Mathematisches Modellieren

Lukas Pottmeyer

10. Juli 2024

Vorwort

Dieses Skript entstand im Laufe der Master-Vorlesung *Mathematisches Modellieren* für das gymnasiale Lehramt an der Universität Duisburg-Essen im Sommersemester 2024. Da ich nur wenig Zeit für die Fertigstellung des Skriptes hatte, wird es von Fehlern aller Art nur so wimmeln. Wer Fehler entdeckt, kann mich sehr gerne per Mail an

lukas.pottmeyer@uni-due.de

darauf hinweisen. Weiter fehlt noch die Einbindung von Referenzen (intern und extern). Daher benutzen Sie dieses Skript bitte auf eigenes Risiko.

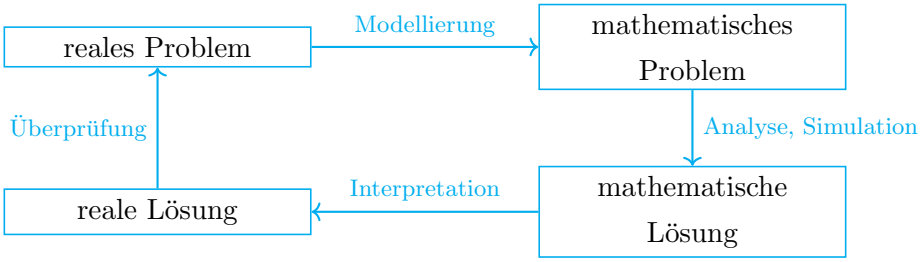
Inhaltsverzeichnis

Einleitung

Wie wird das Wetter morgen in Essen? Wie viele Portionen veganes Essen bereitet die Mensa jeden Tag vor? Wie konnte sich der Coronavirus so schnell verbreiten? Sind meine Daten auf Moodle eigentlich sicher? Was gucke ich als nächstes auf Netflix? Warum muss ich mir diese Fragen durchlesen?
Die Antworten, wie bei sehr vielen Fragen, haben etwas mit Mathematik zu tun. Und um solche Fragestellungen geht es auch in der mathematischen Modellierung. Hier soll versucht werden reale Probleme mathematisch zu beschreiben und, nach Möglichkeit, zu lösen. Ob eine Lösung brauchbar ist, hängt natürlich stark von eigenen Ansprüchen ab. Genügt es, wenn Sie wissen, ob Sie morgen Ihre dicke Winterjacke brauchen, oder möchten Sie die exakte Temperatur, Luftdruck, Niederschlagsmenge, ... pro Minute kennen? Wir wollen also ein mathematisches Modell für den gegebenen Sachverhalt finden. Dieses soll genau genug sein um uns eine brauchbare Antwort zu liefern, aber so einfach dass wir es auch behandeln können. Dabei stoßen wir auf das folgende Bonini-Paradoxon:

> *Nimmt das Maß an Vollständigkeit der komplexe Systeme abbildenden Modelle zu, dann das ihrer Verständlichkeit ab.*

D.h.: Wenn ein Problem mit allen theoretischen Hilfsmitteln modelliert werden soll, dann wird es in der Regel zu komplex um es lösen zu können. Ein grobes Schema sieht folgendermaßen aus:



Wir werden in dieser Vorlesung Modelle mit Differenzen- und Differenzialgleichungen betrachten. Dabei haben wir ein Hauptaugenmerk auf der Beschreibung von Populationsgrößen. Messen wir diese in diskreten Zeitabständen haben wir es mit Differenzen-, anderenfalls mit Differenzialgleichungen zu tun. Dabei betrachten wir insbesondere das Exponentielle und das logistische Wachstum einer Population. Bevor wir uns mit der Interaktion zweier Spezies (die die Größe der jeweils anderen Art beeinflussen) beschäftigen, lernen wir das hilfreiche Konzept de Entdimensionalisierens kennen. Das ermöglicht es Probleme alleine durch eine geeignete Wahl der Einheiten zu vereinfachen. Haben wir dieses Hilfsmittel an der Hand, studieren wir das klassische Räuber-Beute Modell nach Lotka und Volterra, so wie eine Variante in der das logistische Wachstum zugrunde gelegt wird. Damit wir dieses Modell untersuchen können, müssen wir noch einen Abstecher in die Theorie der Differenzialgleichungssysteme machen.

Differenzialgleichungen wurden (hauptsächlich) eingeführt um reale Phänomene zu beschreiben und zu analysieren. Sie wurden kurz nach der Einführung der Analysis durch Newton (1643–1727) und Leibniz (1646–1716) erwähnt. Eine der ersten Erwähnungen betrifft das Problem, welche Kurve durch das hängende Kabel beschrieben wird. Dieses wurde 1690 von Leibniz, Huygens und Jacob Bernoulli gelöst und wir werden es im Kapitel über Entdimensionalisierung kurz behandeln. Die mathematischen Modelle entstehen also bei Differenzialgleichungen „lösungsorientiert“, da es erst das Problem und dann die Mathematik gegeben hat. Mittlerweile werden Differenzialgleichungen neben der angewandten Mathematik natürlich auch Teil der reinen Mathematik studiert.

Danach kommt großer Einschnitt und wir beschäftigen uns im letzten Kapitel mit Kryptographie. Dieses Kapitel kann auch losgelöst von den ersten Kapiteln gelesen werden. Wir beschränken uns auf die klassischen asymmetrischen Kryptosysteme RSA und Elgamal. Am Ende wird kurz auf die Vorteile der Verschlüsselung mit elliptischen Kurven eingegangen. Die Mathematik, die wir im Kapitel zur Kryptographie kennenlernen werden, ist deutlich älter als die Anwendungen, die wir betrachten werden. Primzahlen und elliptische Kurven über endlichen Körpern wurden aus rein mathematischem Interesse studiert. Anwendungen in der realen/greifbaren Welt waren nie angestrebt. Der außermathematische Nutzen wurde also erst nach der

Erforschung bekannt. Die mathematischen Modelle in der Kryptographie sind also eher „problemorientiert“, da es erst die Mathematik gab und dann herausgefunden wurde zu welchem realen Problem diese Mathematik passt. Mittlerweile werden elliptische Kurven neben der reinen Mathematik auch in der angewandten Mathematik studiert.
Dies sind natürlich nur sehr wenige mathematische Modelle, bzw. Methoden die bei Modellierungen eine Rolle spielen. In allem können Sie Mathematik finden! Bereiche, die keinen Platz mehr in dieser Vorlesung gefunden haben, sind unter anderem: Stochastik, dynamische Systeme, Graphentheorie, Kombinatorik, Optimierung, ...
Wir folgen in den ersten Kapiteln grob (und nur in Auszügen) dem Buch [1] von Sebastian Bauer. Ein schönes Buch zur Kryptographie aus dem ich mich oft bedient habe ist [3]. Andere Themen, die auch die gerade genannten Bereiche der Mathematik behandeln, finden Sie unter anderem in den Büchern [2] und [4].

Kapitel 1

Drei erste Beispiele

Wir betrachten in diesem Kapitel erste Beispiele.

1.1 Metallfeder

Problem 1.1.1. Aus einem gegebenen Draht soll eine Metallfeder (z.B. ein Stoßdämpfer) hergestellt werden. Die Feder soll einen Durchmesser von 10cm und eine Höhe von 15cm haben. Auf dieser Höhe soll der Draht gleichmäßig fünf Umdrehungen beschreiben.
Wie lang muss der Draht dafür sein?

Wir möchten die Frage beantworten in dem wir ein mathematisches Modell der Metallfeder erstellen. Dazu sollten wir als erstes geeignete Koordinaten wählen. Da die Feder ein dreidimensionales Objekt ist, arbeiten wir für unser Modell im $\mathbb{R}^3$.

Ziel: Stelle ein Abbild der Feder als Graph einer Funktion φ dar.

In der Aufsicht sieht die Feder aus wie ein Kreis mit Radius 5cm. Wir können das Modell unserer Feder nach Belieben in den $\mathbb{R}^3$ zeichnen. Allerdings sollten wir es so einfach wie möglich halten. Wir wählen unser Modell also so, dass es aus Richtung der z-Koordinate wie ein Kreis vom Radius 5 um den Ursprung aussieht.
In der xy-Ebene ist der Kreis mit Radius 5 um den Ursprung gegeben durch den Graph von

$$\varphi_K : [0, 2\pi] \longrightarrow \mathbb{R}^2 \quad ; \quad t \mapsto (5\cos(t), 5\sin(t)).$$

Da unsere Feder fünf Umdrehungen um den Ursprung machen soll, ist unser gesuchtes φ von der Form

$$\varphi_K : [0, 10\pi] \longrightarrow \mathbb{R}^3 \quad ; \quad t \mapsto (5\cos(t), 5\sin(t), z(t)).$$

Die z-Koordinate beschreibt gerade die Höhe der Feder. Diese soll gleichmäßig (also linear) von 0 auf 15 steigen. Es folgt:

$$\varphi_K : [0, 10\pi] \longrightarrow \mathbb{R}^3 \quad ; \quad t \mapsto (5\cos(t), 5\sin(t), \frac{15}{10\pi} \cdot t). \tag{1.1}$$

Dieses φ ist das „mathematische Modell“ unserer Feder! Wir haben gefragt wie lang der Draht ist, aus dem die Feder hergestellt wurde. Die Übertragung auf das mathematische Modell lautet also:

Wie lang ist der Graph von φ aus (1.1)?

Hier lohnt es sich etwas weiter auszuholen und die Situation allgemeiner zu betrachten. Wir machen also einen Abstecher in die Mathematik um zu klären, wie wir allgemeiner die Länge eines Graphen berechnen können.

Einschub Mathematik

Seien $n \in \mathbb{N}$ und $a, b \in \mathbb{R}$, mit $n \geq 2$ und $a < b$. Sei weiter $f : [a, b] \longrightarrow \mathbb{R}^n$ stetig und differenzierbar auf (a, b). Den Abstand von zwei Punkten im $\mathbb{R}^n$ können wir mit der euklidischen Norm $\| \cdot \|_2$ bestimmen. Diese soll uns auch helfen, die Länge des Graphen von f zu berechnen. Wir schreiben $f = \begin{pmatrix} f_1 \\ \vdots \\ f_n \end{pmatrix}$. Die Idee ist folgende:
Wir unterteilen das Intervall $[a, b]$ in k gleich große Teilintervalle $[a_k(0), a_k(1)]$, $[a_k(1), a_k(2)]$, …, $[a_k(k-1), a_k(k)]$. Dann ist

$$a_k(i) = \frac{b-a}{k} \cdot i + a \qquad \text{für alle } i \in \{0, \ldots, k\}.$$

Für großes k, sollte dann der Wert

$$\sum_{i=0}^{k-1} \|f(a_k(i)) - f(a_k(i+1))\|_2 = \sum_{i=0}^{k-1} \sqrt{\sum_{j=1}^{n} \left(f_j(a_k(i)) - f_j(a_k(i+1))\right)^2} \tag{1.2}$$

eine gute Abschätzung für die Länge des Graphen sein. Um mehr als nur eine Abschätzung zu bekommen, lassen wir k gegen unendlich streben.

Definition 1.1.2. Mit den gerade eingeführten Notationen ist die *Länge* vom Graph von f gegeben durch

$$L(f) = \lim_{k \longrightarrow \infty} \sum_{i=0}^{k-1} \|f(a_k(i)) - f(a_k(i+1))\|_2.$$

Wir betrachten noch einmal Gleichung (1.2) für festes k. Nach dem Mittelwertsatz der Differenzialrechnung existiert für jedes Tupel $(i,j) \in \{0,\dots,k-1\} \times \{1,\dots,n\}$ ein $c_{i,j} \in (a_k(i), a_k(i+1))$, mit

$$f_j(a_k(i)) - f_j(a_k(i+1)) = f_j'(c_{i,j}) \cdot (a_k(i+1) - a_k(i)) = f_j'(c_{i,j}) \cdot \frac{b-a}{k}.$$

Es folgt

$$\begin{aligned} L(f) &= \lim_{k \longrightarrow \infty} \sum_{i=0}^{k-1} \|f(a_k(i)) - f(a_k(i+1))\|_2 \\ &= \lim_{k \to \infty} \sum_{i=0}^{k-1} \frac{b-a}{k} \cdot \sqrt{\sum_{j=1}^{n} \left(f_j'(c_{i,j})\right)^2}, \end{aligned}$$

für ein $c_{i,j} \in (a_k(i), a_k(i+1))$. Da $a_k(i+1) - a_k(i) = \frac{b-a}{k}$, handelt es sich hierbei um eine Riemann-Summe. Aus der Analysis 1 folgt nun

$$L(f) = \int_a^b \sqrt{\sum_{j=1}^{n} f_j'(t)^2} \mathrm{d}t. \tag{1.3}$$

Das halten wir für später fest.

Theorem 1.1.3. *Seien $n \in \mathbb{N}$ und $a, b \in \mathbb{R}$, mit $n \geq 2$ und $a < b$. Sei weiter $f = \begin{pmatrix} f_1 \\ \vdots \\ f_n \end{pmatrix} : [a,b] \longrightarrow \mathbb{R}^n$ stetig und differenzierbar auf (a,b). Dann ist die Länge des Graphen von f gegeben durch $L(f) = \int_a^b \sqrt{\sum_{j=1}^{n} f_j'(t)^2} \mathrm{d}t$.*

Zurück zur Metallfeder

Es bleibt noch die Länge unseres Modells φ aus (1.1) zu berechnen. Dazu

sind wir Dank Theorem 1.1.3 in der Lage. Es ist

$$\begin{aligned} L(\varphi) &= \int_0^{10\pi} \sqrt{(-5\sin(t))^2 + (5\cos(t))^2 + (\frac{3}{2\pi})^2}\mathrm{d}t \\ &= \int_0^{10\pi} \sqrt{25 \cdot (\sin(t)^2 + \cos(t)^2) + \frac{9}{4\pi^2}}\mathrm{d}t \\ &= \int_0^{10\pi} \sqrt{25 + \frac{9}{4\pi^2}}\mathrm{d}t = 10\pi \cdot \sqrt{25 + \frac{9}{4\pi^2}} = 157,7942\ldots \end{aligned}$$

Antwort 1.1.4. Alle Angaben wurden in cm gemacht. Der Draht für die Metallfeder muss also $157,7942\ldots$cm lang sein.

Dieses Ergebnis lässt sich auch anschaulich erklären: Wir benötigen in der xy-Ebene 5 Umrundungen eines Kreises mit Radius 5 cm. Das macht eine Länge von $2\pi \cdot 5 \cdot 5$ cm. Dabei legen wir aber noch eine Höhe von 15 cm zurück. Ziehen wir den Draht also in die Länge, erhalten wir ihn als Hypothenuse eines rechtwinkligen Dreiecks, mit Seitenlängen $2\pi \cdot 5 \cdot 5$ cm und 15 cm. Der Satz des Pythagoras liefert nun die Länge $\sqrt{(2\pi \cdot 5 \cdot 5)^2 + 15^2}$ cm = $157,7942\ldots$ cm. Auf diesem Weg erhalten wir aber nur eine Lösung für genau dieses Problem. Durch unseren allgemeineren Ansatz, können wir nun auch Längen von anderen Kurven als diesen Spiralen berechnen.

Bemerkung 1.1.5. Bei unserem Modell haben wir die Dicke des Drahtes nicht berücksichtigt und angenommen, dass sich die Länge beim Verformen nicht verändert. Das ist vertretbar solange der Draht „dünn“ ist.

Welche Genauigkeit für die Länge angemessen ist, muss der Hersteller entscheiden. Es ist sicher nicht möglich die exakten $10\pi \cdot \sqrt{25 + \frac{9}{4\pi^2}}$ cm abzumessen.

Theorem 1.1.3 kann auf sehr komplizierte Integrale führen. Wenn am Ende ein gerundeter Wert genügt, kann man auch direkt Formel (1.2) für ein großes k benutzen. In unserem Beispiel führt die Wahl von $k = 500$ auf den Wert $157,768\ldots$, was immerhin bis auf $0,4$ Millimeter genau ist. Dies nennen wir eine *numerische* Lösung. Diese kommen immer dann ins Spiel, wenn die exakte mathematische Lösung entweder nicht berechnet werden kann oder nicht gebraucht wird.

1.2 Zinsen

Problem 1.2.1. Auf einem Konto liegen 1000 € Startkapital. Das Konto wird jährlich mit 1% verzinst. Wie viele Euro sind nach n Jahren auf dem Konto?

Der Kontostand ändert sich immer nur zum Jahreswechsel. Das Problem ist also *diskret*. Für die Modellierung gehen wir ein paar Punkte durch.

(a) Namen verteilen: Sei a_n der Betrag auf dem Konto nach n Jahren.

(b) Informationen sammeln: Es ist $a_0 = 1000$, $a_1 = 1000 \cdot \frac{101}{100}$, ...

(c) Beobachtungen verallgemeinern: Für jedes $n \in \mathbb{N}$ gilt $a_n = a_{n-1} \cdot \frac{101}{100}$.

Hier haben wir eine Folge von Zahlen, die *rekursiv* definiert ist. D.h. Die Folgenglieder, die wir schon kennen, liefern uns das nächste Folgenglied. Wenn wir an a_{20} interessiert sind, wollen wir aber nach Möglichkeit nicht alle 19 Folgenglieder davor ausrechnen. Wir suchen also eine Formel, die nur von n abhängt um a_n auszurechnen.

(d) Lösen: Wir erhalten ein a_n in dem wir a_{n-1} mit $\frac{101}{100}$ multiplizieren. Weiter ist $a_0 = 1000$. Wir erhalten also a_n, wenn wir die 1000 genau n-mal mit $\frac{101}{100}$ multiplizieren. Die gesuchte Formel sollte also $a_n = 1000 \cdot \left(\frac{101}{100}\right)^n$ sein. Dies beweisen wir schnell per Induktion.

Induktionsanfang: $\boxed{n = 0}$ Für $n = 0$ stimmt die Formel, denn $a_0 = 1000 \cdot \left(\frac{101}{100}\right)^0 = 1000$.

Induktionsvoraussetzung: Für beliebiges aber festes $n \in \mathbb{N}_0$ gelte $a_n = 1000 \cdot \left(\frac{101}{100}\right)^n$.

Induktionsschritt: $\boxed{n \to n+1}$ Sei n wie in der Voraussetzung. Wir müssen zeigen, dass die postulierte Formel auch für $n+1$ gilt. Dazu rechnen wir

$$a_{n+1} = a_n \cdot \left(\frac{101}{100}\right) \overset{\text{IV}}{=} 1000 \cdot \left(\frac{101}{100}\right)^n \cdot \frac{101}{100} = 1000 \cdot \left(\frac{101}{100}\right)^{n+1}.$$

Damit ist der Induktionsschritt erledigt und die Formel bewiesen.

(e) Auf das reale Problem übertragen: Nach unserem Modell befinden sich nach n Jahren $1000 \cdot \left(\frac{101}{100}\right)^n$ € auf dem Konto. Das ist aber nicht ganz korrekt, denn: Der Kontostand wird nur in Cent angegeben. Damit gibt es

maximal zwei Nachkommastellen. In unserem Modell beträgt der Kontostand nach drei Jahren $1000 \cdot \left(\frac{101}{100}\right)^3 = 1030,301$ €. Tatsächlich sind es $1030,3$ €.

Einschub Mathematik

Wir haben in diesem einfachen Beispiel ein wichtiges mathematisches Objekt kennengelernt, das wir hier nun formal definieren werden.

Definition 1.2.2. Sei $k \in \mathbb{N}$ und $f : \mathbb{R}^k \times \mathbb{N}_0 \longrightarrow \mathbb{R}$ eine Funktion. Eine reelle *Differenzengleichung der Ordnung k* in expliziter Form, ist eine rekursiv definierte Folge $(a_n)_{n\in\mathbb{N}0}$, für die gilt

$$a_n = f(a_{n-1}, \ldots, a_{n-k}, n) \quad \text{für alle } n \geq k.$$

Die Werte $a_0, \ldots, a_{k-1}$ heißen *Anfangswerte* der Differenzengleichung.

Definition 1.2.3. Sei $(a_n)_{n\in\mathbb{N}_0}$ eine Differenzengleichung. Eine *geschlossene Formel* der Differenzengleichung ist eine explizit berechenbare Funktion $g : \mathbb{N}_0 \longrightarrow \mathbb{R}$, so dass $a_n = g(n)$ für alle $n \in \mathbb{N}_0$ gilt.

Beispiel 1.2.4. In unserem Beispiel der Zinsrechnung haben wir eine Differenzengleichung der Ordnung 1, mit Anfangswert $a_0 = 1000$ betrachtet. Die geschlossenen Formel lautet $g(n) = 1000 \cdot \left(\frac{101}{100}\right)^n$.

1.3 freier Fall

Problem 1.3.1. Wir lassen einen Gegenstand aus $10m$ Höhe auf den Boden fallen. Wie lange ist dieser Gegenstand unterwegs?

Wir stellen hier nur ein sehr einfaches Modell für diese Aufgabe vor in dem wir den Luftwiderstand und einige andere Einflüsse vernachlässigen. Wir studieren den freien Fall in einem Vakuum.
Die „Erdanziehungskraft“, also die Gravitationsbeschleunigung, beträgt in Deutschland etwa $g = 9,81 m/s^2$. Das s steht für Sekunde. Warum wir diese merkwürdige Einheit haben erklärt sich in wenigen Augenblicken.
Diese Gravitationsbeschleunigung sagt uns, dass die Beschleunigung eines Objektes im freien Fall (zumindest im Vakuum) immer gleich g ist. Wir

müssen nun die Beschleunigung mit der zurückgelegten Strecke in Verbindung bringen. Das kennen wir aus der Schule (spätestens aus Analysis I). Es gilt nämlich

- $y(t)$ ist die zum Zeitpunkt t zurückgelegte Strecke. Dann ist
- $y'(t)$ ist die Geschwindigkeit zum Zeitpunkt t und
- $y''(t)$ ist die Beschleunigung zum Zeitpunkt t.

Mit dieser Notation können wir auch formulieren, was wir eigentlich suchen. Wir suchen die positive reelle Zahl t für die $y(t) = 10m$ gilt.
Es muss also in unserem Modell $y''(t) = g$ für alle t gelten. Weiter wissen wir $y(0) = y'(0) = 0$, da sich unser Objekt zum Zeitpunkt $t = 0$ – das ist der Zeitpunkt an dem wir das Objekt fallen lassen – nicht bewegt.
Aus diesen Annahmen folgern wir durch Integrieren

$$y'(t) = tg + C \quad \text{und} \quad y(t) = \frac{1}{2}t^2 g + Ct + C'' \quad \text{für } C, C' \in \mathbb{R}.$$

Setzten wir den bekannten Wert $y'(0) = 0$ ein, erhalten wir $C = 0$. Aus $y(0) = 0$ folgt auch $C' = 0$. Damit ist die gesuchte Funktion $y(t) = \frac{1}{2}t^2 g$. Wir suchen die positive reelle Zahl t, mit

$$\begin{aligned} & 10m = y(t) = \frac{1}{2}t^2 g = \frac{1}{2}t^2 9,81 m/s^2 = t^2 4,905 m/s^2 \\ \iff & 2,038\ldots s^2 = t^2 \\ \iff & \pm 1,4278\ldots s = t \end{aligned}$$

Da wir ein positives t brauchen, ist die Antwort: Das Objekt ist ca. $1,4278$ Sekunden unterwegs bevor es auf dem Boden aufkommt.

Bemerkung 1.3.2. Dieses Modell ist, wie angekündigt, nur bedingt realistisch. Das Modell macht insbesondere keinen Unterschied zwischen den Objekten. Es ist also egal ob wir einen Amboss oder eine Feder fallen lassen. Wir beziehen uns hier also tatsächlich nur auf den Fall im Vakuum, wo alle Objekte gleich schnell fallen. Ein Modell mit Luftwiederstand betrachten wir evtl. später in der Vorlesung.

Einschub Mathematik

In diesem Beispiel haben wir die nächste Klasse von wichtigen mathematischen Objekten kennengelernt. Diese werden uns viel in dieser Vorlesung beschäftigen.

Definition 1.3.3. Sei $k \in \mathbb{N}$ und $f : \mathbb{R}^{k+1} \longrightarrow \mathbb{R}$ eine Funktion. Eine gewöhnliche reelle *Differenzialgleichung der Ordnung* k in expliziter Form, ist eine Gleichung der Form

$$y^{(k)}(t) = f(y(t), y'(t), \ldots, y^{(k-1)}(t), t).$$

Eine *Lösung* der Differenzialgleichung auf einem Intervall I, ist eine k-mal differenzierbare Funktion $y : \mathbb{R} \to \mathbb{R}$, mit

$$y^{(k)}(t) = f(y(t), y'(t), \ldots, y^{(k-1)}(t), t) \quad \text{für alle } t \in I.$$

Für ein gegebenes $t_0 \in \mathbb{R}$ heißen die Werte $y(t_0)$, $y'(t_0)$, ..., $y^{(k-1}(t_0)$ die *Anfangswerte* der Differenzialgleichung.

Beispiel 1.3.4. In unserem Beispiel des freien Falls haben wir die Differenzialgleichung der Ordnung 2, $y''(t) = g$, mit den Anfangswerten $y(0) = 0$ und $y'(0) = 0$, betrachtet.

Beachten Sie die Ähnlichkeit zur Definition 1.2.2 der Differenzengleichungen. Der große Unterschied ist, dass Differenzialgleichungen *kontinuierlich* und nicht mehr *diskret* sind. D.h.: die Veränderungen finden durchgehend statt und nicht nur nach festen Zeitintervallen (wie etwa jährlichen Zinszahlungen). Ein weiterer Unterschied ist die Lösbarkeit der Probleme. Bei Differenzengleichungen gibt es zu gegebenen Anfangswerten immer genau eine Lösung. Bei Differenzialgeichungen kann es auch passieren, dass es keine Lösung oder unendlich viele Lösungen gibt. Das werden wir später noch etwas genauer beleuchten.
Die Verbindung von Differenzen- und Differenzialgleichungen betrachten wir im folgenden Kapitel.

Kapitel 2

Von Differenzen- zu Differenzialgleichungen

Im Beispiel der Zinsrechnung aus dem letzten Kapitel haben wir eine geschlossene Formel für die Differenzengleichung

$$a_0 = 1000 \quad \text{und} \quad a_n = a_{n-1} \cdot \left(\frac{101}{100}\right)$$

gesehen oder *geraten*. Oft ist es allerdings gar nicht möglich eine geschlossene Formel zu finden, geschweige denn zu erraten. In den folgenden Abschnitten werden wir ein Verfahren kennenlernen mit dem man geschlossene Formeln für einfache Differenzengleichungen berechnen kann.

2.1 Erzeugendenfunktionen

Definition 2.1.1. Eine *formale Potenzreihe* über $\mathbb{R}$ ist ein Ausdruck der Form

$$f(x) = \sum_{n=0}^{\infty} a_n x^n \qquad , a_n \in \mathbb{R} \ \forall n \in \mathbb{N}_0.$$

Die Menge aller formaler Potenzreihen über $\mathbb{R}$ bezeichnen wir mit $\mathbb{R}[\![x]\!]$.

Bemerkung 2.1.2. Potenzreihen kennen Sie schon sehr gut aus der Analysis. Wichtig ist bei uns das Wort *formale*. Wir betrachten diese formalen Potenzreihen als abstrakte Objekte. Insbesondere setzen wir nie irgendeinen Wert für die Variable x ein. **Formale Potenzreihen sind keine Abbildungen!** Daher interessiert uns auch nicht irgendein Konvergenzverhalten. Darüber brauchen wir uns hier keine Sorgen zu machen.

Wir nennen zwei formale Potenzriehen *gleich*, wenn alle Koeffizienten gleich sind. D.h.

$$\sum_{n=0}^{\infty} a_n x^n = \sum_{n=0}^{\infty} b_n x^n \iff a_n = b_n \ \forall n \in \mathbb{N}_0.$$

Alternativ schreiben wir auch $\sum_{n\geq 0} a_n x^n$ anstatt $\sum_{n=0}^{\infty} a_n x^n$.

Satz 2.1.3. *Die Menge $\mathbb{R}[\![x]\!]$ ist ein kommutativer nullteilerfreier Ring (mit Eins) bezüglich der Verknüpfungen*

$+$*:* $\sum_{n=0}^{\infty} a_n x^n + \sum_{n=0}^{\infty} b_n x^n = \sum_{n=0}^{\infty}(a_n + b_n)x^n$

$\cdot$*:* $(\sum_{n=0}^{\infty} a_n x^n) \cdot (\sum_{n=0}^{\infty} b_n x^n) = \sum_{n=0}^{\infty} (\sum_{k=0}^{n} a_k b_{n-k}))x^n$

für alle $\sum_{n=0}^{\infty} a_n x^n,\ \sum_{n=0}^{\infty} b_n x^n \in \mathbb{R}[\![x]\!]$.
D.h.: Es gelten auf $\mathbb{R}[\![x]\!]$ die Kommutativ-, Assoziativ- und Distributivgesetze. Weiter folgt aus $f(x)\cdot g(x) = 0$, dass $f(x) = 0$ oder $g(x) = 0$ gilt. Hierbei ist das Nullelement die formale Potenzreihe $0 = \sum_{n=0}^{\infty} a_n x^n$, mit $a_n = 0$ für alle $n \in \mathbb{N}_0$. Das Einselement ist die formale Potenzreihe $1 = \sum_{n=0}^{\infty} a_n x^n$, mit $a_0 = 1$ und $a_n = 0$ für alle $n \in \mathbb{N}$.

Beweis. Das folgt alles aus den entsprechenden Rechenregeln auf $\mathbb{R}$. Wir beweisen hier nur die Kommutativität bezüglich der Multiplikation. Seien dazu $\sum_{n=0}^{\infty} a_n x^n,\ \sum_{n=0}^{\infty} b_n x^n \in \mathbb{R}[\![x]\!]$ beliebig. Dann gilt

$$\begin{aligned}
\left(\sum_{n=0}^{\infty} a_n x^n\right) \cdot \left(\sum_{n=0}^{\infty} b_n x^n\right) &= \sum_{n=0}^{\infty} \left(\sum_{k=0}^{n} a_k b_{n-k}\right) x^n \\
&= \sum_{n=0}^{\infty} \left(\sum_{k=0}^{n} b_{n-k} a_k\right) x^n \\
&\overset{k'=n-k}{=} \sum_{n=0}^{\infty} \left(\sum_{k'=0}^{n} b_{k'} a_{n-k'}\right) x^n \\
&= \left(\sum_{n=0}^{\infty} b_n x^n\right) \cdot \left(\sum_{n=0}^{\infty} a_n x^n\right).
\end{aligned}$$

Das wollten wir zeigen. □

Bemerkung 2.1.4. Wir fassen Polynome über $\mathbb{R}$ vom Grad d als formale Potenzreihen auf, bei denen alle Koeffizienten a_n, mit $n > d$, gleich Null sind.
Gilt für zwei formale Potenzreihen $f(x)$ und $g(x)$ die Gleichung $f(x)\cdot g(x) = 1$, so schreiben wir $f(x) = \frac{1}{g(x)}$.

Beispiel 2.1.5. Es gilt $\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}$.[1] Die Gleichung folgt aus der Rechnung

$$(1-x)\cdot\sum_{n=0}^{\infty} x^n = \sum_{n=0}^{\infty} x^n - x\cdot\sum_{n=0}^{\infty} x^n = 1 + \sum_{n=1}^{\infty} x^n - \sum_{n=0}^{\infty} x^{n+1}$$
$$= 1 + \sum_{n=1}^{\infty} x^n - \sum_{n=1}^{\infty} x^n = 1.$$

Es ist also möglich formale Potenzreihen mit Quotienten von Polynomen zu identifizieren. Da wir mit Polynomen schon lange rechnen können (und einige schöne Rechentricks kennen), ist das ein sehr gutes Hilfsmittel. Im folgenden Theorem finden wir einige wichtige Gleichungen.

Theorem 2.1.6 („Vokabelheft“)**.** *Sei $c \in \mathbb{R}\setminus\{0\}$ und $k \in \mathbb{N}_0$ beliebig. Dann gilt*

(a) $\sum_{n=0}^{\infty} c\cdot x^n = \frac{c}{1-x}$,

(b) $\sum_{n=0}^{\infty} \binom{n+k}{n}\cdot x^n = \frac{1}{(1-x)^{k+1}}$,

(c) $\sum_{n=0}^{\infty} \binom{n+k}{n}\cdot c^n\cdot x^n = \frac{1}{(1-cx)^{k+1}} = \frac{1/c^{k+1}}{(1/c-x)^{k+1}}$.

Beweis. Teil (a) folgt sofort aus Beispiel 2.1.5. Kommen wir also direkt zu Teil (b). Wir müssen zeigen, dass die Formel *für alle* $k \in \mathbb{N}_0$ gilt. Damit drängt sich eine Induktion über k auf.
Induktionsanfang: $\boxed{k=0}$ Für $k=0$ wird die Gleichung zu $\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}$. Dass diese Gleichung stimmt, haben wir schon in Beispiel 2.1.5 eingesehen.
Induktionsvoraussetzung: Für beliebiges aber festes k gelte $\sum_{n=0}^{\infty} \binom{n+k}{n}\cdot x^n = \frac{1}{(1-x)^{k+1}}$.
Induktionsschritt: $\boxed{k \to k+1}$ Sei k wie in der Induktionsvoraussetzung. Dann gilt

$$\frac{1}{(1-x)^{(k+1)+1}} = \frac{1}{(1-x)^{k+1}}\cdot\frac{1}{1-x} \overset{\text{IV\&2.1.5}}{=} \left(\sum_{n=0}^{\infty}\binom{n+k}{n}\cdot x^n\right)\cdot\left(\sum_{n=0}^{\infty} x^n\right)$$
$$= \sum_{n=0}^{\infty}\left(\sum_{j=0}^{n}\binom{j+k}{j}\cdot 1\right)\cdot x^n. \tag{2.1}$$

[1] Das ist die geometrische Reihe von der wir wissen, dass Sie für alle $x \in (-1,1)$ gilt. Wir wollen aber auch weiterhin keine Zahlen in unsere formalen Potenzreihen einsetzen. Daher müssen wir die Gleichung noch einmal formal verifizieren.

Wir müssen noch zeigen, dass

$$\sum_{j=0}^{n} \binom{j+k}{j} \cdot 1 = \binom{n+k+1}{n} \tag{2.2}$$

gilt für alle $n \in \mathbb{N}_0$. Wir sind also schon wieder bei einer Induktion gelandet. Diese handeln wir ganz schnell ab. Der Induktionsanfang ist klar, da für $n = 0$ auf beiden Seiten eine 1 steht. Die Induktionsvoraussetzung ist der übliche Schablonensatz. Kommen wir also zum Induktionsschritt. Es gilt

$$\begin{aligned}\sum_{j=0}^{n+1} \binom{j+k}{j} &= \left(\sum_{j=0}^{n} \binom{j+k}{j}\right) + \binom{n+1+k}{n+1} \\ &\overset{\text{IV}}{=} \binom{n+k+1}{n} + \binom{n+1+k}{n+1} = \binom{n+1+k+1}{n+1}.\end{aligned}$$

Damit ist die innere Induktion erledigt und (2.2) ist bewiesen. Es folgt

$$\frac{1}{(1-x)^{(k+1)+1}} \overset{(2.1)}{=} \sum_{n=0}^{\infty} \left(\sum_{j=0}^{n} \binom{j+k}{j}\right) \cdot x^n \overset{(2.2)}{=} \sum_{n=0}^{\infty} \binom{n+(k+1)}{n} \cdot x^n.$$

Das mussten wir zeigen.
Teil (c) folgt aus Teil (b) in dem wir x durch $c \cdot x$ ersetzen. □

Definition 2.1.7. Sei $(a_n)_{n\in\mathbb{N}_0}$ eine reelle Folge. Dann heißt die formale Potenzreihe $\sum_{n=0}^{\infty} a_n \cdot x^n$ die *Erzeugendenfunktion* von $(a_n)_{n\in\mathbb{N}_0}$.

2.2 Zinsen reloaded

Problem 2.2.1. Sie haben auf Ihrem Konto einen Betrag a. Pro Jahr gibt es $p'\%$ Zinsen und jedes Jahr überweisen Sie einen Betrag b auf Ihr Konto. Wie viel Geld haben Sie nach n Jahren auf dem Konto?

Wir gehen davon aus, dass a und p' positiv sind und gehen genauso vor, wie im ersten Beispiel der Zinsen 1.2.1.

(a) Namen vergeben: Sei a_n der Betrag nach n Jahren auf dem Konto.

(b) Informationen sammeln: Es ist $a_0 = a$ und $a_1 = a_0 \cdot \left(1 + \frac{p'}{100}\right) + b$.

(c) Verallgemeinern: Es ist $a_n = a_{n-1} \cdot \left(1 + \frac{p'}{100}\right) + b$ für alle $n \in \mathbb{N}$ und $a_0 = a$.

Wir setzen $p = \left(1 + \frac{p'}{100}\right)$. Wir müssen also, um das Modell zu lösen, eine geschlossenen Formel für die Differenzengleichung

$$a_n = p \cdot a_{n-1} + b \ \forall n \in \mathbb{N} \quad \text{und} \quad a_0 = a \tag{2.3}$$

finden. Eine solche Formel kann man hier noch durch scharfes Hinsehen erkennen. Wir wollen aber lieber ein Verfahren vorstellen, das ohne Raten einer Lösung auskommt. Dafür studieren wir die Erzeugendenfunktion von $(a_n)_{n\in\mathbb{N}_0}$. Für spätere Referenzen bekommt diese Rechnung eine eigene Nummer.

2.2.2. Es gilt

$$\begin{aligned}\sum_{n=0}^{\infty} a_n x^n &= a_0 + \sum_{n=1}^{\infty} a_n x^n \overset{(2.3)}{=} a + \sum_{n=0}^{\infty} (p a_{n-1} + b) x^n \\ &= a + p \cdot \sum_{n=1}^{\infty} a_{n-1} x^n + b \cdot \sum_{n=1}^{\infty} x^n = a + px \cdot \sum_{n=1}^{\infty} a_{n-1} x^{n-1} + bx \cdot \sum_{n=1}^{\infty} x^{n-1} \\ &= a + px \cdot \sum_{n=0}^{\infty} a_n x^n + bx \cdot \sum_{n=0}^{\infty} x^n \overset{2.1.6}{=} a + px \cdot \sum_{n=0}^{\infty} a_n x^n + \frac{bx}{1-x}.\end{aligned}$$

Wir ziehen auf beiden Seiten $px \cdot \sum_{n=0}^{\infty} a_n x^n$ ab und erhalten

$$(1 - px) \cdot \sum_{n=0}^{\infty} a_n x^n = a + \frac{bx}{1-x}.$$

Daraus folgt

$$\sum_{n=0}^{\infty} a_n x^n = \frac{a}{1-px} + \frac{bx}{(1-x)(1-px)}. \tag{2.4}$$

Wir haben also die Erzeugendenfunktion, für die wir eine geschlossene Formel finden möchten, als Quotient von Polynomen geschrieben. Die Koeffizienten dieses Quotienten können wir nun mit Theorem 2.1.6 bestimmen. Insbesondere wissen wir $\frac{a}{1-px} = \sum_{n=0}^{\infty} p^n x^n$. Um $\frac{bx}{(1-x)(1-px)}$ in eine passende Form zu bringen benutzen wir Partialbruchzerlegung. Beachten Sie, dass $p \neq 1$ ist, nach unserer Annahme ganz am Anfang. Partialbruchzerlegung liefert

$$\frac{bx}{(1-x)(1-px)} = -\frac{b/p-1}{1-x} + \frac{b/p-1}{1-px} \overset{2.1.6}{=} -\frac{b}{p-1} \cdot \sum_{n=0}^{\infty} x^n + \frac{b}{p-1} \cdot \sum_{n=0}^{\infty} p^n x^n.$$

Das setzen wir in (2.4) ein und erhalten

$$\begin{aligned}\sum_{n=0}^{\infty} a_n x^n &= \sum_{n=0}^{\infty} p^n x^n - \frac{b}{p-1} \cdot \sum_{n=0}^{\infty} x^n + \frac{b}{p-1} \cdot \sum_{n=0}^{\infty} p^n x^n \\ &= (a + \frac{b}{p-1}) \cdot \sum_{n=0}^{\infty} p^n x^n - \frac{b}{p-1} \cdot \sum_{n=0}^{\infty} x^n \\ &= \sum_{n=0}^{\infty} \left((a + \frac{b}{p-1}) \cdot p^n - \frac{b}{p-1} \right) x^n.\end{aligned}$$

Koeffizientenvergleich liefert nun

$$a_n = (a + \frac{b}{p-1}) \cdot p^n - \frac{b}{p-1} \ \forall n \in \mathbb{N}_0.$$

Wir erhalten: Nach n Jahren beträgt der Betrag auf dem Konto (ungefähr) $(a + \frac{b}{p-1}) \cdot p^n - \frac{b}{p-1}$. (Wir haben hier nie über die Währung gesprochen, da diese Information für die Modellierung vollkommen unerheblich ist.)
Dass der Betrag nur ungefähr dem entspricht, was wir berechnet haben, ist wieder der fehlenden Rundung auf die kleinste Geldeinheit geschuldet.

2.3 Naives Beispiel einer Kaninchen Population

Das folgende Beispiel stammt von Fibonacci (Leonardo von Pisa).

Problem 2.3.1. Wie vermehren sich Kaninchen, wenn wir mit einem neugeborenen Paar aufangen?

Es werden folgende Annahmen getroffen:

- Jedes Kaninchen braucht einen Monat um geschlechtsreif zu werden.
- jedes Paar geschlechtsreifer Kaninchen bringt pro Monat ein Paar Kaninchen zur Welt
- ein Kaninchenpaar besteht immer aus einem männlichen und einem weiblichen Tier.

Diese Annahmen packen wir nun wieder in ein mathematisches Modell. Dazu sei a_n die Anzahl von Kaninchen Paaren nach n Monaten. Dann gilt $a_0 = 1$ und $a_1 = 1$. Zu Beginn haben wir ein neugeborenes Paar, nach einem Monat ist das immer noch das einzige Paar, aber nun ist es geschlechtsreif. Damit

ist $a_2 = 2$ (ein geschlechtsreifes, ein neugeborenes), $a_3 = 3$ (zwei geschlechtsreife, ein neugeborenes), $a_4 = 5$ (drei geschlechtsreife, zwei neugeborene). Verallgemeinern wir dies, dann erhalten wir $a_n = a_{n-1} + a_{n-2}$ für alle $n \geq 2$, da alle Kaninchenpaare, die es schon vor zwei Monaten gab wieder neue Kaninchenpaare bekommen. Wir haben es also mit folgender Differenzengleichung zu tun

$$a_n = a_{n-1} + a_{n-2} \ \forall n \geq 2 \quad \text{und} \quad a_0 = a_1 = 1. \tag{2.5}$$

Diese Lösen wir mit Hilfe der Erzeugendenfunktion.

$$\begin{aligned}
\sum_{n=0}^{\infty} a_n x^n &= a_0 + a_1 x + \sum_{n=2}^{\infty} a_n x^n \overset{(2.5)}{=} 1 + x + \sum_{n=2}^{\infty} (a_{n-1} + a_{n-2}) x^n \\
&= 1 + x + \sum_{n=2}^{\infty} a_{n-1} x^n + \sum_{n=2}^{\infty} a_{n-2} x^n \\
&= 1 + x + x \cdot \sum_{n=2}^{\infty} a_{n-1} x^{n-1} + x^2 \cdot \sum_{n=2}^{\infty} a_{n-2} x^{n-2} \\
&= 1 + x + x \cdot \sum_{n=1}^{\infty} a_n x^n + x^2 \cdot \sum_{n=0}^{\infty} a_n x^n \\
&= 1 + x + x \cdot \left(\sum_{n=0}^{\infty} a_n x^n - a_0 x^0 \right) + x^2 \cdot \sum_{n=0}^{\infty} a_n x^n \\
&= 1 + x \cdot \sum_{n=0}^{\infty} a_n x^n + x^2 \cdot \sum_{n=0}^{\infty} a_n x^n.
\end{aligned}$$

Es folgt

$$(1 - x - x^2) \sum_{n=0}^{\infty} a_n x^n = 1$$

und somit

$$\sum_{n=0}^{\infty} a_n x^n = \frac{1}{1 - x - x^2} = \frac{-1}{x^2 + x - 1} \tag{2.6}$$

Wieder wollen wir die rechte Seite durch Partialbruchzerlegung in eine Form bringen in der wir Theorem 2.1.6 benutzen können. Dazu berechnen wir zunächst die Nullstellen von $x^2 + x - 1$ und finden heraus, dass

$$x^2 + x - 1 = (x - \frac{-1 - \sqrt{5}}{2})(x - \frac{-1 + \sqrt{5}}{2})$$

gilt. Beachten Sie, dass das Polynom in zwei verschiedene Linearfaktoren zerfällt. Es soll also für gewisse $A, B \in \mathbb{R}$ gelten

$$\frac{-1}{x^2+x-1} = \frac{A}{(x-\frac{-1+\sqrt{5}}{2})} + \frac{B}{(x-\frac{-1-\sqrt{5}}{2})} = \frac{A(x-\frac{-1-\sqrt{5}}{2}) + B(x-\frac{-1+\sqrt{5}}{2})}{x^2+x-1}. \tag{2.7}$$

$$\begin{aligned}
&\iff -1 = (A+B)x + (A\frac{1+\sqrt{5}}{2} + B\frac{1-\sqrt{5}}{2}) \\
&\iff 0 = A+B \quad \text{und} \quad -1 = A\frac{1+\sqrt{5}}{2} + B\frac{1-\sqrt{5}}{2} \\
&\iff A = -B \quad \text{und} \quad -1 = -B\frac{1+\sqrt{5}}{2} + B\frac{1-\sqrt{5}}{2} = B(\frac{-1-\sqrt{5}}{2} + \frac{1-\sqrt{5}}{2}) \\
&\iff A = -B \quad \text{und} \quad -1 = B(-\sqrt{5}) \\
&\iff A = -\frac{1}{\sqrt{5}} \quad \text{und} \quad B = \frac{1}{\sqrt{5}}
\end{aligned}$$

Aus (2.6) und (2.7) folgt nun

$$\sum_{n=0}^{\infty} a_n x^n = -\frac{1}{\sqrt{5}} \frac{1}{(x-\frac{-1+\sqrt{5}}{2})} + \frac{1}{\sqrt{5}} \frac{1}{(x-\frac{-1-\sqrt{5}}{2})}.$$

Wir formen furchtlos weiter um

$$\begin{aligned}
\sum_{n=0}^{\infty} a_n x^n &= \frac{1}{\sqrt{5}} \left(\frac{1}{(\frac{-1+\sqrt{5}}{2} - x)} - \frac{1}{(\frac{-1-\sqrt{5}}{2} - x)} \right) \\
&= \frac{1}{\sqrt{5}} \left(\frac{\frac{2}{-1+\sqrt{5}}}{(1 - \frac{2}{-1+\sqrt{5}}x)} - \frac{\frac{2}{-1-\sqrt{5}}}{(1 - \frac{2}{-1-\sqrt{5}}x)} \right) \\
&\overset{2.1.6}{=} \frac{1}{\sqrt{5}} \left(\frac{2}{-1+\sqrt{5}} \cdot \sum_{n=0}^{\infty} \left(\frac{2}{-1+\sqrt{5}} \right)^n x^n - \frac{2}{-1-\sqrt{5}} \cdot \sum_{n=0}^{\infty} \left(\frac{2}{-1-\sqrt{5}} \right)^n x^n \right) \\
&= \sum_{n=0}^{\infty} \frac{1}{\sqrt{5}} \cdot \left(\left(\frac{2}{-1+\sqrt{5}} \right)^{n+1} - \left(\frac{2}{-1-\sqrt{5}} \right)^{n+1} \right) x^n
\end{aligned}$$

Koeffizientenvergleich der beiden formalen Potenzreihen liefert nun

$$a_n = \frac{1}{\sqrt{5}} \cdot \left(\left(\frac{2}{-1+\sqrt{5}} \right)^{n+1} - \left(\frac{2}{-1-\sqrt{5}} \right)^{n+1} \right) \quad \forall n \in \mathbb{N}_0.$$

Damit sind wir eigentlich fertig. Da jedoch Wurzeln im Nenner etwas unschön aussehen, nutzen wir

$$\frac{2}{-1+\sqrt{5}} = \frac{1+\sqrt{5}}{2} \quad \text{und} \quad \frac{2}{-1-\sqrt{5}} = \frac{1-\sqrt{5}}{2} \tag{2.8}$$

um die folgende geschlossene Formel zu erhalten:

$$a_n = \frac{1}{\sqrt{5}} \cdot \left(\left(\frac{1+\sqrt{5}}{2} \right)^{n+1} - \left(\frac{1-\sqrt{5}}{2} \right)^{n+1} \right) \ \forall n \in \mathbb{N}_0. \tag{2.9}$$

Antwort 2.3.2. Nach n Monaten gibt es unter unseren Annahmen genau $\frac{1}{\sqrt{5}} \cdot \left(\left(\frac{1+\sqrt{5}}{2} \right)^{n+1} - \left(\frac{1-\sqrt{5}}{2} \right)^{n+1} \right)$ Kaninchenpaare.

Bemerkung 2.3.3. Unsere Annahmen sind natürlich nicht realistisch. Denn:

- Es wird nicht berücksichtigt, dass Kaninchen sterbliche Wesen sind!
- Es werden sicher nicht immer genau ein männliches und ein weibliches Tier geboren.
- Es werden keine anderen Einflüsse, wie Platzmangel, Nahrungsangebot, Feinde, ... berücksichtigt.

Definition 2.3.4. Die Folge $(a_n)_{n\in\mathbb{N}_0}$, mit $a_0 = a_1 = 1$ und $a_n = a_{n-1} + a_{n-2}$ Für alle $n \geq 2$ heißt *Fibonacci-Folge*. Die Folgenglieder sind die *Fibonacci-Zahlen*.

Genau wie im Beweis der geschlossenen Formel für die Fibonacci-Zahlen, zeigt man auch die folgende allgemeine Formel.

Theorem 2.3.5. *Seien $c_1, \ldots, c_k \in \mathbb{R}$ und sei $(a_n)_{n\in\mathbb{N}_0}$ gegeben durch die Differenzengleichung*

$$a_n = c_1 a_{n-1} + c_2 a_{n-2} + \ldots + c_k a_{n-k} \ \forall n \geq k.$$

Falls das Polynom $x^k - c_1 x^{k-1} - \ldots - c_k$ über $\mathbb{C}$ in paarweise verschiedene Linearfaktoren $(x - \lambda_1), \ldots, (x - \lambda_k)$ zerfällt, dann ist eine geschlossene Formel der Differenzengleichung gegeben durch

$$a_n = C_1 \lambda_1^n + \ldots + C_k \lambda_k^n \ \forall n \in \mathbb{N}_0$$

für gewisse Konstanten $C_1, \ldots, C_k \in \mathbb{C}$. Diese Konstanten werden durch die Anfangswerte $a_0, \ldots, a_{k-1}$ bestimmt.

