Lecture notes for "Introduction to Mathematical Modeling" — Freie Universität Berlin, Winter semester 2017/2018

Carsten Hartmann, Nikki Vercauteren, Ralf Banisch

November 28, 2017

Contents

1 Introduction

Mathematical tools & concepts: basic ODE
Suggested references: [Ari94, Ben00]

By "mathematical model" we mean anything that can be expressed in terms of mathematical formulae and that is amenable to mathematical analysis. Typical mathematical models involve (list non-exhaustive, non-disjoint):

- deterministic ODE models (e.g. population growth, chemical reactions mechanical systems and analogues, climate, . . .)
- stochastic models (e.g. growth of small populations, chemical reactions in cells, asset prices, weather prediction, . . .)
- optimality principles (e.g. principles of utility, properties of materials, minimization of energy consumption, trading strategies, . . .)
- discrete or continuous flow models (e.g. queuing problems, traffic, logistics, load balancing in parallel computers, . . .)
- statistical models (e.g. distribution of votes, change in the precipitation rate, wage justice, criminal statistics, Google PageRank, . . .)

Of course, you can combine any of the aforementioned modelling approaches to obtain what is called a *hybrid model.*

Building a model: *divide et impera.* From the modelling viewpoint, the world is divided into things whose effects are neglected (e.g. planetary configurations in a model of traffic flow on a highway), things whose behaviour the model is designed to study, so-called *observables* or *states* (e.g. the average number of commuters in a traffic flow model), and things that affect the model, but that are not within the scope of the model, called *boundary conditions* (e.g., weather conditions in a model of traffic flows).

The standard way to build a mathematical model then involves the following four steps:

1. **Formulate the modelling problem:** What is the question to be answered? What type of model is appropriate to answer the question?
2. **Outline the model:** Which effects should be included in the model, which are negligible? Write down relations between the states.
3. **Practicability check:** Is the model "solvable", either by analytical methods or numerical simulation? Do I have access to all the parameters in the model? Can the model be used to make predictions?
4. **Reality check:** Make predictions of known phenomena and compare with available data (qualitatively or quantitatively).

Note that there is a trade-off between simplicity of a model (easier to analyze and interpret) and "realism" or accuracy of the model (potentially more complicated, analysis may be through computer simulations only).

year	population in millions
1790	3.93
1810	7.24
1830	12.87
1850	23.19
1870	39.82
1890	62.95
1910	91.97
1930	122.78
1950	150.70
1970	208.00
1990	248.14
2010	308.19

Table 1: U.S. population between 1790 and 2010 (see http://www.census.gov/).

Illustrative example: population growth. Suppose we want to describe the long term growth of a population, specifically we want to predict the growth of the, say, U.S. population over several generations. Here are some numbers from the U.S. Census Bureau:

Now let let $\gamma \in \mathbb{R}$ be the net reproduction rate per individual (birth rate minus death rate, and $N(t)$ the size of the population at time $t \geq 0$. Then, by definition of γ, we have

$$\gamma(t) = \lim_{\Delta t \to 0} \frac{1}{N(t)} \frac{N(t + \Delta t) - N(t)}{\Delta t}$$

(Note that this assumes that the limit exists, which is nonsense given the annual census data.) This suggests the following model for N as a function of t:

$$\dot{N}(t) = \gamma(t) N(t) \,, \quad N(0) = N_0 \tag{1.1}$$

where the dot means differentiation with respect to t. This completes Steps 1 and 2 above. Now suppose that γ is independent of t. The solution of (1.1) then is

$$N(t) = N_0 e^{\gamma t} \,, \tag{1.2}$$

which, depending on the sign of γ, means that the population will either grow (γ positive), die out (γ negative) or stay constant ($\gamma = 0$). Sounds okay! So, let us skip the Step 3 and the question of the how to get γ and directly proceed with Step 4: Assuming $\gamma > 0$ it holds that

$$\lim_{t \to \infty} N(t) = \infty \,, \tag{1.3}$$

which cannot be true. (Make sure you understand why the model must be rejected on the basis of this prediction.) So let's go back to Step 2 and take into account that the growth rate of a population will depend on its size due to limited resources, food supply etc. Specifically, let

$$\tilde{\gamma} \colon [0, \infty) \to \mathbb{R} \,, \; N \mapsto \tilde{\gamma}(N) \,.$$

be strictly decreasing for sufficiently large N, with $\tilde{\gamma}(N) \to -\infty$ for $n \to \infty$, so as to make sure that the reproduction rate becomes negative once a certain population size is exceeded. Getting the precise census data to estimate $\tilde{\gamma}(N)$ will be difficult, but we may be happy with a rough estimate; the simplest possible scenario is

$$\tilde{\gamma}(N) = \gamma(1 - N/K), \quad \gamma, K > 0 \tag{1.4}$$

where γ, K must be determined from data. The resulting differential equation,

$$\dot{N}(t) = \gamma N(t)(1 - N(t)/K)\,, \quad N(0) = N_0\,, \tag{1.5}$$

is called the logistic growth model. It is clear that when N grows, the right hand side of the equation will become negative and so will the reproduction rate. This guarantees that N remains finite for all t. It can be shown that

$$N(t) = \frac{KN_0 e^{\gamma t}}{K + N_0(e^{\gamma t} - 1)}\,, \tag{1.6}$$

which implies that

$$\lim_{t\to\infty} N(t) = K\,.$$

For obvious reasons K is called the systems's *capacity.* This looks much better than before, and we may now see how well this model fits the data given in Table 1. Clearly, the model could be extended in various ways, e.g., by splitting the population into subpopulations according to sex and age, by incoporating additional external factors, such as war, immigration etc.

Problems

Exercise 1.1. *Consider the logistic growth model from above.*

a) Discuss the issue of parameter estimation and model validation: How could the unknown parameters γ, K be computed? How well does the model fit the data? How would you judge the predictive power of the model?

b) Would you trust the model, if N_0 was, say, 2 or 3? If not, explain why and discuss possible ways to improve the model. In case you do trust the model, interpret the role of the parameter γ.

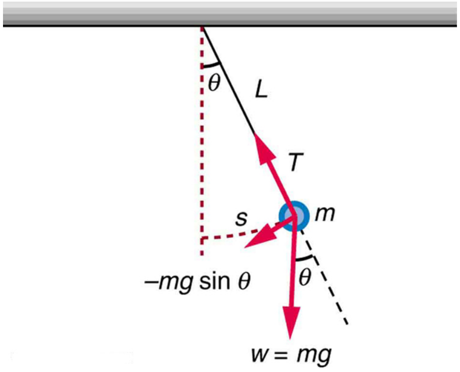


Figure 2.1: The classical pendulum. The radial position at time t is given by the arclength $s(t) = L\theta(t)$, hence the radial force on the mass is $-mL\ddot{\theta}(t)$.

2 Arguments from scale

Mathematical tools & concepts: basic ODE, linear algebra
Suggested references: [Ben00, IBM+05]

Let us start with some motivation and look at the classical pendulum (see Fig. 2.1). The governing equation of motion for the angle θ as a function of t is

$$L\ddot{\theta}(t) = -g\sin\theta(t) \tag{2.1}$$

with g acceleration due to gravity and L the length of the pendulum. When θ is small, $\theta \approx \sin\theta$ and we may replace the last equation by

$$\ddot{\theta}(t) = -\omega^2\theta(t) \tag{2.2}$$

with $\omega = \sqrt{g/L}$. The solution of (2.2) is

$$\theta(t) = A\sin(\omega t) + B\cos(\omega t) \tag{2.3}$$

with A, B depending on the initial conditions $\theta(0)$ and $\dot{\theta}(0)$. Since sine and cosine have a period of 2π, we find that the pendulum has period

$$T = \frac{2\pi}{\omega} = 2\pi\sqrt{\frac{L}{g}}\,, \tag{2.4}$$

which is independent of the mass m of the pendulum and which does not depend on the initial position $\theta(0) = \theta_0$.

Derivation from scale arguments. Let us now derive the essential dependence of T on L and g without using any differential equations. To this end we conjecture that there exists a function f such that

$$T = f(\theta, L, g, m)\,. \tag{2.5}$$

We denote the physical units (a.k.a. *dimensions*) of the variables θ, L, g, m by square brackets. Specifically,

$$[T] = \mathrm{s}\,, \quad [\theta] = \text{dimensionless}\,, \quad [L] = \mathrm{m} \quad [g] = \mathrm{ms}^{-2}\,, \quad [m] = \mathrm{kg}$$

In an equation the physical units must match, so the idea is to combine θ, L, g, m is such a way that the physical units of the formula have the unit of time T. This excludes transcendental functions, such as log or tan for the variables carrying physical units. Using the ansatz

$$f(\theta, L, g, m) \propto L^{\alpha_1} g^{\alpha_2} m^{\alpha_3} \tag{2.6}$$

with unknowns α_1, α_2, $\alpha_3 \in \mathbb{R}$ that must be chosen such that

$$\mathrm{s} = \mathrm{m}^{\alpha_1+\alpha_2}\mathrm{s}^{-2\alpha_2}\mathrm{kg}^{\alpha_3}$$

Note that we have ignored θ as it does not carry any physical units. By comparison of coefficients we then find

$$\alpha_1 = 1/2\,, \quad \alpha_2 = -1/2\,, \quad \alpha_3 = 0\,.$$

which yields

$$T \propto \sqrt{\frac{L}{g}} \tag{2.7}$$

and which is consistent with (2.4). Note, however, that we cannot say anything about a possible dependence of T on the dimensionless angle variable θ. (The unknown dependence on the angle is the constant prefactor 2π.)

2.1 Dimensional analysis

Let y, $x_1, \ldots, x_n$ be physical (measurable) scalar quantities, out of which we want to build a model. The quantities y, x_i come with fundamental physical units $L_1, \ldots, L_m$ with $m \leq n$. Now the model consists in assuming that there is an *a priori* unknown function f such that

$$y = f(x_1, \ldots, x_n) \tag{2.8}$$

In the SI system there are exactly seven fundamental physical units: mass ($L_1 = kg$), length ($L_2 = m$), time ($L_3 = s$), electric current ($L_4 = A$), temperature ($L_5 = K$), amount of substance ($L_6 =$mol) and luminous intensity ($L_7 =$cd), and we postulate that the physical dimension of any measurable scalar quantity can be expressed as a product of powers of the $L_1, \ldots, L_7$.

Example 2.1. *It holds that*

$$[\text{Energy}] = \frac{\mathrm{kg\,m}^2}{\mathrm{s}^2} = L_1 L_2^2 L_3^{-2}\,.$$

Here, the number of fundamental physical units is $m = 3$.

Step 1: Remove redundancies from the model If the unknown function f is a function of n variables $x_1, \ldots, x_n$ with $m \le n$ fundamental physical units, using strategy in the pendulum example may lead to an underdetermined system of equations (there we had 4 variables with only 3 fundamental units).

To remove such redundancies, it is helpful to translate the problem into the language of linear algebra: Let $L_1, \ldots, L_m$ be our fundamental physical units and identify L_i with to the i-th canonical basis vector

$$e_i = (0, \ldots, 0, 1, 0, \ldots, 0)^T \,,$$

of $\mathbb{R}^m$, with the entry 1 in position i. Now pick a subset $\{p_1, \ldots, p_m\}$ of $\{x_1, \ldots, x_n\}$ so that $p_1, \ldots, p_m$ are linearly independent in the sense that no $[p_i]$ can be expressed in terms of the $[p_1], \ldots, [p_{i-1}], [p_{i+1}], \ldots, [p_m]$. With this correspondence, each $[p_i]$ is a linear combination of the canonical basis vectors $e_1, \ldots, e_m$, and so, by construction, there exist $\alpha_{i,1}, \ldots, \alpha_{i,m} \in \mathbb{R}$ with

$$[p_i] = L_1^{\alpha_{i,1}} L_2^{\alpha_{i,,2}} \cdots L_m^{\alpha_{i,m}} \,, \tag{2.9}$$

such that the vectors

$$v_i = (\alpha_{i,1}, \ldots, \alpha_{i,m}) \in \mathbb{R}^m \,, \quad i = 1, \ldots, m \,,$$

are linearly independent and therefore form a basis of $\mathbb{R}^m$.

Example 2.2 (Cont'd)**.** *The dimensional unit of energy has the canonical basis representation*

$$\begin{pmatrix} 1 \\ 2 \\ -2 \end{pmatrix} = e_1 + 2e_2 - 2e_3$$

if considered as a vector in $\mathbb{R}^3$*.*

We call the set $\{p_1, \ldots, p_m\}$ the *set of primary variables* or *primary quantities*.[1] The *secondary variables* are then defined as the set

$$\{s_1, \ldots, s_{n-m}\} = \{x_1, \ldots, x_n\} \setminus \{p_1, \ldots, p_m\} \,. \tag{2.10}$$

By construction, the secondary variables are expressible as linear combinations of the primary variables. In terms of primary and secondary variables our postulated model (2.8) reads (with an abuse of notation)

$$y = f(p_1, \ldots, p_m, s_1, \ldots, s_{n-m}) \,. \tag{2.11}$$

Step 2: Construct dimensionless quantities Having refined our model according to Step 1 above, we construct a quantity z with $[z] = [y]$ such that

$$z = p_1^{\alpha_1} \cdots p_m^{\alpha_m} \,, \tag{2.12}$$

with uniquely defined coefficients $\alpha_1, \ldots, \alpha_m \in \mathbb{R}$. Now call Π the dimensionless quantity given by $\Pi = y/z$, in other words

$$\Pi = \frac{f(p_1, \ldots, p_m, s_1, \ldots, s_{n-m})}{p_1^{\alpha_1} \cdots p_m^{\alpha_m}} \,, \tag{2.13}$$

[1] We assume that the set of primary variables exists, otherwise we have to rethink our postulated model f. (It is important that you understand this reasoning.)

We want to express Π solely as a function of the primary variables. To this end note that we can write

$$[s_j] = [p_1]^{\alpha_{j,1}} \cdots [p_m]^{\alpha_{j,m}}$$

for suitable coefficients $\alpha_{j,1}, \ldots, \alpha_{j,m}$; this can be done for all the s_j. Along the lines of the previous considerations we introduce z_j with $[z_j] = [s_j]$ by

$$z_j = p_1^{\alpha_{j,1}} \cdots p_m^{\alpha_{j,m}}$$

and define the dimensionless quantity $\Pi_j = s_j/z_j$. Note that, by the rank-nullity theorem there are exactly $n-m$ such quantities where $n-m$ is the dimension of the nullspace of the matrix spanned by the $x_1, \ldots, x_n$ Replacing all the s_j by $z_j\Pi_j$, we can recast (2.13) as

$$\Pi = F(p_1, \ldots, p_m, \Pi_1, \ldots, \Pi_{n-m})\,, \tag{2.14}$$

with the shorthand

$$F(p_1, \ldots, p_m, \Pi_1, \ldots, \Pi_{n-m}) := \frac{f(p_1, \ldots, p_m, z_1\Pi_1, \ldots, z_{n-m}\Pi_{n-m})}{p_1^{\alpha_1} \cdots p_m^{\alpha_m}}\,, \tag{2.15}$$

This suggests that we regard F as a function $F\colon P \to \mathbb{R}$ of the primary variables $p_1, \ldots, p_m$ where $P = \operatorname{span}\{p_1, \ldots, p_m\} \subset \mathbb{R}^m$. Note, however, that Π and Π_j are dimensionless (and so is F). Hence F is even independent of the primary variables, for otherwise we could rescale, say, p_1, by which none of $p_2, \ldots, p_m$ or any of the dimensionless quantities Π_j change (they are dimensionless); as a consequence F is a homogeneous function of degree 0 in p_1, and the same is true for any of the other p_j. Therefore F is independent of $p_1, \ldots, p_m$.

Step 3: Find y up to a multiplicative constant The last statement can be rephrased by saying that y can be expressed in terms of a relation between dimensionless parameters. The surprising implication is that the unknown quantity y that we want to model has the functional form of z in (2.12), namely,

$$y = \Pi\, p_1^{\alpha_1} \cdots p_m^{\alpha_m}\,. \tag{2.16}$$

No trigonometric functions, no logarithms or anything like this appear here. We summarize our findings by the famous Buckingham Π Theorem (see [Buc14]).

Theorem 2.3 (Buckingham, 1914)**.** *Any complete physical relation of the form $y = f(x_1, \ldots, x_n)$ can be reduced to a relation between the associated dimensionless quantities, where the number of independent dimensionless quantities is equal to $n-m$, the difference between the number physical quantities $x_1, \ldots, x_n$ and the number of fundamental dimensional units. That is, there exists a function $\Phi\colon \mathbb{R}^{n-m} \to \mathbb{R}$ such that*

$$y = z\,\Phi(\Pi_1, \ldots, \Pi_{n-m}) \tag{2.17}$$

or, in other words,

$$y = z\,\Phi\left(\frac{s_1}{z_1}, \ldots, \frac{s_{n-m}}{z_{n-m}}\right)\,. \tag{2.18}$$

2.2 A historical example

To appreciate the power and usefulness of Buckingham's theorem we have to see it in action. The following example is taken from [IBM$^+$05]; see also [Tay50] for the original article. When the U.S. tested the atomic bomb "Trinity" at Los Alamos in 1945, the British physicist and mathematician Sir Geoffrey Taylor[2] could quite accurately estimate the mass of the bomb based on the dimensional analysis of the radius of the shock wave as a function of time, using only film footage of the explosion. (The data was still classified then.) Taylor assumed that the expanding shock wave R due to the explosion could be expressed as

$$R = f(t, E, \rho, p) \tag{2.19}$$

where t is time, E the released energy (that is a function of the mass of the bomb), ρ is the density of the ambient air and p denotes air pressure. The corresponding physical units are (cf. Example 2.1)

$$[R] = L_2\,, \quad [t] = L_3\,, \quad [E] = L_1 L_2^2 L_3^{-2}\,, \quad [\rho] = L_1 L_2^{-3}\,, \quad [p] = L_1 L_2^{-1} L_3^{-2}\,,$$

with the three fundamental physical units L_1 (mass), L_2 (length) and L_3 (time). The latter implies that there are three primary variables where, without loss of generality we pick t, $E\,\rho$. Then

$$[t] = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \quad [E] = \begin{pmatrix} 1 \\ 2 \\ -2 \end{pmatrix}, \quad [\rho] = \begin{pmatrix} 1 \\ -3 \\ 0 \end{pmatrix}. \tag{2.20}$$

(Remember that Taylor wanted to find out how large E was, so our choice was only to about 67 percent arbitrary.) Expressing $[R]$ in terms of the chosen basis then leads to the linear system of equations

$$Ax = b \tag{2.21}$$

with unknown $x = (\alpha_1,\, \alpha_2,\, \alpha_3)$ and coefficients

$$A = \begin{pmatrix} 0 & 1 & 1 \\ 0 & 2 & -3 \\ 1 & -2 & 0 \end{pmatrix}, \quad b = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}. \tag{2.22}$$

By construction the matrix A has maximum rank and so the unique solution of (2.21)–(2.22) is $x = (2/5,\, 1/5,\, -1/5)^T$, from which we find

$$[R] = \left(\frac{t^2 E}{\rho}\right)^{1/5}. \tag{2.23}$$

Following (2.16) we can thus form a dimensionless quantity by setting

$$\Pi = R\left(\frac{t^2 E}{\rho}\right)^{-1/5}. \tag{2.24}$$

[2] Sir Geoffrey Ingram Taylor F.R.S. (1886–1975) was a British physicist and mathematician, expert on fluid dynamics and part of the British delegation to the Manhattan project between 1944 and 1945. The famous Taylor-Couette instability is named after him.

time (in miliseconds)	radius (in meters)
0.10	11.1
0.24	19.9
0.38	25.4
0.52	28.8
0.66	31.9
⋮	⋮
3.53	61.1
3.80	62.9
4.07	64.3
4.34	65.6
4.61	67.4
⋮	⋮

Table 2: Radius of the shock wave as a function of time (from [Tay50]).

Analogously, we find for the remaining linearly dependent (secondary) quantity

$$\Pi_1 = p\left(\frac{E^2\rho^3}{t^6}\right)^{-1/5}. \tag{2.25}$$

Now combining (2.24)–(2.25), Buckingham's theorem tells us that

$$R = \left(\frac{t^2E}{\rho}\right)^{1/5}\Phi\left(p\left(\frac{E^2\rho^3}{t^6}\right)^{-1/5}\right). \tag{2.26}$$

Here $\Phi(\cdot)$ is a yet unknown function of Π_1 that must be determined from appropriate data.[3] In Taylor's case a reasonable approximation was $\Phi(\Pi_1) \approx \Phi(0)$, simply because t was small compared whereas E was fairly large, in other words:

$$p \ll \left(\frac{E^2\rho^3}{t^6}\right)^{1/5},$$

assuming that the numerical values of p and ρ were of order 1. Taylor did experiments with small explosives and found out that $\Phi(0) \approx 1$, which led him to conclude that [Tay50]

$$R = \left(\frac{t^2E}{\rho}\right)^{1/5} \tag{2.27}$$

describes the radius of the shock wave as a function of time t and the parameters E, ρ. Given measurement data $(t, R(t))$ and the value of the density of air $\rho = 1.25\mathrm{kg/m}^3$, it is then possible to estimate E and hence the mass of the nuclear bomb. Taylor had the following data:

Taking the log on both sides of (2.27) yields

$$\log R = \frac{2}{5}\log t + b\,, \quad b = \frac{1}{5}\log E - \frac{1}{5}\log\rho \tag{2.28}$$

[3] More about this in the next section.

from which $E \approx 8.05 \cdot 10^{13}$ Joules can be obtained by a least squares fit of the data. (See the next section.) Using the conversion factor 1 kiloton $= 4.186 \cdot 10^{12}$ Joules Taylor estimated the weight of the nuclear bomb Trinity as 19.2 kilotons. The true weight of the bomb was about 21 kilotons which was revealed much later. Thus Taylor estimate proved indeed quite accurate.

Problems

Exercise 2.4. *Explain the statement $p \ll (E^2\rho^3/t^6)^{1/5}$ below equation (2.26). Why would a statement like $t^6 \ll E^2$ or $t \approx 0$ be meaningless?.*

Exercise 2.5. *Prove that the $\alpha_1, \ldots, \alpha_m$ in (2.12) are unique.*

Exercise 2.6. *A recurrent theme in both U.S. kitchens and books on mathematical modelling is the question how to cook a turkey. Cookbook sometimes give directions of the form: "Set the oven to $T_0 = 180°\mathrm{C}$ and put it in the oven for 20 minutes per pound of weight." Analyse (and criticise) this rule of thumb based on the following modelling assumptions:*

a) A piece of meat is cooked when its minimum internal temperature has reached a certain value $T_{\min}$ that may depend, e.g., on the type of meat.

b) The cooking time t is a function of the difference ΔT between the oven temperature and the raw meat, the thermal conductivity κ of the meat, its average density ρ and the characteristic size (length) l of the piece of meat.

c) Most mammals and birds obey the law of elastic similarity that says that larger animals have relatively thicker trunks so as to ensure a certain stability against external stress while being efficient regarding the use of material (e.g. bones) [Bie05]. Maximum efficiency is achieved when the vertical thickness t of a body (trunk) scales with its length l as $t \propto l^{3/2}$. Together with the fact that volume is porportional to lA, with $A \propto t^2$ being cross-sectional area, the elastic similarity principle implies that the mass of a body of a bird or mammal is proportional to $lt^2 \propto l^4$.

d) Temperature is measured in units of energy per volume (and so are temperature differences), the thermal conductivity measures the amount of energy crossing a unit cross-sectional area per second divided by the temperature gradient perpendicular to this area, i.e.,

$$[\kappa] = \frac{[\mathit{energy}] \times [\mathit{length}]}{[\mathit{area}] \times [\mathit{time}] \times [\mathit{temperature}]}\,.$$

3 Arguments from data

Mathematical tools & concepts: linear algebra, random variables
Suggested reference: [BTF$^+$99]

Recall the problem of Section 2.2: Determine the size of a nuclear bomb from measurement data $\{(t_i, R(t_i))\colon i = 1, 2, \ldots, N\}$ based on the model (2.27). The equivalent logarithmic representation (2.28) is an equation of the form

$$y(t) = \alpha x(t) + \beta\,,$$

with the new variables $y = \log R$ and $x = \log t$, known parameter α and unknown parameter β. If the measurement data and the model were exact, it would be possible to estimate the unknown coefficient β from a single measurement $(x(t_1), y(t_1)) = (\log t_1, \log R(t_1))$.

If we take into account that measurement data are subject to measurement errors coming, e.g., from the measurement apparatus or from other sources of error that are not part of the measurement model, then an apparently more realistic model could be an equation of the form

$$Y(t) = \alpha X(t) + \beta + \epsilon(t)\,, \tag{3.1}$$

where ϵ is a (typically stationary Gaussian) stochastic process that represents the measurement or, more generally, statistical noise.[4]

3.1 Linear regression models

Generally, we consider linear models of the form

$$Y(t) = \sum_{i=1}^{n} \alpha_i X_i(t) + \beta + \epsilon(t)\,. \tag{3.2}$$

Here $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$ and β are the *a priori* unknown model parameters, Y denotes the dependent variable, that for simplicity is assumed to be scalar and that depends on m independent variables $X_1, \ldots, X_n$. We assume that ϵ is stationary with zero mean and covariance σ^2, i.e. for all t it holds

$$\mathbb{E}[\epsilon(t)] = 0\,, \quad \mathbb{E}[\epsilon(t)^2] = \sigma^2\,. \tag{3.3}$$

Note that the model is called *linear* because of the linear dependence on the parameters; it does not matter whether the function is linear or nonlinear in the X_i, for we can always redefine the X_i by a suitable nonlinear transformation, such as $X_i \mapsto \log X_i$, to obtain an equation of the form 3.2.

