0,T], L^r(\Omega))$-norm (in particular, therefore, for the $L^r(Q_T)$-norm) of the solution u. Such a bound is much more 'valuable' than a bound in $L^\infty([0,T], L^1(\Omega))$; see use of this bound at the end of the following chapter.

[35] At the point (t_0,x_0) itself Δg is not defined

[36] In the case where we proceed as in Footnote 32, we take $\tilde{g}_\delta(t,x) := g_\delta(t,x) - \epsilon t/\theta$ instead of $\tilde{g}$, and we choose ϵ, δ such that $f(\delta) < \epsilon$.

[37] Also here, if we want to be rigorous, in the definition of ψ_r, u must be replaced by $u_\delta := u + \delta$ unless it is a priori clear that u is strictly positive,

7.4 Existence of global solutions

One way to prove existence of solutions of nonlinear PDEs is to reformulate the PDE as a fixed point problem in a suitable function space, and then apply a fixed point theorem which provides the existence of a fixed point and thus a solution of the PDE. The typical conditions for applying a fixed point theorem are (besides some technical requirements, if necessary) the compactness of the fixed point operator or its domain of definiton, and the existence of an a priori bound for fixed points. (For ODEs, the existence of a global solution followed much more simply from the existence of an a priori bound by exploiting the existence of 'maximal' solutions of ODEs. But aalso for ODEs existence is based on a fixed point principle, such as Banach's fixed point theorem.)

A fixed point theorem which has quite few technical requirements and that is suited for PDE problems is Schaefer's fixed point theorem ([Schae55], see also [Evans] p. 504):
Theorem (Schaefer's fixed point theorem). Let X be a real Banach space and let $Z : X \longrightarrow X$ be compact. Furthermore, let the set

$$M := \{x \in X \,|\, \exists\, \lambda \in [0,1] :\ x = \lambda Z(x)\}$$

be bounded. Then Z has (at least) one fixed point.

Proof: See appendix; the theorem is traced back to Schauder's fixed point theorem. □

In the theorem, compactness of mappings is defined as follows.

Def. (compact mapping). A (usually nonlinear) mapping $Z : X \longrightarrow Y$ between two Banach spaces X, Y is called compact if it is continuous and for every bounded set $M \subset X$ it holds true that the set $\overline{Z(M)}$ is compact.[38]
Conversion of the PDE problem into a fixed-point problem. The given PDE system

$$\partial_t u + Lu = SR(u^+)$$

with initial condition $u|_{t=0} = u_0 \geq 0$ and boundary conditions can be written as a fixed point problem for the nonlinear(!) mapping

$$Z : X \longrightarrow X, \quad u \longmapsto v = Z(u),$$

where v solution of the (linear!) problem

$$\partial_t v + Lv = SR(u^+),$$

[38] And a set M, subset of a Banach space, is called compact, if every collection of sets whose union covers M contains a finite subcolllection that already covers M. Only if the Banach space is finite-dimensional, compactness of a set is equivalent to boundedness and closedness of the set.

with the corresponding initial and boundary conditions; the solution space X yet to be chosen. Obviously, an $x \in X$ is a fixed point of Z if and only if u solves the above PDE problem.
Thus, in order to apply Schaefer's fixed point theorem, we must (a) find a bound for the set M and (b) – by choosing a suitable function space X – ensure that Z is compact.

Ad (a): Boundedness of M: Let $u \in X$ and $v = Z(u)$ be solutions of the linear PDE above; let $w := \lambda v$. Thus, the set M is the set for which $u = w$. Multiplication of the linear PDE by λ gives that $\partial_t w + Lw = \lambda SR(u^+)$; further $w|_{t=0} = \lambda u_0$, and also in the boundary condition, if it is inhomogeneous, such a factor λ occurs. Thus, the set M is characterized by

$$M = \{x \in X \mid \exists\, \lambda \in [0,1] : \partial_t u + Lu = \lambda\, SR(u^+),\ u|_{t=0} = \lambda u_0, \text{ and boundary condition}\}.$$

Since λ is to be chosen from a bounded set $[0,1]$, all estimates from the earlier chapters (there, $\lambda = 1$) carry over to the case $\lambda \in [0,1]$. Hence, due to Sec. 7.3 we know that the set M is bounded under the assumptions made there.

Ad (b): Compactness of $Z : X \longrightarrow X$: To get well-definedness and compactness of Z, we use the 'regularizing effect' of solving a linear parabolic PDE

$$\partial_t v + Lv = f\,.$$

For example, we can choose the space $X := C^{2+\alpha,1+\frac{\alpha}{2}}(\overline{Q}_T)$, with $0 < \alpha < 1$. Now let $\alpha < \beta < 1$. Let $u \in X$. Trivially, it follows that $u \in C^{\beta,\frac{\beta}{2}}(\overline{Q}_T)$. It follows that the capped function $u^+ \in C^{\beta,\frac{\beta}{2}}(\overline{Q}_T)$. Thus all powers and polynomial expressions of u^+, hence also $SR(u^+) \in C^{\beta,\frac{\beta}{2}}(\overline{Q}_T)$. According to a theorem from the theory of linear parabolic differential equations, for right-hand side $f = SR(u^+) \in C^{\beta,\frac{\beta}{2}}(\overline{Q}_T)$ the solution $v \in C^{2+\beta,1+\frac{\beta}{2}}(\overline{Q}_T)$ ([WYW] Sec. 8.3.1). This space in turn is compactly embedded in X. (That Hölder spaces are compactly embedded in each other can be shown using Arzela-Ascoli's theorem).
If we take a bounded sequence (u_n) in X, the above argument shows that the associated sequence of solutions (v_n) has a subsequence convergent in X; Z is therefore compact.

The well-definiteness and compactness of Z also follows for the choice $X := W_p^{2,1}(Q_T)$ for $p > n+1$: This space, as already stated at the beginning of Sec. 7, is compactly embedded in $C^0(\overline{Q}_T)$; $u \in X \subset C^0(\overline{Q}_T)$ trivially entails $f = SR(u^+) \in C^0(\overline{Q}_T)$, hence $f \in L^p(Q_T)$. By a theorem on linear parabolic differential equations (see [WYW] Sec. 9.2.3 or [Lady68]), then the solution is $v \in W_p^{2,1}(Q_T) = X$. Taking again a bounded sequence (u_n) in X, the above argument shows that the associated sequence of solutions (v_n) has a subsequence convergent in X; thus Z is compact.
In this choice of space X, note that the a priori estimate for ψ_r (see end of Sec. 7.3) initially yields only an estimate of the solution v in the $L^\infty(0,T;L^r(Q_T))$, not in the

norm of $X = W_p^{2,1}(Q_T)$. However, by choosing $r = r(p)$ sufficiently large, it follows from $u \in L^\infty(0,T;L^r(Q_t))$ that $f = SR(u^+) \in L^p(Q_T)$, and the theory of linear parabolic differential equations thus yields that the solution $v \in W_p^{2,1}(Q_T) = X$ (see [Habil]).

7.5 Reactive transport with equilibrium reactions, derivation of a model, the instantaneous limit

We now want to transfer the problem with reactions in equilibrium, which we had considered in Sec. 5.4 for the batch problem, to the PDE setting. The starting point is the 'kinetic' PDE system

$$\partial_t u + Lu = SR(u) \tag{7.2}$$

for the I concentrations and the assumption that all J reactions are reversible; thus, $R_j(u) = R_j^v(u) - R_j^r(u)$ holds. Let $I < J$.