2.4 Fibonacci-Zahlen in der Natur

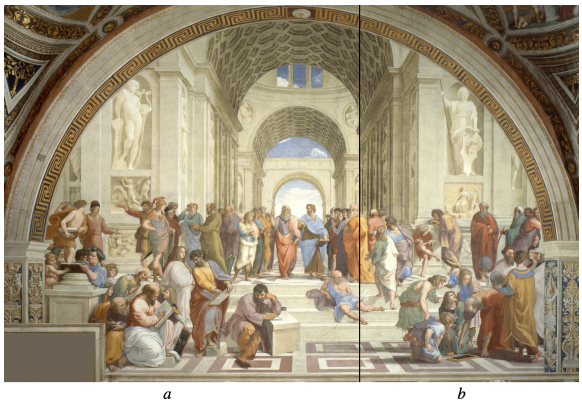
In der geschlossenen Formel für die Fibonacci-Zahlen, ist der Wert $\theta = \frac{1+\sqrt{5}}{2} = 1{,}618\ldots$ aufgetaucht. Wir sind auf diesen Wert gestoßen als Nullstelle des Polynoms $x^2 - x - 1$. Die zweite Nullstelle war $\frac{1-\sqrt{5}}{2} = 1 - \theta < 0$. Es gilt also die Gleichung $\theta^2 - \theta - 1 = 0$. Umstellen dieser Gleichung führt auf zwei hilfreiche Beziehungen:

$$\frac{1}{\theta} = \theta - 1 \quad \text{und} \quad \frac{1}{\theta^2} = \frac{1}{1+\theta} = 1 - \frac{1}{\theta}. \tag{2.10}$$

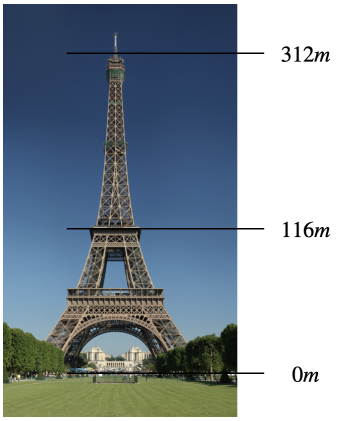
Die Zahl $\theta = \frac{1+\sqrt{5}}{2} = 1{,}618\ldots$ wird auch *goldener Schnitt* genannt. Die Aufteilung eines Bildes oder Gebäudes in zwei Teile der Längen a und b, so dass $\frac{a+b}{a} = \frac{a}{b}$ gilt, wird in der Kunst und der Architektur als besonders ästhetisch angesehen. Eine kurze Umformung der Gleichung liefert

$$\frac{a+b}{a} = \frac{a}{b} \Longleftrightarrow 1 + \frac{1}{a/b} = \frac{a}{b} \Longleftrightarrow \left(\frac{a}{b}\right)^2 - \frac{a}{b} - 1 = 0.$$

Da das Verhältnis $\frac{a}{b}$ positiv sein muss, erhalten wir $\frac{a}{b} = \frac{a+b}{a} = \theta$. Setzen wir $a + b = L$ so folgt, dass damit $a = \frac{L}{\theta}$ gelten muss.



(a) In dem Werk „Die Schule von Athen“ von Raphael kann man den goldenen Schnitt an mehreren Stellen finden. Teilen wir das Bild am rechten Rand des Torbogens, so wird es genau im Verhältnis 1 zu θ – also im goldenen Schnitt – geteilt.



(b) Ursprünglich war der Eiffelturm $312m$ hoch und die zweite Etage befindet sich auf ca. $116m$ Höhe. Diese zweite Etage unterteilt den Eiffelturm also in etwa im goldenen Schnitt.

Teilen wir nicht ein Rechteck sondern einen Kreis im Verhältnis 1 zu θ, erhalten wir den Winkel $\frac{1}{\theta} \cdot 360° = 222{,}492\ldots$ und den Gegenwinkel $(1 -$

$\frac{1}{\theta}) \cdot 360° \overset{(2.10)}{=} \frac{1}{\theta^2} \cdot 360° = 137{,}507\ldots$. Der kleinere dieser beiden Winkel wird *goldener Winkel* genannt.

Wir verlassen nun die Kunst und beschäftigen uns ab jetzt mit folgender sehr allgemeiner Frage.

Frage. Wo wachsen Blätter an Pflanzen?

Wir beschäftigen uns hier nur mit sogenannten *wechselständigen* Pflanzen. D.h. Pflanzen bei denen sich keine zwei Blätter auf der gleichen Höhe des Stiels befinden. Wir können bei diesen Pflanzen an jedem Stiel genau sagen, welches Blatt als erstes (ganz unten) gewachsen ist, welches als zweites, usw. Solche Pflanzen kann man mit einem Blattstellungsdiagramm darstellen. Dabei wird das unterste Blatt im inneren Kreis eingezeichnet, und die restlichen Blätter in der entsprechenden Reihenfolge von innen nach aussen.

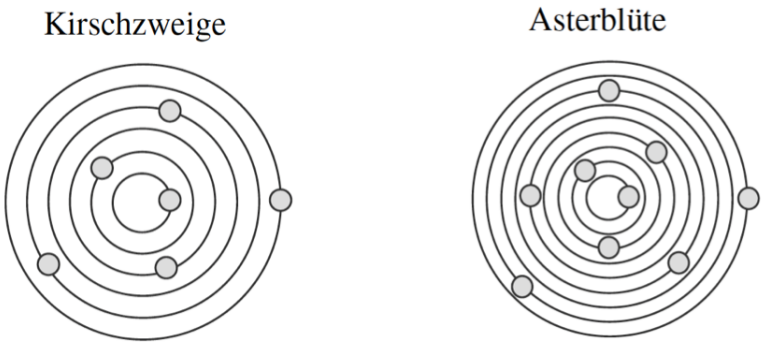


Abbildung 2.2: Blattstellungsdiagramme für Kirschzweige (links) und die Asterblüte (rechts). Die Kirschzweige haben einen Divergenzwinkel von $\frac{2}{5} \cdot 360°$ und die Asterblüte einen Divergenzwinkel von $\frac{3}{8} \cdot 360°$.

Der Winkel, in dem zwei aufeinanderfolgende Blätter stehen, nennen wir den *Divergenzwinkel* der Pflanze. Ist der Divergenzwinkel ein rationales Vielfaches von 360° – sagen wir $\frac{a}{b} \cdot 360°$ – dann werden das erste Blatt und das $b$te Blatt genau übereinander stehen. Dabei wird der Stiel insgesamt a mal „umlaufen“. Das gilt natürlich auch umgekehrt: Stehen das erste und das $b$te Blatt übereinander und wird dabei der Stiel $a$mal umlaufen, dann hat die Pflanze einen Divergenzwinkel von $\frac{a}{b} \cdot 360°$.
In den Beispielen aus Abbildung 2.2 sehen wir, dass Zähler und Nenner der Divergenzwinkel Fibbonacci-Zahlen sind. Schimper und Braun haben

Anfang des 19. Jahrhunderts eine große Studie zu Divergenzwinkeln geführt und sind dabei in den meisten Fällen auf Divergenzwinkel der Form $\frac{a_n}{a_{n+2}}$ gestoßen, wobei a_n die $n$te Fibonacci-Zahl darstellt. Woran kann das liegen? Für die Lichtausbeute der Pflanze ist es ideal, wenn nie zwei Blätter genau übereinanderstehen. Der Divergenzwinkel sollte also nach Möglichkeit ein irrationales Vielfaches von 360° sein. Tatsächlich lässt sich mit Hilfe der Theorie von Kettenbrüchen zeigen, dass die Zahlen $\pm\theta$ und $\pm\frac{1}{\theta}$ schlechter als alle anderen reellen Zahlen durch rationale Zahlen angenähert werden können. Sie sind also in gewisser Weise die ïrrationalsten"reellen Zahlen. Daher erscheint aus Sicht der Lichtausbeute der goldene Winkel, bzw. der Gegenwinkel dazu, für die Pflanzen am besten zu sein.
Sei nun die Fibonacci-Zahlen $a_0, a_1, \ldots$ gegeben. Dann gilt mit der geschlossenen Formel für diese

$$\frac{a_n}{a_{n+2}} = \frac{\frac{1}{\sqrt{5}} \cdot \left(\left(\frac{1+\sqrt{5}}{2}\right)^{n+1} - \left(\frac{1-\sqrt{5}}{2}\right)^{n+1} \right)}{\frac{1}{\sqrt{5}} \cdot \left(\left(\frac{1+\sqrt{5}}{2}\right)^{n+3} - \left(\frac{1-\sqrt{5}}{2}\right)^{n+3} \right)} \xrightarrow{n\to\infty} \frac{1}{\theta^2}.$$

Das sieht man sofort ein nachdem man erkannt hat, dass $|\frac{1-\sqrt{5}}{2}| < 1$ ist. Damit nähern die Winkel der Form $\frac{a_n}{a_{n+2}} \cdot 360°$ tatsächlich den goldenen Winkel an! Da es sehr schwer zu entscheiden ist, ob zwei Blätter genau(!) übereinanderstehen, können möglicherweise die meisten der festgestellten Divergenzwinkel auf den goldenen Winkel zurückzuführen sein.
Etwas schöner sieht man dieses Phänomen bei Sonnenblumen, Ananas und Tannenzapfen. Auch hier findet man stets die Fibonacci-Zahlen wieder. Bei der Sonnenblume wird der nächste Samen im goldenen Winkel vom vorherigen angelegt. Genau, wie bei der Lichtausbeute wird durch diese möglichst irrationale Anordnung der Platz optimal genutzt, vgl. Abbildung 2.3.

2.5 Der Übergang Differenzen- zu Differenzialgleichungen

Problem 2.5.1. Auf einem Bauernhof gibt es 200 Schafe. Die Herde soll ohne Zukauf von Tieren auf 500 Tiere wachsen. Wann ist das Ziel erreicht?

Das lässt sich sicher nicht ohne weitere Angaben lösen. Es wird daher ein Jahr gewartet und die Herde wird noch einmal gezählt. Nach einem Jahr

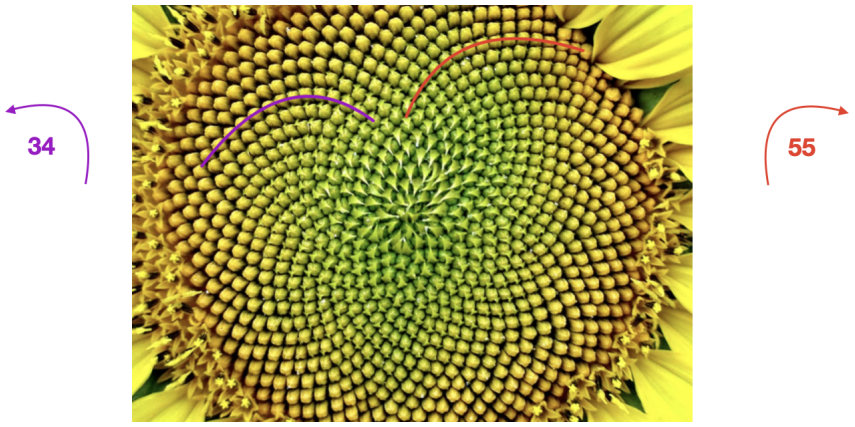


Abbildung 2.3: Bei dieser Sonnenblume (so wie bei fast allen anderen auch) sehen wir genau 34 Bögen, die von der Mitte aus nach links gebogen sind, und 55, die nach rechts gebogen sind. Das sind wieder zwei aufeinanderfolgende Fibonacci-Zahlen. Das ergibt sich wieder durch die Beobachtung $\frac{34}{55} \approx \frac{1}{\theta}$.

sind es bereits 230 Tiere. Die Differenz zwischen den neu geborenen und den verstorbenen Tieren betrug im ersten Jahr also 30.
Der Zuwachs an Tieren wird von der aktuellen Größe der Herde abhängen. Es werden also nicht in jedem Jahr genau 30 Tiere dazukommen, da es bei einer grösseren Gesamtpopulation auch mehr Geburten und mehr Todesfälle geben wird. Gehen wir davon aus, dass es zu Beginn 2000 Tiere gab, dann ist es eine vernünftige Annahme, dass es unter den gleichen Bedingungen nach einem Jahr 2300 Tiere gegeben hätte. D.h.:

Annahme: Das Wachstum ist proportional zur Populationsgröße.

Der Wachstumsfaktor ist in unserem Beispiel $r = \frac{230}{200} = 1 + \frac{3}{20} = 1{,}15$.
Mit dieser Annahme ist der weitere Weg klar vorgegeben.
Wir verteilen Namen und bezeichnen mit a_n die Anzahl von Schafen nach n Jahren. Dann ist $a_0 = 200$, $a_1 = 230$, $a_2 = 1{,}15 \cdot 230$ und allgemein $a_n = 1{,}15 \cdot a_{n-1}$ für alle $n \in \mathbb{N}$.
Diese Differenzengleichung kennen wir schon vom ersten Beispiel mit den Zinsen. Dort haben wir auch eine geschlossene Formel gefunden. Es gilt

$$a_n = (1,15)^n \cdot 200 \qquad \text{für alle } n \in \mathbb{N}_0.$$

Beachten Sie, dass dies (noch) nur für natürliche Zahlen n gilt. Das ist der Kern von unseren diskreten Differenzengleichungen.

Wir finden aber leicht heraus, dass $a_6 < 500 < a_7$ ist. Damit können wir annehmen, dass im Verlauf des siebten Jahres die Herde die Größe von 500 Tieren übersteigen wird.

Bemerkung 2.5.2. Da die Geburten von Schafen gehäuft zu gewissen Zeiten im Jahr stattfinden (Osterlämmer), ist es tatsächlich sinnvoll, die Herdengröße nach genau einem Jahr noch einmal zu ermitteln. Wenn wir nur einen Monat genommen hätten, würde es einen großen Unterschied machen, ob wir den November oder den Februar gewählt hätten.
Andere Dinge werden in unserem Modell nicht berücksichtigt:

- Gibt es überhaupt genug Platz (und Futter, ...) für 500 Tiere?
- Wie ist die Altersstruktur und Geschlechterverteilung der Herde? Wurde die Herde neu mit jungen gesunden Tieren gegründet, dann wird es zwar einige Geburten, aber sicher sehr wenige Todesfälle geben. Wir sind also stillschweigend davon ausgegangen, dass die wesentlichen Merkmale (insbesondere Alter und Geschlecht) gleichmäßig verteilt sind.
- Die Population wächst in unserem Modell immer weiter. Damit kann das Modell nur für begrenzte Zeitintervalle angewendet werden.

Wir verändern das Beispiel minimal.

Problem 2.5.3. In einer Petrischale seien $200mm^2$ von Bakterien besiedelt. Es sollen ohne Hinzufügen von Bakterien von außen, $500mm^2$ werden. Wann ist das Ziel erreicht?

Wieder fehlt etwas. Nach genau einem Tag wird noch einmal gemessen. Nach einem Tag sind $230mm^2$ besiedelt.
Das sind die gleichen Voraussetzungen, wie eben bei den Schafen. Jetzt allerdings ist die Zeiteinheit von einem Tag vollkommen willkürlich gewählt. Wir hätten genau so gut auch eine Stunde oder fünf Minuten bis zur Messung warten können. Wenn wir zusätzlich noch fragen wann das Ziel *genau* erreicht ist, ist unsere diskrete Methode der Differenzengleichung nicht mehr das richtige Mittel. Denn die Bakterienpopulation wächst kontinuierlich.
Da wir eine willkürliche Zeiteinheit von einem Tag gewählt haben, müssen wir bei der Lösung zwei Parameter berücksichtigen. Erstens die gewählte

Zeiteinheit h und zweitens die gemessene Wachstumsrate r. Diese Wachstumsrate hängt natürlich von h ab. In unserem Fall haben wir $h =$ einen Tag, und $r = 1{,}15$. Was können wir über die Abhängigkeit von r und h aussagen? Die Werte sind nicht proportional zu einander, da sonst in einem halben Tag die Wachstumsrate $\frac{1{,}15}{2} < 1$ wäre.
Als nächstes überlegen wir uns, ob es einen linearen Zusammenhang gibt. Wenn wir die Zeitspanne $h = 0$ wählen, dann ist die zugehörige Wachstumsrate gleich 1 (die Populationsgröße verändert sich nicht). Der einzige mögliche lineare Zusammenhang ist daher

$$r = r(h) = 1 + h \cdot p, \qquad \text{für ein } p \in \mathbb{R}. \tag{2.11}$$

Das kann allerdings nicht für alle Zeiteinheiten gelten. Denn für $h = 1$ und $r = 1{,}15$ erhalten wir $p = 0{,}15$. Dann wäre die Wachstumsrate in zwei Tagen gleich $1 + 2 \cdot 0{,}15 = 1{,}3$. Andererseits können wir genau wie bei der Schafspopulation argumentieren, dass die Wachstumsrate gleich $1{,}15^2 = 1{,}3225 \neq 1{,}3$ ist. Damit gilt dieser lineare Zusammenhang im Allgemeinen nicht. Für sehr kleines h können wir allerdings einen linearen Zusammenhang (2.11) annehmen, da die Gleichung für $h = 0$ erfüllt ist. Hier nehmen wir an, dass die Zuordnung $h \mapsto r(h)$ differenzierbar (hinreichend schön) ist. Denn jede differenzierbare Funktion sieht auf winzigen Intervallen aus wie eine lineare Abbildung (Gerade). Bezeichen wir also mit $x(t)$ die Größe der besiedelten Fläche in mm^2 nach t Tagen, dann folgt für *sehr kleines* h

$$x(t+h) \approx (1 + h \cdot p) \cdot x(t) \quad \Longrightarrow \quad x(t+h) - x(t) \approx h \cdot p \cdot x(t).$$

Teilen wir beide Seiten durch h und präzisieren das Symbol $\approx$, erhalten wir

$$\lim_{h \to 0} \frac{x(t+h) - x(t)}{h} = p \cdot x(t).$$

Mit der Definition der Ableitung folgt damit

$$x'(t) = p \cdot x(t), \qquad \text{für ein } p \in \mathbb{R}. \tag{2.12}$$

Wir sind nun auf eine Differenzialgleichung gestoßen, die die Populationsgröße beschreibt und nur noch von dem *einen* Parameter p abhängt.
Wir sehen im nächsten Abschnitt, dass alle Lösungen der DGL (2.12) von der Form $x(t)C \cdot e^{pt}$ sind. Nutzen wir $x(0) = 200$, folgt $C = 200$. Aus $x(1) = 230$ folgt dann

$$230 = 200 \cdot e^{p \cdot 1} \quad \Longleftrightarrow \quad e^p = \frac{230}{200} = 1{,}15.$$

Damit folgt $x(t) = 200 \cdot (1{,}15^t)$. Wir haben also tatsächlich unsere Lösung des diskreten Falles, auch für den kontinuierlichen Fall bestätigt.
Ist $x(t)$ die Populationsgröße nach t Tagen, so finden wir $x(t) = (1{,}15)^t \cdot 200$. Damit verallgemeinert sich unsere diskrete Lösung auch auf den kontinuierlichen Fall. Wenn wir nun wissen möchten, wann genau das Ziel von $500mm^2$ erreicht wird, rechnen wir

$$\begin{aligned} 500 = x(t) = 200 \cdot (1{,}15)^t \quad &\iff \quad (1{,}15)^t = \frac{500}{200} = \frac{5}{2} \\ &\iff \quad t \cdot \ln(1{,}15) = \ln(\frac{5}{2}) \\ &\iff t = \frac{\ln(\frac{5}{2})}{\ln(1{,}15)} = 6{,}5560\ldots \end{aligned}$$

Das Ziel wird also nach etwa 6,556 Tagen erreicht sein.
Das hier beschriebene Wachstumsmodell ist das *exponentielle Wachstum*. Es beschreibt das ungehemmte Wachstum einer Population und hat seinen Ursprung in unserer Annahme, dass das Wachstum proportional zur Größe der Population ist.

$$\underbrace{x'(t)}_{\text{Änderungsrate zum Zeitpunkt } t} = C \cdot \underbrace{x(t)}_{\text{Größe zum Zeitpunkt } t}$$

Ein weiteres wichtiges Beispiel für exponentielles Wachstum ist der Zerfall radioaktiver Substanzen. Auch hier ist der Zerfall proportional zur vorhandenen Masse. Wir sprechen also auch von exponentiellem Wachstum, wenn das Wachstum negativ ist.

2.6 Lineare DGL erster Ordnung

Im letzten Abschnitt haben wir eine DGL kennengelernt, und die Lösung auf später verschoben. Wir wollen direkt eine ganze Klasse von Differenzialgleichungen betrachten.

Definition 2.6.1. Sei I ein Intervall und $p, q : I \longrightarrow \mathbb{R}$ Funktionen. Dann heißt die Differenzialgleichung

$$x'(t) = p(t) \cdot x(t) + q(t)$$

linear von Ordnug 1. Ist die Funktion q konstant Null, so nennen wir die Differenzialgleichung *homogen*, anderenfalls nennen wir sie *inhomogen*.

Satz 2.6.2. *Sei I ein Intervall und $p : I \longrightarrow \mathbb{R}$ stetig. Sei weiter $P : I \longrightarrow \mathbb{R}$ eine Stammfunktion von p. Dann sind alle Lösungen der homogenen linearen Differenzialgleichung $x'(t) = p(t) \cdot x(t)$ gegeben durch*

$$x(t) = c \cdot e^{P(t)} \quad \text{für } c \in \mathbb{R}.$$

BEWEIS. Übung. □

Lemma 2.6.3. *Seien $x_1(t)$ und $x_2(t)$ Lösungen der linearen Differenzialgleichung $x'(t) = p(t) \cdot x(t) + q(t)$ auf einem Intervall I. Dann ist $(x_1 - x_2)(t)$ eine Lösung der homogenen linearen Differenzialgleichung $x'(t) = p(t) \cdot x(t)$.*

BEWEIS. Das ist schnell nachgerechnet. Es gilt:

$$\begin{aligned}(x_1 - x_2)'(t) &= x_1'(t) + x_2'(t) = (p(t) \cdot x_1(t) + q(t)) - (p(t) \cdot x_2(t) + q(t)) \\ &= p(t) \cdot x_1(t) - p(t) \cdot x_2(t) = p(t) \cdot (x_1 - x_2)(t).\end{aligned}$$

Genau das mussten wir zeigen. □

Korollar 2.6.4. *Gegeben sei die lineare Differenzialgleichung $x'(t) = p(t) \cdot x(t) + q(t)$, mit stetigen Funktionen p, q auf einem Intervall I. Sei $P(t)$ eine Stammfunktion von $p(t)$ und sei $x_p(t)$* eine *Lösung der Differenzialgleichung. Dann sind* alle *Lösungen der DGL gegeben durch*

$$x(t) = c \cdot e^{P(t)} + x_p(t) \quad , \text{ mit } c \in \mathbb{R}.$$

BEWEIS. Das folgt sofort aus der Kombination von 2.6.2 und 2.6.3. □

Slogan: Kennen wir eine Lösung, kennen wir alle Lösungen.

Es bleibt die Frage, wie man eine Lösung der DGL findet. Das können wir recht allgemein mit dem Prinzip der *Variation der Konstanten* machen.

Konstruktion 2.6.5. Sei $x'(t) = p(t) \cdot x(t) + q(t)$, mit p, q stetig auf einem Intervall I. Die folgenden Schritte führen zu allen Lösungen dieser DGL. Eine einzige Lösung x_p nennen wir auch *partikuläre Lösung* der DGL. Die Gesamtheit aller Lösungen bezeichnen wir als *allgemeine Lösung.*

(1) Wir lösen erst die homogene Gleichung $x'(t) = p(t) \cdot x(t)$. Diese Lösungen sind nach Satz 2.6.2 gegeben durch

$$x_h(t) = c \cdot e^{P(t)} \quad , \text{ mit } c \in \mathbb{R}. \tag{2.13}$$

Hier ist $P(t)$ (irgend)eine fest gewählte Stammfunktion von $p(t)$.

(2) Betrachte die Konstante c in (2.13) als Funktion $c(t)$.

Als Ansatz für unsere partikuläre Lösung wählen wir nun $x_p(t) = c(t) \cdot e^{P(t)}$. Diese Funktion setzen wir in die DGL ein und erhalten

$$\begin{aligned} & x_p'(t) = p(t) \cdot x_p(t) + q(t) \\ \Longrightarrow\ & c'(t)e^{P(t)} + c(t)p(t)e^{P(t)} = p(t) \cdot c(t)e^{P(t)} + q(t) \\ \Longrightarrow\ & c'(t) = q(t)e^{-P(t)} \end{aligned}$$

Damit ist $x_p(t) = c(t)e^{P(t)}$ genau dann eine Lösung der DGL, wenn $c(t) = \int q(t)e^{-P(t)}\mathrm{d}t$ eine Stammfunktion von $q(t)e^{-P(t)}$ ist. Da $q(t)$ stetig ist (und $e^{-P(t)}$ sowieso), existiert so eine Stammfunktion. Damit haben wir tatsächlich eine partikuläre Lösung gefunden.

(3) Wir setzten $x_h(t)$ und $x_p(t)$ zusammen. Nach Korollar 2.6.4 wissen wir nun, dass die allgemeine Lösung der DGL gegeben ist durch

$$x(t) = c \cdot e^{P(t)} + \left(\int q(t)e^{-P(t)}\mathrm{d}t \right) \cdot e^{P(t)} \quad \text{, mit } c \in \mathbb{R}.$$

Bemerkung 2.6.6. Dass es die Funktion $c(t) = \int q(t)e^{-P(t)}\mathrm{d}t$ gibt, heißt nicht, dass wir sie auch explizit bestimmen/berechnen können. In der Praxis kann man diese Funktion numerisch bestimmen – also eine Approximation der Funktionswerte finden.

Beispiel 2.6.7. Wir bestimmen alle Lösungen von $x'(t) = 4x(t) + (t^2 + 1)$. Dazu führen wir die Schritte von eben aus.

(1) Alle Lösungen der homogenen Gleichung $x'(t) = 4x(t)$ sind $x_h(t) = ce^{4t}$, für $c \in \mathbb{R}$.

(2) $\boxed{c \leftrightarrow c(t):}$ Wir setzen $x_p(t) = c(t)e^{4t}$ in die DGL ein:

$$c'(t)e^{4t} + c(t)4e^{4t} = x_p'(t) = 4x_p(t) + (t^2 + 1) = 4c(t)e^{4t} + (t^2 + 1)$$

Damit ist $x_p(t)$ eine partikuläre Lösung, genau dann wenn $c'(t) = (t^2 + 1) \cdot e^{-4t}$ ist. Im folgenden berechnen wir eine Stammfunktion von $(t^2 +$

$1) \cdot e^{-4t}$ mit partieller Integration. Es ist

$$\begin{aligned}\int (t^2+1)e^{-4t}\mathrm{d}t &= (t^2+1)\cdot(-\frac{1}{4})e^{-4t} - \int 2t\cdot(-\frac{1}{4})e^{-4t}\mathrm{d}t \\ &= -\frac{t^2+1}{4}e^{-4t} + \frac{1}{2}\int te^{-4t}\mathrm{d}t \\ &= -\frac{t^2+1}{4}e^{-4t} + \frac{1}{2}\left([t\cdot(-\frac{1}{4})e^{-4t}] - \int(-\frac{1}{4})e^{-4t}\mathrm{d}t\right) \\ &= -\frac{t^2+1}{4}e^{-4t} - \frac{t}{8}e^{-4t} + \frac{1}{8}\int e^{-4t}\mathrm{d}t \\ &= -e^{-4t}\left(\frac{t^2}{4}+\frac{t}{8}+\frac{9}{32}\right).\end{aligned}$$

Damit ist eine partikuläre Lösung gegeben durch

$$x_p(t) = c(t)e^{4t} = -e^{-4t}\left(\frac{t^2}{4}+\frac{t}{8}+\frac{9}{32}\right)\cdot e^{4t} = -\frac{t^2}{4}-\frac{t}{8}-\frac{9}{32}.$$

(3) Wir fassen zusammen und erhalten, dass alle Lösungen der DGL gegeben sind durch

$$x(t) = x_h(t) + x_p(t) = ce^{4t} - \frac{t^2}{4} - \frac{t}{8} - \frac{9}{32} \quad \text{, mit } c\in\mathbb{R}.$$

Wir sehen, dass die partikuläre Lösung vom gleichen Typ ist, wie die Funktion $q(t)$ – beides sind quadratische Polynome. Wenn (wie in diesem Beispiel) die Funktion $p(t)$ konstant ist, ist das meistens der Fall. Wir können das Beispiel daher alternativ mit dem *Ähnlichkeitsansatz* lösen.

Beispiel 2.6.8. Wir möchten noch einmal eine partikuläre Lösung von $x'(t) = 4x(t)+(t^2+1)$ finden (siehe Beispiel 2.6.7). Da $p(t) = 4$ konstant ist, und die rechte Seite ein quadratisches Polynom ist, machen wir den Ansatz $x_p(t) = at^2+bt+c$. Ziel ist es jetzt a, b, c zu bestimmen, so dass $x_p(t)$ die DGL löst. Dazu müssen wir die Funktion $x_p(t)$ einfach wieder in die DGL einsetzen:

$$2at + b = x_p'(t) = 4x_p(t) + (t^2+1) = (4a+1)t^2 + 4bt + (4c+1).$$

Koeffizientenvergleich liefert die Bedingungen

$$\begin{aligned}0 &= 4a+1 \\ 2a &= 4b \\ b &= 4c+1\end{aligned}$$

Das ist ein sehr einfaches lineares Gleichungssystem, das wir schnell lösen können. Das Ergebnis ist $a = -\frac{1}{4}$, $b = -\frac{1}{8}$, $c = -\frac{9}{32}$. Damit haben wir, wie erwartet, die partikuläre Lösung $x_p(t) = -\frac{1}{4}t^2 - \frac{1}{8}t - \frac{9}{32}$ gefunden.

Definition 2.6.9. Ein *Anfangswertproblem* (AWP) ist gegeben durch eine Differenzialgleichung

$$x^{(k)} = f(x(t), x'(t), \ldots, x^{(k-1)}(t), t) \tag{2.14}$$

und reelle Zahlen $t_0, x_0, x_1, \ldots, x_{k-1}$ für die

$$x^{(i)}(t_0) = x_i \quad \forall i \in \{0, \ldots, k-1\} \tag{2.15}$$

gelten soll. Gesucht ist dann eine Funktion $x(t)$ auf einem Intervall I, mit $t_0 \in I$, die (2.14) für alle $t \in I$ und (2.15) erfüllt. Das Intervall I nennen wir *Lösungsintervall*. Das größtmögliche Lösungsintervall heißt dementsprechend *maximales Lösungsintervall*.

Satz 2.6.10. *Sei $x'(t) = p(t)x(t) + q(t)$ eine lineare DGL der Ordnung 1, mit $p(t), q(t)$ stetig auf einem Intervall I. Dann ist jedes AWP $x(t_0) = x_0$, für $t_0 \in I$, $x_0 \in \mathbb{R}$, eindeutig lösbar. D.h. Es gibt genau eine Funktion, die auf dem Intervall I die Bedingungen $x'(t) = p(t)x(t) + q(t)$ und $x(t_0) = x_0$ erfüllt.*

Beweis. Mit Variation der Konstanten (Konstruktion 2.6.5) haben wir gesehen, dass es eine Lösung $x_p(t)$ der DGL gibt. In Korollar 2.6.4 haben wir gesehen, dass damit *alle* Lösungen der DGL gegeben sind durch $x(t) = c \cdot e^{P(t)} + x_p(t)$, mit $c \in \mathbb{R}$ beliebig. Hier ist $P(t)$ irgendeine Stammfunktion von $p(t)$. Es gibt aber genau eine Konstante c, für die auch das AWP $x(t_0) = x_0$ erfüllt ist. Denn

$$x_0 = x(t_0) \iff x_0 = c \cdot e^{P(t_0)} + x_p(t_0) \iff c = \frac{x_0 - x_p(t_0)}{e^{P(t_0)}}.$$

Dieser Ausdruck ist definiert, da $e^{P(t_0)} \neq 0$ ist. □

Mischungsprozesse

Problem 2.6.11. Ein Tank enthält 1000ℓ Wasser, in dem $50kg$ Salz gelöst sind. Nun werden ein Zu- und ein Abfluss eingeschaltet. Pro Minute fließen

jetzt 10ℓ aus dem Tank heraus und 10ℓ mit einem Salzgehalt von $2kg$ hinzu. Alles wird sofort gut durchmischt. Wie groß ist der Salzgehalt im Tank nach $t \in \mathbb{R}$ Minuten? Was können wir über das Langzeitverhalten des Salzgehaltes aussagen?

Bemerkung 2.6.12. Wir überlegen uns die Antwort auf die zweite Frage ohne Rechnung. Da der Salzgehalt im Wasser, das in den Tank läuft, immer gleich ist, gehen wir davon aus, dass der Salzgehalt im Tank den Wert annähert, den wir ehalten würden wenn wir den Tank von vornherein mit $2kg$ Salz pro 10ℓ gefüllt hätten. Auf lange Sicht erwarten wir also einen Salzgehalt von etwa $2kg \cdot \frac{1000\ell}{10\ell} = 200kg$.

Als nächstes versuchen wir das Problem zu modellieren. Dazu verteilen wir als erstes wieder Namen: Sei $x(t)$ der Salzgehalt im Tank t Minuten nach öffnen des Zu- und Ablaufs. Der Zulauf ist leicht zu handhaben, denn in h Minuten laufen stets $h \cdot 10\ell$ Wasser mit einem Salzgehalt von $h \cdot 2kg$ dazu. Es laufen natürlich auch $h \cdot 10\ell$ Wasser ab, aber den Salzgehalt können wir nicht so einfach bestimmen, da dieser nicht konstant ist.
Für sehr kleines $h > 0$ können wir aber annehmen, dass in etwa $x(t+h) \approx x(t)$ ist. Der Salzgehalt pro 10ℓ Wasser zum Zeitpunkt t ist $h \cdot \frac{10\ell}{1000\ell} \cdot x(t) = h \cdot \frac{x(t)}{10}$. Damit sollten für kleines $h > 0$ in h Minuten etwa $h \cdot \frac{x(t)}{100}$ Salz ablaufen. Es sollte also gelten

$$x(t+h) \approx x(t) + h \cdot 2 - \frac{x(t)}{100} \iff \frac{x(t+h) - x(t)}{h} \approx 2 - \frac{x(t)}{100}.$$

Wir präzisieren das Symbol $\approx$ wieder dadurch, dass es sicher zur Gleichheit wird, wenn h gegen Null strebt. Es ist also

$$x'(t) = \lim_{h \to 0} \frac{x(t+h) - x(t)}{h} = 2 - \frac{x(t)}{100} = -\frac{1}{100} \cdot x(t) + 2.$$

Weiter kennen wir den Salzgehalt zum Zeitpunkt $t = 0$, nämlich $x(0) = 50$. Der Salzgehalt $x(t)$ ist also gegeben durch das AWP

$$x'(t) = -\frac{1}{100} \cdot x(t) + 2 \quad \text{und} \quad x(0) = 50. \tag{2.16}$$

Nach gerade eben wissen wir, dass es genau eine Lösung für dieses AWP gibt. Unsere Modellannahmen sind also in so weit vollständig, dass das mathematische Problem tatsächlich genau eine Lösung besitzt.

Die homogene Lösung der DGL ist $x_h(t) = c \cdot e^{-1/100 \cdot t}$. Eine partikuläre Lösung erhalten wir durch den Ähnlichkeitsansatz. Wir erwarten eine konstante Lösung, die schnell durch lösen von $0 = -\frac{1}{100} \cdot x + 2$ gefunden ist. Damit ist die allgemeine Lösung der DGL $x(t) = c \cdot e^{-1/100 \cdot t} + 200$. Einsetzen des Anfangswertes, liefert sofort $c = -150$. Der Salzgehalt zum Zeitpunkt t ist damit gegeben durch

$$x(t) = -150 \cdot e^{-1/100 \cdot t} + 200.$$

Wir stellen fest, dass diese Funktion die Bedingung $\lim_{t \to \infty} x(t) = 200$ erfüllt, die wir uns in Bemerkung 2.6.12 überlegt haben. Es war also möglich den langfristigen Verlauf zu bestimmen, ohne die tatsächliche Funktion zu bestimmen. Unteranderem damit wollen wir uns im nächsten Abschnitt beschäftigen.

2.7 Trennen der Variablen

Wir betrachten in diesem Abschnitt eine weitere einfache Klasse von Differenzialgleichungen erster Ordnung, in die auch die lineare DGL aus (2.16) fällt.

Definition 2.7.1. Wir sagen, dass eine DGL erster Ordnung *getrennte Variablen* besitzt, wenn es Funktionen $f, g : \mathbb{R} \longrightarrow \mathbb{R}$ gibt, mit

$$x'(t) = f(x(t)) \cdot g(t).$$

Ist g konstant gleich 1, so heißt der DGL *autonom*.

Beispiel 2.7.2. Wir betrachten drei Differenzialgleichungen.

(a) $x'(t) = -x(t)^2$ ist eine autonome DGL.

(b) $x'(t) = -x(t)^2 \cdot t$ ist nicht autonom, besitzt aber getrennte Variablen.

(c) $x'(t) = -x(t)^2 + t$ besitzt keine getrennten Variablen.

Wir lösen zunächst die DGL (a). Man kann sie noch recht leicht durch raten lösen, aber wir stellen eine Idee vor, die ohne Raten auskommt. Dazu nehmen

wir an, dass auf einem kleinen Intervall $x(t) \neq 0$ gilt. Dann ist

$$\begin{aligned} x'(t) = -x(t)^2 &\iff \frac{x'(t)}{x(t)^2} = -1 \\ &\iff \int \frac{x'(t)}{x(t)^2} \mathrm{d}t = \int -1 \mathrm{d}t = -t + c, \quad \text{mit } c \in \mathbb{R}. \end{aligned}$$

Die unbestimmten Integrale stehen für eine Stammfunktion. Das c ist dabei die unbestimmte Integrationskonstante (zwei Stammfunktionen unterscheiden sich nur durch die Addition einer Konstanten). Das linke Integral können wir leicht mit der Substitutionsregel berechnen. Beachten Sie: Ist eine Funktion $H(x(t))$ gegeben und ist $H' = h$, dann ist $(H(x(t)))' = h(x(t)) \cdot x'(t)$. Es ist also $H(x(t))$ eine Stammfunktion von $h(x(t)) \cdot x'(t)$. In unserem Fall ist $h(x(t)) = \frac{1}{x(t)^2}$. Diese Substitution notieren wir kurz durch

$$-t + c = \int \frac{x'(t)}{x(t)^2} \mathrm{d}t = \int \frac{1}{x^2} \mathrm{d}x = \frac{1}{1 + (-2)} x^{1+(-2)} = -\frac{1}{x}.$$

Damit ist natürlich gemeint

$$-t + c = -\frac{1}{x(t)}.$$

Eine Integrationskonstante müssen wir dabei tatsächlich nur auf einer Seite hinzufügen. Umstellen der Gleichung führt auf

$$x(t) = \frac{1}{t - c}, \quad \text{mit } c \in \mathbb{R}.$$

Natürlich ist auch die konstante Nullfunktion eine Lösung der DGL. Damit haben wir alle Lösungen der DGL gefunden. Sie sind gegeben durch $x(t) = 0$ und $x(t) = \frac{1}{t-c}$ für $c \in \mathbb{R}$. Wieder nennen wir die Gesamtheit aller Lösungen die *allgemeine Lösung*. Wenn wir uns die Lösungen etwas genauer betrachten stellen wir fest, dass wir jedes AWP $x(t_0) = x_0$ zu dieser DGL eindeutig lösen können! Ist nämlich $x_0 = 0$, so ist die konstante Nullfunktion die einzige Lösung. Für $x_0 \neq 0$ gibt es genau eine Konstante $c \in \mathbb{R}$, die diese Bedingung erfüllt; nämlich $c = \frac{1}{x_0} - t_0$.

Beispiel (b) lösen wir genauso. Wir kürzen die Schreibweise ab jetzt aber

etwas ab. Für $x(t) \neq 0$ ist

$$\begin{aligned} x'(t) = -x(t)^2 \cdot t &\iff \int \frac{x'(t)}{x(t)^2} \mathrm{d}t = \int -t \mathrm{d}t = -\frac{1}{2}t^2 + c \\ &\iff \int \frac{1}{x^2} \mathrm{d}x = -\frac{1}{2}t^2 + c \\ &\iff -\frac{1}{x(t)} = -\frac{1}{2}t^2 + c \iff x(t) = \frac{1}{t^2 - c}. \end{aligned}$$

Die Lösungen der DGL sind somit die konstante Nullfunktion und die Funktionen $x(t) = \frac{1}{t^2-c}$, mit $c \in \mathbb{R}$. Wieder stellen wir nach einer kurzen Überlegung fest, dass jedes AWP zu dieser DGL eindeutig lösbar ist.

Die DGL in (c) können wir immer noch so umformen, dass alle Komponenten mit x auf einer Seite sind und alle ohne x auf der anderen Seite – etwa $x'(t) + x(t)^2 = t$. Allerdings tritt auf der linken Seite die Ableitung $x'(t)$ nicht als Faktor auf. Daher wird Sie beim Integrieren nicht verschwinden. Das war aber der Kern der Lösungen von (a) und (b): Durch Integration verschwindet der Term $x'(t)$ und es bleibt nur noch eine Gleichung in $x(t)$ übrig. Das funktioniert hier in Teil (c) nicht. Tatsächlich kann man Lösungen dieser DGL nicht mit elementaren Funktionen aufschreiben!

Diese Lösungsstrategie funktioniert immer, wenn wir eine DGL mit getrennten Variablen haben!

Beispiel 2.7.3. Wir betrachten die DGL

$$x'(t) = 2 \cdot \sqrt{|x(t)|}.$$

Die DGL ist autonom und das Verfahren, das wir in Beispiel 2.7.2 kennengelernt haben, wird auch hier funktionieren. Die konstante Nullfunktion löst die DGL. Wir nehmen ab jetzt wieder $x(t) \neq 0$ an. Jetzt starten wir das Prozedere der Trennung der Variablen:

$$\begin{aligned} x'(t) = 2 \cdot \sqrt{|x(t)|} &\iff \frac{x'(t)}{2 \cdot \sqrt{|x(t)|}} = 1 \\ &\iff \int \frac{1}{2 \cdot \sqrt{|x|}} \mathrm{d}x = \int \frac{x'(t)}{2 \cdot \sqrt{|x(t)|}} \mathrm{d}t = \int 1 \mathrm{d}t = t + c \\ &\iff t + c = \begin{cases} \sqrt{x(t)} & \text{falls } x(t) \geq 0 \\ -\sqrt{-x(t)} & \text{falls } x(t) < 0 \end{cases} \end{aligned}$$

Quadrieren wir beide Seiten erhalten wir

$$|x(t)| = (t+c)^2, \quad \text{mit } c \in \mathbb{R}.$$

Wir setzen $x_c(t) = (t+c)^2$ für jedes $c \in \mathbb{R}$. Da Quadrieren keine Äquivalenzrelation ist, überprüfen wir schnell, ob die Funktionen $x_c(t)$ tatsächlich die DGL lösen. Dafür müssen wir die Funktion nur in die DGL einsetzen:

$$2(t+c) = x_c'(t) = 2 \cdot \sqrt{|x(t)|} = 2|t+c|. \quad \text{Passt, für } t \geq -c$$

Die Funktion $x_c(t)$ löst die DGL also auf dem Intervall $[-c, \infty)$. Anders als bei den bisherigen Beispielen, kann die Funktion $x_c(t)$ den Wert 0 annehmen. Das AWP

$$x'(t) = 2 \cdot \sqrt{|x(t)|} \quad \text{und} \quad x(0) = 0$$

besitzt damit die Lösungen $x_0(t) = t^2$ auf $[0, \infty)$ und die konstante Nullfunktion. Es ist also nicht eindeutig lösbar.
Tatsächlich gibt es unendlich viele Lösungen, die alle auf ganz $\mathbb{R}$ definiert sind. Wir sehen, dass für jedes $c \in \mathbb{R}$ die Gleichung $x_c'(-c) = 0$ gilt. Damit können wir diese Lösungen auch mit der konstanten Lösung $x(t) = 0$ kombinieren. Es folgt, dass für jedes $c \in \mathbb{R}$ die (differenzierbare) Funktion

$$\tilde{x_c}(t) = \begin{cases} (t+c)^2 & \text{für } t > -c \\ 0 & \text{sonst} \end{cases}$$

die DGL löst. Insbesondere löst jede Funktion $\tilde{x_c}(t)$, mit $c \leq 0$ das AWP.

Bemerkung 2.7.4. Die *nicht Eindeutigkeit* im vorherigen Beispiel liegt daran, dass wir hier verschiedene Lösungen der DGL gefunden haben, die sich schneiden. Das gilt natürlich auch allgemein.
Sei eine DGL $x'(t) = f(x(t), t)$ gegeben. Schneiden sich in einem Punkt (t_0, x_0) zwei verschiedene Lösungen $x_1(t)$ und $x_2(t)$, dann hat das AWP $x'(t) = f(x(t), t)$ mit $x(t_0) = x_0$ mindestens zwei Lösungen.
Bei einem Modellierungsprozess wollen wir (wie bei Frage 2.6.11) *eine* Lösung für ein gegebenes Problem finden. Daher sollte bei der Modellierung darauf geachtet werden, dass dieses Phänomen nicht auftritt. Wir möchten also Bedingungen für eine DGL finden, die garantieren, dass ein AWP immer nur eine Lösung besitzt. Wie gerade festgestellt, ist das äquivalent dazu, dass sich verschiedene Lösungen einer DGL nicht schneiden können.

Definition 2.7.5. Sei $x'(t) = f(x(t), t)$ eine DGL erster Ordnung. Ein $x^* \in \mathbb{R}$ heißt *stationärer Punkt* oder *stationäre Lösung*, wenn $f(x^*, t) = 0$ für alle $t \in \mathbb{R}$ gilt.
Ein $x^* \in \mathbb{R}$ ist offensichtlich genau dann eine stationäre Lösung, wenn die konstante Funktion $x(t) = x^*$ die DGL $x'(t) = f(x(t), t)$ löst.

Diese stationären Punkte sagen uns schon ziemlich viel über das Verhalten der Lösungen einer autonomen DGL aus.

Lemma 2.7.6. *Sei $x(t)$ eine Lösung der autonomen DGL $x'(t) = f(x(t))$, mit stetigem f. Gilt $\lim_{t\to\pm\infty} x(t) \in \mathbb{R}$, so ist $\lim_{t\to\pm\infty} x(t)$ ein stationärer Punkt der DGL.*

Beweis. Wir betrachten nur den Grenzwert $t \to +\infty$. Im Verlauf des Beweises benutzen wir zwei einfache Folgerungen aus der Dreiecksungleichung. Nämlich $|a| \geq |b| - |a - b|$ und $|a - b| \leq |a - c| + |b - c|$ für alle $a, b, c \in \mathbb{R}$. Jetzt legen wir aber endlich los. Sei also $\lim_{t\to\infty} x(t) = r \in \mathbb{R}$ und $x'(t) = f(x(t))$. Da f stetig ist, gilt

$$\lim_{t\to\infty} f(x(t)) = f\left(\lim_{t\to\infty} x(t)\right) = f(r).$$

Damit folgern wir, dass es für jedes $\varepsilon > 0$ ein t^* gibt, mit

(a) $|x(t) - r| < \varepsilon \quad \forall t \geq t^*$,

(b) $|f(x(t)) - f(r)| < \varepsilon \quad \forall t \geq t^*$.

Sei nun $\varepsilon > 0$ beliebig aber fest, und sei $t^* \in \mathbb{R}$, wie oben. Dann gilt für alle $t > t^*$ die Ungleichung

$$\begin{aligned}
|x(t) - x(t^*)| &= |\int_{t^*}^{t} x'(\tau)\mathrm{d}\tau| = |\int_{t^*}^{t} f(x(\tau))\mathrm{d}\tau| \\
&\geq |\int_{t^*}^{t} f(r)\mathrm{d}\tau| - |\int_{t^*}^{t} f(x(\tau)) - f(r)\mathrm{d}\tau| \\
&= (t - t^*) \cdot |f(r)| - \int_{t^*}^{t} |f(x(\tau)) - f(r)|\mathrm{d}\tau \\
&\overset{\text{(b)}}{\geq} (t - t^*) \cdot |f(r)| - \int_{t^*}^{t} \varepsilon\mathrm{d}\tau \\
&= (t - t^*) \cdot |f(r)| - (t - t^*) \cdot \varepsilon.
\end{aligned}$$

Umstellen dieser Ungleichung führt auf

$$\begin{aligned}(t-t^*)\cdot|f(r)| &\leq |x(t)-x(t^*)|+(t-t^*)\cdot\varepsilon\\ &\leq |x(t)-r|+|x(t^*)-r|+(t-t^*)\cdot\varepsilon \overset{(a)}{<} 2\varepsilon+(t-t^*)\cdot\varepsilon.\end{aligned}$$

Diese Ungleichung glit für alle $t \geq t^*$. Setzen wir nun $t = t^* + 1$ erhalten wir

$$|f(r)| < 3\varepsilon.$$

Da $\varepsilon > 0$ beliebig ist, folgt damit $f(r) = 0$, was nichts anderes bedeutet, als dass $r = \lim_{t\to\infty}$ ein stationärer Punkt der DGL ist. □

Theorem 2.7.7. *Sei D ein Intervall und $f : D \longrightarrow \mathbb{R}$ stetig. Ist $r \in \mathbb{R}$ größer als eine Nullstelle von f, dann bezeichnen wir mit r_- die größte Nullstelle von f, die kleiner r ist. Ist $r \in \mathbb{R}$ kleiner als eine Nullstelle von f, dann bezeichnen wir mit r_+ die kleinste Nullstelle, die größer als r ist. Wir nehmen an, dass immer wenn r_+ und/oder r_- definiert und $f(r) \neq 0$ ist, folgende Bedingung gilt:*

$$\left|\int_r^{r_\pm} \frac{1}{f(z)}\mathrm{d}z\right| = \infty. \tag{INT-INF}$$

Dann gilt für das AWP

$$x'(t) = f(x(t)) \quad \text{und} \quad x(t_0) = x_0 :$$

(a) für jedes Tupel $(t_0, x_0) \in \mathbb{R} \times D$ hat das AWP eine eindeutige Lösung.

(b) Die Lösung des AWP ist entweder konstant oder streng monoton.

(c) Jede beschränkte Lösung ist auf ganz $\mathbb{R}$ definiert.