Example 3.1 (Cobb-Douglas production function)**.** *Another model of the above form is the popular Cobb-Douglas model that is used to express the amount of output of a production as a function of capital and labor [Dou76]. In its most standard form for production of a single good with two factors, the function is*

$$Y = \kappa L^{\alpha_1} C^{\alpha_2} e^{\epsilon}\,, \tag{3.4}$$

[4] That is, $\epsilon(t) = \epsilon(t;\omega)$ is a random variable $\epsilon(t;\cdot)\colon \Omega \to \mathbb{R}$ for each fixed t. We adopt the convention that capital letters, such as X, Y are used to distinguish real-valued random variables (i.e. measurable functions $X, Y\colon \Omega \to \mathbb{R}$ on some probability space $(\Omega, \mathcal{E}, P)$, with $\mathcal{E}$ being a σ-algebra over the set Ω) from their values $x, y \in \mathbb{R}$.

with Y the total production (the real value of all goods produced in a year), L the labor input (the total number of person-hours worked in a year), C the capital input (the real value of all machinery, equipment and buildings), κ the productivity, and ϵ the statistical error. The coefficients α_i are called the output elasticities*; they are a measure for the succeptibility of the output to a change in levels of either labor or capital used. If $\alpha_1 + \alpha_2 = 1$, then doubling the usage of capital C and labor L will also double output Y.*

Taking the logarithm on both sides of (3.4), we have

$$\log Y = \alpha_1 \log L + \alpha_2 \log C + \log \kappa + \epsilon\,, \tag{3.5}$$

where the right hand side of the equation is an affine function of the form

$$f(x) = \boldsymbol{\alpha}^T x + \beta + \epsilon\,, \tag{3.6}$$

with the coefficients $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)^T$ and $\beta = \log \kappa$ and dependent variables $(X_1, X_2) = (\log L, \log C)$. It is commonly assumed that ϵ is a zero-mean Gaussian random variable with variance σ^2 that is independent of C and L.[5]

Measurements. Without loss of generality we can assume that

$$\beta = 0\,.$$

This is so, because we can always treat β as one of the α_i for $X_i(t) = 1$. (See Exercise 3.13) below.) We now want to estimate the unkown model parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$ from, say, N noisy measurements

$$M := \{(\mathbf{x}(t_j), y(t_j)) \colon j = 1, \ldots, N\} \tag{3.7}$$

of the measurement vector $(\mathbf{x}, y) = (x_1, \ldots, x_n, y)$. In other words, we have N observational equations of the form

$$y(t_j) = \sum_{i=1}^{n} \alpha_i x_i(t_j) + \epsilon(t_j)\,, \quad j = 1, \ldots, N\,, \tag{3.8}$$

It is convenient to arrange the data as the matrix

$$\mathcal{X} = \begin{pmatrix} x_1(t_1) & x_2(t_1) & \cdots & x_n(t_1) \\ \vdots & \vdots & & \vdots \\ x_1(t_N) & x_2(t_N) & \cdots & x_n(t_N) \end{pmatrix} \in \mathbb{R}^{N \times n} \tag{3.9}$$

and define the vectors

$$\mathcal{Y} = \begin{pmatrix} y(t_1) \\ \vdots \\ y(t_N) \end{pmatrix}, \quad \boldsymbol{\epsilon} := \begin{pmatrix} \epsilon(t_1) \\ \vdots \\ \epsilon(t_N) \end{pmatrix}. \tag{3.10}$$

Thus (3.8) reads

$$\mathcal{Y} = \mathcal{X}\boldsymbol{\alpha} + \boldsymbol{\epsilon}\,, \tag{3.11}$$

where $\boldsymbol{\alpha}$ is interpreted as a column vector. At this stage it does not matter anymore whether the data are interpreted as N i.i.d. realizations of a random variable, or as a time series with N independent data points.[6] In the following we will regard $\mathcal{X}$ as a measurement of the determinstic variables $X = (X_1, \ldots, X_n)$ where the error in the dependent variable $\mathcal{Y}$ comes from the additive noise $\boldsymbol{\epsilon}$.

[5] The Gaussianity is a consequence of the Central Limit Theorem. Can you explain why?

[6] Here "i.i.d." stands for "independent and identically distributed".

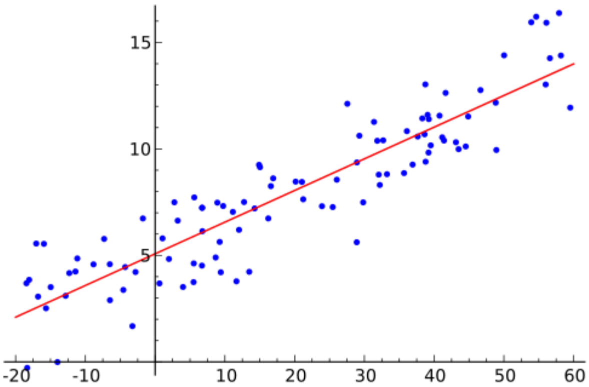


Figure 3.1: Linear regression for $n = 1$: Find the straight line that minimizes the sum of squared deviations from the data points.

3.2 Least squares and maximum likelihood principles

We want to estimate the unknown coefficients $\boldsymbol{\alpha}$ from (independent) observations stored in the data matrices $(\mathcal{X}, \mathcal{Y})$.

Assumption 3.2. *We implement the following standing assumptions for (3.12):*

a) The vector $\boldsymbol{\epsilon} \in \mathbb{R}^N$ is a random variable with mean and covariance

$$\mathbb{E}[\boldsymbol{\epsilon}] = 0\,, \quad \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T] = \sigma^2 I_{N\times N}\,.$$

b) The matrix $\mathcal{X}$ has full column rank $n \leq N$.

Least squares. A general procedure to determine the unknown parameters is to choose $\boldsymbol{\alpha}$, such that the error $\boldsymbol{\epsilon}$ is small in a suitable norm $\|\cdot\|$. A convenient choice is to minimize $\boldsymbol{\epsilon}$ in the ℓ^2 norm, in other words, we seek $\boldsymbol{\alpha}$, such that the sum of squared deviations

$$\sum_{i=1}^{N} |\epsilon(t_i)|^2 = (\mathcal{Y} - \mathcal{X}\boldsymbol{\alpha})^T(\mathcal{Y} - \mathcal{X}\boldsymbol{\alpha})$$

is minimized (cf. Figure 3.1). To this end we define the function

$$S_N \colon \mathbb{R}^n \to \mathbb{R}\,, \quad \boldsymbol{\alpha} \mapsto (\mathcal{Y} - \mathcal{X}\boldsymbol{\alpha})^T(\mathcal{Y} - \mathcal{X}\boldsymbol{\alpha})\,, \tag{3.12}$$

Under above assumptions S_N is differentiable and strictly convex, hence it has a unique minimum. (See Exercise 3.11 below.)

Definition 3.3. *We call the minimizer*

$$\boldsymbol{\alpha}^* = \operatorname*{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^n} S_N(\boldsymbol{\alpha}) \tag{3.13}$$

the least squares estimator (LSE) *of $\boldsymbol{\alpha}$ given the data $(\mathcal{X}, \mathcal{Y})$.*

Theorem 3.4 (Least squares estimator)**.** *The LSE is given by*

$$\boldsymbol{\alpha}^* = (\mathcal{X}^T\mathcal{X})^{-1}\mathcal{X}^T\mathcal{Y} \tag{3.14}$$

Proof. The first and second derivatives of S_N with respect to the parameter vector $\boldsymbol{\alpha}$ are given by

$$\nabla S_N(\alpha) = 2\mathcal{X}^T\mathcal{X}\boldsymbol{\alpha} - 2\mathcal{X}^T\mathcal{Y}\,, \quad \nabla^2 S_N(\alpha) = 2\mathcal{X}^T\mathcal{X}\,.$$

Equating the first derivative to zero we obtain the normal equation

$$\mathcal{X}^T\mathcal{X}\boldsymbol{\alpha} - \mathcal{X}^T\mathcal{Y} \ = 0,$$

that, under the assumption that the data matrix $\mathcal{X} \in \mathbb{R}^{N\times n}$ has maximum rank n, has the unique solution

$$\boldsymbol{\alpha}^* = (\mathcal{X}^T\mathcal{X})^{-1}\mathcal{X}^T\mathcal{Y}\,.$$

Since, by the same argument, $\nabla^2 S_N(\alpha)$ is positive definite and independent of α, it follow that $S_N(\alpha^*)$ is the unique minimum of S_N. □

Remark 3.5. *The LSE can be interpreted as the best approximation of the solution to the linear equation*

$$\mathcal{X}\boldsymbol{\alpha} = \mathcal{Y} \tag{3.15}$$

in the Euclidean norm for given $\mathcal{X} \in \mathbb{R}^{N\times n}$, $\mathcal{Y} \in \mathbb{R}^N$. If $N > n$ the linear system of equations (3.15) does not have a solution, and by construction, the LSE satisfies the best-approximation property

$$\alpha^* = \operatorname*{argmin}_{\alpha\in\mathbb{R}^n} \|\mathcal{X}\boldsymbol{\alpha} - \mathcal{Y}\|^2\,,$$

with $\|\cdot\|$ denoting Euclidean norm. In other words, α^* is the orthogonal projection of $\mathcal{Y}$ onto the column space of $\mathcal{X}$.

Remark 3.6. *It can be shown (see, e.g., [BTF+99, Sec. 3.4]) that, given the assumptions from page 15, the LSE is the linear estimator of α that has minimum variance among all linear estimators (best linear estimator).*

Maximum likelihood principle. So far we have viewed (3.8) as a fitting problem for the unknown parameter vector $\boldsymbol{\alpha}$ without making any assumptions on the very nature of the measurement error ϵ or on the distributions of (X, Y). Specifically, we assume that X is deterministic and that all the $\epsilon(t_i)$ are uncorrelated Gaussian random variables.[7]

Assumption 3.7. *The vector $\boldsymbol{\epsilon} \in \mathbb{R}^N$ is an N-dimensional Gaussian random variable with zero mean and covariance*

$$\mathbb{E}[\epsilon\epsilon^T] = \sigma^2 I_{N\times N}\,.$$

[7] Thanks to the central limit theorem, this is often a reasonable choice.

Then in the linear regression model (3.8), the dependent variables Y are Gaussian with

$$Y \sim \mathcal{N}(\mathcal{X}\boldsymbol{\alpha}, \sigma^2)\,. \tag{3.16}$$

We define the likelihood function of Y as the Gaussian density of Y, but considered as a function of the parameters $\boldsymbol{\alpha}$ and σ^2, i.e.

$$L(\boldsymbol{\alpha}, \sigma^2; \mathbf{x}, y) = (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{|y - \boldsymbol{\alpha}^T\mathbf{x}|^2}{2\sigma^2}\right)\,. \tag{3.17}$$

By Assumption 3.7, the likelihood function of the data vector $\mathcal{Y}$ then is

$$\begin{aligned} L(\boldsymbol{\alpha}, \sigma^2; \mathcal{X}, \mathcal{Y}) &= \prod_{i=1}^{N} L(\boldsymbol{\alpha}, \sigma^2; \mathbf{x}(t_i), y(t_i)) \\ &= (2\pi\sigma^2)^{-N/2} \exp\left(-\frac{1}{2\sigma^2}(\mathcal{Y} - \mathcal{X}\boldsymbol{\alpha})^T(\mathcal{Y} - \mathcal{X}\boldsymbol{\alpha})\right) \end{aligned} \tag{3.18}$$

The idea of the maximum likelihood principle now is to find parameters $\boldsymbol{\alpha}, \sigma^2$ that fit the given data $(\mathcal{X}, \mathcal{Y})$ best. (Remember that we assume that $\mathcal{X}$ is deterministic.) This is to say that we seek $\boldsymbol{\alpha}, \sigma$, such that the joint normal density $f(\mathcal{X}, \mathcal{Y}; \boldsymbol{\alpha}, \sigma^2)$ of the data $(\mathcal{X}, \mathcal{Y})$ attains its maximum, which means that the data is best explained by these parameter values. If we consider the density of the normal distribution as a function of the unknown parameters (over which we want to maximize), we have our likelihood function

$$L(\boldsymbol{\alpha}, \sigma^2; \mathcal{X}, \mathcal{Y}) = f(\mathcal{X}, \mathcal{Y}; \boldsymbol{\alpha}, \sigma^2) \tag{3.19}$$

Definition 3.8. *We call*

$$(\hat{\boldsymbol{\alpha}}, \hat{\sigma}^2) = \underset{(\boldsymbol{\alpha}, \sigma^2) \in \mathbb{R}^n \times \mathbb{R}_+}{\operatorname{argmax}} L(\boldsymbol{\alpha}, \sigma^2; \mathcal{X}, \mathcal{Y}) \tag{3.20}$$

the maximum likelihood estimator (MLE) *of* $(\boldsymbol{\alpha}, \sigma^2)$ *given the data* $(\mathcal{X}, \mathcal{Y})$.

The logarithm is monotonic, and it is often more convenient to maximize the log-likelihood

$$\log L(\boldsymbol{\alpha}, \sigma^2; \mathcal{X}, \mathcal{Y}) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathcal{Y} - \mathcal{X}\boldsymbol{\alpha})^T(\mathcal{Y} - \mathcal{X}\boldsymbol{\alpha}) \tag{3.21}$$

rather than the likelihood function. It can be readily seen that the maximizer of the log-likelihood function is the MLE. The proof of the next theorem is left as an exercise to the reader.

Theorem 3.9 (Maximum likelihood estimator)**.** *The MLE of* $(\boldsymbol{\alpha}, \sigma^2)$ *is given by*

$$\begin{aligned} \hat{\boldsymbol{\alpha}} &= (\mathcal{X}^T\mathcal{X})^{-1}\mathcal{X}^T\mathcal{Y} \\ \hat{\sigma}^2 &= \frac{1}{N}(\mathcal{Y} - \mathcal{X}\hat{\boldsymbol{\alpha}})^T(\mathcal{Y} - \mathcal{X}\hat{\boldsymbol{\alpha}}) \end{aligned} \tag{3.22}$$

Proof. Exercise. □

Note that $\hat{\boldsymbol{\alpha}} = \boldsymbol{\alpha}^*$, i.e., the MLE of $\boldsymbol{\alpha}$ agrees with the LSE. It should be stressed that both LSE and MLE are linear transformations of the random observation $\mathcal{Y}$, hence they are both random variables [BTF+99].

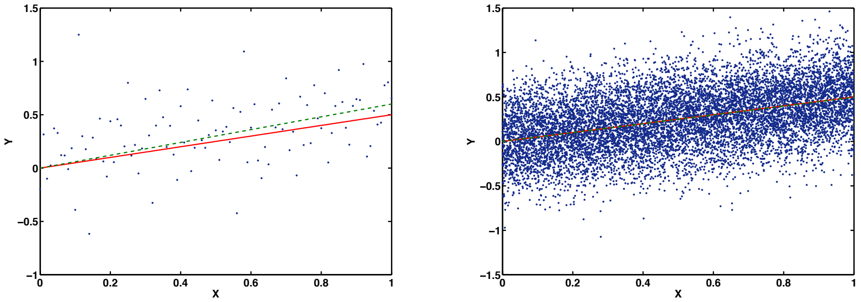


Figure 3.2: LSE and MSE of the one-dimensional model (3.23) with $N = 10^2$ and $N = 10^6$ data points (red: exact model, green: estimated model).

Remark 3.10. *The MLE $\hat{\sigma}^2 = \hat{\sigma}^2(N)$ for the variance σ^2 of the measurement error ϵ is asymptotically unbiased, i.e., it holds*

$$\sigma^2 = \lim_{N\to\infty} \mathbb{E}[\hat{\sigma}^2(N)]\,.$$

Example. As an illustrative example we consider the model

$$Y(t) = \frac{1}{2}X(t) + \epsilon(t) \tag{3.23}$$

with $\epsilon \sim \mathcal{N}(0, \sigma^2)$. We have generated a random realization of ϵ with N data points, from which we generated observation data

$$x(t_i) = i/N\,, \quad y(t_i) = 0.5x(t_i) + \epsilon(t_i)\,, \quad i = 1, \ldots, N\,.$$

We then computed the LSE (equivalently: the MLE) for the given data. Figure 3.2 shows two typical realizations for $N_1 = 10^2$ and $N_2 = 10^5$. The corresponding estimates are $\alpha_1^* = 0.5572$ for $N_1 = 10^2$ and $\alpha_2^* = 0.5000$ for $N_2 = 10^5$. The fact that the estimator is linear in Y, together with the fact that the noise has mean zero, implies that the LSE and MLE estimators are unbiased with

$$\mathbb{E}[\alpha^*] = \mathbb{E}[\hat{\alpha}] = \alpha\,.$$

This is to say that both MLE and LSE estimators will always fluctuate around the true value α, no matter how small N is. By the law of large numbers, their empirical mean converges to α when the estimation is repeated infinitely often.[8]

Generalized linear models (whitening). The linear regression model under Assumption 3.2 is a special case of a model with error variance

$$\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T] = \sigma^2 W\,,$$

[8] Clearly the estimator converges faster when N is larger, because by the central limit theorem its variance decreases with $1/N$.

with $W \in \mathbb{R}^{N\times N}$ being a symmetric and positive definite (s.p.d.) matrix. A noticeable feature of this so-called *generalized linear modeI* is that there are $N(N+1)/2$ additional unknowns in the game, if W is not known *a priori*. Therefore it is impossible to estimate both $\boldsymbol{\alpha}$ and $\sigma^2 W$ without further assumtions on the measurement error, if the sample size N is fixed.

For simplicity we assume that W is given. In this case it is possible to reduce the parameter estimation problem for the generalized model to the estimation problem for (3.11). To this end, recall that both W and its inverse W^{-1} have s.p.d. square roots, i.e., there exist s.p.d. matrices Q, R, such that

$$W = QQ\,, \quad W^{-1} = RR\,.$$

Multiplying (3.11) with $R = W^{-1/2}$ from the left, we obtain

$$R\mathcal{Y} = R\mathcal{X}\boldsymbol{\alpha} + R\boldsymbol{\epsilon}\,, \tag{3.24}$$

which upon redefining

$$\tilde{\mathcal{Y}} = R\mathcal{Y}\,, \quad \tilde{\mathcal{X}} = R\mathcal{X}\,, \tilde{\boldsymbol{\epsilon}} = R\boldsymbol{\epsilon}$$

can be recast as

$$\tilde{\mathcal{Y}} = \tilde{\mathcal{X}}\boldsymbol{\alpha} + \tilde{\boldsymbol{\epsilon}}\,, \tag{3.25}$$

Rescaling the observation variables by the square root of the inverse error covariance is known by the name of *whitening* because now

$$\mathbb{E}[\tilde{\boldsymbol{\epsilon}}\tilde{\boldsymbol{\epsilon}}^T] = \sigma^2 I_{N\times N}\,.$$

As a consequence the last equation is again a standard linear system satisfying Assumption 3.2, and we can apply one of the Theorems 3.4 or 3.9 to find that the best linear estimator for the unknown parameters $\boldsymbol{\alpha}$ reads

$$\tilde{\boldsymbol{\alpha}}^* = (\mathcal{X}^T W^{-1}\mathcal{X})^{-1}\mathcal{X}^T W^{-1}\mathcal{Y}\,. \tag{3.26}$$

The assertion that $\tilde{\boldsymbol{\alpha}}^*$ is unbiased and indeed the best linear estimator for the parameters in the generalized linear model (3.24) is called *Gauss-Markov-Theorem*; the interested reader is referred to [BTF+99, Thm. 4.4] for details.

Problems

Exercise 3.11. *For the linear model (3.1) with unknown scalar coefficients* (α, β)*, we define the LSE* (α^*, β^*) *as the minimizer of the function*

$$S_N(\alpha,\beta) = \sum_{i=1}^N (y(t_i) - \alpha x(t_i) - \beta)^2$$

Prove that S_N *is convex and strictly convex if*

$$\sum_{i=1}^N (x(t_i) - \bar{x}) > 0\,, \quad \bar{x} = \frac{1}{N}\sum_{i=1}^N x(t_i)$$

Exercise 3.12. *If we drop the assumption that the data matrix $\mathcal{X} \in \mathbb{R}^{N \times n}$ has full rank $n \leq N$, the LSE $\boldsymbol{\alpha}^*$ is given by any solution of the normal equations*

$$\mathcal{X}^T \mathcal{X} \boldsymbol{\alpha} - \mathcal{X}^T \mathcal{Y} = 0.$$

In this case α^* is no longer unique. Show that $S_N(\alpha)$ as defined in (3.12) attains its minimum for any solution of the normal equations.

Exercise 3.13. *Consider the linear model (3.1) with unknown scalar coefficients (α, β). Compute LSE and MLE of the parameters (α, β); cf. exercise 3.11.*

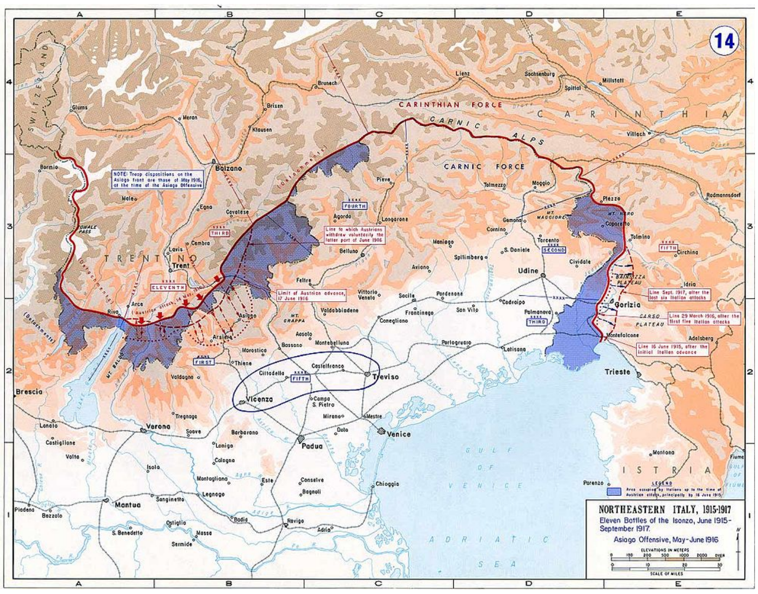

14
CARINTHIAN FORCE
CARINTHIA
CARNIC ALPS
CARNIC FORCE
TRENTINO
Bolzano
Trent
Udine
Gorizia
Trieste
Treviso
Vicenza
Padua
Venice
Verona
Brescia
Mantua
ISTRIA
ADRIATIC SEA
GULF OF VENICE
LEGEND
Area occupied by Italians up to the time of Austrian attack, principally by 16 June 1915
NORTHEASTERN ITALY, 1915-1917
Eleven Battles of the Isonzo, June 1915-September 1917.
Asiago Offensive, May-June 1916
ELEVATIONS IN METERS
0 100 200 500 1000 2000 OVER
0 10 20 30
SCALE OF MILES

Figure 4.1: Italian Front 1915–1917 (source: History Department of the US Military Academy).

4 Population models in biology

Mathematical tools & concepts: ODE systems, eigenvalues
Suggested reference: [IBM+05]

A famous model that describes how biological populations evolve in time is the Lotka–Volterra predator–prey model that is a variant of the logistic equation (1.5). It was derived independently by Alfred J. Lotka in 1910 in the context of chemical reactions [Lot10] and by Vito Volterra in 1926 in order to explain a found paradox in the fish catches in the Adriatic Sea after World War I [Vol26].

The apparent paradox is an interesting one: In 1915 when Italy declared war on Austria and both countries were afraid of being invaded via their sea ports, they set mines in the Adriatic Sea to prevent the other party to reach their harbours. As a consequence, fishing was stopped in the Adriatic Sea during World War I. After the war, when the mines had been removed, the fishermen expected an enourmous fish catch as the fish populations had had more than three years to recover. However, the opposite was true, and it was Volterra who first came up with an explanation using a mathematical model that is nowadays known as the *Lotka-Volterra equation.*

4.1 Lotka-Volterra model

The Lotka-Volterra equation is the logistic equation for a biological population with two species, one being the predator, the other one being the prey, for example a fish predator and its prey. Let $P(t)$ the number of predators and $N(t)$ the number of prey at time t. We assume that there is a enough food available for the prey (e.g. plankton), but they are eaten by the predator. The rate of change for the the prey per capita can be modelled by

$$\frac{\dot{N}(t)}{N(t)} = a - bP(t)\,, \tag{4.1}$$

which describes exponential growth of the prey with effective growth rate $a - bP(t)$. The reproduction rate of the predator population depends on whether there is enough for them to eat; they die without prey. Letting $dN(t) - c$ be the effective growth rate, the size of the predator population is governed by

$$\frac{\dot{P}(t)}{P(t)} = -d + cN(t)\,. \tag{4.2}$$

We assume that a (prey reproduction rate), b (the rate of predation upon the prey), c (growth rate of the predator population) and d (predator mortality) are all strictly positive and that (4.1)–(4.2) are equipped with suitable initial conditions $N(0) = N_0 > 0$ and $P(0) = P_0 > 0$.

It is customary to rescale the free variable t and the dependent variables N, P to recast the equations in dimensionless form.[9] To this end we define

$$\tau = at\,, \quad u = \frac{c}{d}N\,, \quad v = \frac{b}{a}P$$

in terms of which the Lotka-Volterra equations read (see Exercise 4.6)

$$\begin{aligned} \frac{du}{d\tau} &= u(1-v)\,, \quad u(0) = u_0 \\ \frac{dv}{d\tau} &= \mu v(u-1)\,, \quad v(0) = v_0\,, \end{aligned} \tag{4.3}$$

with $\mu = d/a$.