A First Attempt. In the long run, i.e., in the limit $t \to \infty$, one can conjecture (and also show) that a concentration vector u arises in such a way that the forward and backward rates compensate each other. If the reactions are very fast, one can expect such a state to set up very quickly. In an idealization (reaction rates extremely fast) we want to assume that such a state occurs 'instantaneously' (some authors use the expression 'the instantaneous limit' for this), i.e. that at all times and locations the equations $R_j^v(u) = R_j^r(u)$, $j = 1, ..., J$, i.e.

$$R(u) = 0\,, \tag{7.3}$$

are satisfied. In the case of the law of mass action and strict positivity of solutions, this can be expressed as

$$-\ln K + S^T \ln u = 0\,.$$

We thus have J equations for the I unknowns u_i, which certainly does not give a useful model. How can we get more equations? Naively substituting the equation $R(u) = 0$ into the PDE system (7.2) would give additional I homogeneous PDEs $\partial_t u_i + Lu_i = 0$; however, we would then have a total of $I + J$ equations to satisfy for I unknowns, which seems questionable. What is the correct model describing equilibrium?

Heuristic derivation of the equilibrium problem. In order to get the 'right' number of equations for our unknowns, unlike above, we proceed as follows: We first eliminate the fast reactions from as many of the PDEs as possible, i.e., we concentrate the rates on as few PDEs as possible. To achieve this, we adopt from Sec. 5.4 the coordinate transformation $u = S\xi + U\eta$, $\xi = (S^T S)^{-1} S^T u$, $\eta = (U^T U)^{-1} U^T u$. This yields,

analogously to Sec. 5.4, the system equivalent to the above kinetic PDE system[39]

$$\begin{aligned} \partial_t \eta + L\eta &= 0 \\ \partial_t \xi + L\xi &= R(S\xi + U\eta). \end{aligned}$$

Note that we have exploited here that the transport operator L and the stoichiometric matrices commute, which is the case only in the case of species-independent diffusion-dispersion coefficient.[40] The above η-PDEs describe the fact that for kinetic rates the solution never leaves the stoichiometric class $u_0 + \mathcal{S}$. These η-PDEs do not contain the rate coefficients, and also the creation of the η-variables was independent of thee rate coefficients. Hence, these equations hold for arbitrary rate coefficients, even for very large rate coefficients, and therefore it is plausible to assume that this should also be so in the limiting case of 'infinitely fast' reactions. These are $I-J$ equations. Together with the J equilibrium conditions (**??**), this gives I equations for the I unknowns u_i:

$$\begin{aligned} \partial_t \eta + L\eta &= 0 \\ R(S\xi + U\eta) &= 0 \end{aligned}$$

,,

or, equivalently,

$$\partial_t \eta + L\eta = 0 \tag{7.4}$$

$$S^T \ln(S\xi + U\eta) - \ln K = 0\,, \tag{7.5}$$

Note that the equations are *decoupled*, i.e., one can *first* solve the (scalar!, linear!) PDEs for the η_i and *subsequently*, at any point of the computational domain, solve the nonlinear algebraic equations for the ξ_i. The numerical computation of a solution is thus much less time-consuming than that of the nonlinear, fully coupled *kinetic* PDE problem – after all, we only have I–J many PDEs instead of I many, and these are linear instead of nonlinear. And algebraic equations have no couplings among the grid points in the computational domain; especially on parallel computers algebraic equations can be solved very efficiently without any communication among the processors.
The existence and uniqueness of a solution can be shown analogously to the ODE case: The solution to the η-equations exists and is unique, using reasonable assumptions and initial/boundary conditions[41]. (linear parabolic theory). Computing a $\xi(t,x)$ can

[39] This transformation is well-defined only in the case where the columns of S are linearly independent; however, in the other case, as mentioned in Sec. 5.4, a modification of the procedure is possible which allows a matrix A to appear in front of R to handle linear dependencies among the columns of S

[40] This assumption is actually needed whenever one wants to form linear combinations of the PDEs and introduce new variables; indeed, applying matrices to the PDE system is equivalent to forming linear combinations of the PDEs.

[41] We only need to set initial and boundary conditions for η. If we have initial/boundary conditions also for ξ – or in other words, if we have initial/boundary conditions for u – then these should be consistent with the equilibrium conditions

be rewritten as a (finite-dimensional!) optimization problem (in $\mathbb{R}^J$, we don't need a function space!) at every single point of the domain Q_T, and taking advantage of the properties of the objective function φ (including strict convexity), existence and uniqueness of $\xi(t,x)$ can be shown for every $(t,x) \in Q_T$. Alternatively, one can also invoke the theorem on implicit functions, which provides at least the existence of a local resolution function $\xi = \xi(\eta)$, since $\frac{\partial}{\partial \xi}[S^T \ln(S\xi + U\eta) - \ln K] = S^T \mathrm{diag}(u_1^{-1}, ..., u_I^{-1})S$, and this matrix is symmetric positive definite, thus invertible.
To interpret the model: the assumption that all reactions are at local (or: dynamical) equilibrium does not mean that these reactions cease to take place. It also does not mean that at any time at any location the forward and backward reaction rates exactly balance (r is not 0). The equilibrium condition $R(u) = 0$ describes a manifold on which the solution must lie, and if, for example, transport processes threaten to cause the solution to leave this manifold, then instantaneous reactions occur such that the solution remains on the manifold.
If one would like to calculate not only the concentrations u_i, but also the resulting reaction rates, one can do so by

$$r = \partial_t \xi + L\xi \tag{7.6}$$

a posteriori.
By the way, the transformation $u \mapsto (\eta, \xi)$ can now be unwound for the system (7.4)-(7.6); we obtain

$$\begin{aligned} \partial_t u + Lu &= Sr \\ S^T \ln u - \ln K &= 0, \end{aligned}$$

In this formulation, the I concentrations u_i and the J reaction rates r_j are the unknowns obtained as the solution of a system consisting of I linear PDEs and J nonlinear algebraic (equilibrium) equations. Thus, this formulation consists of more equations and unknowns than (7.4)-(7.5).

Generalization: Mixed equilibrium-kinetic problem. One can also consider reactive systems in which some recations are 'fast' and others are 'slow', i.e., one assumes local equilibrium for only part $j = 1, ..., J_{eq}$ of the reactions, and one continues to describe the remaining $J_{kin} = J - J_{eq}$ reactions $j = J_{kin} + 1, ..., J$ kinetically. In this case, one can first describe the initial problem as

$$\partial_t u + Lu = S_{eq} R_{eq}(u) + S_{kin} R_{kin}(u)$$

where we have decomposed $S = (S_{eq}|S_{kin})$ and $R_{kin}^T = (R_{eq}^T | R_{kin}^T)$. We now want to throw out only the fast reactions from as many PDEs as possible. To do this, we can now perform the previously used coordinate transformation but with S replaced by S_{eq}, i.e. $u = S_{eq}\xi + U\eta$, $\xi = (S_{eq}^T S_{eq})^{-1} S_{eq}^T u$, $\eta = (U^T U)^{-1} U^T u$, where now $U \in \mathbb{R}^{I \times (I - J_{eq})}$ is a matrix whose columns span $S_{eq}^{\perp}$; it is now $\xi \in \mathbb{R}^{J_{eq}}$, $\eta \in \mathbb{R}^{I - J_{eq}}$. One obtains

$$\begin{aligned} \partial_t \eta + L\eta &= (U^T U)^{-1} U^T S_{kin} R_{kin}(S_{eq}\xi + U\eta) \\ \partial_t \xi + L\xi &= R_{eq}(S_{eq}\xi + U\eta) + (S_{eq}^T S_{eq})^{-1} S_{eq}^T S_{kin} R_{kin}(S_{eq}\xi + U\eta) \end{aligned}$$

and after assuming equilibrium, that is, replacing the ξ-PDEs with the equilibrium conditions,

$$\begin{aligned} \partial_t \eta + L\eta &= (U^T U)^{-1} U^T \, S_{kin} R_{kin}(S_{eq}\xi + U\eta) \\ R(S_{eq}\xi + U\eta) &= 0 \\ r_{eq} &= \partial_t \xi + L\xi - (S_{eq}^T S_{eq})^{-1} S_{eq}^T \, S_{kin} R_{kin}(S_{eq}\xi + U\eta) \end{aligned}$$