Beweis. Bis ich dazu komme den Beweis zu tippen, verweise ich auf den Beweis von Satz 5.1 aus dem Buch von Sebastian Bauer. □

Bemerkung 2.7.8. Theorem 2.7.7 besagt, dass AWPe zu autonomen DGL $x'(t) = f(x(t))$, mit *schönen* Funktionen f, immer eindeutig lösbar sind. Insbesondere ist Bedingung (INT-INF) für alle Polynome erfüllt, was man ganz leicht mit Partialbruchzerlegung einsieht. Allgemeiner gilt (INT-INF) für alle stetig differenzierbaren Funktionen f.

Die Funktion $f(x) = 2\sqrt{|x|}$ aus Beispiel 2.7.3 erfüllt die Bedingung (INT-INF) natürlich nicht. Die einzige Nullstelle von f ist 0. Es ist also $1_- = 0$, und

$$\int_1^0 \frac{1}{2\sqrt{|x|}}\mathrm{d}x = -\int_0^1 \frac{1}{2\sqrt{x}}\mathrm{d}x = -\left[\sqrt{x}\right]_0^1 = -1.$$

Wie schon in Bemerkung 2.7.4 festgestellt, können sich zwei Lösungen der DGL $x'(t) = f(x(t))$ nicht schneiden, wenn f die Bedingung (INT-INF) erfüllt. Denn falls $x_1(t)$ und $x_2(t)$ verschiedene Lösungen wären und für ein gewisses $t_0 \in \mathbb{R}$ gilt $x_1(t_0) = x_2(t_0)$, dann wären x_1 und x_2 verschiedene Lösungen des AWP $x'(t) = f(x(t))$ und $x(t_0) = x_1(t_0)$. Das ist aber nach Theorem 2.7.7 ausgeschlossen.

Wir geben noch eine etwas handlichere Version von Theorem 2.7.7 für DGLen mit getrennten Variablen.

Theorem 2.7.9. *Seien $f, g : \mathbb{R} \longrightarrow \mathbb{R}$ stetig differenzierbar und seien $x_0, t_0 \in \mathbb{R}$ beliebig. Dann ist das AWP*

$$x'(t) = g(t) \cdot f(x(t)) \quad \text{und} \quad x(t_0) = x_0$$

eindeutig lösbar. Es gilt wieder, dass sich zwei verschiedene Lösungen der DGL nicht schneiden können.

Lemma 2.7.6 und Theorem 2.7.7 zusammen ermöglichen es nun Lösungen von autonomen DGLen zu skizzieren, ohne diese genau zu berechnen.

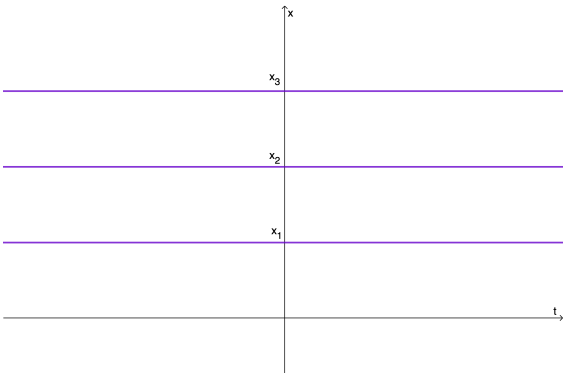
Beispiel 2.7.10. Sei $f(x)$ eine Funktion mit drei reellen Nullstellen $x_1 < x_2 < x_3$, die (INT-INF) erfüllt. Die Vorzeichen von f seien so verteilt

$$f(x) \begin{cases} < 0 & \text{falls } x < x_1 \\ > 0 & \text{falls } x_1 < x < x_2 \\ < 0 & \text{falls } x_2 < x < x_3 \\ > 0 & \text{falls } x_3 < x \end{cases}.$$

Wir skizzieren nun in Abhängigkeit von $x_0 \in \mathbb{R}$ Lösungen des AWP

$$x'(t) = f(x(t)) \quad \text{und} \quad x(0) = x_0.$$

Die stationären Punkt (konstanten Lösungen) der DGL sind genau die Nullstellen x_1, x_2, x_3. Falls also $x_0 \in \{x_1, x_2, x_3\}$ ist, haben wir eine konstante Lösung. Diese konstanten Lösungen können wir schon einmal in ein Koordinatensystem einzeichen.



Hier haben wir für die Skizze angenommen, dass alle Nullstellen positiv sind. Beachten Sie, dass wir eine Funktion $x(t)$ in der Variablen t betrachten. Die Achse in der horizontalen ist also die „t-Achse“ und die Achse in der vertikalen ist die „x-Achse“.

Wir haben in Theorem 2.7.7 und Bemerkung 2.7.8 festgestellt, dass sich zwei Lösungen der DGL nicht schneiden können.

Ist nun $\mathbf{x_0 \in (x_1, x_2)}$, dann ist die zugehörige Lösung $x(t)$ zwischen den stationären Lösungen x_1 und x_2 gefangen. Das bedeutet aber, dass $x(t)$ beschränkt ist und somit nach Theorem 2.7.7 (c) auf ganz $\mathbb{R}$ definiert. Teil (c) von Theorem 2.7.7 besagt, dass $x(t)$ monoton ist. Da wir wissen, dass $x'(x_0) = f(x_0) > 0$ ist, ist also $x(t)$ eine streng monoton wachsende Funktion, die auf ganz $\mathbb{R}$ definiert ist. Sie ist nach oben beschränkt durch x_2 und nach unten durch x_1.

Da $x(t)$ beschränkt und monoton ist, existieren $\lim_{t\to\infty} x(t)$ und $\lim_{t\to-\infty} x(t)$ in den reellen Zahlen. Nach Lemma 2.7.6 sind diese Werte stationäre Punkte der DGL – also in $\{x_1, x_2, x_3\}$. Aufgrund der Monotonie muss damit gelten

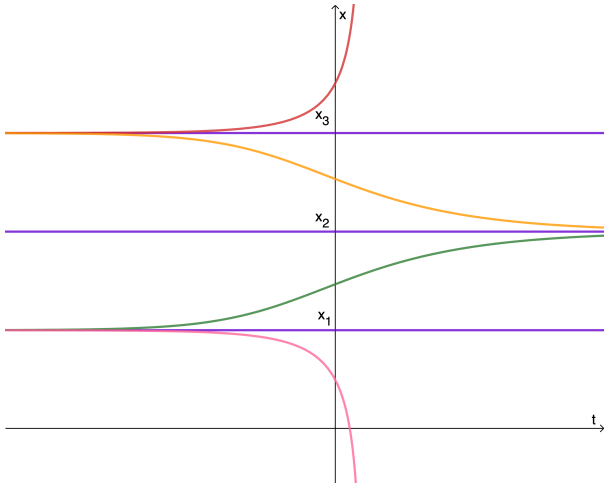
$$\lim_{t\to\infty} x(t) = x_2 \quad \text{und} \quad \lim_{t\to-\infty} x(t) = x_1.$$

Diese Lösung ist im folgenden Bild grün skizziert.

Genau so argumentieren wir, wenn $\mathbf{x_0 \in (x_2, x_3)}$ ist. Dann ist die Lösung $x(t)$ eine streng monoton fallende Funktion mit den Grenzwerten $\lim_{t\to\infty} x(t) = x_2$ und $\lim_{t\to-\infty} x(t) = x_3$. Diese ist orange eingezeichnet.

Ist nun $\mathbf{x_0 > x_3}$, so sehen wir immer noch, dass dass die zugehörige Lösung $x(t)$ streng monoton wächst. Allerdings ist $\lim_{t\to\infty} = \infty$, denn ein reeller Limes wäre eine stationäre Lösung größer als x_3. So eine stationäre Lösung existiert aber nicht! Damit ist $x(t)$ zwar nach unten beschränkt durch x_3

(was wie eben auf $\lim_{t\to-\infty} = x_3$ führt) aber nach oben unbeschränkt. Ob $x(t)$ dennoch auf ganz $\mathbb{R}$ definiert ist, verrät uns Theorem 2.7.7 nicht. Ein möglicher Verlauf dieser Funktion ist rot skizziert.
Im letzten Fall ist $x_0 < x_1$. In diesem Fall ist die Lösung $x(t)$ streng monoton falled, unbeschränkt nach unten, und mit $\lim_{t\to-\infty} = x_1$. Dieser Verlauf ist rosa eingezeichnet.



2.8 Die logistische Differenzialgleichung

Wir möchten das Wachstum einer Population beschreiben. Bisher haben wir dafür das exponentielle Wachstum kennengelernt. Dieses ist gegeben durch die DGL $x'(t) = p \cdot x(t)$, mit $p \in \mathbb{R}$. Hier beschriebt $x(t)$ die Größe einer Population zum Zeitpunkt t. Ist $p < 0$ stirbt die Population aus, ist $p = 0$ dann bleibt die Populationsgröße konstant, ist $p > 0$ dann wächst die Population über alle Maßen. Diese Beschreibung des Populationswachstums ist als nur für kurze Zeitintervalle geeignet.
Wir suchen hier nach einem realistischeren Modell für die Beschreibung des Wachstums einer gegebenen Population. Sei wieder $x(t)$ die Populationsgröße zum Zeitpunkt t.

1. Versuch *Einfachstes mathematisches Modell:* Die Konstante p aus dem exponentiellen Wachstum gibt die Wachstumsrate der Population an. Diese teilen wir nun auf in eine Geburtenrate und eine Sterberate. Die Geburtenrate soll immer noch proportional zur Populationsgröße sein. D.h. verdoppelt

sich die Population, so sollte es auch doppelt so viele Geburten geben.
Bei der Sterberate ist diese Proportionalität nicht mehr sinnvoll. Wenn es mehr Individuen gibt, dann hat jedes einzelne weniger Platz, weniger Futter, ... Damit sollte sich ein Anwachsen der Population zunehmend schlechter auf die Sterberate auswirken. Die Sterberate sollte daher proportional sein zu einem $f(x(t))$, wobei $f(x)$ schneller wächst als eine lineare Funktion. Die einfachste Funktion, die wir dafür wählen können ist $f(x) = x^2$.
Zusammen erhalten wir im ersten Versuch die folgende DGL, die das Wachstum der Population beschreiben soll:

$$x'(t) = \underbrace{r \cdot x(t)}_{\text{Geburtenrate}} - \underbrace{s \cdot x(t)^2}_{\text{Sterberate}}, \quad \text{mit } r, s \in \mathbb{R}. \tag{2.17}$$

2. Versuch *Kapazitätsgrenze:* Wir nehmen an, dass es eine theoretische Obergrenze $K > 0$ für die Populationsgröße gibt. Dann sollte das Wachstum der Population so wohl proportional zur aktuellen Populationsgröße $x(t)$ sein, als auch zum verbleibenden Platz bis zur Kapazitätsgrenze $(K - x(t))$. Diese Idee führt auf die DGL

$$x'(t) = \lambda \cdot x(t) \cdot (K - x(t)), \quad \text{mit } \lambda, K \in \mathbb{R}. \tag{2.18}$$

3. Versuch *Konkurrenz:* Wir starten mit dem exponentiellen Wachstum $x'(t) = p \cdot x(t)$. Wir wollen dieses Modell mit dem negativen Einfluss erweitern, den die Konkurrenz innerhalb der Population verursacht. Dabei ist die Konkurrenz um Futter, Verstecke, und andere Ressourcen gemeint (siehe 1. Versuch). Ganz grob entsteht Konkurrenz dann, wenn zwei Individuen aufeinandertreffen. Für ein festes Individuum (nennen wir es Heinz) gehen wir davon aus, dass die Anzahl von Treffen innerhalb eines festen Zeitintervalls (sagen wir innerhalb eines Tages) proportional zur Populationsgröße ist. Dann ist diese Anzahl gegeben durch $a'' \cdot x(t)$. D.h. wenn es doppelt so viele Individuen gibt, dann trifft Heinz pro Tag auch doppelt so viele Artgenossen. Die Anzahl aller Treffen zweier beliebiger Individuen pro Zeitintervall ist dann $\frac{a'' \cdot x(t) \cdot x(t)}{2} = a' \cdot x(t)^2$. Der negative Einfluss dieser Treffen sollte also proportional zu dieser Größe sein. Damit erhalten wir die DGL

$$x'(t) = p \cdot x(t) - a \cdot x(t)^2, \quad \text{mit } p, a \in \mathbb{R}. \tag{2.19}$$

Vergleichen wir die drei Differenzialgleichungen (2.17), (2.18) und (2.19) stellen wir fest, dass es sich immer um dieselbe DGL handelt! Das halten wir in einer Definition fest.

Definition 2.8.1. Für $\lambda, K \in \mathbb{R}$ heißt die DGL $x'(t) = \lambda \cdot x(t) \cdot (K - x(t))$ die *logistische Differenzialgleichung*. Eine Lösung dieser DGL beschreibt das *logistische Wachstum*.

Bemerkung 2.8.2. Die logistische DGL können wir mit Leichtigkeit analysieren! Es ist eine autonome DGL mit rechter Seite $f(x) = \lambda x(K - x)$; ein quadratisches Polynom. Damit erfüllt f die Bedingung (INT-INF). Die Nullstellen von f sind 0 und K. Das sind somit die stationären Punkte, also die konstanten Lösungen der DGL. Unter der Annahme, dass $\lambda, K > 0$ ist, haben wir

$$f(x) \begin{cases} < 0 & \text{falls } x < 0 \\ > 0 & \text{falls } 0 < x < K \\ < 0 & \text{falls } x > K \end{cases}.$$

Genau wie im letzten Abschnitt folgt damit

$$\lim_{t \to \infty} x(t) = \begin{cases} -\infty & \text{falls } x(0) < 0 \\ K & \text{falls } 0 < x(0) < K \\ K & \text{falls } x > K \end{cases}.$$

Aus der Sicht, dass $x(t)$ ein Populationswachstum beschreibt, ist der erste Fall ausgeschlossen, da es keine negative Populationsgröße $x(0) < 0$ geben kann. Mit diesem Wachstumsmodell wird sich eine Population auf lange Sicht also immer ihrer Kapazitätsgrenze annähern. Das erscheint deutlich realistischer als das ungehemmte exponentielle Wachstum.

Natürlich können wir die Lösungen der logistischen DGL auch explizit berechnen. Sei also das folgende AWP gegeben

$$x'(t) = \lambda \cdot x(t) \cdot (K - x(t)) \quad \text{und} \quad x(t_0) = x_0$$

mit $\lambda, K \in \mathbb{R} \setminus \{0\}$ und $t_0, x_0 \in \mathbb{R}$ beliebig. Falls $x_0 = 0$ oder $x_0 = K$ haben wir nur die konstante Lösung. Wir nehmen also an, dass $0 \neq x_0 \neq K$ gilt. Dann wissen wir insbesondere (Lösungen können sich nicht schneiden), dass

für die Lösung $x(t)$ des AWPs immer $0 \neq x(t) \neq K$ gilt. Jetzt rattern wir das übliche Lösungsverfahren für DGLen mit getrennten Variablen herunter.

$$\begin{aligned}
& x'(t) = \lambda \cdot x(t) \cdot (K - x(t)) \\
\iff \quad & \frac{x'(t)}{x(t) \cdot (K - x(t))} = \lambda \\
\iff \quad & \int \frac{x'(t)}{x(t) \cdot (K - x(t))} \mathrm{d}t = \int \lambda \mathrm{d}t = \lambda \cdot t + c, \quad c \in \mathbb{R} \\
\iff \quad & \int \frac{1/K}{x} + \frac{1/K}{K - x} \mathrm{d}x = \int \frac{1}{x(K - x)} \mathrm{d}x = \lambda \cdot t + c, \quad c \in \mathbb{R} \\
\iff \quad & \frac{1}{K} \ln |x(t)| - \frac{1}{K} \ln |K - x(t)| = \lambda \cdot t + c, \quad c \in \mathbb{R} \\
\iff \quad & \ln \left| \frac{x(t)}{K - x(t)} \right| = \lambda K \cdot t + c, \quad c \in \mathbb{R} \\
\iff \quad & \left| \frac{x(t)}{K - x(t)} \right| = e^{\lambda K \cdot t + c} = e^{\lambda K \cdot t} \cdot c, \quad c > 0.
\end{aligned}$$

Hier sollten wir eine Kleinigkeit erklären. Das c auf der rechten Seite ist nicht immer dasselbe. In der vorletzten Zeile müsste eigentlich Kc stehen, aber auch das ist wieder irgendeine reelle Konstante, die wir einfach wieder c nennen. In der letzten Zeile ist das c eigentlich ein e^c. Der Wert muss nun aber positiv sein, da die Exponentialfunktion nur positive Werte annimmt. Ansonsten ist es genau das gleiche Verfahren, wie immer.
Bis auf die Betragsstriche kann man diese Gleichung leicht nach $x(t)$ auflösen. Wir wissen, dass $x(t)$ nie die konstanten Funktionen 0 und K schneiden kann. Dadurch ist das Vorzeichen von $\left| \frac{x(t)}{K - x(t)} \right|$ immer das gleiche. Der Betrag wird also gebildet in dem entweder mit 1 oder mit -1 multipliziert wird. Diesen Faktor können wir aber einfach wieder in der unbestimmten Konstanten c zusammenfassen. Es folgt also

$$x'(t) = \lambda \cdot x(t) \cdot (K - x(t)) \iff \frac{x(t)}{K - x(t)} = e^{\lambda K \cdot t} \cdot c, \quad c \in \mathbb{R} \setminus \{0\}.$$

Nach $x(t)$ auflösen liefert, dass die Funktion $x(t)$ genau dann die DGL löst, wenn

$$x(t) = \frac{K}{e^{-\lambda K t} \cdot c + 1}, \quad c \in \mathbb{R} \setminus \{0\}.$$

Jetzt finden wir noch die Konstante c für die der Anfangswert $x(t_0) = x_0$ erfüllt ist. Es muss also gelten

$$x_0 = x(t_0) = \frac{K}{e^{-\lambda K t_0} \cdot c + 1}.$$

Umstellen liefert $c = \frac{K-x_0}{x_0} \cdot e^{\lambda K t_0}$. Die eindeutige Lösung des AWPs ist somit

$$x(t) = \frac{K}{\frac{K-x_0}{x_0} \cdot e^{-\lambda K(t-t_0)} + 1}.$$

Das halten wir noch einmal formal fest.

Proposition 2.8.3. *Seien $\lambda, K \in \mathbb{R} \setminus \{0\}$ beliebig. Die nicht konstanten Lösungen der logistischen DGL $x'(t) = \lambda \cdot x(t) \cdot (K - x(t))$ sind gegeben durch $x(t) = \frac{K}{e^{-\lambda K t} \cdot c + 1}$ für $c \in \mathbb{R} \setminus \{0\}$. Die eindeutige Lösung des AWPs*

$$x'(t) = \lambda \cdot x(t) \cdot (K - x(t)) \quad \text{und} \quad x(t_0) = x_0 \notin \{0, K\}$$

ist

$$x(t) = \frac{K}{\frac{K-x_0}{x_0} \cdot e^{-\lambda K(t-t_0)} + 1}.$$

Beispiel 2.8.4. Jetzt können wir uns noch einmal das Beispiel 2.5.3 der Bakterienpopulation in der Petrischale anschauen.

> *In einer Petrischale seien $200mm^2$ von Bakterien besiedelt. Es sollen ohne Hinzufügen von Bakterien von außen, $500mm^2$ werden. Nach einem Tag sind bereits $230mm^2$ besiedelt. Wann ist das Ziel erreicht?*

Diesmal geben wir aber noch die Information an, dass die Petrischale $1000mm^2$ großist. Wir kennen also die Kapazitätsgrenze $K = 1000$ für die Bakterienpopulation $x(t)$. Wir wählen daher das Modell des logistischen Wachstums, mit $K = 1000$ und dem Anfangswert $x(0) = 200$. Mit Proposition 2.8.3 wissen wir, dass dann

$$x(t) = \frac{1000}{\frac{1000-200}{200} \cdot e^{-\lambda \cdot 1000 \cdot t} + 1} = \frac{1000}{4 \cdot e^{-\lambda \cdot 1000 \cdot t} + 1}.$$

Wir haben noch die Information $x(1) = 230$. Damit können wir λ berechnen. Es ist

$$\lambda = -\ln\left(\frac{770}{4 \cdot 230}\right) \cdot \frac{1}{1000}.$$

Jetzt kennen wir die gesuchte Funktion:

$$x(t) = \frac{1000}{4 \cdot \left(\frac{770}{920}\right)^t + 1}.$$

Um die Frage zu beantworten, müssen wir das $t > 0$ finden, mit $x(t) = 500$. Es folgt $t = 7{,}788\ldots$ Mit dem Wissen über die Größe der Petrischale vermuten wir also, dass das Ziel nach ca. 7,78 Tagen erreicht ist.

Bemerkung 2.8.5. In der Realität wird es selten so sein, dass die Population ihre Kapazitätsgrenze monoton annähert. Meistens wird die Population erst merken, was die Populationsgrenze gewesen ist, wenn der Bestand wieder zurückgeht. Daher ist eine oszillierende Annäherung an die Kapazitätsgrenze K oft realistischer.

Bemerkung 2.8.6. Wir sind davon ausgegangen, dass wir die Populationsgröße als stetige Funktion darstellen können. Das ist faktisch nicht möglich, da wir immer Sprünge in der Funktion haben werden. Bei großen Beständen sind diese Sprünge allerdings zu vernachlässigen. Sie können sich auch vorstellen, dass wir (wie ei den Bakterien) die Größe in Volumen oder Biomasse messen.
Wenn wir stattdessen tatsächlich exakt die einzelnen Individuen in Jahren oder Monaten zählen wollen, erhalten wir ein diskretes Modell. Dieselben überlegungen wie eben führen dann auf die *logistische Differenzengleichung*

$$a_n = \lambda \cdot a_{n-1} \cdot (K - a_{n-1}) \quad \forall n \in \mathbb{N}.$$

Eine geschlossene Formel für diese Differenzengleichung lässt sich nicht finden. Tatsächlich führt diese Rekursion zu einem sehr chaotischen Verhalten der Folge $(a_n)_{n\in\mathbb{N}_0}$. Daher ist sie ein guter Startpunkt für das mathematische Studium des Chaos. Das werden wir in dieser Vorlesung aber nicht weiter vertiefen.

Kapitel 3

Entdimensionalisierung

Hier werden wir ein Hilfsmittel kennenlernen, dass man auch „fortgeschrittenes Skalieren“ nennen könnte. Damit werden wir später auch etwas komplexere Populationsmodelle behandeln können.

3.1 Dimensionen und ihre Rechenregeln

Das Beispiel mit den Zinsen (Problem 1.2.1) kann unverändert gelöst werden, wenn wir die Währung € durch eine andere ersetzen. Wir können auch, wie in Problem 2.5.1, die Währung € durch Schafe ersetzen. Die Mathematik kümmert sich nicht um Einheiten. Wenn ein Problem also einfach wird, wenn wir eine andere (auch exotische) Einheit wählen, dann sollten wir mit dieser Einheit rechnen.

Notation 3.1.1. Die Dimension eines Wertes a bezeichnen wir mit $[a]$. Durch die Bestimmung der Dimension wird noch keine Einheit gewählt! Übliche Dimensionen sind in der folgenden Tabelle zusammengefasst.

Name	Bezeichnung	mögliche Einheiten
Länge	L	m, cm, mm, ft, yd, ...
Zeit	T	Sekunden, Tage, π Jahre, ...
Masse	M	kg, oz, ...
Anzahl	A_x	Fische, Studierende, ...

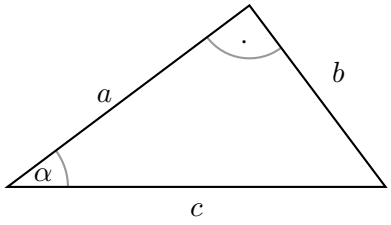
Besitzt ein Wert a keine Dimension setzen $[a] = 1$ und nennen a *dimensionslos*. Die Dimension Anzahl fällt etwas aus der Reihe. Wir wollen die

Dimensionen in erster Linie für Populationsmodelle nutzen, in denen zwei Spezies interagieren. Diese Spezies sollen auch in Ihren Dimensionen klar getrennt sein, weshalb wir nicht einfach beide in der Dimension Masse zusammenfassen möchten. Dadurch bekommt die Anzahl immer noch einen Index, etwa A_F für die Anzahl von Fischen, etc.
Das sind nur die Dimensionen, die für unsere Vorlesung entscheidend sind. In der Physik benötigt man auch die Dimensionen Stromstärke I, Lichtstärke J, Temperatur θ, Stoffmenge N.

Bemerkung 3.1.2. Beim Rechnen mit Werten, die eine Dimension besitzen müssen wir etwas aufpassen. Es gilt:

- Werte a und b können nur dann addiert werden, wenn $[a] = [b]$ gilt. Ausdrücke wie $7kg + 8m$ sind nicht definiert. Gilt hingegen $[a] = [b]$, so ist $[a + b] = [a]$. Es ist also $L + L = L$, $M + M = M$, usw.
- Bei der Multiplikation ist nichts zu berücksichtigen. Es gilt einfach $[a \cdot b] = [a] \cdot [b]$. Weiter setzen wir $[a^{-1}] = \frac{1}{[a]}$. Beides kennen Sie bereits. Es ist $[(7h)^{-1}] = \frac{1}{T}$ und damit $[\frac{8km}{7h}] = \frac{L}{T}$. Die Geschwindigkeit $\frac{L}{T}$ kann also aus den Dimensionen von oben zusammengesetzt werden.
- Daraus folgt, dass Dimensionen alle Potenzgesetze erfüllen.

Beispiel 3.1.3. Diese Dimensionsregeln können benutzt werden um den Satz des Pythagoras zu beweisen. Dazu betrachten wir ein rechtwinkliges Dreieck mit Hypotenuse c und kleinstem Winkel α.

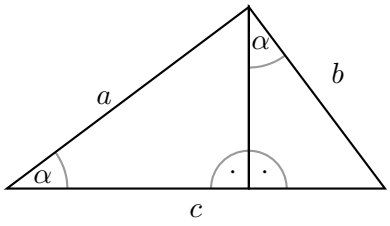


Die Dimension von a, b und c ist klar: es handelt sich um Längenangaben, daher ist $[a] = [b] = [c] = L$. Der Flächeninhalt F von dem Dreieck hat die Dimension $[F] = L^2$.
Was ist die Dimension von α? Dazu überlegen wir uns erst $[\cos(\alpha)] = [\frac{a}{c}] = \frac{L}{L} = 1$. Damit ist aber auch $[\alpha] = [\arccos(\cos(\alpha))] = 1$. Damit ist α – wie jeder Winkel – dimensionslos.

Durch die Angabe von c und α ist das rechtwinklige Dreieck bis auf Spiegelung eindeutig bestimmt. Insbesondere lässt sich in jedem rechtwinkligen Dreieck der Flächeninhalt berechnen durch die Werte c und α. Damit kann F als Funktion in c und α aufgefasst werden. Wir schreiben daher $F = F(c, \alpha)$. Es gibt aber nur eine Möglichkeit aus den Dimensionen von c und α die Dimension von F zu erhalten. Es ist

$$[F] = L^2 = L^2 \cdot 1 = [c]^2 \cdot [\alpha] = [c^2 \cdot \alpha].$$

Wir schließen, dass $F(c, \alpha) = c^2 \cdot f(\alpha)$, für eine gewisse Funktion $f : (0, \frac{\pi}{4}] \longrightarrow (0, \infty)$[1]. Der dimensionsbehaftete Teil c muss quadriert und mit einem dimensionslosen Wert multipliziert werden. Dies gilt natürlcih für alle rechtwinkligen Dreiecke. Insbesondere also für die rechtwinkligen Dreiecke, die wir erhalten wenn wir das Ausgangsdreieck entlang der Höhe zerschneiden.



Es folgt $F(c, \alpha) = F(a, \alpha) + F(b, \alpha)$, was nach unserer Vorüberlegung nichts anderes ist als

$$\begin{aligned} & c^2 \cdot f(\alpha) = a^2 \cdot f(\alpha) + b^2 \cdot f(\alpha) \\ \overset{f(\alpha)\neq 0}{\Longleftrightarrow} \quad & c^2 = a^2 + b^2. \end{aligned}$$

Damit haben wir tatsächlich den Satz des Pythagoras erhalten.

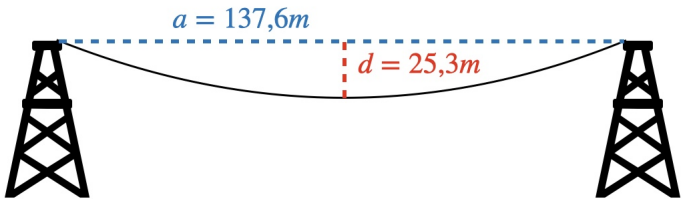
3.2 Das hängende Kabel

Das folgende Problem wird uns etwas länger beschäftigen. Es dient auch der Erklärung, warum Entdimensionalisierung nützlich sein könnte.

[1] Es ist $f(\alpha) = \frac{1}{2} \cdot \sin(\alpha) \cdot \cos(\alpha)$, aber hier geht es nur darum, dass wir aus der Dimensionsanalyse auf die Existenz einer solchen Funktion schließen können

Problem 3.2.1. Ein Kabel soll zwischen zwei Strommasten, die einen Abstand von $a = 137{,}6\ m$ haben aufgespannt werden. Dabei soll es in der Mitte genau $d = 25{,}3\ m$ durchhängen. Wie lang muss das Kabel dafür sein?

Hier ist eine Skizze zum beschriebenen Problem:



Wir machen ein paar Annahmen:

- Die Masten stehen ebenerdig und auf gleicher Höhe.
- Kabeldicke und -festigkeit wird vernachlässigt. Wir nehmen also an, dass alle Verformungen des Kabels möglich sind.
- Die Länge des Kabels nehmen wir als konstant an. Dehnung durch Gewicht und Wetter spielt keine Rolle.

Wir gehen natürlcih davon aus, dass das Kabel nicht gegen die Schwerkraft ankommt. Die Aufhängungen an den Masten stellen daher die einzigen lokalen Maxima des Kabelverlaufes dar. Unter diesen Annahmen vergleichen wir ein paar mathematische Modelle für das Kabel.

Modell 1: (Rechteck) In diesem naiven Modell fällt das Kabel senkrecht auf den Durchhang von 25,3 m, verläuft dann waagerecht zum Boden bis zum anderen Mast und steigt dann wieder senkrecht nach oben. In diesem Modell ist die Kabellänge gleich $(2 \cdot d + a)\ m = 188{,}2\ m$.

Modell 2: (Dreieck) Jetzt nehmen wir an, dass das Kabel durch einen Verlauf beschrieben wird in dem es vom ersten Mast bis zur Mitte linear auf den Durchhang $d = 25{,}3\ m$ fällt und danach wieder linear ansteigt. In diesem Modell beträgt die Länge des Kabels $2 \cdot \sqrt{(a/2)^2 + d^2} = 146{,}6 \ldots m$.

Bemerkung 3.2.2. Diese beiden Modelle sind nicht realistisch. Unter unseren Annahmen haben wir aber trotzdem etwas wichtiges herausgefunden: Die Länge des Kabels liegt zwischen $146{,}6\ldots m$ und $188{,}2\ m$. Wenn uns viel Verschnitt bei den Kabeln nichts ausmacht, können wir mit diesen Angaben gut arbeiten.

Bemerkung 3.2.3. In den kleinen Rechnungen der Modelle 1 und 2 ist uns schon (unbemerkt) eine Entdimensionalisierung über den Weg gelaufen. Wir haben nämlich erst gerechnet (z.B. $2 \cdot \sqrt{(a/2)^2 + d^2} = 146{,}6\ldots$) und dann haben wir einfach die Einheit m hinzugefügt. Über die Rechenregeln mit Dimensionen haben wir uns keine Gedanken gemacht. Was eigentlich passiert ist, ist folgendes.
Wir haben a durch $\frac{a}{1\ m}$ und d durch $\frac{d}{1\ m}$ ersetzt. Diese Werte sind dimensionslos und wir können mit ihnen rechnen ohne auf Dimensionsregeln zu achten. Wir haben nun die „dimensionslose Länge“ berechnet:

$$\frac{\ell}{1\ m} = \frac{2 \cdot \sqrt{(a/2)^2 + d^2}}{1\ m} = 2 \cdot \sqrt{\frac{1}{4}(\frac{a}{1\ m})^2 + (\frac{d}{1\ m})^2} = 146{,}6\ldots = \lambda_D.$$

Daraus folgt natürlich $\ell = 146{,}6\ldots \cdot 1\ m$. Würde diese Aufgabe in den USA gestellt, würde dort so gerechnet

$$\frac{\ell}{1\ ft} = \frac{2 \cdot \sqrt{(a/2)^2 + d^2}}{1\ ft} = 2 \cdot \sqrt{\frac{1}{4}(\frac{a}{1\ ft})^2 + (\frac{d}{1\ ft})^2} = 481{,}00\ldots\ldots = \lambda_{USA}.$$

Es ist also $\ell = 481 \cdot 1\ ft$. Genauso wie mit $1\ ft$ können wir mit jedem Wert der Dimension L rechnen. Sei dazu $\bar{\ell}$ irgendein Wert mit $[\bar{\ell}] = L$. Dann ist

$$\frac{\ell}{\bar{\ell}} = 2 \cdot \sqrt{\frac{1}{4}(\frac{a}{\bar{\ell}})^2 + (\frac{d}{\bar{\ell}})^2}.$$

Für $\bar{\ell} = a$ erhalten wir einfach

$$\frac{\ell}{a} = 2 \cdot \sqrt{\frac{1}{4} + (\frac{d}{a})^2} = \lambda_{d/a}. \tag{3.1}$$

Dieser Wert ist dimensionslos und damit unabhängig von der Wahl der Einheit. Das ermöglicht es uns das Modell zu skalieren. Mit (3.1) folgt, dass die Länge des Kabels *in diesem Modell* gleich $\lambda_{d/a} \cdot a$ ist. Wäre also $a = 1\ m$, so müsste das Modell eine Länge von $\lambda_{d/a}\ m$ liefern. Das könnten wir schnell überprüfen – was wir hier nicht machen werden, da wir schon wissen, dass das Modell nicht realistisch ist.

Die Modelle 1 und 2 sind unter anderem deshalb unrealistisch, da das Kabel in diesen Modellen Knicke hat. Die erwarten wir nicht. Wir gehen also ab jetzt davon aus, dass der Verlauf des Kabels glatt ist, also durch eine differenzierbare Funktion beschrieben wird.

Modell 3: (Parabel) Wir versuchen einen parabelförmigen Verlauf. Dazu setzen wir Koordinaten fest. Der Tiefpunkt soll bei $(0,0)$ angenommen werden. Aus der Symmetrie folgt dann, dass die Aufhängungspunkte des Kabels bei den Punkten $(-a/2, d)$ und $(a/2, d)$ liegen. Es gibt genau eine Parabel, die durch drei gegebene Punkte verläuft. In diesem Fall sehen wir sofort, dass diese gegeben ist durch $f(x) = \frac{4d}{a^2} \cdot x^2$, auf dem Intervall $[-\frac{a}{2}, \frac{a}{2}]$.
Dieses Modell wollen wir nun entdimensionalisieren. Dazu listen wir als erstes die Dimensionen von allen beteiligten Werten auf:

$$[4] = 1 \quad ; \quad [d] = [a] = [x] = [f(x)] = L.$$

Sei nun $\bar{\ell}$ beliebig mit $[\bar{\ell}] = L$. Dann ist die dimensionslose Variable gegeben durch $X = \frac{x}{\bar{\ell}}$. Die dimensionslose Funktion in dieser Variablen ist

$$F(X)\frac{f(x)}{\bar{\ell}} = \frac{f(X \cdot \bar{\ell})}{\bar{\ell}} = \frac{4d\bar{\ell}}{a^2} \cdot X^2, \text{ mit } X \in [-\frac{a}{2\bar{\ell}}, \frac{a}{2\bar{\ell}}].$$

Für $\bar{\ell} = \frac{a}{4}$ ist $F(X) = \frac{d}{a}X^2$, mit $X \in [-2, 2]$. Damit können wir die entdimensionalisierte Länge des Kabels berechnen. Es ist

$$\begin{aligned} \lambda_{d/a} &= \int_{-2}^{2} \sqrt{1^2 + F'(X)^2}\mathrm{d}X = \frac{a}{2d} \cdot \int_{-2}^{2} \sqrt{1 + \left(2\frac{d}{a}X\right)^2} \cdot \frac{2d}{a}\mathrm{d}X \\ &= \frac{a}{2d} \cdot \int_{-\frac{4d}{a}}^{\frac{4d}{a}} \sqrt{1 + X^2}\mathrm{d}X. \end{aligned} \tag{3.2}$$

Eine Stammfunktion von $\sqrt{1 + X^2}$ haben evtl nicht alle sofort zur Hand, daher begnügen wir uns für den Moment mit einer numerischen Berechnung, die uns ein CAS liefert. Es ist

$$\lambda_{d/a} = 4{,}3359\ldots$$

Die Kabellänge in diesem Modell ist also

$$\lambda_{d/a} \cdot \bar{\ell} = \lambda_{d/a} \cdot \frac{a}{4} = 149{,}154\ m.$$

Dieser Wert liegt immerhin in den Grenzen, die wir uns in Bemerkung 3.2.2 überlegt haben.

Bemerkung 3.2.4. Es ist wichtig zu beachten, dass wir den parabelförmigen Lauf des Kabels einfach angenommen haben. Wir wissen nicht, ob das Kabel tatsächlich so verläuft. Das können wir nun aber experimentell überprüfen. Unter der Annahme, dass der Abstand der beiden Masten $a' = 0{,}5\ m$ wäre, so müsste der Durchhang $d' = \frac{d}{a} \cdot a' = 0{,}0919\ldots\ m$ sein (hier brauchen wir wieder, dass $\frac{d}{a}$ dimensionslos ist). Die Länge des Kabels im Modell des parabelförmigen Verlaufs wäre $\lambda_{d/a} \cdot \frac{a'}{4} = 0{,}5419\ldots\ m$. Das können wir aber in einem kleinen Modell überprüfen.
In unserem Experiment in der Vorlesung haben wir eine tatsächliche Kabellänge von 54 *cm* herausbekommen. Wir haben relativ grob gearbeitet und eine millimetergenaue Messung kann nicht vorgenommen worden sein. Dennoch stellen wir fest, dass das Ergebnis nah an der berechneten Lösung liegt. Es kann also in diesem groben Modell nicht ausgeschlossen werden, dass die Form des Kabels tatsächlich parabelförmig ist. Spoiler: Wenn wir ganz genau gearbeitet hätten, hätten wir sicher einen Unterschied bemerkt. Im folgenden versuchen wir zu erklären, wie man auf die tatsächliche Form des Kabels kommt.

Bevor wir uns überlegen, wie ein physikalisch begründetes Modell aussehen kann, möchten wir uns noch um die exakte Lösung des Integrals aus (3.2) kümmern. Dazu machen wir einen kleinen Schlenker

Erinnerung 3.2.5. Die komplexen Zahlen $\mathbb{C}$ bestehen aus allen Elementen der Form $z = a + b \cdot i$, mit $a, b \in \mathbb{R}$ und $i^2 = -1$. Das Element i wird auch *imaginäre Einheit* genannt. Die reelle Zahl a heißt *Realteil von* z und wird mit $\mathrm{Re}(z)$ bezeichnet. Die reelle Zahl b heißt *Imaginärteil von* z und wird mit $\mathrm{Im}(z)$ bezeichnet. Es ist nicht sonderlich überraschend, dass zwei komplexe Zahlen genau dann gleich sind, wenn sie den gleichen Real- und den gleichen Imaginärteil haben.
Mit den folgenden Rechenoperationen werden die komplexen Zahlen zu einem Körper. D.h. man kann mit Ihnen genau so rechnen, wie wir auch auf den reellen Zahlen rechnen können.

$$\begin{aligned} + :&\quad (a + b \cdot i) + (c + d \cdot i) = (a + c) + (b + d) \cdot i \\ \cdot :&\quad (a + b \cdot i) \cdot (c + d \cdot i) = (ac - bd) + (ad + bc) \cdot i \end{aligned}$$

Die Menge $\mathbb{C}$ enthält alle reellen Zahlen (das sind genau die komplexen Zahlen, mit Imaginärteil gleich Null). Anschaulich erweitern sie dadurch

auch die reelle Zahlengerade. Dafür identifizieren wir eine komplexe Zahl z mit dem Punkt $(\mathrm{Re}(z), \mathrm{Im}(z))$. Die horizontale Achse nennen wir reelle-Achse und die vertikale imaginäre-Achse.
Die *komplex-konjugierte Zahl zu* $z = a+b\cdot i$ ist gegeben durch $\bar{z} = a-b\cdot i$. Der *Betrag* von $z = a + b \cdot i$ ist der Abstand von z zum Nullpunkt und wird mit $|z|$ bezeichnet. Mit dem Satz des Pythagoras und der dritten binomischen Formel erhalten wir

$$|z| = \sqrt{a^2 + b^2} = \sqrt{(a + b \cdot i)(a - b \cdot i)} = \sqrt{z \cdot \bar{z}}.$$

Ist $z \in \mathbb{C} \setminus \{0\}$, so gilt damit $1 = z \cdot \frac{\bar{z}}{|z|^2}$.
Verbinden wir eine komplexe Zahl $z \in \mathbb{C} \setminus \{0\}$ mit dem Nullpunkt, schließt diese Verbindungsstrecke einen Winkel α mit der reellen Achse ein (von der reellen Achse gegen den Uhrzeigersinn gemessen). Mit diesem α gilt

$$z = |z| \cdot (\cos(\alpha) + i \cdot \sin(\alpha)) .$$

Das ist die sogenannte Darstellung in Polarkoordinaten. Das alles haben Sie (hoffentlich) in der linearen Algebra schon gelernt. Bei Bedarf gehe ich in der Vorlesung noch weiter darauf ein.

Wir brauchen noch eine Aussage über die Polarkoordinaten, die Sie hoffentlich schon kennengelernt haben. Für spätere Referenzen halten wir diese Aussage jedoch gesondert fest.

Satz 3.2.6. *Sei $\alpha \in \mathbb{R}$ beliebig. Dann gilt $e^{i\cdot\alpha} = \cos(\alpha) + i \cdot \sin(\alpha)$. Hier ist für jede komplexe Zahl z*

$$e^z = \sum_{n=0}^{\infty} \frac{z^n}{n!}.$$

Wenn man die Potenzreihendarstellung von cos und sin als bekannt voraussetzt, ist die Aussage fast offensichtlich. Wenn diese Darstellung nicht bekannt ist, dann glauben Sie mir einfach...
Aus Satz 3.2.6 folgern wir sofort, dass für jedes $\alpha \in \mathbb{R}$ gilt

$$\cos(\alpha) = \frac{1}{2} \cdot \left(e^{i\alpha} + e^{-i\alpha}\right) \qquad \text{und} \qquad \sin(\alpha) = \frac{1}{2i} \cdot \left(e^{i\alpha} - e^{-i\alpha}\right) .$$

Jetzt kommen wir erst einmal wieder zurück zu den reellen Zahlen.

Definition 3.2.7. Wir definieren die beiden folgenden Funktionen.

- $\cosh : \mathbb{R} \longrightarrow \mathbb{R} \quad ; \quad x \mapsto \frac{1}{2} \cdot (e^x + e^{-x})$
- $\sinh : \mathbb{R} \longrightarrow \mathbb{R} \quad ; \quad x \mapsto \frac{1}{2} \cdot (e^x - e^{-x})$

Die Funktion cosh heißt *Cosinus hyperbolicus* und die Funktion sinh heißt *Sinus hyperbolicus*.

Bemerkung 3.2.8. Da wir keine Angst vor der e-Funktion haben, haben wir auch keine Angst vor cosh und sinh. Insbesondere sehen wir ganz leicht, dass für alle $x \in \mathbb{R}$ die Gleichung

$$\cosh(x)^2 - \sinh(x)^2 = 1 \tag{3.3}$$

gilt. Die beiden Funktionen parametrisieren somit die Hyperbel $x^2 - y^2 = 1$. Daher kommt der Name *hyperbolicus*.

Nur mit dem Wissen über de e-Funktion zeigt man ganz leicht die folgenden Aussagen.

Lemma 3.2.9. *Es gilt*

(a) $\cosh'(x) = \sinh(x)$

(b) $\sinh'(x) = \cosh(x)$

(c) $\sinh(x) > 0$ *für alle* $x > 0$, $\sinh(0) = 0$, *und* $\sinh(x) < 0$ *für* $x < 0$

(d) $\cosh(x) > 0$ *für alle* $x \in \mathbb{R}$

(e) $\cosh(x)$ *ist streng monoton steigend auf* $(0, \infty)$ *und streng monoton fallend auf* $(-\infty, 0)$

(f) $\sinh(x)$ *ist streng monoton steigend auf ganz* $\mathbb{R}$

(g) $\sinh(-x) = -\sinh(x)$ *und* $\cosh(-x) = \cosh(x)$ *für alle* $x \in \mathbb{R}$

Definition 3.2.10. Die Umkehrfunktion von sinh auf $\mathbb{R}$ heißt *Areasinus-hyperbolicus* und wird mit arsinh bezeichnet. Die Umkehrfunktion von cosh auf $(0, \infty)$ heißt *Areacosinus-hyperbolicus* und wird mit arcosh bezeichnet.

Bemerkung 3.2.11. Für alle $x \in \mathbb{R}$ gilt $\sinh(\operatorname{arsinh}(x)) = x$. Wir können diese Gleichung somit als Gleichung zwischen Funktionen betrachten. Leiten wir beide Seiten ab erhalten wir

$$\operatorname{arsinh}'(x) \cdot \cosh(\operatorname{arsinh}(x)) = 1.$$

Umstellen der Gleichung (beachten Sie, dass $\cosh(\operatorname{arsinh}(x)) \neq 0$) ergibt nun

$$\operatorname{arsinh}'(x) = \frac{1}{\cosh(\operatorname{arsinh}(x))} \overset{(3.3)}{=} \frac{1}{\sqrt{1+\sinh(\operatorname{arsinh}(x))^2}} = \frac{1}{\sqrt{1+x^2}}.$$

Damit können wir nun das Integral, das in **Modell 3** nur numerisch berechnet wurde, auch explizit berechnen. In 3.2 haben wir festgestellt

$$\lambda_{d/a} = \frac{a}{2d} \cdot \int_{-\frac{4d}{a}}^{\frac{4d}{a}} \sqrt{1+X^2}\mathrm{d}X.$$

Es fehlte eine Stammfunktion für $\sqrt{1+X^2}$. Eine solche können wir jetzt berechnen.

$$\begin{aligned}
\int \sqrt{1+X^2}\mathrm{d}X &= \int (1+X^2)\cdot\frac{1}{\sqrt{1+X^2}}\mathrm{d}X = \int \frac{1}{\sqrt{1+X^2}}\mathrm{d}X + \int \frac{X^2}{\sqrt{1+X^2}}\mathrm{d}X \\
&= \operatorname{arsinh}(X) + \frac{1}{2}\int X\cdot\frac{2X}{\sqrt{1+X^2}}\mathrm{d}X \\
&= \operatorname{arsinh}(X) + \frac{1}{2}\left(\left[X\cdot 2\sqrt{1+X^2}\right] - \int 2\sqrt{1+X^2}\mathrm{d}X\right) \\
&= \operatorname{arsinh}(X) + \left[X\cdot\sqrt{1+X^2}\right] - \int \sqrt{1+X^2}\mathrm{d}X
\end{aligned}$$

Durch umstellen erhalten wir

$$2\int \sqrt{1+X^2}\mathrm{d}X = \operatorname{arsinh}(X) + X\cdot\sqrt{1+X^2},$$

also

$$\int \sqrt{1+X^2}\mathrm{d}X = \frac{1}{2}\operatorname{arsinh}(X) + \frac{1}{2}X\cdot\sqrt{1+X^2}.$$

Einsetzen in die Formel für $\lambda_{d/a}$ ergibt

$$\begin{aligned}
\lambda_{d/a} &= \frac{a}{2d}\cdot\left[\frac{1}{2}\operatorname{arsinh}(X) + \frac{1}{2}X\cdot\sqrt{1+X^2}\right]_{-\frac{4d}{a}}^{\frac{4d}{a}} \\
&= \frac{a}{4d}\cdot\left(\operatorname{arsinh}(\frac{4d}{a}) + \frac{4d}{a}\cdot\sqrt{1+(\frac{4d}{a})^2} - \operatorname{arsinh}(-\frac{4d}{a}) + \frac{4d}{a}\cdot\sqrt{1+(-\frac{4d}{a})^2}\right) \\
&= \frac{a}{4d}\cdot\left(2\operatorname{arsinh}(\frac{4d}{a}) + \frac{8d}{a}\cdot\sqrt{1+(\frac{4d}{a})^2}\right) = 4{,}335\ldots
\end{aligned}$$

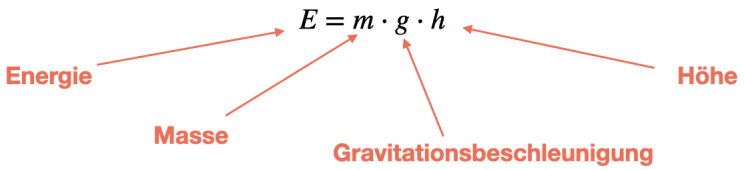
was das erste Ergebnis für $\lambda_{d/a}$ bestätigt.