Vector field and fixed points. Even though there is an explicit solution to (4.3), we will not take advantage of this fact, but rather try to get some qualitative insight into the dynamics of the Lotka-Volterra system by studying the underlying autonomous (i.e. time-independent) vector field. Let

$$F_\mu \colon \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}^2\,, \quad (u,v) \mapsto (u(1-v),\, \mu v(u-1)) \tag{4.4}$$

the family of vector fields associated with the Lotka-Volterra system, i.e. the right hand side of (4.3) parametrized by $\mu > 0$. The Lotka-Volterra vector field is depicted in Figure 4.2 for $\mu = 1$. Since F_μ is locally Lipschitz, the Picard-Lindelöf existence and uniqueness theorem for initial value problems [Tes12] implies that (4.3) has a unique solution. The solutions then are the integral curves of F_μ, i.e., for every $(u_0, v_0) \in \mathbb{R}_+ \times \mathbb{R}_+$ a differentiable curve

$$\gamma \colon D \to \mathbb{R}_+ \times \mathbb{R}_+, \quad \tau \mapsto (u(\tau), v(\tau)) \tag{4.5}$$

with $\gamma(0) = (u_0, v_0)$ is a solution of (4.3) if

$$\frac{d}{dt}\gamma(\tau) = F_\mu(\gamma(\tau))\,, \quad \tau \in D \subset \mathbb{R}\,. \tag{4.6}$$

[9] You should be able to explain the rationale behind the dimensionless scaling.

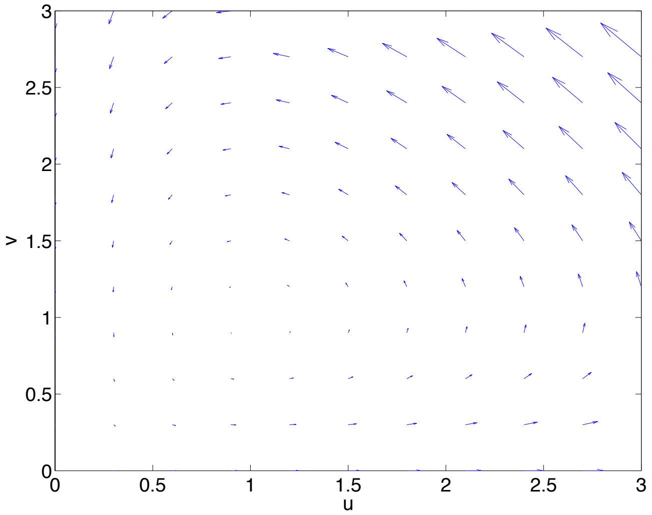


Figure 4.2: Vector field of the Lotka-Volterra system for $\mu = 1$.

In other words, the solution trajectories are everywhere tangential to the vector field. This gives us some idea of how a typical solution of (4.3) could look like. An important property of any vector field are its critical points:

Definition 4.1. *A point* $(u_{\text{eq}}, v_{\text{eq}}) \in \mathbb{R}_+ \times \mathbb{R}_+$ *is called* critical point*,* equilibrium *or* fixed point *of (4.6) if* $F_\mu(u_{\text{eq}}, v_{\text{eq}}) = 0$.

By definition, a solution that goes through a critical point is constant, hence the names "equilibrium" or "fixed point". The Lotka-Volterra system has only two critical points in the positive orthant, namely

$$(u_{\text{eq}}, v_{\text{eq}}) = (0,0) \quad \text{and} \quad (u_{\text{eq}}, v_{\text{eq}}) = (1,1)\,. \tag{4.7}$$

The dynamics when one of the populations is absent at time $\tau = 0$ is relatively easy to understand: if $u_0 = 0$ and $v_0 > 0$, the first equation in (4.3) entails $u(\tau) = u_0$ which, together with the second equation, implies that

$$v(t) = e^{-\mu\tau} v_0\,. \tag{4.8}$$

Thus the predators are bound to die out. Conversely, if $v_0 = 0$ and $u_0 > 0$ it follows by the analogous argument that

$$u(t) = e^{\tau} u_0\,, \tag{4.9}$$

assuming that the prey population has infinite resources available. (They will die out later when they have eaten all the plankton.) If, however, both u_0 and v_0 are different from zero but small, then $u_0 \gg u_0 v_0$ and $v_0 \gg u_0 v_0$, which suggests to neglect the bilinear terms in (4.3) and employ the approximation

$$\begin{aligned} \frac{du}{d\tau} &\approx u\,, \quad u(0) = u_0 \\ \frac{dv}{d\tau} &\approx -\mu v\,, \quad v(0) = v_0\,. \end{aligned} \tag{4.10}$$

This means that the prey population will still grow even though there is a small predator population, while the number of predators will decrease initially; they have not enough to eat, so they die before they can reproduce. Clearly, once the prey population grows so that $u(\tau)v(\tau)$ is no longer small compared to $u(\tau)$ the approximation that is behind (4.9) is no longer valid.

Now let us consider the other equilibrium $(u_{\text{eq}}, v_{\text{eq}}) = (1, 1)$. To linearize F_μ about the point $(1, 1)$, it is convenient to introduce new coordinates by $\xi = u - 1$ and $\eta = v - 1$, in terms of which (4.3) reads

$$\begin{aligned} \frac{d\xi}{d\tau} &= -\eta(1+\xi)\,, \quad \xi(0) = u_0 - 1 \\ \frac{d\eta}{d\tau} &= \mu\xi(\eta+1)\,, \quad \eta(0) = v_0 - 1\,. \end{aligned} \tag{4.11}$$

Upon noting that $u, v \approx 1$ is equivalent to $\xi, \eta \approx 0$, this leads to the following linearized system of differential equations

$$\begin{aligned} \frac{d\xi}{d\tau} &\approx -\eta\,, \quad \xi(0) = u_0 - 1 \\ \frac{d\eta}{d\tau} &\approx \mu\xi\,, \quad \eta(0) = v_0 - 1\,. \end{aligned} \tag{4.12}$$

Upon replacing the "≈" in the linearized equation by equality signs, the latter is equivalent to the differential equation of the pendulum,

$$\frac{d^2\xi}{d\tau^2} = -\mu\xi(\tau)\,,$$

from page 6, with solution

$$\xi(\tau) = A\sin(\mu^{1/2}\tau) + B\cos(\mu^{1/2}\tau)\,. \tag{4.13}$$

The unknown constants $A, B \in \mathbb{R}$ depend only on the initial conditions. Now, since $\eta = -\xi'$, we may suspect that $(\xi(\tau), \eta(\tau))$ and hence $(u(\tau), v(\tau))$ are periodic in the neighbourhood of the equilibrium, with period given by

$$T = 2\pi\mu^{-1/2}\,. \tag{4.14}$$

We will come back to the validity of these kinds of arguments that are based on linearization later on.

Integral curves and periodic orbits. The above reasoning suggests that the solutions of the Lotka-Volterra equation are periodic, at least in the neighbourhood of the critical point $(1, 1)$. We will now show that *all* nonstationary solutions of (4.3), i.e. all solutions away from the critical points are indeed periodic. For this purpose we need the following definition.

Definition 4.2. *A function* $I\colon \mathbb{R}^2 \to \mathbb{R}$ *is called a* first integral *(also:* constant of motion *or* conserved quantity*) if*

$$I(\gamma(\tau)) = I(\gamma(0)) \quad \forall \tau \in D\,.$$

First integrals of an ODE, such as (4.3), are useful in either finding explicit solutions or in finding periodic orbits. Specifically, if an ODE has an integral with compact level sets, then these level sets are candidates for periodic orbits.

Theorem 4.3. *Let $u_0, v_0 > 0$, $(u_0, v_0) \neq (1,1)$. Then $\tau \mapsto (u(\tau), v(\tau))$ is periodic, i.e., there exists a positive number $T \in (0, \infty)$, such that*

$$(u(\tau + T), v(\tau + T)) = (u(\tau), v(\tau)) \quad \forall \tau \geq 0 \,.$$

Proof. Suppose $u \neq 0$ and $v \neq 1$. If we divide the second equation of (4.3) by the first equation and switch to u as the free variable, we obtain

$$\frac{dv}{du} = -\mu \frac{v(1-u)}{u(1-v)} \,, \quad v(u_0) = v_0 \,,$$

which is a separable equation for $v = v(u)$. Separating variables and integrating,

$$\int_{v_0}^{v} \frac{1-\tilde{v}}{\tilde{v}} d\tilde{v} = -\mu \int_{u_0}^{u} \frac{1-\tilde{u}}{\tilde{u}} d\tilde{u} \,,$$

yields

$$\log v - v + C_1 = \mu(u - \log u) + C_2 \,,$$

with integration constants $C_1 = v_0 - \log v_0$ and $C_2 = \mu(\log u_0 - u_0)$. Defining $C = C_1 - C_2$ and switching back to the free variable τ, it follows that

$$C = \mu u(\tau) + v(\tau) - \log(u(\tau)^{\mu} v(\tau)) \quad \forall \tau \,.$$

The function

$$I \colon \mathcal{X} \to \mathbb{R} \,, \quad (u,v) \mapsto \mu u + v - \log(u^{\mu} v)$$

with $\mathcal{X} = \mathbb{R}_+ \times \mathbb{R}_+$ is strictly convex with compact level curves

$$I^{-1}(C) = \{(u,v) \in \mathcal{X} \colon I(u,v) = C\} \subset \mathcal{X}$$

for all

$$C \geq \min_{(u,v) \in \mathcal{X}} I(u,v) = \mu + 1 \,.$$

I is strictly convex, hence it has a unique minimum $I(1,1) = \mu + 1$. It then follows that the solutions of (4.3) with strictly positive initial conditions $u_0, v_0 > 0$, $(u_0, v_0) \neq (1,1)$ lie on the contours $I^{-1}(\cdot)$ having all positive length. Since $\|F_\mu\|$ is finite and bounded away from zero on each contour, it follows that every solution returns to its initial value in finite time $0 < T < \infty$. □

The left panel of Figure 4.3 shows the numerically computed solution for three different initial values that have been computed with the Matlab function `ode15s`. Due to finite precision of the numerical solver, the numerical solution is not exactly periodic and spirals inwards (see the right panel of the figure).

We can say more about the oscillations around the equilibrium $(u_{\text{eq}}, v_{\text{eq}}) = (1,1)$: Their running mean is equal to the equilibrium value.

Lemma 4.4. *It holds that*

$$\frac{1}{T} \int_0^T u(t)\, dt = \frac{1}{T} \int_0^T v(t)\, dt = 1 \,.$$

for all solutions of (4.3) with period $T = T(u_0, v_0)$.

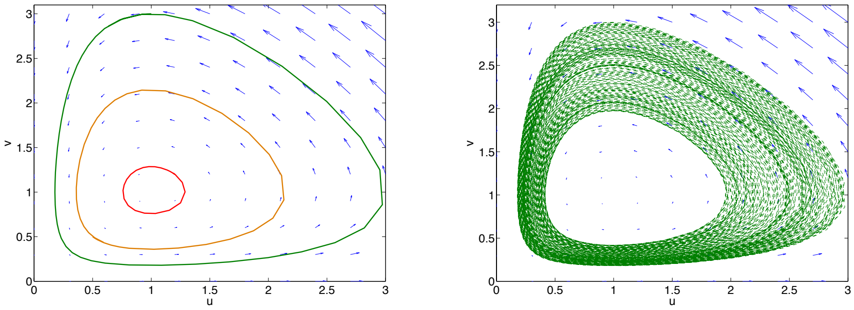


Figure 4.3: Left panel: Solutions of (4.3) for different initial conditions $(u_0, v_0) = (1.2, 1.2)$ (red), $(u_0, v_0) = (0.5, 0.5)$ (orange) and $(u_0, v_0) = (1, 3)$ (green). Right panel: Numerically computed green solution trajectory for 500 time periods.

Proof. Let us prove the rightmost part of the above equality and consider the equation for u only:

$$\frac{du}{d\tau} = u(1 - v)\,.$$

By Theorem 4.3 there exists $T \in (0, \infty)$, such that $u(T) = u(0)$. Separating variables and integrating from 0 to T yields

$$\int_{u(0)}^{u(T)} \frac{du}{u} = \int_0^T (1 - v(t))\, dt\,,$$

which, using that the upper and lower limit in the integral on the left side of the equality coincide, can be recast as

$$T = \int_0^T v(t)\, dt \quad \Leftrightarrow \quad \frac{1}{T}\int_0^T v(t)\, dt = 1\,.$$

The other part of the assertion can be proved in exactly the same way by solving the equation for v. □

The effect of fishing We now want to model the effect that fishing has on fish predators and their prey. For the sake of simplicity we assume that the reduction rate of predator and prey population due to fishing pressure is given by a single parameter $\delta > 0$. The unscaled equations (4.1)–(4.2) then become

$$\begin{aligned} \frac{dN^\delta}{dt} &= N^\delta(a - bP^\delta - \delta)\,, \quad N^\delta(0) = N_0 \\ \frac{dP^\delta}{dt} &= -P^\delta(d - cN^\delta + \delta)\,, \quad P^\delta(0) = P_0\,. \end{aligned} \tag{4.15}$$

In the unscaled form, the nontrivial equilibrium is

$$\left(N^\delta_{\text{eq}}, P^\delta_{\text{eq}}\right) = \left((d + \delta)/c,\, (a - \delta)/b\right)\,, \tag{4.16}$$

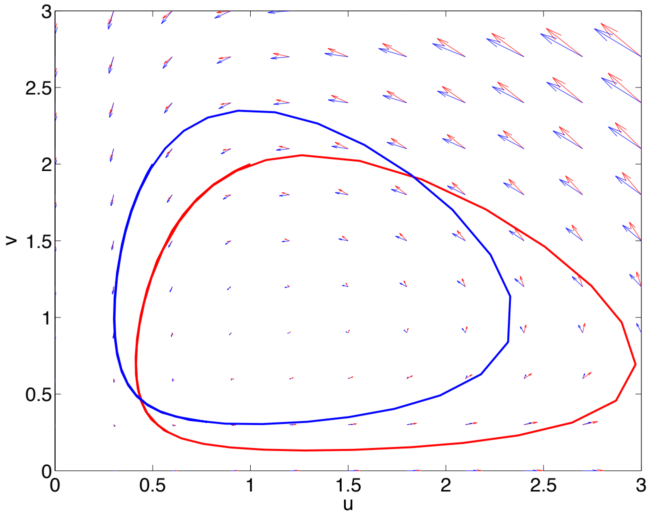


Figure 4.4: The effect of fishing for $\delta/a = \delta/d = 0.3$: typical solutions without (red) and with fishing (blue).

which reduced to the known equilibrium for $\delta = 0$, namely,

$$(u_{\text{eq}}, v_{\text{eq}}) = (1, 1) \quad \Leftrightarrow \quad \left(N^0_{\text{eq}}, P^0_{\text{eq}}\right) = (d/c,\, a/b) \tag{4.17}$$

We observe that the fishing shifts the equilibrium towards smaller values of predators, but larger values of its prey. In other words, the fishing pressure of the predator is higher than on the prey (because its food basis is reduced as well). According to Lemma 4.4 the average catch is equal to (cf. 4.8)

$$N^\delta_{\text{eq}} = \frac{1}{T}\int_0^T N^\delta(t)\,dt\,, \quad P^\delta_{\text{eq}} = \frac{1}{T}\int_0^T P^\delta(t)\,dt\,. \tag{4.18}$$

As a consequence the total average catch is given by

$$\delta\left(N^\delta_{\text{eq}} + P^\delta_{\text{eq}}\right) = \delta\left(\frac{d}{c} + \frac{a}{b}\right) + \delta\left(\frac{\delta}{c} - \frac{\delta}{b}\right) \tag{4.19}$$

Assuming that the effect of the fishing kicks in immediately whereas the equilibria represent long term properties of the ecosystem, the average catch after the recovery phase is smaller than it was before, if and only if

$$b > c\,, \tag{4.20}$$

which is the case when the reduction of the predator population due to fishing leads to a relatively higher survival rate of its prey.

4.2 Stability of fixed points

In the last section we have analysed the local behaviour on the nonlinear Lotka-Volterra model in the neighbourhood of its critical points based on a linearization about these points. As we have seen, the linear model shares some features of the nonlinear model, e.g., periodic oscillations around the critical point $(u_{\mathrm{eq}}, v_{\mathrm{eq}}) = (1, 1)$, or the exponential growth of the prey population close to the origin. The idea behind this kind of analysis is that the solution of the original nonlinear system and its linearization should behave similarly in a small neighbourhood of the fixed point. Under certain assumptions this idea can indeed be justified, and we will give precise statements below.[10]

Linearization about a critical point. We confine our attention to the case of two-dimensional systems. Consider the autonomous ODE

$$\frac{dx}{dt} = F(x) \tag{4.21}$$

where $F\colon \mathbb{R}^2 \to \mathbb{R}^2$ is any smooth (e.g. Lipschitz continuous and continuously differentiable) vector field with an isolated fixed point $x^* \in \mathbb{R}^2$ satisfying $F(x^*) = 0$. By Taylor's theorem, we can expand F about the point x^*:

$$F(x) = F(x^*) + \nabla F(x^*)(x - x^*) + o(\|x - x^*\|)\,, \tag{4.22}$$

with the matrix

$$\nabla F(x^*) = \begin{pmatrix} \frac{\partial F_1}{\partial x_1} & \frac{\partial F_1}{\partial x_2} \\ \frac{\partial F_2}{\partial x_1} & \frac{\partial F_2}{\partial x_2} \end{pmatrix}$$

denoting the 2×2 Jacobian of the function F, evaluated at $x = x^*$. We moreover used the Landau notation $o(\|x - x^*\|)$ to indicate that the remainder goes to zero faster than $\|x - x^*\|$ as $x \to x^*$, i.e.,

$$r(x) = o(\|x - x^*\|) \quad \Leftrightarrow \quad \lim_{x \to x^*} \frac{r(x)}{\|x - x^*\|} = 0\,.$$

Setting $y = x - x^*$ the linearization of (4.21) thus reads

$$\frac{dy}{dt} = Ay\,, \quad A = \nabla F(x^*)\,, \tag{4.23}$$

which is hopefully a good approximation to (4.21), whenever $\|y\|$ is small. We will specify what "good approximation" means later on.

Classification of equilibria. Consider now the initial value problem

$$\frac{dy}{dt} = Ay\,, \quad y(0) = y_0\,, \tag{4.24}$$

for a regular matrix $A \in \mathbb{R}^{2 \times 2}$. The solution of (4.23) is given by

$$y(t) = \exp(At)y_0\,, \tag{4.25}$$

where the exponential of a matrix (called *matrix exponential*) is defined by

$$\exp(B) = \sum_{k=0}^{\infty} \frac{B^k}{k!}\,. \tag{4.26}$$

[10] For the case of the critical point $(u_{\mathrm{eq}}, v_{\mathrm{eq}}) = (1, 1)$ this is not the case, unfortunately, even though the linearized system showed the same qualitative behaviour as the nonlinear model.

Note that this is in accordance with the usual exponential series

$$\exp(z) = \sum_{k=0}^{\infty} \frac{z^k}{k!}$$

for a real number $z \in \mathbb{R}$. Now suppose that A can be diagonalized by solving the corresponding eigenvalue problem

$$Av = \lambda v$$

for some $\lambda \in \mathbb{C}$. It is easy to see that if A has two distinct eigenvalues λ_1, λ_2, then the corresponding eigenvectors v_1, v_2 are linearly independent. To see this, assume the contrary: then there exists a $\alpha \in \mathbb{C}$, such that

$$v_1 - \alpha v_2 = 0\,.$$

But this implies

$$0 = A(v_1 - \alpha v_2) = \lambda_1 v_1 - \lambda_2 \alpha v_2 = (\lambda_1 - \lambda_2)v_1 \neq 0\,,$$

which proves that the assumption that v_1 and v_2 are linearly dependent must be wrong. Calling $V = (v_1, v_2)$ the 2×2 matrix that diagonalizes A, i.e. $V^{-1}AV = \Lambda$ with $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2)$. Then, by definition of the matrix exponential, we have

$$\exp(At) = V \exp(\Lambda t) V^{-1}\,, \tag{4.27}$$

where $\exp(\Lambda t)$ is a diagonal matrix with entries $\exp(\lambda_i t)$. Depending on whether the real part of the λ_i is positive negative or zero the exponential $e^{\lambda_i t}$ will either grow, decay or oscillate. As a consequence we can analyse the solution (4.24) in terms of the eigenvalues of the matrix A.

Depending on whether eigenvalues are real or complex, we can distinguish 5 cases as is illustrated in Figure 4.5; the case when the two eigenvalues $\lambda_{1,2} = \pm i\omega$ are pure imaginary, in which case the the solution of (4.3) is a linear combination of sines and cosines, is not extra mentioned.[11] Critical points with the property that the real part of $\lambda_{1,2} = \pm i\omega$ of the Jacobian matrix is zero are called *elliptic*, otherwise the critical point is called hyperbolic. The following theorem that is stated in an informal way guarantees that the linearization of a nonlinear ODE preserves the properties of hyperbolic equilibria.[12]

Theorem 4.5. *Let $dy/dt = Ay$ be the linearization of $dx/dt = F(x)$ at a critical point x^*. If no eigenvalue of $A = \nabla F(x^*)$ has real part zero then there exists a small neighbourhood of x^*, in which the flow map of the linearized system is topologically conjugate to the flow map of the nonlinear system.*

Informally the theorem states that the solution of a linearized ODE close to a hyperbolic equilibrium is basically a distorted, but otherwise qualitatively equivalent version of the exact solution. In particular, the stability of hyperbolic equilibria is preserved under linearization; for details see [Tes12, Sec. 9.3].

[11] This is a consequence of the famous Euler formula $e^{i\omega t} = \cos(\omega t) + i\sin(\omega t)$. Note that complex eigenvalues of a 2×2 matrix always come in conjugate pairs $\lambda_{1,2} = \alpha \pm i\omega$

[12] The theorem is due to the Russian mathematician David Grobman and the U.S. mathematician Philip Hartman who proved it independently.

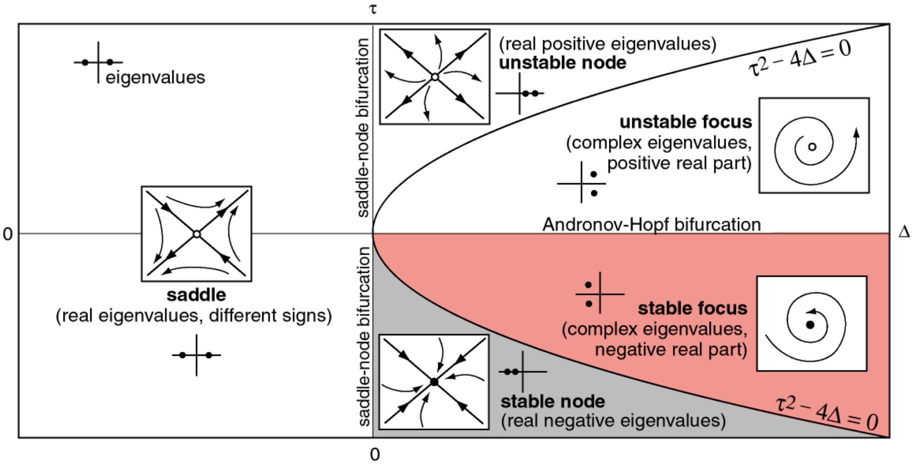

τ
eigenvalues
saddle-node bifurcation
(real positive eigenvalues)
unstable node
$\tau^2 - 4\Delta = 0$
unstable focus
(complex eigenvalues, positive real part)
0
Andronov-Hopf bifurcation
Δ
saddle
(real eigenvalues, different signs)
saddle-node bifurcation
stable focus
(complex eigenvalues, negative real part)
stable node
(real negative eigenvalues)
$\tau^2 - 4\Delta = 0$
0

Figure 4.5: Classification of equilibria of $dy/dt = Ay$ in terms of the eigenvalues of $A \in \mathbb{R}^{2\times 2}$. The eigenvalues are given by $\lambda_{1,2} = \tau/2 \pm \sqrt{\tau^2/4 - \Delta}$ where $\Delta = \det A$ and $\tau = \operatorname{trace} A$ (figure taken from: [Izh07]).

Stability analysis of the Lotka-Volterra model. As an example to illustrate the above we analyse the behaviour of the dimensionless Lotka-Volterra system close to the origin, i.e., we revisit (4.10). The Jacobian at $(0,0)$,

$$A = \begin{pmatrix} 1 & 0 \\ 0 & -\mu \end{pmatrix}, \tag{4.28}$$

has the two real eigenvalues $\lambda_1 = -\mu < 0$ and $\lambda_2 = 1$. Hence the origin is a saddle or *semistable equilibrium*. As a second example we consider the linearization (4.12) around the fixed point $(1,1)$; this equilibrium is *neutrally stable*, because solutions are periodic, which means that any solution in a sufficiently small neighbourhood of $(1,1)$ stays within this neighbourhood without approaching the fixed point. As a consequence the linearization cannot be informative, for otherwise the critical point would be hyperbolic (e.g. asymptotically stable) rather than elliptic. The Jacobian matrix at $(1,1)$ reads

$$A = \begin{pmatrix} 0 & -1 \\ \mu & 0 \end{pmatrix}, \tag{4.29}$$

and indeed the two eigenvalues of A are $\lambda_{1,2} = \pm i\sqrt{\mu}$. Note that even though the linearized system is again neutrally stable, Theorem 4.5 does not apply.[13]

[13] Further note note that elliptic and hyperbolic equilibria are not mutually exclusive and so the fact that both eigenvalues are pure imaginary does not imply that the nonlinear system has an elliptic fixed point too. For instance, it can happen that the nonlinear system has a crititcal point that is unstable in one direction and neutrally stable in the other direction, whereas the linearized system has an elliptic fixed point.