Initial and boundary conditions are again to be required only for η. In contrast to the *mere* equilibrium problem, in the mixed equilibrium-kinetic problem the equations of the $I - J_{eq}$-many η- and the J_{eq}-many ξ-variables are no longer decoupled. Only the r_{eq}-equations are still decoupled and can be dropped or computed a posteriori. Some techniques for ensuring that at least *some* of the η-equations are still decoupled even in the mixed problem can be found in [Kr07], among others. The existence proof given above for the *pure* equilibrium problem (i.e., regarding the problem as a constrained optimization problem) obviously cannot be directly applied to the mixed problem, since the constraint (=the η-PDEs) are now nonlinear due to the occurrence of R_{kin} in general, so the admissible set of the corresponding optimization problem is probably no longer convex. The above ξ-η-r_{eq}-problem can be transformed back into an easier to read form (but numerically more complicated to solve, since now r_{eq} is no longer decoupled):

$$\begin{aligned} \partial_t u + Lu &= S_{GG} r_{GG} + S_{kin} R_{kin}(u) \\ R_{GG}(u) &= 0 \end{aligned} \tag{7.7}$$

In this formulation, one has $I + J_{eq}$ equations and unknowns.

8 Reactions with immobile species (mineral precipitation and dissolution), complementarity problems

So far we have considered only the reactions of species dissolved in the fluid among themselves; these processes are not bound to a porous medium, but they can occur in fluids in general (chemical reactors, combustion of gases,...) In the following we consider also reactions between *mobile* (i.e. dissolved in the fluid) and *immobile* species present in the soil matrix or adhering to the soil matrix. The mobile species are described by PDEs, the immobile by ODEs, and all these equations are generally coupled. Reactions that occur between mobile and immobile species are called *heterogeneous*.
There are, in principle, two classes of heterogeneous reactions as far as the structure of the resulting equations is concerned:

- sorption reactions.

- mineral reactions

8.1 Sorption reactions

A micro-scale model of sorption: The surface of the solid skeleton consists of one or more minerals (e.g., FeOH). The mineral particles located on the surface of the soil matrix can react with ions of the fluid (e.g., with H^+, Ca^{2+}, SO_4^{2-} to form $FeOH_2^+, FeO^-, FeOCa^+, FeSO_4^-$). The FeOH surface particle is called *free sorption site* (=*uncomplexed surface site*), the reaction product (also immobile) as *occupied sorption site* (=*complexed surface site*). The reactions can often be described (approximately) by the LMA, both kinetically and in local equilibrium.
A more detailed model: Double Layer Model (see [Be96] Chap. 8)Ȧssumption: The surface can get a net charge; surface charge density

$$\sigma = \frac{F n_w}{A} \sum_i z_i m_i \, [C/m^2],$$

where $F = 96.48$ Coulomb/mole is Faraday's constant, n_w is the amount of water in kg per volume, $z_i \in \mathbb{Z}$ is the charge of the sorption sites, m_i is the molarity of the sorption sites (in moles per kg of water) and A is the size of the surface area per volume. Assumption: A layer with ions (e.g. H^+, Ca^{2+}, SO_4^{2-}) is formed in the fluid, which compensates this surface charge ($\rightarrow$ double layer). The advective velocity near the edge (Poiseuille flow profile) is very low, i.e., transport in the boundary layer is completely dominated by diffusion (thermal motion) and electrostatic forces. The description of these electrostatic forces as well as some approximations lead to the fact that the 'effective' rate at which the reaction proceeds (which is determined by the rate at which ions advance to the surface) can be described by a reaction rate parameter

$$k = k_0 \exp\Big(\frac{\Psi F}{RT}\Big),$$

where T is the temperature in Kelvin, R is the universal gas constant in joules/(Kelvin times mole)=volts times coulombs per Kelvin times mole, and Ψ is the electrostatic potential of the surface. The argument of the exponential function follows from the relationship between surface charge density and the potential:

$$\sigma = \sqrt{8 \cdot 10^3 RT\epsilon\epsilon_0 I} \, \sinh\Big(\frac{\Psi F}{2RT}\Big)$$

and the above formula which relates the surface charge density to the molarities (i.e., concentrations) z_i. Here I is the ionic strength in the fluid and $\epsilon_0 = 8.85 \cdot 10^{-12}$ and $\epsilon = 78.5$ at 25 degrees Celsius.

General (macroscopic) multicomponent model. In a general multicomponent problem, we have two phenomena that complicate both analysis and numerics:

- both mobile and immobile species.

- both kinetic and equilibrium reactions

A model that includes both of these difficulties has the structure

$$\begin{aligned} \big(\, \partial_t \vec{c} + L\vec{c} \, \partial_t \vec{s} \, \big) &= S_{eq} \, \vec{r}_{eq} + S_{kin} \, \vec{R}_{kin}(\vec{c}, \vec{s}) \\ R_{eq}(\vec{c}, \vec{s}) &= 0 \end{aligned}$$

Where $\vec{c}$ and $\vec{s}$ are the vectors of mobile and immobile species concentrations, respectively, and $\vec{R} = (\vec{R}_{eq}^T, \vec{R}_{kin}^T)^T$ is a vector of given rate functions, e.g., according to the

MWG. If we divide the stoichiometric matrix further according to

$$S = (S_{eq}|S_{kin}) = \left(S^{mob} \right) = \left(\begin{array}{c|c} S_{eq}^{mob} & S_{kin}^{mob} \\ \hline S_{eq}^{immo} & S_{kin}^{immo} \end{array} \right) ,$$

so this can also be written as

$$\begin{aligned} \partial_t \vec{c} + L\vec{c} &= S_{eq}^{mob}\, \vec{r}_{eq} + S_{kin}^{mob}\, \vec{R}_{kin}(\vec{c},\vec{s}) \\ \partial_t \vec{s} &= S_{eq}^{immo}\, \vec{r}_{eq} + S_{kin}^{immo}\, \vec{R}_{kin}(\vec{c},\vec{s}) \\ R_{eq}(\vec{c},\vec{s}) &= 0. \end{aligned}$$

The equilibrium condition (see Sec. 5.4) in the case of the law of mass action reads

$$(S_{GG}^{mob})^T \ln \vec{c} + (S_{GG}^{immo})^T \ln \vec{s} = \ln \vec{K} .$$

Three-species sorption model. Often, instead of the general multi-species model, a three-species model is used with a sorption reaction:

$$\alpha\, \mathrm{C} + \beta\, \mathrm{S} \rightleftarrows \mathrm{CS}$$

Where C is a mobile substance (concentration: c), S denotes the free sorption sites ('concentration': $\tilde{s}$), and CS denotes the complexed sorption sites ('concentration': s). In most cases, $\beta = 1$. We can decide whether to describe the reaction as kinetic or as a equilibrium reaction:

Kinetic model: Assuming the LMA, the reaction rate is

$$R(c,s,\tilde{s}) = k_v\, c^\alpha\, \tilde{s}^\beta - k_r\, s\,,$$

and the mass conservation is described by

$$\begin{aligned} \partial_t c + Lc &= -\alpha R(c,s,\tilde{s}) \\ \partial_t \tilde{s} &= -\beta R(c,s,\tilde{s}) \\ \partial_t s &= \quad R(c,s,\tilde{s}) \end{aligned}$$