Modell 4: (physikalisch begründet) Sei nun $f(x) : [-\frac{a}{2}, \frac{a}{2}] \longrightarrow \mathbb{R}$ eine Modellfunktion für den Verlauf des hängenden Kabels. Das Kabel wird keine

Knickpunkte haben. Wir können also annehmen, dass f differenzierbar ist. Die Länge des Kabels in diesem Modell ist damit

$$L(f) = \int_{-a/2}^{a/2} \sqrt{1^2 + f'(x)^2}\mathrm{d}x \in \mathbb{R}.$$

Das folgt sofort aus Theorem 1.1.3, wenn wir den Graph von f darstellen durch $x \mapsto (x, f(x))$.
Wir wollen ausnutzen, dass das Kabel einen Ruhezustand hat. Egal wie sehr wir an dem Kabel wackeln, irgendwann wird es sich wieder im gleichen hängenden Zustand befinden. Hier kommt die Physik ins Spiel. Dieser Ruhezustand ist der, bei dem die Energie der Kette minimal wird. Die Energie, die auf ein Objekt wirkt, ist in etwa „Masse mal Höhe“. Es gilt also



Die Höhe an der Stelle x ist nach Definition genau $f(x)$. Die Masse des Kabels sei gleichmässig über das Kabel verteilt, mit einem Wert von μ $\frac{kg}{m}$, $\mu > 0$. Was folgt kennen sie bereits von der Metallfeder. Wir unterteilen das Intervall $[-\frac{a}{2}, \frac{a}{2}]$ in n gleichgroße Stücke mit den Randpunkten $x_0 < x_1 < \ldots < x_n$. Dann ist die Energie des Kabels ungefähr

$$E_n(f) = \sum_{i=1}^{n} g \cdot \mu \cdot \text{„Länge von } f \text{ zwischen } x_{i-1} \text{ und } x_i\text{“} \cdot f(x_i). \qquad (3.4)$$

Diese Abschätzung wird genauer, je feiner wir die Unterteilung wählen – also je größer n wird. Die gleichen Ideen, die uns schon Theorem 1.1.3 geliefert haben, können wir nun auch hier anwenden. Die Summe (3.4) lässt sich als Riemann-Summe schreiben. Lassen wir nun n gegen unendlich streben erhalten wir genau die Energie $E(f)$ der Kette. Es ist also

$$E(f) = \lim_{n\to\infty} E_n(f) = \int_{-a/2}^{a/2} g \cdot \mu \cdot \sqrt{1 + f'(x)^2} \cdot f(x)\mathrm{d}x.$$

Das hängende Kabel wird also modelliert durch eine Funktion $f : [-\frac{a}{2}, \frac{a}{2}] \longrightarrow \mathbb{R}$, mit folgenden Eigenschaften:

- f ist differenzierbar,
- $f(-\frac{a}{2}) = f(\frac{a}{2})$ (die Masten sind gleich hoch),
- $E(f) = \int_{-a/2}^{a/2} g \cdot \mu \cdot \sqrt{1 + f'(x)^2} \cdot f(x) \mathrm{d}x$ ist minimal unter allen Funktionen, die die ersten beiden Punkte erfüllen und die feste(!) Länge L besitzen.

Die Länge des Kabels müssen wir festsetzen, da sonst tatsächlich ein rechteckiger Verlauf, bei dem das Kabel gerade auf den Boden fällt, dann auf dem Boden liegend bis zum nächsten Mast führt und dann wieder gerade nach oben führt, die geringste Energie aufweist. Das ist aber ein rein mathematisches Problem! Im wesentlichen ist es eine Extremwertaufgabe, bei der die gesuchte Variable eine Funktion ist. Die Mathematik, die zum lösen notwendig ist, werden wir in der Vorlesung nicht behandeln, sondern direkt die Lösung angeben. Die Lösung ist

$$f(x) = k \cdot \cosh(\frac{x}{k}) + c,$$

mit $k, c \in \mathbb{R}$, die von der Länge L und dem Wert $f(\frac{a}{2})$ abhängen. Da das c nur eine vertikale Verschiebung beschreibt und wir unser Koordinatensystem frei wählen können, setzen wir einfach $c = 0$. Unsere physikalisch begründete Modellfunktion ist somit

$$f : [-\frac{a}{2}, \frac{a}{2}] \longrightarrow \mathbb{R} \quad ; \quad x \mapsto k \cdot \cosh(\frac{x}{k}),$$

wobei wir das k noch bestimmen müssen. Um uns daran zu gewöhnen, entdimensionalisieren wir das Problem wieder. Dazu listen wir erst alle Parameter und ihre Dimensionen auf, was hier sehr einfach ist:

$$[k] = [x] = [a] = [d] = [f(x)] = L.$$

Jetzt führen wir entdimensionalisierte Variablen und Funktionen ein. Dazu setzen wir $X = \frac{x}{\ell}$ und

$$F(X) = \frac{f(x)}{\bar{\ell}} = \frac{f(X \cdot \ell)}{\ell} = \frac{k}{\ell} \cosh(\frac{X \cdot \bar{\ell}}{k}),$$

wobei $\bar{\ell}$ irgendein Wert der Dimension L ist. Das Problem würde am einfachsten, wenn wir $\bar{\ell} = k$ setzen würden, allerdings wissen wir noch nicht

welchen Wert k besitzt. Daher wählen wir wie in den anderen Modellen $\bar{\ell} = a$. Dann ist

$$F(X) = \frac{k}{a}\cosh(\frac{a}{k} \cdot X), \quad \text{mit } X \in [-\frac{1}{2}, \frac{1}{2}],$$

die entdimensionalisierte Modellfunktion. Die entdimensionalisierte Länge des Kabels ist damit

$$\begin{aligned}\lambda_{d/a} &= \int_{-1/2}^{1/2} \sqrt{1 + F'(X)^2}\mathrm{d}X = \int_{-1/2}^{1/2} \sqrt{1 + \sinh(\frac{a}{k} \cdot X)^2}\mathrm{d}X \\ &\overset{(3.3)}{=} \int_{-1/2}^{1/2} \cosh(\frac{a}{k} \cdot X)\mathrm{d}X = [\frac{k}{a}\sinh(\frac{a}{k} \cdot X)]_{-1/2}^{1/2} \\ &= 2 \cdot \frac{k}{a}\sinh(\frac{a}{2k}).\end{aligned}$$

Jetzt sollten wir uns endlich um das k kümmern.
Da wir den Durchhang kennen, wissen wir, dass $f(\frac{a}{2}) - f(0) = d$ gelten muss. Ausgedrückt mit unseren entdimensionalisierten Werten gilt also

$$\frac{k}{a} \cdot \cosh(\frac{a}{2k}) = \frac{k}{a} + \frac{d}{a}. \tag{3.5}$$

Setzen wir $y = \frac{a}{2k}$ wird dies zu

$$\cosh(y) = 1 + 2\frac{d}{a} \cdot y.$$

Das lässt sich algebraisch leider nicht lösen. Geometrisch betrachtet benötigen wir die positive Schnittstelle der Funktionen $\cosh(y)$ und $1 + 2\frac{d}{a} \cdot y$. Die kann man beliebig gut approximieren – sie liegt bei $y = 0{,}7056\ldots$. Damit erhalten wir sofort

$$k = 97{,}49\ldots m \quad \text{und} \quad \lambda_{d/a} = 1{,}085\ldots m.$$

Die Länge des Kabels in diesem Modell ist somit $\lambda_{d/a} \cdot a = 149{,}31\ldots m$ und nur leicht größer als im Modell des parabelförmigen Verlaufes.

3.3 Die logistische DGL entdimensionalisiert

Sei $(a_n)_{n\in\mathbb{N}}$ eine reelle Folge. Falls die Folge konvergiert, so ist der Grenzwert a dieser Folge dadurch charakterisiert, dass für alle $\epsilon > 0$ ein $N \in \mathbb{N}$ existiert mit $|a_n - a| < \epsilon$ für alle $n \geq N$. Es muss also möglich sein die Differenz von Folgengliedern und dem Grenzwert zu bestimmen. Das funktioniert nur, wenn $[a_n] = [a]$ für alle $n \in \mathbb{N}$ gilt.

Lemma 3.3.1. *Sei $x(t)$ eine reellwertige differenzierbare Funktion auf einem Intervall I. Dann ist $[x'(t)] = \frac{[x(t)]}{[t]}$.*

BEWEIS. Es ist für jedes $t_0 \in I$

$$x'(t)_0 = \lim_{t \longrightarrow t_0} \frac{x(t_0) - x(t)}{t_0 - t}.$$

Es ist $[\frac{x(t_0)-x(t)}{t_0-t}] = \frac{[x(t)]}{[t]}$ für alle $t \in I$. Mit unserer Vorüberlegung muss auch der Grenzwert diese Dimension besitzen. □

Wir betrachten noch einmal die logistische DGL in dem üblichen Kontext, dass $x(t)$ die Größ einer Population zum Zeitpunkt t beschreibt. Die Größe kann in verschiedenen Dimensionen gemessen werden wie z.B. L^3 oder M. Wir werden hier die generische Anzahl A benutzen. Da wir nur eine Spezies betrachten, können wir hier auf den Index verzichten.

Bemerkung 3.3.2. Wir werden tatsächlich keinen kontinuierlichen Verlauf erkennen, da das Sterben eines Individuums die Population sprungartig (und somit nicht stetig) verringert. Wenn wir von einer Population ausgehen, die sehr groß ist, sind diese Sprünge aber zu vernachlässigen. Auch, dass wir nicht nur ganze Zahlen als Funktionswerte herausbekommen stört uns nicht besonders. Die Funktion gibt lediglich die ungefähre Entwicklung an. Außerdem haben wir auch gesehen, dass diskrete Modelle ebenfalls zu nicht ganzzahligen Lösungen führen können (vgl. Zinsrechnung).

Wir starten nun mit der logistischen DGL $x'(t) = \lambda \cdot x(t) \cdot (K - x(t))$. Wir führen die folgenden Schritte aus:

1. Parameter und Dimensionen auflisten: Es ist $[t] = T$ und (wie oben beschrieben) $[x(t)] = A$. Die Kapazitätsgrenze K beschreibt ebenfalls eine Populationsgröße, daher ist auch $[K] = A$. Alternativ kann man hier über die Rechenregeln argumentieren, denn der Faktor $K - x(t)$ setzt bereits voraus, dass $[K] = [x(t)]$ ist. Für das λ stellen wir fest

   $$\frac{A}{T} = \frac{[x(t)]}{[T]} = [x'(t)] = [\lambda] \cdot [x(t)] \cdot [K - x(t)] = [\lambda] \cdot A^2.$$

   Damit erhalten wir $[\lambda] = \frac{1}{AT}$. Das deckt sich auch genau mit der Herleitung dieses Terms, da es sich um die Reproduktion pro Individuum, pro Zeiteinheit handelt.

2. dimensionslose Variablen einführen. Seien $\bar{t}$, mit $[\bar{t}] = T$, und $\bar{a}$, mit $[\bar{a}] = A$, beliebig. Wir setzen $\tau = \frac{t}{\bar{t}}$ und erhalten damit

$$\frac{x(t)}{\bar{a}} = \frac{x(\tau \cdot \bar{t})}{\bar{a}} = X(\tau).$$

3. DGL in den neuen Variablen schreiben. Es ist

$$\begin{aligned} X'(\tau) &= \left(\frac{x(\tau \cdot \bar{t})}{\bar{a}}\right)' = \frac{\bar{t}}{\bar{a}} \cdot x'(\tau \cdot \bar{t}) \\ &= \frac{\bar{t}}{\bar{a}} \cdot \lambda \cdot x(\tau \cdot \bar{t}) \cdot (K - x(\tau \cdot \bar{t})) \\ &= \frac{\bar{t}}{\bar{a}} \cdot \lambda \cdot \bar{a} \cdot X(\tau) \cdot (K - \bar{a} \cdot X(\tau)) \\ &= \bar{t} \cdot \bar{a} \cdot \lambda \cdot X(\tau) \cdot (\frac{K}{\bar{a}} - X(\tau). \end{aligned}$$

4. Einheiten wählen: Damit die entdimensionalisierte DGL besonders einfach wird, setzen wir $\bar{a} = K$ und $\bar{t} = \frac{1}{K \cdot \lambda}$. Beachten Sie, dass tatsächlich $[\frac{1}{K\lambda}] = [K]^{-1} \cdot [\lambda]^{-1} = A^{-1}AT = T$ gilt. Mit diesen Einheiten wird die DGL zu

$$X'(\tau) = X(\tau) \cdot (1 - X(\tau)). \tag{3.6}$$

Wir haben es also geschafft, nur durch die Wahl von anderen Einheiten, beide Koeffizienten λ und K aus der DGL zu entfernen. Um die logistische DGL zu lösen, brauchen wir also nur die Lösungen der DGL (3.6), die ohne Koeffizienten auskommt. Wie in Proposition 2.8.3 (nur deutlich einfacher) erhalten wir, dann die Lösungen von (3.6) gegeben sind durch die konstanten Lösungen $X(\tau) = 0$ und $X(\tau) = 1$, sowie durch

$$X(\tau) = \frac{1}{e^{-\tau} \cdot C + 1}, \qquad \text{mit } C \in \mathbb{R} \setminus \{0\}.$$

Jetzt fügen wir die Einheiten wieder hinzu und erhalten die bekannte Lösung

$$x(t) = \bar{a} \cdot X(\tau) = \frac{K}{e^{-\tau} \cdot C + 1} = \frac{K}{e^{-\frac{t}{\bar{t}}} \cdot C + 1} = \frac{K}{e^{-t\lambda K} \cdot C + 1},$$

für $C \in \mathbb{R} \setminus \{0\}$. Es genügte also *eine einzige* DGL zu lösen, um die Lösungen *jeder* logistischen DGL zu erhalten.
Die hier benutzten Schritte lassen sich auf jede dimensionsbehaftete DGL anwenden.

Kapitel 4

Populationsmodelle mit Interaktion

Bisher haben wir nur das exponentielle und das logistische Wachstum einer Population kennengelernt. In beiden Fällen konnte sich die Population ohne Interaktion mit anderen Spezies (inkl. des Menschen) entwickeln. In diesem Kapitel betrachten wir Populationsentwicklungen mit Interaktion.

4.1 Fischfang

In diesem Abschnitt betrachten wir eine Fischpopulation in einem großen See oder in einem Teilbereich eines Meeres. Wir beschäftigen uns mit der Frage:

Wie sieht eine optimale Strategie für den Fischfang aus?

Diese Frage ist sehr grob formuliert und kann sicher keine eindeutige Antwort liefern. Aus Sicht der Fische ist die optimale Fangstrategie sicher keine Fische zu fangen. Wir werden also eher verschiedene Strategen vergleichen, als eine Antwort auf die gestellte Frage zu geben.

Annahme: Ohne Fischfang würde die Fischpopulation einem logistischen Wachstum folgen. Beschreibt also $x(t)$ die Grösse der Fischpopulation zum Zeitpunkt t, dann gilt

$$x'(t) = \lambda x(t)(K - x(t)), \text{ mit reellen Zahlen } \lambda, K > 0.$$

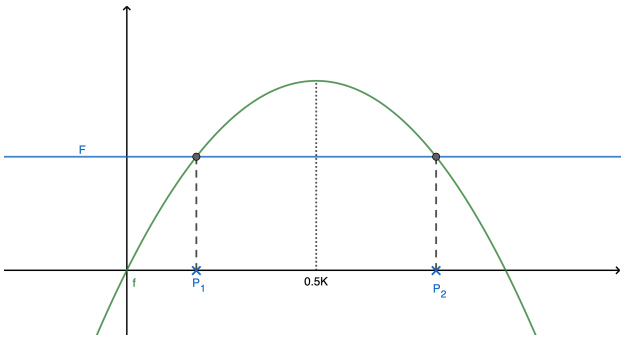
Eine *Fangstrategie* wird beschrieben durch eine Funktion $u : [0, \infty) \longrightarrow \mathbb{R}$, die die Anzahl von gefangenen Fischen pro Zeiteinheit misst. Die Anzahl gefangener Fische im Zeitintervall $[t_0, t_1]$ ist somit $\int_{t_0}^{t_1} u(t)\mathrm{d}t$.
Für die Fischpopulation gilt, bei gegebener Fangstrategie $u(t)$, die DGL

$$x'(t) = \lambda x(t)(K - x(t)) - u(t). \tag{4.1}$$

Strategie 1: (konstante Fangrate) Bei den meisten internationalen Fischereivereinbarungen wird eine Menge von Fischen vorgegeben, die gefangen werden darf. Für den Kabeljau in der Nordsee z.B. liegt diese Menge bei 5060 Tonnen pro Jahr[1]. Das entspricht einer konstanten Fangrate $u(t) = F$ für alle $t \in \mathbb{R}$. Die DGL (4.1), die die Größe der Fischpopulation $x(t)$ zum Zeitpunkt t misst, wird dann zu

$$x'(t) = \lambda x(t)(K - x(t)) - F. \tag{4.2}$$

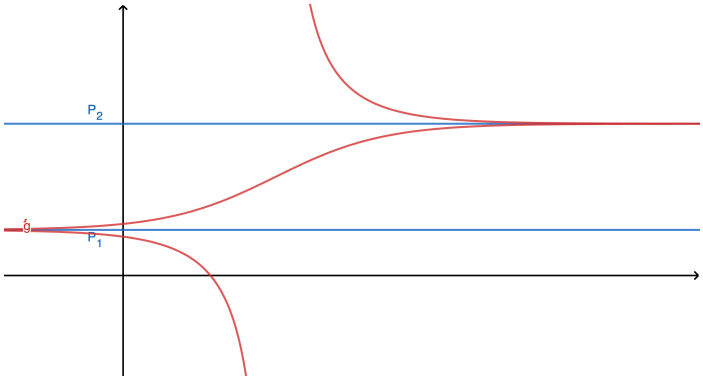
Wir untersuchen das Langzeitverhalten dieser DGL. Dazu setzen wir $f(x) = \lambda x(K - x)$, was der rechten Seite der DGL entspricht, die die Fischpopulation ohne Fischerei modelliert. Die stationären Punkte von (4.2) sind dann die Schnittpunkte von $f(x)$ und F. Für eine bessere Vergleichbarkeit mit anderen Fangstrategien, werden wir die Funktion $f(x)$ stets einzeln betrachten (und zeichnen).



Falls $F > \frac{\lambda K^2}{4}$ ist, so ist die rechte Seite von (4.1) durchgehend negativ und es folgt $\lim_{t\to\infty} x(t) = 0$. Die Population stirbt also aus, was schlecht für die Fische und schlecht für die Fischerei ist. Wir nehmen also im folgenden $\boxed{F \in [0, \frac{\lambda K^2}{4}]}$ an. In diesem Fall gibt es zwei stationäre Punkte

[1] Stand 2022: siehe https://www.bmel.de/SharedDocs/Pressemitteilungen/DE/2021/181-fangquoten.html

$P_1 < P_2$. Betrachtungen des Vorzeichens von $f(x) - F$ ergeben die quantitativen Lösungen



Solange die Populationsgröße zu Beginn des Fischfangs $x(t) > P_1$ erfüllt hat, wird die Populationsgröße stets gegen P_2 streben und die Population überlebt.

Bemerkung 4.1.1. Man könnte annehmen, dass eine Fangrate von $F = \frac{\lambda K^2}{4}$ optimal wäre, da dann der Ertrag maximal ist. In diesem Fall ist der Ertrag pro Zeiteinheit $\int_{t_0}^{t_0+1} u(t)\mathrm{d}t = F = \frac{\lambda K^2}{4}$. Da aber die Werte von λ und K nie exakt bekannt sind und eine Fangrate über $\frac{\lambda K^2}{4}$ zum Aussterben führt, ist es nicht realistisch diese Fangrate genau zu erreichen. Man könnte nur versuchen diese Fangrate anzunähern.

Liegt der Wert von F knapp unter $\frac{\lambda K^2}{4}$, so liegen P_1 und P_2 sehr nah beieinander. Wie eben eingesehen gilt:

- die Größe jeden Fischbestands, mit $x(t) > P_2$, strebt auf lange Sicht gegen P_2.
- jeder Fischbestand, mit $x(t) < P_1$, stirbt aus.

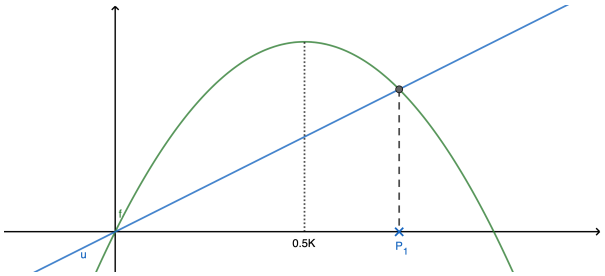
Wenn nun P_1 und P_2 nah beieinander liegen (was die Fangrate vergrössert), führt das auf eine sehr riskante Strategie. Denn kleine Störungen, wie eine Krankheit bei den Fischen, eine Zunahme der Fressfeinde oder eine leichte Überschreitung der Fangrate, führt bei weiterer Befischung zum Aussterben der Population.

Strategie 2: (konstanter Aufwand) Wir nehmen nun an, dass immer mit dem gleichen Aufwand gefischt wird. D.h. Es werden immer die gleichen

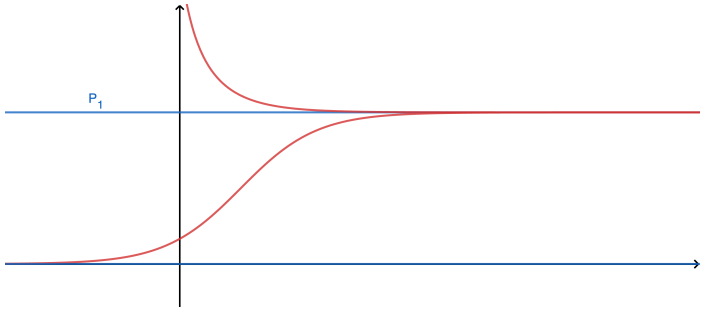
Boote und Netze benutzt, die immer gleich lange in Betrieb sind. Dann sollte die Fangrate proportional zur Populationsgröße $x(t)$ sein. Die Fangstrategie ist in diesem Modell also $u(t) = c \cdot x(t)$ für eine positive konstante $c \in \mathbb{R}$. Dann ist die DGL, die den Fischbestand modelliert gegeben durch

$$x'(t) = \lambda x(t)(K - x(t)) - c \cdot x(t). \tag{4.3}$$

Wieder setzen wir $f(x) = \lambda x(K-x)$. Ist $c \geq f'(0) = \lambda K$, so ist $f(x)-cx < 0$ für alle $x > 0$. Damit folgt, dass in diesem Fall die Fischpopulation ausstirbt. Das soll verhindert werden, und somit muss ein $c \in [0, \lambda K)$ gewählt werden. Wir führen die gleichen Ideen wie eben aus und erhalten das entsprechende Bild



Es gibt nur die stationären Punkte $x = 0$ und $P_1 = K - \frac{c}{\lambda}$. Für alle $x \in (0, P_1)$ ist $f(x) - cx > 0$. Damit konvergiert jede Lösung von (4.3), mit $x(0) > 0$, gegen den stationären Punkt P_1.



Dieser Strategie, solange $c \in [0, \lambda K)$ gilt, führt also nicht zum Aussterben der Population. Auf lange Sicht ist der Ertrag pro Zeiteinheit gegeben durch

$$\int_{t_0}^{t_0+1} u(t)\mathrm{d}t = \int_{t_0}^{t_0+1} cx(t)\mathrm{d}t = \int_{t_0}^{t_0+1} cP_1\mathrm{d}t = c(K - \frac{c}{\lambda}).$$

Wenn wir das als Funktion in der Variablen c betrachten, erhalten wir eine nach unten geöffnete Parabel mit den Nullstellen $c = 0$ und $c = \lambda K$. Das

Maximum wird also angenommen für $c = \frac{\lambda K}{2}$ und beträgt $\frac{K}{4}$. Das ist genau der gleiche maximale Ertrag, wie in der riskanten Strategie der konstanten Fangrate! Grafisch findet man die optimale Fangrate, wenn man die Gerade cx so einzeichnet, dass sie die Parabel $f(x)$ im Maximum schneidet.

Bemerkung 4.1.2. Wie findet man nun die richtige Fangrate? Es ist sicher sehr aufwändig (bis unmöglich) die Parameter λ und K zu berechnen. Stattdessen kann man versuchen, den Aufwand nach und nach leicht zu erhöhen – etwa in dem ein Boot mehr pro Monat ausfährt. Solange wie der Ertrag dadurch steigt, macht man weiter damit den Aufwand zu erhöhen. Sobald sich der Ertrag aber verringert, verringert man auch wieder den Aufwand und lässt z.B. ein Boot im Hafen. Auf diese Weise kann man versuchen die optimale Fangrate zu treffen.

Als Fazit halten wir fest, dass das Modell des konstanten Aufwandes den gleichen optimalen Ertrag liefern kann, wie das Modell der konstanten Fangrate, ohne den Nachteil von irreparablen Schäden an der Population durch leichte Überfischung.

Strategie 3: (konstanter Aufwand - Version 2) Wir möchten das Modell aus Strategie 2 noch ein bisschen verfeinern. Wir hatten in Bemerkung 4.1.2 beschrieben, dass die optimale Fangrate angenähert werden kann, in dem nach und nach den Aufwand erhöht, bzw. verringert. Als Beispiel wurde genannt, dass ein Boot mehr, bzw. weniger, zum Fischen ausfährt. Wir sollten also die Fangrate eines einzigen Bootes betrachten. Diese Fangrate nennen wir $g(t)$.

1. Ist nur ein einziger Fisch im Meer, so wird man diesen niemals fangen können. Wir wollen ab jetzt annehmen, dass es eine Populationsgröße V gibt, die sich immer komplett verstecken kann. Ist also $x(t) \leq V$, so kann das Boot keine Fische fangen und es gilt $g(t) = 0$. Das nennen wir *Schwelleneffekt*.

2. Ab einer gewissen Populationsgröße A, wird das Boot immer voll ausgelastet zurück in den Hafen kommen. Auch wenn sich die Population dann noch vergrößert wird das Schiff nicht noch mehr Fische fangen. Es gilt also $g(t) = S$ für alle t, mit $x(t) \geq A$. Das nennen wir den *Sättigungseffekt*.

3. Liegt die Populationsgröße $x(t)$ zwischen V und A, soll die Fangrate proportional zur „fangbaren Populationsgröße“ $x(t) - V$ sein.

Mit diesen Annahmen erhalten wir die Fangstrategie für ein einzelnes Boot

$$g(t) = \begin{cases} 0 & \text{für } x(t) < V \\ p \cdot (x(t) - V) & \text{für } V \leq x(t) < A \\ S & \text{für } A \leq x(t) \end{cases}.$$

Da diese Funktion natürlich stetig sein soll, muss gelten $p \cdot (A - V) = S$, also $p = \frac{S}{A-V}$. Fahren nun n Boote aus um Fische zu Fangen, so wird die Fangstrategie modelliert durch

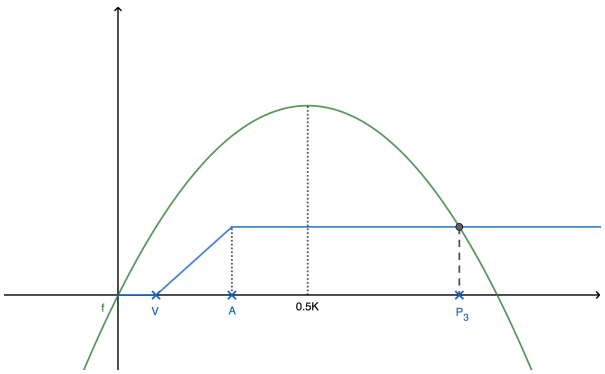
$$u(t) = n \cdot g(t) = \begin{cases} 0 & \text{für } x(t) < V \\ n \cdot \frac{S}{A-V} \cdot (x(t) - V) & \text{für } V \leq x(t) < A \\ n \cdot S & \text{für } A \leq x(t) \end{cases}. \tag{4.4}$$

Ab jetzt ist mit $u(t)$ stets diese Funktion gemeint. Verlassen also immer n Boote den Hafen zum Fischen, dann erhalten wir für die Größe $x(t)$ der Fischpopulation zum Zeitpunkt t das Modell

$$x'(t) = \underbrace{\lambda \cdot x(t) \cdot (K - x(t))}_{=f(x(t))} - u(t). \tag{4.5}$$

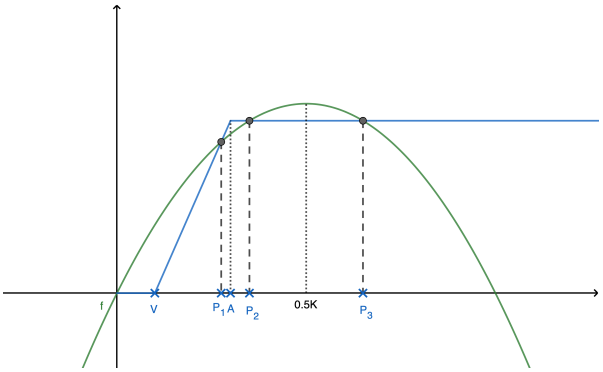
Annahme: Es gilt $A < \frac{K}{2}$. D.h. die Boote kommen immer voll ausgelastet zurück, ab einem Fischbestand, der kleiner ist als die halbe Kapazitätsgrenze. Wir gehen also von kleineren Booten aus.

Der Fall von großen Booten wird in den Übungen behandelt.
Es bleibt noch eine Fallunterscheidung über den Wert von n. Für **kleines** n erhalten wir



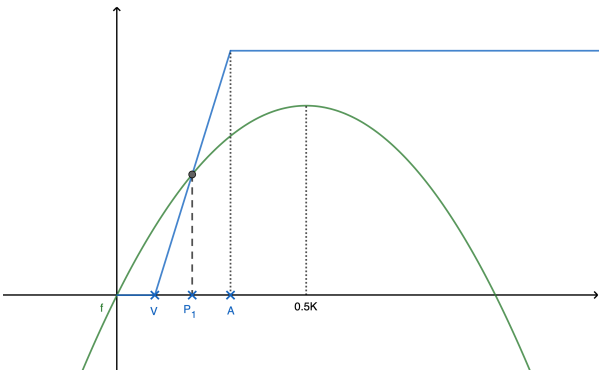
Es gibt nur einen positiven stationären Punkt P_3. Alle Lösungen von (4.5) mit positivem Anfangswert konvergieren somit gegen P_3. Das ist daher die Größe der Fischpopulation, die sich auf lange Sicht einstellen wird.

Für **mittleres** n erhalten wir



Hier gibt es drei positive stationäre Punkte $P_1 < P_2 < P_3$. Da die Differenz $f(x) - u(x)$ an jedem Schnittpunkt das Vorzeichen wechselt, findet man schnell heraus, dass jede Lösung mit einem Anfangswert $x(0) > P_2$ gegen P_3 konvergiert, jede Lösung mit einem Anfangswert $0 < x(0) < P_2$ gegen P_1 konvergiert.

Für **großes** n erhalten wir

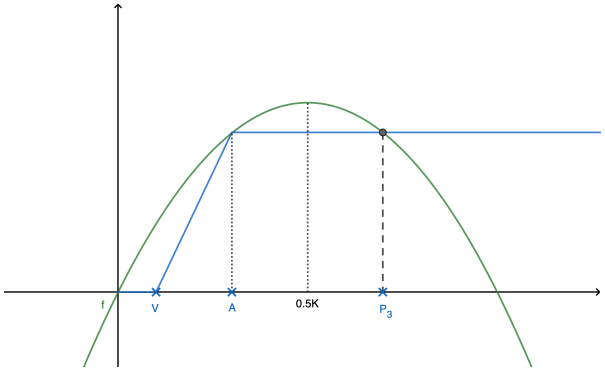


Es gibt wieder nur einen positiven stationären Punkt P_1. Alle Lösungen von (4.5) mit positivem Anfangswert konvergieren gegen P_1. Das ist daher die Größe der Fischpopulation, die sich auf lange Sicht einstellen wird.

Bemerkung 4.1.3. In allen Fällen sind die auf lange Sicht möglichen Erträge pro Zeiteinheit abzulesen an den y-Koordinaten der Schnittpunkte zwischen f und u. Insbesondere sehen wir, dass der Ertrag in der Skizze für

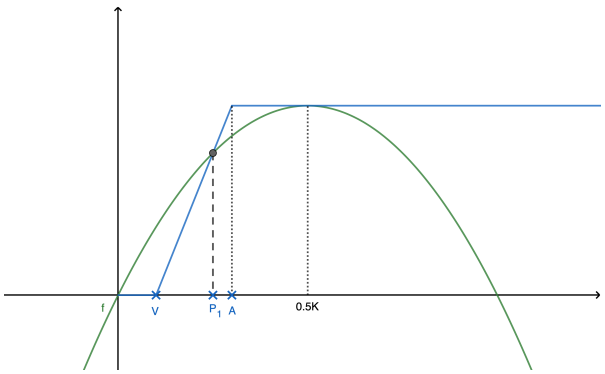
großes n nur geringfügig über dem Ertrag in der Skizze für kleines n liegt. Für die Fischpopulation (und die Wirtschaftlichkeit der Fischerei), ist ein kleines n natürlich von Vorteil.
Man hat also einen besonders großen möglichen Ertrag, wenn wir in dem Fall des mittleren n sind, und P_2 und P_3 sehr nah beieinander liegen. Um diese Situation genauer zu studieren betrachten wir noch zwei Sonderfälle, die (für möglicherweise nicht ganzzahliges n) eintreten können.

Es existiert genau ein $\underline{n} \in \mathbb{R}$ für das A ein stationärer Punkt ist. In diesem Fall gibt es noch einen weiteren positiven stationären Punkt $P_3 > A$. Die Skizze dazu sieht so aus



Hier wechselt das Vorzeichen von $f(x) - u(x)$ nicht am stationären Punkt A. Damit konvergiert jede Lösung von (4.5) gegen A, wenn $x(0) < A$ gilt, und gegen P_3, wenn $x(0) > A$.

Weiter existiert ein kritischer Wert $\bar{n} \in \mathbb{R}$, für den $\frac{K}{2}$ ein stationärer Punkt ist. Dann gibt es genau einen weitern positiven stationären Punkt $P_1 < \frac{K}{2}$. Hier sehen Sie die zugehörige Skizze

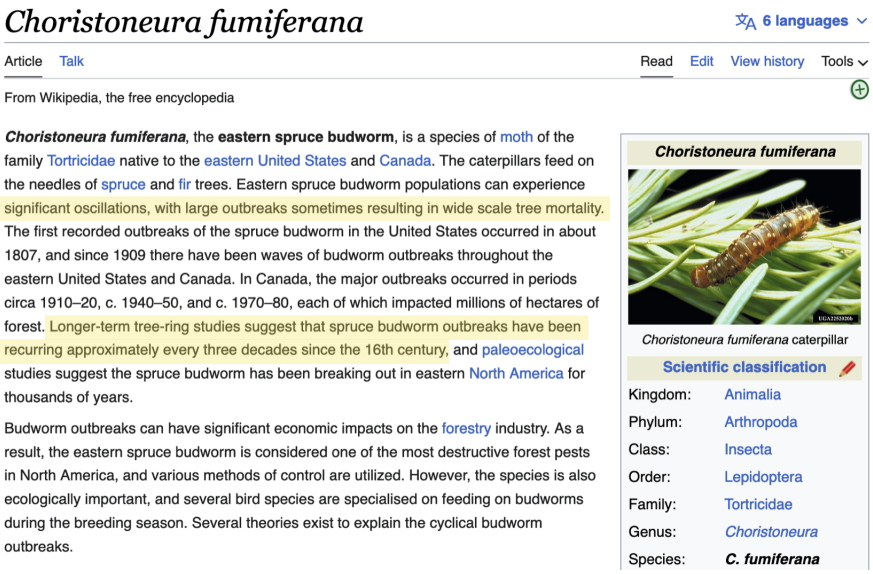


In diesem Fall konvergiert jede Lösung von (4.5) gegen $\frac{K}{2}$, wenn $x(0) > \frac{K}{2}$ gilt. Wenn jedoch $x(0) < \frac{K}{2}$ konvergiert die Lösung gegen den kleineren Wert P_1. Wie in Strategie 1 (konstante Fangrate), ist dieser Zustand auf gewisse Weise ideal. Den der Ertrag wird auf lange Sicht bei $f(\frac{K}{2}) = \bar{n} \cdot S = \frac{\lambda K^2}{4}$ lpro Zeiteinheit liegen. Es gibt allerdings auch das gleiche Risiko. Wird der Wert $\bar{n}$ nur leicht überschritten, bricht die Populationsgröße auf P_1 zusammen. Eine leichte Reduzierung von n genügt dann nicht aus, damit sich die Population wieder erholt. Erst wenn die Anzahl an Booten wieder unter $\underline{n}$ sinkt, kann die Population wieder wachsen.

Bemerkung 4.1.4. Wir stellen fest, dass Strategie 3, mit Schwellen- und Sättigungseffekt und einem Wert $A < \frac{K}{2}$ (also der Annahme, dass die Boote *klein* sind), die gleichen Risiken besitzt, wie die Strategie 1 der konstanten Fangrate: Wenn der maximal mögliche langfristige Ertrag angepeilt wird, können kleine Schwankungen in der Populationsgröße zum Zusammenbruch der Population führen. Den Fall von großen Booten, betrachten Sie in den Übungen.

4.2 Der Tannenwickler

Der Tannenwickler ist eine Schmetterlingsart. Wenn wir nichts über etwas wissen, dann schauen wir bei Wikipedia und finden eine kurze Beschreibung.



Choristoneura fumiferana

6 languages

Article Talk Read Edit View history Tools

From Wikipedia, the free encyclopedia

Choristoneura fumiferana, the **eastern spruce budworm**, is a species of moth of the family Tortricidae native to the eastern United States and Canada. The caterpillars feed on the needles of spruce and fir trees. Eastern spruce budworm populations can experience significant oscillations, with large outbreaks sometimes resulting in wide scale tree mortality. The first recorded outbreaks of the spruce budworm in the United States occurred in about 1807, and since 1909 there have been waves of budworm outbreaks throughout the eastern United States and Canada. In Canada, the major outbreaks occurred in periods circa 1910–20, c. 1940–50, and c. 1970–80, each of which impacted millions of hectares of forest. Longer-term tree-ring studies suggest that spruce budworm outbreaks have been recurring approximately every three decades since the 16th century, and paleoecological studies suggest the spruce budworm has been breaking out in eastern North America for thousands of years.

Budworm outbreaks can have significant economic impacts on the forestry industry. As a result, the eastern spruce budworm is considered one of the most destructive forest pests in North America, and various methods of control are utilized. However, the species is also ecologically important, and several bird species are specialised on feeding on budworms during the breeding season. Several theories exist to explain the cyclical budworm outbreaks.

Choristoneura fumiferana

Choristoneura fumiferana caterpillar

Scientific classification

Kingdom:	Animalia
Phylum:	Arthropoda
Class:	Insecta
Order:	Lepidoptera
Family:	Tortricidae
Genus:	*Choristoneura*
Species:	***C. fumiferana***

Der Tannenwickler zeigt also folgendes Verhalten:

- Für einen langen Zeitraum gibt es nur eine geringe Populationsgröße.
- Ungefähr alle 30 Jahre gibt es einen rasanten Anstieg der Populationsgröße mit beträchtlichem Schaden am Lebensraum.
- Danach bricht die Population wieder zusammen.

Dieses Phänomen möchten wir anhand eines einfachen mathematischen Modells erklären / begründen. Dazu nehmen wir folgendes an:

- Die Hauptfressfeinde der Tannenwickler sind Reviervögel, deren Anzahl hauptsächlich durch die Reviergröße bestimmt ist. Ihr Überleben ist nicht von den Tannenwicklern abhängig.
- Es gibt einen Schwelleneffekt, denn sind zu wenig Tannenwickler vorhanden, dann sind sie als Nahrungsquelle zu unattraktiv und die einzelnen Exemplare können sich gut verstecken.
- Es gibt auch einen Sättigungseffekt, da jeder Vogel nur eine berenzte Anzahl an Tannenwicklern pro Tag fressen kann.

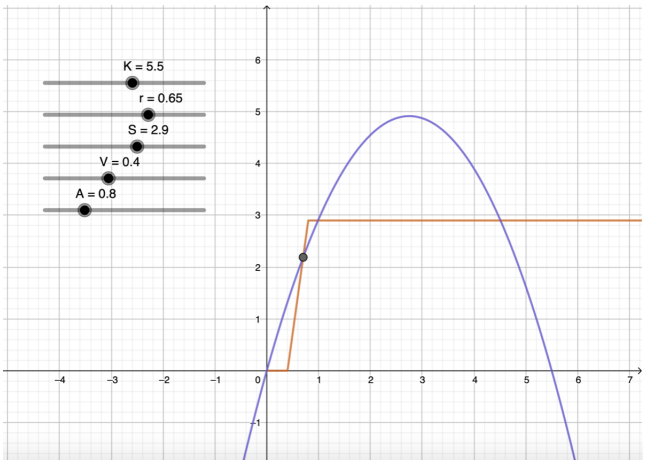
Wir erhalten daher eine Modellfunktion wie in der Strategie 3 aus dem letzten Abschnitt. Das Wachstum der Tannenwicklerpopulation $x(t)$ wird also negativ beeinflusst durch eine Funktion

$$g(t) = \begin{cases} 0 & \text{für } x(t) < V \\ \frac{S}{A-V} \cdot (x(t) - V) & \text{für } V \leq x(t) < A \\ S & \text{für } A \leq x(t) \end{cases} . \tag{4.6}$$

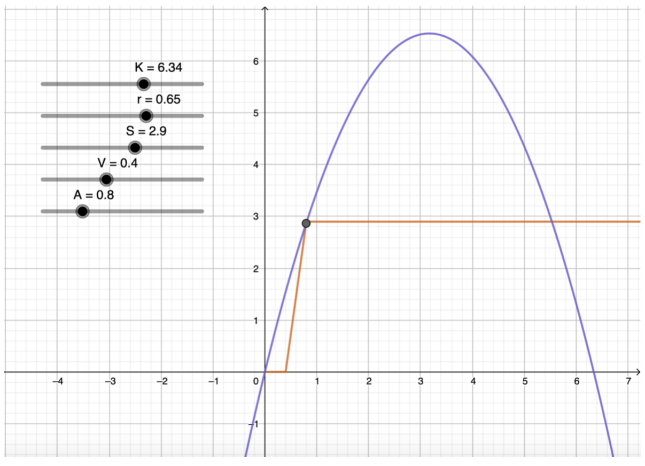
Wie immer nehmen wir auch an, dass sich die Tannenwickler ohne Fressfeinde einem logistischen Wachstum folgen würden. Damit gilt

$$x'(t) = \lambda x(t)(K - x(t)) - g(x(t)),$$

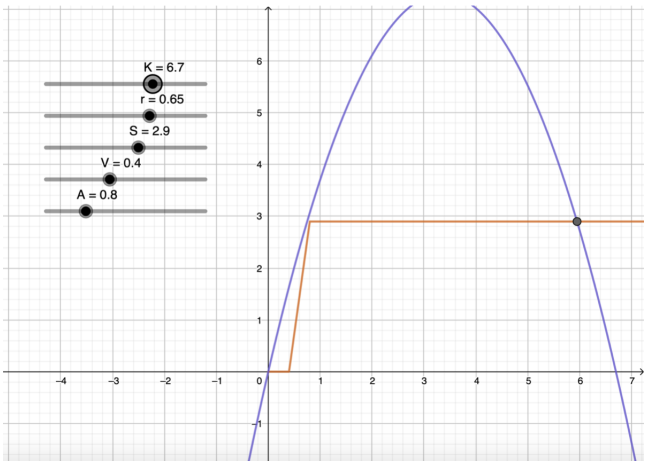
wobei $\lambda > 0$ die Reproduktionsrate der Tannenwickler ist und K die Kapazitätsgrenze. Stellen wir das logistische Wachstum blau dar und die Funktion g grün, so erhalten wir



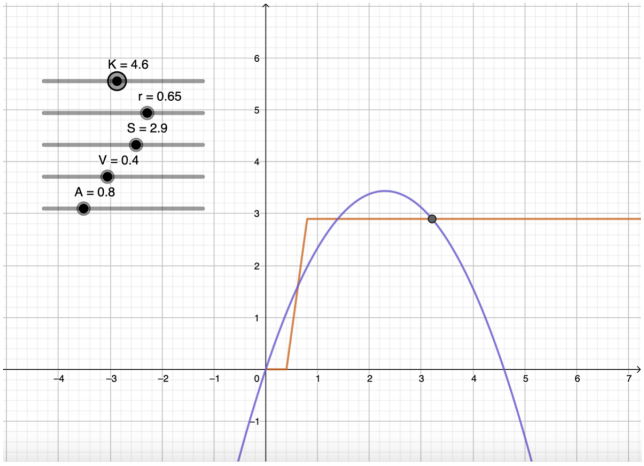
Wenn es nun wenig Exemplare gibt, so strebt die Populationsgröße immer gegen den ersten positiven Schnittpunkt. Die Population bleibt also klein. Das deckt sich nicht mit dem am Anfang des Abschnittes beobachteten Verhalten. Wir müssen also noch etwas anderes beachten.
Die Koeffizienten λ, V, A, S nehmen wir als im wesentlichen konstant an. Der Wert K wird durch den potentiellen Lebensraum der Tannenwickler bestimmt. Dieser Lebensraum der Tannenwickler besteht aus Balsamtannen. Wenn es nun nur wenige Tannenwickler gibt, so scheint das gut für die Balsamtannen zu sein. Diese vermehren sich und wachsen, was wiederum mehr Platz für die Tannenwickler schafft und den Wert K vergrößert. Dadurch erhalten wir zunächst den Zustand



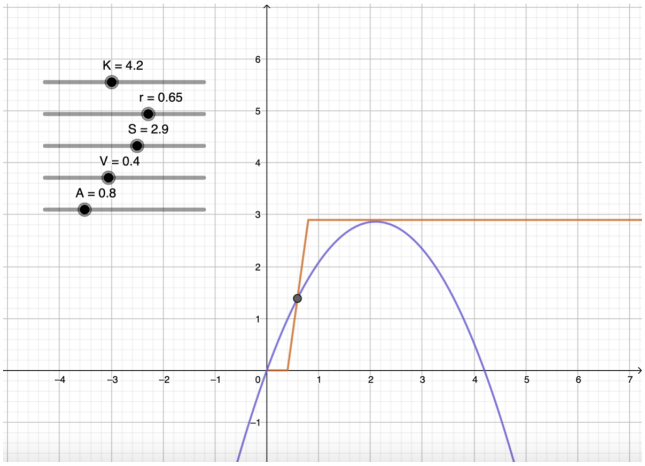
Wenn sich nun der Wert K nur noch leicht vergrößert (es also durch mehr Balsamtannen noch ein bisschen attraktiver für die Tannenwickler wird) erhalten wir



Der stationäre Punkt macht also einen riesigen Sprung, was einen sofortigen rasanten Anstieg von $x(t)$ – der Tannenwicklerpopulation – nach sich zieht. Nun richten die vielen Tannenwickler einen verheerenden Schaden an den Balsamtannen an, was zu einer Reduzierung von K führt. Der stationäre Punkt – und damit die Tannenwicklerpopulation – verringert sich dadurch langsam.



Damit der Tannenwicklerbesand allerdings auf ein geringes Niveau wie zu Beginn fällt, muss K so weit sinken, dass der Graph des logistischen Wachstums vollständig unter der Sättigungsgrenze S liegt.



Erst dann ist der Bestand so gering, dass sich der Balsamtannenbestand wieder erholen kann. Die lange Wuchsdauer von Bäumen, kann die lange Periode der kleinen Tannenwicklerpopulation erklären.

Bemerkung 4.2.1. Wir haben hier anschaulich die Interaktion zweier Spezies betrachtet (Tannenwickler und Balsamtannen). Solche Interaktionen werden wir im folgenden Abschnitt genauer studieren.

Bemerkung 4.2.2. Oft wird anstatt der nicht differenzierbaren Funktion g aus (4.6), die Modellfunktion

$$\tilde{g}(t) = S \cdot \left(1 - \frac{V^2}{V^2 + x(t)^2}\right)$$

benutzt. Diese Funktion hat einen ähnlichen Verlauf und besitzt einen Parameter weniger, wodurch man einfacher mit ihr arbeiten kann. Auch die Tatsache, dass keine Fallunterscheidung mehr nötig ist, macht das Arbeiten mit dieser Funktion recht attraktiv. Da jedoch die Werte von V und A in diesem Modell gekoppelt sind, können sie nicht unabhängig von einander festgelegt werden.
Die zugehörige DGL

$$x'(t) = \lambda x(t)(K - x(t)) - S \cdot \left(1 - \frac{V^2}{V^2 + x(t)^2}\right)$$

lässt sich entdimensionalisieren zu $X'(\tau) = X(\tau)\left(a(1 - bX(\tau)) - \frac{X(\tau)}{1+X(\tau)^2}\right)$, mit den dimensionslosen Parametern $a = \frac{\lambda V K}{S}$ und $b = \frac{V}{K}$. Welche Einheiten auf diese DGL führen, kann als Übung selbst herausgefunden werden[2].

[2]Das bedeutet so viel wie: ich hatte keine Lust die ganze Entdimensionalisierung aufzuschreiben und halte das für einfach genug, dass es alle alleine schaffen.

4.3 Räuber-Beute Beziehungen nach Lotka-Volterra

In diesem Abschnitt beschäftigen wir uns mit der Interaktion zweier Spezies, die in einem Räuber-Beute-Verhältnis stehen. Etwa Füchse und Kaninchen. Wieder modellieren wir die Populationsgrößen kontinuierlich. Ab jetzt beschreibt

$x_B(t)$ die Größe der Beute-Population zum Zeitpunkt t

$x_R(t)$ die Größe der Räuber-Population zum Zeitpunkt t.

Worauf man sich sicher einigen kann ist folgendes: Treffen sich ein Räuber- und ein Beutetier, hat das potentiell schlechte Auswirkungen auf die Beutepopulation und gute Auswirkungen auf die Räuberpopulation. Die Anzahl von Treffen zwischen einem Beute-Tier und einem Räuber pro Zeiteinheit ist in etwa $k \cdot x(t) \cdot y(t)$, für ein $k \in \mathbb{R}$. Hier nehmen wir an, dass die Anzahl von Treffen proportional zur Größe der Beute-Population und zur Größe der Räuber-Population ist. Diese Annahme kennen Sie bereits von der dritten Herleitung des logistischen Wachstums!
Das Wachstum der Beute-Population wird also negativ durch einen Wert $px(t)y(t)$ beeinflusst, wobei das Wachstum der Räuber-Population positiv durch einen Term $qx(t)y(t)$ beeinflusst wird. Wir machen noch zwei weitere Annahmen:

- Die Existenz von Räubern ist die wesentliche Beschränkung des Wachstums der Beute-Population. D.h.: Wenn es keine Räuber mehr gibt, wächst die Beute-Population ungehemmt.

- Die Beutetiere sind essentiell für das Überleben der Räuber. D.h.: gibt es keine Beutetiere mehr, dann stirbt auch die Räuber-Population aus.

Die Beute-Population *ohne Räuber* folgt also einem exponentiellen Wachstum $x'_B(t) = ax_B(t)$, mit $a > 0$. Die Räuber-Population *ohne Beute* folgt einem negativen exponentiellen Wachstum $x'_R(t) = -bx_R(t)$. Diese Annnahmen zusammen ergeben das folgende berühmte Modell.

Definition 4.3.1. Wir nutzen die Bezeichnungen von oben. Das *Lotka-Volterra-Modell* zur Beschreibung eines Räuber-Beute-Verhältnisses ist ge-

geben durch

$$\begin{aligned} x'_B(t) &= sx_B(t) - ax_B(t)x_R(t) \\ x'_R(t) &= -rx_R(t) + bx_B(t)x_R(t), \end{aligned} \tag{LV}$$

mit positiven reellen Zahlen a, b, r und s.