Problems

Exercise 4.6. *Show that (4.1)–(4.2) and (4.3) are equivalent under the substitutions*

$$\tau = at\,, \quad u = \frac{c}{d}N\,, \quad v = \frac{b}{a}P\,.$$

Exercise 4.7. *Consider the linear ODE system*

$$\begin{aligned}\frac{dx}{dt} &= y\,, \quad x(0) = x_0\\ \frac{dy}{dt} &= -x\,, \quad y(0) = y_0\,.\end{aligned}$$

a) Show that $I(x, y) = x^2 + y^2$ is a constant of motion.

b) Consider the forward Euler scheme

$$\begin{aligned}x_{n+1} &= x_n + hy_n\\ y_{n+1} &= y_n - hx_n\,, \quad n = 0, 1, 2, \ldots\end{aligned}$$

for sufficiently small step size $h > 0$ and show that $d_n(x_0, y_0) = x_n^2 + y_n^2$ is strictly increasing for all $(x_0, y_0) \in \mathbb{R}^2 \setminus \{(0, 0)\}$, with

$$\lim_{n\to\infty} d_n = \infty\,.$$

Exercise 4.8. *Prove (4.18).*

Exercise 4.9. *Consider the family of linear systems*

$$\begin{aligned}\frac{dx}{dt} &= y\,, \quad x(0) = x_0\\ \frac{dy}{dt} &= -x - \mu y\,, \quad y(0) = y_0\end{aligned}$$

for a scalar parameter $\mu \in \mathbb{R}$.

a) Characterize the stability of the unique fixed point $(x_{\rm eq}, y_{\rm eq}) = (0, 0)$ as a function of the parameter μ. Plot the eigenvalue(s) of the Jacobian matrix over the range $-3 \le \mu \le 3$ in 0.1 steps.

b) Solve the equation in Matlab over the time interval $I = [0, 10]$ with initial condition $(x_0, y_0) = (1, 0)$ and the parameter values

$$\mu \in \{-2, -1, 0, 1, 2\}\,,$$

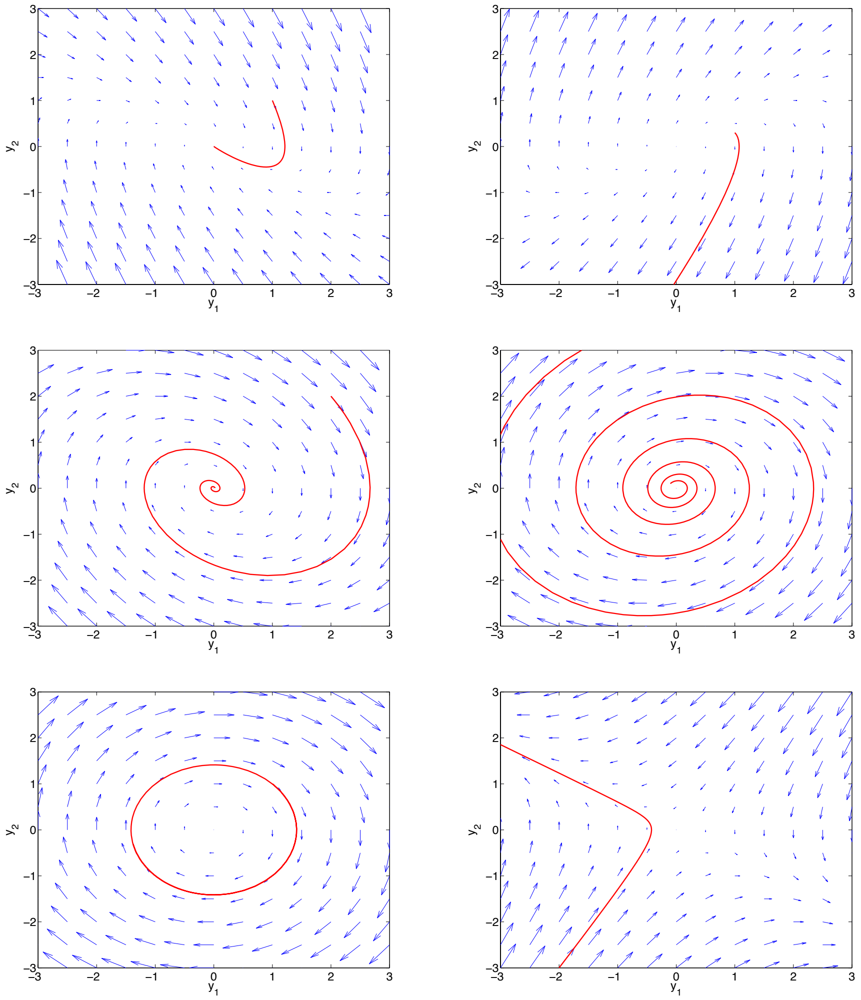
using the function `ode23t`*. Plot the solutions $(x(t; x_0, y_0), y(t; x_0, y_0))$ in the x-y-plane and explain your observation.*
*(**Hint:** Use the Matlab command* `help ode23t` *and modify the example accordingly. It is recommended to use function handles "*`@`*" to define the right hand side of the ODE; for details type* `doc function_handle`*.)*

Exercise 4.10. *Consider the following modification of the Lotka-Volterra model:*

$$\begin{aligned}\frac{dN^\epsilon}{dt} &= N^\epsilon(1 - P^\epsilon - \epsilon N^\epsilon)\,, \quad N^\epsilon(0) = N_0\\ \frac{dP^\epsilon}{dt} &= -P^\epsilon(1 - N^\epsilon)\,, \quad P^\epsilon(0) = P_0\,,\end{aligned}$$

where $\epsilon > 0$. Compute all fixed points and analyse their stability.

Exercise 4.11. *The following plots show linear 2-dimensional vector fields along with some typical solution trajectories (shown in red). Classify the stability of the associated fixed points according to the eigenvalues of the Jacobian:*



5 Basic principles of control theory

Mathematical tools & concepts: ODE, optimization
Suggested reference: [Whi96]

Recall the considerations about the effect of fishing on a population of two species on page 27. We will now modify the question a little bit and ask whether there is an optimal harvesting strategy that maximizes the sustainable catch or that maximizes the profit on a time-horizon of several years or tens of years.

5.1 Fishery management based on the logistic model

Our considerations will be based on the logistic population model for a single species [IBM+05]. In other words, we do not take into account the interaction between different species as in the previous section. Our model must be reasonably simple so that it is amenable to mathematical analysis, but still useful for the given purpose. To this end we introduce the functions

$$x(\cdot) \in \mathbb{R}\,, \quad b(\cdot) \in \mathbb{R}\,, \quad h(\cdot) \in \mathbb{R}\,, \tag{5.1}$$

where $x(t)$ denotes the fish population at time t, $b(t)$ the number of boats operating at time t and $h(t)$ the harvesting rate at time t. For simplicity, we assume that all functions can take real values, even though the number of boats will be an integer number. Our harvesting strategy will be based on controlling the number of boats that are used for the fishing; we call b the *control variable*, even though, strictly speaking, it is a (e.g. piecewise continuous) function $b\colon [0,\infty) \to \mathbb{R}$.

There are clearly other players in the game of finding an optimal harvesting strategy that come in form or parameters or boundary conditions, such as legal requirements, wages or overhead costs of maintaining a fishing fleet. Specifically, we consider the following parameters: $c_B > 0$ the overhead cost per boat and unit of time, n the number of fishermen per boat, w their salary per unit of time, p the market price of one unit of fish. The boundary conditions and available parameters determine what a good harvesting strategy is. For example, maximizing the sustainable catch is different from maximixing the long-term profit, which may be different from maximing the short-term profit.[14]

Constitutive relations, equations of motion and admissible controls It is time to set up the model. The first step consists in relating the harvest rate h with the number of fishes, x, and the number of boats, b. The static relation between these variables that is different from, say, the dynamic relation between different species in the predator-prey model is called a *constitutive relation*. Another example of a constitutive relation is Hooke's law is a kinematic relation between the force exerted by a spring and its extension, in contrast to Newton's law that expresses a dynamical relation between force and acceleration. Here we assume that the harvesting rate is proportional to both the number of fishes and and the number of boats, i.e., we assume the following relation

$$h(t) = qx(t)b(t)\,, \tag{5.2}$$

[14] "The answer depends on the question".

where $q > 0$ is a proportionality constant that depends on the efficacy of the fishing boats (e.g. the nets used etc.). The harvesting rate is the rate by which the growth rate of a fish population is reduced as an effect of fishing; we assume that the fish population evolves according to the logistic equation:

$$\frac{dx}{dt} = \gamma x \left(1 - \frac{x}{K}\right) - h \,, \quad x(0) = x_0 > 0 \tag{5.3}$$

where $\gamma > 0$ is the initial growth rate of the population when x is small and $K > 0$ is the capacity of the ecosystem without fishing (cf. p. 5). Maximizing any given objective, such as sustainable catch or profit under the constraint that the fish population evolves according to the dynamics (5.3) is not possible without further specifying what the admissible controls $b(\cdot)$ are. Here we assume that the only admissible strategies are of the form

$$b\colon [0,\infty) \to \mathbb{R} \,, \quad b(t) = \begin{cases} 0 & t \le t^* \\ b_0 & t > t^* \,, \end{cases} \tag{5.4}$$

with the two adjustable, but *a priori* unkown parameters $t^* \ge 0$ and $b_0 > 0$. As a consequence our harvesting strategy can be controlled by choosing the right time t^* at which fishing is started or resumed and the corresponding number b_0 of boats. The resulting logistic model then is a switched ODE of the form

$$\frac{dx}{dt} = \begin{cases} \gamma x \left(1 - x/K\right) & t \le t^* \\ \gamma x \left(1 - q b_0/\gamma - x/K\right) & t > t^* \,, \end{cases} \tag{5.5}$$

where we have used the constitutive relation (5.2) in the second equation.

Maximizing the sustainable catch Suppose we want to choose b_0 so that the average long-term catch is maximized. This requires that we do not overfish, for otherwise the fish population goes extinct and hence the average long-term catch is zero. For the average long term catch, it does not matter how t^* is chosen, so we can set it equal to zero and ignore it in what follows.

We identify the sustainable population under fishing with the asymptotically stable equilibrium of the system for $b_0 > 0$; asymptotic stability is essential for long-term catchability, because it the asymptotic stability guarantees that the equilibrium is robust under small perturbation, in other words, the population returns to its equilibrium size after a small perturbation that may be, e.g., due to fluctuating environmental conditions. If one is fishing at an unstable equilibrium instead the fluctuations may cause the population to drift away from its equilibrium and eventually go extinct.

Lemma 5.1. *Let $\gamma > q b_0$. Then*

$$x^* = \left(1 - \frac{q b_0}{\gamma}\right) K$$

is the unique stable equilibrium.

Proof. The proof is left as an exercise. □

Remark 5.2. *The assumption $\gamma > q b_0$ makes sure that the fish population, growing with rate γ when sufficiently far away from the capacity limit, is not eaten up by the fishing. For $\gamma = q b_0$ the single stable equilibrium is $x^* = 0$.*

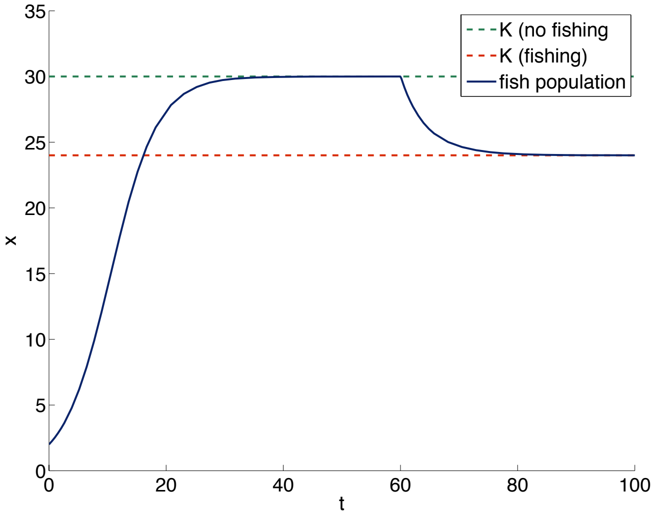


Figure 5.1: Solution of the logistic equation (5.5) with parameters $\gamma = 0.25$, $K = 30$, $q = 0.025$, $b_0 = 2$, $t^* = 60$ and initial value $x_0 = 2$.

Bear in mind that the solution to the logistic equation for $b_0 = 0$ satisfies

$$\lim_{t \to \infty} x(t; b_0 = 0) = K \,.$$

That is, the fishing reduces the capacity of the ecosystem by the factor $1 - qb_0/\gamma$. A representative solution of (5.4) is shown in Figure 5.1. We now define the average long-term catch as

$$J_0(b_0) = \lim_{T \to \infty} \frac{1}{T} \int_0^T h(t)dt \,, \tag{5.6}$$

where the expression for the associated sustainable catch rate follows from (5.2):

$$h(t) = qx^* b_0 \,. \tag{5.7}$$

Hence, with Lemma 5.1,

$$J_0(b_0) = qb_0 K \left(1 - \frac{qb_0}{\gamma}\right) \,. \tag{5.8}$$

The function $J_0(\cdot)$ is strictly concave, which implies that it has a unique maximum. The maximizer $b_0^* = \operatorname{argmax} J_0(b_0)$ is given by

$$b_0^* = \frac{\gamma}{2q} \,, \tag{5.9}$$

which—rounded to the nearest integer—determines the optimal fleet size. (Here in the example, $b_0^* = 5$). The corresponding optimal sustainable population is

$$x^* = \frac{K}{2} \,. \tag{5.10}$$

We observe that the maximum sustainable catch is independent of the efficacy q, which seems counterintuitive, but is understandable if we realize that b_0^* is inversely proportional to q, which makes the optimal harvesting rate independent of q. Rougly speaking, a lower efficacy requires to use more boats and vice versa: With too many boats the fish population is depleated too much, which results in a lower catch; the same happens when too few boats are at work, which conserves the fish population, but is suboptimal in terms of the catch.

5.2 Optimal control

The function J_0 is symmetric about its maximum, so if the optimal number of boats was, say, $b_0^* = 4.6$, the sustainable catch with $b_0 = 5$ boates would be slightly higher than with $b_0 = 4$. If, however, we take into account that fishing boats are costly, $b_0 = 4$ will probably be the more reasonable choice.

Objective functional: maximizing profit We will now seek to maximize profit rather than catch, which requires to take into account the costs of maintaining a fishing fleet, the market price of fish etc. To this end we define profit as revenue minus the total costs; accordingly the profit rate is the revenue rate minus the rate of the total costs. Using that revenue is the catch times the market price of fish, whereas the total cost is the sum of the overhead costs and the salaries of the fishermen. That is:

$$P(t) = ph(t) - (c_B + nw)b(t)\,. \tag{5.11}$$

The total profit until time $t = T$ is then obtained by integrating over the profit rate from 0 to T. To simplify matter we assume that $T = \infty$ and we discount the future profit with a discount rate $\delta > 0$. Together with the constitutive relation (5.2) the overall profit as a function of b turns out to be

$$J(b) = \int_0^\infty e^{-\delta t} b(t)\,(pqx(t) - c)\,dt\,. \tag{5.12}$$

with the shorthand $c = c_B + nw$. The discount factor δ accounts for inflation, interest rates or the fact that future rewards are less profitable than immediate rewards; it also ensures that J is finite for our choice of admissible controls $b(\cdot)$.

Extremum principle We want to maximize (5.12) over all admissible harvesting strategies, i.e. over the switching time t^* and the number of boats b_0. Since the population $x(t)$ depends on this choice, our optimal harvesting problems is of the form of a maximization probem with a constraint:

$$\max_{b(\cdot)} J(b) \tag{5.13}$$

over the set of admissible control strategies defined by (5.4) and subject to

$$\frac{dx}{dt} = \begin{cases} \gamma x\,(1 - x/K) & t \le t^* \\ \gamma x\,(1 - qb_0/\gamma - x/K) & t > t^*\,. \end{cases} \tag{5.14}$$

Generally, problems of the form (5.13)–(5.14) can be solved by the method of Lagrange multipliers or by eliminating the contraint; see [Whi96] for further

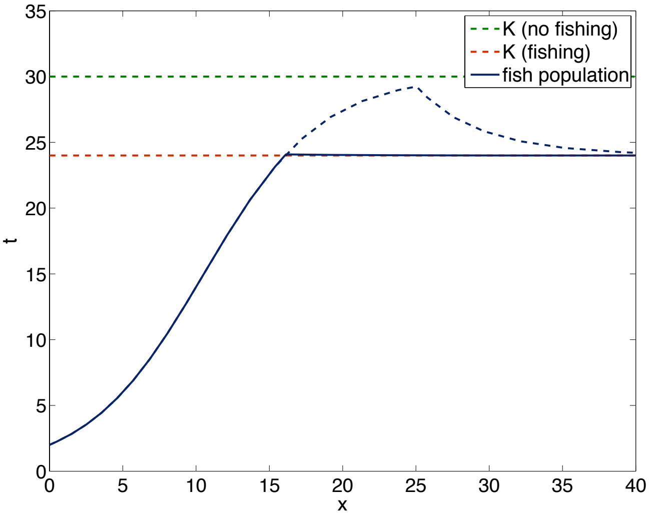


Figure 5.2: Solution of the switched logistic equation. the solution at the switching point t^* is continuous but non-differentiable, because the control variable has a jump discontinuity at t^* and jumps from $b(t^*) = 0$ to $b(t^* + \epsilon) = b_0$.

details and alternative methods for solving optimal control problems. Note that

$$\begin{aligned} J(b) &= \int_0^{t^*} e^{-\delta t} b(t)\,(pqx(t) - c)\,dt + \int_{t^*}^{\infty} e^{-\delta t} b(t)\,(pqx(t) - c)\,dt \\ &= \int_{t^*}^{\infty} e^{-\delta t} b_0\,(pqx(t) - c)\,dt\,. \end{aligned}$$

As a consequence we can solve (5.13)–(5.14) by first determining the optimal swiching time t^* which allows for solving (5.14) analytically and plugging the solution $x(t)$ into (5.13), which then eliminates the constraint and allows us to compute the optimal number of boats.

1. As a first step, we maximize over the switching time t^*. Clearly the optimal swiching time will depend on the initial value x_0: If x_0 is larger than the maximum capacity under fishing, then it pays off resume fishing from the very beginning; if, however, the initial fish population is below the capacity, then one should wait and resume fishing once the fish population has reached the fishable capacity; waiting longer to further increase the population does not pay of, in particular since future profits are discounted; see Figure 5.2. Let us assume that $x_0 < x^*$ and recall that

$$x(t) = \frac{K}{1 + (K/x_0 - 1)\exp(-\gamma t)}\,, \quad t \in [0, t^*] \tag{5.15}$$

is the solution to the logistic equation in the initial period without fishing. The optimal switching time is then determined by the condition

$$x(t^*) = x^*\,. \tag{5.16}$$

Solving the equation for t^* yields

$$t^* = \gamma^{-1}\left(\log\left(\frac{K}{x_0} - 1\right) + \log\left(\frac{\gamma}{qb_0} - 1\right)\right), \tag{5.17}$$

which determines the optimal switching time $t^* = t^*(b_0)$ as a function of the number of boats (via the capacity that is a function of b_0).

2. As a next step we eliminate the constraint from J, by noting that

$$x(t) = x^* \quad \forall t \geq t^* . \tag{5.18}$$

Hence

$$\begin{aligned} J(b) &= \int_{t^*}^{\infty} e^{-\delta t} b_0 \left(pqx(t) - c\right) dt \\ &= b_0 \int_{t^*(b_0)}^{\infty} e^{-\delta t}\left(pqK\left(1 - \frac{qb_0}{\gamma}\right) - c\right) dt \\ &= \frac{b_0}{\delta}\left(pqK\left(1 - \frac{qb_0}{\gamma}\right) - c\right) e^{-\delta t^*(b_0)} . \end{aligned} \tag{5.19}$$

The profit function is nonnegative when $pqK(1 - qb_0/\gamma) > c$, with c denoting the total cost per boat. Then for

$$0 \leq b_0 \leq \frac{\gamma}{q}\left(1 - \frac{c}{pqK}\right) \tag{5.20}$$

the function J is bounded from below by zero and has a unique maximum by Rolle's theorem. (Recall that the optimal fleet size for the maximum sustainable catch was $b_0^* = \gamma/(2q)$.) An example with the parameters

$$\gamma = 0.25,\ K = 30,\ q = 0.025,\ p = 10,\ \delta = 0.2,\ x_0 = 2\,, \tag{5.21}$$

is shown in Figure 5.3.

Problems

Exercise 5.3. *Prove Lemma 5.1.*

Exercise 5.4. *Consider the time-discrete logistic model with seasonal fishing*

$$\begin{aligned} x_{n+1}^* &= x_n + \tilde{\gamma} x_n \left(1 - \frac{x_n}{K}\right) \\ x_{n+1} &= \left(1 - \tilde{q} b_0\right) x_{n+1}^* \end{aligned}$$

that can be interpreted as the forward Euler discretization of (5.5) with $\tilde{\gamma} = \gamma \Delta t_1$ and $\tilde{q} = q\Delta t_2$. Compute the maximum sustainable average catch as a function of Δt_1 (recovery period) Δt_2 (fishing season) and b_0 (number of boats).

Exercise 5.5. *Compute the optimal fleet size for the parameters (5.19) to obtain*

a) the maximum sustainable catch,

b) the maximum long-term profit,

as described in the text. Plot the profit functions in both cases, compare the results and discuss the role of the discount parameter δ.

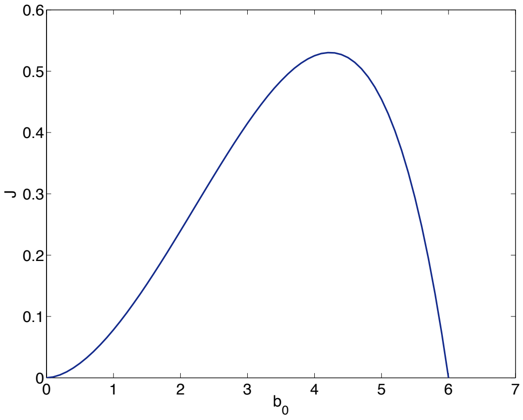


Figure 5.3: Profit as a function of the number of ships operating at time $t \geq t^*$.

6 Basic principles of bifurcation theory

Mathematical tools & concepts: ODE, bifurcations, hysteresis, attractors.
Suggested reference: [KaEn]

An important concept when studying ecological or climate systems is the concept of resilience. Resilience can be defined as the capacity of a system to recover from stress, i.e. to respond to a perturbation by resisting damage and recovering quickly. Disturbances of sufficient magnitude or duration can profoundly affect a system and may force it to reach a threshold, or a tipping point, and enter a different regime. In the framework of dynamical systems theory, the question is then wether abrupt changes in the dynamics will occur and if yes, at which values of parameters those changes will occur. An abrupt change in model dynamics with a (slight) change in parameters is called a bifurcation. The change of a steady state from a stable node into a saddle is an example of such a bifurcation. Bifurcation theory is the part of theory about dynamical systems, that deals with classifying, ordering and studying the regularity in these changes. A famous example is the Earth's radiation budget or global energy balance model (EBM). A global EBM summarises the state of the Earth's climate in a single variable, namely the temperature overaged over the entire globe.

In 1964, Brian Harland at Cambridge University postulated that the Earth had experienced an ice age with global scale glaciation. He pointed out that sedimental glacial deposits, similar in type to those found in Svalbard or Greenland, are widely distributed on virtually every continent. Climate physicists were just developing mathematical models of the Earth's climate, providing a new perspective on the limits to glaciation.

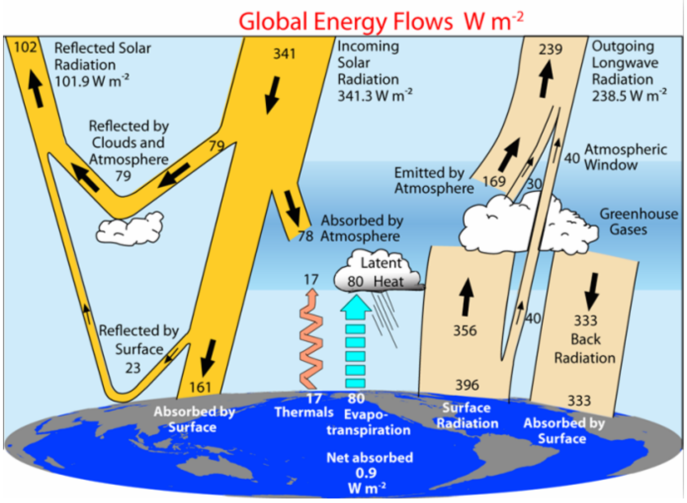


Figure 6.1: Detailed radiative energy balance (source: Trenberth et al., Bulletin of the American Meteorological Society, 2009).