Simplification can be achieved by making further assumptions, e.g. that $\tilde{s}$ is approximately constant[42] or has little effect on the rate. Then

$$R(c,s,\tilde{s}) \approx \tilde{k}_v\, c^\alpha \; - k_r\, s = k_r \left(\frac{k_v}{k_r}\, c^\alpha - s \right) =: K\,(\varphi(c) - s).$$

A function φ in this model is called *isotherm*[43]; the function $\varphi(c) = \frac{k_v}{k_r} c^\alpha$ hei ßt *Friendly isoterms*. The *isoterm model of sorption (kinetic)* is

$$\begin{aligned} \partial_t (c - \alpha s) + Lc &= 0 \\ \partial_t s &= K\,(\varphi(c) - s) \end{aligned}$$

[42] This assumption is justified if $s \ll s + \tilde{s}$; note also: $\beta s + \tilde{s} = \text{const}$

[43] although it has nothing to do with temperature

Equilibrium model: Assuming again that $\tilde{s}$ is approximately constant or that the equilibrium point hardly depends on $\tilde{s}$, we get from

$$\begin{aligned} \partial_t c + Lc &= -\alpha r \\ \partial_t s &= r \\ R(c,s) &= 0 \end{aligned}$$

that

$$\begin{aligned} \partial_t(c+\alpha s) + Lc &= 0 \\ R(c,s) &= 0\,. \end{aligned}$$

Assuming further that the equation $R(c,s)=0$ has a resolution function $s=\psi(c)$, the *isoterm model of sorption (at equilibrium)* is:

$$\partial_t(c+\alpha\psi(c)) + Lc = 0.$$

The function ψ is called *equilibrium isoterm*. There are different models for ψ. The equilibrium isoterm fitting the Freundlich kinetic model is obviously $\psi(c)=\text{const}\cdot c^\alpha$.

Without the assumption $\tilde{s}=\text{const}$ one gets, for $\alpha=\beta=1$, obviously $\tilde{s}+s=\text{const}=:K_S$ (which is also graphically quite clear); the equilibrium condition according to the LMA is $s=K_{eq}c\tilde{s}$. Combination yields $s=K_{eq}c(K_S-s)$, which translates to

$$s = K_S\,\frac{K_{eq}\,c}{1+K_{eq}\,c} =: \psi(c)\,.$$

Above ψ is called *Langmuir isoterm*. Allowing general α, we obtain the *generalized Langmuir isoterm*

$$\psi(c) = K_S\,\frac{K_{eq}\,c^\alpha}{1+K_{eq}\,c^\alpha}\,.$$

8.2 Reactions with minerals

In a three-species model, a mineral reaction ("precipitation-dissolution reaction") consisting of two partial reactions is

$$\begin{aligned} &\alpha\,\mathrm{A}+\beta\,\mathrm{B} \longrightarrow \mathrm{C} \quad \text{(precipitation (German: Ausfällung))}, \\ &\alpha\,\mathrm{A}+\beta\,\mathrm{B} \longleftarrow \mathrm{C} \quad \text{(dissolution (German: Auflösung, in Lösung gehen))}. \end{aligned}$$

Here C is an immobile component of the soil matrix, and A, B are mobile substances (i.e., ions). There are also more complicated mineral reactions:

$$\sum_i \alpha_i\,\mathrm{A_i} \;\rightleftarrows\; \mathrm{C} + \sum_i \beta_i\,\mathrm{B_i}\,,$$

where C is again the mineral and $\mathrm{A_i}, \mathrm{B_i}$ are mobile species. Of course, there are often *multiple* such reactions in a system. However, usually assumed that there is only *one*

reaction in every mineral reaction; thus, the block of mineral reactions, when all mobile species $\mathrm{X_i}$ are placed on one side (i.e., allowing signed $s_{ij} \in \mathbb{R}$) is

$$\sum_i s_{ij}\, \mathrm{X_i} \rightleftarrows \mathrm{C_j}\,, \quad j = ...$$

which in the stoichiometric matrix makes a block $\begin{pmatrix} S_{min} \\ -\mathrm{Id} \end{pmatrix}$.

The main feature of reactions involving minerals is that the activity of minerals is assumed to be constant (w.l.o.g. $=1$), i.e., the rate of the dissolution reaction is independent of the mineral concentration[44]. The LMA for the three-species model formulated in activities

$$R(c_1, c_2, c_3) = k_v \underbrace{a_1(\vec{c})}_{\approx c_1}{}^{\alpha} \underbrace{a_2(\vec{c})}_{\approx c_2}{}^{\beta} - k_r \underbrace{a_3(\vec{c})}_{\approx 1}$$

thus becomes

$$R(c_1, c_2) = \underbrace{k_v\, c_1^{\alpha} c_2^{\beta}}_{=:R_{prec}(c_1,c_2)} - \underbrace{k_r}_{=:R_{diss}}\,.$$

The corresponding equilibrium condition is

$$c_1^{\alpha} c_2^{\beta} = \frac{k_r}{k_v} =: K_{eq}\,.$$

The constant K_{eq} is called *solubility product* (German: Löslichkeitsprodukt) of the mineral C (the term might be familiar from school chemistry lessons). The independence of the rate from the mineral concentration has a weighty implication: For mineral concentration $c_3 \longrightarrow 0$, in general the negatively signed source term of the c_3 differential equation does not go to zero, i.e., the source term for c_3 can be negative even if $c_3 = 0$; the non-negativity of c_3 would thus not be assured (see Fig. 8). Thus, the model must be modified. For $c_3 = 0$, the condition $0 \le R_{diss} \le R_{prec}$ must hold (instead of: $R_{prec} = K_{eq}$). A possible description is therefore (see publications by Knabner & vanDuijn):

$$\begin{aligned} \partial_t c_1 + L c_1 &= -\alpha\, r \\ \partial_t c_2 + L c_2 &= -\beta\, r \\ \partial_t c_3 &= r \\ \text{where } \; r &\in k_v c_1^{\alpha} c_2^{\beta} - k_r\, H(c_3), \end{aligned}$$

und where H is the set-valued Heaviside-'function'

$$H(x) = \begin{cases} \{1\}, & x > 0 \\ [0,1], & x = 0 \\ \{0\}, & x < 0\,. \end{cases}$$

[44] As a justification, it can be argued that the surface size of the mineral is assumed to change little as the amount of the mineral changes. There are also models in which one tries to model the size of the mineral surface (as a function of the amount of mineral) as well and have the rate depend on the surface size

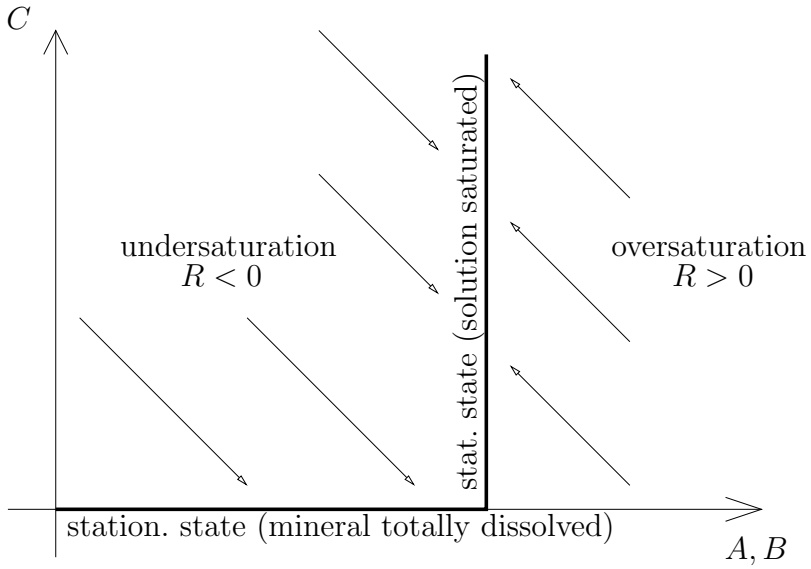


Abbildung 8: The arrows indicate possible changes of the concentration vector (spatial transport & other reactions ignored). The bold L-shaped line indicates the steady states. On the vertical bold line the solution is saturated, on the horizontal the mineral is completely dissolved.