Bemerkung 4.3.2. Mit was für einer Art von mathematischem Objekt haben wir es hier zu tun? Dazu gibt es zwei Sichtweisen. In der ersten Sichtweise, stehen in (LV) zwei DGL, die simultan gelöst werden müssen. Wenn wir eine der beiden Funktionen $x_B : \mathbb{R} \longrightarrow \mathbb{R}$ oder $x_R : \mathbb{R} \longrightarrow \mathbb{R}$ kennen, dann können wir auch die andere berechnen. Leider kennen wir aber noch keine der beiden Funktionen.
In der zweiten Sichtweise, steht in (LV) nur eine DGL in zwei Dimensionen. Gesucht ist also eine Funktion

$$x : \mathbb{R} \longrightarrow \mathbb{R}^2 \quad ; \quad t \mapsto \begin{pmatrix} x_B(t) \\ x_R(t) \end{pmatrix},$$

deren Ableitung $\nabla x(t) = (x'_B(t)\ ,\ x'_R(t))$ eine gewisse Bedingung $\nabla x(t) = f(x(t), t)$ erfüllt für eine Funktion $f : \mathbb{R}^3 \longrightarrow \mathbb{R}^2$.

Bemerkung 4.3.3. Bevor wir uns genauer mit solchen *mehrdimensionalen* DGL beschäfftigen, versuchen wir einfache Lösungen des Lotka-Volterra Modells (LV) zu bestimmen. Am einfachsten sind wie immer konstante Funktionen.

1. Wenn $x_B(t) = x_B$ und $x_R(t) = x_R$ konstant sind, dann gilt

$$\begin{aligned} 0 &= sx_B - ax_Bx_R \\ 0 &= -rx_R + bx_Bx_R \end{aligned}$$

   Die Lösungen sind genau $(0, 0)$ und $(\frac{r}{b}, \frac{s}{a})$. Es gibt also zwei konstante Lösungen des Lotka-Volterra Modells. Erstens $x_B(t) = x_R(t) = 0$ für alle $t \in \mathbb{R}$, und zweitens $x_B(t) = \frac{r}{b}$ und $x_R(t) = \frac{s}{a}$ für alle $t \in \mathbb{R}$. Der erste Fall sagt einfach nur aus, dass die Populationsgrößen stabil sind, wenn es keine Populationen gibt. Das ist für uns eher uninteressant. Deutlich interessanter ist die Existenz einer zweiten Lösung. Es gibt also einen Zustand in dem Räuber und Beute mit konstanter Populationsgröße koexistieren können.

2. Zwei andere Lösungen haben wir schon in unseren Annahmen beschrieben: $x_B(t) = 0$ und $x_R(t) = c \cdot e^{-rt}$, für ein $c \in \mathbb{R}$, beschreibt den Zustand wenn es keine Beutetiere gibt. Genauso beschreibt $x_B(t) = c \cdot e^{st}$ und $x_R(t) = 0$, für ein $c \in \mathbb{R}$, den Zustand ohne Räuber.

Definition 4.3.4. Sei $k \in \mathbb{N}$ und $f : \mathbb{R}^{k+1} \longrightarrow \mathbb{R}^k$ eine Funktion. Ein reelles *Differenzialgleichungssystem* in expliziter Form, ist eine Gleichung der Form

$$\nabla x(t) = f(x(t), t).$$

Eine *Lösung* des Differenzialgleichungssystems auf einem Intervall I, ist eine differenzierbare Funktion $x : \mathbb{R} \longrightarrow \mathbb{R}^k$, mit $\nabla x(t) = f(x(t), t)$ für alle $t \in I$. Zusammen mit einer Bedingung $x(t_0) = x_0$, für gegebene $t_0 \in \mathbb{R}$ und $x_0 \in \mathbb{R}^k$, nennt man das Differenzialgleichungssystem auch *Anfangswertproblem.*

Das ist wieder die einfachste Form. Es ist auch möglich, dass der Definitionsbereich von f eingeschrenkt wird (etwa da Wurzeln oder Logarithmen auftauchen). Auch dann sprechen wir noch von einem Differenzialgleichungssystem. Schreiben wir $x(t) = \begin{pmatrix} x_1(t) \\ \vdots \\ x_k(t) \end{pmatrix}$, mit differenzierbaren Komponentenfunktionen $x_i : \mathbb{R} \longrightarrow \mathbb{R}$, $i \in \{1, \ldots, k\}$, dann sieht das DGL-System folgendermaßen aus:

$$\begin{aligned} x_1'(t) &= f_1(x_1(t), x_2(t), \ldots, x_k(t), t) \\ &\vdots \qquad\qquad \vdots \\ x_k'(t) &= f_k(x_1(t), x_2(t), \ldots, x_k(t), t) \end{aligned}$$

Definition 4.3.5. Ist die Funktion f in einem Differenzialgleichungssystem nicht explizit von t abhängig, dann heißt das System *autonom.*

Genau wie bei gewöhnlichen DGL, können wir auch bei Systemen stationäre Punkte / Lösungen definieren.

Definition 4.3.6. Ist Differenzialgleichungssystem $\nabla x(t) = f(x(t), t)$ gegeben, dann heißen die Elemente $\vec{a} \in \mathbb{R}^k$, mit $f(\vec{a}, t) = 0$ für alle $t \in \mathbb{R}$, die stationären Punkte des DGL-Systems. Das sind genau die konstanten Lösungen des Systems.

Beispiel 4.3.7. Das Lotka-Volterra-Modell (LV) ist ein autonomes DGL-System. Wie in Bemerkung 4.3.3 berechnet, sind die stationären Punkte genau die Punkte $(0,0)$ und $(\frac{r}{b}, \frac{s}{a})$.

Bemerkung 4.3.8. Im Lotka-Volterra-Modell kommen vier Parameter vor. Das macht es recht Aufwendig damit zu arbeiten. Daher entdimensionalisieren wir das System erst einmal.

1. Parameter und Dimensionen auflisten:

   Die Parameter sind r, s, t, a, b, $x_B(t)$ und $x_R(t)$. Es gilt sicher $[t] = T$. Die Größe der Populationen wollen wir in Anzahl messen. Es gilt damit auch $[x_B(t)] = A_B$ und $[x_R(t)] = A_R$.

   Die Parameter r und s sind die Wachstumsraten im exponentiellen Wachstum. Es gilt also $[r] = [s] = \frac{1}{T}$. Der Wert a misst den negativen Effekt auf die Beutepopulation pro Räuber und pro Zeit. Es sollte also $[a] = \frac{1}{TA_R}$ gelten. Genauso gilt $[b] = \frac{1}{TA_B}$.

   In jedem Fall sollten wir die Konsistenz mit den Dimensionsrechenregeln überprüfen. Diese können auch zur Herleitung der Dimensionen herangezogen werden. Es ist $x_B'(t) = sx_B(t) - ax_B(t)x_R(t)$. Mit den Rechenregeln für Dimensionen gilt also

   $$\begin{aligned} & [x_B'(t)] = [sx_B(t)] = [ax_B(t)x_R(t)] \\ \iff \quad & \frac{A_B}{T} = [s] \cdot A_B = [a] \cdot A_B A_R \\ \iff \quad & [s] = \frac{1}{T} \quad \text{und } [a] = \frac{1}{TA_R}. \end{aligned}$$

   Genauso sehen wir

   $$\begin{aligned} & [x_R'(t)] = [-rx_R(t)] = [bx_B(t)x_R(t)] \\ \iff \quad & \frac{A_R}{T} = [r] \cdot A_R = [b] \cdot A_B A_R \\ \iff \quad & [r] = \frac{1}{T} \quad \text{und } [b] = \frac{1}{TA_B}. \end{aligned}$$

   Damit haben wir alle auftretenden Dimensionen herausgefunden.

2. Dimensionslose Variablen einführen:

   Seien $\bar{t}$, $a_{\bar{B}}$ und $a_{\bar{R}}$ beliebig, mit $[\bar{t}] = T$, $[a_{\bar{B}}] = A_B$ und $[a_{\bar{R}}] = A_R$. Dann setzen wir

(i) $\tau = \frac{t}{\bar{t}}$

(ii) $X_B(\tau) = \frac{x_B(t)}{\bar{a_B}} \overset{\text{(i)}}{=} \frac{x_B(\tau \cdot \bar{t})}{\bar{a_B}}$

(iii) $X_R(\tau) = \frac{x_R(t)}{\bar{a_R}} \overset{\text{(i)}}{=} \frac{x_R(\tau \cdot \bar{t})}{\bar{a_R}}$

als dimensionslose Variablen fest.

3. Das DGL-System mit den dimensionslosen Variablen formulieren:

   Es ist

$$\begin{aligned} X'_B(\tau) &\overset{\text{(ii)}}{=} \left(\frac{x_B(\tau \cdot \bar{t})}{\bar{a_B}}\right)' = \frac{\bar{t}}{\bar{a_B}} \cdot x'_B(\tau \cdot \bar{t}) \\ &= \frac{\bar{t}}{\bar{a_B}} \cdot (s x_B(\tau \cdot \bar{t}) - a x_B(\tau \cdot \bar{t}) x_R(\tau \cdot \bar{t})) \\ &\overset{\text{(ii)+(iii)}}{=} \frac{\bar{t}}{\bar{a_B}} (\cdot s \bar{a_B} X_B(\tau) - a \bar{a_B} X_B(\tau) \bar{a_R} X_R(\tau)) \\ &= \bar{t} \cdot s X_B(\tau) - \bar{t} \cdot a \bar{a_R} X_B(\tau) X_R(\tau) \end{aligned}$$

   und analog

$$X'_R(\tau) = -\bar{t} \cdot r X_R(\tau) + \bar{t} \cdot b \bar{a_B} X_B(\tau) X_R(\tau).$$

4. Einheiten wählen:

   Wir versuchen zunächst die erste Zeile des DGL-Systems zu vereinfachen. Dazu wählen wir $\bar{t} = \frac{1}{s}$ und $\bar{a_R} = \frac{1}{\bar{t} \cdot a} = \frac{s}{a}$. Beachten Sie, dass das tatsächlich die Dimensionsbedingenen $[\bar{t}] = T$ und $[\bar{a_R}] = A_R$ erfüllt. Mit dieser Wahl erhalten wir

$$\begin{aligned} X'_B(\tau) &= X_B(\tau) - X_B(\tau) X_R(\tau) \\ X'_R(\tau) &= -\frac{r}{s} X_R(\tau) + \frac{b \bar{a_B}}{s} X_B(\tau) X_R(\tau). \end{aligned}$$

   Wir setzen nun $\bar{a_B} = \frac{r}{b}$ (was wieder der Dimensionsbedingung entspricht) und erhalten

$$\begin{aligned} X'_B(\tau) &= X_B(\tau) - X_B(\tau) X_R(\tau) = X_B(\tau) \cdot (1 - X_R(\tau)) \\ X'_R(\tau) &= \frac{r}{s} \cdot (-X_R(\tau) + X_B(\tau) X_R(\tau)) = \frac{r}{s} \cdot X_R(\tau) \cdot (X_B(\tau) - 1). \end{aligned}$$

Definition 4.3.9. Das *entdimensionalisierte Lotka-Volterra-Modell* ist gegeben durch

$$\begin{aligned} X_B'(\tau) &= X_B(\tau) \cdot (1 - X_R(\tau)) \\ X_R'(\tau) &= \delta \cdot X_R(\tau) \cdot (X_B(\tau) - 1), \end{aligned} \tag{entLV}$$

mit einem dimensionslosen Parameter $\delta > 0$.

4.4 Picard-Lindelöf und Phasenportraits

Wir haben schon konstante und „halbkonstante“ Lösungen des Lotka-Volterra-Modells gefunden. Aber gibt es auch noch andere? Bei der Beantwortung dieser Frage werden wir in diesem Kapitel mit Kanonen auf Spatzen schießen und einen sehr starken Satz aus der Theorie der DGL-Systeme kennenlernen.

Definition 4.4.1. Sei $k \in \mathbb{N}$ und $f : \mathbb{R}^k \times \mathbb{R} \longrightarrow \mathbb{R}^k$ eine Funktion. Dann heißt f *lokal Lipschitz-stetig (in der ersten Variablen)*, wenn es für jede kompakte Teilmenge $K \subseteq \mathbb{R}^k \times \mathbb{R}$ eine Konstante $L_K > 0$ gibt, mit

$$|f(x,t) - f(\tilde{x},t)| \leq L_K \cdot |x - \tilde{x}| \quad \text{für alle } (x,t),\ (\tilde{x},t) \in K.$$

Bemerkung 4.4.2. Die Definition besagt, dass das Wachstum einer lokal Lipschitz-stetigen Funktion f in Richtung der ersten Variablen beschränkt ist auf kompakten Teilmengen. Damit sieht man sofort, dass die Funktion $f : \mathbb{R} \times \mathbb{R} \longrightarrow \mathbb{R}$; $(x,t) \mapsto \sqrt{|x|}$ nicht lokal Lipschitz-stetig ist.
Das erinnert an das Gegenbeispiel zur (INT-INF)-Bedingung (Bsp. 2.7.3). Tatsächlich gilt auch, dass jede lokal Lipschitz-stetige Funktion $f : \mathbb{R} \times \mathbb{R} \longrightarrow \mathbb{R}$, mit $f(x,t) = f(x,t')$ für alle $x,t,t' \in \mathbb{R}$ die (INT-INF)-Bedingung erfüllt. Mit Theorem 2.7.7 gilt also, dass mit so einer Funktion f, jedes AWP zur DGL $x'(t) = f(x(t),t)$ eine eindeutige Lösung auf einem maximalen Lösungsintervall besitzt. Diese Aussage verallgemeinert der Satz von Picard-Lindelöf auf DGL-Systeme.

Satz 4.4.3. *Jede stetig partiell-differenzierbare Funktion $f : \mathbb{R}^k \times \mathbb{R} \longrightarrow \mathbb{R}^k$ ist lokal Lipschitz-stetig.*

Beweis. Der Beweis wird in der Analysis geführt. Wir zeigen nur, dass Abbildungen der Form

$$f : \mathbb{R}^k \longrightarrow R^n \quad ; \quad \vec{x} \mapsto A \cdot \vec{x},$$

mit einer $k \times n$-Matrix A, (lokal) Lipschitz-stetig sind. Da auf dem $\mathbb{R}^{\text{irgendwas}}$ alle Normen äquivalent sind, können wir uns die Norm frei aussuchen. Wir werden hier mit der 1-Norm $\left| \begin{pmatrix} x_1 \\ \dots \\ x_k \end{pmatrix} \right| = |x_1| + \dots + |x_k|$. Setzen wir dann

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nk} \end{pmatrix}$$

und es folgt

$$\begin{aligned} |f(\vec{x}) - f(\vec{y})| = |A \cdot (\vec{x} - \vec{y})| &= \left| \sum_{i=1}^{n} \sum_{j=1}^{k} a_{ij} \cdot (x_j - y_j) \right| \\ &\leq \max_{1 \leq i \leq n \text{ und } 1 \leq j \leq k} |a_{ij}| \cdot \left| \sum_{i=1}^{n} \sum_{j=1}^{k} (x_j - y_j) \right| \\ &\underbrace{\leq \max_{1 \leq i \leq n \text{ und } 1 \leq j \leq k} |a_{ij}| \cdot n}_{L} \cdot |\vec{x} - \vec{y}|. \end{aligned}$$

Das wollten wir zeigen. Da diese Ungleichung für alle $\vec{x}, \vec{y} \in \mathbb{R}^k$ gilt, gilt sie auch für alle kompakten Teilmengen. □

Theorem 4.4.4 (Satz von Picard-Lindelöf)**.** *Sei $k \in \mathbb{N}$ und $f : \mathbb{R}^k \times \mathbb{R} \longrightarrow \mathbb{R}^k$ stetig und lokal Lipschitz-stetig (in der ersten Variablen). Sei weiter $(x_0, t_0) \in \mathbb{R}^k \times \mathbb{R}$. Dann hat das Anfangswertproblem*

$$\nabla x(t) = f(x(t), t) \quad \text{und} \quad x(t_0) = x_0$$

eine eindeutige Lösung $x(t)$ auf einem maximalen Definitionsintervall $I = (t_-, t_+) \subseteq \mathbb{R}$. Diese Lösung läuft von „Rand zu Rand“. D.h.: es ist

$$t_- = -\infty \text{ oder } \lim_{t \to t_-} |x(t)| = \infty \text{ (oder beides), und}$$

$$t_+ = \infty \text{ oder } \lim_{t \to t_+} |x(t)| = \infty \text{ (oder beides).}$$

Bemerkung 4.4.5. Die Lösungen eines DGL-Systems sind Funktionen ï¿¼$x : \mathbb{R} \longrightarrow \mathbb{R}^k$. Für $k > 1$ können wir diese als Kurven im $\mathbb{R}^k$ darstellen. Diese Darstellungen sind invariant unter Verschiebung von t. D.h.: die Kurve von $x(t)$ ist das gleiche wie die Kurve von $x(t + t_0)$ für jedes $t_0 \in \mathbb{R}$.

Definition 4.4.6. Die verschiedenen Kurven, die die Lösungen eines DGL-Systems darstellen, heißen die *Trajektorien* des DGL-Systems.

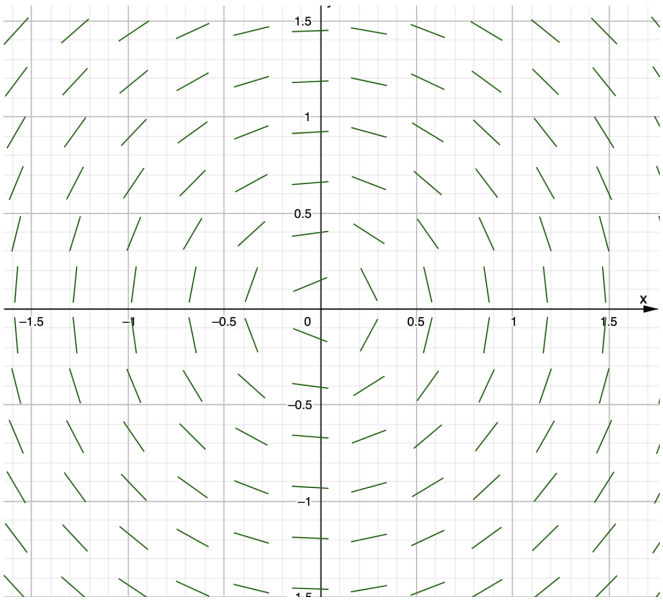
Satz 4.4.7. *Unter den Voraussetzungen des Satzes von Picard-Lindelöf, können sich verschiedene Trajektorien eines* autonomen *DGL-Systems nicht schneiden.*

Beweis. Beschreiben $x(t)$ und $y(t)$ zwei Trajektorien, die sich schneiden, dann ist $x(t_0) = y(t_1) = x_0$ für gewisse $t_0, t_1 \in \mathbb{R}$ und $x_0 \in \mathbb{R}^k$. Die Funktion $\tilde{y}(t) = y(t - t_0 + t_1)$ beschreibt die gleiche Trajektorie wie $y(t)$ und ist ebenfalls eine Lösung des DGL-Systems. Denn: $\tilde{y}'(t) = y'(t - t_0 + t_1) = f(y(t - t_0 + t_1)) = f(\tilde{y}(t))$ (hierfür brauchen wir, dass das System autonom ist). Damit sind nun aber $x(t)$ und $\tilde{y}(t)$ unterschiedliche Lösungen des DGL-Systems, mit Anfangswert $t_0 \mapsto x_0$, im Widerspruch zum Satz von Picard-Lindelöf. □

Beispiel 4.4.8. Wir betrachten das DGL-System

$$\begin{pmatrix} x'(t) \\ y'(t) \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \cdot \begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = \begin{pmatrix} y(t) \\ -x(t) \end{pmatrix}.$$

Da die Abbildung $\begin{pmatrix} x \\ y \end{pmatrix} \mapsto f(x, y) = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \cdot \begin{pmatrix} x \\ y \end{pmatrix}$ (lokal) Lipschitz-stetig ist (siehe Satz 4.4.3), besitzt das DGL-System nach dem Satz von Picard-Lindelöf 4.4.4 zu jedem Anfangswert genau eine Lösung.
Wir möchten die Lösungskurven (Trajektorien) des DGL-Systems in den $\mathbb{R}^2$ einzeichnen, *bevor wir eine Lösung berechnen*! Dazu überlegen wir uns zunächst, was uns die Aussage $f(1, 0) = \begin{pmatrix} 0 \\ -1 \end{pmatrix}$ über Lösungen des DGL-Systems aussagt. Das sagt doch nichts anderes, als dass die Ableitung der Lösung, die durch den Punkt $(1, 0)$ verläuft, an dieser Stelle gleich $(0\ ,\ -1)$ ist. Die Steigung in x-Richtung ist also gleich 0. Die Trajektorie schneidet also die x-Achse senkrecht im Punkt $(1, 0)$.
Zeichnen wir nun an ganz vielen Stellen die Steigung der Trajektorien an diesem Punkt als Strecke mit gleicher Länge ein, erhalten wir das *Phasenportrait* des DGL-Systems. Dieses gibt ein grobes Bild der Trajektorien an. In diesem Fall erhalten wir



Dieses Phasenportrait deutet darauf hin, dass die Trajektorien Kreise um den Ursprung sind. Wir suchen nun alle Lösungen des gegebenen DGL-Systems

$$x'(t) = y(t)$$
$$y'(t) = -x(t)$$

Wir können das DGL-System umschreiben zu $x''(t) = -x(t)$. Wir haben also aus einem DGL-System in zwei Dimensionen eine DGL der Ordnung zwei gemacht! Zwei Lösungen dieser DGL kennen wir alle: $x_1(t) = \cos(t)$ und $x_2(t) = \sin(t)$. Damit ist aber auch $x(t) = c_1 \cos(t) + c_2 \sin(t)$ eine Lösung für alle $c_1, c_2 \in \mathbb{R}$. Damit sind Lösungen des DGL-Systems gegeben durch

$$\begin{aligned} x(t) &= c_1 \cos(t) + c_2 \sin(t) \\ y(t) &= x'(t) = -c_1 \sin(t) + c_2 \cos(t) \end{aligned} \tag{4.7}$$

mit $c_1, c_2 \in \mathbb{R}$ beliebig. Das sind sogar alle Lösungen des DGL-Systems. Denn:
Für jede Wahl von $((x_0, y_0), t_0) \in \mathbb{R}^2 \times \mathbb{R}$ gibt es genau ein $(c_1, c_2) \in \mathbb{R}^2$, mit

$$\begin{aligned} x_0 &= c_1 \cos(t_0) + c_2 \sin(t_0) \\ y_0 &= x'(t_0) = -c_1 \sin(t_0) + c_2 \cos(t_0), \end{aligned}$$

da die Matrix $\begin{pmatrix} \cos(t_0) & \sin(t_0) \\ -\sin(t_0) & \cos(t_0) \end{pmatrix}$ regulär ist. Nach dem Satz von Picard-Lindelöf 4.4.4 ist damit (4.7), mit diesen $c_1, c_2 \in \mathbb{R}$, die eindeutige Lösung des AWP. Wie das Phasenportrait vermuten ließ, sind damit alle Trajektorien Kreise um den Ursprung.

Bemerkung 4.4.9. Die Lösungen im Beispiel 4.4.8 sind 2π-*periodisch*. Dass die Lösungen periodisch sind erkennt man auch daran, dass die Trajektorien geschlossenen Kurven sind: Irgendwann sind wir dort angekommen, wo wir schon einmal waren und alles beginnt von vorne...

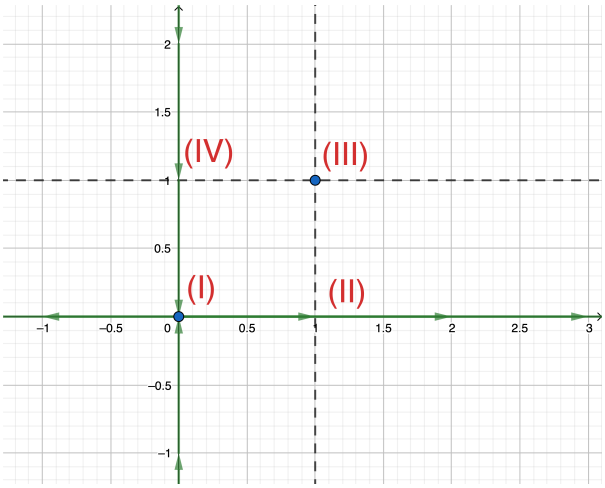
4.5 Zurück zum Lotka-Volterra-Modell

Das Lotka-Volterra-Modell beschreibt das Wachstum zweier Spezies in einem Räuber-Beute-Verhältnis. Um etwas sparsamer mit Indizes zu sein, beschreibt ab jetzt $x(t)$ die Größe der Beute-Population und $y(t)$ die Grösse der Räuber-Population zum Zeitpunkt t. Das entdimensionalisierte Lotka-Volterra-Modell (entLV) ist damit

$$\begin{aligned} X'(\tau) &= X(\tau) \cdot (1 - Y(\tau)) \\ Y'(\tau) &= \delta \cdot Y(\tau) \cdot (X(\tau) - 1), \end{aligned}$$

mit $\delta > 0$. Die stationären Punkte sind $(0, 0)$ und $(1, 1)$. Weitere Lösungen sind $X(\tau) = c_1 e^{-\delta\tau}$ und $Y(\tau) = 0$ sowie $X(\tau) = 0$ und $Y(\tau) = c_2 e^{\tau}$. Es gibt also Lösungen, die vom Nullpunkt startend, die beiden Koordinaten Achsen (in beide Richtungen) durchlaufen. Auf der rechten Seite des Modells stehen nur Polynome, und somit ist insbesondere alles stetig differenzierbar und daher auch lokal Lipschitz-stetig.

Bemerkung 4.5.1. Der Satz von Picard-Lindelöf 4.4.4 ist anwendbar und daher können sich zwei verschiedene Lösungskurven nicht schneiden. Mit den Lösungen, die wir bereits gefunden haben, folgt damit, dass eine Lösung, die im ersten Quadranten startet für immer im ersten Quadranten bleibt (siehe Skizze unten). Natürlich gilt das auch für den zweiten, dritten und vierten Quadranten, aber mit dem Hintergrund der Modellierung, ist für uns nur der erste Quadrant von Interesse, da es keine negativen Populationsgrößen gibt.



Die Lösungskurven / Trajektorien mit $X(\tau) = 0$, bzw. $Y(\tau) = 0$, haben wir in grün gezeichnet. Wir möchten untersuchen, in welchen Bereichen eine Trajektorie in X- und Y-Richtung steigt oder fällt. Wir wolle nalso eine ganz grobe Idee des Phasenportraits bekommen. Dazu überlegen wir uns zunächst, wann X' bzw Y' verschwindet und erhalten neben den Koordinatenachsen noch zwei weitere Geraden (gestrichelt eingezeichnet). Der Schnittpunkt dieser Geraden ist natürlich der stationäre Punkt, da das genau der Punkt ist, in dem beide Ableitungen gleich Null sind. Der erste Quadrant wird also in vier Abschnitte (I), (II), (III), (IV) unterteilt. Genauer

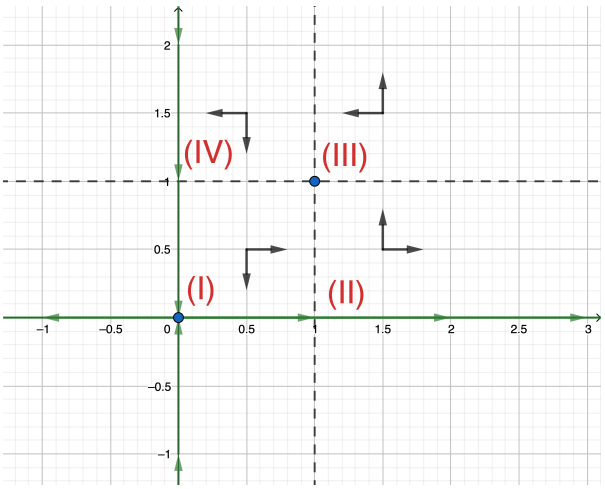
- (I) $= \{(x, y) | 0 \leq x < 1,\ 0 \leq y < 1\}$
- (II) $= \{(x, y) | 1 \leq x,\ 0 \leq y < 1\}$
- (III) $= \{(x, y) | 1 \leq x,\ 1 \leq y\}$
- (IV) $= \{(x, y) | 0 \leq x < 1,\ 1 \leq y\}$

Ist nun X, Y irgendeine Lösung des entdimensionalisierten Lotka-Volterra-Modells und ist $\tau_0 \in \mathbb{R}$ gegeben, mit $(X(\tau_0), Y(\tau_0)) \in$ (I), so ist

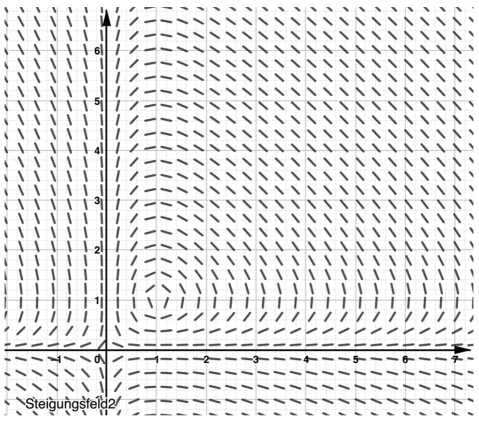
$$X'(\tau_0) = X(\tau_0) \cdot (1 - Y(\tau_0)) > 0 \quad \text{und } Y'(\tau_0) = \delta \cdot Y(\tau_0) \cdot (X(\tau_0) - 1) < 0.$$

Damit verläuft jede Trajektorie im Bereich (I), von links nach rechts (in X-Richtung positiv) und von oben nach unten (in Y-Richtung negativ). Genau so sehen wir, dass in (II) jede Trajektorie von links nach rechts und

von unten nach oben verläuft, und entsprechendes für die Bereiche (III) und (IV). Das markieren wir mit einem Pfeilepaar in jedem Bereich.



Weiter wissen wir, dass die gestrichelten Linien immer senkrecht geschnitten werden. Mögliche Verläufe für die Trajektorien wären zum Beispiel Kreisel um den stationären Punkt $(1, 1)$ – entweder auf den Punkt zu, oder von diesem Punkt weg bewegend. Auch geschlossene Kurven um den Punkt $(1, 1)$ würden zu diesen Überlegungen passen. Das Phasenportrait (für den Fall $\delta = 1$) sieht aus



Das würde immer noch zu Kreiseln und geschlossenen Kurven passen.

Definition 4.5.2. Sei $\nabla x(t) = f(x(t))$ ein autonomes DGL-System, mit $x(t) = \begin{pmatrix} x_1(t) \\ \vdots \\ x_k(t) \end{pmatrix}$. Eine *Lyapunov-Funktion* für das DGL-System, ist eien

Funktion $V : \mathbb{R}^k \longrightarrow \mathbb{R}$, so dass für jede Lösung $x(t)$ des DGL-Systems auf einem Intervall I, eine Konstante $c \in \mathbb{R}$ existiert, mit $V(x(t)) = c$ für alle $t \in I$.

Bemerkung 4.5.3. Die Bezeichnung *Lyapunov-Funktion* für eine solche Funktion ist nicht ganz einheitlich. Oft wird bei einer *Lyapunov-Funktion* nur $(V \circ x)'(t) \leq 0$ verlangt. In unserer Definition bedeutet das, dass die Lösungskurven des DGL-Systems alle in einer Niveau-Linie der Funktion V gefangen sind. Diese können dann untersucht werden um Aussagen über die Form der der Lösungskurven zu tätigen. Das ist nur dann hilfreich, wenn es auch wirklich Höhenlinien gibt. Wenn wir also einfach eine Konstante Funktion f ür V wählen, so ist zwar die Eigenschaft einer Lyapunov-Funktion gegeben, allerdings sagt diese nichts über die Verläufe der Lösungskurven aus.

Es ist oft nicht möglich eine Lyapunov-Funktion explizit zu bestimmen. Beim entdimensionierten Lotka-Volterra-Modell, versuchen wir im folgenden unser Glück.

Sei also das entdimensionalisierte Lotka-Volterra-Modell

$$\begin{aligned} X'(\tau) &= X(\tau) \cdot (1 - Y(\tau)) \\ Y'(\tau) &= \delta \cdot Y(\tau) \cdot (X(\tau) - 1), \end{aligned}$$

gegeben und sei $(X(\tau), Y(\tau))$ eine Lösung auf einem maximalen Definitionsintervall I, deren Kurve im ersten Quadranten liegt. Dann ist eine Lyapunov-Funktion eine Funktion $V : (0, \infty)^2 \longrightarrow \mathbb{R}$, mit $V(X(\tau), Y(\tau)) = c$ für alle $\tau \in I$ und ein festes $c \in \mathbb{R}$.

Wir nehmen an, dass es so eine Funktion gibt. Dann gilt für alle $\tau \in I$

$$\begin{aligned}
& V(X(\tau), Y(\tau)) = c \\
\iff & (V(X(\tau), Y(\tau)))' = 0 \\
\iff & \nabla V(X(\tau), Y(\tau)) \cdot \begin{pmatrix} X'(\tau) \\ Y'(\tau) \end{pmatrix} = 0 \\
\iff & \left(\frac{\partial V}{\partial X}(X(\tau), Y(\tau)) \ , \ \frac{\partial V}{\partial Y}(X(\tau), Y(\tau)) \right) \cdot \begin{pmatrix} X(\tau)(1 - Y(\tau)) \\ \delta Y(\tau)(X(\tau) - 1) \end{pmatrix} = 0 \\
\iff & \frac{\partial V}{\partial X}(X(\tau), Y(\tau)) \cdot X(\tau)(1 - Y(\tau)) + \frac{\partial V}{\partial Y}(X(\tau), Y(\tau)) \cdot \delta Y(\tau)(X(\tau) - 1) = 0 \\
\overset{\cdot \frac{1}{X(\tau)Y(\tau)}}{\iff} & \frac{\partial V}{\partial X}(X(\tau), Y(\tau)) \cdot \frac{1 - Y(\tau)}{Y(\tau)} + \frac{\partial V}{\partial Y}(X(\tau), Y(\tau)) \cdot \frac{\delta(X(\tau) - 1)}{X(\tau)} = 0
\end{aligned}$$

Der Faktor, der mit der Ableitung von V in X-Richtung multipliziert wird, hängt nur noch von Y ab. Der Faktor, der mit der Ableitung von V in Y-Richtung multipliziert wird, hängt nur noch von X ab. Damit kann nun ziemlich einfach eine passende Funktion V berechnet werden. Denn die letzte Aussage (und damit auch die erste) ist wahr, wenn gilt

$$\frac{\partial V}{\partial X}(X(\tau), Y(\tau)) = \delta \frac{X(\tau) - 1}{X(\tau)} \quad \text{und} \quad \frac{\partial V}{\partial Y}(X(\tau), Y(\tau)) = -\frac{1 - Y(\tau)}{Y(\tau)}.$$

Dies ist wiederum der Fall, wenn $V(x, y) = F(x) + G(y)$ gilt, mit $F'(x) = \delta \frac{x-1}{x}$ und $G'(y) = -\frac{1-y}{y}$. Eine Stammfunktion von F und eine von G ist aber ganz einfach berechnet. Damit erhalten wir

$$V(x, y) = F(x) + G(y) = \delta \left(x - \ln(x) \right) + \left(y - \ln(y) \right). \tag{4.8}$$

Diese Funktion erfüllt alle äquivalenten Aussagen. Insbesondere handelt es sich hierbei um eine Lyapunov-Funktion des entdimensionalisierten Lotka-Volterra-Modells.

Bemerkung 4.5.4. Kommt Ihnen das bekannt vor? Im wesentlichen haben wir hier wieder die Variablen getrennt! Wann immer es definiert ist, ist

$$\frac{X'(\tau)}{X(\tau) \cdot (1 - Y(\tau))} = 1 = \frac{Y'(\tau)}{\delta Y(\tau) \cdot (X(\tau) - 1)}.$$

Umstellen liefert

$$X'(\tau) \cdot \frac{X(\tau) - 1}{X(\tau)} = Y'(\tau) \cdot \frac{1 - Y(\tau)}{\delta Y(\tau)}.$$

Wir integrieren beide Seiten nach τ und erhalten mit der Substitutionsregel (und den üblichen Notationen)

$$\begin{aligned} & \int X'(\tau) \cdot \frac{X(\tau)-1}{X(\tau)} \mathrm{d}\tau = \int Y'(\tau) \cdot \frac{1-Y(\tau)}{\delta Y(\tau)} \mathrm{d}\tau \\ \iff & \int \frac{X-1}{X} \mathrm{d}X = \int \frac{1-Y}{\delta Y} \mathrm{d}Y \\ \iff & X(\tau) - \ln|X(\tau)| = \delta^{-1} \cdot (\ln|Y(\tau)| - Y(\tau)) + C, \quad C \in \mathbb{R}. \end{aligned}$$

Das ist exakt die Bedingung $V(X(\tau), Y(\tau)) = C$, mit V aus (4.8).

Diese Lyapunov-Funktion V aus (4.8) besitzt eine einzige Extremstelle, nämlich ein lokales Minimum im stationären Punkt $(1,1)$. Weiter strebt $V(x,y)$ gegen unendlich, wenn (x,y) gegen den Rand von $(0,\infty)^2$ konvergiert. Unsere Anschauung sagt uns (ACHTUNG, DASS IST KEIN BEWEIS!), dass der Graph von V in etwa aussieht, wie ein Paraboloid (ein Becher). Die Niveau-Linien von V sind somit geschlossene Kurven um den Punkt $(1,1)$. Damit liegt jede Lösungskurve in einer geschlossenen Kurve.

Satz 4.5.5. *Die Lösungskurven / Trajektorien des entdimensionalisierten Lotka-Volterra-Modells im ersten Quadranten, sind geschlossene Kurven um den stationären Punkt* $(1,1)$.

Beweis. Die Idee des Beweises ist ganz einfach: Die Lösungskurven sind entweder geschlossenen Kurven oder „gebogene“ offene Intervalle. Nun können offene Intervalle aber nicht lückenlos ohne überschneidung aneinander gesetzt werden. Da auf einer geschlossenen Kurve um $(1,1)$ im ersten Quadranten aber keine stationären Punkte liegen, haben wir nur solche offenen Intervalle zur Verfügung um die Niveau-Linie zu überdecken.
Nun ein tatsächlicher Beweis: Sei $(X(\tau), Y(\tau))$ eine Lösung des Systems, auf dem maximalen Intervall $I = (t_-, t_+)$. Wie wir gerade gesehen haben, liegt die zugehörige Lösungskurve in einer geschlossenen Niveau-Linie von V. Insbesondere ist die Lösungskurve beschränkt und es gilt $|(X(\tau), Y(\tau))| < c$, für ein $c \in \mathbb{R}$. Mit dem Satz von Picard-Lindelöf 4.4.4 ist damit $I = (-\infty, \infty)$. Wenn die Lösungskurve keine geschlossene Kurve wäre, würde sie in einem der Bereiche (I), (II), (III), oder (IV) enden. Wir haben aber schon festgestellt, dass das Wachstum in X- und in Y-Richtung monoton ist in jedem einzelnen dieser Bereiche. Damit wären X und Y monotone beschränkte

Funktion und damit folgt $\lim_{\tau\to\infty}(X(\tau), Y(\tau)) = (a, b)$ für einen Punkt aus dem ersten Quadranten. Dieser Punkt liegt in einem der vier Bereiche (I), (II), (III), oder (IV), ist aber sicher nicht gleich $(1, 1)$, da sich der Punkt auf einer geschlossenen Kurve *um* den Punkt $(1, 1)$ befindet. Da nun $X(\tau)$ gegen die konstante Funktion a und $Y(\tau)$ gegen die konstante Funktion b konvergieren, folgt $\lim_{\tau\to\infty} X'(\tau) = 0 = \lim_{\tau\to\infty} Y'(\tau)$. Damit folgt

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} = \lim_{\tau\to\infty} \begin{pmatrix} X(\tau) \\ Y(\tau) \end{pmatrix} = \lim_{\tau\to\infty} \begin{pmatrix} X(\tau)(1 - Y(\tau)) \\ \delta Y(\tau)(X(\tau) - 1) \end{pmatrix} = \begin{pmatrix} a(1-b) \\ \delta b(a-1) \end{pmatrix}.$$

Das impliziert aber $(a, b) \in \{(0, 0), (1, 1)\}$, was ausgeschlossen ist. Damit haben wir endlich den gesuchten Widerspruch und in jeder Niveau-Linie liegt genau eine Trajektorie des Lotka-Volterra-Modells. □

Wir haben in dem Beweis eine bekannte Aussage wiederentdeckt. Nämlich, dass der Grenzwert einer Lösung des autonomen DGL-Systems ein stationärer Punk des Systems ist. Das kennne wir bereits von den Differenzialgleichungen aus Lemma 2.7.6. Das gilt auch in dieser Situation.

Lemma 4.5.6. *Sei $\nabla\vec{x}(t) = f(\vec{x}(t))$ ein autonomes DGL-System, mit lokal Lipschitz-stetigem $f : \mathbb{R}^n \longrightarrow \mathbb{R}^n$. Ist nun $\vec{x}(t)$ eine Lösung auf $\mathbb{R}$ und existiert $\lim_{t\to\infty} \vec{x}(t) \in \mathbb{R}^n$, so ist $\lim_{t\to\infty} \vec{x}(t)$ ein stationärer Punkt. Existiert $\lim_{t\to-\infty} \vec{x}(t) \in \mathbb{R}^n$, so ist auch dies ein stationärer Punkt.*

Aus Satz 4.5.5 folgt die

1. Lotka-Volterra-Regel: Die Beute- und die Räuberpopulation verhalten sich periodisch. Dabei folgen die Schwankungen der Räuberpopulation phasenverzögert denen der Beutepopulation.

Gibt es viele Beutetiere und wenig Räuber, finden die Räuber immer genug zu fressen und können sich weiter vermehren. Dadurch steigt die Größe der Räuberpopulation. Erst wenn ein kritischer Wert an Raubtieren erreicht ist, beginnt die Anzahl von Beutetieren zu sinken (Übergang vom Sektor (II) in Sektor (III)). Ab dann sinkt die Zahl der Beutetiere und die Zahl der Raubiere steigt weiter, bis der nächste kritische Wert an Beutetieren erreicht ist (Übergang von Sektor (III) in Sektor (IV)). Dann finden die Räuber nicht mehr genug zu fressen um sich weiter auszubreiten. Als Folge sinken sowohl die Anzahl der Raubtiere als auch die Anzahl der Beutetiere. Erst wenn sich

die Anzahl von Raubtieren genügend reduziert hat (Übergang von Sektor (IV) in Sektor (I)) fängt die Beutepopulation an sich zu erholen. Sind dann wieder genug Beutetiere da (Übergang von Sektor (I) in Sektor (II)) geht alles von vorne los.

Bemerkung 4.5.7. Das Gleiche Phänomen ist auch in der Marktwirtschaft zu beobachten, wenn $x(t)$ den Preis eines Produkten und $y(t)$ das Angebot dieses Produktes zum Zeitpunkt t beschreibt. Auch hier folgt das Angebot zeitlich verzögert den Schwankungen des Preises. Da dies zum ersten mal in der Agrarwissenschaft anhand von Preisen für Schweine beschrieben wurde, wird die erste Lotka-Volterra-Regel manchmal auch *Schweinezyklus* genannt.

2. Lotka-Volterra-Regel: Die über genügend lange Zeiträume gemittelten Größen der Räuber- bzw. Beutepopulation sind konstant. Die Größe der Mittelwerte hängt nur von den Wachstums- und Rückgangsraten der Populationen, nicht aber von den Anfangsbedingungen ab.

Diese zweite Regel müssen wir noch beweisen. Sei dazu $(X(\tau), Y(\tau))$ eine nicht-konstante Lösung des entdimensionalisierten Lotka-Volterra-Modells. Wir wissen bereits, dass die Funktion $(X(\tau), Y(\tau))$ periodisch ist. Eine Periode sei gegeben durch den Wert $T > 0$. D.h. $(X(\tau_0 + T), Y(\tau_0 + T)) = (X(\tau_0), Y(\tau_0))$ für alle $\tau_0 \in \mathbb{R}$. Wir nutzen $Y'(\tau) = \delta Y(\tau)(X(\tau) - 1) \Longrightarrow X(\tau) = \frac{1}{\delta}\frac{Y'(\tau)}{Y(\tau)} + 1$. Dann ist die mittlere Populationsgröße der Beutepopulation in einem Zeitintervall der Länge T gegeben durch den Mittelwert

$$\begin{aligned}\frac{1}{T}\int_{\tau_0}^{\tau_0+T} X(\tau)\mathrm{d}\tau &= \frac{1}{T}\int_{\tau_0}^{\tau_0+T} \frac{1}{\delta}\frac{Y'(\tau)}{Y(\tau)} + 1\mathrm{d}\tau = \frac{1}{T}\cdot\left[\frac{1}{\delta}\ln(Y(\tau)) + \tau\right]_{\tau_0}^{\tau_0+T} \\ &= 1 + \frac{1}{T\delta}\left(\ln(Y(\tau_0+T) - \ln(Y(\tau_0+T))\right) = 1.\end{aligned}$$

Das letzte Gleichheitszeichen folgt, da Y eine T-periodische Funktion ist. Genauso zeigt man auch für den Mittelwert der Räuberpopulation

$$\frac{1}{T}\int_{\tau_0}^{\tau_0+T} Y(\tau)\mathrm{d}\tau = 1.$$

Das zeigt, dass die Mittelwerte unabhängig vom Anfangswert immer gleich sind. Weiter sehen wir, da wir im entdimensionalisierten Fall sind, dass die mittlere Populationsgröße immer eine Einheit ist. Die Einheit für A_B, die

Anzahl der Beutetiere, war $\frac{r}{b}$ und die Anzahl, und die Einheit für A_R, die Anzahl der Raubtiere, war $\frac{s}{a}$ (siehe Bemerkung 4.3.8). Damit sind die mittleren Populationsgrößen im Lotka-Volterra-Modell (LV) gegeben durch $\frac{r}{b}$ für die Beutetiere und $\frac{s}{a}$ für die Raubtiere.

4.6 Lineare DGL-Systeme der mit konstanten Koeffizienten

In diesem Abschnitt betrachten wir kurz DGL-System der Form $x'(t) = A \cdot x(t)$, mit einer 2×2-Matrix A und einer Funktion $x : \mathbb{R} \longrightarrow \mathbb{R}^2$.

Definition 4.6.1. Ein *homogenes lineares DGL-System der Ordnung n mit konstanten Koeffizienten* ein DGL-System der Form

$$\begin{pmatrix} x_1'(t) \\ \vdots \\ x_n'(t) \end{pmatrix} = A \cdot \begin{pmatrix} x_1(t) \\ \vdots \\ x_n(t) \end{pmatrix}, \tag{4.9}$$

wobei A eine $n \times n$-Matrix mit reellen Koeffizienten ist.

Bemerkung 4.6.2. Wieder kommt das Wort *linear* in der Definition von den bekannten Eigenschaften aus Abschnitt 2.6. Seien nun $x(t)$ und $y(t)$ zwei Lösungen von (4.9) und sei $\lambda \in \mathbb{R}$ beliebig. Dann gilt, dass auch $(x + y)(t)$ und $\lambda \cdot x(t)$ Lösungen von (4.9) sind. Das sieht man ganz einfach ein:

- $(x + y)'(t) = x'(t) + y'(t) = A \cdot x(t) + A \cdot y(t) = A \cdot (x + y)(t)$ und
- $(\lambda \cdot x(t))' = \lambda \cdot x'(t) = \lambda \cdot A \cdot x(t) = A \cdot (\lambda \cdot x(t))$.

In den Übungen wurde der folgende Satz bereits vermutet.

Satz 4.6.3. *Alle Lösungen von (4.9) sind gegeben durch*

$$x(t) = \exp(A \cdot t) \cdot \vec{c}, \text{ mit } \vec{c} \in \mathbb{R}^n.$$

Hier ist $\exp(A \cdot t) = \sum_{k=0}^{\infty} \frac{1}{k!} \cdot (A \cdot t)^k$.

Beweis. Die Funktionen $\exp(A \cdot t) \cdot \vec{c}$ lösen natürlich für jedes $\vec{c} \in \mathbb{R}^n$ das DGL-System (4.9). Weiter ist $\exp(A \cdot t)$ stets eine reguläre Matrix, da wie

für reelle Zahlen gilt

$$\exp(A \cdot t) \cdot \exp((-1) \cdot A \cdot t) = \exp\left(\begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{pmatrix}\right) = E_n,$$

wobei E_n die Einheitsmatrix mit n Zeilen bezeichnet.
Für jeden Anfangswert $x(t_0) = x_0 \in \mathbb{R}^n$, $t_0 \in \mathbb{R}$, gibt es daher genau einen Vektor $\vec{c} \in \mathbb{R}^n$, mit $\exp(A \cdot t_0) \cdot \vec{c} = x_0$. Die Funktion $t \mapsto \exp(A \cdot t) \cdot \vec{c}$ ist somit eine Lösung des AWPs bestehend aus (4.9) und $x(t_0) = x_0$. Nach Satz 4.4.3 ist die rechte Seite des DGL-Systems lokal Lipschitz-stetig. Damit besagt der Satz von Picard-Lindelöf 4.4.4, dass die Funktion $t \mapsto \exp(A \cdot t) \cdot \vec{c}$ die *einzige* Lösung des AWPs ist. Damit ist wie behauptet jede Lösung ders DGL-Systems von der angegebenen Form. □

Beispiel 4.6.4. Wir berechnen $\exp\left(\begin{pmatrix} 0 & t \\ -t & 0 \end{pmatrix}\right)$, für $t \in \mathbb{R}$. Wenn wir $\begin{pmatrix} 0 & t \\ -t & 0 \end{pmatrix}^n = \begin{pmatrix} a_n & b_n \\ c_n & d_n \end{pmatrix}$ setzen, dann finden wir per Induktion schnell heraus, dass für alle $n \in \mathbb{N}_0$ gilt

$$a_n = d_n = \begin{cases} 0 & \text{falls } n \text{ ungerade} \\ (-1)^{n/2} t^n & \text{falls } n \text{ gerade} \end{cases}$$

$$b_n = -c_n = \begin{cases} 0 & \text{falls } n \text{ gerade} \\ (-1)^{(n-1)/2} t^n & \text{falls } n \text{ ungerade} \end{cases}.$$

Es folgt

$$\exp\left(\begin{pmatrix} 0 & t \\ -t & 0 \end{pmatrix}\right) = \begin{pmatrix} \sum_{k=0}^{\infty} \frac{(-1)^k \cdot t^{2k}}{(2k)!} & \sum_{k=0}^{\infty} \frac{(-1)^k \cdot t^{2k+1}}{(2k+1)!} \\ -\sum_{k=0}^{\infty} \frac{(-1)^k \cdot t^{2k+1}}{(2k+1)!} & \sum_{k=0}^{\infty} \frac{(-1)^k \cdot t^{2k}}{(2k)!} \end{pmatrix}$$
$$= \begin{pmatrix} \cos(t) & \sin(t) \\ -\sin(t) & \cos(t) \end{pmatrix}.$$

Es folgt (schon wieder, vgl. Bsp. 4.4.8), dass die Lösungen von $\begin{pmatrix} x'(t) \\ y'(t) \end{pmatrix} =$

$\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \cdot \begin{pmatrix} x(t) \\ y(t) \end{pmatrix}$ gegeben sind durch

$$\begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = \exp\left(\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \cdot t\right) \cdot \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} \cos(t) & \sin(t) \\ -\sin(t) & \cos(t) \end{pmatrix} \cdot \begin{pmatrix} c_1 \\ c_2 \end{pmatrix},$$

für $c_1, c_2 \in \mathbb{R}$ beliebig.