6.1 Solar radiation

The Earth's climate is fundamentally controlled by the way that solar radiation interacts with the Earth's surface and atmosphere. We receive around 343 watts per square meter of radiation from the Sun. Some of this is reflected back to space by clouds and by the Earth's surface, but approximately two thirds is absorbed by the Earth's surface and atmosphere, increasing the average temperature. Earth's surface emits radiation at longer wavelengths (infrared), balancing the energy of the radiation that has been absorbed. If more of the solar radiation were reflected back to space, then less radiation would be absorbed at the surface and the Earth's temperature would decrease. The surface albedo is a measure of how much radiation is reflected; snow has a high albedo (~ 0.8), seawater has a low albedo (~ 0.1), and land surfaces have intermediate values that vary widely depending mainly on the types and distribution of vegetation. When snow falls on land or ice forms at sea, the increase in the albedo causes greater cooling, stabilizing the snow and ice. This is called ice-albedo feedback, and it is an important factor in the waxing (and waning) of ice sheets.

6.2 Energy balance models

A simple observation is that the global average temperature at the Earth's surface increases if the amount of energy reaching the Earth from the Sun exceeds the amount of energy emitted by the Earth and released into the stratosphere, and decreases if the converse is true. To consider the simplest scenario, we think of the Earth as a solid sphere and ignore all spatial differences to characterize the state of the system by a single variable, namely the global mean surface temperature T. We are interested in the evolution of T over time t. We merge all components that can exchange heat with outer space into one single element to obtain the *heat capacity* C of the system. This is the energy needed to raise the temperature $T(t)$ by one kelvin (the heat capacity varies a lot between land, water, etc, but we consider again the global average value C). The energy needed to increase T by an amount ΔT after a time Δt, i.e. $T(t+\Delta t) = T(t) + \Delta T$ is thus $AC\Delta T$, where A is the surface area of the planet.

Now let E_{in} be the average amount of solar energy reaching one square meter of the Earth's surface per unit time, and E_{out} be the average amount of energy emitted by one square meter of the Earth's surface and released into the stratosphere per unit time. Then we have

$$AC\Delta T = A(E_{in} - E_{out})\Delta t\,.$$

Letting Δt tend to zero, we obtain the global energy balance model describing the evolution of T,

$$C\frac{dT}{dt} = E_{in} - E_{out}\,. \tag{6.1}$$

We assume that no forcing modifies the solar energy or radiative properties; E_{in} and E_{out} thus do not depend explicitly on time (but they depend on T). If the incoming energy balances the outgoing energy, the Earth's temperature remains constant and the planet is said to be in thermal equilibrium. To specify E_{in} and E_{out}, we consider the following:

- Viewed from the Sun, the Earth is a disk of area πR^2, where R is the radius of the Earth.

- The energy flux density is S_0.

- The amount of energy flowing through the disk (i.e. reaching the Earth) is $E_{in} = (1-\alpha)\pi R^2 S_0$, where α is the average albedo of the Earth.

- All bodies radiate energy in the form of electromagnetic radiation. The amount of energy radiated (black body radiation) depends on temperature according to the Stefan-Boltzmann law,

  $$F_{SB}(T) = \sigma T^4.$$

  Here σ is Stefan's constant, $\sigma = 5.67.10^{-8} \cdot \mathrm{Wm}^{-2}\mathrm{K}^{-4}$

- The amount of energy radiated out by the Earth is distributed uniformly across the globe (area $4\pi R^2$), such that

  $$E_{out}(T) = 4\pi R^2 \sigma T^4.$$

With these expressions, the differential equation 6.1 becomes

$$C\frac{dT}{dt} = (1-\alpha)Q - \sigma T^4\,, \tag{6.2}$$

where we have used $Q = \frac{1}{4}S_0$.

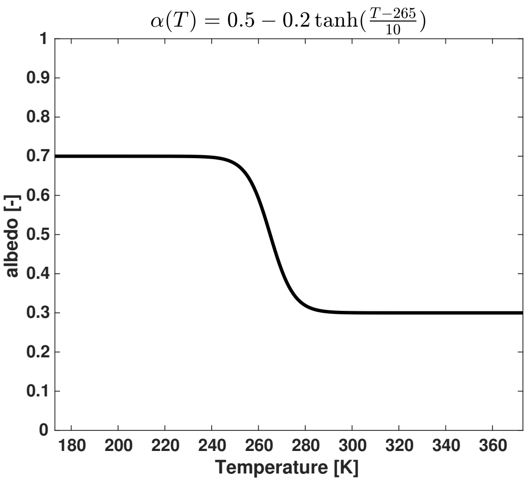


Figure 6.2: Albedo dependance on temperature.

Greenhouse effect We can now calculate the steady state temperature

$$T^* = \left(\frac{(1-\alpha)Q}{\sigma}\right)^{1/4} .$$

Considering a typical albedo value of $\alpha = 0.3$, the solar energy flux density $S_0 = 1368\text{Wm}^{-2}$ (thus $Q = 342\text{Wm}^{-2}$), the equilibrium temperature is $T^* = 254.8\text{K}$. However, the actual value of the surface average temperature is $T^* = 287.7\text{K}$. The difference is largely explained by the *greenhouse effect* of the Earth's atmosphere, that is, the effect of gases like CO_2, water vapor and methane. The greenhouse gases only affect the infrared (long wavelengths) part of the energy spectrum, and thus only the energy radiated by the Earth. We include this effect through a factor $0 < \epsilon < 1$ which reduces the outgoing energy, leading to

$$C\frac{dT}{dt} = (1-\alpha)Q - \epsilon\sigma T^4 , \tag{6.3}$$

Albedo dependence on temperature The albedo depends on the amount of ice and snow cover and therefore on the temperature, $\alpha = \alpha(T)$. The energy balance equation should therefore be written as

$$C\frac{dT}{dt} = (1-\alpha(T))Q - \epsilon\sigma T^4 , \tag{6.4}$$

It is reasonable to assume that it has a low value at high temperature (no ice), a high value at low temperature (Earth completely frozen), and some continuous variation in between (Fig. 6.2):

$$\alpha(T) = 0.5 - 0.2 \cdot \tanh\left(\frac{T-265}{10}\right) .$$

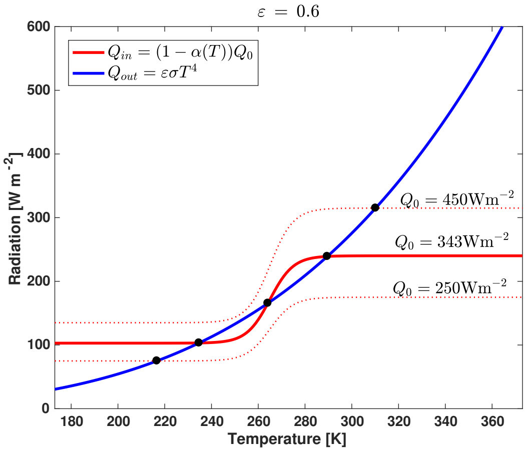


Figure 6.3: Incoming and outgoing energy as a function of globally averaged temperature.

Qualitative solution Steady state or fixed points are obtained when the left hand side of 6.4 vanishes. There are three equilibria. Two of them are stable and one is unstable. The leftmost is a snowball solution, the rightmost is a ice-free solution.

Bifurcation and hysteresis As the strength of the solar input S_0 or the emissivity ϵ is changed, the two curves of Figure 6.3 move with respect to each other. As a result, the number of crossing points, and thus of fixed points, changes. If the greenhouse effect increases, the emissivity becomes lower and the outgoing energy curve moves down. The same happens if the solar input becomes larger. The two lower equilibria representing cold and moderate climates disappear, leaving only the warmer climate solution. In the opposite scenario where the outgoing energy curve moves up, only the cold (snowball) climate is possible. The number and character of the solutions as function of such a *bifurcation* parameter may be represented schematically in a bifurcation diagram as in 6.4. Notice that the bifurcation diagram is shown using the *dimensionless* bifurcation parameter Q/Q_0 (Why?). It is common practice to represent the stable solutions by solid curves and unstable solutions by dashed curves. The bifurcation diagram shows that, with decreasing solar input, the mean temperature decreases until it reaches a *tipping point.* The climate system then transits to the lower branch, the planet turns frozen, and the temperature equilibrates at its frozen stable equilibrium. A reverse scenario is also possible, however the transition from frozen to unfrozen equilibrium occurs at a different tipping point. Since the paths for increasing and decreasing value of the bifurcation parameter are distinct, we see that *hysteresis* occurs in our climate model.

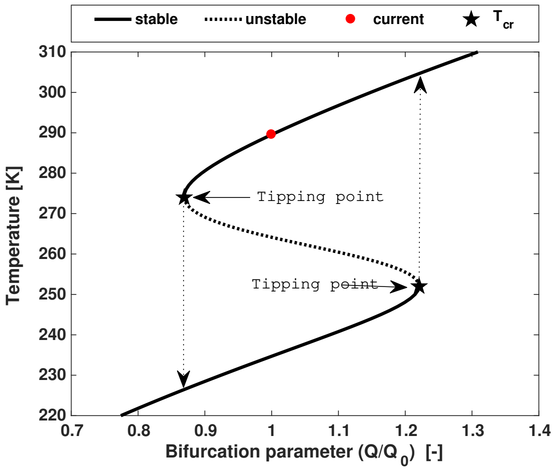


Figure 6.4: Mean surface temperature at equilibrium as a function of the solar constant, normalized by its present value.

6.3 Bifurcation theory

In general, the notion of a bifurcation refers to a qualitative change in the behaviour of a dynamical system as some parameter on which the system depends varies continuously,

$$\dot{x} = f(\lambda, x) \tag{6.5}$$

where f is a smoth function which depends on x and on one or more real parameters $\lambda_1, \dots, \lambda_m$. The qualitative dynamical behaviour of a dynamical system is determined by its equilibria and their stability, so all bifurcations are associated with bifurcations of equilibria. A possible definition is

Definition 6.1. *A point (x^*, λ_0) is a bifurcation point of equilibria for 6.5 if the number of solutions of the equation $f(\lambda, x) = 0$ for x in every neighbourhood of (x^*, λ_0) is not a constant independant of λ.*

In the following we will present some typical examples of bifurcations. In those examples, the vector field is a simple, low order (quadratic or cubic) polynomial function with one or two real parameters. These examples are quite generic despite their simplicity, in the sense that often, when the nonlinearity of the vector field is more complicated, it can be approximately described by one of the following examples. The higher the order of the polynomial, the more complicated the bifurcation gets, but also the less likely.

Transcritical bifurcation Consider the ODE

$$\dot{x} = \lambda x - x^2. \tag{6.6}$$

This has two equilibria at $x = 0$ and $x = \lambda$, which coincide if $\lambda = 0$. For $f(\lambda, x) = \lambda x - x^2$, we have

$$\frac{\partial f}{\partial x}(\lambda; x) = \lambda - 2x, \quad \frac{\partial f}{\partial x}(\lambda; 0) = \lambda, \quad \frac{\partial f}{\partial x}(\lambda; \lambda) = -\lambda.$$

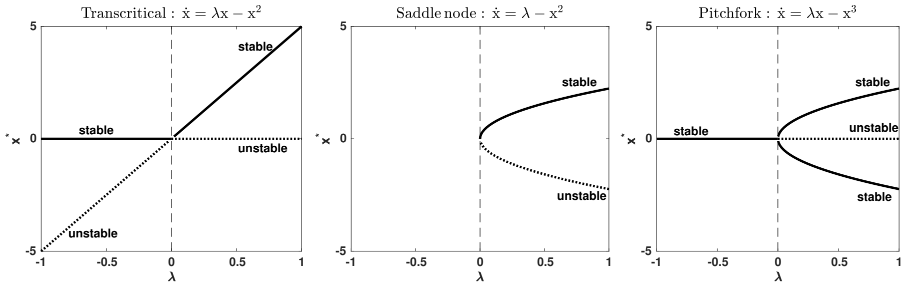


Figure 6.5: Bifurcation diagram for example bifurcations

Thus, the equilibrium $x = 0$ is stable for $\lambda < 0$ and unstable for $\lambda > 0$, while the equilibrium $x = \lambda$ is unstable for $\lambda < 0$ and stable for $\lambda > 0$. This transcritical bifurcation arises in systems where there is some "trivial" solution branch (here corresponding to $x = 0$), which exists for all values of the parameter λ. There is a second branch $x = \lambda$ that crosses the first one at the bifurcation point $(x, \lambda) = (0, 0)$. When the branches cross one solution goes from stable to unstable while the other goes from unstable to stable (Fig. 6.5).

Saddle-node bifurcation Consider the ODE

$$\dot{x} = \lambda - x^2. \tag{6.7}$$

If $\lambda < 0$, the equation has no equilibrium solution. If $\lambda > 0$, it has two equilibrium solutions $x = \pm\sqrt{\lambda}$, one stable and one unstable. These two solutions coincide if $\lambda = 0$, and the single trivial solution is unstable. This bifurcation is called a saddle-node bifurcation. A pair of hyperbolic equilibria, one stable and one unstable, emerge out of nowhere (Fig. 6.5).

Hysteresis Consider the ODE

$$\dot{x} = \lambda + x - x^3. \tag{6.8}$$

The equation $f(\lambda, x) = 0$ has one solution for $\lambda < -\lambda^*$, three solutions for $-\lambda^* < \lambda < \lambda^*$, and one solution for $\lambda > \lambda^*$. For $\lambda = \pm\lambda^*$, there are two solutions.

The dynamics of 6.8 have a particularity: suppose we have the capability to continuously vary the value of the parameter λ, for example by changing some of the physics (e.g. emitting CO_2 in the atmosphere, releasing large amounts of freshwater into the ocean by melting the polar ice cap, etc.). If we start with a large negative value of λ, the system will eventually reach an equilibrium on the lower branch of the bifurcation diagram. As we keep increasing λ, the system will "slide" along the stable branch of the bifurcation diagram until it exceeds the value λ^*, where it will transition quickly to the upper stable branch of the bifurcation diagram (why quickly?). As we further increase λ, the system will now slide along the upper branch. The reverse change from the upper to the lower branch will however necessitate a large decrease of λ, since it will occur at the value $-\lambda^*$. That is, the parameter value at which the transition occurs depends on the direction in which the parameter is varied. This phenomenon is called a hysteresis.

Pitchfork bifurcation Consider the ODE

$$\dot{x} = \lambda x - x^3. \tag{6.9}$$

For $\lambda \leq 0$, the system has one stable equilibrium point at $x = 0$. At $\lambda = 0$, a bifurcation occurs. For $\lambda > 0$, the system has three equilibrium states at $x = 0$ (unstable) and $x = \pm\sqrt{\lambda}$ (stable). Thus the stable equilibrium 0 looses stability at the bifurcation point, and two new stable equilibria appear. The pitchfork shape bifurcation diagram gives its name (Fig. 6.5). The pitchfork bifurcation in which a stable solution bifurcates into two new stable solutions is called a supercritical pitchfork bifurcation. Up to change in the signs of x and λ, the other possibility is the subcritical pitchfork bifurcation, described by

$$\dot{x} = \lambda x + x^3. \tag{6.10}$$

In this case, we have three equilibria $x = 0$ (stable) and $x = \pm\sqrt{-\lambda}$ (unstable) for $\lambda < 0$, and one unstable equilibrium $x = 0$ for $\lambda > 0$. A supercritical pitchfork bifurcation leads to a "soft" loss of stability, in which the system can go to nearby stable equilibria $x = \pm\sqrt{\lambda}$ when the equillibrium at $x = 0$ looses stability as λ passes through 0. On the other hand, a subcritical pitchfork bifurcation leads to a "hard" loss of stability, in which there are no nearby equilibria and the system goes to some far-off dynamics (perhaps to infinity) when the equilibrium at $x = 0$ looses stability.

Finding bifurcation points A solution branch is a set of equilibrium points x^* which can be written as a smooth function of the bifurcation parameter λ, i.e. $f(\lambda, x^*) = 0$. The equilibrium points will depend smoothly on λ as long as

$$\frac{\partial f}{\partial x}(\lambda, x^*) \neq 0.$$

This is a consequence of the implicit function theorem. Therefore, solution branches are expected to meet at points where $f(\lambda, x^*) = 0$ and $\frac{\partial f}{\partial x}(\lambda, x^*) = 0$. These are candidates for bifurcation points.

7 Modelling of chemical reactions

Mathematical tools & concepts: conditional probabilities, ODE
Suggested reference: [Hig08]

The main theme of this section will be the stochastic modelling of chemical reactions. Nonetheless the reader may replace *chemical reaction* by *evolutionary game* or the alike. To begin with, we mention two prototypical examples:

Enzyme kinetics Enzyme-catalysed reactions with single-substrate mechanisms due to Michaelis and Menten are systematically written as

$$S + E \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} SE \xrightarrow{k_2} E + P$$

where it is assumed that the back reaction $S + P \to ES$ is negligible. If the concentration of the substrate S is high, the enzyme E is entirely saturated and only exists in its complex form ES. This entails that, after a short relaxation time depending upon the initial conditions, the concentrations of both the enzyme and the complex quickly converge to a steady-state.

Depletion of ozone in the stratosphere An important environmental process is the catalytic destruction of ozone by atomic halogens, the main source of which is photodissociation of halocarbon refrigerants and foam-blowing agents, such as CFCs or freons. One such example is the cyclic reaction

$$Cl + O_3 \xrightarrow{k_1} ClO + O_2$$

and

$$ClO + O_3 \xrightarrow{k_2} Cl + 2O_2\,.$$

The second reaction recreates the original chlorine atom, which can repeat the first reaction and continue to destroy ozone (i.e., chlorine acts as a catalyst).

7.1 The chemical master equation

Generally, we consider a situation with N different molecular species $A, B, C, \ldots$ that can interact via M different reactions, such as $A + B \to C$. We call

$$X(t) = (X_1(t), \ldots, X_N(t)) \in \mathbb{N}_0^N \tag{7.1}$$

the *state vector*, with $X_i(t) \in \{0, 1, 2, 3, \ldots\}$ being the number of molecules at time $t \geq 0$. If any of the M reactions fires at time t, say, the j-th reaction, the state vector is updated according to the rule

$$X(t) \mapsto X(t) + \nu_j\,. \tag{7.2}$$

The vector $\nu_j \in \mathbb{Z}^N$ is called the *stoichiometric vector* of the j-th reaction.

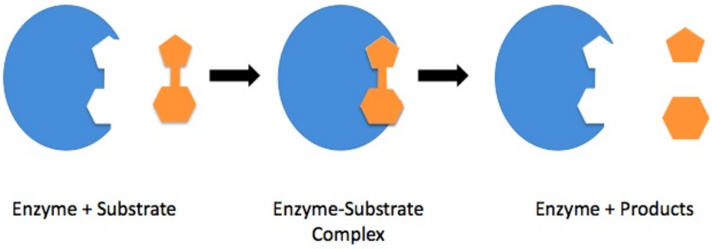


Figure 7.1: Chemical reaction catalysed by an enzyme.

Probability of a reaction We assume that the system of chemical species is spatially homogeneous ("well-stirred"), so the chemical species abundances do not vary in space. When two molecules collide they can react with a certain probability, depending on the boundary conditions like temperature or pressure; by the above homogeneity assumption the probability that the reageants collide is linear in the number of each molecular species involved in a particular reaction (cf. Exercise 7.9). Specifically, we define the probability that the j-th reaction fires over the infinitesimal time interval $[t, t + dt)$ given that $X(t) = x$ as

$$P(j\text{-th reaction fires over } [t, t + dt) \,|\, X(t) = x) = a_j(x)dt\,. \tag{7.3}$$

The function $a_j : \mathbb{N}_0^N \to \mathbb{R}_+$ is called the propensity of the reaction. The exact functional form of a_j depends on the type o the reaction.

Example 7.1. *As an example that will guide us through this section consider three species A, B, C with the single binary reaction*

$$A + B \xrightarrow{c} C\,.$$

Since the reaction turns one A and one B into one C, the stoichiometric vector is $\nu_1 = (-1, -1, 1)$. Now suppose that initially we have an initial mixture consisting of four molecules of type A, three molecules of type B and zero C molecules, i.e. $X(0) = (4, 3, 0)$. Then, since the total number of particles is finite, the set of possible states at time $t > 0$ is

$$S = \{(4, 3, 0),\, (3, 2, 1),\, (2, 1, 2),\, (1, 0, 3)\}\,.$$

Note that the state $X(t) = (1, 0, 3)$ is a fixed point, also called absorbing state, *since all the B molecules are eaten up and no further reaction can happen. The propensity of the reaction results from the consideration that, in a well-stirred system, the probability of a reaction happening per unit of time must be proportional to the number of both A and B molecules, which implies that*

$$a_1(x_1, x_2, x_3) = cx_1x_2$$

It is unrealistic to assume that the number of molecules in a test tube can be counted. Hence we seek a more coarse grained description of the number of molecules at time t. The idea is to derive an differential equation for the *probability* to have $x = (x_1, \ldots, x_N)$ molecules at time t,

$$\rho(x,t) = P(X(t) = x)\,, \tag{7.4}$$

given that we know the probability distribution of states at time $t = 0$. (Note that this entails the situation that $X(0)$ is known exactly.)

Interlude: Basic probability theory.

Definition 7.2 (probability space)**.** *A probability space $(\Omega, \mathcal{E}, P)$ consists of*

- *A non-empty set Ω. This is the sample space or the space of possible outcomes.*
- *A σ-algebra $\mathcal{E} \subset 2^\Omega$. $\mathcal{E}$ is a set of subsets of Ω and models the 'event space' (i.e. the things we can assign probabilities to).*
- *A probability measure $P : \mathcal{E} \to [0,1]$ which satisfies $P(\Omega) = 1$ and the so-called σ-additivity property:*

  $$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

  if $A_i \in \mathcal{E}$ are pairwise disjoint.

Example 7.3. *The probability space $\Omega, \mathcal{E}, P)$ that models a fair six-sided die consists of*

- $\Omega = \{1, 2, 3, 4, 5, 6\}$
- $\mathcal{E} = 2^\Omega = \{\emptyset, \{1\}, \{2\}, \ldots, \{1,2\}, \{1,3\}, \ldots, \Omega\}$
- $P(\{1\}) = \ldots = P(\{6\}) = \frac{1}{6}$.

Probabilities of other events can be computed using the σ-additivity property, for example $P\{1,2\}) = P(\{1\}) + P(\{2\}) = \frac{1}{3}$.

We need two more definitions:

Definition 7.4 (conditional probability)**.** *Let $(\Omega, \mathcal{E}, P)$ be a probability space and $A, B \in \mathcal{E}$ with $P(B) > 0$. The conditional probability of A given B is defined as*

$$P(A|B) = \frac{P(A \cup B)}{P(B)}.$$

Definition 7.5 (partitions)**.** *The collection $\{B_0, \ldots, B_M\}$ is called a partition of Ω if $B_i \subset \Omega$ and $B_i \cap B_j = \emptyset$ for al $i \neq j = 0, \ldots, M$ and $B_0 \cup B_1 \cup \ldots \cup B_M = \Omega$.*

Now we are ready to state and prove a Lemma that we will need later to derive the chemical master equation:

Lemma 7.6 (Law of total probability)**.** *Let $(\Omega, \mathcal{E}, P)$ be a probability space and let $\{B_0, B_1, \ldots, B_M\} \in \mathcal{E}$ be a partition of Ω such that $P(B_j) > 0$ for all $j = 0, \ldots, M$ Then, for any $A \in \mathcal{E}$, it holds*

$$P(A) = \sum_{j=0}^{M} P(A|B_j)P(B_j)\,.$$

Proof. Since the $\{B_j\}_{j=0,\ldots,M}$ are a partition of Ω, we can write $A \subset \Omega$ as

$$A = A \cap \left(\bigcup_{j=0}^{M} B_j \right) = \bigcup_{j=0}^{M} (A \cap B_j) \,,$$

where we have used de Morgan's rule in the second equality. Since any probability measure P is countably additive (σ-additive) and all the $A \cap B_j$ are disjoint, we have

$$\begin{aligned} P(A) &= P\left(\bigcup_{j=0}^{M} (A \cap B_j) \right) \\ &= \sum_{j=0}^{M} P(A \cap B_j) \\ &= \sum_{j=0}^{M} P(A|B_j)\, P(B_j) \,. \end{aligned}$$

The last equality is a direct consequence of the definition $P(A|B) = P(A \cap B)/P(B)$ of conditional probabilities. □

Recurrence equation for the state probability We will now derive an equation for the $\rho(x, t + dt)$ of the molecular state vector $X(t + dt)$ at time $t + dt$, assuming that we know the distribution $\rho(x, t)$ of $X(t)$ at time t. To compute $\rho(x, t + dt)$ from $\rho(x, t)$, it is helpful to realize that, in a chemical system with M possible reactions there are exactly $M + 1$ different scenarios that can lead to the situation $X(t + dt) = x$:

- $X(t) = x$ and no reaction fired over $[t, t + dt)$,
- $X(t) = x - \nu_j$ for any $j = 1, \ldots, M$ and the j-th reaction fired.

Here we assume that dt is sufficiently small (in fact: infinitesimally small), so that at most one reaction can occur between t and $t + dt$.

Now let A above the event $\{X(t + dt) = x\}$, so that $P(A) = \rho(x, t + dt)$ is exactly the probability that we want to compute.[15] Then, with $B_j = \{X(t) = x - \nu_j\}$ the conditional probability $P(A|B_j)$ is exactly the probability that the j-th rection fires over $[t, t + dt)$, given that $X(t)$ is $x - \nu_j$ at time t. Moreover, $P(A|B_0)$ is one minus the probability that any of the M reaction fires over $[t, t + dt)$ given that $X(t) = x$. By definition of the propensities, we thus have

- $P(A|B_0) = 1 - \sum_{j=1}^{M} a_j(x) dt$,
- $P(A|B_j) = a_j(x - \nu_j) dt$

Furthing noting that $P(B_j) = \rho(x - \nu_j, t)$ for $j = 1, \ldots, M$ the law of total probability, Lemma 7.6, implies that

$$\rho(x, t + dt) = \left(1 - \sum_{j=1}^{M} a_j(x) dt \right) \rho(x, t) + \sum_{j=1}^{M} a_j(x - \nu_j) dt\, \rho(x - \nu_j, t) \,. \quad (7.5)$$

[15] More precisely, we define $A = \{\omega \in \Omega \colon X(t + dt, \omega) = x\} \subset \Omega$ where $X(t, \cdot) \colon \Omega \to \mathbb{N}_0^N$ is a stochastic process with sample paths (realizations) $X(\cdot) = \{X(t, \omega) \colon t \geq 0\}$.