The solution generally has quite low regularity (see exercise); $t \mapsto c_3(t,x)$ can be discontinuous (jump)![45] In the thesis of J. Hoffmann (Univ. Erlangen, 2009) the model

$$\begin{aligned}
&\partial_t c_1 + Lc_1 = -\alpha\, r \\
&\partial_t c_2 + Lc_2 = -\beta\, r \\
&(c_3 = 0 \ \wedge\ \partial_t c_3 - k_v c_1^\alpha c_2^\beta + k_r \geq 0) \quad \vee \quad (c_3 \geq 0 \ \wedge\ \partial_t c_3 - k_v c_1^\alpha c_2^\beta + k_r = 0)
\end{aligned}$$

is used and equivalence to the above representation is proven.
A condition of the above form, i.e.,

$$(f_i(\vec{x}) \geq 0 \ \wedge x_i = 0) \quad \vee \quad (f_(\vec{x}) = 0 \ \wedge x_i \geq 0)$$

is called *complementarity condition* (CC). Such a condition can always be equivalently rewritten as

$$f_i(\vec{x}) \cdot x_i = 0 \ \wedge\ f_i(\vec{x}) \geq 0 \ \wedge\ x_i \geq 0.$$

A problem containing (a) complementarity conditions(s) is called *complementarity problem* (CP). A CP is called linear if the f_i(s) is/are linear. CPs can be rewritten into so-called *variational inequalities* (VI); there is a whole theory on CPs/VIs (see the book *Kinderlehrer, Variational Inequalities*) as well as on numerical solution procedures for these problems. These solution methods were developed in the field

[45] In Knabner & vanDuijn at jump points the derivative $\partial_t c_3$ is understood in the sense of $\lim_{\epsilon \to 0, \epsilon > 0} \partial_t c(t+\epsilon)$.

of *optimization* (because optimization problems with inequality constraints can be written as CPs via their 'KKT' condition).
Regarding the associated equilibrium problem: There are (see also sketch) two cases of equilibria:[46] 1. $R_{prec} = R_{diss}$ (and $c_3 \geq 0$), corresponds to: solution is saturated
2. $c_3 = 0$ and $R_{prec} \leq R_{diss}$, corresponds to: mineral is totally dissolved

This can be summarized as the CC

$$(c_3 = 0 \ \wedge \ c_1^\alpha c_2^\beta \leq K_{eq}) \ \ \vee \ \ (c_3 \geq 0 \ \wedge \ c_1^\alpha c_2^\beta = K_{eq})$$

or equivalently

$$(K_{eq} - c_1^\alpha c_2^\beta) \cdot c_3 = 0 \ \wedge \ c_3 \geq 0 \ \wedge \ K_{eq} - c_1^\alpha c_2^\beta \geq 0 \,,$$

which has to be solved together with the equilibrium conditions

$$\begin{array}{lcl} \partial_t c_1 + L c_1 & = & -\alpha\, r \\ \partial_t c_2 + L c_2 & = & -\beta\, r \\ \partial_t c_3 & = & r \end{array}$$

or, for short,

$$\begin{array}{lcl} \partial_t (c_1 + \alpha c_3) + L c_1 & = & 0 \\ \partial_t (c_2 + \beta c_3) + L c_2 & = & 0 \,. \end{array}$$

Alternatively, the PDEs can be rewritten with the introduction of a reaction invariant $\eta := \beta c_1 - \alpha c_2$ to be

$$\begin{array}{lcl} \partial_t \eta + L \eta & = & 0 \\ \partial_t (\underbrace{c_1}_{= \frac{1}{\beta}(\eta + \alpha c_2)} + \alpha c_3) + L \underbrace{c_1}_{= \frac{1}{\beta}(\eta + \alpha c_2)} & = & 0 \end{array}$$

which together with the CC represents three equations for the three unknowns η, c_2, c_3. The advantage of this last representation is that the equation for η is decoupled. For the decoupling it is exploited that η was formed as linear combination of only mobile species.
Numerical Solving. Numerical solving of nonlinear PDEs generally requires a discretization in space and time and the application of Newton's method (or a similar iterative method) for the reduction to linear systems of equations. But how does one treat complementarity constraints, i.e., *inequalities*? There are various methods developed in the field of optimization for solving CPs numerically. A rather simple one is the following: One chooses a function $\varphi : \mathbb{R}^2 \longrightarrow \mathbb{R}$ with the property

$$\varphi(a, b) = 0 \iff ab = 0 \ \wedge \ a \geq 0 \ \wedge \ b \geq 0 \,.$$

[46] You can observe this yourself if you pour salt into a glass of water: If you take little salt (so little that the product of sodium and chlorine ions is less than the solubility product), then the salt dissolves completely. Otherwise, only as much dissolves as results from the solubility product (i.e., the amount of dissolved ions does not increase further if the amount of salt is increased); the solution is then saturated, and the rest remains undissolved as a solid on the bottom of the glass.

Examples for functions with this property are $\varphi(a,b) = \min\{a,b\}$ and $\varphi(a,b) = a+b-\sqrt{a^2+b^2}$. One can then rewrite the CC (which contains *inequalities*) equivalently into an *equation(!)*

$$\varphi(f_i(\vec{x}), x_i) = 0\,;$$

note that the inequalities have 'disappeared'. The price to pay for this is: The above functions φ are not very smooth, they do not fulfill the regularity requirements which are classically needed for Newton's method (i.e., function differentiable, derivative Lipschitz continuous). However: In the 1990s it turned out that one can extend the theory of Newton's method to functions of quite low regularity, to so-called *strongly semismooth functions* (the definition of this class of functions is quite technical); the Newton method for such functions is, like the classical Newton method, locally quadratically convergent; it is called *semismooth Newton method*. Regarding the implementation of the semismooth Newton method, it is exactly the same as the classical Newton's method, up to the fact that if you need the Jacobian at a point of no differentialbility, just take the right-hand or the left-hand derivative – as you like.
Other methods for nonlinear CPs include so-called *interior point methods* and *active set strategies* (see books on Numerical Methods for Restricted Optimization Problems).

General multispecies model with mineral reactions.
The scope of the model:

- several minerals / mineral reactions, but all in equilibrium,
- besides further equilibrium reactions (divided into sorption reactions and reactions within the mobile phase (aqueous reactions)),
- further kinetic reactions (both sorption reactions and reactions within the mobile phase, but no mineral reactions).