Lemma 4.6.5. *Seien S und A zwei $n \times n$-Matrizen, mit S invertierbar. Dann gilt für alle $t \in \mathbb{R}$*

$$\exp(S^{-1} \cdot A \cdot S \cdot t) = S^{-1} \cdot \exp(A \cdot t) \cdot S.$$

Beweis. Da Multiplikation mit einem Skalar kommutativ ist, und sich die inversen Matrizen beim Potenzieren aufheben, gilt für alle $k \in \mathbb{N}_0$

$$\left(S^{-1} \cdot A \cdot S \cdot t\right)^k = \left(S^{-1} \cdot (A \cdot t) \cdot S\right)^k = S^{-1} \cdot (A \cdot t)^k \cdot S.$$

Damit folgt nun

$$\begin{aligned} \exp(S^{-1}ASt) &= \lim_{n\to\infty} \sum_{k=0}^{n} \frac{1}{k!} \left(S^{-1} \cdot A \cdot S \cdot t\right)^k = \lim_{n\to\infty} \sum_{k=0}^{n} \frac{1}{k!} S^{-1} \cdot (A \cdot t)^k \cdot S \\ &= \lim_{n\to\infty} S^{-1} \cdot \left(\sum_{k=0}^{n} \frac{1}{k!} (A \cdot t)^k\right) \cdot S \\ &= S^{-1} \cdot \left(\lim_{n\to\infty} \sum_{k=0}^{n} \frac{1}{k!} (A \cdot t)^k\right) \cdot S \\ &= S^{-1} \cdot \exp(A \cdot t) \cdot S. \end{aligned}$$

□

Satz 4.6.6. *Sei A eine diagonalisierbare reelle 2×2-Matrix mit Eigenwerten λ_1 und λ_2. Seien weiter v_1 und v_2 linear unabhängige Eigenvektoren von A, so dass v_i Eigenvektor zu λ_i ist für $i \in \{1, 2\}$. (Diese Eigenvektoren existieren, da A diagonalisierbar ist.) Dann sind alle Lösungen von*

$$\begin{pmatrix} x'(t) \\ y'(t) \end{pmatrix} = A \cdot \begin{pmatrix} x(t) \\ y(t) \end{pmatrix}$$

gegeben durch

$$\begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = c_1 \cdot e^{\lambda_1 \cdot t} \cdot v_1 + c_2 \cdot e^{\lambda_2 \cdot t} \cdot v_2, \quad c_1, c_2 \in \mathbb{R}.$$

BEWEIS. Ist S die Matrix mit den Spaltenvektoren v_1 und v_2 (in dieser Reihenfolge), dann gilt $S^{-1} \cdot A \cdot S = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$ und somit

$$A = S \cdot \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \cdot S^{-1}.$$

Damit ist nach Lemma 4.6.5

$$\exp(A \cdot t) = S \cdot \exp\left(\begin{pmatrix} \lambda_1 t & 0 \\ 0 & \lambda_2 t \end{pmatrix}\right) \cdot S^{-1} = S \cdot \begin{pmatrix} e^{\lambda_1 t} & 0 \\ 0 & e^{\lambda_2 t} \end{pmatrix} \cdot S^{-1}.$$

Die Lösungen des DGL-Systems sind nach Satz 4.6.3 gegeben durch

$$\begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = \exp(A \cdot t) \cdot \begin{pmatrix} c_1' \\ c_2' \end{pmatrix} = S \cdot \begin{pmatrix} e^{\lambda_1 t} & 0 \\ 0 & e^{\lambda_2 t} \end{pmatrix} \cdot S^{-1} \cdot \begin{pmatrix} c_1' \\ c_2' \end{pmatrix},$$

für $c'1, c_2' \in \mathbb{R}$. Da S^{-1} regulür ist, durchläuft auch $S^{-1} \cdot \begin{pmatrix} c_1' \\ c_2' \end{pmatrix}$ alle Elemente des $\mathbb{R}^2$. Wir setzen daher $\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = S^{-1} \cdot \begin{pmatrix} c_1' \\ c_2' \end{pmatrix}$ und erhalten, dass alle Lösungen des DGL-Systems gegeben sind durch

$$\begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = S \cdot \begin{pmatrix} e^{\lambda_1 t} & 0 \\ 0 & e^{\lambda_2 t} \end{pmatrix} \cdot \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = c_1 \cdot v_1 \cdot e^{\lambda_1} + c_2 \cdot v_2 \cdot e^{\lambda_2 t}.$$

Hier haben wir nur noch ausgenutzt, dass die beiden Spaltenvektoren von S gegeben sind durch v_1 und v_2. □

Beispiel 4.6.7. Wir erinnern uns noch einmal an Differenzengleichungen der Form $a_{n+2} = aa_{n+1} + ba_n$. Um eine geschlossene Formel für diese zu finden berechnen wir die Lösungen der charakteristischen Gleichung $x^2 - ax - b = 0$. Im konkreten Fall $a_{n+2} = -a_{n+1} + 2a_n$ erhalten wir die Gleichung $x^2 + x - 2 = 0$, mit den Lösungen $x = 1$ und $x = -2$. eine geschlossene Formel der Differenzengleichung ist damit $a_n = c_1 \cdot 1^n + c_2 \cdot (-2)^n$, für $c_1, c_2 \in \mathbb{R}$ (siehe Theorem 2.3.5).
Wir betrachten nun die Differenzialgleichung $x''(t) = -x'(t) + 2x(t)$. Beachten Sie dabei die Ähnlichkeit zur Differenzengleichung von oben! Um diese zu lösen setzen wir $y(t) = x'(t)$ und erhalten das DGL-System

$$\begin{pmatrix} x'(t) \\ y'(t) \end{pmatrix} = \begin{pmatrix} y(t) \\ -x'(t) + 2x(t) \end{pmatrix} = \begin{pmatrix} y(t) \\ 2x(t) - y(t) \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 2 & -1 \end{pmatrix} \cdot \begin{pmatrix} x(t) \\ y(t) \end{pmatrix}.$$

Dieses System lösen wir, wie eben beschrieben. Die Eigenwerte erhalten wir aus

$$0 = \det \begin{pmatrix} -\lambda & 1 \\ 2 & -1-\lambda \end{pmatrix} = \lambda^2 + \lambda - 2.$$

Das ist exakt die charakteristische Gleichung der zugehörigen Differenzengleichung! Die Nullstellen – und somit die Eigenwerte – sind daher $\lambda_1 = 1$ und $\lambda_2 = -2$. Die Eigenvektoren sind die Lösungen der Gleichungssysteme

$$\left(\begin{array}{cc|c} -1 & 1 & 0 \\ 2 & -1-1 & 0 \end{array}\right) \quad \text{und} \quad \left(\begin{array}{cc|c} -(-2) & 1 & 0 \\ 2 & -1-(-2) & 0 \end{array}\right).$$

Eigenvektoren sind somit $v_1 = \begin{pmatrix} 1 & 1 \end{pmatrix}^T$ (zum Eigenwert $\lambda_1 = 1$) und $v_2 = \begin{pmatrix} 1 & -2 \end{pmatrix}^T$ (zum Eigenwert $\lambda_2 = -2$). Damit erhalten wir alle Lösungen des Systems durch die Funktionen

$$\begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = c_1 \cdot e^{1\cdot t} \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} + c_2 \cdot e^{-2t} \cdot \begin{pmatrix} 1 \\ -2 \end{pmatrix}, \quad c_1, c_2 \in \mathbb{R}.$$

Insbesondere folgt $x(t) = c_1 e^{1\cdot t} + c_2 e^{-2\cdot t}$, was das stetige Pendant der Lösung der zugehörigen Differenzengleichung ist.

Jetzt fehlt noch der Fall, der eintritt, wenn A nicht diagonalisierbar ist. Das ist der Fall, wenn es nur einen reellen Eigenwert λ gibt, zu dem es aber nur einen Eigenvektor gibt. Um die Jordanform von A zu bekommen, brauchen wir noch einen Hauptvektor von A zu λ. Das ist ein Vektor $w \in \mathbb{R}^2$, mit $(A - \lambda E_2)^2 \cdot w = 0$, der kein Eigenvektor ist.

Satz 4.6.8. *Sei A eine reelle 2×2-Matrix, die nicht diagonalisierbar ist. Sei λ der Eigenwert von A und seien v und w ein Eigenvektor und ein Hauptvektor. Dann sind alle Lösungen von*

$$\begin{pmatrix} x'(t) \\ y'(t) \end{pmatrix} = A \cdot \begin{pmatrix} x(t) \\ y(t) \end{pmatrix}$$

gegeben durch

$$\begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = c_1 \cdot e^{\lambda\cdot t} \cdot v + c_2 \cdot e^{\lambda\cdot t} \cdot (v \cdot t + w), \quad c_1, c_2 \in \mathbb{R}.$$

BEWEIS. Die Matrix S, deren Spalten gegeben sind durch v und w, ist regulär und es gilt

$$S^{-1} \cdot A \cdot S = \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}.$$

Jetzt brauchen wir nur noch $\exp\left(\begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}\right) = \begin{pmatrix} e^{\lambda} & t \cdot e^{\lambda} \\ 0 & e^{\lambda} \end{pmatrix}$ und der Rest des Beweises folgt exakt dem Beweis von Satz 4.6.6. □

Bemerkung 4.6.9. Es sieht so aus als hätten wir damit alle Lösungen gefunden, da wir die Lösungen beschreiben mit der Unterteilung ob A diagonalisierbar ist oder nicht. Allerdings haben sich im diagonalisierbaren Fall auch komplexe Lösungen eingeschlossen, da die Eigenwerte – und damit auch die Eigenvektoren – komplexwertig sein können. Sei aber $g(t) = \begin{pmatrix} g_1(t) \\ g_2(t) \end{pmatrix}$ eine Funktion $\mathbb{C} \longrightarrow \mathbb{C}^2$, die die Bedingung $\nabla g(t) = A \cdot g(t)$, für eine reelle 2×2-Matrix A, erfüllt. Dann besteht $g(t)$ aus einem Real- und einem Imaginärteil. Beide beschreiben reelle Funktionen $t \mapsto \mathrm{Re}(g(t))$ und $t \mapsto \mathrm{Im}(g(t))$. Diese sind differenzierbar und es gilt

$$\begin{aligned} \nabla \mathrm{Re}(g(t)) + i \cdot \nabla \mathrm{Im}(g(t)) &= \nabla(\mathrm{Re}(g(t)) + i \cdot \mathrm{Im}(g(t))) = \nabla g(t) \\ &= A \cdot g(t) = A \cdot (\mathrm{Re}(g(t)) + i \cdot \mathrm{Im}(g(t)) \\ &= A \cdot \mathrm{Re}(g(t)) + i \cdot A \cdot \mathrm{Im}(g(t)). \end{aligned}$$

Da A reell ist, folgt daraus,

$$\nabla \mathrm{Re}(g(t)) = A \cdot \mathrm{Re}(g(t)) \quad \text{und} \quad \nabla \mathrm{Im}(g(t)) = A \cdot \mathrm{Im}(g(t)).$$

Damit sind die Funktionen $\mathrm{Re}(g(t))$ und $\mathrm{Im}(g(t))$ zwei reelle Lösungen des DGL-Systems. Alle reellen Lösungen sind daher gegeben durch $c_1 \cdot \mathrm{Re}(g(t)) + c_2 \cdot \mathrm{Im}(g(t))$, mit $c_1, c_2 \in \mathbb{R}$, sofern $\mathrm{Re}(g(t)) \neq c \cdot \mathrm{Im}(g(t))$.
Besitzt nun A einen Eigenwert $\lambda \in \mathbb{C} \setminus \mathbb{R}$ so ist der zweite Eigenwert $\bar{\lambda}$. Ist v ein Eigenvektor zu λ, so ist $\bar{v}$ ein Eigenvektor zu $\bar{\lambda}$. Da A zwei verschiedene Eigenwerte besitzt, können wir Satz 4.6.6 anwenden und erhalten, mit $c_1 = 1$ und $c_2 = 0$, dass $g(t) = e^{\lambda t} v$ eine komplexe Lösung des DGL-Systems ist. Wir schreiben nun $\lambda = a + b \cdot i$ und $v = v_1 + v_2 \cdot i$, mit $a, b \in \mathbb{R}$ und $v_1, v_2 \in \mathbb{R}^2$.

Dann ist

$$\begin{aligned} g(t) = e^{\lambda t}v = e^{at+bti} \cdot (v_1 + i \cdot v_2) &= e^{at} \cdot e^{bti} \cdot (v_1 + i \cdot v_2) \\ &\overset{3.2.6}{=} e^{at} \cdot (\cos(bt) + i \cdot \sin(bt)) \cdot (v_1 + i \cdot v_2) \\ &= e^{at} \cdot ((\cos(bt)v_1 - \sin(bt)v_2) + i \cdot (\cos(bt)v_2 + \sin(bt)v_1)) \end{aligned}$$

Für reelles t ist somit

$$\mathrm{Re}(e^{\lambda t}v) = e^{at} \cdot (\cos(bt)v_1 - \sin(bt)v_2) \quad \text{und} \quad \mathrm{Im}(e^{\lambda t}v) = e^{at} \cdot (\cos(bt)v_2 + \sin(bt)v_1).$$

Es ist leicht eingesehen, dass $\mathrm{Re}(e^{\lambda t}v)$ kein Vielfaches von $\mathrm{Im}(e^{\lambda t}v)$ ist. Mit unseren Vorüberlegungen können wir nun endlich schließen, dass im Betrachteten Fall alle *reellen* Lösungen gegeben sind durch

$$\begin{aligned} \begin{pmatrix} x(t) \\ y(t) \end{pmatrix} &= c_1 \cdot \mathrm{Re}(e^{\lambda t} \cdot v + c_2 \cdot \mathrm{Im}(e^{\lambda t} \cdot v \\ &= c_1 e^{at}(\cos(bt) \cdot v_1 - \sin(bt) \cdot v_2) + c_2 e^{at}(\cos(bt) \cdot v_2 + \sin(bt) \cdot v_1) \\ &= e^{at} \cdot (c_1 \cdot (\cos(bt) \cdot v_1 - \sin(bt) \cdot v_2) + c_2 \cdot (\cos(bt) \cdot v_2 + \sin(bt) \cdot v_1)) \\ &= e^{at} \cdot (\cos(bt) \cdot (c_1 v_1 + c_2 v_2) + \sin(bt) \cdot (c_2 v_1 - c_1 v_2)) \end{aligned}$$

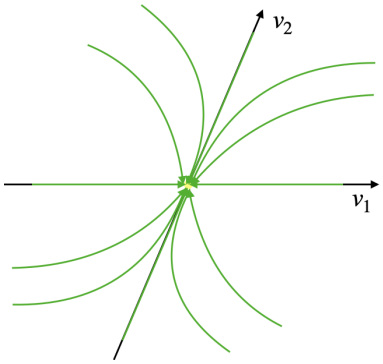
Bemerkung 4.6.10. Wir fassen alle möglichen Trajektorien des DGL-Systems (4.9), mit einer reellen 2×2-Matrix A zusammen. Seien λ_1 und λ_2 Eigenwerte von A. Wir nehmen weiter an, dass A regulär ist, woraus sofort folgt, dass $(0, 0)$ der einzige stationäre Punkt von (4.9) ist und 0 kein Eigenwert von A ist.

1. Fall: A ist diagonalisierbar, $\lambda_1, \lambda_2 \in \mathbb{R}$, mit linear unabhängigen Eigenvektoren v_1, v_2. Dann sind die Lösungen von (4.9):

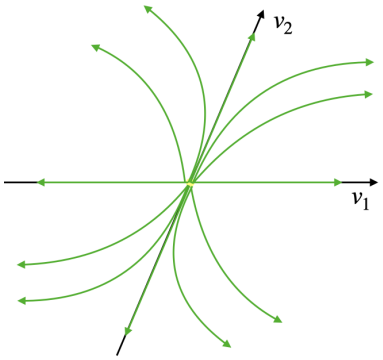
$$\begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = c_1 \cdot e^{\lambda_1 \cdot t} \cdot v_1 + c_2 \cdot e^{\lambda_2 \cdot t} \cdot v_2, \quad c_1, c_2 \in \mathbb{R}.$$

Je nach dem welches Vorzeichen λ_1, λ_2 besitzen erhalten wir ein anderes Bild der Trajektorien:

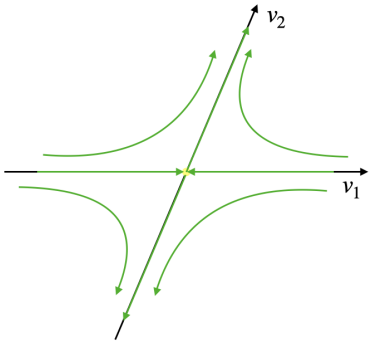
$\boxed{\lambda_1, \lambda_2 < 0}$ Dann strebt jede Lösung gegen den stationären Punkt $(0,0)$. Das nennen wir einen *stabilen Knoten*.



$\boxed{\lambda_1, \lambda_2 > 0}$ Dann strebt jede Lösung weg vom stationären Punkt $(0,0)$. Das nennen wir einen *instabilen Knoten*.



$\boxed{\lambda_1 < 0 < \lambda_2}$ Dann strebt jede Lösung in v_1-Richtung zum stationären Punkt $(0,0)$ hin und in v_2-Richtung vom stationären Punkt $(0,0)$ weg. Das nennen wir einen *Sattelpunkt*.

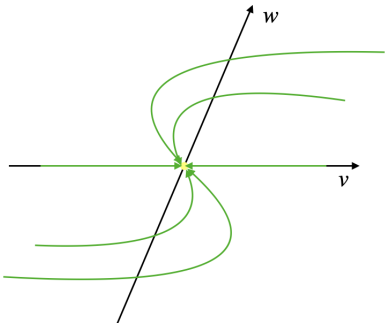


2. Fall: A ist nicht diagonalisierbar. Das ist der Fall, wenn $\lambda = \lambda_1 = \lambda_2 \in \mathbb{R}$ ist und es keine zwei linear unabhängige Eigenvektoren zu λ gibt. Sei v ein Eigenvektor und w ein Hauptvektor zu λ. Dann sind alle Lösungen von (4.9)

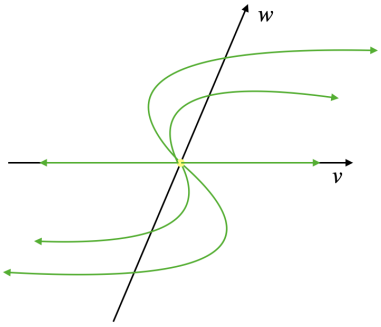
$$\begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = c_1 \cdot e^{\lambda \cdot t} \cdot v + c_2 \cdot e^{\lambda \cdot t} \cdot (v \cdot t + w), \quad c_1, c_2 \in \mathbb{R}.$$

Wieder ist das Vorzeichen von λ relevant.

$\boxed{\lambda < 0}$ Dann strebt jede Lösung gegen den stationären Punkt $(0,0)$. Das nennen wir wieder einen *stabilen Knoten.*



$\boxed{\lambda > 0}$ Dann strebt jede Lösung weg vom stationären Punkt $(0,0)$. Das nennen wir wieder einen *instabilen Knoten.*

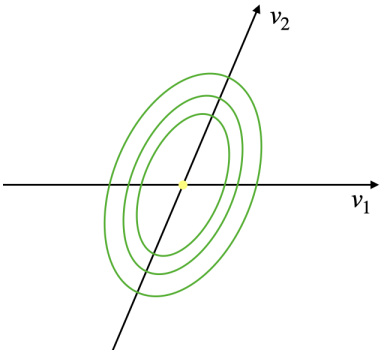


3. Fall: A besitzt einen komplexen Eigenwert $\lambda \in \mathbb{C} \setminus \mathbb{R}$ – sagen wir $\lambda = a + b \cdot i$, mit $a, b \in \mathbb{R}$. Sei $v \in \mathbb{C}^3$ ein Eigenvektor zu λ. Dann schreiben wir $v = v_1 + v_2 \cdot i$, mit $v_1, v_2 \in \mathbb{R}^3$. Dann sind alle Lösungen von (4.9) gegeben durch

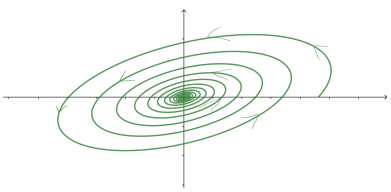
$$\begin{pmatrix} x(t) \\ y(t) \end{pmatrix} e^{at} \cdot (\cos(bt) \cdot (c_1 v_1 + c_2 v_2) + \sin(bt) \cdot (c_2 v_1 - c_1 v_2)), \quad c_1, c_2 \in \mathbb{R}.$$

Diesmal ist das Vorzeichen von a (dem Realteil von λ entscheidend.

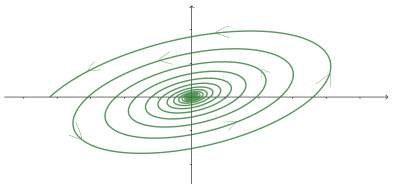
$\boxed{a = 0}$ Dann sind die Lösungen periodisch und wir erhalten *geschlossene Kurven* um den stationären Punkt $(0,0)$.



$\boxed{a < 0}$ Dann drehen sich die Kurven unendlich oft um den stationären Punkt $(0,0)$ und nähern ihn dabei immer weiter an. Dies nennen wir einen *stabilen Strudel.*



$\boxed{a > 0}$ Dann drehen sich die Kurven unendlich oft um den stationären Punkt $(0,0)$ und entfernt sich dabei immer weiter von ihm. Dies nennen wir einen *instabilen Strudel*.



Wir erinnern noch kurz an die Spur $\mathrm{tr}(A)$ einer Matrix A. Diese ist gegeben durch die Summe der Einträge auf der Diagonalen von A.

Lemma 4.6.11. *Sei A eine reelle 2×2-Matrix. Dann sind die Eigenwerte von A gegeben durch*

$$\lambda_{1/2} = \frac{\mathrm{tr}(A)}{2} \pm \sqrt{\left(\frac{\mathrm{tr}(A)}{2}\right)^2 - \det(A)}.$$

Beweis. Schreibe $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. Dann ist das charakteristische Polynom von A gegeben durch

$$\det \begin{pmatrix} a-\lambda & b \\ c & d-\lambda \end{pmatrix} = (a-\lambda)(d-\lambda) - bc = \lambda^2 - \mathrm{tr}(A)\lambda + \det(A).$$

Die Nullstellen von diesem Polynom sind die Eigenwerte, was das Lemma bereits beweist. □

Satz 4.6.12. *Gegeben sei ein homogenes lineares DGL-System* $\begin{pmatrix} x'(t) \\ y'(t) \end{pmatrix} = A \cdot \begin{pmatrix} x(t) \\ y(t) \end{pmatrix}$, *mit einer regulären reellen 2×2-Matrix A. Dieses System hat genau einen stationären Punkt bei $(0,0)$ und es gilt*

(a) $\det(A) < 0 \implies$ *der stationäre Punkt ist ein Sattelpunkt.*

(b) $\det(A) > 0,\ \mathrm{tr}(A) < 0 \implies$ *der stationäre Punkt ist stabil.*

(c) $\det(A) > 0,\ \mathrm{tr}(A) > 0 \implies$ *der stationäre Punkt ist instabil.*

Beweis. Die Determinante ist immer das Produkt der beiden Eigenwerte. Ist also $\det(A) < 0$ gibt es einen positiven und einen negativen Eigenwert (insbesondere sind beide Eigenwerte reell). Es folgt aus der Fallunterscheidung aus Bemerkung 4.6.10, dass wir es mit einem Sattelpunkt zu tun haben.

Ist $0 < \det(A) \leq \left(\frac{\operatorname{tr}(A)}{2}\right)^2$, so ist $0 \leq \sqrt{\left(\frac{\operatorname{tr}(A)}{2}\right)^2 - \det(A)} < \frac{|\operatorname{tr}(A)|}{2}$. Mit Lemma 4.6.11 sind damit die Eigenwerte reell und haben das gleiche Vorzeichen, wie $\operatorname{tr}(A)$. Ist hingegen $\left(\frac{\operatorname{tr}(A)}{2}\right)^2 < \det(A)$, so sind die Eigenwerte komplex mit Realteil $\frac{\operatorname{tr}(A)}{2}$. Die Realteile der Eigenwerte haben somit das gleiche Vorzeichen wie $\operatorname{tr}(A)$.
Insgesamt haben wir festgestellt, dass im Fall $\det(A) > 0$ das Vorzeichen der Realteile der Eigenwerte immer das gleiche ist wie das Vorzeichen von $\operatorname{tr}(A)$. Mit der Fallunterscheidung aus Bemerkung 4.6.10 wissen wir, dass der stationäre Punkt instabil ist, falls die Realteile positiv sind und stabil, falls die Realteile negativ sind. Damit ist die Aussage des Satzes bewiesen. □

Können wir dieses Wissen nutzen um nicht notwendigerweise lineare DGL-Systeme zu studieren? In Maßen funktioniert das tatsächlich. Da der Satz von Grobman-Hartman aussagt, dass sich Lösungen eines beliebigen zweidimensionalen DGL-Systems in der nähe von stationären Punkten in etwa so verhält wie sich die Linearisierung des Systems an diesem Punkt verhält. Die präzise Vormulierung des Satzes ist bereits sehr anspruchsvoll. Daher begnügen wir uns hier mit einer unpräzisen Version.

Theorem 4.6.13 (Satz von Grobman-Hartman (unpräzise Version))**.** *Sei ein DGL-System gegeben durch*

$$\begin{pmatrix} x'(t) \\ y'(t) \end{pmatrix} = f(x(t), y(t)),$$

mit f stetig partiell differenzierbar. Sei (x_0, y_0) ein stationärer Punkt des DGL-Systems und sei $J_f(x_0, y_0) = f'(x_0, y_0)$ die Jacobi-Matrix von f an (x_0, y_0). Seien weiter die Realteile der Eigenwerte von $J_f(x_0, y_0)$ nicht Null (insbesondere ist Null kein Eigenwert). Dann verhalten sich die Trajektorien des DGL-Systems in der Nähe von (x_0, y_0) genau wie die Trajektorien des DGL-Systems

$$\begin{pmatrix} x'(t) \\ y'(t) \end{pmatrix} = J_f(x_0, y_0) \cdot \begin{pmatrix} x(t) \\ y(t) \end{pmatrix}$$

um den stationären Punkt $(0, 0)$.

Beispiel 4.6.14. Gegeben sei das DGL-System

$$\begin{pmatrix} x'(t) \\ y'(t) \end{pmatrix} = \begin{pmatrix} x(t) - x(t)^2 y(t) \\ -y(t) + x(t)y(t) \end{pmatrix} = f(x(t), y(t)).$$

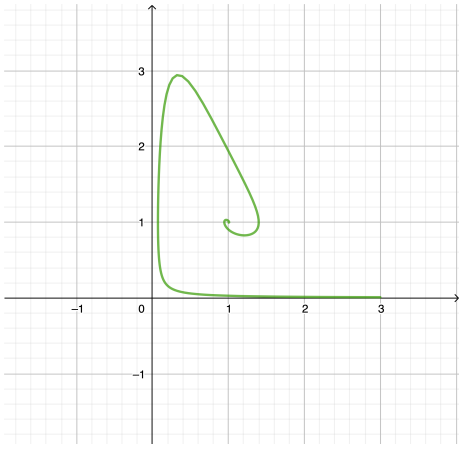
Wir möchten das Verhalten von Lösungen in der Nähe der stationären Punkte dieses Systems bestimmen. Dazu berechnen wir als erstes die stationären Punkte, durch lösen des Gleichungssystems

$$\begin{pmatrix} x - x^2 y \\ -y + xy \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Dies hat die beiden Lösungen $(0,0)$ und $(1,1)$. Das sind daher die stationären Lösungen. Um den Satz von Grobman-Hartman anzuwenden, benötigen wir die Jacobi-Matrix von f. Diese ist

$$J_f(x,y) = \begin{pmatrix} 1 - 2xy & -x^2 \\ y & -1 + x \end{pmatrix}.$$

Am stationären Punkt $(0,0)$ ausgewertet erhalten wir $J_f(0,0) = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$. Die Eigenwerte sind $1 > 0$ und $-1 < 0$. Damit handelt es sich beim stationären Punkt $(0,0)$ um einen Sattelpunkt.
Werten wir die Jacobimatrix am stationären Punkt $(1,1)$ aus, erhalten wir $J_f(1,1) = \begin{pmatrix} -1 & -1 \\ 1 & 0 \end{pmatrix}$, mit den Eigenwerten $-\frac{1}{2} \pm \frac{\sqrt{3}}{2} \cdot i$. Diese sind komplex mit negativem Realteil. Beim stationären Punkt $(1,1)$ entsteht daher ein stabiler Strudel. Die Skizze einer Lösungskurve startend am Punkt $(3, 0.1)$ passt zu diesem Resultat:

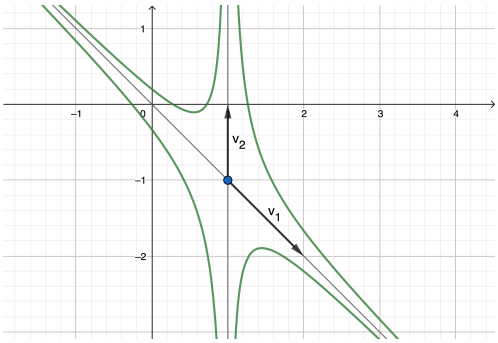


Dass die Lösungskurve am Sattelpunkt die beiden Achsen annähert liegt daran, dass die Eigenvektoren von $J_f(0,0)$ in Richtung der beiden Achsen zeigen.

Beispiel 4.6.15. Das DGL-System

$$\begin{pmatrix} x'(t) \\ y'(t) \end{pmatrix} = \begin{pmatrix} -x(t)+1 \\ y(t)+x(t)^2 \end{pmatrix} = f(x(t), y(t))$$

hat den stationären Punkt $(1,-1)$ und es gilt $J_f(1,-1) = \begin{pmatrix} -1 & 0 \\ 2 & 1 \end{pmatrix}$. Die Eigenverte sind $\lambda_1 = -1$ und $\lambda_2 = 1$, woraus wir folgern, dass es sich um einen Sattelpunkt handelt. Die zugehörigen Eigenvektoren sind $v_1 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$ und $v_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$. Daher sehen die Lösungen in der Nähe des stationären Punktes etwa so aus:



4.7 Bäuber-Beute-Modell mit beschränkten Ressourcen

Ohne eine Räuber-Population führt das Lotka-Volterra-Modell zu einem unbeschränkten Wachstum der Beutepopulation. Das ist auf lange Sicht unrealistisch. Die einfachste Idee dies zu verhindern, ist ein Räuber-Beute-Modell, dass das logistische Wachstum bei der Beutepopulation zugrunde legt. Beschreibt also $x(t)$ die Größe der Beutepopulation zum Zeitpunkt t und $y(t)$ die Größe der Räuberpopulation zum Zeitpunkt t, dann erhalten wir das Modell

$$\begin{aligned} x'(t) &= sx(t)(K - x(t)) - ax(t)y(t) \\ y'(t) &= -ry(t) + bx(t)y(t) \end{aligned} \tag{4.10}$$

wobei wieder s, r, a, b, K positive reelle Zahlen beschreiben. Wir sehen sofort, dass die Beute ohne Räuber logistisch wächst und die Räuber ohne Beute aussterben.
Dieses Modell wurde in den Übungen bereits behandelt. Dort wurde es entdimensionalisiert zu

$$\begin{aligned} X'(\tau) &= X(\tau)(1-\gamma X(\tau)) - X(\tau)Y(\tau) \\ Y'(\tau) &= \delta Y(\tau)(X(\tau)-1), \end{aligned} \tag{4.11}$$

mit $\gamma = \frac{r}{bK}$ und $\delta = \frac{r}{sK}$. Die stationären Punkte dieses Modells sind

$$P_1 = (0,0) \quad , \quad P(\gamma^{-1},0) \quad , \quad P_3 = (1, 1-\gamma).$$

Wir kennen die Lösungen $X(\tau) = 0$ und $Y(\tau) = c \cdot e^{-\delta}$, $c \in \mathbb{R}$, und $Y(\tau) = 0$ und $X(\tau) = \frac{\gamma^{-1}}{e^{-\tau} \cdot c + 1}$, $c \in \mathbb{R}$, (siehe Proposition 2.8.3). Da sich Lösungskurven nicht schneiden können, gilt damit wieder, dass Lösungen im ersten Quadranten auch im ersten Quadranten bleiben. Das passt zu unserem Modell, da es keine negativen Populationsgrößen gibt. Wir möchten also nur Lösungen im ersten Quadranten betrachten.
Mit Blick auf den stationären Punkt P_3 fällt dabei eine nötige Fallunterscheidung ins Auge, da der Punkt P_3 genau dann im ersten Quadranten liegt, wenn $\gamma < 1$ ist. Nur in diesem Fall ist also eine konstante Koexistenz beider Spezies möglich. Lösen von $X' = 0$ liefert, dass Lösungen unterhalb der Geraden $Y = 1 - \gamma X$ in X-Richtung steigen und oberhalb dieser Geraden in X-Richtung fallen.
Bei der Y-Richtung kennen wir die Unterteilung bereits vom Lotka-Volterra-Modell: Die Lösungen links von $X = 1$ sind in Y-Richtung fallend, die Lösungen rechts von $X = 1$ sind in Y-Richtung monoton steigend. Da P_3 genau der Schnittpunkt der Geraden $X = 1$ und $Y = 1 - \gamma X$ ist, wird der erste Quadrant in vier „Monotoniebereiche“ unterteilt, wenn $\gamma < 1$ und in drei, wenn $\gamma \geq 1$.
Um mehr Informationen über den Verlauf der Lösungskurven zu bekommen, möchten wir den Satz von Grobman-Hartman benutzen. Im Fall von (4.11) ist

$$f(X,Y) = \begin{pmatrix} X(1-\gamma X) - XY \\ \delta Y(X-1) \end{pmatrix}$$

und somit

$$J_f(X,Y) = \begin{pmatrix} 1-2\gamma X - Y & -X \\ \delta Y & \delta(X-1) \end{pmatrix}.$$

Insbesondere ist

$$J_f(P_1) = \begin{pmatrix} 1 & 0 \\ 0 & -\delta \end{pmatrix}.$$

Diese Matrix hat die Eigenwerte $1 > 0$ und $-\delta < 0$. Damit ist $\boxed{P_1 \text{ ein Sattelpunkt}}$.
Für P_2 erhalten wir

$$J_f(P_2) = \begin{pmatrix} -1 & -\gamma^{-1} \\ 0 & \delta(\gamma^{-1}-1) \end{pmatrix}.$$

Die Eigenwerte sind -1 und $\delta(\gamma^{-1}-1)$. Es ist also $\boxed{P_2 \text{ ein Sattelpunkt, wenn } \gamma < 1}$ und $\boxed{P_2 \text{ ist stabil, wenn } \gamma > 1}$.
Zuletzt sehen wir

$$J_f(P_3) = \begin{pmatrix} -\gamma & -1 \\ \delta(1-\gamma) & 0 \end{pmatrix}.$$

Diese Matrix hat die Determinante $\delta(1-\gamma)$ und die Spur $-\gamma < 0$. Der Punkt ist somit stabil oder ein Sattelpunk, je nachdem welches Vorzeichen die Determinante besitzt. Genauer ist $\boxed{P_3 \text{ ein Sattelpunkt, wenn } \gamma > 1}$ und $\boxed{P_3 \text{ ist stabil, wenn } \gamma < 1}$.
Da wir in diesem Abschnitt bisher ohne jede Struktur (Satz, Definition, Bemerkung, ...) gearbeitet haben, halten wir kurz inne um zu sortieren, was wir bisher herausgefunden haben.

Fall $\gamma > 1$:	**Fall $\gamma < 1$:**
Es gibt nur zwei mögliche (nicht negative) stationäre Punkte $P_1 = (0,0)$ und $P_2 = (\gamma^{-1}, 0)$.	Es gibt drei mögliche (nicht negative) stationäre Punkte $P_1 = (0,0)$, $P_2 = (\gamma^{-1},0)$ und $P_3 = (1, 1-\gamma)$. Dabei beschreibt P_3 eine Koexistenz beider Spezies.
P_1 ist ein Sattelpunkt.	P_1 ist ein Sattelpunkt.
P_2 ist stabil.	P_2 ist ein Sattelpunkt.
	P_3 ist stabil.

Vermutung 1: Im Fall $\gamma > 1$ scheint also jede Lösung gegen den stabilen stationären Punkt P_2 zu streben. Das bedeutet, dass in diesem Fall (und in unserem gewählten Modell) die Räuberpopulation immer ausstirbt.

Vermutung 2: Im Fall $\gamma < 1$ scheint es eine Lösung zu geben, die vom Sattelpunkt P_2 (möglicherweise mit einigen Umrundungen) zum stabilen Punkt P_3 führt. Dann gilt, dass jede Lösung, die einmal einen Punkt (x_0, y_0) mit $x_0 \in (0, 1)$ durchläuft, zwangsläufig gegen P_3 konvergiert. Wir vermuten also in diesem Fall, dass sich bei vorhandener Räuber- und Beutepopulation immer die Koexistenz im stationären Fall P_3 einstellen wird.

Bemerkung 4.7.1. Genau wie im eindimensionalen Fall gilt auch bei autonomen DGL-Systemen, dass Grenzwerte von Lösungen (wenn sie im eigentlichen Sinn existieren) immer stationäre Punkte sind. Das gilt zumindest, wenn die definierende Funktion f schön genug – etwa lokal Lipschitz-stetig – ist. Die präzise Aussage habe ich in Lemma 4.5.6 formuliert, da es dort rückblickendbetrachtet besser passt.
Für uns bedeutet das, dass Funktionen nur an stationären Punkten enden oder starten können.

Damit ist Vermutung 1 bewiesen: Die Funktion $X(\tau)$ nimmt irgendwann einen Wert kleiner als 1 an, denn bereits ohne den Term $-X(\tau)Y(\tau)$ wird die Kapazitätsgrenze $K = \gamma^{-1} < 1$ angenähert. Aufgrund der Monotonie kann die Trajektorie den Bereich $x < 1$ nicht mehr verlassen und ist streng monoton fallend. Mit Lemma 4.5.6 folgt nun, dass die Trajektorie gegen den stationären Punkt P_2 konvergieren muss.
Im Fall $\gamma < 1$ wollen wir auch zeigen, dass jede Trajektorie irgendwann in den Bereich $x < 1$ läuft. Diese können den Bereich zwar auch wieder verlassen, aber darum kümmern wir uns später.

Satz 4.7.2. *Im Fall $\gamma < 1$: Sei $(X(\tau), Y(\tau))$ eine maximale Lösung von (4.11) im ersten Quadranten. Dann nimmt $X(\tau)$ einen Wert < 1 an.*

Beweis. Wir nehmen eine Lösung mit Anfangswert $X(0) > 1$ und $Y(0) > 0$. Wenn Y beschränkt ist, sind wir fertig, da dies nur möglich ist, wenn die Lösungskurve in y-Richtung wieder fällt, was nur im Bereich $x < 1$ passieren kann. Wenn die Lösung unbeschränkt ist, können wir die Autonomie nutzen um $Y(0) \geq 2$ anzunehmen.
Weiter fassen wir Y als Funktion in X (bzw $X(\tau)$) auf. Dann ist diese Funktion Y auf dem Intervall $(1, X(0))$ streng monoton fallend. Wenn die Trajektorie nicht in den Bereich $x < 1$ übergeht, kann die Steigung von

Y nicht beschränkt sein. Das werden wir nutzen um einen Widerspruch zu erzeugen. Dazu berechnen wir die Ableitung von $Y(X)$. Bevor wir das können, müssen wir uns überlegen, wie wir Y als Funktion in X auffassen können. Dazu nehmen wir die Umkehrabbildung $X^{-1} : X(0, t_+) \to (0, t_+)$ von X, wobei t_+ die maximale obere Grenze des Definitionsintervalles von X ist.Jetzt haben wir jedem $X(\tau)$ den zugehörigen Wert τ zugeordnet. Zu diesem τ suchen wir nun den passenden Y-Wert. Damit ist $Y(X(\tau)) = Y(X^{-1}(X(\tau)))$, mit der Variablen $X(\tau)$. Es folgt

$$\begin{aligned} Y'(X) = Y'(X(\tau)) = (Y(X^{-1}(X(\tau))))' &= (X^{-1}(X(\tau))' \cdot Y'(X^{-1}(X(\tau)) \\ &= \frac{\delta Y(X^{-1}(X(\tau)))(X(\tau) - 1)}{X(\tau)(1 - \gamma X(\tau) - Y(X^{-1}(X(\tau))))} \\ &= \frac{\delta Y(X)(X-1)}{X(1 - \gamma X - Y(X))}. \end{aligned}$$

Per Annahme ist stets $X > 1$ und $Y(X) \geq 2$. Damit ist

$$\begin{aligned} |Y'(X)| = \frac{\delta Y(X)(X-1)}{X(-1 + \gamma X + Y(X))} &\leq \frac{\delta Y(X)X}{-1 + \gamma + Y(X)} \\ &\leq \frac{\delta Y(X)X(0)}{-1 + \gamma + Y(X)} = \delta X(0) \cdot \frac{1}{1 - \frac{1-\gamma}{Y(X)}} \\ &\leq \delta X(0) \cdot \frac{1}{1 - \frac{1-\gamma}{2}}. \end{aligned}$$

Damit ist die Ableitung von Y (aufgefasst als Funktion in X) beschränkt, was – wie oben erklärt – einen Widerspruch dazu darstellt, dass Die Lösungskurve im Bereich $x > 1$ verbleibt. Damit ist der Satz bewiesen. □

Genau so zeigt man auch, dass es keine Lösungen im ersten Quadranten gibt, die in negativer Zeitrichtung im Bereich $x < 1$ verbleiben.
Wir folgern, dass es eine Trajektorie gibt, die vom Sattelpunkt P_2 zum stabilen Strudel P_3 führt. Da alle Lösungen irgendwann in den Bereich $x < 1$ übergehen, sind sie ab dann durch die Verbindung von P_2 zu P_3 beschränkt. Damit münden alle Lösungen im stabilen Strudel P_3, was bedeutet, dass immer der konstante Koexistenzzustand beider Spezies angenähert wird. Das beweise Vermutung 2.

Bemerkung 4.7.3. Das hier beschriebene Verhalten passt nicht zu den Daten der Hudson-Bay-Company, das wir uns zum klassischen Lotka-Volterra-

Modell betrachtet haben. Dort waren die periodischen Schwankungen zu beobachten, die es in dem Modell mit beschränkten Ressourcen nicht gibt. Das logistische Wachstum wird insbesondere dann zugrunde gelegt, wenn wir Wachstum unter Laborbedingungen betrachten. In einem Räuber-Beute-Modell mit Pantoffeltierchen im Labor, wurden einfach alle Beutetiere aufgefressen, so dass danach auch die Räuberpopulation starb. Eine Erklärung dafür könnten fehlende Verstecke für die Beutetiere sein. Ein passenderes Modell für Spezies in der Natur sollte also auch einen Schwellen- und einen Sättigungseffekt berücksichtigen. Ein solches Modell wird unter anderem im Buch von Bauer studiert.

Kapitel 5

Kryptographie

Kryptographie (altgriechisch etwa: verborgen schreiben) ist die Theorie von verschlüsselter Kommunikation. Diese hat eine sehr lange Tradition. Eine große Herausforderung wurde in den 1970er Jahren gemeistert: Geheime Datenübertragung zwischen zwei Parteien die sich noch nie getroffen haben und deren gesamte Kommunikation mitgelesen wird. Dies scheint auf den ersten Blick vollkommen unmöglich, wir werden aber ein Verfahren kennenlernen, welches genau das leistet.

5.1 Erste Beispiele

Am sichersten ist die Kommunikation, wenn ausschließlich Sender und Empfänger wissen, dass überhapt eine Kommunikation stattfindet. Das kann zB erreicht werden, wenn die Nachrichten versteckt werden (in Bildern, mit Zitronensaft schreiben, ...). Dann spricht man von *Steganographie*. Bei der Steganographie wird also der Kanal, über den kommuniziert wird, versteckt in der Kryptographie wird nur die Nachricht (und nicht der Kanal) verschlüsselt. Damit wollen wir uns hier beschäftigen.

Setting 5.1.1. *Im folgenden haben wir immer das gleiche Setting: Person 1 (Mia) möchte eine geheime Nachricht an Person 2 (Ben) schicken und Person 3 (Kim) möchte diese Nachricht mitlesen.*
Die Nachrichten sollen also vor dem Versenden verschlüsselt *werden. Die eigentliche Nachricht nennen wir* Klartext *und die Verschlüsselung überführt den Klartext in eine* Chiffre*. Dabei soll natürlich folgendes erfüllt sein:*

- *Jede Chiffre muss eindeutig einem Klartext zugeortnet werden können.*
- *Ben soll aus der Chiffre sehr einfach den zugehörigen Klartext auslesen können.*
- *Kim soll den Klartext höchstens unter enormem Zeitaufwand generiren können.*

Definition 5.1.2. Ein *symmetrisches Kryptosystem* ist gegeben durch ein Tupel (M, C, K, E, D), wobei M, C und K nicht-leere endliche Mengen sind. Weiter sind $E = \{e_k : M \to C | k \in K\}$ und $D = \{d_k : C \to M | k \in K\}$ Mengen von Abbildungen für die gilt

$$d_k(e_k(m)) = m \quad \forall m \in M \; \forall k \in K.$$

Wir nennen M die Menge der Klartexte, C die Menge der Chiffren und K die Menge der Schlüssel.

Beispiel 5.1.3. Die berühmte *Ceasar-Verschlüsselung* besteht aus einer Verschiebung des Alphabets um eine vorgegebene Anzahl von Stellen. Wenn wir etwa im Klartext MATHEISTTOLL jeden Buchstaben durch den ersetzen, der vier Positionen weiter hinten im Alphabet vorkommt erhalten wir die Chiffre RFYMJXYYTQQ. Dabei fangen wir nach dem Z wieder bei A an.
Das passende Kryptosystem identifiziert jeden Buchstaben als eigene Nachricht. Dann ist

$$M = C = \{\text{A,B,C,D,}\ldots\text{,Y,Z}\}$$

und $K = \{1, 2, 3, 4, \ldots, 26\}$. Die Ver- und Entschlüsselung besteht nun aus

$$e_k(m) = \text{Buchstabe } k \text{ nach } m \text{ im ABC}$$
$$d_k(m) = \text{Buchstabe } k \text{ vor } m \text{ im ABC.}$$

Bemerkung 5.1.4. Da wir nach dem Z wieder bei A anfangen, haben wir das ABC an den Enden zusammengeklebt. Das kennen Sie bereits von gewissen mathematischen Objekten: Den Restklassenringen! (Sie nennen es wahrscheinlich das „modulo Rechnen“.) An diese Ringe wollen wir kurz erinnern.

Für jedes $n \in \mathbb{N}$ definieren wir $\mathbb{Z}/n\mathbb{Z} = \{[\alpha]_n | \alpha \in \mathbb{Z}\}$, wobei $[\alpha]_n = \{\alpha + qn | q \in \mathbb{Z}\}$ eine Restklasse von n ist. Für $\alpha, \beta \in \mathbb{Z}$ gilt

$$\begin{aligned}[\alpha]_n = [\beta]_n &\iff n \mid \alpha - \beta \\ &\iff \alpha = \beta + qn \text{ für ein } q \in \mathbb{Z} \\ &\iff \alpha \equiv \beta \mod n.\end{aligned}$$

Es folgt, dass $\mathbb{Z}/n\mathbb{Z}$ aus genau n Elementen besteht. Auf $\mathbb{Z}/n\mathbb{Z}$ ist durch die folgenden Verknüpfungen eine Ringstruktur definiert:

$$[\alpha]_n + [\beta]_n = [\alpha + \beta]_n \quad \text{und} \quad [\alpha]_n \cdot [\beta]_n = [\alpha \cdot \beta]_n.$$

Substituieren wir nun $A \hat{=} [1]_{26}, \ldots, Z \hat{=} [26]_{26} = [0]_{26}$ so erhalten wir die ganz einfache Gestalt der Caesar-Verschlüsselung: $M = C = K = \mathbb{Z}/26\mathbb{Z}$, $e_k(m) = m + k$ und $d_k(c) = c - k$.

Definition 5.1.5. Verschlüsselungen bei denen jedes Symbol der Nachricht als eigener Klartext aufgefasst wird, nennen wir *monoalphabetische Substitutionen*. D.h. Jeder Buchstabe wird (nach der Schlüsselwahl) immer gleich verschlüsselt.

Die Caesar-Verschlüsselung ist eine monoalphabetische Substitution. Wie man solche monoalphabetischen Substitutionen knackt erklärt uns Edgar Allen Poe in seiner Geschichte „Der Goldkäfer“.

Bemerkung 5.1.6. Wir fassen die Lösungsstrategie für monoalphabetische Substitutionen zusammen:

- Haben wir eine Chiffre vorliegen, müssen wir als erstes erkennen, dass es sich um eine Chiffre handelt. Zweitens müssen wir uns überlegen wie die Nachricht verschlüsselt wurde. Drittens müssen wir erraten in welcher Sprache der Klartext verfasst ist.

- Gibt es Leerzeichen? Wenn ja, analysiere die kürzesten Wörter. Häufige Wörter mit 3 Buchstaben sind *der*, *die*, *ein*. Weiter liefern oft übliche Strukturen, wie Anrede oder Grußformel, starke Hinweise auf die Lösung.

- Vergleiche die Häufigkeiten der Symbole mit den Häufigkeiten mit denen die Buchstaben in der gewählten Sprache vorkommen. Für Deutsch sind die 8 häufigsten Buchstaben mit prozentualem Vorkommen

e	n	i	s	r	a	t	d
17,4%	9,78%	7,55%	7,27%	7%	6,51%	6,15%	5,08%

- Ersetzte das vorherrschende Symbol der Chiffre durch ein `e` und versuche daraus neue Erkenntnisse über den Klartext zu erlangen. Bei langen Texten (oder mehreren abgefangenen Nachrichten) wird dies fast immer die richtige Übersetzung sein.