Rearranging the terms and dividing by dt yields

$$\frac{\rho(x,t+dt)-\rho(x,t)}{dt} = \sum_{j=1}^{M} a_j(x-\nu_j)\rho(x-\nu_j,t) - \sum_{j=1}^{M} a_j(x)\rho(x,t)\,. \tag{7.6}$$

Chemical master equation We take advantage of the fact that the right hand side of (7.6) is independent of dt and that the expression on the left is a finite-difference approximation of the partial derivative with respect to t: Letting $dt \to 0$, we obtain the chemical master equation (CME)[16]

$$\frac{\partial}{\partial t}\rho(x,t) = \sum_{j=1}^{M} a_j(x-\nu_j)\rho(x-\nu_j,t) - \sum_{j=1}^{M} a_j(x)\rho(x,t)\,. \tag{7.7}$$

The CME is a discrete linear partial differential equation on a countable spatial domain $S \subset \mathbb{N}_0^N$, which size depends on whether the number of molecules is finite or not (cf. Example 7.1). If we introduce the (possibly infinite) vector $u(t) = \rho(\cdot,t)$ with entries $u = (u_x)_{x\in\mathbb{N}_0^N}$ the CME is equivalent to a linear system of ordinary differential equations of the form

$$\frac{du}{dt} = Au(t)\,, \quad u(0) = \rho(x,0)\,, \tag{7.8}$$

where entries of the square matrix A are the propensities (cf. Exercise 7.10). If the initial value $X(0) = x_0$ is known then $\rho(x,0) = \delta(x-x_0)$ and the CME yields the distribution of $X(t)$ for any $t > 0$.

Example 7.7 (Example 7.1, cont'd)**.** *Recalling that $a(x) = cx_1x_2$ and $\nu = (-1,-1,1)$, the CME is readily found to be*

$$\begin{aligned}\frac{\partial}{\partial t}\rho(x_1,x_2,x_3,t) &= c(x_1+1)(x_2+1)\rho(x_1+1,x_2+1,x_3-1,t)\\ &\quad - cx_1x_2\rho(x_1,x_2,x_3,t)\,.\end{aligned}$$

To solve it, it must be endowed with the initial condition

$$\rho(x_1,x_2,x_3,0) = \delta(x_1-4)\delta(x_2-3)\delta(x_3)\,,$$

with (x_1,x_2,x_3) being any state from the state space

$$S = \{(4,3,0),\ (3,2,1),\ (2,1,2),\ (1,0,3)\}\,.$$

To see that the CME is indeed equivalent to a linear ODE system, let us define the vector $u = (u_1,\ldots,u_4)$ with $0 \le u_i \le 1$ given by

$$u_1(t) = \rho(4,3,0,t)\,, \quad u_2 = \rho(3,2,1,t)\,, \quad u_3 = \ldots\,.$$

[16] The reader may be surprised that the right hand side of (7.6) does not depend on dt at all. The reason is that the propensities were only defined in terms of the infinitesimal time increment dt, i.e., the right hand side of (7.5) is already a linearization in dt, which is why after dividing by the time increment it becomes constant (i.e. independent of dt).

In terms of the new state vector u the CME can be recast as

$$\begin{aligned}\dot{u}_1 &= -12cu_1\\ \dot{u}_2 &= 12cu_1 - 6cu_2\\ \dot{u}_3 &= 6cu_2 - 2cu_3\\ \dot{u}_4 &= 2cu_3\end{aligned}$$

where the dot denotes the derivative with respect to t. In other words, we have rewritten the CME as the linear system of equations

$$u'(t) = Au(t)\,, \quad A = \begin{pmatrix} -12c & 0 & 0 & 0\\ 12c & -6c & 0 & 0\\ 0 & 6c & -2c & 0\\ 0 & 0 & 2c & 0\end{pmatrix}$$

with real eigenvalues $\lambda \in \{0, -2c, -6c, -12c\}$. The simple eigenvalue $\lambda_1 = 0$ corresponds to the asymptotically stable equilibrium point of the CME, which is the stationary probability of the absorbing state $x^* = (1, 0, 3)$, i.e.,

$$\lim_{t\to\infty} u(t) = (0,0,0,1) \quad \Longleftrightarrow \quad \lim_{t\to\infty} \rho(x^*, t) = 1\,.$$

7.2 Stochastic simulation algorithm

We have shown that the CME can be solved in pretty much the same way as one would solve a linear ODE. Note, however, that the ODE system thus obtained can be fairly large, possibly even infinitely large.

We will now present an alternative method to simulate the CME in terms of the continuous-time Markov jump process $(X(t))_{t\geq 0}$ that is furnished by the CME. In fact, the matrix A that is obtained by rewriting the CME as an equivalent ODE system is the transpose of the generator matrix of this Markov jump process. This property entails that the sum over all row-entries of A^T is zero, which implies that A has at least one eigenvalue $\lambda = 0$ corresponding to an equilibrium of the chemical reaction system; for details see [And11].

The idea of the stochastic simulation algorithm (SSA), also known as *Gillespie algorithm* is to simulate the random reaction events by

- drawing a random time until the next reaction occurs,
- drawing randomly one of the M reactions to occur.

Once a reaction has fired, the number of molecules is updated according to the corresponding stiochoimetric vector.

Time until next reaction To make this intuitive idea precise, let $X(t) = x$ and consider the time τ until the next reaction fires. Call

$$p_0(\tau; x, t) = P(\text{no reaction fires in } [t, t+\tau) | X(t) = x) \tag{7.9}$$

the probability that no reaction happens in the finite interval $[t, t+\tau)$ with $\tau > 0$. Further let us suppose that whatever happens in $[t, t+\tau)$ is independent of what happens in $[t+\tau, t+\tau+s)$ for all $s > 0$, in other words, the system is memoryless. The, by independence, the probability that no reaction fires in $[t, t+\tau+d\tau)$ can be written as

$$p_0(\tau + d\tau; x, t) = p_0(\tau; x, t)\left(1 - \sum_{j=1}^{M} a_j(x)d\tau\right), \tag{7.10}$$

where, by definition of the propensities, the term in the parenthesis is the probability of no reaction between $t+\tau$ and $t+\tau+d\tau$, given that $X(t+\tau) = x$. Rearranging terms, dividing by $d\tau$ and letting $d\tau \to 0$, it follows that

$$\frac{dp_0}{d\tau} = -a_{\text{tot}} p_0 \,, \tag{7.11}$$

with

$$a_{\text{tot}}(x) = \sum_{j=1}^{M} a_j(x) \,. \tag{7.12}$$

Solving (7.11) with the initial condition $p_0(\tau = 0; x, t) = 1$ it follows that

$$p_0(\tau; x; t) = \exp(-a_{\text{tot}}(x)\tau) \,, \tag{7.13}$$

which is to say that τ is an exponential waiting time with parameter a_{tot}.[17] This implies that the average waiting time between two reactions is

$$\mathbb{E}(\tau | X(t) = x) = 1/a_{\text{tot}}(x) \,. \tag{7.14}$$

Next reaction index To determine the next reaction, we define $p_1(\tau, j; x, t)d\tau$ to be the probability that no reaction happens in the interval $[t, t+\tau)$ and the j-th reaction fires in $[t+\tau, t+\tau+d\tau)$, given that $X(t) = x$. Then, by independence and definition of the propensities, we have

$$p_1(\tau, j; x, t) = p_0(\tau; x, t) a_j(x) \,. \tag{7.15}$$

Using (7.13), the latter can be recast as the product of two probability densities:

$$p_1(\tau, j; x, t) = \left(\frac{a_j(x)}{a_{\text{tot}}(x)}\right)\left(a_{\text{tot}}(x)\exp(-a_{\text{tot}}(x)\tau)\right) \,. \tag{7.16}$$

We notice that p_1 is the joint probaility density of the time until the next reaction τ and the reaction index j. The (conditional) probability of the reaction index j is proportional to a_j with a_{tot} as normalization constant. Since p_1 is a product density, both random variables τ and j are independent and can be drawn independently. The following algorithm goes back to [Gil77].

The following Lemma is helpful for generating exponentially distributed random variables from a uniformly distributed random variable.

Lemma 7.8. *Let Z be a random variable with cumulative distribution function $F(z) = P(Z \leq z)$. If $U \in [0, 1]$ is a uniformly distributed random variable, then*

$$\tilde{Z} = F^{-1}(U)$$

[17] Waiting times are memoryless iff they are exponentially distributed.

Algorithm 1 Stochastic Simulation Algorithm

```
Given $X(0) = x$, define $a_{\text{tot}}(x) = \sum_j a_j(x)$.
while $t < T$ do
    Generate exponential waiting time $\tau \sim \text{Exp}(a_{\text{tot}}(X(t)))$.
    Pick reaction index $j$ randomly with probability $a_j(X(t))/a_{\text{tot}}(X(t))$
    Set $t \mapsto t + \tau$ and update state vector $X(t) \mapsto X(t) + \nu_j$.
end while
```

is distributed according to F where

$$F^{-1}(u) = \inf\{z \in \mathbb{R} \colon F(z) \geq u\}$$

is the generalized inverse of F. In particular,

$$\tau = -\frac{\log(1-U)}{a_{\text{tot}}}$$

is exponentially distributed with parameter $a_{\text{tot}} > 0$.

Proof. Let the random variable $\tilde{Z} \colon [0,1] \to \mathbb{R}$ be defined by $U \mapsto F^{-1}(U)$. Then

$$\begin{aligned} P(\tilde{Z} \leq z) &= P(\{U \in [0,1] \colon \tilde{Z}(U) \leq z\}) \\ &= P(\{U \in [0,1] \colon F^{-1}(U) \leq z\}) \\ &= P(\{U \in [0,1] \colon U \leq F(z)\}) \\ &= F(z)\,, \end{aligned}$$

where we have used the monotonicity of F in the third equality and the fact that $u \in [0,1]$ is uniformly distributed in the last equality. This shows that the distribution function of $\tilde{Z}$ is F. The rest of the proof is left as an exercise. □

If the cumulative distribution function is a continuous monotonic function, then the generalized inverse agrees with the standard inverse function.

Problems

Exercise 7.9. *Let $x = (x_1, x_2, \ldots)$ be the state vector of a system with species $A, B, \ldots$. Construct propensity functions $a(x)$ for the following reactions.*

a) zeroth-order reactions: $\emptyset \xrightarrow{c_0} A$

b) first-order reactions: $A \xrightarrow{c_1} B$

c) dimerization: $A + A \xrightarrow{c_2} B$

Justify your choice in each case.

Exercise 7.10. *Consider a reversible reaction between three species of the form*

$$A + B \underset{c_-}{\overset{c_+}{\rightleftharpoons}} C\,,$$

with rate constants $c_\pm > 0$. Let $X(t) = (X_A(t), X_B(t), X_C(t))$ be the state vector of the system at time $t \geq 0$.

a) Derive the CME for the probability density function (pdf)

$$\rho(x,t) = P(X(t) = x)\,.$$

b) Let $X(0) = (3,2,1)$. Recast the CME as an equivalent system of linear odes and compute its equilibrium state as a function of $c_\pm$. Explain the meaning of the equilibrium (Hint: use that $\rho(x,t)$ is a pdf).

Exercise 7.11. *Consider the Michaelis-Menten system*

$$S + E \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} SE \xrightarrow{k_2} E + P$$

with initial values

$$X_S(0) = 5 \cdot 10^{-7} N_A \text{vol}\,,\; X_E(0) = 2 \cdot 10^{-7} N_A \text{vol}\,,\; X_{SE}(0) = X_P(0) = 0\,.$$

and kinetic parameters

$$k_1 = 10^6/(N_A \text{vol})\,,\; k_{-1} = 10^{-4}\,,\; k_2 = 10^{-1}\,,\; \text{vol} = 10^{-15}\text{liters}\,.$$

Here $N_A = 6.023 \cdot 10^{23}$ is Avogadro's number.

a) Derive the propensities and stiochiometric vectors.

b) Simulate the Michaelis-Menten system using Gillespie's SSA from [Hig08]. Generate additional realizations using 10 times smaller and 10 times larger initial values. Plot typical realizations and explain your observation.

8 Modelling of traffic flow

Mathematical tools & concepts: delay differential equations, Euler method, scalar conservation laws, partial differential equations.
Suggested reference: [MeG07, BCD02]

We will discuss two different ways to model traffic flows: a microscopic approach that is based on the dynamics of single cars and a mean field approach that employs an analysis on the level of fluxes and densities of vehicles.

8.1 From individual vehicles to vehicle densities

Suppose there are N vehicles in one traffic lane, all of equal length l and mass m. The vehicles are labelled $j = 1, \dots, N$ where $j = 1$ corresponds to the leading vehicle. Let the front car have the distance $x_j(t)$ to the beginning of the road at time $t \geq 0$. We assume that there is only one traffic lane, so vehicles cannot overtake each other.

A delay differential equation for the vehicle positions On a busy road the vehicles have to brake depending on both the distance between vehicles and their relative velocities.

Suppose that the average values of $|x_{j+1}(t) - x_j(t)|$ are relatively small for all $j = 1, \dots, N-1$, i.e., we consider a busy road, and vehicles avoid collisions by braking when they come too close. It is reasonable to assume that the braking force of, say, vehicle $j+1$ will be higher the smaller the distance $|x_{j+1}(t) - x_j(t)|$ to the j-th vehicle and the faster it approaches the j-th vehicle, i.e., the larger the relative velocity $d/dt(x_{j+1}(t) - x_j(t))$. Let us further assume that the response of the driver of vehicle $j+1$ is delayed by $\tau > 0$, where for simplicity we assume that the reaction time τ is constant for all drivers. Letting F_{j+1} denote the braking force, the simplest model to account for this situation is

$$F_{j+1}(t+\tau) = k \frac{\dot{x}_{j+1}(t) - \dot{x}_j(t)}{|x_{j+1}(t) - x_j(t)|}\,, \tag{8.1}$$

where, as usual, the dot denotes derivative with respect to t and $k > 0$ is constant. Using Newton's law, $F = ma$, equation (8.1) implies that

$$\begin{aligned} m\frac{d^2}{dt^2} x_{j+1}(t+\tau) &= k \frac{\dot{x}_{j+1}(t) - \dot{x}_j(t)}{|x_{j+1}(t) - x_j(t)|} \\ &= k \frac{d}{dt} \log |x_{j+1}(t) - x_j(t)|\,, \end{aligned} \tag{8.2}$$

which can be integrated to yield

$$\frac{d}{dt} x_{j+1}(t+\tau) = \frac{k}{m} \log |x_{j+1}(t) - x_j(t)| + a_{j+1}\,, \tag{8.3}$$

for $j = 1, \dots, N-1$. Equation (8.3) is a system of $N-1$ delay differential equations (DDE) where the positions $x_1(t)$ (and the velocities) of the first vehicle are given. There is no way to solve (8.3) analytically, but we can find a numerical solution as we will discuss later on in this section (cf. Exercise 8.8). However there is some hope that the DDE will admit steady state solutions or equilibria,

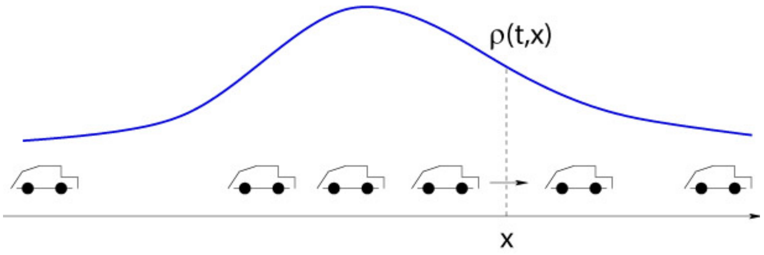


Figure 8.1: Density of vehicles and micro-macro passage.

for, in the real world, nothing can grow or decay forever. One special case of the DDE is its *Markovian limit* $\tau \to 0$, in which case we obtain a nonlinear system of $N-1$ ordinary time-inhomogeneous differential equations

$$\begin{aligned} x_1(t) &= \phi(t) \\ \dot{x}_{j+1}(t) &= \frac{k}{m} \log|x_{j+1}(t) - x_j(t)| + a_{j+1}\,, \quad j = 1, \ldots, N-1\,. \end{aligned} \tag{8.4}$$

A micro-macro passage: densities and fluxes It is known that the velocities of cars decrease when their density increases. To arrive at a description of (8.3) in terms of densities and fluxes, consider a street section of length $2s \gg l$ and define the density of vehicles at x at time t to be

$$\rho(x,t) = \frac{\#\text{ vehicles in } (x-s, x+s) \text{ at time } t}{2s}\,, \tag{8.5}$$

where we assume that the street section is symmetric around the position $x \in \mathbb{R}$. We regard ρ as a macroscopic variable that replaces the detailed microscopic description in terms of the positions of single vehicles by a *coarse-grained* description in terms of (average) numbers of cars per street section; clearly ρ depends on the length of the street section over which we average, but it can be shown that $\rho = \rho_{N,s}$ converges to a limit as $N \to \infty$ and $l \ll 2s \to 0$ with $Nl \to \text{const}$ [BCD02]. Here we assume the birds-eye perspective (e.g. seen from a traffic surveillance helicopter) and busy traffic conditions and, as a consequence, we may safely ignore the dependence of ρ on N, s.

Example 8.1. *One situation in which ρ is independent of $N < \infty$ and s is when all vehicles are at equal distance d at any time (which implies that they are all moving at the same speed), in which case*

$$\rho(x,t) = (d+l)^{-1} \tag{8.6}$$

(cf. Exercise 8.9). As the vehicle length is equal to l, the denominator is bounded from below by l; therefore the maximum achievable density is the density of bumber-to-bumber traffic at constant speed, with

$$\rho(x,t) = l^{-1}\,. \tag{8.7}$$

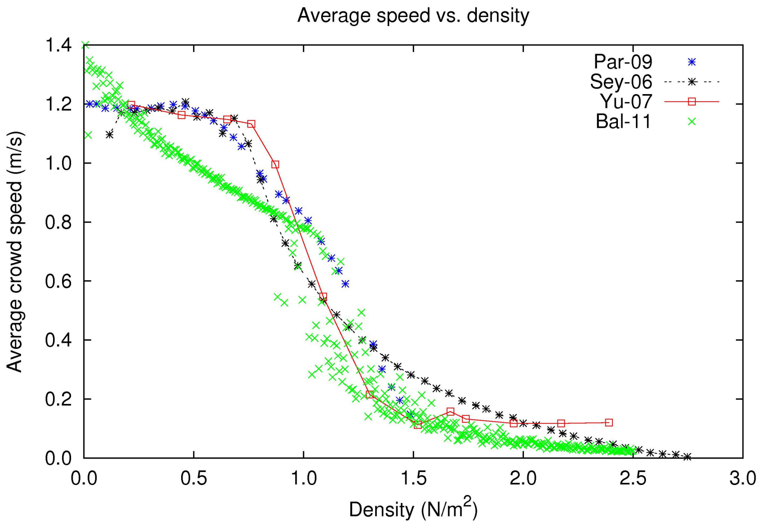


Figure 8.2: Fundamental diagram of pedestrian flows (from: [BMR11]).

We want to analyse the maximum capacity of the traffic lane under equilibrium conditions. To this end we assume that the observed speed v of vehicles at (x,t) depends only on the density ρ. In an abuse of notation, we write

$$v(x,t) = v(\rho(x,t))\,. \tag{8.8}$$

It is known from empirical data of traffic flows that there exists a critical density ρ_{crit}, below which the vehicles move at the maximum possible speed $v_{\max}$, and that there is a maximum density $\rho_{\max}$, at which the flow stops. From Example 8.1 it readily follows that $\rho_{\max} \le 1/l$. From the critical to the maximum density, v decays towards zero where it is also known from experimental data that v is a decreasing function of the density, i.e.

$$v'(\rho) \le 0\,. \tag{8.9}$$

Figure 8.2 shows experimental and simulation data of pedestrian flows under various environmental conditions that shows the universal signature of almost all traffic flows; the graphical relation $v(\rho)$ is called *fundamental diagram.*

Steady state and equilibrium flow We suppose that all vehicles (cars, pedestrians, ...) are separated by a distance $d > 0$ and move at the same constant speed v. The equilibrium density corresponding to this situation is (cf. Exercise 8.1)

$$\rho(x,t) = (d+l)^{-1}\,, \quad (x,t) \in \mathbb{R} \times [0,\infty)\,. \tag{8.10}$$

In equilibrium, all vehicles move at the same speed $v_j = dx_j/dt$, hence together with the DDE (8.3) it follows that

$$v = \lambda \log(d+l) + a\,, \tag{8.11}$$

where we have introduced the shorthands $\lambda = k/m$ and $a = a_{j+1}$ for $j = 1, \ldots, N-1$. Combining the last two equations, we find the fundamental equilibrium relation between the speed v and density ρ:

$$v = -\lambda \log \rho + a\,, \tag{8.12}$$

with the yet unknown parameters a and λ that must be determined from data; by definition of $\rho_{\max}$, it holds that $v(\rho_{\max}) = 0$, which is equivalent to

$$a = \lambda \log \rho_{\max}\,. \tag{8.13}$$

Hence

$$v = -\lambda \log\left(\frac{\rho}{\rho_{\max}}\right)\,. \tag{8.14}$$

An expression for λ is easily obtained by requiring that v is continuous as a functional of ρ. Setting $v_{\max} = v(\rho_{\text{crit}})$, the last equation entails

$$\lambda = v_{\max}\left\{\log\left(\frac{\rho_{\max}}{\rho_{\text{crit}}}\right)\right\}^{-1}\,, \tag{8.15}$$

which, together with the empirical finding that $v(\rho)$ equals $v_{\max}$ below the critical vehicle density yields the surprisingly general relation

$$v(\rho) = \begin{cases} v_{\max}\,, & \rho \le \rho_{\text{crit}} \\ v_{\max}\left\{\log\left(\frac{\rho_{\max}}{\rho_{\text{crit}}}\right)\right\}^{-1} \log\left(\frac{\rho_{\max}}{\rho}\right)\,, & \rho > \rho_{\text{crit}}\,. \end{cases} \tag{8.16}$$

Maximum traffic flux at equilibrium We now define the instantaneous traffic flux J as the number of vehicles passing through a street sector $[x, x+\Delta x)$ in the time interval $[t, t+\Delta t)$, in other words,

$$J = \left(\frac{\#\text{ vehicles in } [x, x+\Delta x) \text{ at time } t}{\Delta x}\right)\left(\frac{\Delta x}{\Delta t}\right)$$

In mathematical terms, letting Δx and Δt tend to zero, we have the following

Definition 8.2 (Density flux). *We define the flux J to be the functional*

$$J(\rho) = \rho v(\rho)\,.$$

With (8.16) it readily follows that

$$J(\rho) = \begin{cases} \rho v_{\max}\,, & \rho \le \rho_{\text{crit}} \\ \rho v_{\max}\left\{\log\left(\frac{\rho_{\max}}{\rho_{\text{crit}}}\right)\right\}^{-1} \log\left(\frac{\rho_{\max}}{\rho}\right)\,, & \rho > \rho_{\text{crit}}\,, \end{cases} \tag{8.17}$$

which, provided that $\rho_{\max} \ge e \cdot \rho_{\text{crit}}$, attains its unique maximum at

$$\rho^* = \frac{\rho_{\max}}{e}\,, \tag{8.18}$$

with $e = 2.71828\ldots$ being the base of the natural logarithm (see Exercise 8.10).

8.2 Traffic jams and propagation of perturbations

We want to study what happens when the first vehicle brakes, i.e., we want to study the effect of a perturbation of the lead vehicle on the pursuing vehicles, when the traffic flows close to the maximum flux point.

To this end, let us go back to the microscopic picture again and consider a platoon of vehicles under maximum flux conditions as described in the previous section. We suppose that all vehicles move at constant speed

$$v(\rho^*) = v_{\max}\left\{\log\left(\frac{\rho_{\max}}{\rho_{\text{crit}}}\right)\right\}^{-1} \tag{8.19}$$

where we have used (8.16) with $\rho^* = \rho_{\max}/e$ and have tacitly assumed that $\rho^* > \rho_{\text{crit}}$. Let us further assume that we can extend the time $t \geq 0$ to the whole real axis, $t \in \mathbb{R}$, and that the lead vehicle crosses the origin $x = 0$ at time $t = 0$, i.e. $x_1(0) = 0$. With the sign convention

$$x_{j-1} - x_j \geq l > 0 \tag{8.20}$$

and the shorthand $v^* = v(\rho^*)$, equation (8.3) becomes

$$\begin{aligned}\frac{d}{dt}x_{j+1}(t+\tau) &= \lambda\log|x_{j+1}(t) - x_j(t)| + a\\ &= v^*\log(x_j(t) - x_{j+1}(t)) + v^*\log\rho_{\max}\\ &= v^*\log\left(\rho_{\max}(x_j(t) - x_{j+1}(t))\right)\end{aligned} \tag{8.21}$$

where we have used that $v^* = \lambda$, which follows from (8.15) and (8.20) and which together with (8.13) entails the relation $a = v^*\log\rho_{\max}$.