The stochiometric matrix has the block structure:

$$S = \left(\begin{array}{c|c|c||c} S^1_{mob} & S^1_{sorp} & S^1_{min} & S^1_{kin} \\ \hline\hline 0 & S^2_{sorp} & 0 & S^2_{kin} \\ \hline 0 & 0 & -\mathrm{Id} & 0 \end{array}\right)$$

The first three columns contain equilibrium reactions, in the fourth column kinetic reactions. The first row belongs to mobile species, the other two rows to immobile species, the first the sorption sites, then the minerals. The vector of (unknown) equilibrium reation rates is accordingly divided into $r_{eq} = (r^T_{mob}, r^T_{sorp}, r^T_{min})^T$. The system for the unknown $c_{mob}, c_{sorp}, c_{min}, r_{mob}, r_{sorp}, r_{min}$ is (readable from the block structure of S)

$$\begin{array}{lcll} \partial_t c_{mob} + L c_{mob} & = & S^1_{mob}\, r_{mob} + S^1_{sorp}\, r_{sorp} + S^1_{min}\, r_{min} & + S^1_{kin} R_{kin}(c_{mob}, c_{sorp}) \\ \partial_t c_{sorp} & = & S^2_{sorp}\, r_{sorp} & + S^2_{kin} R_{kin}(c_{mob}, c_{sorp}) \\ \partial_t c_{min} & = & -r_{min} & \end{array}$$

with the equilibrium conditions, which in the case of mass action kinetics ("$S_{eq}^T \ln \vec{c} = \ln \vec{K}$") read

$$\begin{array}{ll} (S^1_{mob})^T \ln c_{mob} - \ln K_{mob} = 0 & \text{(equil. of the aqueous reactions)} \\ (S^1_{sorp})^T \ln c_{mob} + (S^2_{sorp})^T \ln c_{sorp} - \ln K_{sorp} = 0 & \text{(equil. of sorption reactions)} \\ \varphi(\ln K_{min} - S^1_{min} \ln c_{mob}, c_{min}) = 0 & \text{(equil. of mineral reactions)} \end{array}$$

where the complementarity function φ is to be applied component-wise.
In the 1980s to the 2010s, many different transformations of this system, or mostly of models of somewhat smaller scope, have been published. A method to transform this system into a system of reduced size (decoupling of equations, elimination of r_{eq},...), which can then be solved numerically faster, can be found in [Habil, Kr21], variants of it in the dissertation of J. Hoffmann, Erlangen, 2009.[47]

9 Appendix: Fixed-point theorems

In Sec. **??** we needed a fixed point theorem – Schaefer's fixed point theorem – to show the existence of a solution for the PDE model. Here in the appendix, in a chain of (fixed point) theorems, this fixed point theorem is derived. Note that the FP theorems appearing here always only provide the existence and never the uniqueness of a fixed point (unlike Banach's FP theorem).

9.1 Fixed-point theorem in finite dimensions

Theorem (Brouwer)[48] Let B be the closed unit sphere in $\mathbb{R}^n$ with respect to the Euclidean norm and let $Z : B \longrightarrow B$ in $C^\infty(B)$. Then Z has (at least) one fixed point, that is, a $x \in B$ with $Z(x) = x$.

Remark. In the case $n = 1$ the validity of the statement of the theorem is graphically obvious (make a sketch of the function, or use the intermediate value theorem on $Z - \mathrm{id}$). It is an elementary corollary of the intermediate value theorem (Zwischenwertsatz). The proof that follows can be given for all $n \in \mathbb{N}$; however, we will restrict ourselves to the case $n = 2$ in one proof step to make the exposition less technical.

Proof. Suppose there is no fixed point, i.e., $x \neq Z(x)$ for all $x \in B$. For each $x \in B$ we consider the straight line $g_x(\alpha) := x + \alpha \underbrace{(x - Z(x))}_{\neq 0}$ that goes through the points $x, Z(x) \in B$. The line g_x has exactly two intersections with ∂B (for this, realize that a tangent orientation of g_x is not possible). So there exist two different $\alpha_1, \alpha_2 \in \mathbb{R}$ with

[47] For this method there was a prize for special efficiency of the method in connection with the solution of an international large-scale benchmark problem [CKK10, MoMaS10] for reactive transport.
[48] see [GilbTrud] p. 236

$1 = \|x+\alpha(x-Z(x))\|^2$. This condition can be written as a quadratic equation with the solution

$$\alpha_{1,2} = \alpha_{1,2}(x) = \frac{\langle x, Z(x)-x\rangle}{\|x-Z(x)\|^2} \pm \sqrt{\frac{|\langle x, Z(x)\rangle - x|^2 + (1-\|x\|^2)\|x-Z(x)\|^2}{\|x-Z(x)\|^4}}\,.$$

One of the solutions is nonnegative, the other nonpositive (is obvious because fpr $\alpha = 0$ we are at point $x \in B$). Let $\alpha_1(x)$ be the nonnegative solution. Let $r(x) := g_x(\alpha_1) \in \partial B$ be denoted "the boundary point belonging to x". The mapping

$$x \longmapsto r(x), \quad B \longrightarrow \partial B$$

is smooth since the discriminant is strictly positive and Z is smooth. Now consider the mapping

$$F : [0,1]\times B \longrightarrow B, \quad F(t,x) := \underbrace{x + t\,\alpha_1(x)(x-Z(x))}_{\in B}\,.$$

It is $F(0,x) = x$, so $F(0,\cdot) = \mathrm{id} : B \to B$.
It is $F(1,x) = r(x)$, so $F(1,\cdot) = r : B \to \partial B$.
We now compute the volume $V(t) := \mathrm{Vol}(F(t,B))$ of the set $F(t,B)$, $t \in [0,1]$:
Let $J_x F(t,x)$ be the Jacobian matrix of F formed with respect to the argument x. It is

$$\begin{aligned} V(t) &= \int\limits_B \det J_x F(t,x)\, dx\,, \\ V(0) &= \int\limits_B \det J_x(\mathrm{id})\, dx = \int\limits_B 1\, dx = \mathrm{Vol}(B)\,, \\ V(1) &= \mathrm{vol}(\partial B) = 0\,. \end{aligned}$$

In order to obtain a contradiction to these findings, we will prove that $\frac{d}{dt}V(t) \equiv 0$:
We start with

$$\frac{d}{dt}V(t) = \int\limits_B \frac{\partial}{\partial t} \det\left[\frac{\partial}{\partial x_1}F(t,x), ..., \frac{\partial}{\partial x_n}F(t,x)\right] dx\,.$$

To simplify the representation, we assume here $n = 2$; it is
$\frac{\partial}{\partial t} \det\left[\frac{\partial}{\partial x_1}F(t,x_1,x_2), \frac{\partial}{\partial x_2}F(t,x_1,x_2)\right] = \frac{\partial}{\partial x_1} \det\left[\frac{\partial}{\partial t}F(t,x_1,x_2), \frac{\partial}{\partial x_2}F(t,x_1,x_2)\right] - \frac{\partial}{\partial x_2} \det\left[\frac{\partial}{\partial x_1}F(t,x_1,x_2), \frac{\partial}{\partial t}F(t,x_1,x_2)\right]$,
which can be elementarily checked by expansion (=Entwickung) of the determinants. [49] Now

$$\begin{aligned} \int\limits_B \frac{\partial}{\partial x_1} \det\left[\frac{\partial}{\partial t}F, \frac{\partial}{\partial x_2}F\right] dx &= \int\limits_{-1}^{1} \left(\int\limits_{x_1=-\sqrt{1-x_2^2}}^{x_1=+\sqrt{1-x_2^2}} \frac{\partial}{\partial x_1} \det\left[\frac{\partial}{\partial t}F, \frac{\partial}{\partial x_2}F\right] dx_1 \right) dx_2 \\ &= \int\limits_{-1}^{1} \left(\det\left[\frac{\partial}{\partial t}F(t,x_1,x_2), \frac{\partial}{\partial x_2}F(t,x_1,x_2)\right]\Big|_{x_1=-\sqrt{1-x_2^2}}^{x_1=+\sqrt{1-x_2^2}} \right) dx_2\,. \end{aligned}$$

[49] A generalization of this equation to $n \in \mathbb{N}$ can be found in [GilbTrud], Lemma 10.12, p. 235.