- Suche Symbole, die oft doppelt vorkommen und vergleiche diese ebenfalls mit den vorherrschenden Doppellbuchstaben der gewählten Sprache. In Deutsch gibt es als Doppelbuchstaben mit Abstand am häufigsten *ss*, *nn* und *ll*.

- Ist der Anfang gemacht, so kann man durch Ausprobieren schnell auf den Klartext kommen.

Bemerkung 5.1.7. Bei der Caesar-Verschlüsselung ist es sogar noch viel einfacher. Wenn ein einziger Buchstabe richtig entschlüsselt wurde, kennen wir den Schlüssel und können damit die gesamte Nachricht entschlüsseln. Es ist also egal wie man Sicherheit für ein Kryptosystem definiert: Die Caesar-Verschlüsselung ist sehr unsicher! Im schlimmsten Fall muss man die 26 möglichen Schlüssel ausprobieren.
Dabei – wie ab jetzt immer – wenden wir den *Kerkhoff'schen Grundsatz* an: Die Sicherheit eines Kryptosystems darf nur in der Geheimhaltung des Schlüssels liegen und nicht in der Geheimhaltung des Algorithmus.
Wir müssen also immer davon ausgehen, dass Kim weiß mit welcher Methode verschlüsselt wurde.

Bemerkung 5.1.8. Die einfachste Attacke um eine Chiffre zu entschlüsseln ist die *only ciphertext*-Attacke. Hier wird nur das Wissen über die Chiffre c und – wie immer – das Kryptosystem vorausgesetzt. Es wird nun $d_k(c)$ für alle Schlüssel k berechnet und die sinnvollste Nachricht wird als Klartext ausgewählt. Die Schlüsselmenge K sollte also sehr groß sein um diese Attacke abzuwehren.

Definition 5.1.9. Bei der *Vigenere-Verschlüsselung*, werden immer r-Tupel von Buchstaben als eigene Nachricht aufgefasst. Nach Identifizierung des Alphabets mit $\mathbb{Z}/26\mathbb{Z}$ setzen wir $M = C = K = (\mathbb{Z}/26\mathbb{Z})^r$. Die Ver- und

Entschlüsselung verläuft nun wieder per Addition bzw. Subtraktion. Ist $m = ([m_1]_{26}, \ldots, [m_r]_{26}) \in M$ und $k = ([k_1]_{26}, \ldots, [k_r]_{26}) \in K$, so ist

$$c = e_k(m) = m + k = ([m_1 + k_1]_{26}, \ldots, [m_r + k_r]_{26}).$$

Beispiel 5.1.10. Zur Veranschaulichung betrachten wir wieder ein Beispiel. Anna und Ben einigen sich auf den Schlüssel KRYPT (die Schlüssellänge ist also 5). Es soll der folgende Klartext verschlüsselt werden

DER KUCKUCK UND DER ESEL DIE HATTEN EINEN STREIT
WER WOHL AM BESTEN SAENGE ZUR SCHOENEN MAIENZEIT

Nun wird der erste Buchstabe (das D) mit einem K addiert, der zweite Buchstabe mit einem R, der dritte Buchstabe mit einem Y, der vierte Buchstabe mit einem P, der fünfte Buchstabe mit einem T. Dann geht es von vorne los und der sechste Buchstabe wird wieder mit K addiert, usw. Dies lässt sich sehr übersichtlich in einer Tabelle anordnen, indem wir den Schlüssel so oft über den Klartext schreiben, wie es passt. D.h.:

```
KRY PTKRYPT KRY PTK RYPT KRY PTKRYP TKRYP TKRYPT
DER KUCKUCK UND DER ESEL DIE HATTEN EINEN STREIT

KRY PTKR YP TKRYPT KRYPTK RYP TKRYPTKR YPTKRYPTK
WER WOHL AM BESTEN SAENGE ZUR SCHOENEN MAIENZEIT
```

Jetzt muss Anna nur noch die jeweils übereinanderstehenden Buchstaben addieren und sie erhält die Chiffre c.

```
KRY PTKRYPT KRY PTK RYPT KRY PTKRYP TKRYP TKRYPT
DER KUCKUCK UND DER ESEL DIE HATTEN EINEN STREIT
OWQ AONCTSE FFC TYC WRUF OAD XUELDD YTFDD MEJDYN

KRY PTKR YP TKRYPT KRYPTK RYP TKRYPTKR YPTKRYPTK
WER WOHL AM BESTEN SAENGE ZUR SCHOENEN MAIENZEIT
HWQ MISD ZC VPKSUH DSDDAP RTH MNZNUHPF LQCPFYUCE
```

Diese Chiffre kann Ben nun genauso entschlüsseln, da er ja den Schlüssel KRYPT kennt. Aber was ist mit Kim, die diese Nachricht heimlich abgefangen hat? Es gibt 26^r verschiedene Schlüssel und r ist unbekannt. Aber selbst für recht kleines r ist das eine große Zahl. Für $r = 5$ gibt es etwa mehr als 11 Millionen Schlüssel.

Durch stures Ausprobieren der Schlüssel wird es also schwierig. Auch eine einfache Häufigkeitsanalyse der vorkommenden Buchstaben hilft hier nicht

weiter, da der gleiche Buchstabe im Klartext durch unterschiedliche Buchstaben in der Chiffre dargestellt werde kann. Der Klartext KUCKUCK besteht nur aus drei verschiedenen Buchstaben. Das Wort wird in der Chiffre aber durch lauter verschiedene Buchstaben dargestellt (AONCTSE).
Wie kann Kim nun trotzdem an den Klartext kommen?

- Sie könnte zunächst versuchen, den Schlüssel zu erraten. Da wir uns hier im Kapitel „Kryptographie“ befinden, ist der Schlüssel KRYPT tatsächlich nicht sehr originell gewählt. Das Erraten ist allerdings unmöglich, wenn der Schlüssel zufällig gewählt wird.

- Sie könnte auch testen, ob sie einzelne Wörter zuordnen kann. Fängt die Nachricht mit einer üblichen Anrede an, oder hört sie mit einer üblichen Grußformel auf? Dann hat sie einige Teile des Schlüssels schnell konstruiert und kann damit die Nachricht knacken.

- Ist auch das nicht der Fall, gibt es trotzdem einen Trick zum entschlüsseln! Dieser Trick basiert darauf, dass manche Buchstabenpaare gehäuft vorkommen. Z.B.: gibt es viele Wörter, die eines der Paare EN, EI, ER, CH enthalten. Kommt eine Buchstabenfolge zweimal in einem Text vor und werden beide Buchstabenfolgen gleich verschlüsselt, so muss der Abstand zwischen den Buchstabenfolgen ein Vielfaches der Schlüssellänge sein!

  Also sucht Kim nach Buchstabenpaaren, die mehrfach in der Chiffre vorkommen. In unserer Chiffre kommt dreimal das Paar DD vor. Die Abstände zwischen diesen Paaren sind 5 und 25. Auch das Paar WQ kommt zweimal vor, mit einem Abstand von 40. Alle diese Abstände sind durch 5 teilbar. Kim vermutet also, dass die Schlüssellänge gleich 5 ist. Damit ist sie der Lösung sehr nahe gekommen. Denn nun weiß sie, dass alle fünften Buchstaben mit dem selben Buchstaben verschlüsselt wurde. Sie schreibt also den ersten, sechsten, 11,ten, 16-ten, ... Buchstaben der Chiffre in eine Liste und macht mit dieser Liste eine Häufigkeitsanalyse. Ist die Nachricht viel länger als der Schlüssel, so kann sie genau wie bei der Caesar-Verschlüsselung dadurch schnell den ersten Buchstaben des Schlssels generieren. Genauso verfährt sie auch mit den restlichen Buchstaben und kann damit den gesamten

Schlüssel erzeugen. Kennt sie den Schlüssel, kann sie natürlich auch die Nachricht entschlüsseln.

Bemerkung 5.1.11. Man kann nie sicher sein, dass ein Angreifer die Attacke fährt, die man erwartet. Eine Möglichkeit *Sicherheit* zu messen ist die erwartete Anzahl von Rechenschritten zu bestimmen, die Kim braucht um die Nachricht zu knacken. Erwarten wir, dass die Nachricht mit wenigen Rechenschritten geknackt wird, ist das System nicht sicher. Erwarten wir, dass die Nachricht nur mit extrem vielen Rechenschritten zu knacken ist, sehen wir das System als sicher(er) an.

Definition 5.1.12. Seien $g, f : \mathbb{R} \longrightarrow \mathbb{R}$ Funktionen. Wir schreiben $g \in O(f)$, falls $\limsup_{x\to\infty} \frac{g(x)}{f(x)} < \infty$.

Die Aussage $O(f)$ besagt also, dass die Funktion g höchstens so schnell wächst wie die Funktion f. Dabei ignorieren wir Konstanten. Anstelle von g und f schreiben wir meistens nur die Funktionsvorschrift. Wir benutzen diese Notation um die Anzahl von Rechenschritten abzuschätzen, die für gewisse Aufgaben nötig sind. Daher wird unsere Variable oft n, und nicht x, heißen.

Beispiel 5.1.13. Wir benutzen diese Notation um die Anzahl von Rechenschritten abzuschätzen, die für gewisse Aufgaben nötig sind. Daher wird unsere Variable oft n, und nicht x, heißen. Etwa

- $10^{10}n \in O(n)$
- $10^{-10}n^2 \notin O(n)$
- $\sqrt{n} \in O(n)$
- $\ln(n) \in O(\sqrt{n})$
- $\sqrt{n} \notin O(\ln(n))$
- $\frac{10^{10}}{n} \in O(1)$

Die anzahl von Rechenschritten um eine Caesar-Verschlüsselung der Länge n zu knacken ist $< 26n$ – also in $O(n)$. Da n genau die Anzahl von Rechenschritten ist, die man braucht um die Nachricht zu entschlüsseln *wenn man den Schlüssel schon kennt*, ist dies eine etwas formalere Begründung für die Unsicherheit der Caesar-Verschlüsselung.

Jedes symmetrische Kryptosystem lebt davon, dass Anna und Ben einen gemeinsamen Schlüssel haben mit dem Anne die Nachricht ver- und Ben die Nachricht entschlüsseln kann. Aber wie kommt der Schlüssel von Anna zu Ben, oder andersherum? Wenn Sie im Internet einkaufen, dann bezahlen Sie per Kreditkarte, auch wenn Sie sich vorher nicht persönlich mit der IT-Abteilung des Shops zusammengesetzt haben. Nun ist das Internet aber eine potentiell sehr unsichere Leitung. Es muss also davon ausgegangen werden, dass die gesamte Kommunikation überwacht wird. Wie ist es dennoch möglich Informationen sicher auszutauschen?

5.2 Etwas Zahlentheorie

Wir stellen uns nun auf die Annahme, dass es sich um digitale Kommunikation handelt, die verschlüsselt werden soll. Dann genügt es immer Zahlen zu verschlüsseln, da in einem PC alles durch Bit-Stellungen – also Zahlen im Binärsystem – gespeichert ist. Um Text in Bit-Stellungen zu verwanden wird meistens die Codierung ASCII oder Unicode benutzt. Dabei besteht ASCII aus den nötigsten Satzzeichen und Unicode enthält auch Emojis und hat noch einige freie Plätze, die in Zukunft vergeben werden können. Weiter ist alles von praktischer Relevanz endlich. Wenn Sie sehr sehr viele Symbole verschicken möchten, müssen Sie das in mehreren Nachrichten vorgegebener Größe machen, da irgendwann der Speicher voll ist. Wir wollen daher endliche Zahlbereiche – wie $\mathbb{Z}/n\mathbb{Z}$ – genauer betrachten.

Bemerkung 5.2.1. Wir werden zunächst versuchen eine Möglichkeit zu finden, wie Anna und Ben sich auf einen geheimen Schlüssel einigen können obwohl die gesamte Kommunikation überwacht wird. Mit diesem Schlüssel kann dann ein symmetrisches Kryptosystem gestartet werden. Dazu müssen Anna und Ben irgendetwas geheimes kennen, was sonst niemand kennt. Da die gesamte Kommunikation überwacht wird, müssen dies zunächst zwei verschiedene Dinge sein. Sagen wir Anna wählt eine Zahl a (geheim) und Ben wählt eine Zahl b (geheim).
Da wir davon ausgehen, dass die gesamte Kommunikation überwacht wird, ist es nicht möglich a und/oder b zu verschicken. Stattdessen, sollen Werte verschickt werden, diie man aus a bzw. b berechnen kann. Das kann man sich als „verschlüsseln“ von a und b vorstellen. Es soll also eine Funktion

f geben, und der Wert $f(a)$ wird an Ben geschickt und $f(b)$ wird an Anna geschickt.
Dann kennt Anna a und $f(b)$. Ben kennt b und $f(a)$. Kim kennt $f(a)$ und $f(b)$. Das sind erst einmal drei verschiedene Informationen. Wichtig ist nur, dass Kim aus $f(a)$ nicht a und aus $f(b)$ nicht b berechnen können darf!
Wir brauchen also eine Funktion f, die ganz einfach (schnell) berechnet werden kann, damit Anna und Ben ihre geheimen Zahlen verschlüsseln können. Es muss aber unheimlich schwierig sein, Urbilder von f zu berechnen! Das ist in jedem Fall nötig für eine geheime Kommunikation. Dass Anna und Ben aus a und $f(b)$, bzw. aus b und $f(a)$, den gleichen Wert berechnen können müssen stellen wir erst einmal hinten an.
Bisher hatten wir als Verschlüsselungsfunktion nur die Addition kennengelernt, die durch die Subtraktion (die genauso schwierig/einfach ist wie die Addition) wieder rückgängig gemacht werden konnte. Unsere Funktion f soll nun eine große Diskrepanz im Aufwand von Berechnungen der Bilder, bzw. der Urbilder, aufweisen.

Wie schon angekündigt, beschäftigen wir uns zunächst mit den Restklassenringen $\mathbb{Z}/n\mathbb{Z}$. Beachten Sie beim Studium, dass die Mathematik, die wir hier behandeln werden, viel älter als jeder Computer ist. Die Mathematik ist hier also – anders als in den vorherigen Abschnitten – nicht extra für Probleme in der realen Welt konstruiert worden. Sie entstand vollständig in der reinen Mathematik.

Bemerkung 5.2.2. Sind zwei ganze Zahlen a und b gegeben, von denen mindestens eine ungleich Null ist, haben diese beiden Zahlen immer einen größten gemeinsamen Teiler. Das ist die größte natürliche Zahl $\mathrm{ggT}(a, b)$, die sowohl a als auch b teilt. Diese Zahl ist immer eindeutig bestimmt.

Satz 5.2.3 (Division mit Rest)**.** *Seien $a, b \in \mathbb{Z}$, mit $b \neq 0$. Dann gibt es eindeutige Elemente $r, q \in \mathbb{Z}$, mit*

(i) $a = q \cdot b + r$ *und*

(ii) $0 \leq r < |b|$.

Den Beweis von diesem Satz haben Sie (hoffentlich) schon einmal gesehen. Daher verzichten wir hier darauf.

Lemma 5.2.4. *Seien $a, b, q, r \in \mathbb{Z}$ mit $b \neq 0$ und $a = q \cdot b + r$. Dann gilt* $\mathrm{ggT}(a, b) = \mathrm{ggT}(b, r)$.

Beweis. Sei $c \in \mathbb{Z}$ mit $c \mid a$ und $c \mid b$. Dann ist auch $c \mid (-1) \cdot q \cdot b$, und damit $c \mid a - q \cdot b = r$. D.h.: Jeder gemeinsame Teiler von a und b ist auch ein gemeinsamer Teiler von b und r. Genauso gilt für ein $c \in \mathbb{Z}$ mit $c \mid b$ und $c \mid r$ auch $c \mid q \cdot b + r = a$. Es sind also die gemeinsamen Teiler von a und b genau die gemeinsamen Teiler von b und r. Insbesondere folgt $\mathrm{ggT}(a, b) = \mathrm{ggT}(b, r)$. □

Kombinieren wir dieses Lemma mit der Division mit Rest erhalten wir ein gutes Verfahren um den ggT von zwei Zahlen auszurechnen. Da stets $\mathrm{ggT}(a, b) = \mathrm{ggT}((-1) \cdot a, b)$ gilt, genügt es wenn wir den ggT von zwei natürlichen Zahlen berechnen können.

Euklidischer Algorithmus 5.2.5. Sei $a_1 \in \mathbb{N}_0$ und $a_2 \in \mathbb{N}$. Folgendes Verfahren liefert den größten gemeinsamen Teiler von a_1 und a_2.

1. Schritt: Bestimme $q, a_3 \in \mathbb{N}$ mit $a_1 = q \cdot a_2 + a_3$ und $0 \leq a_3 < a_2$.
2. Schritt: Ist $a_3 = 0$, so ist $\mathrm{ggT}(a_1, a_2) = \mathrm{ggT}(a_2, 0) = a_2$ und wir sind fertig. Ist $a_3 \neq 0$ so gehe zum 1. Schritt zurück mit (a_2, a_3) anstatt (a_1, a_2).

Dieses Verfahren liefert eine Folge $a_2 > a_3 > \ldots \in \mathbb{N}_0$. Da es nur endlich viele natürliche Zahlen kleiner als a_2 gibt, endet das Verfahren nach endlich vielen Schritten. Weiter haben wir in Lemma 5.2.4 gesehen, dass für alle $i \in \mathbb{N}$ gilt $\mathrm{ggT}(a_i, a_{i+1}) = \mathrm{ggT}(a_{i+1}, a_{i+2})$. Damit liefert das Verfahren auch wirklich $\mathrm{ggT}(a_1, a_2)$.

Beispiel 5.2.6. Wir berechnen den $\mathrm{ggT}(a, b)$ für

(a) $a = 748$ und $b = 528$:

$$\begin{aligned}
&(1) \quad 748 = 1 \cdot 528 + 220 \\
&(2) \quad 528 = 2 \cdot 220 + 88 \\
&(3) \quad 220 = 2 \cdot 88 + 44 \\
&(4) \quad 88 = 2 \cdot 44 + 0 \quad \Longrightarrow \quad \mathrm{ggT}(748, 528) = 44
\end{aligned}$$

Es genügt also immer wieder Division mit Rest durchzuführen und danach die Einträge nach links zu shiften. Das machen wir solange, bis wir

den Rest Null erhalten. Dann hören wir auf und der letzte positive Rest gibt uns den gesuchten ggT an.

(b) $a = 34$ und $b = 21$:

$$
\begin{aligned}
&(1) \quad 34 = 1 \cdot 21 + 13 \\
&(2) \quad 21 = 1 \cdot 13 + 8 \\
&(3) \quad 13 = 1 \cdot 8 + 5 \\
&(4) \quad 8 = 1 \cdot 5 + 3 \\
&(5) \quad 5 = 1 \cdot 3 + 2 \\
&(6) \quad 3 = 1 \cdot 2 + 1 \\
&(7) \quad 2 = 2 \cdot 1 + 0 \quad \Longrightarrow \ \mathrm{ggT}(34, 21) = 1
\end{aligned}
$$

Bemerkung 5.2.7. In Beispiel (a) haben wir den ggT trotz großer Zahlen recht schnell gefunden. In Beispiel (b) hat die Berechnung des ggT sehr lange gedauert, obwohl die Zahlen viel kleiner waren. Der Euklidische Algorithmus braucht am längsten, wenn in jedem Schritt $q = 1$ ist, da in diesem Fall die Reste immer so groß wie möglich sind.
Wir suchen nun die kleinsten natürlichen Zahlen a_{n+1} und a_n für die der Euklidische Algorithmus genau n Schritte benötigt. Dann muss nach der Vorbemerkung in jedem bis auf den letzten Schritt $q = 1$ sein. Es gilt also

$$
\begin{aligned}
&(1) \quad a_{n+1} = 1 \cdot a_n + a_{n-1} \\
&(2) \quad a_n = 1 \cdot a_{n-1} + a_{n-2} \\
&(3) \quad a_{n-1} = 1 \cdot a_{n-2} + a_{n-3} \\
&\quad \vdots \quad \vdots \\
&(\mathrm{n}) \quad a_2 = q \cdot a_1 + a_0
\end{aligned}
$$

Da der Euklidische Algorithmus hier abbrechen soll, gilt $a_0 = 0$. Da die Ausgangswerte so klein wie möglich sein sollen, muss $a_1 = 1$ gelten (was $q = a_2$ impliziert). Wieder da die Werte so klein wie möglich sein sollen, und $a_2 > a_1 = 1$ gilt, muss $a_2 = 2$ gelten. Weiter sehen wir in jedem Schritt die wohlbekannte Gleichung $a_{n+1} = a_n + a_{n-1}$, mit den Anfangswerten $a_1 = 1$ und $a_2 = 2$. Damit ist a_k die k-te Fibonacci-Zahl für alle $k \in \mathbb{N}$. Das passt genau zu den Zahlen aus Beispiel (b).

Wir haben gezeigt: Braucht der Euklidische Algorithmus mit Anfangswerten a und b mindestens n Schritte, dann ist $a \geq f_{n+1}$ und $b \geq f_n$, wobei f_k jeweils die k-te Fibonacci-Zahl bezeichnet. Anders ausgedrückt bedeutet das, dass der Euklidische Algorithmus am langsamsten ist, wenn wir mit zwei aufeinanderfolgenden Fibonacci-Zahlen starten. Das kan man nutzen um zu zeigen, dass der Euklidische Algorithmus immer schnell ist. Die genaue Formulierung halten wir in einem Satz fest, den Sie in den Übungen beweisen.

Satz 5.2.8. *Seien $a, b \in \mathbb{N}$, mit $a \geq b > 0$. Dann ist die Anzahl von Schritten, die der Euklidische Algorithmus benötigt in $O(\ln(b))$.*

Bemerkung 5.2.9. Wir betrachten noch einmal das Beispiel (a) aus 5.2.6. Wen wir die Zeilen rückwärts, also von unten nach oben lesen, erhalten wir

$$\begin{aligned} 44 &\overset{(3)}{=} 220 - 2 \cdot 88 \overset{(2)}{=} 220 - 2 \cdot (528 - 2 \cdot 220) = 5 \cdot 220 + (-2) \cdot 528 \\ &\overset{(1)}{=} 5 \cdot (748 - 1 \cdot 528) - 2 \cdot 528 = 5 \cdot 748 - 7 \cdot 528. \end{aligned}$$

Wir können also $\mathrm{ggT}(748, 528)$ als $\mathbb{Z}$-Linearkombination von 748 und 528 schreiben.

Satz 5.2.10 (Lemma von Bézout)**.** *Seien $a, b, d \in \mathbb{Z}$ mit $(a, b) \neq (0, 0)$. Die Gleichung $a \cdot x + b \cdot y = d$ ist genau dann für ganze Zahlen x, y lösbar, wenn $\mathrm{ggT}(a, b) \mid d$ gilt.*

Beweis. Wir müssen zwei Implikationen zeigen.

- $\Rightarrow$ Seien $x, y \in \mathbb{Z}$ mit $a \cdot x + b \cdot y = d$. Jeder Teiler von a ist ein Teiler von $a \cdot x$ und jeder Teiler von b ist ein Teiler von $b \cdot y$. Damit gilt $\mathrm{ggT}(a, b) \mid a \cdot x$ und $\mathrm{ggT}(a, b) \mid b \cdot y$. Insbesondere ist damit auch $\mathrm{ggT}(a, b) \mid d$.

- $\Leftarrow$ Genau wie in der vorstehenden Bemerkung sehen wir, in dem wir den Euklidischen Algorithmus von unten nach oben betrachten, dass $\mathrm{ggT}(a, b)$ eine $\mathbb{Z}$-Linearkombination von a und b ist. (Hier hat sich natürlich eine Induktion über die Anzahl von Schritte im Euklidischen Algorithmus versteckt.) D.h. es gibt $x', y' \in \mathbb{Z}$, mit

$$\mathrm{ggT}(a, b) = a \cdot x' + b \cdot y'. \tag{5.1}$$

Sei nun $d \in \mathbb{Z}$ mit $\mathrm{ggT}(a,b) \mid d$. Dann existiert ein $k \in \mathbb{Z}$ mit $k \cdot \mathrm{ggT}(a,b) = d$. Multiplizieren wir die Gleichung (5.1) auf beiden Seiten mit k erhalten wir

$$d = a \cdot \underbrace{(k \cdot x')}_{=x} + b \cdot \underbrace{(k \cdot y')}_{=y}.$$

Das war zu zeigen.

□

Fast alles, was Sie über die ganzen Zahlen $\mathbb{Z}$ wissen, folgt aus dem Lemma von Bézout. Insbesondere impliziert dieses Lemma die eindeutige Primfaktorzerlegung auf $\mathbb{Z}$.

Definition/Satz 5.2.11. *Die* Einheitengruppe modulo n *ist gegeben durch die Menge*

$$\begin{aligned}(\mathbb{Z}/n\mathbb{Z})^* &= \{[a]_n | \exists [b]_n \in \mathbb{Z}/n\mathbb{Z},\ [a]_n \cdot [b]_n = [1]_n\} \\ &= \{[a]_n | \ \mathrm{ggT}(a,n) = 1\}\end{aligned}$$

und der Multiplikation auf $\mathbb{Z}/n\mathbb{Z}$.
Die Eulersche-φ-Funktion *bildet alle natürlichen Zahlen* n *ab auf* $\varphi(n) = |(\mathbb{Z}/n\mathbb{Z})^*|$.

Beweis. Der Beweis ist zum Glück mit unserem Wissen ganz einfach:

$$\begin{aligned}\mathrm{ggT}(a,n) = 1 &\iff ax + ny = 1 \text{ für gewisse } x, y \in \mathbb{Z} \\ &\iff ax \equiv 1 \mod n \text{ für gewisses } x \in \mathbb{Z} \\ &\iff [a]_n \cdot [x]_n = [1]_n \text{ für gewisses } x \in \mathbb{Z}\end{aligned}$$

□

Da der Euklidische Algorithmus sehr schnell ist, findet man auch sehr schnell die Zahlen x und y aus dem Lemma von Bézout. Wie im Beweis gerade gesehen, kann man daher sehr schnell (mit wenigen Rechenschritten) das Inverse von $[a]_n \in (\mathbb{Z}/n\mathbb{Z})^*$ berechnen.

Beispiel 5.2.12. Wir berechnen das Inverse von $[9]_{13}$. Dazu führen wir erst den Euklidischen Algorithmus mit 9 und 13 durch.

$$\begin{aligned} (1) \quad & 13 = 1 \cdot 9 + 4 \\ (2) \quad & 9 = 2 \cdot 4 + 1 \\ (3) \quad & 4 = 4 \cdot 1 + 0 \end{aligned}$$

Damit haben wir bestätigt, dass $\mathrm{ggT}(13, 9) = 1$ ist, was wir natürlich vorher schon wussten. Allerdings können wir nun mit dem Lemma von Bézout das Inverse von $[9]_{13}$ berechnen. Es ist

$$1 \overset{(2)}{=} 9 - 2 \cdot 4 \overset{(1)}{=} 9 - 2 \cdot (13 - 1 \cdot 9) = 3 \cdot 9 - 2 \cdot 13.$$

Es folgt $1 \equiv 3 \cdot 9 \mod 13$ und somit ist $[3]_{13}$ das Inverse von $[9]_{13}$. Wie üblich schreiben wir dafür $[3]_{13} = [9]_{13}^{-1}$.

Bemerkung 5.2.13. Wir halten kurz inne und besprechen ein paar Dinge.

- Wir haben die Einheiten*gruppe* definiert. Das ist natürlich tatsächlich eine Gruppe, wie sie auch in der linearen Algebra definiert wurde. D.h. es gibt eine Verknüpfung, die assoziativ ist, ein neutrales Element bzgl. dieser Verknüpfung und jedes Element besitzt ein Inverses. Man sieht schnell ein, dass $(\mathbb{Z}/n\mathbb{Z})^*$ alle diese Eigenschaften besitzt. Besteht eine Gruppe nur aus endlich vielen Elementen, so sprechen wir von einer endlichen Gruppe.

- Sei $(G, \circ)$ eine endliche Gruppe mit neutralem Element e. Für jedes $g \in G$ setzen wir $g^0 = e$ und bezeichnen mit g^{-1} das Inverse von g. Dann setzen wir

$$g^n = \underbrace{g \circ \ldots \circ g}_{n\text{-mal}} \quad \text{und} g^{-n} = (g^n)^{-1} \quad \text{für alle } n \in \mathbb{N}.$$

  Der Satz von Lagrange (der wird in der linearen Algebra behandelt) gilt $g^{|G|} = e$ für jedes $g \in G$. Insbesondere gibt es für jedes $g \in G$ ein kleinstes Element $\mathrm{ord}(g) \in \mathbb{N}$, mit $g^{\mathrm{ord}(g)} = e$. Diese natürliche Zahl $\mathrm{ord}(g)$ heißt die *Ordnung von* g. Erfüllt $k \in \mathbb{N}$ die Bedingung $g^k = e$, so folgt $\mathrm{ord}(g) \mid k$. Denn: Aufgrund der Minimalität der Ordnung, ist $k \geq \mathrm{ord}(g)$. Wir teilen k mit Rest durch $\mathrm{ord}(g)$ und erhalten $k =$

$q\,\mathrm{ord}(g) + r$, mit $0 \leq r < \mathrm{ord}(g)$. Damit gilt $e = g^k = (g^{\mathrm{ord}(g)})^q \circ g^r = e^q \circ g^r = g^r$. Da $r < \mathrm{ord}(g)$ ist, muss $r = 0$ gelten, was nichts anderes als $\mathrm{ord}(g) \mid k$ bedeutet.

Die Gruppe G heißt *zyklisch*, wenn es (mindestens) ein $g \in G$ gibt, mit $G = \{e, g, g^2, \ldots\}$. Das ist genau dann der Fall, wenn es ein $g \in G$ gibt, mit $\mathrm{ord}(g) = |G|$. In diesem Fall heißt g ein *Erzeugendenelement von G*.

Lemma 5.2.14. *Seien $p_1, \ldots, p_r$ verschiedene Primzahlen, dann gilt*

$$\varphi(p_1 \cdot \ldots \cdot p_r) = (p_1 - 1) \cdot \ldots \cdot (p_r - 1).$$

Beweis. Es ist $\varphi(n)$ die Anzahl von Elementen aus $\{1, 2, \ldots, n\}$, die keinen gemeinsamen Teiler mit n haben. Wenn nun $n = p$ eine Primzahl ist, sind das genau die Elemente $1, \ldots, p-1$. Damit gilt $\varphi(p) = p-1$. Der allgemeine Fall folgt aus dem chinesischen Restsatz, den Sie in der linearen Algebra kennengelernt haben.
Der letzte Satz ist sicher etwas unbefriedigend, wenn Sie sich nur noch dunkel daran erinnern können, dass es etwas mit dem Namen *chinesischer Restsatz* gab. Daher werden wir noch kurz den Beweis im Fall $r = 2$ darstellen, der für uns besonders wichtig ist. Dazu seien p und q verschiedene Primzahlen. Welche Zahlen aus $\{1, \ldots, pq\}$ sind nun durch p teilbar?
Es sind genau die Elemente $p, 2p, 3p, \ldots, qp$. Genauso sind die Elemente aus $\{1, \ldots, pq\}$ die durch q teilbar sind die Elemente $q, 2q, \ldots, pq$. Das sind zusammen genau $q + p - 1$ verschiedene Elemente. Damit sind genau $pq - p - q + 1 = (p-1)(q-1)$ der Elemente aus $\{1, \ldots, pq\}$ weder durch p noch durch q teilbar. Da p und q Primzahlen sind, sind das genau die Elemente, die keinen gemeinsamen Teiler mit pq haben. Es folgt, wie gewünscht, $\varphi(pq) = (p-1)(q-1)$. □

Theorem 5.2.15 (Satz von Euler)**.** *Sei $n \in \mathbb{N}$ und $a \in \mathbb{Z}$, mit $\mathrm{ggT}(a, n) = 1$, dann gilt $a^{\varphi(n)} \equiv 1 \mod n$. Insbesondere gilt für jede Primzahl p und jedes $a \in \mathbb{Z}$ die Bedingung $a^p \equiv a \mod p$.*

Beweis. Wir können die Aussage umformulieren zu $[a]_n^{\varphi(n)} = [1]_n$ für alle $[a]_n \in (\mathbb{Z}/n\mathbb{Z})^*$. Da $\varphi(n) = |(\mathbb{Z}/n\mathbb{Z})^*|$ ist, ist dies wahr nach dem Satz von Lagrange. (Dieser besagt, dass in jeder endlichen Gruppe G die Bedingung $g^{|G|}$ das neutrale Element ist.)

Ist nun p eine Primzahl, so gilt $a^{\varphi(p)} \equiv a^{p-1} \equiv 1 \mod p$ für alle $a \in \mathbb{Z}$, mit $\text{ggT}(a,p) = 1$. Multiplizieren wir nun alles mit a erhalten wir $a^p \equiv 1 \mod p$.
Falls $\text{ggT}(a,p) \neq 1$, bleibt nur noch $p \mid a$ übrig. In diesem Fall ist daher ohnehin $a^p \equiv 0^p \equiv 0 \equiv a \mod p$. □

Beispiel 5.2.16. Es ist also $3^{60} \equiv 1 \mod 77$, da $\varphi(77) = \varphi(7 \cdot 11) = 6 \cdot 10 = 60$ und $\text{ggT}(3,77) = 1$ ist. Können wir das auch rechnerisch überprüfen, und wenn ja, wie viele Rechenschritte wären dafür nötig?

$$\begin{aligned} 3^2 &\equiv 9 \mod 77 \\ 3^4 \equiv 9^2 &\equiv 81 \equiv 4 \mod 77 \\ 3^8 \equiv 4^2 &\equiv 16 \mod 77 \\ 3^{16} \equiv 16^2 &\equiv 256 \equiv 25 \mod 77 \\ 3^{32} \equiv 25^2 &\equiv 625 \equiv 9 \mod 77 \end{aligned}$$

Damit folgt

$$\begin{aligned} 3^{60} &\equiv 3^{32} \cdot 3^{16} \equiv 3^8 \cdot 3^4 \equiv 9 \cdot 25 \cdot 16 \cdot 4 \\ &\equiv 400 \cdot 36 \equiv 15 \cdot 36 \equiv 540 \equiv 1 \mod 77. \end{aligned}$$

Wir haben also genau acht Multiplikationen gebraucht um 3^{60} modulo 77 zu berechnen!

Bemerkung 5.2.17. Dieses Verfahren nennt man *schnelles Exponentieren*. Soll $[a]_n^k$ in $(\mathbb{Z}/n\mathbb{Z})^*$ berechnet werden, für $k, n \in \mathbb{N}$ und $a \in \mathbb{Z}$, mit $\text{ggT}(a,n) = 1$, dann

1. Schreibe $k = a_0 + a_1 \cdot 2 + a_2 \cdot 2^2 + \ldots + a_r \cdot 2^r$ in Dualschreibweise mit $a_i \in \{0,1\}$ für alle $i \in \{0,\ldots,r\}$ und $a_r = 1$. Ihr Smartphone kann diesen Schritt überspringen, da dort bereits alles in Dualzahlen dargestellt ist.

2. Berechne sukzessive $[a]_n^2$, $([a]_n^2)^2$, …, $([a]_n^{2^{r-1}})^2$. Das sind r Multiplikationen in $\mathbb{Z}/n\mathbb{Z}$.

3. Berechne nun $[a]_n^k = [a]_n^{a_0 + a_1 \cdot 2 + a_2 \cdot 2^2 + \ldots + a_r \cdot 2^r} = [a]_n^{a_0} \cdot ([a]_n^2)^{a_1} \cdot \ldots \cdot ([a]_n^{2^r})^{a_r}$. Das sind maximal r Multiplikationen in $\mathbb{Z}/n\mathbb{Z}$.

Weiter ist $[a]_n^{\varphi(n)} = [1]_n$. Damit können wir immer ohne Einschränkung $k \leq \varphi(n) < n$ annehmen. Insbesondere ist somit $2^r \leq k < n$, was $r < \frac{\ln(n)}{\ln(2)}$ impliziert. Es folgt, dass die Anzahl von Rechenschritten die wir brauchen um $[a]_n^k$ zu berechnen in $O(\ln(n))$ ist. Das ist – wie der Name es versprochen hat – tatsächlich schnell!

Bemerkung 5.2.18. Seien nun $[g]_n \in (\mathbb{Z}/n\mathbb{Z})^*$ fest gewählt. Wir betrachten die Funktion

$$f : \mathbb{N} \longrightarrow (\mathbb{Z}/n\mathbb{Z})^* \quad ; \quad k \mapsto [g]_n^k.$$

Diese Funktion ist wohldefiniert, da natürlich $[g]_n^k$ wieder in $(\mathbb{Z}/n\mathbb{Z})^*$ ist. Weiter haben wir gerade gesehen, dass die Funktion ganz einfach (also schnell) berechnet werden kann.
Wir betrachten eine solche Funktion f mit $[g]_n = [5]_{23}$. Wenn nun $f(k) = [17]_{23}$ ist, was ist dann k? Wir müssen die Gleichung $[5]_{23}^k = [17]_{23}$ lösen. Da uns spontan nichts besseres einfällt, probieren wir erstmal aus:

$$\begin{aligned}
5^1 &\equiv 5 \mod 23 \\
5^2 &\equiv 25 \equiv 2 \mod 23 \\
5^3 &\equiv 10 \mod 23 \\
5^4 &\equiv 50 \equiv 4 \mod 23 \\
5^5 &\equiv 20 \mod 23 \\
5^6 &\equiv 100 \equiv 8 \mod 23 \\
5^7 &\equiv 40 \equiv 17 \mod 23
\end{aligned}$$

Damit ist $k = 7$ eine Lösung von $f(k) = [17]_{23}$. Dieses Ausprobieren scheint sehr aufwendig zu sein, insbesondere für große Zahlen. Wenn uns nichts einfällt, was viel besser funktioniert, haben wir hier eine Funktion gefunden, die schnell berechnet ist, aber Urbilder nur sehr schwierig zu berechnen sind. Das ist genau wonach wir gesucht haben.
Mit dem Satz von Euler 5.2.15 sind mit $k = 7$ auch $7 + 22$, $7 + 2 \cdot 22$, ... Lösungen. Daher wollen wir die Funktion etwas präzisieren. Dazu brauchen wir einen Hauch Gruppentheorie.

Satz 5.2.19. *Sei $(G, \circ)$ eine endliche Gruppe und $g \in G$ fest gewählt. Die Abbildung*

$$f_g : \mathbb{Z}/\mathrm{ord}(g)\mathbb{Z} \longrightarrow G \quad ; \quad [k]_{\mathrm{ord}(g)} \mapsto g^k \tag{5.2}$$

ist injektiv und erfüllt $f_g([k]_{\mathrm{ord}(g)}+[\ell]_{\mathrm{ord}(g)}) = f_g([k]_{\mathrm{ord}(g)})\circ f_g([\ell]_{\mathrm{ord}(g)})$ *(d.h.* f_g *ist ein Gruppenhomomorphismus).*

BEWEIS. Wir zeigen als erstes, dass f_g wohldefiniert ist. Sei also $k, k' \in \mathbb{Z}$, mit $k \geq k'$. Nun gilt

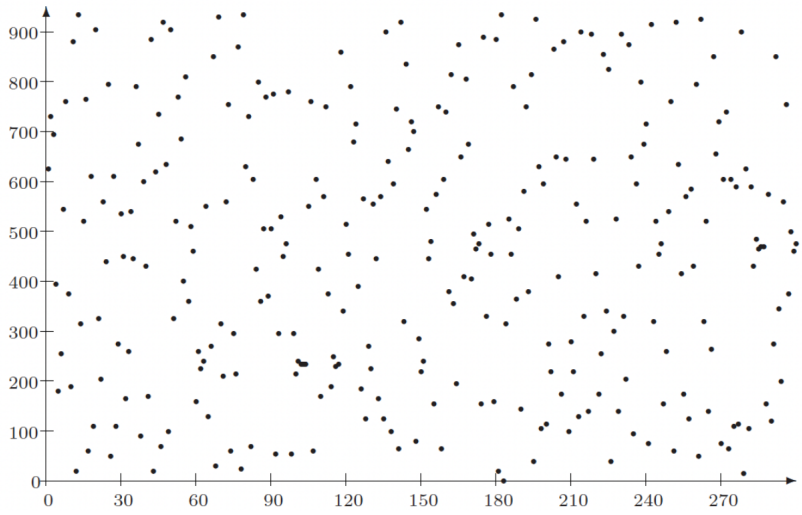
$$\begin{aligned} g^k = g^{k'} &\iff g^{k-k'} = e \iff \mathrm{ord}(g) \mid k-k' \\ &\iff [k]_{\mathrm{ord}(g)} = [k']_{\mathrm{ord}(g)}. \end{aligned}$$

Es ist also $f_g([k]_{\mathrm{ord}(g)}) = f_g([k']_{\mathrm{ord}(g)}) \iff [k]_{\mathrm{ord}(g)} = [k']_{\mathrm{ord}(g)}$. Die Richtung „$\Leftarrow$“ zeigt, dass f_g wohldefiniert ist. Die Richtung „$\Rightarrow$“ zeigt, dass f_g injektiv ist. Die zweite Eigenschaft ist (mit den üblichen Potenzgesetzen) offensichtlich. □

Definition 5.2.20. Wenn G zyklisch ist und g ein Erzeugendenelement von G ist, dann ist $|G| = \mathrm{ord}(g)$ und f_g aus (5.2) ist bijektiv. Insbesondere gibt es in diesem Fall eine Umkehrfunktion $\log_g : G \longrightarrow \mathbb{Z}/_{|G|\mathbb{Z}}$. Diese Funktion nennen wir den *diskreten Logarithmus* zur Basis g auf G.

Bemerkung 5.2.21. Aus der linearen Algebra wissen wir, dass für jede Primzahl p die Einheitengruppe $(\mathbb{Z}/_{p\mathbb{Z}})^* = \{[1]_p, \ldots, [p-1]_p\}$ zyklisch ist. Im folgenden schreiben wir kurz $[a]$ für die Restklasse $[a]_p$. Sei nun $[g]$ ein Erzeugendenelement von $(\mathbb{Z}/_{p\mathbb{Z}})^*$, dann ist insbesondere $\mathrm{ord}([g]) = p-1$. Für jedes $[k] \in \mathbb{Z}/_{(p-1)\mathbb{Z}}$ können wir $[g]^k$ in $O(\ln(p))$ Rechenschritten berechnen.

Die Aufgabe aus den gegebenen Werten p, $[g]$ und $[g]^k$ den Wert $[k]_{p-1}$ zu berechnen ist das berühmte *diskrete Logarithmus Problem* (DLP). Betrachten wir die Werte von $627^k \mod 941$ für die ersten 300 natürlichen Zahlen k, erhalten wir das Bild



Das Bild deutet darauf hin, dass sich die Werte scheinbar unvorhersehbar verhalten. Wenn wir also k suchen, mit $627^k \equiv 377 \mod 941$, so hilft es nicht wenn wir wissen, dass $627^9 \equiv 374 \mod 941$ und $627^{112} \equiv 751 \mod 941$ ist.[1] Dieses DLP scheint sehr schwierig zu sein. Damit haben wir mit f_g eine Funktion gefunden, die einfach zu berechnen ist, aber die Berechnung der Umkehrfunktion ist ein eigenständiges mathematisches Problem. Wir haben also einen ganz heißen Kandidaten für unsere gesuchte Funktion vom Anfang des Abschnittes gefunden!

Bemerkung 5.2.22. Wenn wir beim DLP einfach nur Ausprobieren, in dem wir alle möglichen Werte für k einsetzen, brauchen wir $O(p)$ viele Rechenschritte. Es gibt Algorithmen, die schneller sind als das. In den Übungen lernen wir ein Verfahren kennen, das den diskreten Logarithmus in $O(\sqrt{p})$ Rechenschritten berechnet. Dieser Algorithmus birgt in der Praxis aber andere Schwierigkeiten.
Wenn wir eine 1024-Bit Primzahl p wählen – also ein Primzahl in der Größenordnung $p \approx 2^{1024}$ – dann ist $\ln(p) \approx 709$ und $\sqrt{p} \approx 1{,}34 \cdot 10^{154}$. Die Berechnung von $[g]^k$ erfolgt also selbst auf einem alten Smartphone innerhalb weniger Millisekunden. Die Berechnung von k ist hingegen (mit dem hier angedeuteten Wissen) selbst mit aktuellen Super-Computern sehr zeitaufwendig.

[1] Tatsächlich ist $k = 113$.

Das DLP können wir auch dann noch formulieren, wenn $[g]_p$ kein Erzeugendenelement ist. Dann ändert sich nur der Definitions und der Wertebereich der zugrunde liegenden Funktionen. Die Aufgabenstellung bleibt aber identisch.

Diffie-Hellman Schlüsseltausch 5.2.23. Wie immer wollen Anna und Ben geheim kommunizieren.

- Anna und Ben einigen sich auf eine Primzahl p und ein $g \in \{1, \ldots, p-1\}$, so dass die Ordnung von $[g]_p \in (\mathbb{Z}/p\mathbb{Z})^*$ sehr großist. Diese Zahlen werden ganz öffentlich kommuniziert; z.B. stellt Anna diese Informationen auf ihrer Homepage bereit.

- Nun wählt Anna einen geheimen Schlüssel $a \in \mathbb{N}$ und Ben einen geheimen Schlüssel $b \in \mathbb{N}$. Beide Schlüssel werden niemandem verraten und sind somit nur Anna (a) und Ben (b) selbst bekannt.

- Nun berechnet Anna den Wert $[A]_p = [g]_p^a$ und Ben berechnet $[B]_p = [g]_p^b$. Anna sendet A an Ben und Ben sendet B an Anna.

- Nun berechnet Anna B^a und Ben berechnet A^b. Es ist

$$B^a \equiv (g^b)^a \equiv (g^{a \cdot b}) \equiv (g^a)^b \equiv A^b \mod p$$

  Insbesondere ist $[B]_p^a = [A]_p^b$.

- Der gemeinsame Schlüssel von Anna und Ben ist nun $[B]_p^a$ und beiden Parteien bekannt.

Beispiel 5.2.24. Wir geben ein Beispiel mit sehr kleinen Zahlen an:

Anna geheimes Wissen	öffentliches Wissen	Bens geheimes Wissen
	$p = 17$ und $g = 5$	
$a = 6$		$b = 10$
	$A = 5^6 \mod 17 = 2$ $B = 5^{10} \mod 17 = 9$	
$k = 9^6 \mod 17 = 4$		$k = 2^{10} \mod 17 = 4$

Damit haben sich Anna und Ben auf den Schlüssel $k = 4$ geeinigt. Es ist zu beachten, dass keiner von beiden am Anfang wusste was der Schlüssel am Ende sein wird.
Angenommen Kim konnte die gesamte Kommunikation aus 5.2.23 mithören. Dann kennt Sie p, g, A und B. Aus diesen Daten kann sie aber nur unter großem Aufwand den Schlüssel $B^a \mod p$ generieren.
Jeder der geheimen Schlüssel a oder b würde sofort zum Ziel führen, aber um an a zu gelangen muss $\log_g(A)$ berechnet werden. Es muss also das DLP gelöst werden. In unserem Beispiel muss Kim also folgende Rechnungen anstellen:

$$5^1 \equiv 5 \not\equiv 2 \mod 17 \quad 5^2 \equiv 8 \not\equiv 2 \mod 17 \quad 5^3 \equiv 6 \not\equiv 2 \mod 17$$
$$5^4 \equiv 13 \not\equiv 2 \mod 17 \quad 5^5 \equiv 14 \not\equiv 2 \mod 17 \quad 5^6 \equiv 2 \mod 17$$

Erst nach diesen Rechnungen hat Kim den Wert $a = \log_5([2]_{17}) = [6]_{16}$ gefunden und kann nun $k = B^a \mod p$ berechnen.
ACHTUNG: Es ist nicht bekannt, ob das DLP tatsächlcih gelöst werden muss, um den Diffie-Hellman Schlüsseltausch zu knacken. Möglicherweise lässt sich der gemeinsame Schlüssel auch anders aus Kims Wissen generieren.

Bemerkung 5.2.25.

- In der Praxis wird empfohlen, mindestens eine 1024-Bit Primzahl zu nutzen um eine akzeptable Sicherheit zu erlangen.

- Es gibt eine Klasse von *Index-Calculus-Algorithmen*, die das DLP in voraussichtlich $O(e^{\sqrt{\ln(p)\ln(\ln(p))}})$ Rechenschritten lösen. Das kann man etwas einfacher als $O(p^\varepsilon)$, mit $\varepsilon > 0$, schreiben. Je kleiner dabei das ε, desto größer die implizite Konstante. Diese Algorithmen benutzen die eindeutige Primfaktorzerlegung auf $\mathbb{Z}$.

- Bei einem dieser Algorithmen, das Zahlkörper-Sieb, basiert der Großteil der Rechenschritte nur auf der Primzahl p. Es kann also eine Primzahl p vorbereitet werden. Wenn diese dann bei einer Kommunikation benutzt wird, kann man hoffen mit einigem Aufwand und extrem guter Hardware, jedes DLP in $(\mathbb{Z}/p\mathbb{Z})^*$ lösen.[2]

[2] Einige Author*innen, zB hier, spekulieren, dass die NSA diesen Aufwand aufbringen kann.

Das ist in sofern gefährlich, da einige Primzahlen sehr häufig verwendet werden. Das liegt daran, dass es Primzahlen gibt die unsicherer sind als andere. Wenn man also weiß, dass eine Primzahl „sicher“ ist, dann möchte man sie gerne nutzen.