Braking of the lead vehicle and perturbation of the pursuing vehicles For $t > 0$, we consider the DDE system

$$\begin{aligned}\frac{d}{dt}x_1(t) &= \phi(t)\\ \frac{d}{dt}x_j(t+\tau) &= v^*\log\left(\rho_{\max}(x_{j-1}(t) - x_j(t))\right)\,, \quad j = 2,\ldots,N\,.\end{aligned} \tag{8.22}$$

where we assume that the system in equilibrium for $t \leq 0$:[18]

$$x_j(t) = v^*t - (j-1)(d+l)\,, \quad j = 1,\ldots,N \tag{8.23}$$

Let us assume that the first vehicle with position $x_1(t)$ brakes at time $t = 0$ and releases the brake after a short time $t_b > 0$. Specifically,

$$\phi(t) = \begin{cases} v^*\,, & t \leq 0\\ v^*\left(1 - b(t)\right)\,, & t > 0\,.\end{cases} \tag{8.24}$$

where we set $b(t) = kt\exp((t_b - t)/t_b)$. Solving the ODE for $x_1(t)$, using (8.22)–(8.24) we find

$$x_1(t) = \int_0^t \phi(s)\,ds = v^*t - v^*\int_0^t b(s)ds\,, \quad t > 0\,, \tag{8.25}$$

[18] The reader should think of v^* as a model parameter, rather than the instantaneous velocity of individual vehicles that is given by $v_j = dx_j/dt$.

with

$$\int_0^t b(s)ds = ekt_b \left(t_b - (t + t_b)e^{-t/t_b} \right) . \tag{8.26}$$

We call $y_j(t)$ the hypothetical position of the j-th vehicle, if the lead vehicle had not braked, i.e. without the perturbation. We further call

$$z_j(t) = x_j(t) - y_j(t) , \quad j = 1, \dots, N , \tag{8.27}$$

the *perturbation displacement* due to the perturbation of the lead vehicle's motion. The perturbation displacement of the first vehicle then is

$$z_1(t) = -v^* \int_0^t b(s)ds , \quad t > 0 . \tag{8.28}$$

By (8.23), it follows for the pursuing vehicles with indices $j = 2, \dots, N$,

$$z_j(t) = x_j(t) - v^* t + (j-1)(d+l) , \quad t > 0 . \tag{8.29}$$

Note that $z_j(t) = 0$ for $t \leq 0$ and all $j = 1, \dots, N$. Further note that the *non-collision constraint*

$$x_{j-1}(t) - x_j(t) \geq l \quad \forall t \in \mathbb{R} . \tag{8.30}$$

entails

$$z_j(t) - z_{j-1}(t) \leq d \quad \forall t \in \mathbb{R} . \tag{8.31}$$

The latter follows from (8.29), together with

$$l \leq x_{j-1}(t) - x_j(t) = z_{j-1}(t) - z_j(t) + d + l \quad \forall t \in \mathbb{R} . \tag{8.32}$$

Reaction time and the onset of traffic jam Equations (8.27)–(8.31) allow us to recast (8.22)–(8.23) as a DDE for the perturbation displacement. Bearing in mind that $d + l = e/\rho_{\max}$ holds under maximum flow conditions, the perturbation displacement (8.29) of the pursuing vehicles reads

$$z_j(t) = x_j(t) - v^* t + \frac{(j-1)e}{\rho_{\max}} , \quad t > 0 . \tag{8.33}$$

Plugging the last equation into (8.22), using that (8.32), we obtain a closed DDE system for the perturbation displacement for $t > 0$:

$$\frac{d}{dt} z_j(t + \tau) = v^* \log \left(\rho_{\max} \left\{ \frac{e}{\rho_{\max}} + z_{j-1}(t) - z_j(t) \right\} \right) - v^* , \tag{8.34}$$

with $j = 2, \dots, N$, the lead vehicle displacement

$$z_1(t) = -v^* \int_0^t b(s)\, ds \tag{8.35}$$

and the initial conditions

$$z_j(0) = 0 , \quad j = 2, \dots, N . \tag{8.36}$$

Note that (8.34) is equivalent to

$$\frac{d}{dt} z_j(t) = v^* \log \left(\left\{ 1 + \frac{\rho_{\max}}{e} \left(z_{j-1}(t - \tau) - z_j(t - \tau) \right) \right\} \right) , \tag{8.37}$$

which follows from shifting the independent variable t according to $t \mapsto t - \tau$ and moving the rightmost term $-v^*$ in (8.34) under the logarithm.

Figure 8.3 shows a simulation of (8.34)–(8.36) for different reaction times τ; cf. Exercise 8.8 for the parameter values used and for further details.

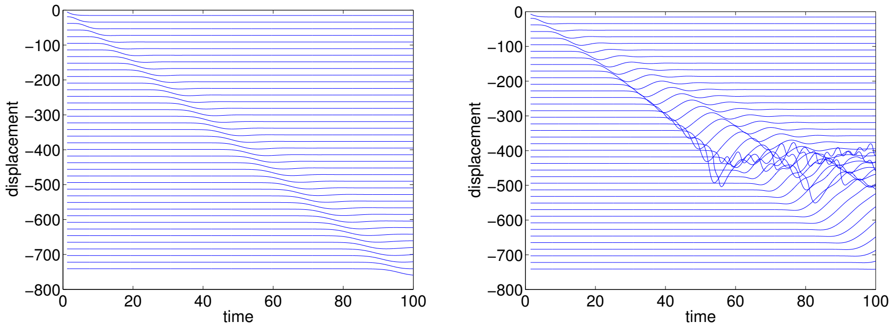


Figure 8.3: The left panel shows how a perturbation propagates in case of short reaction time (no accident), the right panel shows the case of a too long reaction time; the 11th car cannot brake anymore and crashes into the 12th car.

8.2.1 Numerical solution

First vehicle. The equation for the displacement of the first vehicle

$$\frac{d}{dt}z_1(t) = -v^* b(t) \tag{8.38}$$

is an ODE. The easiest way to solve it numerically is to use the *forward Euler scheme*, which approximates an ODE of the form $\dot{x} = f(x,t)$ with the iteration

$$\tilde{x}(t+h) = \tilde{x}(t) + hf(\tilde{x}(t),t), \qquad \tilde{x}(0) = x_0.$$

The scheme is derived by approximating $\frac{d}{dt}x(t)$ with the finite difference $h^{-1}(x(t+h) - x(t))$ for some finite step size $h > 0$, which is a parameter of the algorithm and must be chosen by the user. Here the tilde denotes numerical approximation. The forward Euler scheme applied to (8.38) gives

$$\tilde{z}_1(t+h) = \tilde{z}_1(t) - hv^* b(t), \tag{8.39}$$

an equation we can iterate N times (starting from $t = 0$) in order to obtain the numerical solution vector $(\tilde{z}_1(0), \tilde{z}_1(h), \ldots, \tilde{z}_1(hN))$. Here $\tilde{z}_1(nh)$ is the numerical approximation to the true solution $z_1(t)$ for $t = nh$.

2nd vehicle. The equations (8.37) for vehicles 2 to N are DDEs, and care must be taken with their numerical integration.

If we apply the forward Euler scheme to (8.37) for $j = 2$, we obtain

$$\tilde{z}_2(t+h) = \tilde{z}_2(t) + hv^* \log\left(\left\{1 + \frac{\rho_{\max}}{e}\left(\tilde{z}_1(t-\tau) - \tilde{z}_2(t-\tau)\right)\right\}\right) \tag{8.40}$$

The initial condition $\tilde{z}_2(0) = 0$ is not enough to solve (8.40) due to the presence of the time delay $\tau > 0$: In order to compute the first iterate $\tilde{z}_2(h)$, we need $\tilde{z}_1(-\tau)$ and $\tilde{z}_2(-\tau)$. In general, the initial condition $\tilde{z}_1(t) = \tilde{z}_2(t) = 0$ for $t \in [-\tau, 0)$ (which we luckily have) is needed to iterate (8.40) forward in time.

In addition, $\tilde{z}_1(t)$ from equation (8.39) is needed for $t = 0, h, \ldots, hN$ in order to compute $\tilde{z}_2(t)$ for $t = 0, h, \ldots, hN$ using (8.40). Thus we can solve the equation for the 2nd vehicle after we solved the equation for the 1st.

Iterating this argument, we can solve the equation for the $(j+1)$st vehicle after we solved the equation for the jth vehicle using forward Euler. This eventually leads to the numerical solution $(\tilde{z}_j(0), \tilde{z}_j(h), \ldots, \tilde{z}_j(hN))$ being available for all $j = 1, \ldots, N$.

8.3 Flow modeling: macroscopic modeling of traffic flows

In the microscopic model, each driver reacts only to the car in front. This is realistic in a tunnel, or in very dense traffic. However on open road, the driver will probably look further ahead and react according to changes in density. We want to model the overall flow in this setting, using a so-called continuum model of traffic flow.

Again we consider only one traffic lane, without entrances or exits. If we select some stretch of the road between two points denoted x_1 and x_2, the total number of cars to be found between x_1 and x_2 will depend on the time t. Specifically, if more cars flow into the segment $[x_1, x_2]$ than flow out of it, the number of cars in the segment will increase. This can be expressed mathematically through the *conservation of total number of cars* in $[x_1, x_2]$:

$$\text{Rate of change of traffic} = \text{Traffic inflow} - \text{Traffic outflow}$$

Scalar conservation laws In the following, we will make use of variables such as the density $\rho = \rho(x,t)$, the flux $J = J(x,t)$ and the average speed $v = v(x,t)$. Recall that the three are related by the simple identity $J = v\rho$. Considering our stretch of road in space and time, we have

$$J(x,t) := v(x,t)\rho(x,t)$$

where v is the observed speed at location x and time t. We assume that J and ρ are nonnegative functions. We also make the simplistic assumption that the speed is a function of density alone, *i.e.* $v = v(\rho)$. In the following, we will keep track of the total number of cars in $[x_1, x_2]$ during a time $[t_1, t_2]$.

The number of cars entering $[x_1, x_2]$ through the point x_1 during the time interval $[t_1, t_2]$ is given by $\int_{t_1}^{t_2} J(x_1, t)dt$, and the number of cars leaving $[x_1, x_2]$ through the point x_2 is $\int_{t_1}^{t_2} J(x_2, t)dt$. The number of cars to be found in the space interval $[x_1, x_2]$ at time t_1 is given by $\int_{x_1}^{x_2} \rho(x, t_1)dx$ and the number at time t_2 is given by $\int_{x_1}^{x_2} \rho(x, t_2)dx$.

Since the total number of cars in $[x_1, x_2]$ during a time $[t_1, t_2]$ is conserved (no cars disappear or appear), we have

$$\int_{x_1}^{x_2} \rho(x,t_2)dx - \int_{x_1}^{x_2} \rho(x,t_1)dx = \int_{t_1}^{t_2} J(x_1,t)dt - \int_{t_1}^{t_2} J(x_2,t)dt \tag{8.41}$$

Supposing that ρ and J are continuously differentiable with respect to x and t, the left-hand side of 8.41 can be expressed as

$$\int_{x_1}^{x_2} \left[\rho(x,t_2) - \rho(x,t_1)\right] dx = \int_{x_1}^{x_2} \int_{t_1}^{t_2} \frac{\partial}{\partial t}\rho(x,t)dtdx. \tag{8.42}$$

The right-hand side can be rewritten similarly, leading to

$$\int_{x_1}^{x_2}\int_{t_1}^{t_2}\frac{\partial}{\partial t}\rho(x,t)dtdx = -\int_{t_1}^{t_2}\int_{x_1}^{x_2}\frac{\partial}{\partial x}J(x,t)dxdt.$$

The latter is equivalent to

$$\int_{x_1}^{x_2}\int_{t_1}^{t_2}\left[\frac{\partial}{\partial t}\rho(x,t)+\frac{\partial}{\partial x}J(x,t)\right]dtdx = 0, \tag{8.43}$$

which is true for any choice of rectangle $[x_1, x_2] \times [t_1, t_2]$. By the Fundamental Theorem of the Calculus of Variation (see the following Lemma for a simple version) this implies that

$$\frac{\partial}{\partial t}\rho(x,t)+\frac{\partial}{\partial x}J(x,t) = 0, \tag{8.44}$$

which is a first-order conservation law.

Lemma 8.3. *If $f(x,t)$ is a continuous function defined on $\mathbb{R}^2$ such that*

$$\iint_R f(x,y)dxdy = 0\,.$$

for each rectangle $R \subseteq \mathbb{R}^2$, then $f(x,y) \equiv 0$ for all (x,y).

Proof. Suppose that there exists a pair of coordinates (x_0, y_0) such that $f(x_0, y_0) \neq 0$. Without loss of generality assume that $f(x_0, y_0) > 0$. Since f is continuous, there is a $\delta > 0$ such that $f(x,y) > f(x_0,y_0)/2$ whenever $|x - x_0| < \delta$ and $|y - y_0| < \delta$. Therefore if we let

$$R_\delta := \{(x,y) : |x - x_0| < \delta \text{ and } |y - y_0| < \delta\}\,,$$

then

$$\iint_{R_\delta} f(x,y)dxdy \geq \frac{1}{2}\iint_{R_\delta} f(x_0,y_0)dxdy.$$

By assumption the left hand side is zero, consequently we obtain

$$0 \geq 2\delta^2 f(x_0,y_0),$$

which is a contradiction. Thus $f(x,y) \equiv 0$ □

Simplifying the scalar conservation law The conservation law 8.44 is a first order partial differential equation with two unknowns. To be able to solve it, we need one more equation - a *state equation* - relating the unknowns ρ and J. For this, recall that the microscopic analysis provided us with an expression for the traffic flux $J(\rho)$ in equilibrium conditions (8.17):

$$J(\rho) = \begin{cases} \rho v_{\max}\,, & \rho \leq \rho_{\text{crit}} \\ \rho v_{\max}\left\{\log\left(\frac{\rho_{\max}}{\rho_{\text{crit}}}\right)\right\}^{-1}\log\left(\frac{\rho_{\max}}{\rho}\right)\,, & \rho > \rho_{\text{crit}}\,. \end{cases}$$

The state equation has the right behavior, *i.e.* the traffic flux J increases linearly for small density ρ, levels off until a maximum is reached and then decreases until J becomes zero at bumper-to-bumper traffic. However, the derivative has a jump discontinuity at $\rho = \rho_{\text{crit}}$, which is not so realistic. Moreover, it was derived under equilibrium conditions which will not necessarily be satisfied.

We will now create a similar state equation for $J = J(\rho)$ that is differentiable for all admissible ρ. A simple choice is $J(\rho) = a\rho(b-\rho)$ with a parameter $a > 0$ and $b = \rho_{max}$. When doing this, we assume that $J = J(\rho)$ even outside of equilibrium conditions, meaning that the flux adjusts smoothly to a change in density.

With this choice, the conservation law 8.44 takes the form

$$\frac{\partial}{\partial t}\rho(x,t) + J'(\rho)\frac{\partial}{\partial x}\rho(x,t) = 0, \tag{8.45}$$

This is a first-order partial differential equation (PDE) for ρ. Together with initial and boundary conditions, this model is solvable using the method of characteristics.

Remark 8.4. *From the definition of the density flux, we have that $J(\rho) = v(\rho)\rho$. Therefore the derivative $J'(\rho)$ has the dimension of velocity. We will see later that the PDE expresses that "traffic waves" propagate with a velocity given by $J'(\rho)$.*

Linear traffic waves Before going to the method of characteristics, we will consider the simpler case of linear traffic waves. Our PDE 8.44 will be examined in more detail using the relation $J(\rho) = v_{\text{max}}\rho\left(1 - \frac{\rho}{\rho_{\text{max}}}\right)$. This is a simple quadratic function of ρ.

We will use the form of the equation given by (8.45) to investigate the propagation of *linear traffic waves.*

Let us suppose that $\rho = \rho_0 + \delta\rho$ in (8.45), where $\delta\rho \ll \rho_0$. In other words, we consider a case where the traffic density is slightly perturbed from a constant density ρ_0. To put this into (8.45), we can use the Taylor expansion

$$J'(\rho_0 + \delta\rho) = J'(\rho_0) + J''(\rho_0)\delta\rho + \cdots,$$

but we see that the terms in δ can be dropped since both partial derivatives in (8.45) are of order δ. Thus we obtain the linearized form of the PDE

$$\frac{\partial}{\partial t}\rho(x,t) + J'(\rho_0)\frac{\partial}{\partial x}\rho(x,t) = 0, \tag{8.46}$$

Notice here that $J'(\rho_0)$ is a *constant* and has the dimension of velocity. We can call it v_0 and get

$$\frac{\partial}{\partial t}\rho(x,t) + v_0\frac{\partial}{\partial x}\rho(x,t) = 0. \tag{8.47}$$

Now by substitution, one can easily see that $\rho = f(x - v_0 t)$ is solution for any differentiable $f(x)$. Note that $\rho = f(x - v_0 t)$ describes a wave moving with velocity v_0. For $v_0 > 0$, the wave moves to the right, the opposite sign moving to the left.

Example 8.5. *If $f(x) = \sin x$, then $\rho = \sin(x - v_0 t)$, the point (x,t) such that $x - v_0 t = \pi/2$ is at the crest of the wave and it moves in the $x - t$ plane along* *the straight line* $x = v_0 t + \pi/2$*. Thus the solutions of (8.46) represent* linear traffic waves *. The velocity* v_0 *is given by*

$$v_0 = J'(\rho_0) = v_{\max}\left(1 - \frac{2\rho_0}{\rho_{\max}}\right).$$

It is important to realise that this velocity is relative to the road surface.Note that when $\rho_0 \approx 0$ *we have* $v_0 \approx v_{\max}$*. This says that density changes are propagating with the velocity of the cars when there are few cars, which is reasonable. It also means that traffic waves move with the traffic (again reasonable for light traffic). At the other extreme, when* $\rho \approx \rho_{\max}$*, we have* $v_0 \approx -v_{\max}$*. In this case cars are moving slowly, and the waves move backward relative to the car's motionat the high speed of* v_{max}*. This happens when cars move slowly in a packed traffic and one car suddenly stops. The wave of red brake lights caused by many rear-end cars braking can move towards a driver quickly.*

Method of characteristics We will now learn how to solve initial value problems for first-order PDE's using the method of *characteristic curves.* The idea of the method is to discover curves (the characteristic curves) along which the PDE becomes an ODE. Once the ODE is found, it can be solved along the characteristic curves and transformed into a solution for the PDE. Consider a function $f(x,t)$ satisfying a first-order linear PDE of the form

$$\frac{\partial}{\partial t} f(x,t) + v(x,t)\frac{\partial}{\partial x} f(x,t) = 0. \tag{8.48}$$

We will view this equation as saying that f is not changing along a curve $x = x(t)$, which means

$$\frac{d}{dt} f(x(t),t) = 0. \tag{8.49}$$

Using the chain rule we get

$$0 = \frac{\partial f}{\partial t} + \frac{dx}{dt}\frac{\partial f}{\partial x}. \tag{8.50}$$

By virtue of (8.48) and (8.50) we must have

$$\frac{dx}{dt} = v(x,t), \tag{8.51}$$

which is an ODE for $x(t)$. A solution $\phi(x(t),t)$ satisfies

$$\frac{d}{dt}\phi(x(t),t) = \frac{\partial \phi}{\partial t} + x'(t)\frac{\partial \phi}{\partial x} = \frac{\partial \phi}{\partial t} + v(x,t)\frac{\partial \phi}{\partial x} = 0, \tag{8.52}$$

which implies that $\phi(x(t),t) = \phi_0(x_0)$. Each value of x_0 determines a unique characteristic base curve if v is such that the initial value problems for the ODE (8.51) are uniquely solvable (we assume v smooth enough for that). On any of the integral curves $\phi(x(t),t) = \phi_0(x_0)$, f will also be constant (see (8.49) and (8.52)). Since the curves of constant ϕ and constant f coincide, f has to be a function of ϕ alone

$$f(x(t),t) = F(\phi(x,t)). \tag{8.53}$$

We will consider for example an initial condition $f(x,0) = f_0(x)$ such that $f(x,0) = F(\phi(x,0))$. This equation can be solved for x, which then leads to $f(x,t) = f_0(x(\phi(x,t)))$.

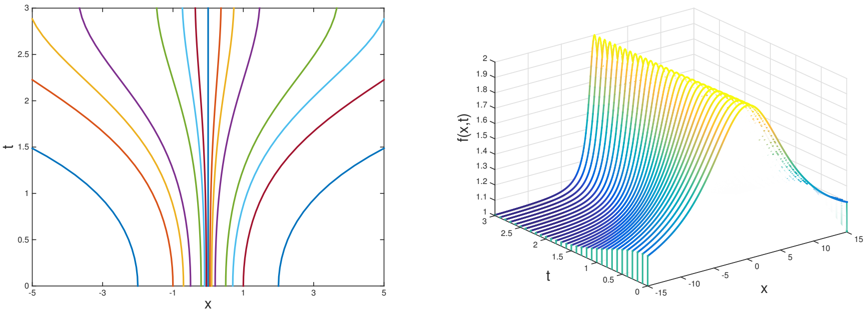


Figure 8.4: The left panel shows the characteristic base curves for example 8.6, the right panel shows the corresponding solution $f(x,t)$

Example 8.6. *Consider the initial value problem*

$$\frac{\partial f}{\partial t} + x\sin(t)\frac{\partial f}{\partial x} = 0, \quad f_0(x) = 1 + \frac{1}{1+x^2}.$$

Here we have $v(x,t) = x\sin t$. Characteristic base curves for this problem are solutions of

$$\frac{dx}{dt} = x\sin t, \quad x(0) = x_0.$$

By separation of variables we get

$$\int \frac{1}{x}dx = \int \sin t dt.$$

Hence

$$\ln x = -\cos t + c,$$

and using the initial condition

$$x(t) = x_0 e^{1-\cos t}.$$

The function f is preserved along the characteristic base curves

$$f(x(t),t) = f_0(x_0), \quad x_0 = x(t)e^{-1+\cos t}.$$

Since we know that

$$f_0(x_0) = 1 + \frac{1}{1+x_0^2},$$

we find that

$$f(x,t) = 1 + \frac{1}{1+x^2e^{-2+2\cos t}},$$

The characteristic base curves and the solution $f(x,t)$ are illustrated in Figure 8.4.

Back to nonlinear scalar conservation laws The method of characteristics can be used for nonlinear conservation equations like our traffic model.

Consider the nonlinear scalar conservation law given by

$$\frac{\partial \rho}{\partial t} + J'(\rho)\frac{\partial \rho}{\partial x} = 0, \tag{8.54}$$

with characteristic base curves such that

$$\frac{dx}{dt} = J'(\rho(x,t)), \quad x(0) = x_0.$$

Assuming a smooth enough ρ, we get a solution $x(t)$ and can rewrite the conservation law as

$$\frac{d}{dt}\rho(x(t),t) = 0.$$

Thus as before, $\rho(x(t),t) = \rho_0(x_0)$, meaning that ρ is contant along characteristic curves. The corresponding characteristic ODE is

$$\frac{dx}{dt} = J'(\rho(x(t),t)) = J'(\rho_0(x_0)),$$

and since $\rho_0(x_0)$ is constant we can integrate to get

$$x(t) = x_0 + J'(\rho(x_0))t.$$

In conclusion, the PDE (8.54) has characteristic base curves that are straight lines and explicitly computable.

Each line has a slope of $[J'(\rho(x_0))]^{-1}$ corresponding to a propagation speed for the density of $J'(\rho(x_0))$.

8.4 Traffic flow when the light turns green

We want to study what happens to the traffic flow when a light turns green. That is, there is a (red) traffic light at $x = 0$, where cars are stopped and standing bumper to bumper behind the traffic light (for $x \leq 0$), the road ahead of the light is empty and the light turns green at time 0. Mathematically at time zero we have

$$\rho_0(x) = \begin{cases} \rho_{\max}\,, & x \leq 0 \\ 0\,, & x > 0\,. \end{cases}$$

For simplicity we assume the normalisation $\rho_{\max} = 1$ and can write

$$J(\rho) = \begin{cases} \rho(1-\rho)\,, & \rho \in [0,1] \\ 0\,, & \rho > 1\,. \end{cases} \tag{8.55}$$

At time $t = 0$, the light turns green and thus

$$\rho_0(x) = \begin{cases} 1, & x \leq 0 \\ 0\,, & x > 0\,. \end{cases}$$

The characteristic base lines satisfy

$$x'(t) = J'(\rho_0) = 1 - 2\rho_0 = \begin{cases} -1, & x \leq 0 \\ 1\,, & x > 0\,. \end{cases},$$

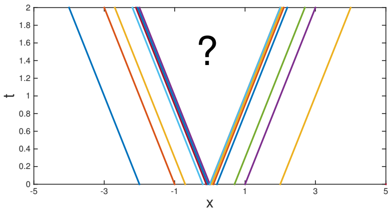


Figure 8.5: Characteristics for $J(\rho)$ from 8.55

and are thus (Figure 8.5)

$$x(t) = \begin{cases} -t + x_0, , & x_0 \leq 0 \\ t + x_0 \,, & x_0 > 0 \,. \end{cases} .$$

The figure reveals a gap region which is not reached by any characteristic base line. There is an inadequacy in our approach: this gap originates from our discontinuous change in density, which we will correct for. In other words, we will now modify the problem in order to have a continuous change in density.