The two points (x_1, x_2) occurring in that expression lie on the boundary of B, but there $x = r(x)$ holds, so $\alpha_1(x) = 0$, so $F(\cdot, x) = x$ for these points x, so $\frac{\partial}{\partial t}F(t,x) = 0$ for $x = (\pm\sqrt{1-x_2^2}, x_2)$. Thus it follows that $\int\limits_B \frac{\partial}{\partial x_1} \int \left[\frac{\partial}{\partial t}F, \frac{\partial}{\partial x_2}F\right] dx = 0$. Similarly, one shows that $\int\limits_B \frac{\partial}{\partial x_2} \det\left[\frac{\partial}{\partial x_1}F, \frac{\partial}{\partial t}F\right] dx = 0$. Thus it follows that $\frac{d}{dt}V(t) \equiv 0$, which is a contradiction to $V(1) \neq V(0)$. □

Relaxation of the regularity requirement:

Satz (Brouwer's fixed point theorem) (actually, P. Bohl, 1904, was the author of this version)
Let B be as above, let $Z : B \longrightarrow B$ be continuous. Then Z has a fixed point.

Proof. Let $Z = (Z_1, ..., Z_n)$, that is $Z_i : B \longrightarrow [-1, 1]$. According to the Weierstraß approximation theorem there is sequence of polynomials $Z_i^k : B \longrightarrow \mathbb{R}$ with $|Z_i^k(x) - Z_i(x)| \leq \epsilon \,\forall\, x \in B \,\forall\, k \geq K(\epsilon)$.[50] Because of the equivalence of norms in $\mathbb{R}^n$, it follows that $\|Z^k(x) - Z(x)\| \leq \sqrt{n}\,\epsilon$. We set $\tilde{Z}^k(x) := (1 - \sqrt{n}\,\epsilon)\, Z^k(x)$. Then $\tilde{Z}^k : B \longrightarrow \mathbb{R}^n$ is a C^∞-function, and its image is, for $\epsilon \leq \frac{1}{\sqrt{n}}$, in B, since

$$\begin{aligned} \|\tilde{Z}^k(x)\| &= \underbrace{(1-\sqrt{n}\,\epsilon)}_{\geq 0} \|Z^k(x)\| \leq (1-\sqrt{n}\,\epsilon)\, \big(\underbrace{\|Z(x)\|}_{\leq 1} + \underbrace{\|Z^k(x) - Z(x)\|}_{\leq \sqrt{n}\epsilon} \big) \\ &\leq (1-\sqrt{n}\,\epsilon)\,(1+\sqrt{n}\,\epsilon) = 1 - n\epsilon^2 \leq 1 \end{aligned}$$

Hence, the C^∞-version of the FP theorem can be applied to $\tilde{Z}^k$. Let x_k be the resulting fixed point of $\tilde{Z}^k$. By the Bolzano–Weierstraß theorem, the sequence (x_k) has a subsequence, denoted again by (x_k), which converges to an $x \in B$ (since B is closed and bounded). It follows

$$\|Z(x) - x\| \leq \underbrace{\|Z(x) - Z(x_k)\|}_{=:\text{(I)}} + \underbrace{\|Z(x_k) - \tilde{Z}^k(x_k)\|}_{=:\text{(II)}} + \underbrace{\|\tilde{Z}^k(x_k) - x_k\|}_{=0} + \underbrace{\|x_k - x\|}_{\to 0}$$

Term (I) converges to zero for $k \to \infty$ since Z is continuous and $x_k \to x$. Term (II) converges to zero since $\tilde{Z}_k$ converges uniformly to Z. It follows $Z(x) = x$. □

Theorem (further generalization). Instead of a ball B in the above sentences, one can take any set $\tilde{B} \subset \mathbb{R}^n$ which is homeomorphic to B (i.e., there exists a bijective continuous mapping from B to $\tilde{B}$ whose inverse mapping is also continuous). [51].

[50] The direct application of the above C^∞-version of the FP-theorem to these C^∞-functions Z^k fails because in general these polynomials do not map into B.

[51] In an illustrative and somewhat simplifying diction, this means that one can take closed sets which have no holes, and whose edges can even have corners/bends with angles strictly between 0 and 2π

Proof. Let $H : B \longrightarrow \tilde{B}$ be a homeomorphism and let $Z : \tilde{B} \longrightarrow \tilde{B}$ be continuous. Then $\tilde{Z} := H^{-1} \circ Z \circ H : B \longrightarrow B$ is continuous, so by the above theorem it has a fixed point $x \in B$. It follows that $Z \circ H(x) = H(x)$, i.e., Z has the fixed point $H(x) \in \tilde{B}$. □

9.2 Fixed point theorems in Banach spaces

It is now our goal to extend the above existence result for fixed points to arbitrary (in general infinite-dimensional) Banach spaces.

Theorem (Schauder's fixed point theorem), (Schauder 1930, see [Evans] p. 502). Let X be a real Banach space. Let $K \subset X$ be compact and convex. Let $Z : K \longrightarrow K$ be continuous. Then Z has a fixed point.

Proof. Let $\epsilon > 0$. The set of all open balls $B(x, \epsilon)$, where $x \in K$, trivially forms an open covering of K. Since K is compact, there must be a finite covering $B(x_i, \epsilon)$, $i = 1, ..., m = m(\epsilon)$ of K:

$$K \subseteq \bigcup_{i=1}^{m} B(x_i, \epsilon)$$

Let K_ϵ be defined as the convex hull of the set $\{x_1, ..., x_m\}$:

$$K_\epsilon := \left\{ x = \sum_{i=1}^{m} \lambda_i x_i \,|\, \lambda_i \geq 0\,,\ \sum_{i=1}^{m} \lambda_i = 1 \right\}$$

We have $K_\epsilon \subseteq K$, since $x_i \in K$ and since K is convex.
$K_\epsilon \subset X$ is homeomorphic[52] to a polyhedron (:=intersection of half-spaces) in $\mathbb{R}^n$. Thus, by the last theorem of Chap. 9.1, every continuous mapping from K_ϵ to K_ϵ has a fixed point.
First we define $f_\epsilon : K \longrightarrow K_\epsilon$ by

$$f_\epsilon(x) := \frac{\sum_{i=1}^{m} \mathrm{dist}(x, K \setminus B(x_i, \epsilon))\, x_i}{\sum_{i=1}^{m} \mathrm{dist}(x, K \setminus B(x_i, \epsilon))} \in K_\epsilon \quad \forall\, x \in K\,.$$

The mapping is well-defined, since in the denominator for arbitrary x not all summands can vanish at the same time. The mapping is continuous, since 'dist' is continuous. It holds

$$\|f_\epsilon(x) - x\| \leq \frac{\sum_{i=1}^{m} \mathrm{dist}(x, K \setminus B(x_i, \epsilon))\, \|x_i - x\|}{\sum_{i=1}^{m} \mathrm{dist}(x, K \setminus B(x_i, \epsilon))} \leq \epsilon\,;$$

[52] If one assumes that the x_i are linearly independent in X (which one can do without loss of generality), then one can take as homeomorphism the mapping $H : \mathbb{R}^m \to X$ which maps the vector $\lambda = (\lambda_1, ..., \lambda_m)$ to the associated x (see definition of K_ϵ).

where the last inequality holds because for each x and i either $x \in B(x_i, \epsilon)$, thus $\|x_i - x\| \leq \epsilon$, or $\operatorname{dist}(x, K \setminus B(x_i, \epsilon)) = 0$.
Now let us define

$$Z_\epsilon : K_\epsilon \longrightarrow K_\epsilon, \quad Z_\epsilon(x) := \underbrace{f_\epsilon(\overbrace{Z(\underbrace{x}_{\in K_\epsilon})}^{\in K})}_{\in K_\epsilon}.$$

Being a composition of continuous functions, Z_ϵ is continuous. According to the above considerations about K_ϵ, Z_ϵ has a fixed point $x_\epsilon \in K_\epsilon \subseteq K$. Let (ϵ_j) be a positive sequence that converges to zero. Since K is compact, there exists a subsequence, again denoted by (ϵ_j), such that (x_{ϵ_j}) is convergent to an $x \in K$. We show that this x is a fixed point of Z:

$$\|x - Z(x)\| \leq \underbrace{\|x - x_{\epsilon_j}\|}_{\to 0} + \underbrace{\|x_{\epsilon_j} - Z_{\epsilon_j}(x_{\epsilon_j})\|}_{=0} + \overbrace{\|\underbrace{Z_{\epsilon_j}(x_{\epsilon_j})}_{=f_{\epsilon_j}(Z(x_{\epsilon_j}))} - Z(x_{\epsilon_j})\|}^{\leq \epsilon_j} + \underbrace{\|Z(x_{\epsilon_j}) - Z(x)\|}_{\to 0 \text{ since } Z \text{ continuous}} \overset{(j\to\infty)}{\longrightarrow} 0$$

Hence, $x = Z(x)$. □

Definition (compact mapping). Let X, Y be Banach spaces. A (possibly nonlinear) mapping $Z : X \longrightarrow Y$ is called compact if Z is continuous and for every bounded set $M \subset X$ it holds that $\overline{Z(M)}$ is compact.