Beispiel 5.2.26. Für Elemente $[g]_p$ mit sehr kleiner Ordnung, ist das DLP sehr einfach, da man nur sehr wenig Exponenten ausprobieren muss. Angenommen $p = 2^n + 1$ für $n \in \mathbb{N}$ (etwa $p = 17$). Wir möchten für gegebene Werte $[a]_p$ und $[b]_p$ den Wert $\log_{[a]}([b])$ berechnen. Wir suchen also ein $x \in \mathbb{N}$ mit $a^x \equiv b \mod p$.
Wir schreiben dieses unbekannte x in Dualdarstellung

$$x = x_0 + x_1 \cdot 2 + x_2 \cdot 2^2 + \ldots + x_{n-1} \cdot 2^{n-1}, \text{ mit } x_i \in \{0, 1\}.$$

Jetzt möchten wir jedes x_i einzeln berechnen. Wir nutzen aus, dass für alle $d \in \mathbb{N}$, mit $p \nmid d$, gilt $d^{2^n} \equiv d^{\varphi(p)} \equiv 1 \mod p$ (siehe 5.2.15). Wir berechnen nun ohne Mühe die Werte

$$b^2, b^{2^2}, \cdot b^{2^{n-1}} \quad \text{und} \quad a^2, a^{2^2}, \cdot a^{2^{n-1}}$$

alles modulo p. Ist nun $\operatorname{ord}([a]) = 2^n = p - 1$, so ist $\operatorname{ord}([a^{2^{n-1}}]) = 2$. Den diskreten Logarithmus zur Basis $[a^{2^{n-1}}]$ ist somit trivial! Die Idee ist nun

$$\begin{aligned} b^{2^{n-1}} &\equiv (a^x)^{2^{n-1}} \equiv a^{2^{n-1}x_0} \cdot a^{2^n(x_1 + x_2 2 + \ldots + x_{n-1}2^{n-2}} \\ &\equiv (a^{2^{n-1}})^{x_0} \mod p. \end{aligned}$$

Damit ist x_0 ganz einfach berechnet. Dieses Verfahren können wir iterieren um nach und nach die anderen x_i zu berechnen. Der nächste Schritt wäre

$$\begin{aligned} b^{2^{n-2}} &\equiv (a^x)^{2^{n-2}} \equiv a^{2^{n-2}x_0} \cdot a^{2^{n-1}x_1} \cdot a^{2^n(x_2 + x_3 2 + \ldots + x_{n-1}2^{n-3}} \\ &\equiv (a^{2^{n-2}})^{x_0} \cdot (a^{2^{n-1}})^{x_1} \mod p. \end{aligned}$$

Wieder ist nun x_1 ganz einfach zu berechnen. So machen wir weiter und erhalten nach $n - 1$ Iterationen alle x_i und somit den gesuchten Wert x. Fazit:

Zerfällt $p - 1$ in ein Produkt lauter kleiner Primzahlen, dann ist der Diffie-Hellman Schlüsseltausch nicht sicher!

Bemerkung 5.2.27. Wählt man nun für den Diffie-Helmann Schlüsseltausch eine zufällige Primzahl in der richtigen Größenordnung, dann wird die noch nicht für eine Attacke vorbereitet sein. Allerdings besteht dann das Risiko, dass es sich um eine unsichere Primzahl handelt.
Wählt man allerdings eine Primzahl von der bewiesen ist, dass $p-1$ große Primteiler besitzt, dann besteht die Gefahr, dass diese Primzahl bereits vorbereitet wurde und der Diffie-Hellman Schlüsseltausch (zumindest mit einem Super-Computer) geknackt werden kann.
Über die (zukünftige?) Rolle von Quantencomputern sprechen wir in dieser Vorlesung nicht.

Bemerkung 5.2.28. Eine Möglicheit den Diffie-Hellman Schlüsseltausch anzugreifen, ist die *Man-in-the-middle*-Attacke. Dabei gibt sich Kim gegenüber Anna als Ben aus und gegenüber Ben als Anna. Nun tauscht Sie einen Schlüssel mit Ben und einen mit Anna, und kann danach beliebig in die Kommunikation eingreifen. Da dieses Verfahren gerade dafür gemacht ist, dass Parteien, die sich noch nie getroffen haben, sicher kommunizieren sollen, ist das ein echtes Problem. Es sollte also entweder ein Kanal benutzt werden, in dem keine Nachrichten verändert werden können, oder es muss eine Authentifizierung stattfinden. Damit werden wir uns später noch kurz beschäftigen.

5.3 Asymmetrische Kryptosysteme

Den Anfang der asymmetrischen Kryptographie haben wir gerade kennengelernt. Anna und Ben sind mit dem Diffie-Hellman Schlüsseltausch in der Lage mit unterschiedlichem Wissen, den gleichen Schlüssel zu generieren. Durch die *Einwegfunktion* f, die in eine Richtung sehr einfach und in die andere Richtung unfassbar schwierig zu berechnen ist, wird dieses Verfahren als sicher angesehen.
Mit diesem Wissen können wir bereits ohne Umschweife eine neue Verschlüsselungsmethode angeben, die auch dann funktioniert, wenn die gesamte Kommunikation mitgehört wird.

Elgamal-Kryptosystem 5.3.1. Anna möchte Ben eine Nachricht schicken und Kim möchte diese Nachricht lesen.

1. Anna und Ben einigen sich auf eine Primzahl p und ein Element $[g]_p \in (\mathbb{Z}/p\mathbb{Z})^*$ mit großer Ordnung. Beides wird öffentlich bereitgestellt. Annas Nachricht ist nun ein Element $m \in (\mathbb{Z}/p\mathbb{Z})^*$.

2. Ben wählt geheim ein geheimes $[b]_{\mathrm{ord}(g)} \in \mathbb{Z}/\mathrm{ord}(g)\mathbb{Z}$, berechnet $B = [g]_p^b \in (\mathbb{Z}/p\mathbb{Z})^*$ und schickt B an Anna.

3. Anna wählt ein geheimes $[a]_{\mathrm{ord}(g)} \in \mathbb{Z}/\mathrm{ord}(g)\mathbb{Z}$, berechnet $A = [g]_p^a \in (\mathbb{Z}/p\mathbb{Z})^*$ und $c = B^a \cdot m$. Danach schickt Sie das Tupel (A, c) an Ben.

4. Ben berechnet erst A^b und dann $(A^b)^{-1}$. Als letztes berechnet er $(A^b)^{-1} \cdot c = m$.

Bemerkung 5.3.2.

- Da $A^b = B^a$ ist, erhällt Ben im letzten Schritt den Klartext m. Dabei werden von Anna und Ben nur Operationen Exponentieren und Invertieren benutzt. Beides funktioniert sehr schnell.

- Tatsächlich ist das Elgamal-Kryptosystem nichts anderes als ERST einen Schlüssel $k = A^b = B^a$ mit Diffie-Hellman erzeugen. DANN mit diesem Schlüssel k die symmetrische Verschlüsselung $e_k(m) = k \cdot m$ benutzen.

- Angenommen Kim hätte alles mitangehört. Dann kennt Sie A, B und c. Wie in den Übungen gesehen, muss Sie den Schlüssel kennen um die Chiffre zu knacken. Daher muss Sie den Diffie-Hellman Schlüsseltausch 5.2.23 knacken.

- Ben muss nur einmal seinen Wert $B = [g]_p^b$ bereitstellen. Dieser kann auch für viele Kommunikationsparteien gleichzeitig verwendet werden.

- Der Schlüssel ist doppelt so lang, wie der Klartext. Das verbraucht mehr Ressourcen, als man eigentlich gerne hätte.

Beispiel 5.3.3. Auch wenn das Verfahren halbwegs klar sein sollte, gehen wir es noch einmal mit Zahlwerten durch. Anna und Ben einigen sich auf $p = 467$ und $g = 3$. Dann ist $\mathrm{ord}([3]_p) = 233$ ein „großer“ Primteiler von $p - 1$.
Ben wählt nun $b = 123$ und berechnet $B = [3]^b = [406]$. Dieses B wird öffentlich bereitgestellt.
Anna wählt $a = 74$ und berechnet $A = [3]^a = [58]$ und $B^a = [332]$.

Sie möchte die Nachricht 295 verschicken. Dazu berechnet sie $c = m \cdot B^a = [295] \cdot [332] = [337]$ und schickt an Ben die Chiffre $(A, c) = ([58], [337])$.
Ben berechnet erst $A^b = [332]$, dann $(A^b)^{-1} = [339]$ und schließlich $(c \cdot A^b)^{-1} = [295]$. Jetzt hat er Annas Nachricht $m = 295$ entschlüsselt.

Die Idee die hinter dieser Art der Verschlüsselung steckt ist folgende: Es gibt nicht mehr einen Schlüssel der sowohl ver- als auch entschlüsseln kann, sondern zwei verschiedene Schlüssel. Einen, den jeder kennen darf, mit dem man verschlüsselt – den werden wir den *öffentlichen Schlüssel* nennen; im Elgamal-Verfahren ist das Bens Wert B. Ein anderer Schlüssel ist streng geheim und wird zum entschlüsseln der Chiffre benutzt. Das ist der *private Schlüssel*; im Elgamle-Verfahren ist das b. Diese Idee kennen Sie schon von Ihrem Briefkasten. Alle können die Klappe öffnen um eine NAchricht in den Briefkasten hineinzuwerfen. Aber nur Sie können mit Ihrem privaten Schlüssel den Brifkasten öffnen um die Nachrichten herauszunehmen.

Definition 5.3.4. Ein *asymmetrisches Kryptosystem* besteht aus endlichen nicht-leeren Mengen M, C, $K_{\ddot{o}}$ und K_p, einer Abbildung $s : K_p \longrightarrow K_{\ddot{o}}$, sowie Mengen $E = \{e_{k_{\ddot{o}}} : M \to C | k_{\ddot{o}} \in K_{\ddot{o}}\}$ und $D = \{e_{k_p} : C \to M | k_p \in K_p\}$. Dabei muss für alle $k_p \in K_p$ die Bedingung $d_{k_p}(e_{s(k_p)}(m)) = m$ für alle $m \in M$ gelten.

Bemerkung 5.3.5. Das Elgamal-Kryptosystem ist ein asymmetrisches Kryptosystem, mit

- $M = (\mathbb{Z}/p\mathbb{Z})^*$, $C = (\mathbb{Z}/p\mathbb{Z})^* \times (\mathbb{Z}/p\mathbb{Z})^*$
- $K_p = \mathbb{Z}/\mathrm{ord}(g)\mathbb{Z}$, $K_{\ddot{o}} = (\mathbb{Z}/p\mathbb{Z})^*$
- $s(k_p) = g^{k_p}$ für alle $k_p \in K_p$

Wir kommen nun zum bekanntesten asymmetrischen Kryptosystem – dem RSA-Verfahren. Dazu brauchen wir noch ein bisschen Vorarbeit.

Satz 5.3.6. *Seien p und q zwei verschiedene Primzahlen und sei $k \in \mathbb{N}$ so, dass $k \equiv 1 \mod \varphi(p \cdot q)$ ist. Dann gilt für alle $m \in \mathbb{Z}$ die Gleichung $m^k \equiv m \mod p \cdot q$.*

Beweis. Wir betrachten drei unterschiedliche Fälle, je nachdem was $\mathrm{ggT}(a, pq)$ ist.

Falls $\boxed{\text{ggT}(m, pq) = 1}$, so gilt mit dem Satz von Euler 5.2.15 $m^{\varphi(pq)} \equiv 1 \mod pq$. Ist nun $k \equiv 1 \mod \varphi(pq)$, so ist $k = r \cdot \varphi(pq) + 1$ für ein $r \in \mathbb{Z}$. Damit folgt

$$m^k \equiv (m^{\varphi(pq)})^r \cdot m \equiv 1 \cdot m \equiv m \mod pq.$$

Sei nun $\boxed{\text{ggT}(m, pq) = pq}$. Dann ist $pq \mid m$, also $m \equiv 0 \mod pq$. Damit ist aber natürlich auch $m^k \equiv 0 \equiv m \mod pq$.
Jetzt fehlt nur noch der Fall $\boxed{1 \neq \text{ggT}(m, pq) \neq pq}$. Da p und q Primzahlen sind, bedeutet dies $\text{ggT}(m, pq) \in \{p, q\}$. Es genügt natürlich nur $\text{ggT}(m, pq) = p$ zu betrachten, da beide Primzahlen absolut gleichberechtigt sind. Dann ist $p \mid m$ und somit folgt wieder $m^k \equiv 0 \equiv m \mod p$. Das bedeutet nichts anderes als $p \mid m^k - m$.
Weiter ist $\text{ggT}(q, a) = 1$. Miit dem Satz von Euler 5.2.15 ist daher $m^{q-1} \equiv 1 \mod q$. Damit ist auch $m^{r\varphi(pq)+1} \equiv (m^{(q-1)})^{r(p-1)} \cdot m^1 \equiv m \mod q$, was nichts anderes bedeutet als $q \mid m^k - m$. Da aber p und q verschiedene Primzahlen sind, folgt aus $p \mid m^k - m$ und $q \mid m^k - m$ sofort $pq \mid m^k - m$. Das ist äquivalent zu $m^k \equiv m \mod pq$ und der Satz ist bewiesen. □

Bemerkung 5.3.7. Sei nun das Produkt von zwei Primzahlen pq gegeben. Weiter ein Wert $e \in \{1, \ldots pq - 1\}$ und $[b]_{pq} = [m]_{pq}^e$ für ein unbekanntes $[m]_{pq} \in \mathbb{Z}/pq\mathbb{Z}$. Wie können wir $[m]_{pq}$ berechnen? Wir kennen zwei Möglichkeiten:

1. Ausprobieren! Teste nach und nach alle $m \in \{1, \ldots, pq - 1\}$ ob sie die Kongruenz $m^e \equiv b \mod pq$ erfüllen. Das ist natürlich extrem aufwendig, wenn pq sehr groß ist.

2. Bestimme ein $d \in \mathbb{Z}$, mit $d \cdot e \equiv 1 \mod \varphi(pq)$. Dann ist nach Satz 5.3.6 $(m^e)^d \equiv m^{ed} \equiv m \mod pq$. So ein d existiert nur dann, wenn $\text{ggT}(e, \varphi(pq)) = 1$ ist. Das müssen wir für diesen Ansatz also voraussetzen.

Weitere Möglichkeiten an das m heranzukommen sind nicht bekannt! Hier haben wir wieder einen Rechenweg der ganz einfach ist (wir brauchen schnelles Exponentieren und Invertieren), und einen Rechenweg der extrem aufwendig ist. Das scheint gut zu unserer Grundidee der asymmetrischen Kryptosysteme zu passen. Beachten Sie, dass man für die Möglichkeit 2 den Wert

$\varphi(pq)$ kennen muss. Falls wir p und q kennen, ist das ganz einfach, aber wir kennen nur pq.

Lemma 5.3.8. *Ist das Produkt von zwei verschiedenen Primzahlen pq gegeben. Wenn man $\varphi(pq)$ berechnen kann, dann kann man auch p und q berechnen.*

Beweis. Seien also pq und $\varphi(pq)$ bekannt. Dann auch der Wert $pq-\varphi(pq)+1 = pq-(p-1)(q-1)+1 = p+q$. Wenn man aber pq und $p+q$ kennt, kennt man auch das Polynom $x^2-(p+q)x+pq$. Die Nullstellen sind schnell berechnet und – na klar – p und q. □

Nach Bemerkung 5.3.7 beruht also die einzig bekannte Methode $e$te Wurzeln modulo pq zu ziehen, auf der Zerlegung von pq in die beiden Primteiler p und q. Wenn wir einfach nur ausprobieren, und durch immer größere Zahlen teilen, brauchen wir dafür $\min\{p,q\}$ Rechenschritte. Auf der Schwierigkeit dieses Faktorisierungsproblems basiert das nächste Kryprosystem.

RSA-Kryptosystem 5.3.9. Wie immer möchte Anna eine Nachricht an Ben schicken.

1. Ben wählt ganz geheim zwei verschiedene Primzahlen p und q. Diese bilden Bens *privaten Schlüssel*. Dann bildet Ben $N = p\cdot q$ und wählt ein $e \in \{2,\dots,N\}$ mit $\mathrm{ggT}(\varphi(N),e) = 1$. Hier ist es wichtig zu beachten, dass Ben natürlich $\varphi(N) = (p-1)\cdot(q-1)$ kennt.

2. Die Werte N und e werden nun ganz öffentlich bereitgestellt. Sie bilden Bens *öffentlichen Schlüssel*.

3. Anna transferiert ihre Nachricht in ein Element $m \in \mathbb{Z}/N\mathbb{Z}$ und berechnet $c = m^e$. Diesen Wert schickt sie Ben.

4. Ben weiß, dass $\varphi(N) = (p-1)\cdot(q-1)$ ist. Damit kann er ganz einfach ein Element $d \in \mathbb{N}$ berechnen mit $d\cdot e \equiv 1 \mod \varphi(N)$. Dieses d nennen wir Bens *Dechiffrierzahl*. Aus $d\cdot e \equiv 1 \mod \varphi(p\cdot q)$ und $N = p\cdot q$ folgt aus Lemma 5.3.6:
$$c^d = m^{e\cdot d} = m.$$
Damit hat Ben Annas Nachricht rekonstruiert.

Bemerkung 5.3.10. Beachten Sie, dass Anna und Ben nur den Euklidischen Algorithmus, das Lemma von Bézout und schnelles Exponentieren benötigen. Hat Kim hingegen alles mitgehört, dann kennt sie N, e, c und weiß dass $c = m^e$ ist. Nach den Vorbemerkungen führt der einzige bekannte Weg, daraus m zu berechnen, über die Faktorisierung von N.
Diese Faktorisierung ist vermutlich genauso schwierig, wie das DLP zu lösen. Zumindest sind die schnellsten Algorithmen die das jeweilige Problem lösen sehr ähnlich und haben daher auch eine ähnliche Laufzeit.
Im Jahr 1991 wurden verschiedene Produkte von zwei Primzahlen veröffentlicht (RSA-Zahlen genannt) um zu überprüfen, ob man sie faktorisieren kann. Für eine erfolgreiche Faktorisierung gab es Geldpreise, allerdings nur bis 2007. Die größte dieser RSA-Zahlen, die faktorisiert wurde, ist

2140324650240744961264423072839333563008614715144755017797754920881418023447140136643345519095804679610992851872470914587687396261921557363047454770520805119056493106687691590019759405693457452230589325976697471681738069364894699871578949749759374979 37

Diese hat 829-Bits und wurde 2020 faktorisiert. Die kleinste nicht fartorisierte RSA-Zahl ist

2211282552952966643528108525502623092761208950247001539441374831912882294140200198651272972656974659908590033003140005117074220456085927635795375718595429883895870922923849100670303412462054578456641366454068421436129301769402084639106587591479425143514445819 9

mit 862-Bits. In der Praxis sollte man mindestens zwei 512-Bit Primzahlen nehmen, besser zwei 1024-Bit Primzahlen.

Bemerkung 5.3.11. Wir werden noch kurz auf einfache Ideen eingehen, die man bei der Wahl der Primzahlen p und q beachten sollte.

- Es ist genau eine der beiden Primzahlen p oder q kleiner als $\sqrt{pq}$. Wir können also versuchen einen Primteiler in der Nähe von $\sqrt{pq}$ zu finden, in dem wir nach und nach die Elemente $\lfloor\sqrt{pq}\rfloor$, $\lfloor\sqrt{pq}\rfloor - 1, \ldots$ überprüfen. Die Beiden Primzahlen p und q müssen in der Praxis also einen Abstand aufweisen, der so groß ist, dass diese Idee nicht zum Ziel führt.

- Das *quadratische Sieb* nutzt aus, dass jede ungerade Zahl die Differenz zweier Quadratzahlen ist. Das sieht man ganz einfach durch die

Gleichung

$$2n + 1 = (n + 1)^2 - n^2.$$

Kennen wir nun aber $x, y \in \mathbb{N}$, mit $pq = x^2 - y^2$, dann ist $pq = (x - y)(x + y)$ und wir haben eine Faktorisierung gefunden. Das führt auf die Idee zu überprüfen, ob $\sqrt{pq + 1^2}$, $\sqrt{pq + 2^2}$, ... natürliche Zahlen sind. Sobald dies der Fall ist, sind wir fertig.

Bemerkung 5.3.12. Man kann natürlich auch versuchen an die Nachrichten zu kommen, ohne sich mit der Mathematik im Hintergrund zu beschäftigen. Weiß Kim etwa, dass Annas Nachricht entweder „Ja“ oder „Nein “ lautet, dann kann Sie einfach selbst diese beiden Nachrichten verschlüsseln und mit Annas Chiffre vergleichen. Das ist die *Chosen-Plaintext*-Attacke.
Beim Elgamal-Verfahren ist das nicht möglich, da dort die Nachricht, mit Annas Schlüssel *a randomisiert* wurde. So eine Randomisierung wird auch bei RSA-Verfahren eingebaut. Auf die Details verzichten wir hier.
Das RSA-Verfahren wird benutzt, wenn Sie https in der Adressleiste Ihres Browsers sehen. Meistens wird es aber benutzt um einen Schlüssel zu übertragen, mit dem man dann ein wesentlich schnelleres symmetrisches Kryptosystem benutzen kann.

5.4 Signaturen

Früher wurden manche Nachrichten mit einem Siegel eines Siegelringes unterzeichnet. Diese Siegel waren so bekannt, dass der Empfänger genau wusste, wem der Siegelring gehört, der zur Unterzeichnung benutzt wurde. Dabei galt natürlich die Annahme, dass es nur einen dieser Siegelringe gibt.
So ähnlich kann man sich die digitale Signatur auch vorstellen. Es ist allen der öffentliche Schlüssel von Ben bekannt. Wenn Ben also seine Identität bestätigen soll, so wird von ihm verlangt, dass er nachweist, dass er auch den zugehörigen privaten Schlüssel kennt. Der öffentliche Schlüssel ist also der Abdruck des Siegelrings, der private Schlüssel ist der Siegelring selbst.
Wir kümmern uns zunächst um eine Authentifizierung zu Beginn einer Kommunikation. Der Ablauf ist ganz einfach: Anna verschlüsselt eine zufällig generierte Nachricht mit Bens öffentlichem Schlüssel. Das Ergebnis schickt sie an Ben und bittet ihn, die Nachricht zu entschlüsseln. Kann er das, dann

geht Anna davon aus, dass sie tatsächlich mit Ben kommuniziert. Kann er es nicht, ist die Authentifizierung fehlgeschlagen.

Bemerkung 5.4.1. Das öffnet wieder Möglichkeiten für Angriffe auf das Kryptosystem. Denn durch diesen Authentifizierungsprozess ist es potentiellen Angreifern möglich, an Paare (m, c) zu kommen, bei denen c die Chiffre zum Klartext m ist. Eine weitere Sicherheitsanforderung an Kryptosysteme ist also, dass das Wissen solcher Paare (m, c) nicht dazu führt, dass man den Schlüssel berechnen kann.

Bei der Signatur ist es ganz ähnlich wie bei der Authentifizierung. Wir setzen dazu ein asymmetrisches Kryptosystem voraus, bei dem neben $d_{k_p}(e_{s(k_p)}(m)) = m$ auch gilt

$$e_{s(k_p)}(d_{k_p}(c)) = c \quad \text{für alle } c \in C. \tag{5.3}$$

Unter dieser Voraussetzung läuft die Signatur folgendermaßen ab.

- Ben soll ein Dokument unterschreiben. Dieses Dokument wird in ein Element $c \in C$ umgewandelt. Das passiert auf eine transparente Art, die von vornherein bekannt ist.
- Bens Unterschrift unter dieses Dokument ist dann der Wert $u = d_{k_p}(c)$.
- Anna überprüft die Unterschrift, in dem sie mit Bens öffentlichem Schlüssel den Wert $e_{k_ö}(u) = e_{k_ö}(d_{k_p}(c))$ berechnet.
- Gilt $e_{k_ö}(u) = c$ so ist die Unterschrift gültig. Ist $e_{k_ö}(u) \neq c$ so ist die Unterschrift gefälscht.

Die Zuordnung Dokument $\longrightarrow C$ findet über sogenannte *Hash-Funktionen* statt. Das sind fest vorgegebene Funktionen, die man effektiv berechnen kann, die Dokumente in Zahlen von vorgegebener Größenordnung transferieren.
Da $(m^e)^d = (m^d)^e$, ist (5.3) ohne weitere Maßnahmen für das RSA-Kryptosystem erfüllt. Dieses eignet sich in der Theorie also besonders gut für Signaturen.

Bemerkung 5.4.2. Bei allen asymmetrischen Kryptosystemen brauchten wir große Primzahlen. Aber woher kommen die? Es gibt kein „großes Buch der Primzahlen“ in dem alle Primzahlen kleiner als 2^{1024} stehen. Was man daher tut, ist folgendes:

- Es gibt Listen mit ausgewählten Primzahlen p und Elementen großer Primzahlordnung $g \in (\mathbb{Z}/p\mathbb{Z})^*$. Diese sind öffentliches bekannt und werden in sehr vielen Protokollen genutzt.

- Man sucht sich selbst eine Primzahl in der gewünschten Größe. Das ist ein eigenes mathematisches Problem, dass wir hier nicht tiefergehend behandeln. Wir versuchen nur ganz grob den Ablauf darzustellen:

  Man wÃ¤hlt zufällig eine Zahl a in der gewünschten Größenordnung. Dann prüft man, ob a eine Primzahl ist. Wenn nicht, dann wählt man das nächste a. Der Primzahlsatz sagt, dass es etwa $\frac{x}{\ln(x)}$ Primzahlen $\leq x$ gibt. Zwischen 2^{1023} und 2^{1024} gibt es daher ungefähr $\frac{2^{1024}}{\ln(2^{1024})} - \frac{2^{1023}}{\ln(2^{1023})}$ Primzahlen. Das entspricht einem Anteil von etwa 0,14%. Wir können also erwarten, dass jede 700te zufällig gewählte Zahl eine Primzahl ist.

  Wie man eine Zahl auf die Eigenschaft „Primzahl“ überprüft, behandeln wir nicht in der Vorlesung.

5.5 Elliptische Kurven

Die Zahlen, die beim Diffie-Hellman-Schlüsseltausch und dem RSA-Verfahren aufgetaucht sind, sind wirklich riesig. Das liegt im wesentlichen daran, dass die Rechenoperationen auf $\mathbb{Z}$ – und somit in $\mathbb{Z}/p\mathbb{Z}$ so einfach sind. Diese einfachen Rechnungen sind zwar sehr gut für Anna und Ben, aber leider auch für Kim. Wenn nun Chipkarten in der Lage seien sollen Dinge zu verschlüsseln, wird das mit diesen großen Zahlen schwierig. Wir brauchen also ein System mit vergleichbarer Sicherheit, das mit deutlich weniger Speicherplatz auskommt.

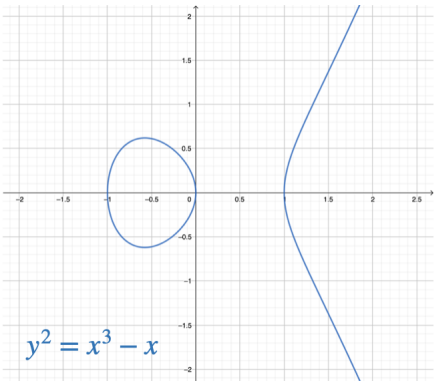
Bemerkung 5.5.1. Den diskreten Logarithmus hatten wir in Definition 5.2.20 für alle zyklischen endlichen Gruppen definiert. Wie später gesehen ist der Zusatz *zyklisch* unerheblich. Wir schränken uns einfach auf Untergruppen ein, die von einem Element erzeugt werden. Wir brauchen also nur eine endliche Gruppe mit einem Element großer Ordnung um einen diskreten Logarithmus definieren zu können. Wir können also versuchen Kryptographie mit anderen Gruppen zu betreiben.

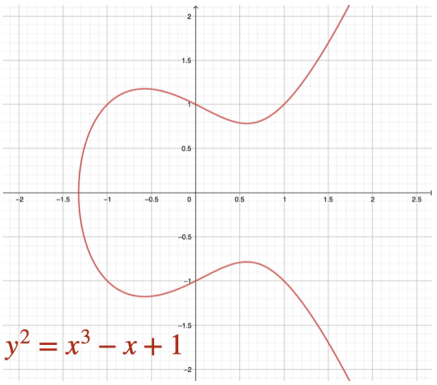
Beispiel 5.5.2. Wir betrachten die Gruppe $\mathbb{Z}/n\mathbb{Z}$ mit der Addition $+$. Dann entspricht g^k genau $[a] + \ldots + [a] = k[a] = [k] \cdot [a]$. Die Umkehrfunktion davon ist aber trivial, da wir einfach mit $[k]^{-1}$ multiplizieren können. Diese Gruppe eignet sich daher überhapt nicht. Das liegt unteranderem daran, dass $\mathbb{Z}/n\mathbb{Z}$ ein Ring ist. Die Multiplikation ist ein zusätzliches Hilfsmittel für Angreifer. Daher haben wir beim Diffie-Hellman Schlüsseltausch auch in der Gruppe $(\mathbb{Z}/p\mathbb{Z})^*$ gearbeitet, in der es keine weitere Verknüpfung neben der Multiplikation gibt.

Bevor wir Gruppen vorstellen, die sich ebenfalls für die Verschlüsselung eignen, brauchen wir ein bisschen Geometrie.

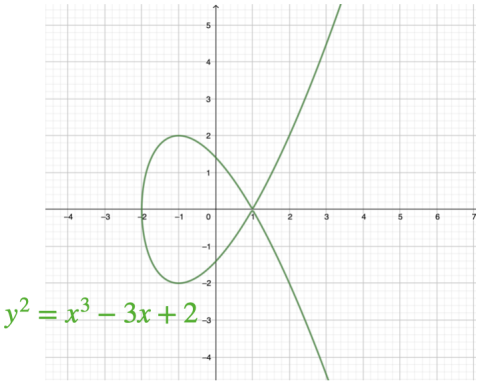
Definition 5.5.3. Eine ebene Kurve mit der Gleichung $y^2 = x^3 + ax + b$, mit $a, b \in \mathbb{R}$, heißt *elliptische Kurve über* $\mathbb{R}$, falls $x^3 + ax + b$ keine mehrfache Nullstelle besitzt. Das ist genau dann der Fall, wenn $4a^3 + 27b^2 \neq 0$ ist.

Die Kurven sind natürlich gegeben durch die Lösungen der Gleichung im $\mathbb{R}^2$. Typische Beispiele sehen so aus:





Da $x^3 - 3x + 2 = (x-1)^2(x+2)$ die doppelte Nullstelle 1 besitzt, ist $y^2 = x^3 - 3x + 2$ keine elliptische Kurve. Die doppelte Nullstelle sorgt dafür, dass sich die Kurve selbst schneidet. Das Bild der Kurve (die KEINE elliptische Kurve ist) ist



Bemerkung 5.5.4. Da die rechte Seite einer elliptischen Kurve $E : y^2 = x^3+ax+b$ ein Polynom vom Grad drei ist, gelten die folgenden Eigenschaften:

(1) Seien $P = (x_P, y_P)$ und $Q = (x_Q, y_Q)$ zwei Punkte auf E mit $x_P \neq x_Q$. Sei weiter g die Gerade durch die Punkte P und Q. Dann schneidet g die elliptische Kurve E in genau einem weiteren Punkt $R = (x_R, y_R)$.

(2) Sei $P = (x_P, y_P)$ ein Punkt auf E mit $x_P \neq 0$. Sei weiter g die Tangente von E an P. Dann schneidet g die elliptische Kurve E in genau einem weiteren Punkt $R = (x_R, y_R)$.

(3) Seien $P = (x_P, y_P)$ und $Q = (x_Q, y_Q)$ zwei verschiedene Punkte auf E mit $x_P = x_Q$. Dann ist $y_P = -y_Q$. Die Gerade g durch die Punkte P und Q verlüft parallel zur y-Achse. Wir sagen, dass diese Gerade g die elliptische Kurve E in einem *unendlichen Punkt* ∞ schneidet.

(4) Sei $P = (x_P, y_P)$ ein Punkt auf E. Verbinden wir P mit dem unendlichen Punkt ∞, zeichnen wir wieder eine vertikale Gerade durch P. Diese schneidet E genau im Punkt $(x_P, -y_P)$.

Alle diese Schnittpunkte sind mit Vielfachheiten versehen. Die Punkte R müssen also nicht zwangsläufig verschieden von P oder Q sein. Diese Bemerkung wird klarer, wenn wir weiter unten ein Beispiel berechnen.

Ab jetzt ist eine elliptische Kurve immer noch ausgestattet, mit einem unendlichen Punkt ∞.

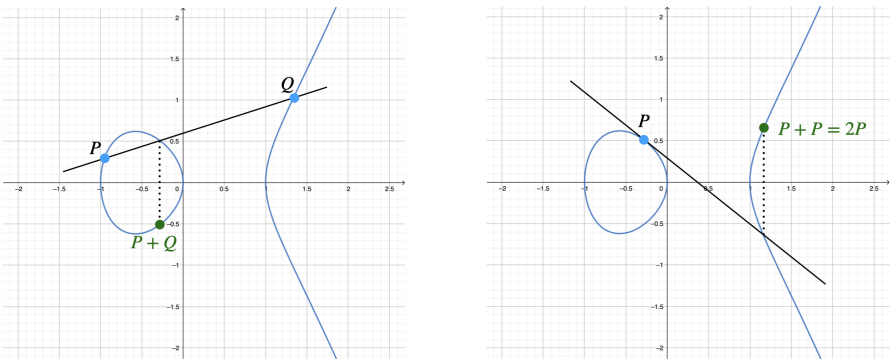
Bemerkung 5.5.5. Was wir in Bemerkung 5.5.4 gesehen haben, ist dass wir aus zwei gegebenen Punkten auf einer elliptischen Kurve (egal ob diese verschieden sind oder nicht) immer einen eindeutig bestimmten dritten

Punkt konstruieren können. Die vorgestellte Konstruktion erinnert also an eine Verknüpfung wie „+ “ oder „· “. Dabei soll $(x_P, -y_P)$ die Rolle des Inversen von $P = (x_P, y_P)$ spielen. Die Gerade durch diese beiden Punkte führt auf den Punkt ∞ (siehe Bemerkung 5.5.4(3)). Damit muss ∞ die Rolle des neutralen Elementes übernehmen.
Da die Verbindung von P und ∞ allerdings auf den Punkt $(x_P, -y_P)$ und nicht auf P führt, müssen wir die Verknüpfung noch mit der Spiegelung an der x-Achse versehen. Wir setzen also, mit R aus Bemerkung 5.5.4:

$$P + Q = (x_R, -y_R)$$

für alle Elemente P und Q der elliptischen Kurve E.

Eine Visualisierung der Addition von zwei verschiedenen Punkten, bzw. von einem Punkt mit sich selbst, sehen Sie hier



Satz 5.5.6. *Mit der oben definierten Verknüpfung wird jede elliptische Kurve über $\mathbb{R}$ zu einer kommutativen Gruppe mit neutralem Element ∞.*

Beweis. Die Eigenschaft des neutralen Elementes und die Existenz von Inversen haben wir bereits hergeleitet. Die Verknüpfung ist kommutativ, da die Gerade durch die Punkte P und Q das gleiche ist, wie die Gerade durch die Punkte Q und P. Die Assoziativität nachzuweisen ist aufwendig, aber nicht besonders schwierig. Das lassen wir in dieser Vorlesung weg. □

Beispiel 5.5.7. Sei die elliptische Kurve $E : y^2 = x^3 - 15x + 18$ mit den beiden Punkten $P = (7, 16)$ und $Q = (1, 2)$ gegeben. Beachten Sie, dass diese Punkte tatsächlich auf der elliptischen Kurve liegen.

(a) Wir berechnen $P + Q$. Die Gerade, die die Punkte P und Q verbindet ist

$$y = g(x) = \frac{16-2}{7-1}x - \frac{2}{6} = \frac{7}{3}x - \frac{1}{3}.$$

Da wir die Punkte suchen, die sowohl auf g als auch auf E liegen, setzen wir g in E ein und erhalten

$$\begin{aligned} &\left(\frac{7}{3}x - \frac{1}{3}\right)^2 = x^3 - 15x + 18 \\ \iff &\frac{49}{9}x^2 - \frac{14}{9}x + \frac{1}{9} = x^3 - 15x + 18 \\ \iff &0 = 9x^3 - 49x^2 - 121x + 161 \end{aligned} \tag{5.4}$$

Da P und Q auf g und E liegen, wissen wir, dass $x = 7$ und $x = 1$ diese Gleichung lösen. Damit ist $(x-7)(x-1) = x^2 - 8x + 7$ ein Teiler der rechten Seite von (5.4). Mit Polynomdivision erhalten wir

$$(9x^3 - 49x^2 - 121x + 161) \div (x^2 - 8x + 7) = 9x + 23.$$

Damit erfüllt auch $x_R = -\frac{23}{9}$ die Gleichung (5.4). Die zugehörige y-Koordinate erhalten wir durch einsetzen in g. Es ist $y_R = \frac{7}{3}x_P - \frac{1}{3} = -\frac{170}{27}$. Es folgt

$$P + Q = (x_R, -y_R) = (-\frac{23}{9}, \frac{170}{27}).$$

(b) Wir berechnen $Q+Q = 2Q$. Dazu benötigen wir zunächst eine Tangente an E. Wir fassen y als Funktion in x auf und erhalten durch ableiten

$$y' \cdot 2y = 3x^2 - 15.$$

Hier erhalten wir die linke Seite mit der Kettenregel, wenn wir y^2 ableiten. Es folgt, für $y \neq 0$, die Gleichung $y' = \frac{3x^2-15}{2y}$. Die Steigung von E am Punkt Q ist damit $\frac{3\cdot 1^2-15}{2\cdot 2} = -3$. Es folgt, dass die Tangente von E am Punkt $Q = (1, 2)$ gegeben ist durch

$$y = g(x) = -3x + 5.$$

Jetzt setzen wir, wie in Teil (a), einfach g in E ein und erhalten

$$\begin{aligned} &(-3x + 5)^2 = x^3 - 15 + +18 \\ \iff &9x^2 - 30x + 25 = x^3 - 15x + 18 \\ \iff &0 = x^3 - 9x^2 + 15x - 7 \end{aligned} \tag{5.5}$$

Wir wissen, dass $Q = (1, 2)$ auf g und E liegt. Da wir die Tangente von E in Q betrachten, liegt Q „doppelt“ auf E. Damit ist $x - 1$ eine doppelte Nullstelle von (5.5). Damit ist $(x-1)^2 = x^2 - 2x + 1$ ein Teiler der rechten Seite von (5.5). Mit Polynomdivision erhalten wir

$$(x^3 - 9x^2 + 15x - 7) \div (x^2 - 2x + 1) = x - 7.$$

Damit ist $x_R = 7$ die gesuchte dritte Lösung. Es folgt $y_R = g(x_R) = -3x_R + 5 = -16$ und somit

$$2Q = (x_R, -y_R) = (7, 16).$$

Es ist also $2Q = P$ und niemand hat das am Anfang bemerkt. Das ist sehr gut für uns! Es scheint also schwierig zu sein aus gegebenem P und Q herauszufinden ob $a \cdot P = Q$ ist. Wenn wir nun schon wissen, dass so ein a existiert, scheint es schwierig zu sein, dieses a zu bestimmen. Das ist genau das gleiche wie beim diskreten Logarithmus!

Führen wir genau die gleichen Schritte wie im Beispiel mit allgemeinen Parametern aus, dann erhalten wir Formeln für die Addintion von zwei Punkten auf einer elliptischen Kurve. Diese fassen wir im folgenden Theorem zusammen.

Theorem 5.5.8. *Sei $E : y^2 = x^3 + ax + b$ eine elliptische Kurve über $\mathbb{R}$. Seien weiter $P = (x_P, y_P)$ und $Q = (x_Q, y_Q)$ zwei Punkte auf E, mit $P \neq \infty \neq Q$. Ist $Q = -P$, dann ist $P + Q = \infty$. Ansonsten setzen wir*

$$\lambda = \begin{cases} (y_Q - y_P) \cdot (x_Q - x_P)^{-1} & \text{falls } P \neq Q \\ (3x_P^2 + a) \cdot (2y_P)^{-1} & \text{falls } P = Q. \end{cases}$$

Dann ist $P + Q = (\lambda^2 - x_P - x_Q, \lambda(2x_P + x_Q - \lambda^2) - y_P)$.

Da wir über den reellen Zahlen $\mathbb{R}$ arbeiten, besitzt jede elliptische Kurve überabzählbar unendlich viele Punkte. Für die Anwendung in der Kryptographie benötigen wir aber endliche Mengen. Betrachten wir aber die Formeln aus Theorem 5.5.8, stellen wir fest, dass diese in jedem Körper berechnet werden können. Wir ersetzen also einfach $\mathbb{R}$ durch $\mathbb{Z}/p\mathbb{Z}$ für eine Primzahl p und erhalten eine endliche Gruppe. Der Ursprung der Verknüpfung bleibt geometrischer Natur, aber die Formeln selbst sind rein algebraisch.

Definition 5.5.9. Sei p eine Primzahl. Eine *elliptische Kurve über* $\mathbb{Z}/p\mathbb{Z}$ ist gegeben durch die Lösungen einer Gleichung $E : y^2 = x^3 + ax + b$, mit $a, b \in \mathbb{Z}/p\mathbb{Z}$, so dass $x^3 + ax + b$ keine mehrfache Nullstelle besitzt, und einem unendlichen Punkt ∞. Wieder besitzt $x^3 + ax + b$ keine mehrfachen Nullstellen, wenn $4a^3 + 27b^2 \neq [0]_p$ ist. Die Menge aller Punkte auf E bezeichnen wir mit $E(\mathbb{Z}/p\mathbb{Z})$.

Theorem 5.5.10. *Sei p eine Primzahl und E eine elliptische Kurve über $\mathbb{Z}/p\mathbb{Z}$. Dann ist $E(\mathbb{Z}/p\mathbb{Z})$ eine endliche kommutative Gruppe bezüglich der Verknüpfung aus Theorem 5.5.8.*

Die Verknüpfung ist also etwas komplizierter als Addition und Multiplikation auf $\mathbb{Z}$. Das erschwert Kim die Arbeit, wenn sie Kryptosysteme knacken möchte. Die Verknüpfung ist noch einfach genug, dass es klare Formeln gibt mit denen wir zwei Punkte addieren können. Wir klären noch kurz wie groß die Gruppe $E(\mathbb{Z}/p\mathbb{Z})$ ist.

Satz 5.5.11. *Für jede Primzahl p und jede elliptische Kurve E über $\mathbb{Z}/p\mathbb{Z}$ gilt*

$$p + 1 - \sqrt{p} \leq |E(\mathbb{Z}/p\mathbb{Z})| \leq p + 1 + \sqrt{p}.$$

Kurz und unpräzise: $|E(\mathbb{Z}/p\mathbb{Z})| = O(p)$.

Wir kommen nun direkt zum Diffie-Hellman-Schlüsseltausch mit elliptischen Kurven.

EC Diffie-Hellman Schlüsseltausch 5.5.12. Wie immer wollen Anna und Ben geheim kommunizieren.

- Anna und Ben einigen sich auf eine Primzahl p, eine elliptische Kurve E über $\mathbb{Z}/p\mathbb{Z}$ und ein $P \in E(\mathbb{Z}/p\mathbb{Z})$, mit großer Ordnungt. Diese Daten werden ganz öffentlich kommuniziert; z.B. stellt Anna diese Informationen auf ihrer Homepage bereit.

- Nun wählt Anna einen geheimen Schlüssel $a \in \mathbb{N}$ und Ben einen geheimen Schlüssel $b \in \mathbb{N}$. Beide Schlüssel werden niemandem verraten und sind somit nur Anna (a) und Ben (b) selbst bekannt.

- Nun berechnet Anna den Punkt $A = aP$ und Ben berechnet $B = bP$. Anna sendet A an Ben und Ben sendet B an Anna.

- Nun berechnet Anna aB und Ben berechnet bA. Es ist

$$aB = a(bP) = (ab)P = b(aP) = bA.$$

- Der gemeinsame Schlüssel von Anna und Ben ist nun $(ab)P$ und beiden Parteien bekannt.

Bemerkung 5.5.13. Anna und Ben müssen nur Werte der Form aP mit bekannten Werten $a \in \mathbb{N}$ und $P \in E\left(\mathbb{Z}/p\mathbb{Z}\right)$. Das können wir mit der gleichen Idee wie beim schnellen Exponentieren machen: Erst berechnet man die Werte $2P$, $4P$, $8P$,... und setzt daraus den Wert aP zusammen. Ganz genau wie beim schnellen Exponentieren sind dafür $O(\ln\left(\left|E\left(\mathbb{Z}/p\mathbb{Z}\right)\right|\right) = O(\ln(p))$ Rechenoperationen in $E(\mathbb{Z}/p\mathbb{Z})$ nötig. Alle nötigen Werte sind also schnell berechnet.

Bemerkung 5.5.14. Angenommen Kim hätte alles aus dem EC Diffie-Hellman-Schlüsseltausch mitgehört, dann kennt sie p, E, P, aP und bP. Die einzige *bekannte* Methode um daraus abP zu berechnen besteht darin einen der Werte a oder b zu berechnen. Sie steht also vor der Aufgabe

Bestimme aus den Werten P und aP die Zahl a.

Das ist die elliptische Kurven Version des diskreten Logarithmus! Da wir bei elliptischen Kurven keinen Ring im Hintergrund haben, gibt es nichts vergleichbares, wie eindeutige Primfaktorzerlegung. Damit gibt es keinen Index-Calculus-Algorithmus und der schnellste bekannte Algorithmus zum lösen dieses Problems benötigt $O(\sqrt{p})$ Rechenschritte. Das erledigt zum Beipspiel der Babystep-Giantstep Algorithmus, der in jeder Gruppe funktioniert.
Es wird empfohlen, dass beim EC Diffie-Hellman-Schlüsseltausch mindestens 160-Bit Primzahlen p benutzt werden. Das ist eine viel kleinere Zahl als die 1023-Bit Primzahlen aus dem klassischen Diffie-Hellman-Schlüsseltausch!

Bemerkung 5.5.15. Ähnlich wie bei den Faktoriesierungsaufgaben, wurden 1997 auch Challenges für die elliptische Kurven Version des diskreten Logarithmus veröffentlicht. Es wurden also elliptische Kurven $E : y^2 = x^3 + [a]_p x + [b]_p$ und zwei Punkte $P = ([x_P]_p, [y_P]_p)$, $Q = ([x_Q]_p, [y_Q]_p)$ auf E veröffentlicht. Die Aufgabe besteht nun darin ein $k \in \mathbb{N}$ zu finden, mit

$Q = kP$. Eine dieser Aufgaben, die immer noch ungelöst ist, ist folgende:

$$\begin{aligned}
p &= 155003179783434785924857641481313994241 1 \\
a &= 139926757376357881587790523597115331671 0 \\
b &= 100929654219153246407626036752581629397 6 \\
x_P &= 131795376323959588846552414558987269569 0 \\
y_P &= 434829348619031278460656303481105428081 \\
x_Q &= 124739221131790715130324772148964069924 0 \\
y_Q &= 207534858442090452193999571026315995117
\end{aligned}$$

Im Vergleich zu den Primzahlen, die man für das klassische DLP und für das RSA-Kryptosystem benötigt, sind diese Zahlen winzig!

Da wir durch Nutzung von elliptischen Kurven, viel weniger Speicherplatz und viel weniger Rechenoperationen benötigen, kann die Kryptographie auch auf Chipkarten, die nur extrem geringe Rechenkapazitäten besitzen, umgesetzt werden. Ihr Personalausweis etwa, ist digital mit Hilfe elliptischer Kurven unterschrieben. Wie das funktioniert besprechen wir jetzt.
Vorher erinnern wir uns noch einmal daran, dass wir für Signaturen so genannte *Hash-Funktionen* brauchen. Dies sind fest vorgegebene Funktionen H, die einem Dokument D einen Wert $H(D)$ zuordnen, mit dem man dann weiterarbeiten kann. Die Elliptische Kurven Signatur (kurz: EC-Signatur) benutzt als zugrundeliegendes Kryptosystem das Elgamal-Verfahren.

EC Signatur 5.5.16. Damit uns nicht langweilig wird, soll diesmal Anna ein Dokument D unterzeichnen. Dazu wird öffentlich eine Primzahl p, eine elliptische Kurve $E(\mathbb{Z}/p\mathbb{Z})$ und ein Element $P \in E(\mathbb{Z}/p\mathbb{Z})$ mit (großer) Ordnung q bereitgestellt. Hier ist q wieder eine Primzahl. Die gewählte Hash-Funktion H liefert nun $H(D) = d \in \{1, \ldots, q-1\}$.

- Anna erzeugt ihren *öffentlichen Schlüssel* $A = aP$, wobei $a \in \{2, \ldots, q-1\}$ ihr *privater Schlüssel* ist.
- Anna veröffentlicht A. Es ist wichtig, dass dieser Schlüssel allen bekannt ist und Anna zugeordnet werden kann.
- Nun wählt Anna noch ein zufällig gewähltes $b \in \{2, \ldots, q-1\}$ und berechnet $B = bP = ([x_B]_p, [y_B]_p)$.

- Anna berechnet ein $[c]_q = [d + ax_B]_q \cdot [b]_q^{-1} \in \mathbb{Z}/q\mathbb{Z}$. Falls $[c]_q = [0]_q$ ist, dann muss Anna ein anderes b wählen.
- Anna unterschreibt das Dokument D mit (x_B, c).
- Ben Berechnet $[v_1]_q = [d]_q \cdot [c]_q^{-1}$ und $[v_2]_q = [x_B]_q \cdot [c]_q^{-1}$.
- Jetzt berechnet Ben den Punkt $Q = v_1 P + v_2 A \in E(\mathbb{Z}/p\mathbb{Z})$.
- Die Kongruenz $x_Q \equiv x_B \mod p$ bestätigt die Unterschrift. Hierbei ist $Q = ([x_Q]_p, [y_Q]_p)$.

Für die Erstellung der Unterschrift benutzt Anna den privaten Schlüssel a. Überprüft wird die Unterschrift mit Ihrem öffentlichen Schlüssel A.

Um einzusehen, dass dieses Verfahren funktioniert, überlegen wir uns zunächst, warum hier soviel modulo q gerechnet wird obwohl wir eine elliptische Kurve über $\mathbb{Z}/p\mathbb{Z}$ betrachten. Es ist q die Ordnung von P. Damit ist $q \cdot P = \infty$ (also gleich dem neutralen Element). Für beliebiges $k, r \in \mathbb{N}_0$ ist somit $(kq + r)P = k(qP) + rP = \infty + rP = rP$. Damit hängt der Wert von nP nur von der Restklasse $[n]_q$ ab. Wir können also auch problemlos $[n]_q \cdot P$ definieren.

Bemerkung 5.5.17. Jetzt zeigen wir schnell, dass diese Signatur funktioniert. Es ist

$$\begin{aligned} Q = v_1 P + v_2 A = v_1 P + v_2 aP = (v_1 + v_2 a)P \\ = \left([d]_q \cdot [c]_q^{-1} + [ax_B]_q \cdot [c]_q^{-1} \cdot\right) P \\ = [d + x_B a]_q \cdot [c]_q^{-1} P = [b]_q P = B. \end{aligned}$$

Genau das wird durch $x_B \equiv x_Q \mod p$ bestätigt.

Literaturverzeichnis

[1] S. Bauer; *Mathematisches Modellieren* als fachlicher Hintergrund für die Sekundarstufe I+II. Springer Spektrum, 2021

[2] C. Eck, H. Garke, P. Knabner; *Mathematische Modellierung* (3. Auflage). Springer Spektrum, 2017

[3] J. Hoffstein, J. Pipher, J.H. Silverman; *An Introduction to Mathematical Cryptography.* Springer, 2008

[4] C.P. Ortlieb, C. v. Dresky, I. Gasser, S. Günzel; *Mathematische Modellierung* Eine Einführung in zwölf Fallstudien (2. Auflage). Springer Spektrum, 2013