Let us define a modified initial density

$$\rho^\epsilon(x,0) = \begin{cases} 1\,, & x \leq -\epsilon \\ \frac{1}{2} - \frac{x}{2\epsilon}\,, & -\epsilon < x \leq \epsilon \\ 0\,, & x > \epsilon \end{cases} \tag{8.56}$$

The characteristic base curves satisfy

$$x'(t) = J'(\rho_0^\epsilon(x_0)) = \begin{cases} -1\,, & x_0 \leq -\epsilon \\ \frac{x_0}{\epsilon}\,, & -\epsilon < x_0 \leq \epsilon \\ 1\,, & x_0 > \epsilon \end{cases}$$

The set of these characteristic base curves is known as the rarefaction fan and is shown in Figure 8.6.

Now in the transition region we get

$$x(t) = (1 - 2\rho_0)(x_0)t + x_0 = 1 - 2(\frac{1}{2} - \frac{x_0}{2\epsilon})t + x_0 = x_0\left(1 + \frac{t}{\epsilon}\right).$$

Solving for x_0 we get

$$x_0 = \frac{x}{1 + t/\epsilon}.$$

With x_0 we get the density in the transition zone

$$\rho^\epsilon(x,t) = \rho_0^\epsilon(x_0) = \frac{1}{2} - \frac{x}{2\epsilon(1 + t/\epsilon)} = \frac{1}{2}\left(1 - \frac{x}{\epsilon + t}\right),$$

for which we can take the limit as $\epsilon \to 0$

$$\lim_{\epsilon \to 0} \rho^\epsilon(x,t) = \rho(x,t) = \frac{1}{2}\left(1 - \frac{x}{t}\right), \quad t > 0, \tag{8.57}$$

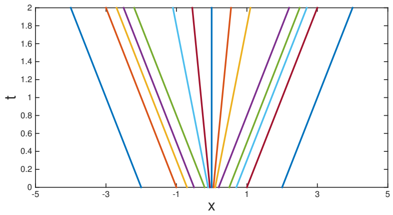


Figure 8.6: The expansion fan fills in the gap of Figure 8.5

where t is taken as a fixed value as shown in figure

As can be seen in the figure, the limit function is linear in x in the rarefaction fan. For $x = -t$, we have $\rho = 1$ as it should be. For $x = t$, we have $\rho = 0$ as we should, and there is a smooth variation in between.

This function tells us how the density varies smoothly from 1 (or $\rho_{\max}$) to 0 as the cars accelerate when the light turns green.

8.5 Some properties of traffic flow from a red light

We will now ask questions associated with motion from a red light. The questions are, how long do you have to wait before you start moving? What is the path of your car once you begin to move? How close do you have to be to go through the light in one cycle?

To understand the answers better, it is easier to restore units for the next analyses. Therefore we go back to $J(\rho) = \rho v_{\max}\left(1 - \frac{\rho}{\rho_{\max}}\right)$.

Waiting time The left end of the expansion fan is the traffic wave corresponding to $\rho_{\max}$ and has a velocity of $-v_{\max}$. If we consider a car at a distance D behind the light, the time until the wave reaches the car is thus $t = D/v_{\max}$. In city traffic $v_{\max}$ might be 50 km/h. If we assume a car spacing of 6 m, the waiting time per car is $t = (6/1000).(3600/50) = .43$ seconds. However, typically large human reaction time has to be added to that value.

Vehicle path Once the car located at a distance D encounters the traffic wave (which moves with velocity $v_{\max}$), it will begin to move. The car's movement after that is completely independent of the traffic wave and can be calculated. The density in the expansion fan is

$$\rho(x,t) = \rho_{\max}\left(\frac{v_{\max}t - x}{2v_{\max}t}\right). \tag{8.58}$$

We already know that the velocity of the car is $v = v_{\max}\left(1 - \rho/\rho_{\max}\right)$. Inserting (8.58) in this expression, we obtain the velocity of a car as a function of x and t. That gives us

$$\frac{dx}{dt} = v_{\max}\left(1 - \frac{1}{\rho_{\max}}\left[\rho_{\max}\frac{v_{\max}t - x}{2v_{\max}t}\right]\right). \tag{8.59}$$

By integrating (8.59), we get the path of a car after it starts moving.

Which cars get through Let us say that the light stays green for a time t_G. The last car to get through the light is the one starting from a distance D_{last} such that its position at time t_G is $x_{last}(t_G) = 0$.

Problems

Exercise 8.7. *Simulate the Markovian approximation (8.4) of the DDE (8.3) for pedestrians in Matlab using forward Euler.*

a) Obtain a rough estimate of the equilibrium parameters $\lambda = k/m$ and $a = a_{j+1}$ from Figure 8.2. Use reasonable assumptions about the size of adults and the legroom necessary for walking to estimate d and l.

b) Simulate (8.4) starting from random initial conditions $x_1(0), \dots, x_N(0)$, with $N = 10$ and various different choices of the motion $x_1(t)$ of the first pedestrian in the row (e.g. constant speed, slowing down). Let it run sufficiently long and explain your observation.

Think about a sensible visualization of the simulation data.

Exercise 8.8. *Simulate the DDE (8.34)–(8.36) for the perturbation displacement, with the parameters $v^* = 28\,\mathrm{m/s}$, $\rho_{\max} = 40$ cars per kilometer, $t_b = 1\,\mathrm{s}$, $k = 0.2\,\mathrm{s}^{-1}$ and $d = 19\,\mathrm{m}$, and the braking law*

$$b(t) = kt\exp((t_b - t)/t_b)\,.$$

a) Implement an iterative solver in Matlab using forward Euler, with time step $h = 0.01\,\mathrm{s}$ and $\tau = Rh$, $R \in \mathbb{N}$ being a multiple of the time step.

b) Simulate the dynamics and estimate the maximum reaction time before an accident occurs when the number of cars in the platoon is $N = 40$. Think about a sensible data visualization.

Exercise 8.9. *The macroscopic density of N vehicles at x at time t is defined as*

$$\rho_{N,s}(x,t) = \frac{\#\ \textit{vehicles in}\ (x-s, x+s)\ \textit{at time}\ t}{2s}\,. \tag{8.60}$$

Consider a situation where all N vehicles are at equal distance at any time. Show that then $\rho_{N,s}$ is independent of N and s and given by

$$\rho(x,t) = |x_{j+1}(t) - x_j(t)|^{-1}\,.$$

Why is the right-hand side of this equation independent of j?

Exercise 8.10. *Compute the density ρ^* that maximizes the flux (8.15). Read off the values for ρ_{crit}, $v_{\max}$ and $\rho_{\max}$ from Figure (7.2) and plot $J(\rho)$ against $v(\rho)$, using (8.15). Explain your findings. For what density does the maximum flux occur if $\rho^* > \rho_{\mathrm{crit}}$? And if $\rho^* < \rho_{\mathrm{crit}}$? If the density is $\rho_{\max}/e$, what is the approximate distance between cars? Would drivers want to drive the speed limit then?*

9 Formal justice

Mathematical tools & concepts: functional equations
Suggested reference: [Ill90]

In this section, we will give a mathematical analysis of the Aristotelian dictum that being just means treating equals equally and unequals unequally. The formal concept of justice means the consistent and continuous application of the same norms and rules to each and every member of a population. Mathematically, formal justice means that that there is a *relation* between, e.g., merit and qualification and the compensation that professionals receive for their work.

9.1 Functional equations

A functional equation is an equation that, just like a differential equation, defines a function in implicit form. Here these implicit function definitions follow from abstract considerations about equality and inequality. The Aristotelian idea of proportionality, that says that any member of a population must be treated proportionately according to merit or excellence, refers precisely to this kind of formal concept of equality and inequality.

In more mathematical terms, this means that, if we can measure or express, e.g., merit by a single non-negative real number x, then then compensation should be a function $m\colon [0,\infty) \to [0,\infty)$ of x. If we agree with Aristotle that compensation (e.g., wage) and merit (e.g., professional qualification) should be proportional, then the function m should satisfy the functional equation

$$\frac{x}{y} = \frac{m(x)}{m(y)}\,, \quad x, y \geq 0\,. \tag{9.1}$$

It is easy to see that (9.1) has the solution

$$m(x) = cx \tag{9.2}$$

for some constant $c \geq 0$, where the positivity follows from the fact that m takes only positive (non-negative) values. To see that $m(x)$ is indeed proportional to x, set $y = 1$ which implies that $m(x) = m(1)x$. As we will show below, this is in fact the *only* solution of the functional equation (9.1).

Formalizing justice Aristotle's concept of proportional justice is rather restrictive and has obvious shortcomings. For example, when applied to crime and punishment it implies that someone who has robbed 10 banks should receive two times the punishment of someone who has robbed "only" 5 banks.[19]

However let us stick to the problem of just wages. Following [Ill90], we shall call a wage system *formally just* if

a) there is a (mathematical) relation between the group of people involved and the set of possible wages,

b) the system is *reliable*, in that it is invariant with respect to the user,

[19] Think about other examples and discuss the shortcomings of the concept of formal justice.

c) the system is *accurate*, i.e., relations between objects (people) are reflected by the relations between the assigned numbers (wages), and

d) there is an accurate inverse in the sense that for any two different wages there exit well-defined prototypes of people who qualify for these wages.

Aristotle's concept of proportional justice meets the first two requirements. Translating the third requirement, condition c), into mathematics, it entails that the function m should be a homomorphism, i.e. a structure-preserving map between relations between people an wage relations. This homomorphism, by the condition d) on its inverse, is specified to be invertible, hence an isomorphism. We stipulate that the function m should be a homomorphism with respect to to the ratio scale, in other words, we measure relations between two people with qualification $x > 0$ and $y > 0$ by the ratio x/y. This is to say that we seek a function $m\colon [0,\infty) \to [0,\infty)$ that satisfies the equation

$$m\left(\frac{x}{y}\right) = \frac{m(x)}{m(y)}\,, \quad x, y \geq 0\,. \tag{9.3}$$

Theorem 9.1. *Let $m\colon [0,\infty) \to [0,\infty)$ satisfy the functional equation*

$$m\left(\frac{x}{y}\right) = \frac{m(x)}{m(y)}\,, \quad x, y \geq 0\,.$$

If m is continuous for some $z \in [0,\infty)$, then m is of the form

$$m(x) = x^s\,, \quad s \in \mathbb{R}\,.$$

Proof. It is clear that $m(x) = x^s$ solves (9.3), however we have to prove the converse statement, namely that *all* solutions of (9.3) that have a point of continuity are of the form $m(x) = x^s$. The proof proceeds in three steps:

1. We first observe that (9.3) is equivalent to

$$m(xy) = m(x)m(y)\,, \quad x, y \geq 0\,, \tag{9.4}$$

which follows from the fact that

$$m(xy) = m\left(\frac{x}{1/y}\right) = \frac{m(x)}{m(1/y)} = m(x)\frac{m(y)}{m(1)}\,,$$

with

$$m(1) = m\left(\frac{x}{x}\right) = \frac{m(x)}{m(x)} = 1\,.$$

2. Define the function $h\colon \mathbb{R} \to \mathbb{R}$ by $h(u) = \log(m(e^u))$. This function solves Cauchy's functional equation

$$h(u+v) = h(u) + h(v)\,, \tag{9.5}$$

as can be seen by noting that

$$\begin{aligned} h(u+v) &= \log(m(e^u e^v)) \\ &= \log(m(e^u)m(e^v)) \\ &= \log(m(e^u)) + \log(m(e^v)) \\ &= h(u) + h(v)\,. \end{aligned}$$

Obviously, (9.5) is satisfied by any linear function

$$h(u) = cu\,, \quad c \in \mathbb{R}\,. \tag{9.6}$$

We will show that Cauchy's equation has no other solutions. By induction, $h(qu) = qh(u)$ for all $q \in \mathbb{Z}$; in particular $h(q) = qh(1)$, and we set $c = h(1)$. Then, with $h(1) = h(p/p) = ph(1/p)$, $p \in \mathbb{N}$, it follows that

$$h\left(\frac{q}{p}\right) = qh\left(\frac{1}{p}\right) = c\frac{q}{p}\,, \quad q \in \mathbb{Z},\, p \in \mathbb{N}\,,$$

which proves that (9.6) holds for all rational arguments $u = q/p$. By the above assumptions m is continuous at $z \geq 0$, as a consequence h is continuous at $w = \log z$. But then, since

$$h(\epsilon) = h(w + \epsilon - w) = h(w + \epsilon) - h(w) \to 0$$

as $\epsilon \to 0$, we can conclude that h is continuous at $u = 0$. Iterating the last argument, it follows that h is continuous everywhere on its domain. Hence all solutions of Cauchy's functional equation (9.5) that have at least one point of continuity are linear functions of the form (9.6).

3. All we have to do now is to invert (9.6) which, using that $x = e^u$ gives

$$m(x) = \exp(h(\log x)) = \exp(\log(x^c)) = x^c\,.$$

But since $c \in \mathbb{R}$ is arbitrary, the assertion follows.

□

Remark 9.2. *There are other solutions to Cauchy's functional equation (9.5), but they represent pathological cases; in particular they are nowhere continuous; for details we refer to [Kuc09].*

Testing whether wages are formally just The solution to the functional equation does not say anything about how the wages should grow with the qualification of an employee, because (9.3) does not tell us what s is (or what it should be in an ideal world). Nevertheless we can use it to test whether a wage system is consistent, i.e. wether all employees in a company or in a country receive payment that follows the same "law". Accordingly (9.3) can be used as a means for decision making, for example, when negotiating the salary with a potential future employee or when considering to cap bankers' bonuses by law.

We suppose that a fair wage scale is given by

$$m(x) = rx^s\,, \tag{9.7}$$

with $r > 0$ being a scaling factor that accounts for, e.g., the currency in which wages are paid; cf. (9.9) below. Figure 9.1 shows possible qualification-wage curves for different values of s. Note that the function $m(x) = rx^s$ with $s < 0$ mets the requirement of formal justice, however paying the more qualified candidate the lower salary does not appear to be just by any sensible standard.

As an illustration of how our model of formal justice can be used to test wage scales consider the situation of three employees working for company X, who

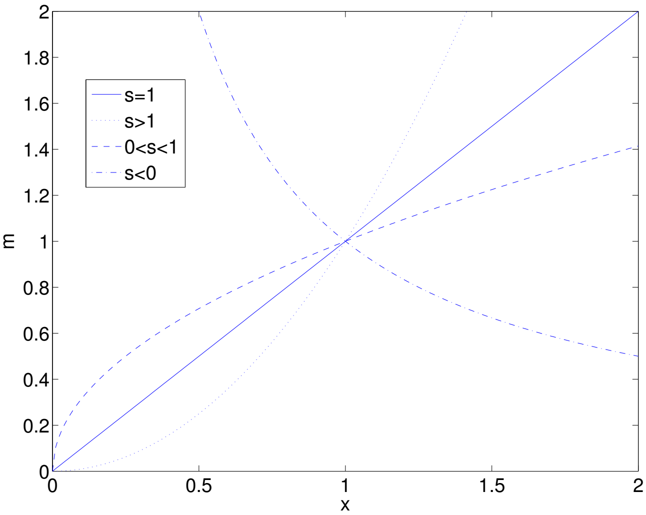


Figure 9.1: Continuous solutions $m(x) = x^s$ of the functional equation $m(x/y) = m(x)/m(y)$ for various values of $s \in \mathbb{R}$.

are paid according to their seniority: Alice has been working for her company for 25 years and makes €2,000,000 per year, Bob who joined the company 16 years ago earns €60,000 per year and Carol after only 3 years gets €40,000. We do a least squares fit of the linear model

$$M = sX + b\,, \quad \text{with } M = \log m,\ X = \log x,\ b = \log r\,. \tag{9.8}$$

In a logarithmic scale a fair wage scale is a straight line with slope s and, when the payment is fair, all data points should lie roughly on this straight line.[20] Figure 9.2 that shows the least square fit of the data suggests—not very surprising—that Alice is significantly overpaid, whereas Bob is underpaid.

9.2 Criticism and possible extensions

Note that there is no rigorous (mathematical or other) argument for using ratios, rather than any other relation between people's qualifications and their wages, e.g., differences such as in

$$m(x - y) = m(x) - m(y)\,, \quad x, y \geq 0\,.$$

(You may think of other sensible choices. By the proof of Theorem 9.1, all solutions of this functional equation are linear.) Nonetheless using ratios to compare measurable quantities is a well-established approach in sociology and quantitative research, hence we will stick to the ratio scale; moreover it is in line with the historical notion of proportional justice.

[20] Clearly three points do not give sufficient statistics, but the example is just meant to illustrate the idea.

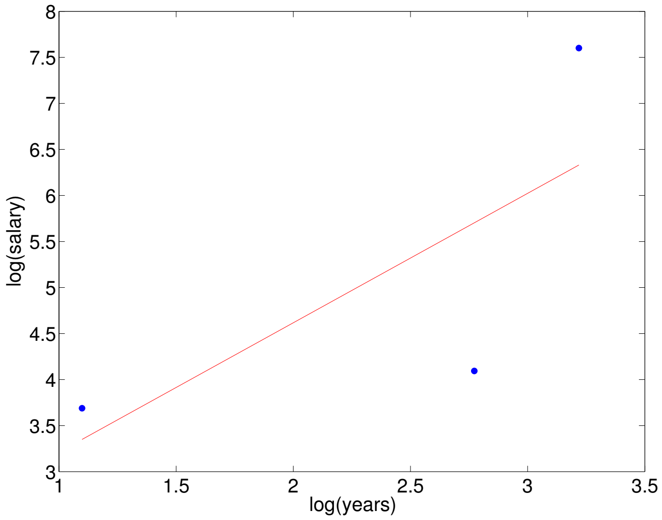


Figure 9.2: Least square fit of a wage system (log scale), with exponent $s \approx 1.4$.

Scaling invariance A more severe drawback of (9.3) is that the equation is not scale invariant. If we change the payment from, say, EUR to CHF, then m scales according to $m \mapsto rm$, with $r > 0$ being the currency exchange rate between Euros and Swiss francs. Specifically, calling $\tilde{m} = rm$, we have

$$\tilde{m}\left(\frac{x}{y}\right) = r\frac{\tilde{m}(x)}{\tilde{m}(y)}\,, \quad x, y \geq 0\,, \tag{9.9}$$

which is different from our original functional equation (9.3). We can, however, account for this lack of scale invariance by simply replacing our model (9.3) by the rescaled model (9.9), which then has solutions of the form

$$\tilde{m}(x) = rx^s\,, \tag{9.10}$$

with $r > 0$ now being a general scale-dependent prefactor.

Formal justice with multiple objectives One may argue that the previous concept of formal justice is "too one-dimensional", in that it tacitly assumes that qualification or merit can be measured by a single parameter. It is more reasonable to assume that the regular payment that an employee receives depends on various independent parameters $x, y, z \ldots$, such as formal degree of education, seniority, extra professional qualifications and so on, which means that the compensation will be a function $m(x, y, z, \ldots)$.

As an example we consider the case $m = m(x, y)$. The idea is that the rule of formal justice, i.e. (9.3) or (9.9) should apply to each qualification measure separately. That is, we require that $m\colon [0,\infty) \times [0,\infty) \to [0,\infty)$ solves the

group	annual salary in £
1	42,803 – 57,520
2	44,971 – 61,901
3	48,505 – 66,623
4	52,131 – 71,701
5	57,520 – 79,081
6	61,901 – 87,229
7	66,623 – 96,166
8	73,480 – 106,148

following system of two coupled functional equations

$$\begin{aligned} m\left(x, \frac{y_1}{y_2}\right) &= r_1(x)\frac{m(x,y_1)}{m(x,y_2)}\,, \quad \forall y_1, y_2 \geq 0 \\ m\left(\frac{x_1}{x_2}, y\right) &= r_2(y)\frac{m(x_1,y)}{m(x_2,y)}\,, \quad \forall x_1, x_2 \geq 0 \end{aligned} \tag{9.11}$$

for each combination of possible qualifications $x, y \geq 0$. Here $r_1(x)$ and $_2(y)$ are qualification dependent scaling factors, similarly the factor r in the modified equation (9.9). It can be shown (see [Ill90] for details) that the only continuous functions m that admit joint representations of the form

$$\begin{aligned} m(x,y) &= r_1(x)y^{u(x)} \\ m(x,y) &= r_2(y)x^{v(y)} \end{aligned} \tag{9.12}$$

are given by

$$m(x,y) = rx^u y^v e^{w \log x \log y}\,, \quad u, v, w \in \mathbb{R}\,. \tag{9.13}$$

Note that on a logarithmic scale, the multi-variable model of formal justice turns into a bilinear model, rather than a linear one:

$$\log m(x,y) = \log r + u \log x + v \log y + w \log x \log y\,, \quad u, v, w \in \mathbb{R}\,. \tag{9.14}$$

Problems

Exercise 9.3. *Let $h\colon \mathbb{R} \to \mathbb{R}$ solve the functional equation $h(t+s) = h(t)+h(s)$ for all $s, t \in \mathbb{R}$ Prove that*

a) $h(-t) = -h(t)$, and

b) $h(t-s) = h(t) - h(s)$ for all $s, t \in \mathbb{R}$.

Exercise 9.4. *Salaries for headteacher in England and Wales (excluding London) range from £42,803 to £106,148 based upon a performance group index that involves, e.g., school leadership, management or pupil progress. The 2013 pay ranges for headteachers are recorded in the following table:*

For comparison, the following table shows the 2003 base salaries of players in the U.S. National Football League, depending on their match experience:

Compare the two salary scales, and explain the rationale behind your comparison. Would you rate any of the above salary scales as fair?
(Hint: Determine the exponent s in the qualification-salary relation $m(x) = cx^s$ by a least square fit of the data given.)

group	annual salary in $
Rookies	225,000
2 yrs	300,000
3 yrs	375,000
4 – 6 yrs	450,000
7 – 9 yrs	655,000
10 yrs	755,00

References

[And11] D.F. Anderson and T.G. Kurtz. Continuous Time Markov Chain Models for Chemical Reaction Networks. In: *Design and Analysis of Biomolecular Circuits*, H. Koeppl, G. Setti, M. di Bernardo, and D. Densmore (eds.), pp. 3–42, Springer, New York, 2011.

[Ari94] R. Aris. *Mathematical Modelling Techniques.* Dover, Mineola, 1994.

[BMR11] A.L. Ballinas-Hernández, A. Muñoz-Meléndez, A. Rangel-Huerta. Multiagent System Applied to the Modeling and Simulation of Pedestrian Traffic in Counterflow. *J. Artif. Soc. Soc. Simulat.* **14**(3), 2, 2011.

[BCD02] N. Bellomo, V. Coscia, M. Delitala. On the Mathematical Theory of Vehicular Traffic Flow I. Fluid Dynamic and Kinetic Modelling. *Math. Mod. Meth. App. Sc.* **12**, 1801–1843, 2002.

[Ben00] E.A. Bender. *An Introduction to Mathematical Modeling.* Dover, Mineola, 2000.

[Bie05] A.A. Biewener. Biomechanical consequences of scaling. *J. Exp. Biol.* **208**, 1665–1676, 2005.

[Buc14] E. Buckingham. On Physically Similar Systems; Illustrations of the Use of Dimensional Equations. *Phys. Rev.* **4**, 345–376, 1914.

[Dou76] P.H. Douglas. The Cobb-Douglas Production Function Once Again: Its History, Its Testing, and Some New Empirical Values. *J. Polit. Econ.* **84**, 903–916, 1976.

[Gil77] D.T. Gillespie. Exact Stochastic Simulation of Coupled Chemical Reactions. *J. Phys. Chem.* **81**, 2340–2361, 1977.

[Hig08] D.J. Higham. Modeling and Simulating Chemical Reactions. *SIAM Review* **50**, 347–368, 2008.

[Ill90] R. Illner. Formal justice and functional equations. *Technical Reports (Mathematics and Statistics), University of Victoria* **DMS-541-IR**, 1990.

[IBM$^+$05] R. Illner, C.S. Bohun, S. McCollum, and T. van Roode. *Mathematical Modelling: A Case Studies Approach.* AMS, Providence, 2005.

[Izh07] E.M. Izhikevich. *Dynamical Systems in Neuroscience: The Geometry of Excitability and Bursting.* MIT Press, Cambridge, 2007.

[KaEn] H. Kaper and H. Engler. *Mathematics and Climate.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, Pennsylvania (2013).

[Kuc09] M. Kuczma. *An introduction to the theory of functional equations and inequalities. Cauchy's equation and Jensen's inequality.* Birkhäuser, Basel, 2009.

[Lot10] A.J. Lotka. Contribution to the Theory of Periodic Reactions. *J. Chem. Phys.* **14**, 271–274, 1910.

[MeG07] M. Mesterton-Gibbons. *A Concrete Approach to Mathematical Modelling.* Wiley, Hoboken, 2007.

[Met87] N. Metropolis. The beginning of the Monte Carlo method. *Los Alamos Science* **15**(584), 125–130, 1987.

[BTF$^+$99] C.R. Rao, H. Toutenburg, A. Fieger, C. Heumann, T. Nittner, and S. Scheid. *Linear Models: Least Squares and Alternatives.* Springer, Berlin, 1999.

[Tay50] G.I. Taylor. The formation of a blast wave by a very intense explosion II: The atomic explosion of 1945. *Proc. Roy. Soc. A* **201**, 175–186, 1950.

[Tes12] G. Teschl. *Ordinary Differential Equations and Dynamical Systems.* AMS, Providence, 2012.

[Vol26] V. Volterra. Variazioni e fluttuazioni del numero d'individui in specie animali conviventi. *Mem. Acad. Lincei Roma* **2**, 31–113, 1926.

[Whi96] P. Whittle. *Optimal control: Basics and beyond.* Wiley & Sons, Chichester, 1996.