Theorem. A mapping $Z : X \longrightarrow Y$ is compact if and only if every bounded sequence (x_n) in X has a subsequence (x_{n_k}) such that $(Z(x_{n_k}))$ is convergent in Y. (In short, under Z bounded sequences become convergent sequences.)

From Schauder's fixed point theorem one can deduce Schaefer's fixed point theorem:[53]

Theorem (Schaefer's fixed point theorem [Schae55]). Let $Z : X \longrightarrow X$ be compact and the set

$$M := \{x \in X \mid \exists\, \lambda \in [0, 1] : x = \lambda Z(x)\}$$

be bounded. Then Z has a fixed point.

Motivation/application: Schaefer's theorem, unlike Schauder's, does not require identifying a suitable compact convex set. Instead, the compactness of an operator has to be shown, which can often be done easily in the context of PDE problems by recourse to known results about the compact embedding of function spaces into other

[53] Schaefer's fixed point theorem can be regarded as a special case of the fixed point theorem of Leray–Schauder. The latter is considerably better known than the former, although the former is generally more convenient to use.

function spaces.

Proof. Let c be a bound of the set M, but not the smallest bound. Set

$$\tilde{Z} : X \longrightarrow X, \quad \tilde{Z}(x) := \begin{cases} Z(x), & \text{if } \|Z(x)\| \leq c \\ \frac{c\,Z(x)}{\|Z(x)\|}, & \text{if } \|Z(x)\| > c \end{cases}$$

("truncation of Z"). It follows $\|\tilde{Z}(x)\| \leq c \,\forall\, x \in X$, so in particular

$$\tilde{Z}(\overline{B(0,c)}) \subseteq \overline{B(0,c)} \quad \text{and therefore} \quad \tilde{Z}(\tilde{Z}(\overline{B(0,c)})) \subseteq \tilde{Z}(\overline{B(0,c)}). \qquad (*)$$

We now consider the restriction of $\tilde{Z}$ to the domain of definition $\tilde{Z}(\overline{B(0,c)})$: We have, c.f. (*),

$$\tilde{Z} : \tilde{Z}(\overline{B(0,c)}) \longrightarrow \tilde{Z}(\overline{B(0,c)})\,.$$

Now let K be the closure of the convex hull of $\tilde{Z}(\overline{B(0,c)})$. Then we have

$$\tilde{Z}(\overline{B(0,c)}) \subseteq K \subseteq \overline{B(0,c)},$$

where the first inclusion is trivial, and the second holds since $\overline{B(0,c)}$ is already a convex closed set containing (after (*)) $\tilde{Z}(\overline{B(0,c)})$, and K is *the smallest* convex closed set containing $\tilde{Z}(\overline{B(0,c)})$. Thus, considering $\tilde{Z}$ restricted to K we get,

$$\tilde{Z} : K \longrightarrow \tilde{Z}(K) \subseteq \tilde{Z}(\overline{B(0,c)}) \subseteq K\,.$$

Since Z is a compact mapping, $\tilde{Z}$ is also a compact mapping. Furthermore, since $\overline{B(0,c)}$ is a bounded set, $\tilde{Z}(\overline{B(0,c)})$ (by definition of compact mapping) is a compact set. From this, using the definition of K as a closed convex hull, we can conclude that K is compact. We can apply Schauder's fixed point theorem to this $\tilde{Z} : K \to K$, because K is convex and compact, and $\tilde{Z} : K \to K$ is continuous. It follows that there exists a fixed point $x \in K$: $\tilde{Z}(x) = x$.

It is now also $Z(x) = x$. To see this, suppose that this is false. Then by definition of $\tilde{Z}$ it must hold that $\|Z(x)\| > c$ (for otherwise $Z(x) = \tilde{Z}(x) = x$). So it is $x = \tilde{Z}(x) = \frac{c\,Z(x)}{\|Z(x)\|}$. From this follows on the one hand $\|x\| = c$, but on the other hand $x = \lambda Z(x)$ with $\lambda := \frac{c}{\|Z(x)\|} \in [0,1]$, hence $x \in M$, so $\|x\| < c$ by definition of the bound c. Contradiction. □

Literatur

[Ada03] R. Adams, J. Fournier, *Sobolev spaces*, Elsevier Science, Oxford, (2nd ed.), 2003.

[Be96] C. Bethke, *Geochemical reaction modeling*, Oxford University Press, 1996.

[MoMaS10] J. Carrayrou, J. Hoffmann, P. Knabner, S. Kräutle, C. de Dieuleveult, J. Erhel, J. Van der Lee, V. Lagneau, K.U. Mayer, K.T.B. MacQuarrie, *Comparison of numerical methods for simulating strongly nonlinear and heterogeneous reactive transport problems-the MoMaS benchmark case*, Comp. Geosci., 14 (2010), 483–502.

[CKK10] J. Carrayrou, M. Kern, P. Knabner, *Reactive transport benchmark of MoMaS*, Comput. Geosci., 14 (2010), 385–392.

[Evans] L.C. Evans, *Partial differential equations*, American Mathematical Society, Providence, 1998.

[GilbTrud] D. Gilbarg, N. Trudinger, *Elliptic partial differential equations of second order*, Springer, 1977.

[Ha64] P. Hartman, *Ordinary differential equations*, John Wiley & Sons, 1964.

[Kr21] S. Kräutle, J. Hodai, P. Knabner, *Robust simulation of mineral precipitation-dissolution problems with variable mineral surface area*, Journal of Engineering Mathematics 129, doi:10.1007/s10665-021-10132-4, 2021.

[Kr07] S. Kräutle, P. Knabner, *A reduction scheme for coupled multicomponent transport-reaction problems in porous media: Generalization to problems with heterogeneous equilibrium reactions*, Water Resour. Res., 43, W03429, doi:10.1029/2005WR004465, 2007.

[Habil] S. Kräutle, *General multi-species reactive transport problems in porous media: Efficient numerical approaches and existence of global solutions*, Habilitation thesis, University of Erlangen-Nuremberg, Germany, 2008.

[Lady68] O.A. Ladyženskaja, V.A. Solonnikov, N.N. Uralceva, *Linear and quasi-linear equations of parabolic type*, American Mathematical Society, 1968.

[Logan] J. Logan, *Transport modeling in hydrogeochemical systems*, Springer, 2001.

[MiSi04] M. Mincheva, D. Siegel, *Stability of mass action reaction-diffusion systems*, Nonlinear Analysis, 56 (2004), 1105–1131.

[Schae55] H. Schaefer, *Über die Methode der a priori-Schranken*, Math. Annalen, 129 (1955), 415–416.

[WYW] Z. Wu, J. Yin, C. Wang, *Elliptic and parabolic equations*, World Scientific Publishing, 2006.